

Predictive Capability of the Partial Least Squares Structural Equation Modeling (PLS-SEM) with Application

Fatma Abdelkhalek¹ and Marianna Bolla²

- ¹ Institute of Mathematics, Budapest University of Technology and Economics, Hungary.
(E-mail: fatma@math.bme.hu)
- ² Institute of Mathematics, Budapest University of Technology and Economics, Hungary.
(E-mail: marib@math.bme.hu)

Abstract. Partial Least Squares Structural Equation Modeling (PLS-SEM) is formulated for latent random variables, which linearly depend on observed random variables. Herman Wold developed a three stage iterative algorithm to estimate the PLS-SEM model parameters. These estimates are considered to have predictive potential. The PLS-SEM estimation and evaluation are distribution free. PLS-SEM lacks of the global standard fit function. Therefore, different criteria are considered to assess the estimated model parameters and its predictive relevance. In this paper, we illustrate the PLS-SEM framework. We use the Survey of Young People in Egypt data to estimate and predict the direct effect of work characteristics, work satisfaction, and institutional corruption on workers' anxiety. The results show the extent to which each exogenous factor affects workers' anxiety and suggest a good predictive capability.

Keywords: PLS-SEM Estimation, Prediction, Evaluation, Holdout Sample.

1 Introduction

Partial Least Squares Structural Equation Modeling (PLS-SEM) is formulated for two kinds of latent random variables (LVs), endogenous and exogenous which linearly depend on dependent and independent observed random variables, respectively. PLS-SEM is a method of structural equation modeling which allows to constructing and examining the casual relationships between the LVs in a single statistical model. By definition, LVs are hidden variables that are not measured directly from the data but are composed from one or more observed variables. This concept was introduced by the psychologist Spearman, who studied how to measure what he called "human intelligence" in 1904 [1]. Generally, SEM is seen as a combination of the factor analysis of Spearman and path analysis of Wright [2]. However, in SEM the LVs are organic part of the model, but they are introduced and given a meaning based on the factor loadings in the traditional factor analysis.

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



Figure 1 represents a simple PLS-SEM abstract model in which there are 9 observed variables (6 independent: x 's and 3 dependent: y 's) that correspond to 3 LVs (2 exogenous: ξ 's and 1 endogenous: η). The PLS-SEM model consists of two components (submodels), the measurement (outer) and structural (inner) models. Both models can be transformed into a Boolean design matrix or what is called an adjacency matrix, where the rows and columns are labeled with the graph vertices. The entries of this matrix are set to 1, if the two vertices are connected; and 0, otherwise [3]. In the measurement model, each group of the observed variables (called block) is associated with exactly one LV. There are two ways to construct the measurement model. The first way, called mode A, is reflective. In this case, the manifest indicators are influenced by the underlying LV. Moreover, the outer relations are called loadings (λ_i 's) and obtained from a simple regression model between each indicator of the block and the corresponding LV separately. Hence, a change in the LV implies a change in each indicator of the corresponding block (see η_1 in Figure 1). The second way, called mode B, is formative. In mode B the manifest variables are viewed as causes rather than effects and the outer weights (w_i 's) are obtained from a multiple regression model between the LV and all indicators of the corresponding block. Here, contrary to mode A, a change in the LV does not necessarily lead to a change in all the indicators of the corresponding block (consider ξ_1 and ξ_2 in Figure 1). For more details about constructing the outer model modes, see [4]. The inner model examines the causality structures between the exogenous and endogenous LVs, which are represented by what is called path coefficients (p_j 's).

PLS-SEM exhibits three linear equations. Let the data matrix consist of the p -dimensional dependent \mathbf{Y} and q -dimensional independent \mathbf{X} observed variables. Accordingly, the PLS-SEM model equations are:

1. The measurement model equations,

$$\mathbf{X} = \mathbf{C}\xi + \varepsilon; \quad \mathbf{Y} = \mathbf{F}\eta + \delta, \quad (1)$$

where ξ is an n - and η is an m - dimensional latent variable; ε and δ are the measurement error vectors in \mathbf{X} and \mathbf{Y} , respectively; \mathbf{C} and \mathbf{F} are $q \times n$ and $p \times m$ loadings/weights matrices that link the observed variables with the latent variables. Typically, $n \leq q$ and $m \leq p$.

2. The structural model equation,

$$\mathbf{B}\eta = \mathbf{A}\xi + \zeta, \quad (2)$$

where \mathbf{B} and \mathbf{A} are $m \times m$ and $m \times n$ path coefficient matrices from endogenous to endogenous and from exogenous to endogenous latent variables, respectively; ζ is a random vector of residuals which is uncorrelated with the exogenous LVs and with the measurement errors. Moreover, the matrix \mathbf{B} is non-singular and upper triangular in the recursive models with 1's along its main diagonal.

PLS-SEM is a distribution-free approach and applicable in the case of small sample sizes [5,6]. Its primary goal is explanatory. It focuses on estimating the

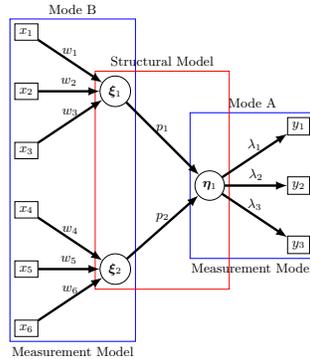


Fig. 1. PLS-SEM Model Representation Example

latent variable scores that maximize the explained variance of the endogenous variables. Herman Wold invented a sequential three stage algorithm to estimate the PLS-SEM model in which the LV scores and the model parameters are estimated [3,5]. These estimates have predictive potential as well [6]. Wold was the first to introduce the predictive capability of the PLS-SEM - which he called “*causal-predictive*” [7]. Thus, PLS-SEM fits both the explanatory and the predictive modeling objectives. There are some differences between these two objectives in terms of the data level. The explanatory modeling aims at estimating and evaluating the relationships between the LVs so as to generalize the estimated results to the population level. The predictive modeling, on the other hand, focuses on predicting the results of the test data using the estimates of the training sample. Fundamentally, PLS-SEM relaxes the maximum likelihood estimation assumptions that are usually applicable to Gaussian data. Therefore, it is considered as a “*soft modeling*” method [5]. Moreover, it is appropriate for estimating complex causal models in both theoretical and empirical situations. Accordingly, researchers named it with a “*silver bullet*” technique [6]. In the last decades, due to the good properties of the PLS-SEM, there was a massive increase in using it in many disciplines, such as marketing, accounting, business, management information systems, operations management, strategic management, tourism, and bioinformatics [8]. However, a few of them aims at exploring the PLS-SEM predictive techniques [9].

The paper is organized as follows. In Section 2, we discuss the PLS-SEM model estimation, evaluation and prediction techniques. In section 3, we apply the three stage algorithm of Wold to a real life data in which we examine the effect of the following factors: 1) work characteristics; 2) work satisfaction; and 3) institutional corruption on workers’ anxiety, using the 2014 Survey of Young People in Egypt (SYPE 2014) [10]. The aim is to: 1) estimate and evaluate the effect of these factors on the level of anxiety using a training sample; 2) use the estimated model to predict the anxiety levels for new observations (holdout sample).

2 PLS-SEM Estimation, Evaluation and Predictive Relevance Criteria

PLS-SEM model estimation and evaluation processes do not require any distributional assumptions. Wold invented a three sequential stage algorithm to estimate the model parameters. In the following, we briefly discuss this algorithm according to [3, 5].

- Stage I is a fixed point iterative convergence process in which the latent variable scores are estimated. It consists of the following steps:
 - i. Initializing the LV score for each case as the weighted sum of its corresponding indicators according to the measurement model adjacency matrix.
 - ii. Updating the so obtained scores with the LVs inner weights, which are computed for each LV to indicate the association strength to its adjacent LVs. It is calculated by either centroid, path or factorial schemes.
 - iii. Updating the LV scores with the outer relation estimates (weights/loadings) which depend on the mode of the measurement model.

The first stage of the algorithm iterates between these steps until convergence. The convergence occurs when the maximum difference between the obtained LV scores at steps t and $t+1$ are less than the tolerance (usually set to a very small e.g., 10^{-5}).

- Stage II uses the final LV scores obtained at the convergence iteration of Stage I to estimate the structural model parameters (path coefficients) and the measurement model parameters (loadings/weights) by OLS regressions. Note that the measurement model estimates that are obtained here are actually equivalent to those (loadings/weights) that are obtained at the convergence iteration of Stage I. Unlike Stage I, this stage is not iterative.
- Stage III estimates the location parameters for the observed variables and the LVs.

For more details about PLS-SEM estimation algorithm, see [5]. PLS-SEM does not have a global standard model fit function as the traditional family of SEM that is known as covariance based-SEM. Therefore, different statistical criteria are used to assess the estimated model parameters such as, communality; redundancy; R-squared (R^2); and the goodness of fit (GoF). In addition to the bootstrapping approach to obtain the significance of the estimated model parameters [4, 6, 11].

Communality assesses the quality of each block of the measurement model that is corresponding to a specific latent variable. It is equivalent to the average variance extracted (AVE) that is the grand mean value of the squared loadings. It is computed as the average of the squared correlation coefficient between each indicator of the measurement block (x_{kj}) and the corresponding LV (l_j),

$$Com_j = \frac{1}{p_j} \sum_{k=1}^{p_j} [cor(x_{kj}, l_j)]^2,$$

where p_j is the number of indicators in the j th LV block, and x_{kj} is the specific k indicator of that block. Furthermore, the communality of the overall model is the mean communality that is obtained by:

$$\overline{Com} = \frac{1}{p} \sum_{j=1}^J p_j \times Com_j.$$

Redundancy also evaluates the outer model, where each formatively measured latent variable is correlated with an alternative reflective measurement of the same construct. It gathers data on the alternative measures at the same time as the former latent variable. It is considered for each endogenous LV, where it measures the amount of variance in the indicators measuring the variable that is explained by the exogenous latent variables that predict the endogenous variable (l_j). It is calculated as follows,

$$Red_j = Com_j \times R^2(l_j, l_j^{pred}).$$

If there are more than one endogenous LV in the model, then one can also calculate the average redundancy for all of the endogenous LVs.

R^2 measures the proportion of the explained variance in the endogenous latent variable by the associated exogenous latent variables.

The *GoF* evaluates the quality of the hypothesized model. It considers the validation of the outer and inner relations and computed as the geometric mean of average communality and average R^2 ,

$$GoF = \sqrt{\overline{Com} \times \overline{R^2}}.$$

Finally, evaluating the significance of the model parameters (path coefficients and outer weights). To do so, the bootstrapping technique is used since the PLS-SEM does not assume normal distribution. This is a resampling approach in which a large number of subsamples is drawn from the original data, then for each subsample the model parameters are estimated. This enables researchers to compute the standard error for each model parameter which is required to determine the significance of each parameter.

After assessing the estimated model, we move to validate the model predictive relevance. Different measures are used to do so: 1) Stone-Geisser Q^2 measure that depends on the blindfolding procedures; 2) effect size f^2 [3, 5, 12, 13]; 3) Root mean square error (*RMSE*), where the prediction errors are calculated from the holdout sample; and 4) Bayesian Information Criterion (BIC) to select the model with low prediction error. These measures are effective especially in the case of small sample size. However, in the case where there is a sufficient data to be divided into two parts: training and testing samples, it is possible to utilize the step by step holdout sample cross validation assessment. The last approach avoids the “overfitting” problem that may happen if the predictive performance is measured on the same sample that is used to estimate the model [14].

The Stone-Geisser Q^2 measure is known as a “blindfolding” technique. It depends on the reuse of the model sample by omitting every i th observation

from the data, where i is the omitted distance, usually between 5 to 10. For example, if the omitted distance i is defined as 6, this means that every 6th observation is omitted from the data. The remaining data after omission is used to estimate the model parameters. Then, the estimated model is reused to predict the omitted observations [15]. The formula of the Q^2 measure is

$$Q^2 = 1 - \frac{\sum_D E_D}{\sum_D O_D},$$

where D is the omission distance in blindfolding, E is the sum of squares of the prediction error, and O is the sum of squares of the observations. Note that the blindfolding procedure is only applied to the endogenous LVs and a value greater than zero indicates a good model predictive capability. Practically speaking, values of .02, .25, and .35 indicate that the exogenous LVs has a weak, medium or strong capability, respectively, of predicting the associated endogenous LV.

The effect size criterion f^2 assesses the change in the R^2 measure. It measures the relative impact of the inclusion of a specific exogenous LV to the predictive performance of estimating the endogenous LV in the following way,

$$f^2 = \frac{R_{included}^2 - R_{excluded}^2}{1 - R_{included}^2}.$$

In the same manner as f^2 is obtained, the relative impact of the structural model on the endogenous LVs can be calculated with respect to the Q^2 measure as follows,

$$q^2 = \frac{Q_{included}^2 - Q_{excluded}^2}{1 - Q_{included}^2}.$$

The benchmark of both measures f^2 and q^2 follows the same criteria of the blindfolding Q^2 measure, i.e. the greater the value, the better the model predictive performance is. Unlike Q^2 , software packages such as SmartPLS and R do not automatically compute f^2 and q^2 measures. Researchers calculate them gradually, a certain exogenous LV is determined to be deleted from the model and run the model then apply the formula by comparing the obtained results of the complete and the reduced models. In some cases, deleting a LV may cause an issue where the model falls apart and the effect size cannot be computed.

The following is an explicit explanation for the step by step holdout sample method to assess the predictive model performance according to [16].

1. Randomly divide the data into two parts, a training sample (around 80% of the total sample size) and the test sample (the remaining cases/holdout). Prior to the analysis, scale the data to have a zero mean and unit variance.
2. Use the training sample to estimate the hypothesized model in which the PLS-SEM parameters are obtained.
3. Use the estimated measurement model parameters to calculate the endogenous and exogenous LV scores for the holdout test sample.
4. Standardize the obtained LV scores of the holdout sample.

5. Calculate the predicted scores for the endogenous latent variables using the estimated path coefficient.
6. Calculate the coefficient of determination for the holdout sample $R_{holdout}^2$ which is the square of the multiple correlation coefficient between the endogenous standardized scores of step 4 and the predicted ones of step 5.
7. Compare the coefficients of determination for the training sample $R_{training}^2$ and the one of the holdout test sample $R_{holdout}^2$.

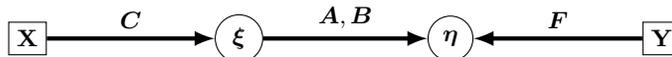


Fig. 2. Holdout Sample Technique Estimation

In sum, Figure 2 visualizes the estimation process in the holdout sample technique. Recall that R^2 measures the proportion of variance of the endogenous latent variable explained by its exogenous ones. In the holdout sample, if the value of the $R_{training}^2$ and the $R_{holdout}^2$ are close to each other. This indicates that the estimated model can be used to predict the future results.

3 Application

Based on the literature, we build the hypothesized model that examines to what extent: 1) work characteristics; 2) work satisfaction; and 3) institutional corruption affect workers' anxiety. Table 1 shows the labels of the observed indicators and the associated latent variables. There are 11 independent and 4 dependent observed variables correspond to 3 exogenous and 1 endogenous LVs, respectively. Note that, all the observed indicators are ordinal scaled, from "1" lowest to "5" highest, and each observed indicator is linked to exactly one latent variable. Figure 3 visualizes the hypothesized model. As shown in Figure 3, mode B is utilized in order to construct the three exogenous latent variables (Wrk, Satis, and Corrup), while mode A is used to construct the endogenous latent variable (Anx).

A sample of 1222 wage workers from the 2014 SYPE data [10] is specified to examine the hypothesized model. The model hypotheses are stated as follows:

1. There is a direct effect of work characteristics on workers' anxiety;
2. There is an inverse effect of work satisfaction on workers' anxiety; and
3. There is a direct effect of institutional corruption on workers' anxiety.

In order to answer the research questions, we used Wold algorithm to estimate and evaluate the PLS-SEM model parameters. Then, we used the estimated model to validate the predictive model relevance. For the last purpose, we followed the step by step holdout sample technique as discussed in Sec. 2. Accordingly, the analysis is processed into two phases. Phase I aims at estimating and evaluating the model, while phase II measures the predictive performance of the estimated model.

Latent Variables	Observed Indicators	Model Label
<i>Exogenous</i>	<i>Independent</i>	
Work		Wrk
Characteristics	Experiencing maltreatment from co-workers	x_1
	Experiencing long working hours	x_2
	Little pay	x_3
	Exhausting work load	x_4
Work		Satis
Satisfaction	The way people treat each other	x_5
	Feeling about belonging to a work community	x_6
	Involvement in the work community	x_7
	Receiving social support	x_8
	Communication at work	x_9
Institutional		Corrup
Corruption	Public institutions corruption	x_{10}
	Getting jobs via backdoor not based on the skills	x_{11}
<i>Endogenous</i>	<i>Dependent</i>	
Workers' Anxiety		Anx
	Worry about yourself or your family	y_1
	Worry about losing your future	y_2
	Worry about losing your home	y_3
	Worry about losing your job	y_4

Table 1. Labels and Abbreviations of The Observed Measurements and Its Corresponding Latent Variables.

3.1 Phase I: Model Estimation and Evaluation

Prior to the analysis, we standardized the total sample and divided it randomly into a training sample of size 977 and a test sample of the remaining cases. The algorithm of Wold is applied on the training sample in order to estimate the hypothesized model. The R statistical package (semPLS) is used to run the analysis. Figure 3 shows the estimated model parameters. In accordance with the PLS-SEM model equations (See Eq. 1 and Eq. 2), the measurement model weights and loadings estimates are summarized into the matrices \mathbf{C} and \mathbf{F} , respectively as follows,

$$\mathbf{C} = \begin{bmatrix} .34 & 0 & 0 \\ .15 & 0 & 0 \\ .31 & 0 & 0 \\ .51 & 0 & 0 \\ 0 & .27 & 0 \\ 0 & .46 & 0 \\ 0 & .58 & 0 \\ 0 & .55 & 0 \\ 0 & .30 & 0 \\ 0 & 0 & .29 \\ 0 & 0 & .95 \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} .90 \\ .89 \\ .86 \\ .77 \end{bmatrix},$$

where \mathbf{C} contains the outer weights between the independent observed indicators (x_i 's) and the exogenous LVs (Wrk, Satis, and Corrup); and \mathbf{F} includes the outer loadings between the dependent observed indicators (y_i 's) and the endogenous LV (Anx). The estimated structural model matrices \mathbf{B} and \mathbf{A} are the following,

$$\mathbf{B} = [1], \quad \mathbf{A} = [.17 \ - .21 \ .24],$$

where \mathbf{A} contains the estimated path coefficients from the exogenous LVs (Wrk, Satis, and Corrup) to endogenous LV (Anx); and \mathbf{B} which maintain the estimated path coefficients from endogenous to endogenous LVs. In this case, \mathbf{B} is the unit matrix of size 1 since the hypothesized model contains only one endogenous LV.

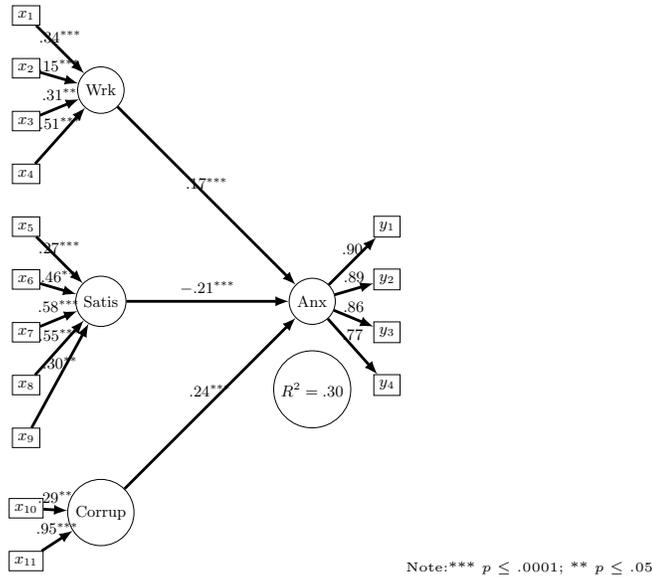


Fig. 3. The Estimated Hypothesized Model

To validate the model estimates, Table 2 presents the evaluation criteria of estimated model. The results indicate a well constructed measurement model blocks according to the communality, redundancy and AVE measures. Moreover, the model explains about 30% of the variation in workers' anxiety levels ($R^2 = .30$) and it has an acceptable overall fit with a GoF of .35. Table 3 shows the correlation coefficients between the LVs. It shows that the exogenous LVs are not highly correlated with each other, while they are sufficiently correlated with the endogenous LV. Consistence with our assumptions, the work satisfaction has an inverse relationship with the worker's anxiety. Moreover, the bootstrapping technique is considered to examine the significance of the estimated path coefficients, with subsamples of size 5000. The results indicate little differences between the path coefficients from the original sample and the estimated ones from the subsamples. Based on the standard errors of the

estimated parameters from all subsamples, the significance of each estimator is obtained. All the estimated path coefficients and the weights of the outer model are statistically significant, see Figure 3. More specifically, the model confirmed a positive direct effect of both work characteristics and institutional corruption on workers' anxiety ($\beta = 0.17, 0.24$, respectively). At the same time, there is an inverse effect of work satisfaction on workers' anxiety ($\beta = -0.21$). That is, as people feel more satisfied with their work, their anxiety levels decrease. Also, as the work characteristics and corruption get worse, the workers' anxiety increases.

	Type	Communality	Redundancy	AVE
Work Characteristics	Exogenous	0.493	0.000	0.000
Work Satisfaction	Exogenous	0.408	0.000	0.000
Institutional Corruption	Exogenous	0.505	0.000	0.000
Workers' Anxiety	Endogenous	0.732	0.122	0.732
R^2		.30		
GoF		.35		

Table 2. Estimated Model Measures Summary

	ξ_1	ξ_2	ξ_3	η_1
ξ_1 : Work Characteristics	1			
ξ_2 : Work Satisfaction	-.131	1		
ξ_3 : Institutional Corruption	.117	-.162	1	
η_1 : Workers' Anxiety	.426	-.470	.395	1

Table 3. Correlations Between Latent Variable Scores

3.2 Phase II: Predictive Model Relevance

Once the estimated model is validated, we measure the predictive performance of it. The semPLS package of R software gives the value of Stone-Geisser Q^2 measure according to the specified omission distance. In our case, the omitted distance is set to 7. The results reported a value of Q^2 measure equal to .26, which indicates that the estimated model reveals a medium ability to forecast the results of the omitted observations.

In addition, we compute the step by step cross validation by means of the holdout sample approach in the following way,

1. For the 245 cases of the holdout sample, we computed the scores of the exogenous (Wrk, Satis, and Corrup) and endogenous (Anx) latent variables by using the estimated measurement matrices C and F , respectively. Then we standardized the obtained LV scores.

2. For the endogenous LV (Anx), we computed its predicted scores by using the estimated path coefficient matrices \mathbf{A} and \mathbf{B} .
3. At last step, we calculated the correlation coefficient between the the estimated scores of the LV (Anx) that we obtained from Step 1 and Step 2 which was .58. Hence, the coefficient of determination for the holdout sample was $R_{holdout}^2 = .33$.

Next, we turn to compare the coefficients of determination of both samples. The results show that the $R_{holdout}^2 = .33$ and the $R_{training}^2 = .30$ are close to each other. In sum, the two predictive relevance criteria that we investigated indicate that the estimated model is appropriate for predicting future results. To illustrate, for new cases, we don't need to re-estimate the model, but use the estimated model parameters. As summarized in Figure 2, when the data for the independent observed indicators are available, the weights matrix \mathbf{C} is used to calculate the scores of the exogenous LVs (Wrk, Satis and Corrup), then based on the estimated path coefficients matrix \mathbf{A} and \mathbf{B} we predict the level of workers' anxiety.

4 Conclusion

This study gives a brief overview of the PLS-SEM model. In the context of Wold [5, 7] and Hair [4, 6], our research summarizes how to: 1) estimate and evaluate a hypothesized model; 2) assess the model's predictive performance by different criteria; 3) use the estimated model to predict future results by using the holdout sample approach. In addition, it applies the PLS-SEM to a real life data in which the direct effect of work characteristics, work satisfaction, and institutional corruption on workers' anxiety levels is examined. The results specified the extent to which each of the underlined factor affects workers' anxiety. Moreover, the estimated model shows a good predictive capability for future results.

In multivariate statistical analysis, there are many similarities and linkages between various approaches. For instance, in the Gaussian case, the parameterization of the inverse of the model covariance matrix Σ^{-1} can be obtained by the block Cholesky matrix decomposition $\mathbf{LD}^{-1}\mathbf{L}^T$ of Kiiveri. In PLS-SEM, on ther other hand, there are no distributional or other restrict assumptions are required. Accordingly, the soft algorithm of Wold is used to estimate the model parameters. However, we notice some relations between the Wold algorithm and the matrix decomposition of Kiiveri. For further research, we aim at developing a free flexible algorithm that combines these approaches to estimate the PLS-SEM model parameters.

Acknowledgments

The research reported in this paper was supported by the Higher Education Excellence Program of the Ministry of Human Capacities in the frame of Artificial Intelligence research area of Budapest University of Technology (BME

FIKP-MI/SC). The research was also supported by the National Research, Development and Innovation Fund (TUDFO/51757/2019-ITM, Thematic Excellence Program).

References

1. Spearman CE. General intelligence, objectively determined and measured. *American Journal of Psychology*, 15:201–293, 1904.
2. Kenneth A Bollen. Structural equations with latent variables. 1989.
3. Jan-Bernd Lohmöller. *Latent variable path modeling with partial least squares*. Springer Science & Business Media, 2013.
4. Joseph F Hair Jr, G Tomas M Hult, Christian Ringle, and Marko Sarstedt. *A primer on partial least squares structural equation modeling (PLS-SEM)*. Sage publications, 2016.
5. Herman Wold. Partial least squares. s. kotz and NL johnson (eds.), encyclopedia of statistical sciences (vol. 6), 1985.
6. Joe F Hair, Christian M Ringle, and Marko Sarstedt. PLS-SEM: Indeed a silver bullet. *Journal of Marketing theory and Practice*, 19(2):139–152, 2011.
7. Herman Wold. Soft modeling: the basic design and some extensions. *Systems under indirect observation*, 2:343, 1982.
8. Roman Rosipal and Nicole Krämer. Overview and recent advances in partial least squares. In *International Statistical and Optimization Perspectives Workshop "Subspace, Latent Structure and Feature Selection"*, pages 34–51. Springer, 2005.
9. G Cepeda, J Henseler, C Ringle, and JL Roldán. Prediction-oriented modeling in business research by means of partial least squares path modeling. *J. Bus. Res.*, 69(10):4545–4551, 2016.
10. Population Council. *Survey of Young People in Egypt dataset, SYPE [Computer file]*. Cairo, Egypt: OAMDI; Economic Research Forum (distributor), 2014.
11. Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*, volume 1. Cambridge university press, 1997.
12. Mervyn Stone. Cross-validated choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974.
13. Pratyush Nidhi Sharma, Galit Shmueli, Marko Sarstedt, Nicholas Danks, and Soumya Ray. Prediction-oriented model selection in partial least squares path modeling. *Decision Sciences*, 2019.
14. Joseph F Hair, Marko Sarstedt, Christian M Ringle, and Jeannette A Mena. An assessment of the use of partial least squares structural equation modeling in marketing research. *Journal of the academy of marketing science*, 40(3):414–433, 2012.
15. Wynne W Chin. How to write up and report PLS analyses. In *Handbook of partial least squares*, pages 655–690. Springer, 2010.
16. Gabriel Cepeda Carrión, Jörg Henseler, Christian M Ringle, and José Luis Roldán. Prediction-oriented modeling in business research by means of PLS path modeling: Introduction to a jbr special section. *Journal of business research*, 69(10):4545–4551, 2016.

Topological Principal Component Analysis

Rafik Abdesselam

ERIC-COACTIS Laboratories, University of Lyon, Lumière Lyon 2
Campus Berges du Rhône, 69635 Lyon Cedex 07, France
(E-mail: rafik.abdesselam@univ-lyon2.fr)
(<http://perso.univ-lyon2.fr/~rabdesse/fr/>)

Abstract. Topological Principal Component Analysis (TPCA) is a multidimensional descriptive method which studies a homogeneous set of continuous variables defined on the same set of individuals. It is a topological method of data analysis that consists of comparing and classifying proximity measures from among some of the most widely used measures for continuous data. It proposes an adjacency matrix associated to a proximity measure according to the data under consideration, then analyzes and visualizes, with graphic representations, the relationship structure of the variables relating to, the known problem of Principal Component Analysis (PCA). Based on the notion of neighborhood graphs, some of these proximity measures are more-or-less equivalent. A topological equivalence index between two measures is defined and statistically tested according to the topological correlation between the variables. The principle of the proposed TPCA is illustrated using a real data set.

Keywords: Proximity measure, neighborhood graph, adjacency matrix, topological equivalence, correlation matrix, MDS graphical representations.

1 Introduction

Similarity measures play an important role in many areas of data analysis. The results of any operation involving structuring, clustering or classifying objects are strongly dependent on the proximity measure chosen. The user has to select one measure among many existing ones. Yet, according to the notion of topological equivalence chosen, some measures are more-or-less equivalent. The concept of topological equivalence uses the basic notion of local neighborhood. We define the topological equivalence between two proximity measures, in the context of association between several categorical variables, through the topological structure induced by each measure.

Principal Component Analysis (PCA) [17], [11], [6], [19] is an important methodology among factorial techniques due to the extent of its field of application. It allows us, among other things, to describe continuous data tables.

This method concerns the relations between or within a set of quantitative variables simultaneously observed on a sample of individuals. Generally the variables are homogeneous in the sense that they revolve around a particular theme.

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



PCA is statistically considered as a widely used multivariate method for dimension reduction and as a technique of representing data. It aims to find common factors, the so-called principal components, in form of linear combinations of the variables under investigation. It allows to have an idea of the correlations structure of the set of variables, as well as possible similarities of behavior between individuals.

In order to understand and act on situations that are represented by a set of objects, very often we are required to compare them. Humans perform this comparison subconsciously using the brain. In the context of artificial intelligence, however, we should be able to describe how the machine might perform this comparison. In this context, one of the basic elements that must be specified is the proximity measure between objects.

Certainly, application context, prior knowledge, data type and many other factors can help in identifying the appropriate measure. For instance, if the objects to be compared are described by Boolean vectors, we can restrict our comparisons to a class of measures specifically devoted to this type of data. However, the number of candidate measures may still remain quite large. Can we consider that all those measures remaining are equivalent and just pick one of them at random? Or are there some that are equivalent and, if so, to what extent? This information might interest a user when seeking a specific measure. For instance, in information retrieval, choosing a given proximity measure is an important issue. We effectively know that the result of a query depends on the measure used. For this reason, users may wonder which one is more useful? Very often, users try many of them, randomly or sequentially, seeking a "suitable" measure. If we could provide a framework that allows the user to compare proximity measures in order to identify those that are similar, they would no longer need to try out all measures.

The present study proposes a new framework for comparing proximity measures in order to choose the best one in the context of association between a set of quantitative variables. The aim is to establish a PCA.

We deliberately ignore the issue of the appropriateness of the proximity measure, as it is still an open and challenging question currently being studied. The comparison of proximity measures can be analyzed from various angles.

The comparison of objects, situations or ideas is an essential task in order to assess a situation, to rank preferences, to structure a set of tangible or abstract elements, and so on. In a word, to understand and act, we have to compare. These comparisons that the brain naturally performs, however, must be clarified if we want them to be done by a machine. For this purpose, we use proximity measures. A proximity measure is a function which measures the similarity or dissimilarity between two objects within a set. These proximity measures have mathematical properties and specific axioms. But are such measures equivalent? Can they be used in practice in an undifferentiated way? Do they produce the same learning database that will serve to find the membership class of a new object? If we know that the answer is negative, then how do we decide which one to use? Of course, the context of the study and the type of data being considered can help in selecting a few possible proximity measures,

but which one should we choose from this selection as the best measure for summarizing the correlation structure of the variables?

The topological correlation structure of the variables partly depends on the data being used. The results of TPCA are different according to the selected proximity measure.

Several studies on the topological equivalence of proximity measures have been proposed, [4,5] [18] [14] [25], also in contexts of discrimination [3] and correspondences [2,1], but none of these propositions has an objective of the correlations synthesis of a set of quantitative variables.

Therefore, this article focuses on how to construct the best adjacency matrix induced by a proximity measure, taking into account the association between all the modalities of the qualitative variables.

In this paper we compare different proximity measures in an aim to synthesize the relationships of a set of continuous variables in the topological context. Comparison of these measures show that the results are different and depending on the proximity measure chosen. The rest of the paper is organised as follows. In section 2, we discuss topological equivalence between two proximity measures and show how to build an adjacency matrix associated with a proximity measure, how to compare and statistically test the degree of topological equivalence between proximity measures and how to select the best measure to describe topologically the structure of the correlations of the variables. Section 3 presents an illustrative example and surveys existing proximity measures on continuous data and presents a comparison between them. This comparison helps the reseachers to take quick decision about which measure to use for considered data. A conclusion of this work is given in section 4.

Table 7 in Appendix summarizes some classic proximity measures used for continuous data [24], we give on \mathbb{R}^n the definition of 15 of them.

We assume that we have at our disposal $\{x^k; k = 1, \dots, p\}$ a set of p homogeneous quantitative variables measured on n individuals. The interest is to analyze the topological structure of all these variables.

2 Topological Correlation

Topological equivalence is based on the concept of the topological graph also referred to as the neighborhood graph. The basic idea is actually quite simple: two proximity measures are equivalent if the corresponding topological graphs induced on the set of objects remain identical. Measuring the similarity between proximity measures involves comparing the neighborhood graphs and measuring their similarity. We will first define more precisely what a topological graph is and how to build it. Then, we propose a measure of proximity between topological graphs that will subsequently be used to compare the proximity measures.

Consider a set $E = \{x^1, x^2, \dots, x^k, \dots, x^p\}$ of p objects in \mathbb{R}^n , associated with the p variables. We can, by means of a proximity measure u , define a

neighborhood relationship V_u to be a binary relationship on $E \times E$. There are many possibilities for building this neighborhood binary relationship.

Thus, for a given proximity measure u , we can build a neighborhood graph on a set of objects-modalities, where the vertices are the modalities and the edges are defined by a property of the neighborhood relationship.

Many definitions are possible to build this binary neighborhood relationship. One can choose the Minimal Spanning Tree (MST) [12], the Gabriel Graph (GG) [16] or, as is the case here, the Relative Neighborhood Graph (RNG) [22].

For any given proximity measure u , we construct the associated adjacency binary symmetric matrix V_u of order p , where, all pairs of neighboring variables (x^k, x^l) , where $k, l = 1, p$, satisfy the following RNG definition.

Definition 1: Relative Neighborhood Graph (RNG)

$$\begin{cases} V_u(x^k, x^l) = 1 & \text{if } u(x^k, x^l) \leq \max[u(x^k, x^r), u(x^r, x^l)] ; \forall x^k, x^l, x^r \in E, \\ & x^r \neq x^k \text{ and } x^r \neq x^l \\ V_u(x^k, x^l) = 0 & \text{otherwise} \end{cases}$$

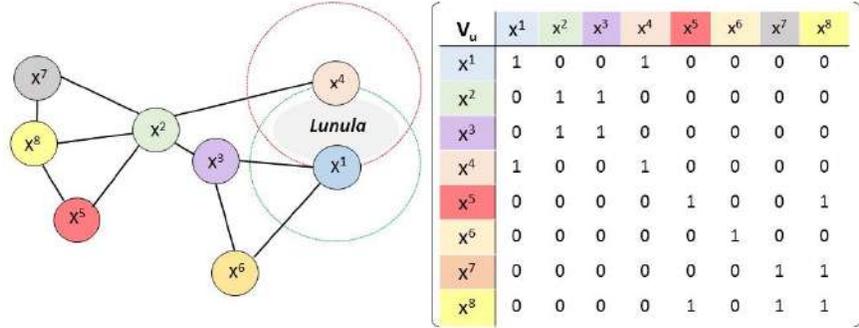


Fig. 1. RNG example with eight variables - Associated adjacency matrix

This means that if two variables x^k and x^l which verify the RNG property are connected by an edge, the vertices x^k and x^l are neighbors.

Thus, for any proximity measure given, u , we can associate an adjacency matrix V_u , of binary and symmetrical order p . Figure 1 illustrates an example of RNG in \mathbb{R}^2 of a set of $p = 8$ objects-variables.

For example, for the first and four variables, $V_u(x^1, x^4) = 1$, it means that on the geometrical plane, the hyper-Lunula (intersection between the two hyperspheres centered on the two variables x^1 and x^4) is empty.

For a given neighborhood property (MST, GG or RNG), each measure u generates a topological structure on the objects in E which are totally described by the adjacency binary matrix V_u . In this paper, we chose to use the Relative Neighbors Graph (GNR).

2.1 Comparison and selection of proximity measures

First we compare different proximity measures according to their topological similarity in order to regroup them and to better visualize their resemblances.

To measure the topological equivalence between two proximity measures u_i and u_j , we propose to test if the associated adjacency matrices V_{u_i} and V_{u_j} are different or not. The degree of topological equivalence between two proximity measures is measured by the following definition of concordance.

Definition 2: Topological equivalence between two adjacency matrices

$$S(V_{u_i}, V_{u_j}) = \frac{1}{p^2} \sum_{k=1}^p \sum_{l=1}^p \delta_{kl}(x^k, x^l)$$

$$\text{with } \delta_{kl}(x^k, x^l) = \begin{cases} 1 & \text{if } V_{u_i}(x^k, x^l) = V_{u_j}(x^k, x^l) \\ 0 & \text{otherwise.} \end{cases}$$

Then, in our case, we want to compare these different proximity measures according to their topological equivalence in a context of association. So we define a criterion for measuring the spacing from the independence or no association position.

A contingency table is one of the most common ways to summarize categorical data. Generally, interest lies in whether there is an association between the row variable and the column variable that produce the table; sometimes there is further interest in describing the strength of that association. The data can arise from several different sampling frameworks, and the interpretation of the hypothesis of no association depends on the framework. The question of interest is whether there is an association between the two variables.

We construct the adjacency matrix denoted by V_{u_\star} , which corresponds to the correlation matrix. Thus, to examine the correlation structure between the variables, we examine the significance of their linear correlation coefficient. This adjacency matrix can be written as follows using the t-test of the linear correlation coefficient ρ of Bravais-Pearson:

Definition 3: Adjacency matrix V_{u_\star} associated to reference measure u_\star

$$\begin{cases} V_{u_\star}(x^k, x^l) = 1 & \text{if } p\text{-value} = P[|T_{n-2}| > t\text{-value}] \leq 5\% ; \forall k, l = 1, p \\ V_{u_\star}(x^k, x^l) = 0 & \text{otherwise} \end{cases}$$

Where p-value is the significance test of the correlation coefficient for the two-sided test of the null and alternative hypotheses, $H_0 : \rho(x^k, x^l) = 0$ vs. $H_1 : \rho(x^k, x^l) \neq 0$.

The p-value is the evidence against a null hypothesis. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis which means that there is no correlation between x^k and x^l variables in the population

Formula for the Student t-test for significance of correlation: $t = r \sqrt{\frac{n-2}{1-r^2}}$ with $\nu = n - 2$ degrees of freedom and $r = r(x^k, x^l)$ is the linear correlation coefficient observed between the variables x^k and x^l .

Let T_{n-2} be a t-distributed random variable of Student with $\nu = n - 2$ d.f. In this case, the null hypothesis is rejected with a p-value less or equal

a significance level of 5%. Using linear correlation test, if the p-value be very small, it means that there is very small opportunity that null hypothesis is correct, and consequently we can reject it. Statistical significance in statistics is achieved when a p-value is less than the significance level of 5% for example. The p-value is the probability of obtaining results which acknowledge that the null hypothesis is true.

The binary and symmetric adjacency matrix build V_{u_\star} , is associated with an unknown proximity measure denoted u_\star and called a reference measure. Thus, with this reference proximity measure we can establish $S(V_{u_i}, V_{u_\star})$, the topological equivalence between the two proximity measures u_i and u_\star , by measuring the percentage of similarity between the adjacency matrix V_{u_i} and the reference adjacency matrix V_{u_\star} .

In order to graphically describe the similarities between proximity measures, we can for example apply the notion of themascope [13], which is a methodological sequence of a clustering method on the results of a factorial method. In this case, a Principal Component Analysis (PCA) followed by a Hierarchical Ascendant Classification (HAC) were performed upon the 15 component dissimilarity matrix defined by $[D]_{ij} = D(V_{u_i}, V_{u_j}) = 1 - S(V_{u_i}, V_{u_j})$ to partition them into homogeneous groups and to view their similarities in order to see which measures are close to one another.

We can use any classic visualization techniques to achieve this. For example, we can build a dendrogram of hierarchical clustering of the proximity measures. We can also use multidimensional scaling or any other technique, such as Laplacian projection, to map the 15 proximity measures into a two dimensional space.

Finally, in order to evaluate and determine the closest class of proximity measures to the reference measure u_\star , we project the latter as a supplementary element into the two data analysis methods, positioned by the dissimilarity vector with 15 components $[D]_{*i} = 1 - S(V_{u_\star}, V_{u_i})$.

2.2 Statistical comparisons between two proximity measures

In this section, we use the Fisher's Exact Test [10] which is an alternative to the Chi-square test when the samples are small. The principle of this test is to determine if the configuration observed in the contingency table is an extreme situation compared to the possible situations taking into account the marginal distributions. Fisher's exact test is an exact statistical test used for the analysis of contingency tables. It is a test qualified as exact because the probabilities can be calculated exactly rather than relying on an approximation which becomes correct only asymptotically as for the chi-square test used in the contingency tables. It is not based on a test statistic whose law is known when n is large enough but it calculates, as its name suggests, the exact p-value directly. To test statistically the topological equivalence between two proximity measures. This non parametric test compares these measures based on their associated adjacency matrices. Two proximity measures are statistically in topological equivalence if the null hypothesis H_0 of independence is rejected.

The comparison between indices of proximity measures has also been studied by [20], [21] and [8] from a statistical perspective. The authors proposed an approach that compares similarity matrices obtained by each proximity measure, using Mantel's test [15], in a pairwise manner.

Fisher's exact test is the statistical test best suited to compare matched binary data, the Cohen's Kappa test [7] also but it is in general an asymptotic test. The Kendall or Spearman coefficient compares matched continuous data.

It makes it possible in this context to measure the agreement or the concordance of the binary values of two adjacency matrices associated with two proximity measures. The Fisher's exact test between two adjacency matrices evaluates the topological equivalence between their proximity measures.

Let V_{u_i} and V_{u_j} be adjacency matrices associated with two proximity measures u_i and u_j . To compare the degree of topological equivalence between these two measures, we propose to test if the associated adjacency matrices are statistically different or not, using a non-parametric test of paired data. These binary and symmetric matrices of order p , are unfolded in two vector-matched components, consisting of $\frac{p(p+1)}{2}$ values: the p diagonal values and the $\frac{p(p-1)}{2}$ values above or below the diagonal.

The degree of topological equivalence between two proximity measures is evaluated from the Fisher's exact test, computed on the 2×2 contingency table formed by the two binary vectors of order $\frac{p(p+1)}{2}$.

We also test the topological equivalence between each proximity measure $u_{i=1,15}$ and the reference measure u_* by comparing the adjacency matrices V_{u_i} and V_{u_*} .

2.3 Graphical representations - Variables and Individuals

In order to represent graphically the possible topological links between the p quantitative variables, we use MultiDimensional Scaling (MDS) which makes it possible to find, for any distance matrix (similarity or dissimilarity) of size $p \times p$, a set of p points identified by their Euclidean coordinates whose distance matrix is equal to or very close to the given distance matrix. We propose to carry out the classical MDS, namely factorial analysis on similarity V_{u_*} or dissimilarity $D_{u_*} = U - V_{u_*}$ table [6]. The topological Principal Component Analysis (TPCA) returns to perform the following PCA:

Definition 4:

TPCA consist to perform the PCA of the triple $\{V_{u_*} ; M ; D_p\}$, where, V_{u_*} is the adjacency matrix associated with the proximity measure u_* , the most appropriate measure for the considered data, $M = I_p$ is the identity matrix of order p and $D_p = \frac{1}{p}I_p$ is the weighted diagonal matrix of variable weights.

The TPCA can be performed from any adjacency matrix V_{u_i} associated with each of the 15 proximity measures u_i considered.

Aid for the interpretation of TPCA results are those of PCA. Graphical representations on factorial plans allow to visualize and identify the topological

structure of the variables. As in PCA, for representations of variables, we consider the most significant variables on the axes, that is the variables highly correlated with factors, having a strong contribution and a good quality of representation, measured by the square cosine of the angle between main axes and initial axes.

For representations of active individuals, these are projected as illustrative elements. The quality of representation of these individuals on the factorial axes is measured by their squared cosine.

3 Illustrative example and Empirical results

To illustrate the TPCA, we use Eurostat data [9] on government finance of the 28 European Union (EU) countries in 2017. We examine how key government finance statistics have developed in the EU-28. Specifically, it considers general government gross debt, deficit/surplus, total revenue and total expenditure. Simple statistics of the considered variables are displayed in Table 1.

Table 1. Summary statistics of public finances

Variable	Frequency	Mean	Standart Deviation (N)	Coefficient of variation (%)	Min	Max
Debt	28	68.043	36.539	53.70	8.70	176.10
Deficit	28	-0.264	1.692	640.07	-3.10	3.50
Revenues	28	42.579	6.654	15.63	26.00	53.80
Expenditures	28	42.850	6.793	15.85	26.30	56.50

In a metric and classical context, we simply have to apply a standardised PCA on the homogeneous set of the 4 characteristics of the government finance of the EU-28.

In a topological context, the main results of the proposed method are presented in the following tables and graphs, which allow us to visualize proximity measures close to each other and to select the one that best describes and synthesis, the government finance of the EU-28.

The objective here is to give a topological synthesis of the public finances of the EU countries in 2017.

An HAC algorithm based on the Ward criterion [23] was used in order to characterize classes of proximity measure relative to their similarities. The reference measure u_* is projected as a supplementary element. The dendrogram of Figure 2 represents the hierarchical tree of the 15 proximity measures considered. Table 2 describes the final composition of each class of proximity measures, the results of the chosen partition into three homogeneous classes, obtained from the cut of the hierarchical tree of Figure 2.

Aggregation based on the criterion of the loss of minimal inertia.

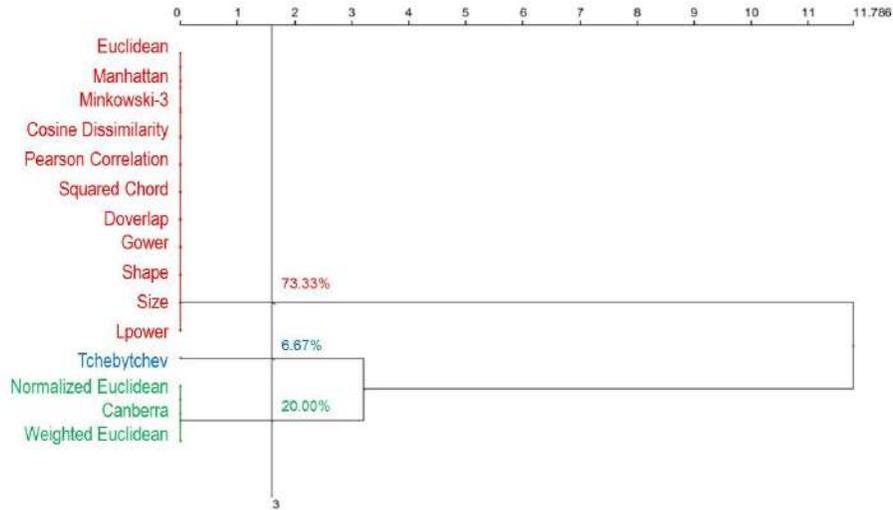


Fig. 2. Hierarchical tree of the proximity measures

Moreover, in view of the results in Table 2, the reference measure u_* is closer to the third class consisting of Normalized Euclidean, Canberra and Weighted Euclidean measures for which there is a strong topological association between the variables of government finance of EU-28 among the 15 proximity measures considered.

Table 2. Clusters composition - Assignment of the reference measure

Class number	Class 1	Class 2	Class 3
Frequency	11	1	3
Proximity measure	<i>Euclidean</i> <i>Manhattan</i> <i>Minkovski - 3</i> <i>Cosine Dissimilarity</i> <i>Pearson Correlation</i> <i>Squared Chord</i> <i>Doverlap, Gower</i> <i>Shape, Size, Lpower</i>	<i>Tchebytchev</i>	<i>Canberra</i> <i>Normalized Euclidean</i> <i>Weighted Euclidean</i>
Reference measure			u_*

It was shown in [25], by means of a series of experiments, that the choice of proximity measure has an impact on the results of a supervised or unsupervised classification.

In a topological framework, Table 3 summarizes all the results of Table 8 given in the Appendix, the similarities and Fisher's Exact p-values between all the $C_{15}^2 = 105$ pairs of proximity measures formed with the 15 measures considered and the 15 pairs formed with the unknown reference measure u_* .

Table 3. Similarities and Fisher’s Exact Test

u_i	u_j	$S(u_i, u_j)$	$p - value$
Class 1	Class 1	1.0000	0.0083**
Class 1	Class 2	0.7500	0.1833
Class 1	Class 3	0.7500	0.1833
Class 2	Class 2	1.0000	0.0083**
Class 2	Class 3	0.5000	1.0000
Class 3	Class 3	1.0000	0.0083**
u_*	Class 1	0.7500	0.1833
u_*	Class 2	0.6250	0.5000
u_*	Class 3	0.8750	0.0333*

Significance level α ; ** $\alpha \leq 1\%$; * $\alpha \in]1\%; 5\%$ **Table 4.** 2×2 Contingency Table - Similarity - Fisher’s Exact Test

Class 2		Class 1 : Euclidean		Mesure		Class 1 : Euclidean	
Tchebytchev	$V_{u_2} = 0$	$V_{u_2} = 1$	Reference	$V_{u_2} = 0$	$V_{u_2} = 1$		
$V_{u_1} = 0$	2	1	$V_{u_*} = 0$	3	1		
$V_{u_1} = 1$	1	6	$V_{u_*} = 1$	0	6		
$S(V_{u_2}, V_{u_1}) = 75\%$; $p - value = 0.1833$				$S(V_{u_*}, V_{u_1}) = 75\%$; $p - value = 0.183$			
Class 3		Class 2 : Tchebytchev		Mesure		Class 2 : Tchebytchev	
Canberra	$V_{u_2} = 0$	$V_{u_2} = 1$	Reference	$V_{u_2} = 0$	$V_{u_2} = 1$		
$V_{u_1} = 0$	1	2	$V_{u_*} = 0$	2	2		
$V_{u_1} = 1$	2	5	$V_{u_*} = 1$	1	5		
$S(V_{u_3}, V_{u_2}) = 50\%$; $p - value = 1.000$				$S(V_{u_*}, V_{u_2}) = 62.50\%$; $p - value = 0.500$			
Class 1		Class 3 : Canberra		Mesure		Class 3 : Canberra	
Euclidean	$V_{u_2} = 0$	$V_{u_2} = 1$	Reference	$V_{u_2} = 0$	$V_{u_2} = 1$		
$V_{u_1} = 0$	2	1	$V_{u_*} = 0$	3	1		
$V_{u_1} = 1$	1	6	$V_{u_*} = 1$	0	6		
$S(V_{u_1}, V_{u_3}) = 75\%$; $p - value = 0.1833$				$S(V_{u_*}, V_{u_3}) = 87.50\%$; $p - value = 0.0333^*$			

Significance level α ; ** $\alpha \leq 1\%$; * $\alpha \in]1\%; 5\%$

The values below the diagonal correspond to the similarities $S(V_{u_i}, V_{u_j})$ and the values above the diagonal are the Fisher’s Exact test p-values.

The similarities in pairs between the 15 proximity measures differ somewhat: some are closer than others. Some measures are in perfect topological equivalence $S(V_{u_i}, V_{u_j}) = 1$ with a significant Fisher’s exact test p-value $< 5\%$; these are therefore identical for the data considered, as is the case with the measures in each class of the partition presented in Table 2.

The table 4 illustrates the contingency tables 2×2 between the measures of each class: Euclidean, Tchebytchev, Canberra and reference measure u_* for the calculation of Fisher’s exact test.

Only the topological equivalence between the reference measure and the Canberra measure is significant, p-value = 0.0034 $< \alpha = 5\%$, the null hypothesis H_0 of independence is rejected.

Table 5. Pearson correlation matrix (p-values)

Variables	Debt	Deficit	Revenues	Expenditures
Debt	1.000			
Deficit	-0.3403 (0.076)	1.000		
Revenues	0.3071 (0.112)	0.0393 (0.8428)	1.000	
Expenditures	0.3845 (0.0434*)	-0.2092 (0.2853)	0.9689 (0.0001**)	1.000

Significance level α ; ** $\alpha \leq 1\%$; * $\alpha \in]1\%; 5\%$

$$V_{u_*} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix}$$

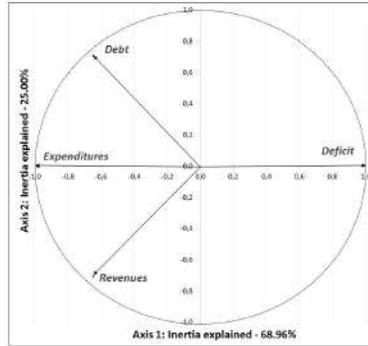


Fig. 3. TPCA - Adjacency matrix and Representation of public finance variables on the first principal plane

The adjacency matrix V_{u_*} associated to the adapted proximity measure u_* to the considered data, is build from the correlations matrix Table 5 according to definition 3. Figure 5 shows on the main first TPCA plane, the topological correlation between the Government finance variables.

The corresponding representation for individuals is given in Figure 4. It is thus possible to suggest which are the variables - government finance are responsible for the proximities between the individuals - the 28 EU countries.

The main numerical and graphical results of the proposed TPCA are given in the following Tables and Figures, and are compared to those of the classical PCA.

Figure 5 presents, for comparison on the first factorial plane, the correlations between principal components - Factors and the original variables. We can see that these graphical representations of the variables are slightly different. Effectively, The percentage of inertia explained on the first principal plane of the Topological PCA is greater than that of Classical PCA and The significant correlations variables-factors are also different.

Table 6 shows that the two first largest eigenvalues of TPCA together account for 93.96% of the variance while the classical PCA summarize 84.88%.

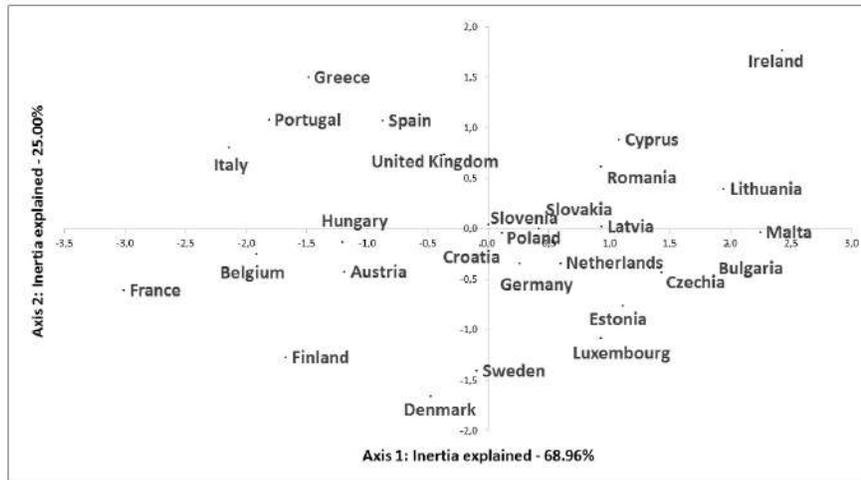


Fig. 4. TPCA - Representation of the EU-28 countries on the first principal plane

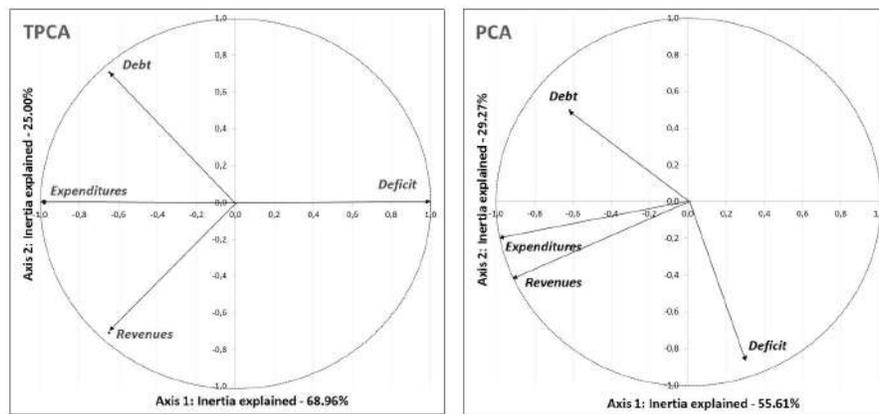


Fig. 5. TPCA - PCA: Representations of public finance variables on the first principal plane

Thus, the first two factors provide an adequate summary of the data, i.e. of government finance of UE-28 countries, we restrict the comparison of the graphical representations to the first factorial plane.

The correlation tables show that the original variables are strongly correlated with the factors, those that contribute the most to the achievement of this principal component.

While the first PCA factor (55.61%) is strongly correlated with three of the original variables, expenditures, revenues and debt, the first TPCA factor (68.96%) opposes these three variables to the deficit. As for the second PCA (29.27%) and TPCA (25.00%) factors, they oppose the debt to revenues.

Table 6. TPCA and PCA - Eigenvalues and Correlations Variables & Factors

TPCA - Eigenvalue	Proportion	Cumulative	Correlations		Factors	
			Variables	F1	F2	
2.758	68.96%	68.96%	Debt	0.645	0.707	
1.000	25.00%	93.96%	Deficit	0.982	0.000	
0.242	6.04%	100.00%	Revenues	0.645	-0.707	
0.000	0.00%	100.00%	Expenditures	0.982	0.000	
4	100.00%	100.00%				

PCA - Eigenvalue	Proportion	Cumulative	Correlations		Factors	
			Variables	F1	F2	
2.224	55.61%	55.61%	Debt	-0.615	0.497	
1.171	29.27%	84.88%	Deficit	0.307	-0.845	
0.605	15.12%	100.00%	Revenues	-0.907	-0.414	
0.000	0.00%	100.00%	Expenditures	-0.964	-0.196	
4	100.00%	100.00%				

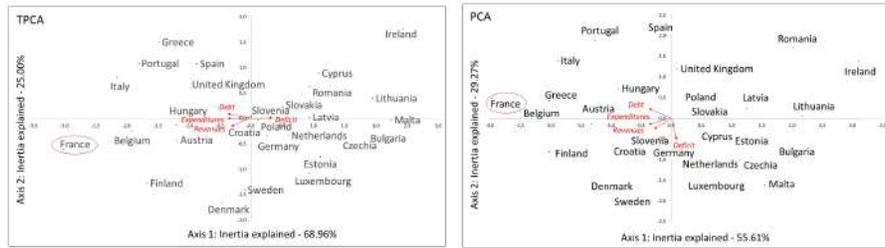


Fig. 6. TPCA and PCA - Representation of the EU countries on the first principal plane

The representations of the countries presented in Figure 6 are of course slightly different, indeed, for example, for France which contributes to the realization of the first TPCA axis, it is characterized by high Debts, high Expenditures, high Revenues and a low Deficit. France also contributes on the first PCA axis, its characterized by high Debts, high Expenditures and high Revenues, but the Deficit does not characterize the first factorial axis of the PCA.

We can represent the topological analysis of each of the 15 proximity measures considered, for example see the Euclidean TPCA in Figure 3. One can moreover give Figure 8, the graphical representation associated with a perfect no correlation between variables.

$$V_{u_{Euclidean}} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix}$$

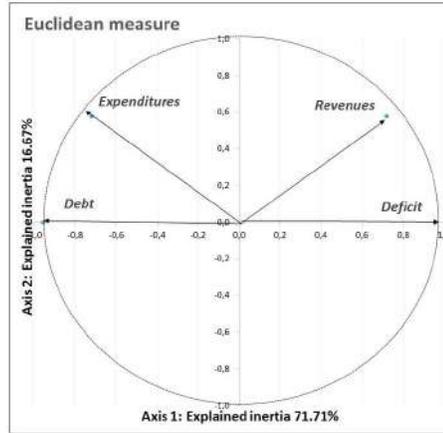


Fig. 7. Euclidean TPCA - Adjacency matrix and Representation of public finance variables on the first principal plane

$$V_{u_o} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

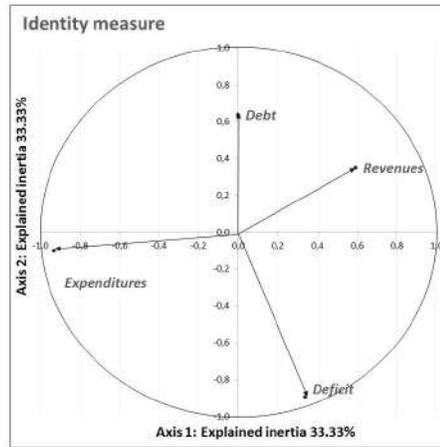


Fig. 8. Adjacency identity matrix - Total absence of correlation between variables

4 Conclusion

This research work proposes a new approach that allows to synthesize and describe a set of quantitative variables in a topological context. Like PCA, the proposed TPCA is a multidimensional topological exploratory method that can be useful for dimension reduction, it enriches the conventional quantitative data analysis methods. Future work involves extending this topological approach to synthesize the relations existing between a set of mixed (quantitatives & qualitatives) variables - Topological Mixed Principal Component Analysis - TMPCA, between two groups of continuous variables - Topological Canonical Analysis - TCA and between several multidimensional data tables - Topological Analysis of Evolutionary Data - TAED.

5 Appendix

Table 7. Some proximity measures for continuous data

Measures	Formula : Distance - Dissimilarity
Euclidean	$u_{Euc}(x, y) = \sqrt{\sum_{j=1}^p (x_j - y_j)^2}$
Manhattan	$u_{Man}(x, y) = \sum_{j=1}^p x_j - y_j $
Minkowski	$u_{Min_\gamma}(x, y) = (\sum_{j=1}^p x_j - y_j ^\gamma)^{\frac{1}{\gamma}}$
Tchebychev	$u_{Tch}(x, y) = \max_{1 \leq j \leq p} x_j - y_j $
Normalized Euclidean	$u_{NE}(x, y) = \sqrt{\sum_{j=1}^p \frac{1}{\sigma_j^2} [(x_j - \bar{x}_j) - (y_j - \bar{y}_j)]^2}$
Cosine dissimilarity	$u_{Cos}(x, y) = 1 - \frac{\sum_{j=1}^p x_j y_j}{\sqrt{\sum_{j=1}^p x_j^2} \sqrt{\sum_{j=1}^p y_j^2}} = 1 - \frac{\langle x, y \rangle}{\ x\ \ y\ }$
Canberra	$u_{Can}(x, y) = \sum_{j=1}^p \frac{ x_j - y_j }{ x_j + y_j }$
Pearson Correlation	$u_{Cor}(x, y) = 1 - \frac{(\sum_{j=1}^p (x_j - \bar{x})(y_j - \bar{y}))^2}{\sum_{j=1}^p (x_j - \bar{x})^2 \sum_{j=1}^p (y_j - \bar{y})^2} = 1 - \frac{(\langle x - \bar{x}, y - \bar{y} \rangle)^2}{\ x - \bar{x}\ ^2 \ y - \bar{y}\ ^2}$
Squared Chord	$u_{Cho}(x, y) = \sum_{j=1}^p (\sqrt{x_j} - \sqrt{y_j})^2$
Overlap measure	$u_{Dev}(x, y) = \max(\sum_{j=1}^p x_j, \sum_{j=1}^p y_j) - \sum_{j=1}^p \min(x_j, y_j)$
Weighted Euclidean	$u_{WEu}(x, y) = \sqrt{\sum_{j=1}^p \alpha_j (x_j - y_j)^2}$
Gower's Dissimilarity	$u_{Gow}(x, y) = \frac{1}{p} \sum_{j=1}^p x_j - y_j $
Shape Distance	$u_{Sha}(x, y) = \sqrt{\sum_{j=1}^p [(x_j - \bar{x}_j) - (y_j - \bar{y}_j)]^2}$
Size Distance	$u_{Siz}(x, y) = \sum_{j=1}^p (x_j - y_j) $
Lpower	$u_{Lp\gamma}(x, y) = \sum_{j=1}^p x_j - y_j ^\gamma$

Where, p is the dimension of space, $x = (x_j)_{j=1, \dots, p}$ and $y = (y_j)_{j=1, \dots, p}$ two points in R^p , \bar{x}_j the mean, σ_j the Standard deviation, $\alpha_j = \frac{1}{\sigma_j^2}$ and $\gamma > 0$.

Table 8. Similarities $S(V_{u_i}, V_{u_j})$ & Fisher's Exact Test p-values

Measure	Euclidean	Manhattan	Minkovski	Cosine dissimilarity	Pearson Correlation	Squared Chord	Overlap measure	Gower's Dissimilarity	Shape Distance	Size Distance	Lpower	Tchebytchev	Normalized Euclidean	Canberra	Weighted Euclidean
u_* measure	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.500	0.033	0.033	0.033
$u_{Euclidean}$	1														
$u_{Manhattan}$	0.008	1													
$u_{Minkovski}$	0.008	0.008	1												
u_{Cosine}	0.008	0.008	0.008	1											
$u_{Pearson}$	0.008	0.008	0.008	0.008	1										
u_{Chord}	0.008	0.008	0.008	0.008	0.008	1									
u_{Dover}	0.008	0.008	0.008	0.008	0.008	0.008	1								
u_{Gower}	0.008	0.008	0.008	0.008	0.008	0.008	0.008	1							
u_{Shape}	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	1						
u_{Size}	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	1					
u_{Lpouce}	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	1				
$u_{Tchebytch}$	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	1			
$u_{Neuclidean}$	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.500	1		
$u_{Canberra}$	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.500	0.008	1	
$u_{W_{euclidean}}$	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.500	0.008	0.008	1
u_* measure	0.875	0.875	0.875	0.625	0.875	0.875	0.875	0.875	0.875	0.875	0.875	0.875	0.875	0.875	0.875

Similarity: $S(u_{Tchebytchev}, u_{Euclidean}) = 0.750$

Fisher's Exact Test : $p - value(u_{Euclidean}, u_{Tchebytchev}) = 0.1833 > \alpha = 5\%$: not significant

References

1. Abdesselam, R.: A Topological Multiple Correspondence Analysis. *Journal of Mathematics and Statistical Science*, Science Signpost Publishing Inc., USA, Vol.5, Issue 8, 175–192, 2019.
2. Abdesselam, R.: Selection of proximity measures for a Topological Correspondence Analysis. *In a Book Series*, 5th Stochastic Modeling Techniques and Data Analysis, International Conference, Chania, Greece, C.H. Skiadas (Ed), 11–24, 2018.
3. Abdesselam, R.: A Topological Discriminant Analysis. *In book Chapter, Volume 3, Data Analysis and Applications 2: Utilization of Results in Europe and Other Topics*, J. Bozeman and C. Skiadas Editors, ISTE Science Publishing, Wiley, 167–178, 2018.
4. Batagelj, V., Bren, M.: Comparing resemblance measures. In Proc. International Meeting on Distance Analysis (Distancia'92), 1992.
5. Batagelj, V., Bren, M.: Comparing resemblance measures. *In Journal of classification*, 12, 73–90, 1995.
6. Caillez, F. and Pagès, J.P.: Introduction à l'Analyse des données", *S.M.A.S.H.*, Paris, 1976.
7. Cohen, J.: A coefficient of agreement for nominal scales. *Educ Psychol Meas*, Vol 20, 27–46, 1960.
8. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. *The journal of Machine Learning Research*, Vol. 7, 1–30, 2006.
9. Eurostat, Data source: Government finance statistics - Statistics explained, https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Government_finance_statistics, 1 – –15, 2018.
10. Fisher, R-A.: The Interpretation of χ^2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, Published by Wiley, 85, 1, 87–94.
11. Hotelling H.: Analysis of a Complex of Statistical Variables into Principal Components. *In J. Educ. Psy.*, vol. 24, 417–441, 498–520, 1933.
12. Kim, J.H. and Lee, S.: Tail bound for the minimal spanning tree of a complete graph. *In Statistics & Probability Letters*, 4, 64, 425–430, 2003.
13. Lebart, L.: Stratégies du traitement des données d'enquêtes. *La Revue de MODULAD*, 3, 21–29, 1989.
14. Lesot, M. J., Rifqi, M. and Benhadda, H.: Similarity measures for binary and numerical data: a survey. *In IJKESDP*, 1, 1, 63–84, 2009.
15. Mantel, N.: A technique of disease clustering and a generalized regression approach. *In Cancer Research*, 27, 209–220, 1967.
16. Park, J. C., Shin, H. and Choi, B. K.: Elliptic Gabriel graph for finding neighbors in a point set and its application to normal vector estimation. *In Computer-Aided Design Elsevier*, 38, 6, 619–626, 2006.
17. Pearson K.: On lines and Planes of Closest Fit to Systems of Points in Space. *In Phil, Mag.*, vol. 2, 11, 559–572, 1901.
18. Rifqi, M., Detyniecki, M. and Bouchon-Meunier, B.: Discrimination power of measures of resemblance. *IFSA'03 Citeseer*, 2003.
19. Saporta, G.: *Probabilités, analyse des données et Statistique*, Editions TECHNIP, 2011.
20. Schneider, J. W. and Borlund, P.: Matrix comparison, Part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results. *In Journal of the American Society for Information Science and Technology*, 58, 11, 1586–1595, 2007.

21. Schneider, J. W. and Borlund, P.: *Matrix comparison, Part 2: Measuring the resemblance between proximity measures or ordination results by use of the Mantel and Procrustes statistics*. In *Journal of the American Society for Information Science and Technology*, 11, 58, 1596–1609, 2007.
22. Toussaint, G. T.: *The relative neighbourhood graph of a finite planar set*. In *Pattern recognition*, 12, 4, 261–268, 1980.
23. Ward, J. R.: *Hierarchical grouping to optimize an objective function*. In *Journal of the American statistical association JSTOR*, 58, 301, 236–244, 1963.
24. Warrens, M. J.: *Bounds of resemblance measures for binary (presence/absence) variables*. In *Journal of Classification*, Springer, 25, 2, 195–208, 2008.
25. Zighed, D., Abdesselam, R., and Hadgu, A.: *Topological comparisons of proximity measures*. In the 16th PAKDD 2012 Conference. In P.-N. Tan et al., Eds. Part I, LNAI 7301, Springer-Verlag Berlin Heidelberg, 379–391, 2012.

The flexible beta-binomial regression model

Roberto Ascari and Sonia Migliorati

Department of Economics, Management and Statistics (DEMS) P.zza dell'Ateneo Nuovo, 1,
Università di Milano-Bicocca, Milano, Italy
(E-mail: roberto.ascari@unimib.it; sonia.migliorati@unimib.it)

Abstract. The overdispersion problem is a typical issue in binomial regression models which arises when data show a larger variance than the one assumed by the model. This excess of variability is often due to violation of the i.i.d. assumption of the binary variates forming the binomial observations, and it may mislead statistical analyses. Several strategies exist to deal with the overdispersion problem, one of these being the use of a regression model based on a compound distribution. A standard choice is the beta-binomial distribution (BB), obtained compounding the binomial distribution with the beta one. Despite the significant advantages shown by the beta-binomial regression model (BBReg) when compared with the binomial regression one, in some scenarios the BBReg can still be affected by overdispersion issues. A recent work [3] introduced the flexible beta, a special mixture with two beta components. In the present work we define a new distribution, the flexible beta-binomial (FBB), obtained compounding the binomial with the flexible beta. The FBB can be expressed as a finite mixture of beta-binomial distributions. Then, following a GLM-type approach, we implement a new regression model based on the FBB distribution (FBBReg) that shows a richer parametrization than existing models, and that is also characterized by a form of the variance and of the intraclass correlation coefficient easily interpretable in terms of overdispersion. This work is particularly focused on comparing the three regression models (binomial, BBReg, and FBBReg) through a simulation study, with the aim of identifying those scenarios where the FBBReg shows better performances than the other two models. Inferential issues are dealt with a Bayesian approach through a Hamiltonian Monte Carlo algorithm.

Keywords: Binomial data, Mixtures, Bayesian inference, Simulation study, Regression.

1 Introduction

In many applications, the interest focuses on counting the number Y of ‘successes’ among a fixed number n of trials. More precisely, each trial can be thought as a binary outcome U_j ($j = 1, \dots, n$) and the sum $Y = \sum_{j=1}^n U_j$ is the variable of interest. The binomial regression (BinReg) model [2] is the typical tool to model this kind of data but it can be affected by the overdispersion issue, meaning that data show a larger variance than the one assumed by the underlying binomial distribution. This excess of variability is often due to violation of the i.i.d. assumption of the variables U_1, \dots, U_n forming the binomial variate, and it may mislead statistical analyses (in particular underestimation of the standard error of regression coefficients). Several strategies have been developed to treat overdispersion, the most popular one being the use of the beta-binomial (BB) [13], a compound distribution characterized by an additional

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain



parameter, which can be interpreted as an intraclass correlation coefficient among U_1, \dots, U_n . Despite the significant advantages provided by this distribution, there are situations where a single additional parameter is unable to account for the entire extra-variation shown by overdispersed data. The aim of this work is to generalize the BB by proposing a new compound distribution, namely the flexible beta-binomial (FBB). Thanks to its finite mixture structure, this distribution is capable of handling situations where the overdispersion is due to the presence of two unobserved groups in the population (i.e. an unobserved qualitative explanatory variable). The rest of the paper is organized as follows. In Section 2 we present the main results regarding the binomial, BB and FBB distributions. In Section 3 we describe a Bayesian approach suitable for regression models based on the proposed distributions and a technique to detect overdispersion in a Bayesian framework (i.e. where classical tools based on deviance are not applicable). Finally, we present an application of the regression models to a real dataset (Section 4) and several simulation studies (Section 5) aimed to evaluate the behavior of the regression models in different challenging scenarios, namely the excess of zeros and the presence of outliers.

2 Distributions for binomial counts

The most basic random variable (rv) interpreting the number of successes among n trials is the binomial one, whose probability mass function (pmf) is given by:

$$f_{Bin}(y|n, \pi) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}, \quad (1)$$

where $y \in \{0, 1, \dots, n\}$, $n \in \mathbb{N}$ and $\pi \in (0, 1)$ represents the probability of success in a single trial. The binomial distribution is obtained assuming the binary outcomes U_1, \dots, U_n to be independent and with a common Bernoulli distribution with probability parameter π . The expected value and the variance of a binomial distribution are:

$$\begin{cases} \mathbb{E}[Y] = n\pi \\ \text{Var}(Y) = n\pi(1 - \pi). \end{cases} \quad (2)$$

Due to the strict connection between mean and variance, the binomial distribution is typically affected by the overdispersion issue. A popular way to deal with overdispersion is to use a compound distribution, i.e. a distribution obtained assuming the parameter π as random. Compound distributions for binomial data are obtained with the following steps:

1. Assume that $Y|n, \pi \sim Bin(n, \pi)$.
2. Assume that π follows some distribution $f_\pi(\pi|\boldsymbol{\theta})$ defined on the interval $(0, 1)$.
3. Compute the marginal distribution of Y : $f(y|n, \boldsymbol{\theta}) = \int f_{Bin}(y|n, \pi) f_\pi(\pi|\boldsymbol{\theta}) d\pi$.

The binomial distribution is obtained considering a degenerate distribution for π . A very popular choice is to assume that π is beta distributed with probability density function (pdf):

$$f_{Beta}(\pi|\mu, \phi) = \frac{1}{B(\phi\mu, \phi(1-\mu))} \pi^{\phi\mu-1} (1-\pi)^{\phi(1-\mu)-1}, \quad (3)$$

where $\mu = \mathbb{E}[\pi]$, ϕ is a precision parameter, $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ is the Beta function and $\Gamma(\cdot)$ denotes the usual Gamma function. Then, Y is marginally distributed according to the BB [13] distribution, defined by the following pmf:

$$f_{BB}(y|n, \mu, \phi) = \binom{n}{y} \frac{B(\phi\mu + y, \phi(1-\mu) + n - y)}{B(\phi\mu, \phi(1-\mu))}. \quad (4)$$

The mean and the variance of a BB distributed rv have the form:

$$\begin{cases} \mathbb{E}[Y] = n\mu \\ \text{Var}(Y) = n\mu(1-\mu) \left[1 + \frac{(n-1)}{\phi+1} \right]. \end{cases} \quad (5)$$

Note that $\theta = \frac{1}{\phi+1}$ can be interpreted as an overdispersion parameter, since $\text{Var}(Y)$ is an increasing function of θ and the form of $\text{Var}(Y)$ approximates the binomial variance as θ goes to zero. Moreover, θ can be interpreted as the common intraclass correlation coefficient among the binary outcomes forming the binomial count [9].

Despite this additional parameter, the variance of the BB can still be too poorly parametrized to model overdispersed data (see the application in Section 4). Thus, an alternative distribution for the parameter π is the flexible beta (FB), the univariate case of the flexible Dirichlet (FD) [34,6]. The FB is a special mixture of two beta components with a common precision parameter ϕ and two arbitrary means $\lambda_1 > \lambda_2$. Its pdf can be written as:

$$f_{FB}(\pi|\lambda_1, \lambda_2, \phi, p) = pf_{Beta}(\pi|\lambda_1, \phi) + (1-p)f_{Beta}(\pi|\lambda_2, \phi), \quad (6)$$

where $0 < \lambda_2 < \lambda_1 < 1$ and $p \in (0, 1)$ is the mixture weight. A reparametrization of the FB, which is very useful in a regression perspective, is given by:

$$\begin{cases} \mu = p\lambda_1 + (1-p)\lambda_2 \\ w = \frac{\lambda_1 - \lambda_2}{\min\{\mu/p, (1-\mu)/(1-p)\}} \\ \phi = \phi \\ p = p \end{cases} \quad (7)$$

where μ is the marginal mean of the FB distribution and w is a normalized measure of distance between the component means. Then, it is possible to compound the binomial distribution with an $FB(\mu, \phi, w, p)$. The resulting distribution is a new discrete random variable for binomial counts, the FBB(n, μ, ϕ, w, p). Its pmf can be expressed as:

$$f_{FBB}(y|n, \mu, \phi, w, p) = pf_{BB}(y|n, \lambda_1, \phi) + (1-p)f_{BB}(y|n, \lambda_2, \phi), \quad (8)$$

where $f_{BB}(y|n, \mu, \phi)$ is given by (4) and

$$\begin{cases} \lambda_1 = \mu + (1-p)w \min\left(\frac{\mu}{p}, \frac{1-\mu}{1-p}\right) \\ \lambda_2 = \mu - pw \min\left(\frac{\mu}{p}, \frac{1-\mu}{1-p}\right). \end{cases} \quad (9)$$

Equation (8) points out that the FBB distribution is a finite mixture with two BB components. The mean and the variance of the FBB take the form:

$$\begin{cases} \mathbb{E}[Y] = n\mu \\ \text{Var}(Y) = n\mu(1-\mu) \left\{ 1 + \frac{(n-1)}{\phi+1} + \frac{(n-1)}{\phi+1} \phi w^2 m(\mu, p) \right\} \end{cases} \quad (10)$$

where $m(\mu, p) = \min\left(\frac{\mu(1-p)}{p(1-\mu)}, \frac{(1-\mu)p}{(1-p)\mu}\right)$. Comparing the form of the variance assumed by the FBB and the BB distribution, some considerations can be made. The variance in Equation (10) can be decomposed into three terms. The first term is the ‘baseline’ binomial variance and it depends only on the mean parameter. The second term is included also in the BB variance (see equation (5)) and it describes the additional amount of variation due to overdispersion. The third term depends also on w and p , i.e. two parameters involved in the mixture structure, meaning that it describes an extra-variation due to the (eventual) presence of two unobserved clusters in the population.

3 Regression models for binomial counts

Let $\mathbf{Y} = (Y_1, \dots, Y_N)^\top$ be a vector of independent responses collected on a sample of size N . Y_i can be thought as the number of events among n_i trials ($i = 1, \dots, N$). Moreover, let \mathbf{X} be the $N \times (K+1)$ design matrix containing the values of K covariates collected on the N sample observations. Instead of modelling $\mathbb{E}[Y_i] = n_i \mu_i$, one can model μ_i , as n_i is fixed for every $i = 1, \dots, N$. Following a GLM-type approach [2], the logit function is a suitable choice to link the parameter $\mu_i = \mathbb{E}[Y_i/n_i]$ to the linear predictor:

$$\text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \mathbf{x}_i^\top \boldsymbol{\beta} = \eta_i, \quad i = 1, \dots, N, \quad (11)$$

where $\mathbf{x}_i = (1, x_{i1}, \dots, x_{iK})^\top$ is the vector of covariates for unit i , and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_K)^\top$ is the vector of regression coefficients.

We can define a flexible beta-binomial regression (FBBReg) model assuming that $Y_i \sim FBB(n_i, \mu_i, \phi, w, p)$. The parametrization of the FBB defines a variation independent parametric space, meaning that no constraints exist among μ_i, ϕ, w and p . In a Bayesian framework, this greatly simplifies the specification of the joint prior distribution, since it allows to assume prior independence. The major consequence of the prior independence is that we can specify a prior distribution for each parameter separately. In order to use non- or weakly informative priors to induce minimum impact on the posterior distribution, we select the following priors:

1. $\boldsymbol{\beta} \sim N_{K+1}(\mathbf{0}, \Sigma)$, where $\mathbf{0}$ is the $(K+1)$ -vector with zero elements, and Σ is a diagonal matrix with ‘large’ variance values.
2. The distribution of ϕ can be derived by imposing a uniform distribution on $\theta = \frac{1}{\phi+1}$, i.e. $\theta \sim Unif(0, 1)$.
3. $w \sim Unif(0, 1)$.

4. $p \sim Unif(0,1)$.

Note that a subset of these priors can be adopted for the BinReg (namely prior 1.) and for the beta-binomial regression (BBReg) (priors 1. and 2.) models. Inferential issues are dealt with a Bayesian approach through a Hamiltonian Monte Carlo (HMC) algorithm [5], which is a generalization of the Metropolis-Hastings combining classical Markov Chain Monte Carlo (MCMC) and deterministic simulation methods. The Stan modeling language [10] allows implementing an HMC method to obtain a simulated sample from the posterior distribution.

To compare the fit of the models we use the Watanabe-Akaike information criterion (WAIC) [11,12], a fully Bayesian criterion that balances between goodness-of-fit and complexity of a model: lower values of WAIC indicate a better fit.

3.1 Detection of overdispersion

Classical tools based on the deviance are not suitable in a Bayesian framework. For this reason, we take advantage of the posterior predictive checks [18]. The main idea of this technique is that replicated data generated under the fitted model should behave similarly to the observed data: any differences between simulated and observed data suggest a potential lack of fit of the model.

Let $\theta^{(b)}$ ($b = 1, \dots, B$) be an element of the MCMC sample from the posterior distribution and let $\mathbf{y}^{(b)}$ be a sample generated from $f_Y(y|\theta^{(b)})$ (i.e. from the posterior predictive distribution). Furthermore, let $T(\cdot)$ be a function of data and model parameters. Then, it is possible to compare the empirical distribution of $T(\mathbf{y}^{(b)})$ ($b = 1, \dots, B$) with $T(\mathbf{y})$ (i.e. the value of $T(\cdot)$ computed on the observed data). This comparison can be fulfilled via plots as well as posterior predictive p -values defined as $P(T(\mathbf{y}^b) \geq T(\mathbf{y})|\mathbf{y})$ (the closer to 0.5 the better). Figure 1 summarizes the posterior predictive checks procedure.

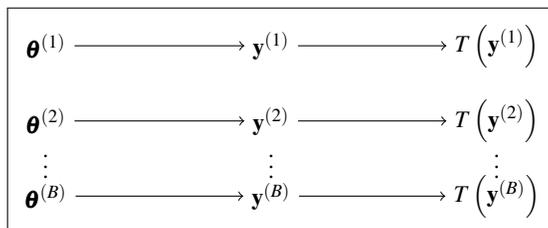


Fig. 1. Algorithm of posterior predictive checks.

In a regression perspective, reasonable choices for $T(\cdot)$ are the mean and the variance. If the observed variance is far away from the center of the distribution, then we can believe that data are overdispersed with respect to the considered model.

4 Real data application

Let us consider a dataset proposed by Otake and Prentice [7]. For a large number of survivors of the atomic bomb in Hiroshima and Nagasaki, 100 cells have been ana-

lyzed and the number Y of cells with chromosomal abnormalities has been recorded. Furthermore, for each subject, the estimated radiation exposure level X (in rads) has been collected. In Figure 2 we plot the sample proportion of events (y-axis) and the dose variable (x-axis), which has been jittered with a normal white noise for better visualization. Each point has been plotted with different colors according to the bomb the subject survived.

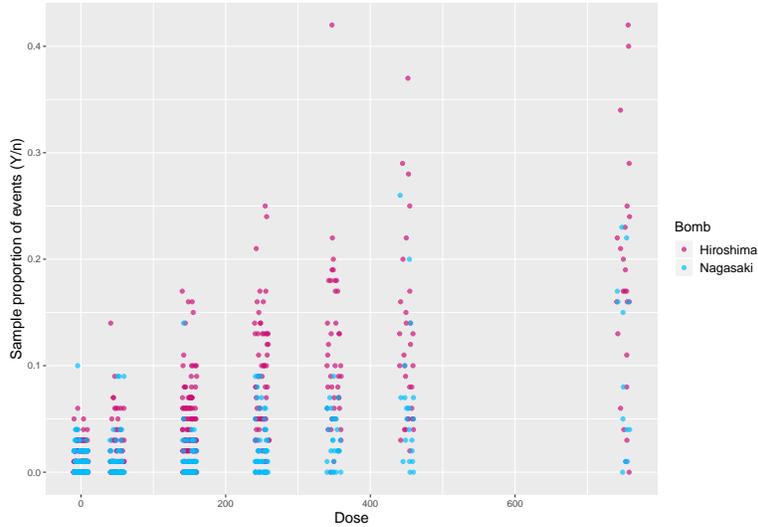


Fig. 2. Atomic Data

We used the `rstan` package in R software to estimate the BinReg, BBReg and FBBReg’s parameters, running a chain of length 30000 and discarding the first 50%. Table 1 reports the posterior mean and 95% credible set (CS) for each estimated parameter as well as the WAIC index for each model. Note that the effect of the dose

Param.	BinReg	BBReg	FBBReg
β_0	-4.130 (-4.184; -4.081)	-4.002 (-4.104; -3.898)	-4.122 (-4.223; -4.019)
β_1	0.0037 (0.0036; 0.0039)	0.0033 (0.0031; 0.0036)	0.0038 (0.0036; 0.0041)
ϕ	(—)	24.396 (21.071; 28.193)	32.975 (27.695; 38.958)
p	(—)	(—)	0.853 (0.745; 0.926)
w	(—)	(—)	0.763 (0.609; 0.879)
WAIC	6163.2	4418.1	4378.6

Table 1. Posterior means and 95% credible sets for the parameters of the three models.

level on the probability of chromosomal abnormalities (β_1) is significant and positive for all models, with the FBBReg showing the highest estimate value (stronger impact). The WAICs reported in Table 1 point the FBBReg as the best model. Looking at the posterior predictive checks in Figure 3 we can see that the means of the replicated

datasets agree with the mean of the observed data for all the three models considered. From the right panel of Figure 3 it is possible to note that the BinReg clearly suffers from the overdispersion problem since the distribution of the variance of the replicated data is far away from the observed variance. Both the BBReg and the FBBReg handle the variance better than the BinReg, with the posterior predictive p -values reported in Table 2 pointing the FBBReg as the preferable model. Looking at Figure 4 it is possible to observe that the FBBReg is able to detect two latent clusters (Hiroshima and Nagasaki groups) and to build a ‘marginal’ regression curve (black curve) that is a weighted mean of the clusters’ regression curves (red and blue curves).

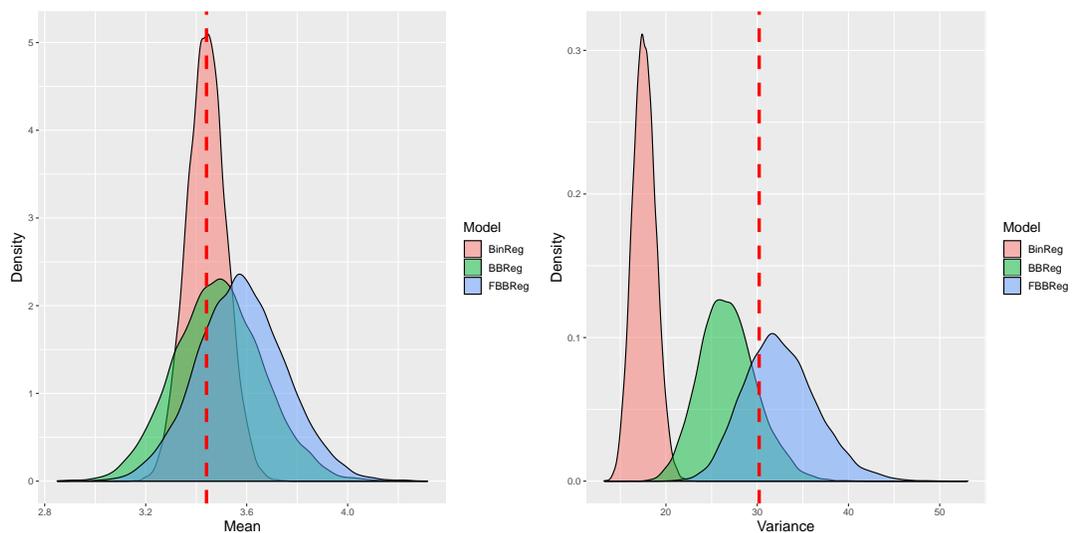


Fig. 3. Posterior Predictive checks for mean and variance for the BinReg, BBReg, and FBBReg models.

	Mean	Variance
BinReg	0.50025	0
BBReg	0.6075	0.1395
FBBReg	0.7829	0.7161

Table 2. Estimated posterior predictive p -values.

5 Simulation studies

This section presents the results of three simulation studies. The first one aims at comparing the fitting abilities of the considered models in different scenarios. The second and third simulation studies compare the regression models in some challenging scenarios, namely excess of zeros and outliers contamination. In all studies, the inference is based on three chains of length 10000 (having set a warm-up period of 5000).

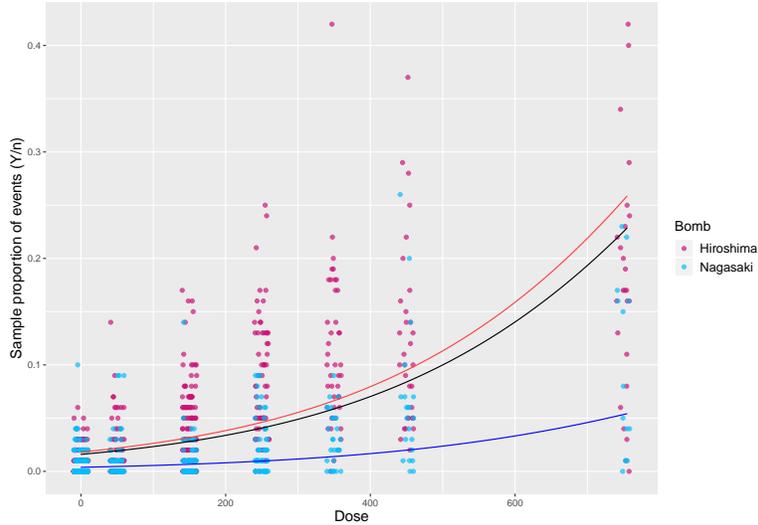


Fig. 4. Jittered Atomic Data. Black line represents FBB regression curve (μ) whereas red and blue lines represent the cluster regression curves (λ_1 and λ_2 , respectively).

5.1 Fitting study

In this Section we consider four data generating processes: (1) an FBBReg model with well-separated clusters and a non-null ‘overdispersion’ parameter $\frac{1}{\phi+1} = 0.3$; (2) a BBReg model; (3) a mixture of BBReg models with different mean and precision parameters (i.e. a mixture of BBReg which is not an FBBReg); (4) an FBBReg model with well-separated clusters and an almost-null ‘overdispersion’ parameter $\frac{1}{\phi+1} = 0.05$. Table 3 summarises the parameters adopted for each scenario.

Scenario	Generating Model	β_0	β_1	ϕ	ϕ_2	w	p
1	FBBReg	-1	2	2.333	(—)	0.75	0.5
2	BBReg	-0.5	1	100	(—)	(—)	(—)
3	Mixture of BBReg	1	2	10	20	0.7	0.5
4	FBBReg	-1	2	19	(—)	0.75	0.5

Table 3. Parameter configurations under each scenario.

For each scenario we simulated a sample of size $N = 200$. In particular, we generated a single covariate x from a uniform distribution on the interval $(-1.5, 2)$ and $\mathbf{n} = (n_1, \dots, n_N)^T$ from i.i.d. Poisson(200) rv’s. A logit link function has been adopted to link the mean parameter and the covariate: $\text{logit}(\mu_i) = \beta_0 + \beta_1 x_i$ ($i = 1, \dots, N$), for fixed β_0 and β_1 . We replicated each scenario 1000 times. Table 4 shows the bias and the mean square error resulting from having estimated parameters through their posterior mean. The last column of Table 4 contains the mean of the WAIC index over the 1000 replications. A graphical representation of the estimates is provided in Figure 5.

Both the FBBReg and the BBReg models show a far better performance (lower WAIC values) than the BinReg model under all scenarios.

Scenario	Model	β_0	β_1	WAIC
1	BinReg	-0.0095 (0.0319)	0.0278 (0.0369)	12187.99 (100%)
	BBReg	0.1087 (0.0331)	-0.1989 (0.0586)	1048.98 (96.3%)
	FBBReg	0.0030 (0.0214)	0.0003 (0.0254)	1038.27 (—)
2	BinReg	0.0004 (0.0006)	-0.0013 (0.0006)	1272.97 (100%)
	BBReg	0.0005 (0.0006)	-0.0015 (0.0006)	1139.80 (10.8%)
	FBBReg	0.0010 (0.0006)	-0.0032 (0.0006)	1140.29 (—)
3	BinReg	0.0038 (0.0125)	0.0112 (0.0173)	6181.97 (100%)
	BBReg	-0.1544 (0.0319)	-0.3919 (0.1652)	1208.84 (100%)
	FBBReg	0.0845 (0.0144)	-0.0556 (0.0108)	1149.21 (—)
4	BinReg	-0.002 (0.0163)	0.0117 (0.0135)	7167.598 (100%)
	BBReg	0.1769 (0.0415)	-0.3053 (0.1015)	1277.939 (100%)
	FBBReg	-0.0008 (0.0087)	0.0079 (0.0067)	1217.341 (—)

Table 4. Bias and Mean Square Error (in parenthesis) of the estimates of the regression parameters. Mean values of WAIC in the last column, together with % of selection of the FBBReg model versus BinReg and BBReg models (in parenthesis).

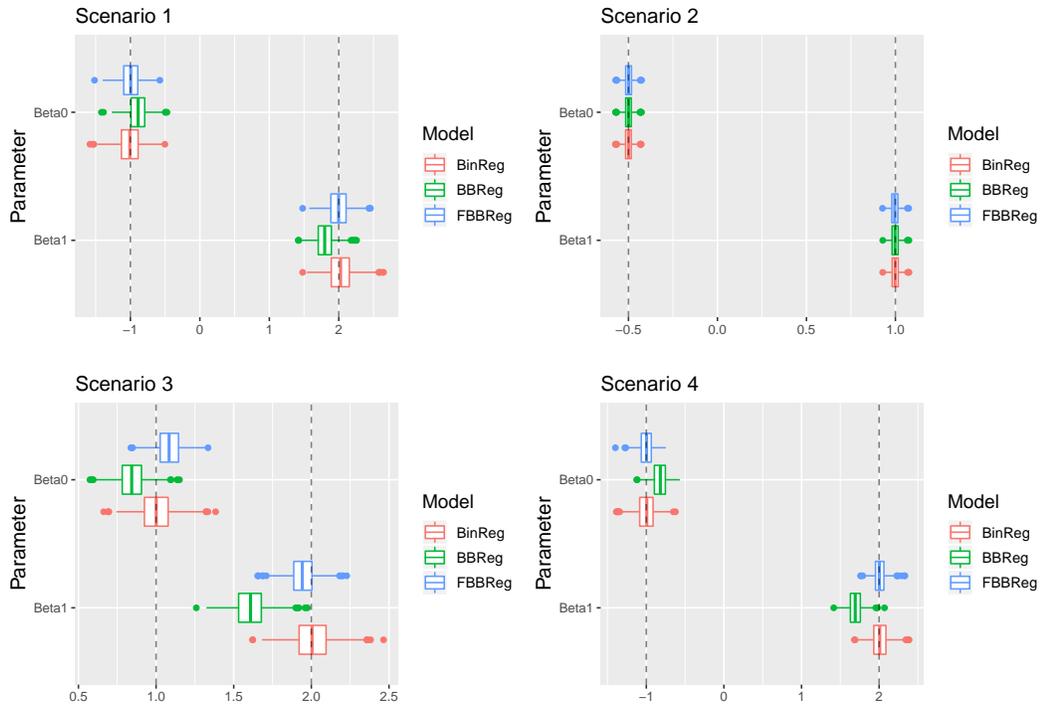


Fig. 5. Distribution of the posterior means of β_0 and β_1 under each scenario for each model. Vertical dashed lines represent the true parameter values.

Under scenario 2 (data generated from the BBReg model), all three models provide accurate estimates, as it is clearly shown by the upper-right panel of Figure 5.

Moreover, the FBBReg model displays a goodness of fit (WAIC mean) very similar to the BBReg’s one. This means that the FBBReg model succeeds in capturing the overdispersion due to a common correlation among the binary outcomes forming the binomial count.

Conversely, whenever data are generated in the presence of two latent clusters (scenarios 1, 3 and 4) all parameter estimates of the BBReg model are highly inaccurate (and even less accurate than the BinReg ones), as it emerges from the upper-left and from the two lower panels of Figure 5. This happens regardless of whether the FBBReg is the data generating model or not, and regardless of the presence of high (scenario 1) or low (scenario 4) overdispersion as well. Moreover, in all three scenarios the FBBReg model displays lower WAIC values than the BBReg ones.

The above remarks highlight that the FBBReg model can recognize (and adapt to) a wider spectrum of scenarios than the simpler BBReg model; though, it preserves accuracy in the estimates and goodness of fit even when the latter is the data generating model.

5.2 Excess of zeros

To study the behavior of the regression models in the presence of a higher percentage of zeros than the one assumed by the data generation process, we designed ad hoc scenarios (250 replications for each one) by artificially creating increasing percentages of zeros. More precisely, for each replication we generated a sample of size $N = 100$ such that $Y_i \sim \text{Bin}(n_i, \text{logit}^{-1}(1 + 2x_i))$, where n_i is generated from a Poisson(50), and x_i is generated from a uniform distribution on the interval $(-1.5, 2)$. Then, in each scenario we randomly selected a different percentage of observations (5%, 10%, 20% and 50%) and set their outcome Y to zero.

Figure 6 shows the mean of the WAIC over the replications. It is noteworthy that the performance of the BinReg gets worse as the percentage of zeros increases, while both the BBReg and the FBBReg models handle the excess of zeros in an adequate way. In particular, the FBBReg performs better than the BBReg, probably due to one mixture component dedicated to the group of zeros.

5.3 Outliers contamination

Outliers often arise in many kinds of applications. In binomial regression contexts, outliers can be (part of) the reason for overdispersion. To study the robustness of the regression models to outlier contamination, we replicated each of the following scenarios 500 times. Namely, we adopted the parameter configuration described in Subsection 5.2 then we artificially modified the outcome of a randomly selected subset of observations as:

$$y_r^{\text{New}} = n_r - y_r^{\text{Old}}, \quad \text{for some randomly selected } r.$$

Scenarios are characterized as follows:

- Scenario 1: we randomly selected 3 observations (3%) with $x < x_{0.15}$ (i.e. the 15th empirical percentile)
- Scenario 2: we randomly selected 3 observations (3%) with $x > x_{0.85}$

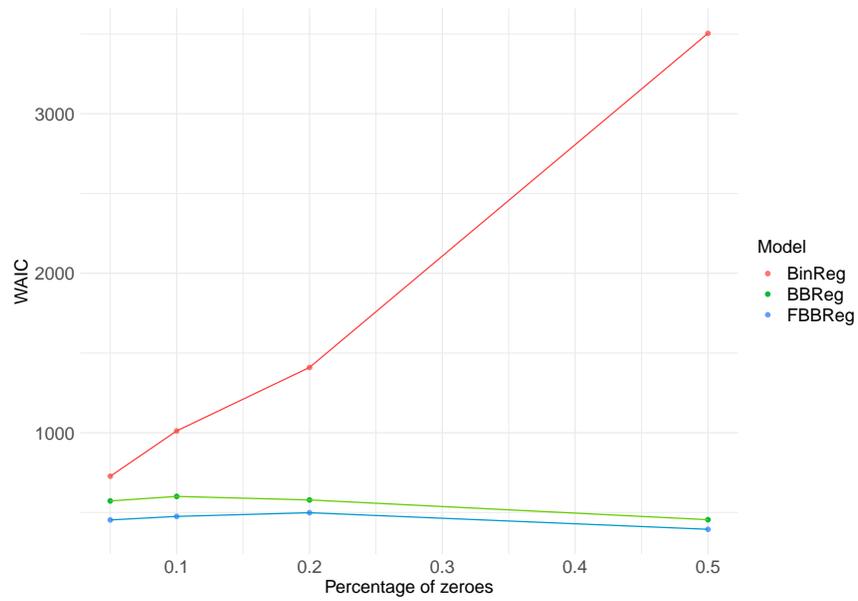


Fig. 6. Mean of WAIC as function of the percentage of zeroes artificially created in the dataset.

- Scenario 3: we randomly selected 3 observations (3%) with $x < x_{0.15}$ and 3 observations (3%) with $x > x_{0.85}$

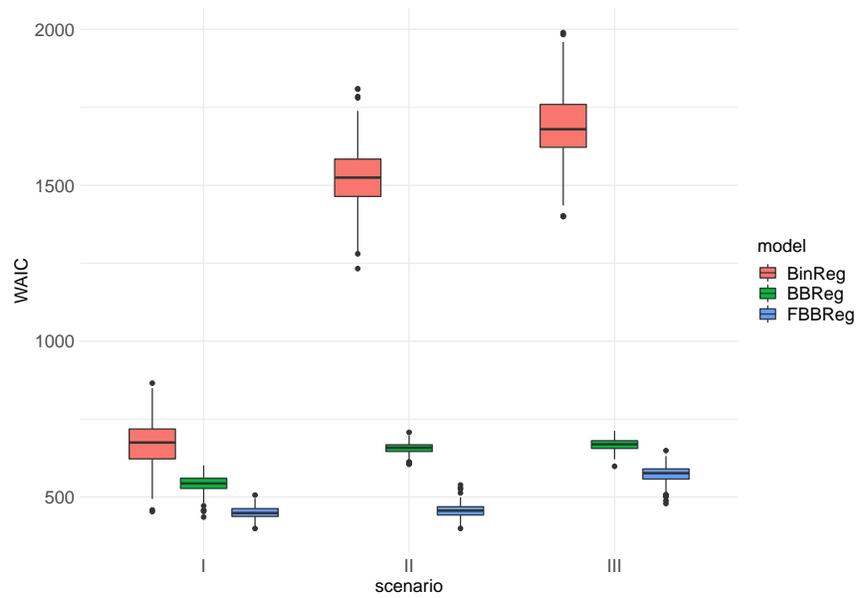


Fig. 7. Distribution of log(WAIC) for each outlier configuration.

In Figure 7 we can see the distribution of the WAIC for every model and every replication. In all scenarios the FBB performs better than simpler models, owing to the fact that one mixture component can be dedicated to a particular (even small) group of outlying observations.

References

1. A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. CRC Press, third edition, 2014.
2. P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, second edition, 1989.
3. S. Migliorati, A. M. Di Brisco, and A. Ongaro. A new regression model for bounded responses. *Bayesian Analysis*, 13(3):845–872, 2018.
4. S. Migliorati, A. Ongaro, and G. S. Monti. A structured Dirichlet mixture model for compositional data: inferential and applicative issues. *Statistics and Computing*, 27(4):963–983, 2017.
5. R. M. Neal. An improved acceptance procedure for the hybrid monte carlo algorithm, 1994.
6. A. Ongaro and S. Migliorati. A generalization of the dirichlet distribution. *Journal of Multivariate Analysis*, 114(1):412–426, 2013.
7. M. Otake and R. L. Prentice. The Analysis of Chromosomally Aberrant Cells Based on Beta-Binomial Distribution. *Radiant Research*, 98(3):456–470, 1984.
8. C. C.M. Parafba, C. A.R. Diniz, and R. M. Pires. Bayesian analysis and diagnostic of overdispersion models for binomial data. *Brazilian Journal of Probability and Statistics*, 29(3):608–639, 2015.
9. R. L. Prentice. Binary Regression Using an Extended Beta-Binomial Distribution with Discussion of Correlation Induced by Covariate Measurement Errors. *Journal of the American Statistical Association*, 81(394):321–327, 1986.
10. Stan Development Team. Stan Modeling Language Users Guide and Reference Manual, 2017.
11. Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, 2017.
12. Sumio Watanabe. A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, 14(1):867–897, 2013.
13. D. A. Williams. The Analysis of Binary Responses from Toxicological Experiments Involving Reproduction and Teratogenicity. *Biometrics*, 31(4):949–952, 1975.

Fitting Heavy Tail Distributions With Mixture Models

Jorge Basilio^{1;3} and Amilcar Oliveira^{1;2}

¹ Universidade Aberta, Lisboa, Portugal
(E-mail: 1600751@estudante.uab.pt)

² CEAUL - Centro de Estatística e Aplicações da Universidade de Lisboa, Portugal

³ National College of Ireland, Dublin, Ireland

Abstract. The normal probability distribution as assumption for financial returns have been recognized as inappropriate, and a source of inaccurate estimation of Value at Risk. Empirical evidence also have been shown that financial returns shows a more accentuated leptokurtic distribution when compared with a Normal distribution and also skewed. This is usually a cause of underestimated values of VaR , specially when the quantiles are very low. Therefore it is necessary to focus on the tail of the distribution and identify models to fit that behavior. We will highlight the differences between the quality of fitting in the tails of the distribution and the fitting for all the distribution.

This work compares and interprets the results obtained by applying mixture models as a method to estimate the behavior on the extremes for heavy tail data distributions. This results will be then used to describe an analytical solution of VaR under mixture models.

Keywords: Mixture Models, Extreme Values, VaR , Risk Analyses .

1 Introduction

Extreme value theory is used to model unusually low or high value data that is observed in the tail of the distribution.

These unusual events represented by extreme data points are often complex to model and requiring advanced techniques to fit a distribution that includes the heavy tail with satisfactory results.

In statistics the concept of mixture distribution concerns the combination of two or more distributions.

Mixture distributions are of importance in order to model complex processes allowing for a more flexible approach than using a single distribution.

Gaussian mixtures are a possible choice to model this type of complex processes and are formed by linear combinations of two or more Gaussian distributions as a weighted sum of that Gaussian distributions in order to form a new distribution.

One promising use case to apply mixture models is to model heavy tailed distributions. Even though financial returns are usually modeled as normally distributed this assumption proved to be inconsistent with empirical evidence.

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST

Asset returns are considered heavy tailed which means that extreme values are more likely to happen in practise than suggested by a normal distribution. This discrepancies could cause estimation errors of major impact if for example we are using those assumption to build risk measures such as *VaR*. Then normality assumption can lead to inappropriate risk management measures [10].

Some research work involving mixture models assumes that the several distributions are drawn from the same probability density functions. On other hand heterogeneous mixture models can be a valid option for modelling more complex phenomenons and improving modelling capabilities.

An option is to model asset returns distribution with a mixture of distributions, a mixture of normal, a mixture of Gaussian and Gumbel and also Gaussian and GEV as extreme value distributions.

The modelling approach presented explores the capabilities of a mixture model in order to fit the financial returns [14].

In this work we will fit the financial returns with different models and analyse the impact of each model to estimate Value at Risk (*VaR*).

2 Value at Risk

VaR is a central tool in risk, asset, and portfolio risk management and it also plays a key role in systemic risk. *VaR* is defined as the maximum loss an asset/portfolio/institution can incur in a given time period at a defined significance level α .

This probability represents a quantile for risk. With a random variable X and a distribution function F that model losses, verified for an asset in a time period. VaR_α is defined as:

$$VaR_\alpha = F^{-1}(1 - \alpha) \tag{1}$$

A *VaR* of d days at $\alpha\%$ significance level means that on $\alpha\%$ of d days, we won't see a loss higher than the *VaR*, but for the $(1 - \alpha)\%$ of times the loss will be higher.

For example, a 1 day *VaR* at 99% confidence level of 5% means that only 1 of every 100 days we will see a loss higher that 5% of the initial capital [9].

The concept of *VaR* becomes central to the study of systemic risk, becoming also a standard concept for the definition of several of the most significant systemic risk measures mentioned in the literature. Yet this method faces challenges dealing with the risk associated with events involving a volatility component with dependencies between extremes values in distinct data sets and modelling extreme values with volatility.

3 Financial Institutions Returns

3.1 The Data Set

Institution financial details frequently are not available on the public domain and are informed only on a periodic basis, so this methodology is an option

to obtain *VaR* based on market public data. As an assumption, the market value, the market capitalization of each institution reflects the book value of the assets.

The process could be summarised as follows:

- Obtain the data:
 - collect stock prices, we will use weekly based stock prices (Friday's price)
- market value of equity (MVE)
 - stock price \times shares outstanding
- Assume market value of assets (MVA)
- getting returns as: $X_t^i = \frac{(MVA_t^i - MVA_{t-}^i)}{MVA_{t-}^i}$, for each financial institutions

Based on the list of banks that are part of the *STOXX Europe 600* Banks index, information on daily quotations for each title, public available for consultation at <https://finance.yahoo.com/>.

3.2 Normal distribution assumption

There are an open ongoing discussion about the application of normal distribution to model financial related data.

In fact, the use of of normal distribution in order to model financial returns is considered a traditional assumption in finance since Markowitz developed his portfolio theory in 1952, as it is the backbone of traditional (mean-variance) premise [4].

Despite that, some recent research work have been rejecting this assumption based on the study of skewness, kurtosis and in special in heavy tail in the distribution of financial returns [8].

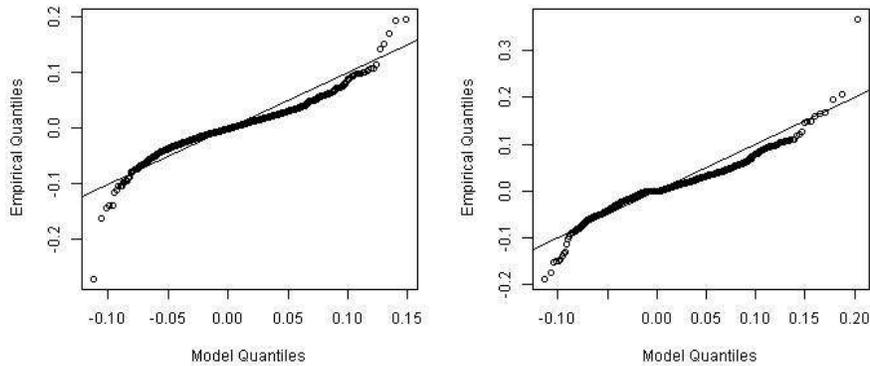
Even though is noticeable that financial returns distributions are at least close to a bell shaped curve, even if this does not translate directly for a normal distribution.

By using the graphs bellow, we can notice a bell shape pattern on the returns of financial institutions. We also can notice an heavy tail behavior, best described in this case, in terms of extreme tails by the *t*-Student curve than by the normal curve.

An additional complementary graphical analysis in order to compare the behavior of the returns across the distinct quantiles and compare that behavior with the expect behavior for a Normal population is to use a Q-Q plot.

The Q-Q plot, or quantile-quantile plot, is a type of graph that allow us assess if it is plausible to assume that a data set came from some theoretical distribution such as a Normal. Even if it is just a visual check, and somewhat subjective, it allows us to see at-a-glance if our assumption is plausible, and how the assumption is eventually violated and which data points cause that violation.

By applying a simple normality test to financial institution returns series, it seems to show up very clear that we should strongly consider other options to model the financial institution returns.



(a) HSBA returns Q-Q plot vs normal distribution
 (b) BBVA returns Q-Q plot vs normal distribution

Lets start by applying some conventional statistical tests for normality such as Shapiro-Wilk’s test and Kolmogorov-Smirnov (K-S) normality test.

Financial Institution	W statistic	$K-S$ statistic	W p-value	$K-S$ p-value
HSBA	0.94344	0.06652	2.2e-16	0.00017
BBVA	0.93228	0.07789	2.2e-16	5.249e-06
BARC	0.77149	0.11839	2.2e-16	2.565e-13

Table 1: Normality testing

Based on the above results, we should consider the financial institutions returns (and financial system returns too) as not normally distributed.

However it is also know that normality tests are in fact very sensitive to what happens on in the extreme tails. This fact can then restrain all the conclusions based on those type of tests.

As we have a relatively large sample of data on our data set it will also worthwhile try a visual approach to investigate normality. Lets then compare the histogram of returns for some financial institutions and the system.

It become also clear we have a different behavior on extreme tails of returns distribution, and in the tail it is not following a normal behavior. This pattern must be included in the modelling.

3.3 Heavy Tails Distributions

In statistics the term heavy tail is associated to distributions with a relatively height probability of extreme outcomes.

Even though there is not a definitive and formal definition of heavy tail usually it is assumed that a distribution has a heavy tail when the probability in the tail is thicker when compared with a normal distribution.

Taken as an example Cauchy distribution we can notice a thicker tail in Cauchy when compared with a normal.

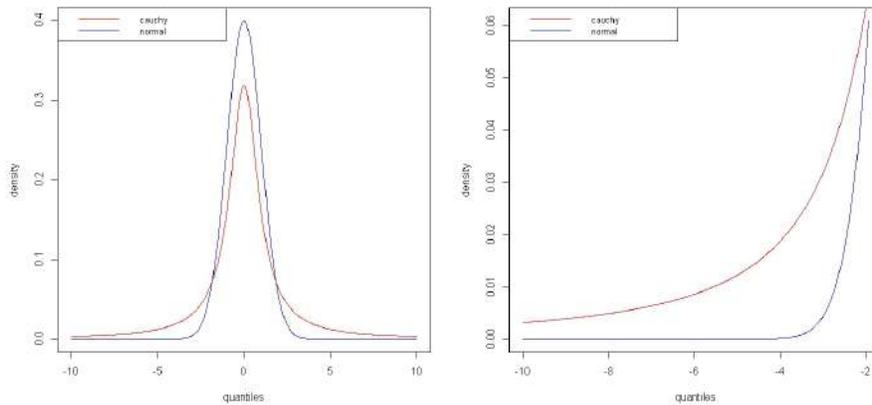


Fig. 2: Heavy tail distribution versus normal

Cauchy among others, like t -Student for example are known as heavy tailed distributions.

In order to close the gap in terms of modelling the extreme tail of financial institution returns as a first approach we will study an approach of univariate fitting where we will include the modelling of the tails of the distribution.

To evaluate how heavy tails impact our financial data set, we will compare results from fitting three theoretical distributions: Normal, Cauchy and t -Student.

3.4 Cauchy distribution

In probability theory, Cauchy distribution is the probability distribution whose probability density function is defined by:

$$f(x) = \frac{1}{\pi(1 + x^2)}$$

with $x \in \mathbb{R}$ in Cauchy standard form, when its mean is 0, and the first and third quantiles are 1 and -1 respectively.

Cauchy distribution then is defined as any distribution that belongs to this family and, if X is a random variable with standard Cauchy distribution then, let $\mu \in \mathbb{R}$ be an arbitrary value and $\sigma > 0$. The random variable Y , defined as:

$$Y = \mu + \sigma X$$

also follows a Cauchy distribution with median μ and whose first quantile is $\mu - \sigma$ and the third quantile is $\mu + \sigma$. The probability density function is therefore defined as:

$$f(y) = \frac{1}{\pi\sigma\left(1 + \frac{(x-\mu)^2}{\sigma^2}\right)}$$

Standard Cauchy distribution can also be defined as a ratio of two normal distributions. Let X and Y be two independent random variables. If $X \sim N(0, 1)$ and $Y \sim N(0, 1)$ then:

$$\frac{X}{Y} \sim \text{Cauchy}(0, 1)$$

In addition to its application in physics, Cauchy distribution is commonly used in models in finance to represent deviations in returns from the predictive model [7]. The reason for this is that practitioners in finance are wary of using models that have light-tailed distributions as Normal, on their returns, and they generally prefer to go the other way and use a distribution with very heavy tails as Cauchy. The history of finance has a vast record of catastrophic predictions based on models that did not have heavy enough tails in their distributions. The Cauchy distribution has sufficiently heavy tails as its moments does not exist, and so it is an ideal candidate to give an error term with extremely heavy tails [5].

3.5 *t*-Student

Also *t*-Student is an option to deal with heavy tails and have been an option for researchers too [16].

The *t*-Student distribution can be defined as a variable $Z \sim N(0, 1)$ and a variable $W \sim X_v^2$, then the standardized quotient of the two follows a *t*-Student distribution with v degrees of freedom:

$$T = \sqrt{v} \cdot \frac{Z}{W} \sim t$$

t-Student distribution could be very useful for financial analysis as we can adapt to the tail behavior of the data. In its conventional form, *t*-Student could not be a very flexible model because of the absence of a location and a scale parameter. An alternative definition can then be described as:

$$\text{If } T \sim t_v \implies S = \mu + \lambda T \sim t_v(\mu, \lambda^2)$$

with μ as location parameter and $\lambda^2(\frac{v}{v-2})$ as scale parameter. The tail decay is therefore polynomial, that is, the density function goes to zero proportional to $x^{-(v+1)}$ for $x \rightarrow \infty$. For low values of v this is a much slower rate than for the Gaussian [17].

Since the typical assumption involving the Gaussian distribution had failed and is now hardly accepted due to the probabilities at the extremes are much larger than those supported by Gaussian distributions. This invalid assumption is specially dangerous for risk management related application.

4 Best fit for returns

If it has been established that financial returns shows heavier tail and peaked than Gaussian it was not yet established which distribution better fit those financial return, remaining as research field. In the process of fitting the data to a theoretical distribution, one could found that usually more than one distribution would be of interest.

The importance of correctly model financial returns and identify distributions that are able to adjust and fit financial return series have been putted at the spotlight every time a new financial crises is faced. Recent crises were not exception and it become even more clear that the typical assumption with Gaussian distributions was anymore acceptable.

The financial return series, like the example below, has phases where it exhibits different volatility. While small returns occur more frequently, there are phases were positive and negative returns are persistently larger as well. This is mentioned as volatility clusters and is understood as characteristic of financial data [18].

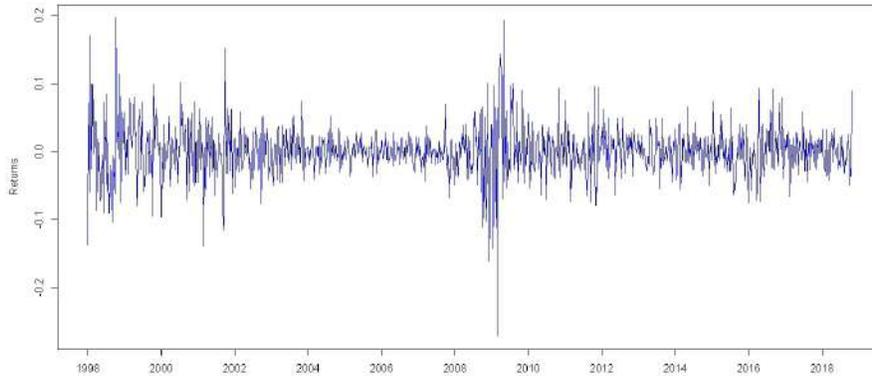


Fig. 3: HSBA weekly returns

However, when working with longer return series, we neglect the exact times of occurrence of each return, and focus on the distribution of returns only.

Starting with HSBA returns series as example, and fitting this series, after normalization to a the three above mentioned distributions: Normal, Cauchy and t -Student, the resulted fit are as showed in the figure:

By visually analyse this results one can notice some differences in the quality of the adjustment, depending on the quantile we take. Due to this fact it becomes harder to identify a single distribution that performs well across all quantiles.

In this particular case, if we take the global results of the fit, we can see that even though, Normal and Cauchy have clear distinct shapes, in terms of

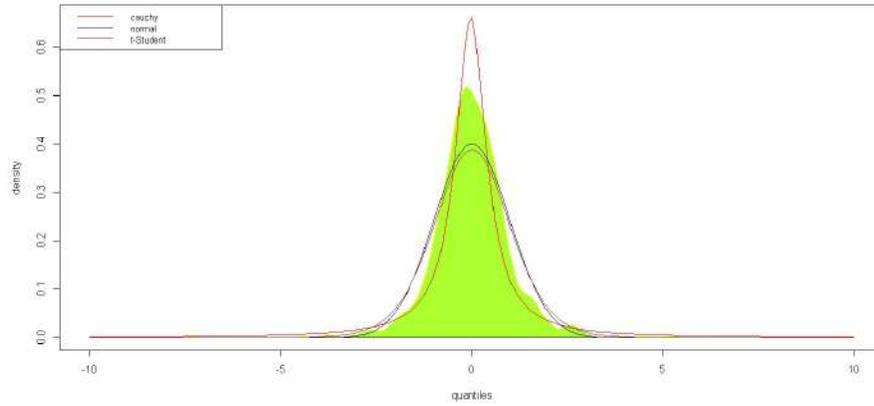


Fig. 4: Fit of HSBA weekly returns

Goodness-of-fit statistics	Cauchy	Normal	<i>t</i> -Student
Kolmogorov-Smirnov	0.0552	0.0652	0.0739
Cramer-von Mises	0.7421	1.6626	2.2699
Anderson-Darling	8.8937	10.2788	13.9965
Goodness-of-fit criteria			
BIC	3043.091	3013.615	2931.064

Table 2: Goodness-of-fit

Goodness-of-fit it turns that the results are quite similar only with a short advantage over Cauchy. However our concerns are more with the behavior on the tail of the distribution, and the following image can proportionate a closer look at the more extreme quantiles in detail:

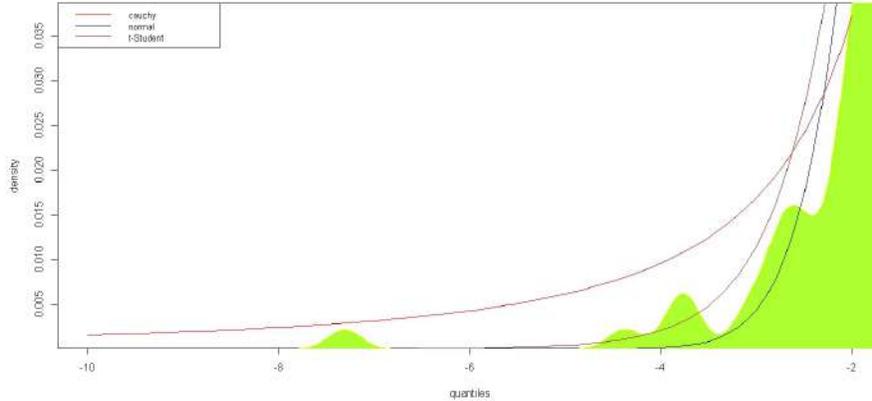


Fig. 5: Fit of HSBA weekly returns

Using the image above, is now easier to identify some limitation on this approach to model extreme values in the tails. There a relevant density on returns series that is not captured by Normal (neither by t -Student), and Cauchy estimate it far for excess.

4.1 Modelling financial returns with a mixture of normal distributions

As empirical evidence and results suggests, the normality assumption of financial institution returns is not verified as it is heavy tailed, and modelling this behavior using only one distribution have shown limitation, one option is consider an approach that uses more than one distribution.

Heavy tailed distributions can be modeled instead by a mixture of distribution. In this case we will first approach the problem by applying a mixture of normal distributions.

Assuming the returns are following a stochastic process for a financial institution i as:

$$R_{it} = \lambda_{it}R_{it}^{\alpha} + (1 - \lambda_{it})R_{it}^{\beta} \quad (2)$$

where $R_{it}^{\alpha} \sim N(\mu_{\alpha}, \sigma_{\alpha})$, $R_{it}^{\beta} \sim N(0, \sigma_{\beta})$, and λ_{it} is 1 with probability p and 0 otherwise.

These three random variables R_{it}^{α} , R_{it}^{β} and λ_{it} are independent each other.

Depending on λ and with probability p the distribution to apply will be $N(\mu_{\alpha}, \sigma_{\alpha})$, for example for the most normal situations. With probability $(1 - p)$, λ will be equal to 0 and the distribution to apply is $N(0, \sigma_{\beta})$ and it could be interpreted as an exceptional case.

The challenge now is with the estimation of the parameters involve; $p, \mu_{\alpha}, \sigma_{\alpha}, \sigma_{\beta}$.

Despite several alternative methods that are possible to use to estimate the parameters of a mixture of normal distribution, if we consider the traditional maximum likelihood method, it could then be formulated as:

$$l((p, \mu_\alpha, \sigma_\alpha, \sigma_\beta) | R_{it}) = \sum_t \log \left[\frac{p}{\sigma_\alpha} \exp\left(-\frac{(R_t - \mu_\alpha)^2}{2\mu_\alpha^2}\right) + \frac{1-p}{\sigma_\beta} \exp\left(-\frac{(R_t^2)}{-\mu_\beta}\right) \right]$$

Due the existence of both poles and saddle points, the maximization of the mixture of normals likelihood could be challenging and the global maximum for that function could not exist [6].

This problem however could be described as an incomplete data problem since the data we observe in our sample can be viewed as a subset of the “complete” data.

4.2 Expectation-Maximization Algorithm

The Expectation-Maximization (EM) Algorithm is an appropriate tool for that type of problems. EM Algorithm is an approach for maximum likelihood estimation in the presence of latent variables and can be used to predict the latent variables values with the condition that the general form of the probability distribution governing those latent variables is know.

The algorithm is implemented as an iterative procedure given a set of incomplete data and considering a set of starting parameters will iterate on two steps

- Expectation step (E – step). Using the available observed data and the current model parameters the missing or latent variables are estimated by.
- Maximization step (M – step). After estimate missing values, this step will be used to update the parameters by compute the parameters that maximize the expected log-likelihood of the model based on the values estimated on E-step.

EM Algorithm includes statistical considerations to compute the maximum-likelihood (ML), source distribution that would have created the observed data, including the effects of counting statistics. Specifically, it assigns greater weight to high-count elements of a profile and less weight to low-count regions [3].

4.3 EM Algorithm and mixture of Gaussians

Taken the case of a mixture of Gaussians let $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ be a sample of n i.i.d. observations of a mixture of two Gaussian and $\mathbf{z} = (z_1, z_2, \dots, z_n)$ the latent variables that determine the component where the observation originates[15].

$$X_i | (Z_i = 1) \sim \mathcal{N}_d(\boldsymbol{\mu}_1, \Sigma_1) \quad \text{and} \quad X_i | (Z_i = 2) \sim \mathcal{N}_d(\boldsymbol{\mu}_2, \Sigma_2),$$

where

$$P(Z_i = 1) = \tau_1 \quad \text{and} \quad P(Z_i = 2) = \tau_2 = 1 - \tau_1$$

The goal of this process is to estimate the parameters for the mixture of Gaussians:

$$\theta = (\boldsymbol{\tau}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$$

The likelihood function therefore is:

$$L(\theta; \mathbf{x}, \mathbf{z}) = \exp \left\{ \sum_{i=1}^n \sum_{j=1}^2 \mathbb{I}(z_i = j) \left[\log \tau_j - \frac{1}{2} \log |\boldsymbol{\Sigma}_j| - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_j) - \frac{d}{2} \log(2\pi) \right] \right\}.$$

The result of the application of EM Algorithm to HSBA financial returns come as a mixture of two normal distributions:

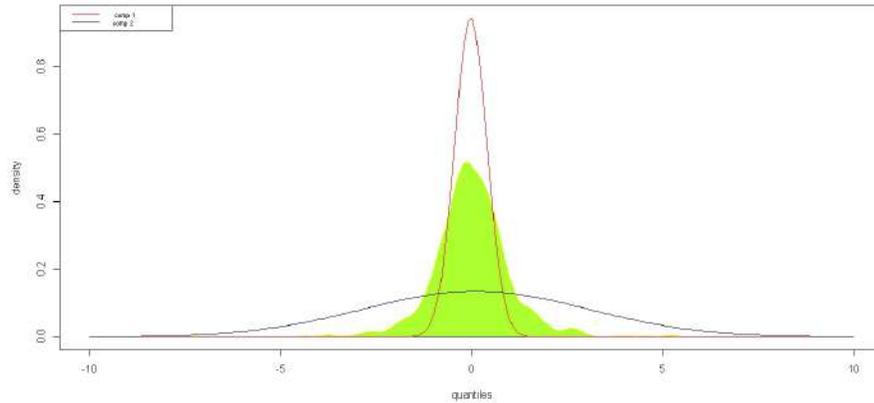


Fig. 6: Fit of HSBA weekly returns: the two components of the mixture

In order to represent correctly the joint density of the mixture distribution we have to build the mixture distribution according to τ parameter as estimated by EM Algorithm. In this case the estimated τ value was 0.761.

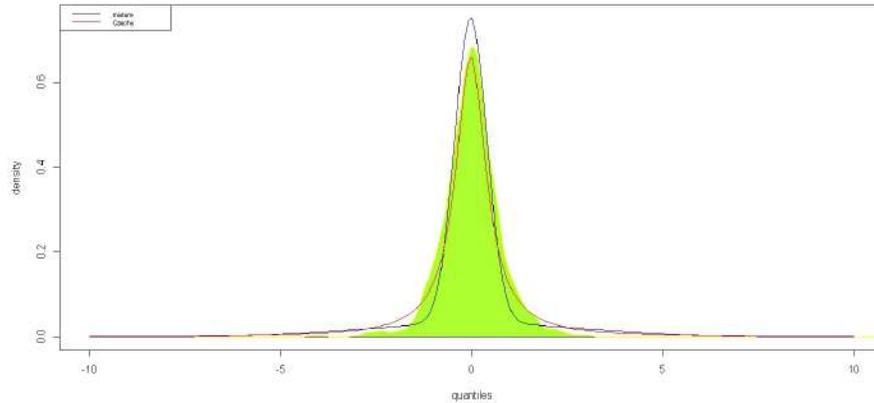


Fig. 7: Fit of HSBA weekly returns with a mixture of Normal

With the two normal mixture model we obtained a BIC criterion of 2871.93, what makes this fit slightly better than the fit provided by Cauchy, Normal and *t*-Student, according to BIC criterion. The results on the tail still not being the as good as desirable,

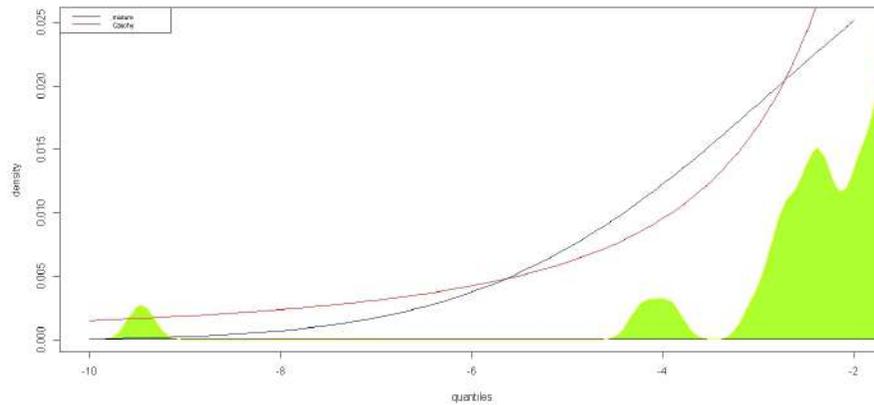


Fig. 8: Fit of HSBA weekly returns with a mixture of Normal

4.4 Extreme Value Mixture Models

The idea behind the use of Mixture of Extreme Value distribution is to combine the flexibility of using a distribution to capture the main component (the bulk distribution), that could be for example a Normal, and also the tails, as extreme

values. With this mixture model one will get an entire distribution function by splitting the distribution in a bulk component and a tail components.

There are several approaches that consider only one tail and also approaches that consider both upper and lower tail. In this case the mixture function will be compounded potentially by a mixture of distribution from distinct families.

In our case we are specially interested in explore a mixture of a Normal distribution as bulk distribution with two Gamma tail distribution in both upper and lower tail. MacDonald et al. (2011) [12] proposed a two tailed mixture model where the standard kernel density estimator is spliced with two extreme value tail models.

This model uses a kernel density estimators to estimate the non-extreme value distribution and Generalized Pareto distribution (GPD) to estimate the tail distribution. A boundary-corrected kernel density estimator is also used in the case of a population with bounded support. This kernel density estimator assumes a particular kernel, in this case the normal density, which is centered at each data point, and uses only one parameter to define bandwidth. The model uses also the standard cross-validation likelihood to define bandwidth, combined with the likelihood for the peaks over threshold tail model, to give a full likelihood for all of the observations. The term tail fraction refers to the proportion of the distribution above the threshold. This parameter will be identified by Φ_u and u represents the threshold.

The distribution function comes as:

$$F(x|\Theta) = \begin{cases} \phi_{ul}1 - G(-x| - u_l, \sigma_{ul}, \epsilon_l) & x < u_l, \\ H(x|\mu, \sigma) & u_l \leq x \leq u_r \\ (1 - \phi_{ur}) + \phi_{ul}G(x|u_r, \sigma_{ur}, \epsilon_r) & x > u_r \end{cases}$$

where $\phi_{ul} = H(u_l|\mu, \sigma)$ and $\phi_{ur} = 1 - H(u_r|\mu, \sigma)$ and $H(\cdot|\mu, \sigma)$ is the normal distribution with mean μ and standard deviation σ . $G(\cdot| - u_l, \sigma_{ul}, \epsilon_l)$ and $G(\cdot| - u_r, \sigma_{ur}, \epsilon_r)$ are GPD distributions for lower and upper tails respectively.

By applying the methodology to HSBA returns series we obtained the following estimates for the parameters:

The graph bellow shows the results for a mixture of a normal $N(-0.00084, 0.023)$ bounded at left by parameter $u_l = -0.0354$ and on right by parameter $u_r = 0.025$.

The Gamma parameters obtained are respectively:

left tail	right tail
$\phi_{ul} = 0.120$	$\phi_{ur} = 0.119$
$\mu_l = 0.133$	$\mu_r = 0.0743$
$\sigma_l = 0.593$	$\sigma_r = 0.728$

The goodness of fit for this model is also slight better than previous ones with an BIC criterion value estimated as 2860.661. The advantage and flexibility of this mixture model is essentially in the tails of the distribution as it is

able to take advantage of the capabilities of Gamma distribution to adapt to the tail. The graphs obtained for the extreme value mixture are as follows:

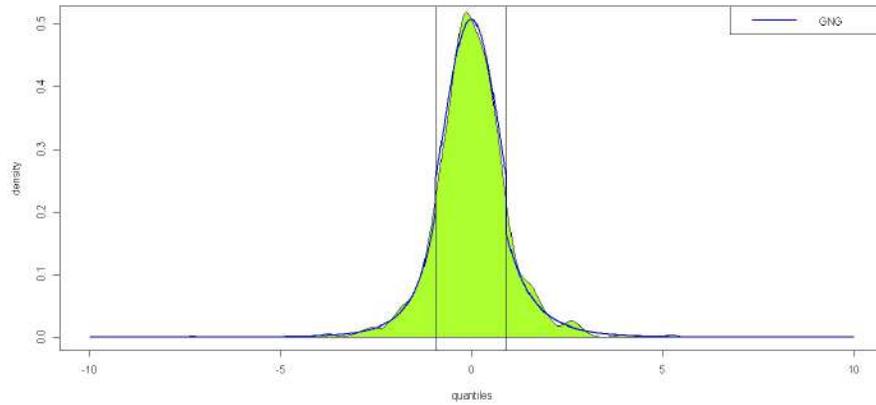


Fig. 9: Fit of HSBA weekly returns with a mixture of Gamma-Normal-Gamma

By visual analyse of the graph obtained with the Gamma-Normal-Gamma (GNG) fitting it is possible to identify a very close adjustment.

Also in the tail of the distribution we can notice a good approximation including a decay of the distribution function when it goes further on the left.

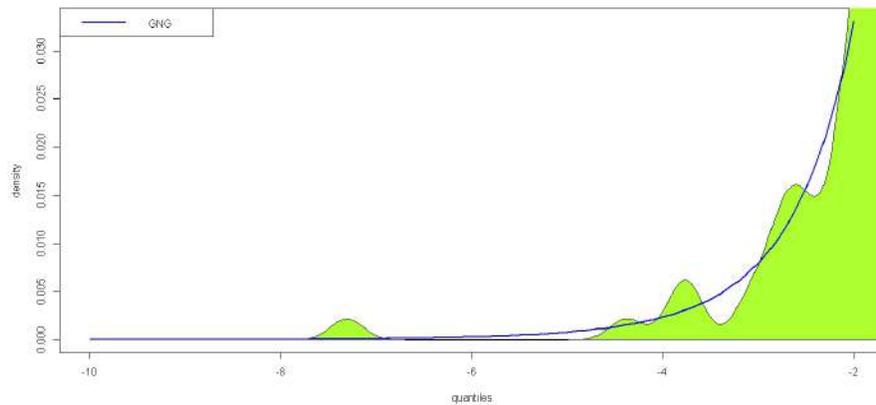


Fig. 10: Fit of HSBA weekly returns with a mixture of GNG on the left tail

5 Value at Risk Estimation

Based on the results we will compare the impact of the different models tested in terms of the estimates obtained for VaR .

Based on the quantiles for each model, Cauchy, Normal, t -Student and extreme value mixture GNG we estimate the VaR for several levels using several α values in a range between 0.001 and 0.1 as showed in the table bellow.

	0.1%	1%	2.5%	5%	10%
Cauchy	-5.7016	0.5702	-0.2278	-0.1133	-0.0554
Normal	-0.1139	-0.0856	-0.0720	-0.0603	-0.0468
t - Student	-0.1541	-0.1023	-0.0823	-0.0667	-0.0503
GNG	-0.1820	-0.0992	-0.0728	-0.0547	-0.0383
Historical VaR	-0.1610	-0.0975	-0.0717	-0.0548	-0.0384

Table 3: VaR estimates

The values showed in the table above represent the percentage of the value of each institution at risk for each one of the risk levels (0.1% to 10%).

In this example is also visible the issue by using Cauchy as model in extreme values as the VaR estimate is too high and value is not reasonable. In general Cauchy based estimates are too excessive when compared with historical VaR value. In this case we can conclude Cauchy consistently overestimates the risk and is not also a good option for risk management application.

6 Conclusion

The results obtained by comparing the goodness of fit obtained by applying distinct statistics modelling techniques, highlighted a concern regarding the quality of the global adjustment versus the quality of the adjustment on the tails of the distribution. In certain applications the analyses of the tail of the distributions is of major importance, as for example in risk analyses. The results obtained for VaR estimates for each model implemented also showed that more complex model could be advantageous as they are more flexible in adapt to the tail of the distribution providing better adjustments.

Complex phenomena requires also more complex models and the complexity of certain phenomena like behavior of financial returns requires more complex and versatile models.

7 Acknowledgements

This research is based on the partial results of the Funded by FCT - Fundação para a Ciência e a Tecnologia, Portugal, through the project UID/MAT/00006/2019.

References

1. Adrian, Tobias and Brunnermeier, Markus K, *CoVaR*, National Bureau of Economic Research, 2011
2. Bingham, Nicholas H and Kiesel, R diger and Schmidt, Rafael and others, *A semi-parametric approach to risk management* Quantitative Finance, 3, 6 , 426-441, Taylor & Francis, 2003
3. Dempster, Arthur P and Laird, Nan M and Rubin, Donald B, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society: Series B (Methodological) 39, 1, 1-22, Wiley Online Library 1977,
4. Fabozzi, Frank J and Gupta, Francis and Markowitz, Harry M, *The legacy of modern portfolio theory*. The Journal of Investing, 11,3, 7-22, 2002
5. Guerrero-Cusumano, José-Luis, *An asymptotic test of independence for multivariate t and Cauchy random variables with applications*. Information sciences, 92,1-4, 33-45,Elsevier, 1996
6. Hamilton, James D, *A quasi-Bayesian approach to estimating parameters for mixtures of normal distributions*. Journal of Business & Economic Statistics, 9, 1, 27-39, Taylor & Francis, 1991
7. Harris, David E, *The distribution of returns*. Available at SSRN 2828744, 2017
8. Jondeau, Eric and Poon, Ser-Huang and Rockinger, Michael, *Financial modelling under non-Gaussian distributions*. Springer Science & Business Media, 2007
9. Kaplanski, Guy and Kroll, Yoram , *VaR risk measures versus traditional risk measures: an analysis and survey*. (Journal of Risk, 2002) volume 4, number 3, pp. 1-27
10. Kimball, Ralph C., et al. *Failures in risk management*. New England Economic Review, 2000, 3-12.
11. Lakshmi, R. Vani, and V. S. Vaidyanathan. *Parameter estimation in gamma mixture model using normal-based approximation*. Journal of Statistical Theory and Applications 15, 1, 25-35, Atlantis Press, 2016
12. MacDonald, A and Scarrott, Carl John and Lee, D and Darlow, B and Reale, Marco and Russell, G . *A flexible extreme value mixture model*. Computational Statistics & Data Analysis, 55, 6, 2137-2157, Elsevier, 2011
13. Padoan, Simone A *Multivariate extreme models based on underlying skew- t and skew-normal distributions*. Journal of Multivariate Analysis, 102,5,977-991, Elsevier, 2011
14. Razzaghi, Mehdi and Kodell, Ralph L *Risk assessment for quantitative responses using a mixture model*. Biometrics 56.2 Wiley Online Library 2000 519-527
15. Reynolds, Douglas A *Gaussian Mixture Models* Encyclopedia of biometrics 741, Berlin, Springer, 2009
16. Terzić, Ivica and others *Empirical estimation and comparison of normal and student t linear var on the belgrade stock exchange*, Sinteza 2014-Impact of the Internet on Business Activities in Serbia and Worldwide 298-302, Singidunum University ,2014
17. Theodossiou, Panayiotis *Financial data and the skewed generalized t distribution*, Management Science, 44, 12-part-1, 1650-1661, INFORMS, 1998
18. Tseng, Jie-Jun and Li, Sai-Ping *Asset returns and volatility clustering in financial time series*, Physica A: Statistical Mechanics and its Applications, 390, 7, 1300-1314 Elsevier, 2011

Automobile Insurance Fraud Detection

Mark Anthony Caruana¹ and Liam Grech²

¹ Department of Statistics and Operations Research, Faculty of Science, University of Malta, Msida, Malta
(E-mail: mark.caruana@um.edu.mt)

² Phoenix GSB, Malta
(E-mail: liamgrech3@gmail.com)

Abstract. The risk of incurring financial losses from fraudulent claims is an issue concerning all insurance companies. The detection of such claims is not an easy task. Moreover, a number of old-school methods have proven to be inefficient. Statistical techniques for predictive modelling have been applied to detect fraudulent claims. In this paper we compare two techniques: Artificial Neural Networks and the Naïve Bayes classifier. The theory underpinning both techniques is discussed and an application of these techniques to a data set of labelled automobile insurance claims is then presented. Fraudulent claims only constitute a small percentage of the total number of claims. As a result, data sets tend to be unbalanced. This in turn causes a number of problems. To overcome such issues, techniques which deal with unbalanced data sets are also discussed. The suitability of Neural Networks and the Naïve Bayes classifier to the data set is discussed and the results are compared and contrasted by using a number of performance measures including ROC curves, Accuracy, AUC, Precision, and Sensitivity.

Keywords: Fraud Detection, Insurance Claims, Artificial Neural Networks, Naive Bayes Classifier, Unbalanced Data sets.

1 Introduction

Insurance fraud refers to a claim made for a loss, where the policyholder (insured) has reported an accident that has not occurred or has exaggerated the extent of the circumstances. Moreover, it ranges from professional fraud, where accidents are planned by organised crime rings, to opportunistic fraud. In the US, the property and causality insurance industry estimates that, each year, 5-10% of claim costs are due to fraud, with almost a third of companies believing that this figure is as high as 20%. In Europe, it is estimated that 10% of all claim expenditure is due to fraudulent claims and particularly, in the UK roughly €2.2 bn is lost each year due to undetected fraud. Insurance fraud is also present in the Maltese Islands. In 2016, an organised automobile insurance fraud ring was uncovered by Maltese police, where twenty two people faced charges of defrauding eight insurance companies of hundreds of thousands of euros by planning traffic accidents and repeating claims under different names between the years 2009 and 2013.

There are a number of different models and algorithms that have been studied for various types of fraud detection. Bolton et al.[3] gave a review of the techniques for statistical fraud detection, while emphasising the need for companies to explore the vast amounts of data that is available to them. Clifton et

al.[4] comment that there is too much emphasis on complex techniques such as Artificial Neural Networks (ANNs), and suggested that faster algorithms such as Naive Bayes (NB) should be explored for fraud detection. Ngai et al. [12] provided an exhaustive overview of the literature for fraud detection, where classification was found to be the most popular method for tackling the fraud detection problem, while Logistic models, Artificial Neural Networks, Bayesian Belief Networks and Decision Trees were observed to be the preferred techniques. For automobile insurance fraud, it was noted that Logistic models were the most popular, followed by Artificial Neural Networks, Naive Bayes, Bayesian Belief Networks and Probit models respectively. The aim of this paper is to compare the performance of Artificial Neural Networks and the Naive Bayes classifier on automobile insurance fraud detection.

The rest of the paper is structured as follows: in section 2 we introduce some basic notation, in sections 3 and 4 the theory behind Neural Networks and Naive Bayes is discussed, in section 5 we apply the two techniques to a data set and compare the results, finally, section 6 contains some concluding remarks.

2 Context

Let $\mathbf{X} = (X_1, \dots, X_p)$ be a random vector where each random variable X_i represents an attribute of a claim (e.g. age of policy holder, past number of claims, presence of a police report, etc..) and can be continuous or discrete. Moreover, let \mathbf{x} denote an observed value of the random vector \mathbf{X} , which can also be referred to as the input of a given model. Let t denote what is known as the target variable. An observation can belong to one of K classes, represented by the class labels $\mathcal{C}_0, \dots, \mathcal{C}_{K-1}$. C shall denote the random variable representing a general class label \mathcal{C}_k . The goal is that of formulating a classifier that can assign observations to the correct class by applying a classification rule. Note that $P(C = \mathcal{C}_k | \mathbf{X} = \mathbf{x})$ denotes the posterior probability of the class \mathcal{C}_k . Throughout this paper, this may be replaced by $P(\mathcal{C}_k | \mathbf{X} = \mathbf{x})$ for ease of notation. For fraud detection, the aim is to classify a set of claims correctly. When $K = 2$, the scenario is reduced to a *two-class* problem which is appropriate for fraud detection. Therefore, the target variable t corresponds to the ‘fraudulent’ class \mathcal{C}_1 when $t = 1$ and the ‘legitimate’ class \mathcal{C}_0 , when $t = 0$. We will assume that we are in possession of the data set which contains n observations. A section of the this data set, which contains N observations will be called the *training set* and will be used to fit a model to this data set. The quality of the model is then assessed using the other section of the data set consisting of $n - N$ observations. This is called the testing set.

3 Artificial Neural Networks

Neural networks (NN) consist of one input layer and one output layer, with an arbitrary number of hidden layers. Within each hidden layer we have an arbitrary number of hidden units and a number of continuous activation functions

to approximate a continuous mapping between the input and output space. In this paper we will primarily consider a neural network with a single hidden layer. Figure 1 below illustrates a NN with a single hidden layer and an output layer.

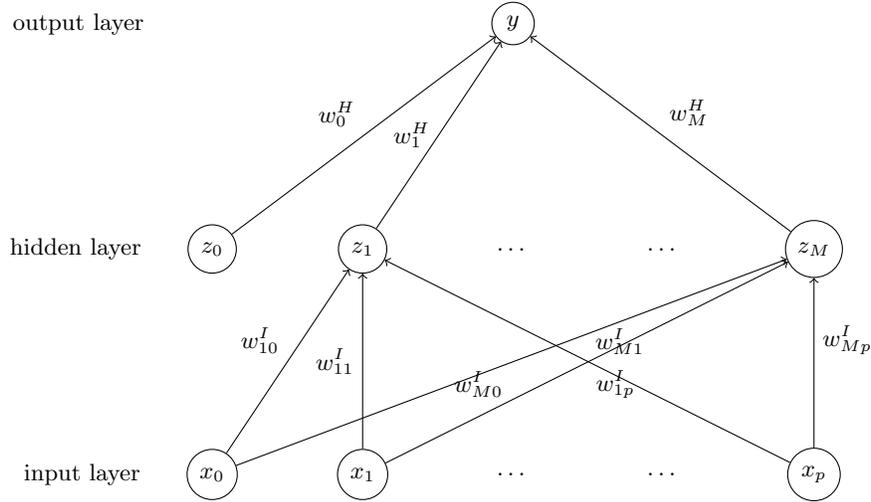


Fig. 1. Neural Network with one hidden layer and one output unit

In our context, the output y is a function of the input variables x_1, \dots, x_p represented by the vector \mathbf{x} and all the weights (parameters) of the network are represented by the vector \mathbf{w} . The nodes in the input layer represent the p variables of some observation from the training set. The first step is to form M weighted sums in terms of the input variables, where the j^{th} denoted by a_j is expressed as:

$$a_j = w_{j0}^I + \sum_{i=1}^p w_{ji}^I x_i = \sum_{i=0}^p w_{ji}^I x_i \quad (1)$$

for $j = 1, \dots, M$ and the weights w_{ji}^I are associated with the links between the input layer and the first hidden layer. The bias term w_{j0}^I is absorbed into the summation and is given a fixed dummy input value of 1. The subscript ji implies that the weight belongs to the link that starts from the i^{th} node in the input layer and ends at the j^{th} node in the hidden layer. The hidden units (activations) of the hidden layer are denoted by z_j and are calculated by applying an activation function h^H for the hidden layer to (1) as follows

$$\text{Hidden Layer : } z_j = h^H \left(\sum_{i=0}^P W_{ji}^I x_i \right) \quad (2)$$

Should one require, it is possible to insert more hidden layers. However, in this paper we will primarily concentrate on NN with only one hidden layer.

For the case of one output node, the output y can be expressed as

$$\text{Output layer (1 unit) : } y = h^O \left(\sum_{s=0}^S w_s^H z_s \right) \quad (3)$$

where one weighted sum has been constructed in this case and w_s^H denotes the weights between the hidden layer and the single output unit. In general, the activation functions for hidden layers and for the output layer need not be the same and the choice of function depends on a number of factors, such as the type of data or the problem to be solved by the network.

In literature one finds a number of activation functions. However, the most frequently used are the Logistic sigmoid functions and the hyperbolic tangent. The former is usually denoted by σ and is defined as follows:

$$\sigma(a) = \frac{1}{1 + \exp(-a)}, \text{ for } a \in \mathbb{R} \quad (4)$$

while the latter is defined as follows:

$$\tanh(a) = \frac{\exp(a) - \exp(-a)}{\exp(a) + \exp(-a)}, \text{ for } a \in \mathbb{R} \quad (5)$$

These activation functions are used often because they are both continuous and differentiable functions. Hence they fully satisfy the assumptions of the theorem below which is central in the theory of Neural Networks.

Theorem 1 (Universal Approximation theorem).

Let h be a continuous sigmoidal function. Then finite sums of the form

$$G(\mathbf{x}) = \sum_{j=1}^M \alpha_j h(\mathbf{w}_j^T \mathbf{x} + w_{j0})$$

where $\mathbf{x}, \mathbf{w}_j \in \mathbb{R}^P$ and $\alpha_j, w_{j0} \in \mathbb{R}$ are fixed, are dense in $C[I^P]$. That is, given any $f \in C[I^P]$ and $\varepsilon > 0$, $G(\mathbf{x})$ exists such that

$$|G(\mathbf{x}) - f(\mathbf{x})| < \varepsilon \quad \forall \mathbf{x} \in I^P$$

The proof of this theorem can be found in Cybenko [5].

It is important to note that the implication of this theorem is that a NN with a single hidden layer and continuous sigmoidal activation functions in the hidden layer, can be shown to be mathematically sufficient to give an approximation of the required non-linear mapping; however, an exact representation is not attainable. As discussed in Hornik [11], the interpretation of this is that a model with a single hidden layer will only be unsuccessful if not enough hidden neurons are used or if the target variable does not depend on the given data in the first place. Also the output for the theorem is taken as a single linear output; however, this can be generalised to sigmoidal output activation functions as well as multiple output nodes as discussed in Ripley [13], and Hassoun [9].

Now that we have defined with a certain degree of clarity a Neural network and its components, we will next discuss how the weights can be estimated by using the data in the training set. The algorithm used is commonly known as *error back-propagation*. This supervised learning technique makes use of the available training set $\mathcal{D} = \{(\mathbf{x}^{(n)}, t^{(n)})\}_{n=1}^N$ in order to obtain an error value or loss for each observation with respect to its target value, which shall be used to construct an error function with respect to the weights and biases of the model. This method sheds light on why so much emphasis was made on the differentiability of the activation functions in each layer, since differentiation plays an important role in the algorithm. In this algorithm the following terms will be used:

- **Input Signals:** The attributes $(x_1^{(n)}, \dots, x_p^{(n)})$ of the n^{th} observation $\mathbf{x}^{(n)}$ are called input signals.
- **Output Signals:** The output for a network with one output unit gives an output signal $y^{(n)}$.
- **Forward Pass:** The input signals are propagated forward through the network, from the input layer, to the hidden layers and finally reaching the output layer.
- **Loss function:** The output signals obtained from the forward pass and the corresponding target variable for that observation are used to obtain the loss function which shall be denoted by L and is discussed below.
- **Error Signal & Backward Pass:** The loss function L obtained after the forward pass is then differentiated with respect to the weights and biases between the final hidden layer and the output layer and by means of the error back-propagation algorithm discussed below, the gradients (error signals) are said to be propagated backwards through the network.
- **Error Function:** This shall be denoted by E and will be discussed below.

As a consequence of theorem 1, a NN with one hidden layer and a finite number of hidden units is sufficient to solve a number of problems. This can be extended by adding more hidden layers, each consisting of an arbitrary number of hidden units and continuous non-linear activation functions in order to approximate a continuous mapping between the input space and output space. Therefore if in general an error function E is expressed in terms of the output signals of the network, which in turn can be expressed in terms of the input signals and parameters of the network, then E can be expressed simply as a function of the parameters of the model. Recall the target variable for binary classification $t \in \{0, 1\}$ for one output node. Generally for the output of a NN, the error function can be expressed as

$$E(\mathbf{w}) = E(y(\mathbf{x}, \mathbf{w}), t) \tag{6}$$

where $\mathbf{w} = (w_1, \dots, w_W)$ is a vector of all the parameters of a network consisting of W weights. This representation of the error function suggests that the output signals and the target variable shall be used to calculate a form of error of the model, while keeping in mind that the aim shall be to minimise this error with respect to the parameters of the network.

Let $\nabla E(\mathbf{w})$ denote the gradient of the error function such that

$$\nabla E(\mathbf{w}) = \begin{bmatrix} \frac{\partial E(\mathbf{w})}{\partial w_1} \\ \frac{\partial E(\mathbf{w})}{\partial w_2} \\ \vdots \\ \frac{\partial E(\mathbf{w})}{\partial w_W} \end{bmatrix} \quad (7)$$

The objective function may have local minima as well as a global minimum at which $\nabla E(\mathbf{w}) = 0$ is satisfied since it is a non-convex function. Therefore, there is not a unique value for \mathbf{w}^* that satisfies $\nabla E(\mathbf{w}) = 0$, which corresponds to a global minimum of the objective function. This is an important property to take into consideration when choosing a suitable algorithm for adjusting the parameters of the network.

The method of *error back-propagation* provides an efficient algorithm for calculating the derivatives of the error function with respect to the weights and depends heavily on the properties of the Neural Network. Bishop [2] considered the output of the model as an overall network function and by implementing differentiable activation functions at each layer of the network, it is possible to differentiate this output function with respect to the weights and biases. If the error function is expressed in terms of the output function, then it is also differentiable with respect to the weights and biases, which allows us to use the gradient descent method. Consider a NN with an arbitrary number of hidden layers and a single output node for the two-class problem. Consider a general differentiable error function E in terms of the output function y and the target variable t . The derivation shall focus on the loss function L . For the n^{th} observation, this can be defined as

$$L = - \left(t^{(n)} \ln y^{(n)} + (1 - t^{(n)}) \ln (1 - y^{(n)}) \right) \quad (8)$$

Since it can be shown that E can be expressed as the sum of L over all observations of the training set through the equation:

$$E = \frac{1}{N} \sum_{n=1}^N L(y^{(n)}, t^{(n)}) \quad (9)$$

and each observation completes a forward pass and a backward pass through the network. It can be shown that the derivatives for the overall error function E with respect to the weights of the output layer for a network with a single output unit can be given as

$$\frac{\partial E}{\partial w_s} = \sum_{n=1}^N \frac{\partial L(y^{(n)}(\mathbf{x}^{(n)}, \mathbf{w}), t^{(n)})}{\partial w_s} \quad (10)$$

while for the hidden layer this can be expressed as

$$\frac{\partial E}{\partial w_{sc}} = \sum_{n=1}^N \frac{\partial L(y^{(n)}(\mathbf{x}^{(n)}, \mathbf{w}), t^{(n)})}{\partial w_{sc}} \quad (11)$$

Finally, to obtain \mathbf{w}^* one may use techniques such as the stochastic gradient descent method.

4 Naive Bayes Classifier for Categorical Attributes

In this section we assume that the attributes X_1, \dots, X_p are conditionally independent of each other. Under this assumption, Bayes' theorem can be applied to obtain

$$P(C_k | \mathbf{X} = \mathbf{x}) = \frac{q(C_k) \prod_{i=1}^p P(x_i | C_k)}{P(\mathbf{x})} \quad (12)$$

or alternatively, since $P(\mathbf{x})$ is the same for all classes, it can be considered as a constant of proportionality such that (12) becomes

$$P(C_k | \mathbf{X} = \mathbf{x}) \propto q(C_k) \prod_{i=1}^p P(x_i | C_k) \quad (13)$$

This assumption is considered to be 'naive' since in practice this is not always the case. More often than not, there may in fact be correlations between the attributes such that this conditional independence does not hold. Nonetheless, it has been shown by Domingos and Pazzani [6] and Hand and Yu [8] that despite the presence of dependencies between the attributes, the NB classifier can perform as well as other more complex classifiers in terms of classification accuracy.

The training phase for the NB classifier corresponds to maximum likelihood estimation of the required parameters. The required parameters in this case are the prior probabilities for each class and the univariate marginal class-conditional probabilities for each observation and each attribute of the training set \mathcal{D} . The form of the class-conditional probabilities depends on the type of variables making up the training set. Since the available data consists only of discrete-valued variables, it is enough to consider the cases when the attributes are categorical, i.e. where an attribute $X_i \in \{1, \dots, M\}$. The NB model can be extended for continuous attributes by considering a Normal distribution for the class-conditional probabilities as discussed in Heckerman and Geiger [10] or by discretising the values as discussed in Dougherty et al. [7]; however, this is not required for the available data since the attributes for the claims are all discrete-valued.

Consider the case when an attribute X_i is a categorical variable such that the attribute can take $M > 2$ possible values, i.e. $X_i \in \{1, \dots, M\}$. Many algorithms (such as NN) handle categorical attributes by using a one-of- M coding scheme such that the single variable is transformed into M binary attributes, one for each possible value of the categorical attribute. For example, if a categorical variable can take values $\{1, 2, 3\}$ and for a specific observation the observed value is 2, then this is coded as $(0, 1, 0)$, i.e. one unity and the rest zero. However as stated in Barber [1], the issue with this coding scheme is that it violates the independence assumption of the NB model, since the three derived attributes are clearly dependent. Therefore, for the NB classifier, multi-level factors shall not be coded using a one-of- M coding scheme, but by the usual coding scheme. Since a categorical variable X_i can take M possible values, the likelihood that, conditioned on the class, the observed attribute x_i

takes a value m from the M possible values is denoted as

$$P(X_i = m | t) = \pi_{it}^{(m)} \quad (14)$$

such that $\sum_{m=1}^M P(X_i = m | t) = 1$.

Consider that there are $p_2 \leq p$ categorical attributes in the given training set $\mathcal{D} = \{(\mathbf{x}^{(n)}, t^{(n)})\}_{n=1}^N$ where p is again the total number of attributes in the training set. Conditioning on the target variable t , each observation is i.i.d. such that the likelihood is given as

$$\mathcal{L} = \prod_{n=1}^N P(\mathbf{x}^{(n)} | t^{(n)}) = \prod_{n=1}^N \prod_{i=1}^{p_2} \prod_{m=1}^M \prod_{t=0}^1 (\pi_{it}^{(m)})^{\mathbb{I}\{x_i^{(n)}=m\}\mathbb{I}\{t^{(n)}=t\}} \quad (15)$$

and the class-conditional log-likelihood is given as

$$\ell = \sum_{n=1}^N \sum_{i=1}^{p_2} \sum_{m=1}^M \sum_{t=0}^1 \mathbb{I}\{x_i^{(n)} = m\} \mathbb{I}\{t^{(n)} = t\} \ln \pi_{it}^{(m)} \quad (16)$$

Now to obtain the required estimates for $\pi_{it}^{(m)}$, the Lagrange multiplier method may be used as carried out by Barber [1] to ensure that probabilities sum to one such that

$$\ell = \sum_{n=1}^N \sum_{i=1}^{p_2} \sum_{m=1}^M \sum_{t=0}^1 \mathbb{I}\{x_i^{(n)} = m\} \mathbb{I}\{t^{(n)} = t\} \ln \pi_{it}^{(m)} + \sum_{t=0}^1 \sum_{i=1}^{p_2} \lambda_{it} \left(1 - \sum_{m=1}^M \pi_{it}^{(m)} \right)$$

and differentiate with respect to $\pi_{it}^{(m)}$ and equate to 0 to obtain

$$\lambda_{it} = \sum_{n=1}^N \frac{\mathbb{I}\{x_i^{(n)} = m\} \mathbb{I}\{t^{(n)} = t\}}{\pi_{it}^{(m)}} \quad (17)$$

Then the maximum likelihood estimators for $\pi_{it}^{(m)}$ are given as

$$\hat{\pi}_{it}^{(m)} = \frac{\sum_n \mathbb{I}\{x_i^{(n)} = m\} \mathbb{I}\{t^{(n)} = t\}}{\sum_{m', n'} \mathbb{I}\{x_i^{(n')} = m'\} \mathbb{I}\{t^{(n')} = t\}} \quad (18)$$

which corresponds to the number of times a categorical variable X_i takes the m^{th} value out of the M possible values for a certain class.

Similarly, it can be shown that for binary attributes, such that $x_i^{(n)} \in (0, 1)$ we have that the estimator for π_{it} is given as

$$\hat{\pi}_{it} = \hat{P}(X_i = 1 | t) = \frac{\sum_n \mathbb{I}\{x_i^{(n)} = 1, t^{(n)} = t\}}{\sum_n \{\mathbb{I}\{x_i^{(n)} = 0, t^{(n)} = t\} + \mathbb{I}\{x_i^{(n)} = 1, t^{(n)} = t\}}} \quad (19)$$

which corresponds to the number of times $X_i = 1$ for a specific value of t (i.e. for a specific class \mathcal{C}_k , with $k \in \{0, 1\}$) divided by the total number of observations in class \mathcal{C}_k . The estimator for the prior probabilities is given as

$$\hat{q}_k = \frac{N_k}{N} \quad (20)$$

which corresponds to the ratio of the number of observations in class \mathcal{C}_k , denoted by N_k , and the total number of observations in the training set N . In practice, this is calculated by counting the number cases from each class by observing the target value $t^{(n)}$ for each observation, hence the target variable is used in the derivation.

Consider the case when X_i is a binary attribute. As an example, suppose that for the class \mathcal{C}_0 , the observations in the training set all have a value of 0 for the specific attribute. This implies that the class-conditional probability $P(X_i = 1 | \mathcal{C}_0) = 0$. When plugging this probability into the NB model defined in (13), this would imply that, whenever a new observation \mathbf{x} has a value of 1 for that attribute, then the probability of that observation belonging to class \mathcal{C}_0 is 0, while the probability of that observation belonging to \mathcal{C}_1 is 1. This prediction is unlikely to be very accurate. The same can be said for the case when the attribute X_i can take more than 2 possible values. The more possible values M that the attribute can take, the more likely that this may become an issue, since it is more likely that certain possibilities never occur in the training set.

The most popular way of overcoming this issue is known as the Laplacian correction or Laplace estimator. It is assumed that for the training set \mathcal{D} , the number of observations N is large enough such that adding a small number of cases for each level of a factor for both classes would be negligible. Let $b \in \mathbb{N}$ denote the number of cases to be added for an attribute-class pair. It is important to note that b cases shall be added for each possible value of the attribute, so for binary attributes, $2b$ must be added to the denominator such that from (19) the Laplace estimator is

$$\hat{\pi}_{it}^L = \frac{\sum_n \mathbb{I}\{x_i^{(n)} = 1, t^{(n)} = t\} + b}{\sum_n \{\mathbb{I}\{x_i^{(n)} = 0, t^{(n)} = t\} + \mathbb{I}\{x_i^{(n)} = 1, t^{(n)} = t\}\} + 2b} \quad (21)$$

For categorical attributes, recall that there are $M > 2$ possible values such that Mb shall be added to the denominator. Therefore, from (18) the Laplace estimator is

$$\hat{\pi}_{it}^{(m)L} = \frac{\sum_n \mathbb{I}\{x_i^{(n)} = m\} \mathbb{I}\{t^{(n)} = t\} + b}{\sum_{m', n'} \mathbb{I}\{x_i^{(n')} = m'\} \mathbb{I}\{t^{(n')} = t\} + Mb} \quad (22)$$

Usually, b is taken to be 1 such that there is at least one count for each attribute-class pair in the training set. This prevents the issue of obtaining values of 0 for the marginal probabilities. Since under the NB assumption the marginal probabilities are multiplied, this would also annihilate the other marginal probabilities so it is an issue that needs to be dealt with, especially when the distribution of an attribute is skewed.

5 Data Analysis

The data set was provided by Datawatch Angoss and consists of 15,419 claims over the span of two years in the U.S. and contains 25 categorical variables.

There are a total of 923 fraudulent claims and 14496 legitimate claims. This class imbalance poses a number of problems when fitting any classification technique. Before splitting the data into the test set and train set the dimension of the data was reduced by applying a number of Chi-square tests between the classification variable and each of the other 24 categorical variables. The only variables which were kept are: Accident area, Sex, Fault, Vehicle category, Vehicle Price, Deductible, Past Number of Claims, Vehicle age, Age of Policy Holder, Police Report Filed, Agent type, Address Change to Claim, Base Policy and Number of Supplements.

The data set was next randomly divided between the test set and the training set. To reduce the class imbalance in the latter, a random under-sampling technique is carried out on the majority class. The training sample was thus reduced to 2753 claims with a 70:30 legitimate-to-fraud ratio.

5.1 NN for Fraud Detection

One hidden layer with a sigmoidal activation function turned out to be sufficient to provide a good classification for the data set. More hidden layers were attempted however, no improvement in the model accuracy was detected. Thus, a NN with one hidden layer using the hyperbolic tangent activation function and a single output node with a logistic sigmoid activation function was used on the training set. A grid search was carried out on three important hyper-parameters: the learning rate η , the number M of hidden layers, and the number of epochs. It was noticed that for $\eta = 0.01$, a network with 35 hidden neurones for 500 epochs achieved the lowest generalisation error and the highest mean AUC.

Figure 2 shows the training error superimposed on the validation error against the number of epochs for the network with 35 hidden neurons. Each of the five plots corresponds to each randomly selected subset held out as the validation set during stratified 5-fold cross-validation. These plots show that further training beyond 500 epochs is unjustified as it would lead to the model overfitting the training data. This is shown by the increase in validation error, corresponding with a decrease in training error.

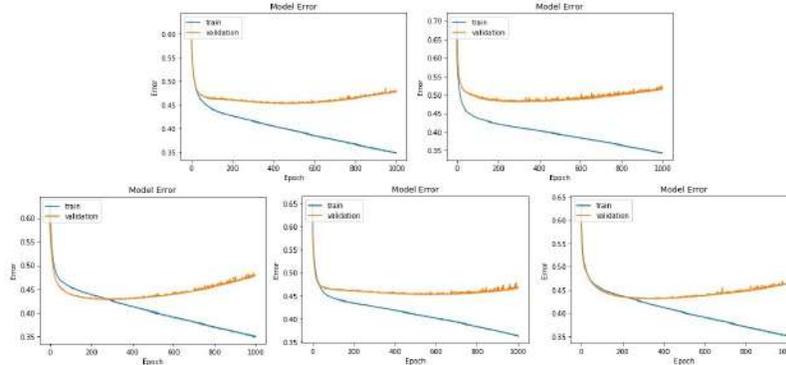


Fig. 2. Training error and Validation errors vs Epochs during cross-validation.

Next, the NN was applied to obtain predictions for the training data. Table 1 gives the confusion matrix for predictions of the chosen model on the training data. The NN classifies 87.2% of legitimate claims correctly, while classifying 62.5% of fraudulent claims correctly for the training data.

		Predicted		Total
		Fraudulent	Legitimate	
True	Fraudulent	$TP = 516$	$FN = 310$	826
	Legitimate	$FP = 246$	$TN = 1681$	1927
Total		762	1991	2753

Table 1. 2×2 contingency table for the NN on the training set

This confusion matrix can be used to calculate the following performance measures in Table 2. The NN performs well on the training data, with an accuracy of 79.8% and an AUC of 0.866. The similar scores for precision, sensitivity and F1 score suggest that the classifier has been trained to find a balance between minimising the number of false negatives and the number of false positives.

Accuracy	AUC	Precision	Sensitivity	F1 Score
0.798	0.866	0.677	0.625	0.650

Table 2. Performance measures for the NN on the training data

The ROC curve for the NN on the training data can give a visual representation of the performance of the classifier, for different cut-off values. Figure 3 reflects the value for the AUC on the training data since the curve is closer to the left top corner of the ROC space than to the diagonal.

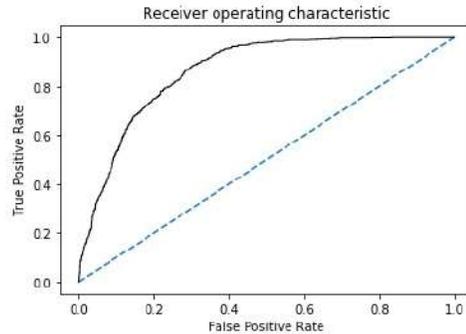


Fig. 3. ROC curve for the NN

5.2 Naive Bayes for Fraud Detection

The Naive Bayes model was fit on the training set in order to obtain the posterior distribution of the two classes. The predictions for the training set can be analysed by means of a confusion matrix. Table 3 shows that 78.6% of legitimate claims have been classified correctly, while 60.8% of fraudulent claims were classified correctly.

		Predicted		Total
		Fraudulent	Legitimate	
True	Fraudulent	$TP = 502$	$FN = 324$	826
	Legitimate	$FP = 413$	$TN = 1514$	1927
Total		915	1838	2753

Table 3. 2×2 contingency table for NB on the training set

From the confusion matrix, the performance measures shown in Table 4 can be derived. The classifier performs reasonably well, giving an AUC of 0.801.

Accuracy	AUC	Precision	Sensitivity	F1 Score
0.732	0.801	0.549	0.608	0.577

Table 4. Performance measures for NB on the training data

The ROC curve in Figure 4 backs up the value of the AUC, with the curve well above the line of no discrimination.

5.3 NN vs Naive Bayes

In this section we will compare the results obtained from NN and the Naive Bayes when applied to the test sets. Firstly, the confusion matrix for the NN is obtained from the predictions on the test data. Table 5 shows that 82.9% of legitimate claims were correctly classified, while 54.6% of fraudulent claims

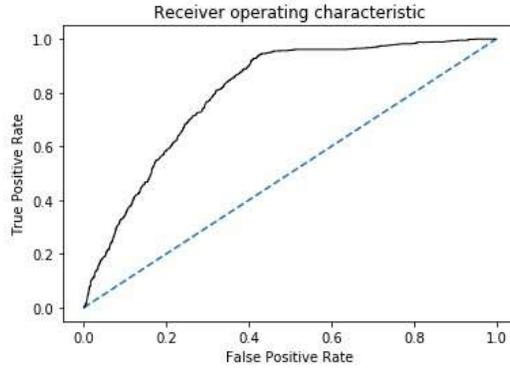


Fig. 4. ROC curve for NB

		Predicted		Total
		Fraudulent	Legitimate	
True	Fraudulent	$TP = 53$	$FN = 44$	97
	Legitimate	$FP = 262$	$TN = 1183$	1445
Total		315	1227	1542

Table 5. 2×2 contingency table for NN on the test set

were classified correctly for the test data.

Secondly, the NB model is used to obtain predictions for the test data which can be summarised by the confusion matrix in Table 6. 77.1% of legitimate claims were correctly classified, while 58.8% of fraudulent claims were correctly classified. The NB classifier achieved 4 more true positives than the NN; however, also predicted 69 more false positives.

		Predicted		Total
		Fraudulent	Legitimate	
True	Fraudulent	$TP = 57$	$FN = 40$	97
	Legitimate	$FP = 331$	$TN = 1114$	1445
Total		388	1154	1542

Table 6. 2×2 contingency table for NB on the test set

The ROC curves for the NN and NB models for the test data are superimposed to obtain a visual comparison of the two classifiers' performance on new data points. Figure 5 shows that the performance of the two classifiers is comparable, with the ROC curve for NN slightly dominating the ROC curve for the NB classifier. The NB classifier slightly outperformed the NN in terms of correctly classifying fraudulent claims; however, this comes at a cost of the

precision of the classifier due to the larger number of false positives.

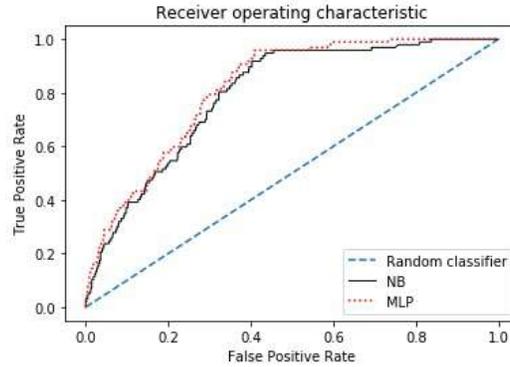


Fig. 5. ROC curves for NN and NB on test data

Furthermore, Table 7 gives a comparison of the performance of the two classifiers on the test set. The NN model and NB model were also fitted on the training data with the original 94 : 6 legitimate-to-fraudulent distribution, before the under-sampling of the majority class was carried out, in order to highlight the need to cater for this class imbalance in the data. Note that for the models trained on the under-sampled data (70 : 30 legitimate:fraudulent ratio), the NN model slightly outperformed the NB model with respect to all performance measures except for sensitivity. The higher sensitivity for the NB model came at a cost of predicting more false positives as mentioned.

Model (training set ratio)	Accuracy	AUC	Precision	Sensitivity	F1 Score
NN (70:30)	0.802	0.817	0.168	0.546	0.257
NB (70:30)	0.759	0.795	0.147	0.588	0.235
NN (94:6)	0.937	0.832	0.500	0.010	0.020
NB (94:6)	0.923	0.503	0.038	0.010	0.016

Table 7. Performance of NN & NB on the test data with different training ratios

Upon examining the performance measures for the models fitted on data with the original class imbalance (94 : 6 legitimate:fraud ratio), the accuracy scores for both the NN and NB model are very high; however, as expected both models failed to detect the majority of fraudulent claims. This confirms that the under-sampling technique used was justified and effective.

6 Conclusion

In this paper the performance of NN and Naive Bayes classifiers was compared when applied to an Insurance fraud detection data set. Before implementing these classifiers, the data was pre-processed by selecting significant variables

for the study and under-sampling the majority class to overcome the class imbalance issue. The two classifiers were then compared by considering a number of performance measures as well as the ROC curve. Model selection by cross-validation was carried out for the NN. Model selection and training time for the more complex NN proved to be a lengthy process, in contrast to the simpler NB classifier. The NN performed better on the training data than the NB model; however, for new instances (i.e. the test data), the performance of the two classifiers was comparable. The NB model slightly outperformed the NN in terms of classifying new fraudulent instances correctly; however, this was at the cost of predicting many more false positives than the NN, which is undesirable due to the possible cost of investigating claims. The under-sampling technique proved to be effective for this application as shown by comparing the results from training the models on different class distributions.

References

1. D. Barber, Bayesian reasoning and machine learning, Cambridge University Press, 2010.
2. C.M. Bishop, Neural Networks for Pattern Recognition, Oxford university press, New York, 1995.
3. R. Bolton, J. Richard and D.J. Hand, Statistical fraud detection: A review, *Statistical Science*, 235-249, 2002.
4. P. Clifton, L. Vincent, K.Smith and G. Ross, A comprehensive survey of data mining-based fraud detection research, 2010.
5. G. Cybenko, Approximation by superpositions of a sigmoidal function, *Mathematics of control, signals and systems*, Springer, 2,4,303-314, 1989.
6. P. Domingos and M. Pazzani, On the Optimality of the Simple Bayesian Classifier Under Zero-One Loss, *Springer*, 29,2-3,103-130,1997.
7. J. Dougherty, R. Kohavi and M. Sahami, Supervised and unsupervised discretization of continuous features, *Machine Learning Proceedings*, 194-202,1995.
8. D. J. Hand and K. Yu, Idiot's Bayes—not so stupid after all?, *International statistical review*, 69,3,385-389, Wiley Online Library, 2001.
9. M.H. Hassoun, Fundamentals of artificial neural networks, MIT Press, 1995.
10. D. Heckerman and D. Geiger, Learning Bayesian networks: a unification for discrete and Gaussian domains, *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 274-284,1995.
11. K. Hornik, Approximation capabilities of multilayer feedforward networks, *Neural networks*, Elsevier, 4,2, 251-257,1991.
12. E. W. T. Ngai, Y. Hu, and Wong, Yiu Hing, Y. Chen, X. Sun, The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature, *Decision Support System*, 50, 3, 559-569, Elsevier 2011.
13. B. D. Ripley, Neural networks and related methods for classification, *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, 409-456,1994.

Examining Items' Suitability as the Marker Indicator in Testing Measurement Invariance

Anastasia Charalampi¹, Catherine Michalopoulou², and Clive Richardson³

¹ Postdoctoral Fellow, Department of Social Policy, Panteion University of Social and Political Sciences

(E-mail: acharalampi@panteion.gr)

² Professor of Statistics, Department of Social Policy, Panteion University of Social and Political Sciences (E-mail: kmichal@panteion.gr)

³ Emeritus Professor of Applied Statistics, Department of Economic and Regional Development, Panteion University of Social and Political Sciences, Athens, Greece (E-mail: crichard@panteion.gr)

Abstract. In testing measurement invariance, researchers must choose which item of the observed construct will serve as the marker indicator. In the literature, little consideration is given to this rarely-reported decision and an item is often selected automatically by software defaults. However, in many cases, the choice of marker indicator may influence the interpretation of the model. In this paper, we explore empirically the suitability of items by repeatedly performing multiple-groups Confirmatory factor analysis of the same data using a different marker indicator each time. The investigation is based on an eleven-item unidimensional scale measuring emotional wellbeing from the European Social Survey of 2006 and 2012. Measurement invariance is tested for gender and employment status groups in a combined sample of eight European countries: Belgium, France, Germany, Netherlands, Poland, Portugal, Russian Federation and Spain.

Keywords: measurement invariance, multiple-groups Confirmatory factor analysis, European Social Survey, wellbeing.

1 Introduction

A prerequisite for meaningful comparisons of constructs across different demographic and social groups, within and across nations is the establishment of their measurement invariance or equivalence (Davidov [15]; Missine *et al.* [29]; Davidov *et al.* [16]; Raudenská [32]). Measurement invariance ensures that a measurement instrument, e.g. an attitude scale, “measures the same concept in the same way across various subgroups of respondents” (Davidov *et al.* [16: 9]; see also, Davidov [15]) or across repeated measurements, time points and social categories (Putnick and Bornstein [30]; Davidov *et al.* [16]). Without measurement invariance, conclusions based on comparisons of different groups

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



or across time would be ambiguously interpreted (Cheung and Rensvold [12]; Steenkamp and Baumgartner [33]; Xu and Tracey [40]).

Although the importance of testing for measurement invariance was pointed out in the literature more than 50 years ago (Putnick and Bornstein [30]), appropriate statistical methods were developed under a structural equation modeling framework relatively recently (e.g. Cheung and Rensvold [12]; Steenkamp and Baumgartner [33]; Vandenberg [38]; Vandenberg and Lance [39]). One of the most popular methods for testing measurement invariance is multiple-groups Confirmatory factor analysis (MGCFA; Brown [3]; Cheung and Rensvold [12]; Han *et al.* [22]; Jung and Yoon [24]). As is the case in any structural equation model, researchers need to “assign a unit of measurement to the latent factor so that the latent factors will have a scale and so that the model is identified” (Han *et al.* [22: 1488]; see also, Cheung and Rensvold [12]; Cheung and Lau [10]). The most commonly used approach is to select an item to serve as a marker (or referent) indicator (reference variable) for each factor and fix its factor loading at one (Brown [3]; Cheung and Rensvold [11]; Cheung and Lau [11]; Han *et al.* [22]; Millsap and Yun-Tein [28]). In the literature, little consideration is given to this decision (Jung and Yoon [24]), which is but rarely reported, or an item is chosen automatically by software defaults (Brown [3]; Bowen and Masa [2]). However, in many cases, the marker indicator may influence the interpretation of the model (Jung and Yoon [24]).

Choosing the marker indicator is considered to be complicated especially when there are many items (Cheung and Rensvold [12]; Cheung and Lau [11]; Han *et al.* [22]; Jung and Yoon [24]). Steenkamp and Baumgartner [33] suggested that the item selected to serve as marker indicator should demonstrate metric invariance. If a noninvariant item is selected, then the parameter estimates could be distorted leading to inaccurate conclusions (Vandenberg [38]; French and Finch [21]). However, without testing for measurement invariance there is no other way of knowing beforehand which items are invariant (Cheung and Lau [11]). In this respect, Liu *et al.* [27: 18] proposed that a “marker variable should have a meaningful metric, or be an indicator of the latent common factor with a high factor loading. For evaluating longitudinal measurement invariance, it is crucial to choose a marker variable whose loading is invariant at all occasions. The model identification strategy requires that two of the thresholds for the marker variable be constrained to be invariant across measurement occasions. Therefore, the marker variable should not only have an invariant factor loading across all measurement occasions, but also have at least two invariant thresholds.” In the cases where there is no obvious item with a meaningful metric, Brown [3] proposed performing MGCFA using different items as marker indicator (see also, Han *et al.* [22]).

In this paper, we empirically explore the suitability of items to serve as marker indicator by performing MGCFA using a different item each time as suggested by Brown [3]. The investigation is based on an eleven-item unidimensional scale measuring emotional wellbeing from the European Social

Survey of 2006 and 2012. Measurement invariance is tested for gender and employment status groups of a combined sample of eight European countries.

1.1 Emotional wellbeing scale

As mentioned in our previous work (Charalampi [6]; Charalampi *et al.* [7, 8, 9]), the personal and social wellbeing module was first included in the questionnaire of Round 3 (2006) of the ESS and was repeated with certain changes in Round 6 (2012) of the survey (European Social Survey [18]; Jeffrey *et al.* [23]). Combining theoretical models and evidence from statistical analyses, six key dimensions were defined for the 2012 ESS measurement of personal and social wellbeing as follows (European Social Survey [18]; Jeffrey *et al.* [23]): evaluative wellbeing, emotional wellbeing, functioning, vitality, community and supportive relationships.

Four variables from emotional wellbeing, three items from vitality and one from the supportive relationships dimensions comprise the eight-item version of the Center for Epidemiologic Studies Depression scale (CES-D 8), which was proposed by Van de Velde *et al.* [36, 37]. CES-D 8 is a shortened version of the original CES-D that was developed by Radloff [31], designed to measure symptomatology of depression in the general population and non-clinical settings (Andresen *et al.* [1]; Carleton *et al.* [5]; Cole *et al.* [14]; Karim *et al.* [25]). In addition to the CES-D 8 scale, the ESS contains three additional items that can be considered as another set of similar questions based on personal feelings (Raudenská [32]). In sum, there are eleven common items included in both Rounds of the ESS (Raudenská [32]). This eleven-item scale measures emotional wellbeing as it consists of seven items measuring the existence of negative emotions and four measuring the existence of positive emotions.

2 Method

2.1 Participants

The analysis was based on the European Social Survey Round 3 Data [19] and the European Social Survey Round 6 Data [20] for the following eight countries: Belgium, France, Germany, Netherlands, Poland, Portugal, Russian Federation and Spain. These countries were selected from the 29 participants in Round 6 because they had also participated in Round 3 (2006), when the wellbeing module including the items measuring emotional wellbeing was first introduced into the questionnaire. The ESS implements all the strict methodological prerequisites for comparability over time and cross-nationally (Kish [26]; Carey [4]) by applying probability sampling, minimum effective

achieved sample sizes in all participating countries and a maximum target non-response rate of 30% (The ESS Sampling Expert Panel [35]). Face-to-face interviewing is used for data collection. The survey population is defined as all persons aged 15 and over residing within private households in each country, regardless of their nationality, citizenship or language; this definition applies to all rounds of the survey. The samples of the eight countries included more women than men with the exception of Germany in Round 6. The mean age ranged from 44 to 52 years in every country. More than 42.9% of the participants were married, the majority had completed at most secondary education with the exception of Russian Federation in Round 6 and at least 39.1% were in paid work.

2.2 Measures

The eleven items used for the emotional wellbeing scale were worded as follows (European Social Survey [18]): felt depressed: how often during the past week (E1); felt that everything I did was an effort: how often during the past week (E2); Sleep was restless: how often during the past week (E3); was happy: how often during the past week (E4); felt lonely: how often during the past week (E5); enjoyed life: how often during the past week (E6); felt sad: how often during the past week (E7); could not get going: how often during the past week (E8); had a lot of energy: how often during the past week (E9); felt anxious: how often during the past week (E10); felt calm and peaceful: how often during the past week (E11). The scale is comprised of seven negatively (E1-E3, E5, E7-E8 and E10) and four positively (E4, E6, E9 and E11) worded items. The scoring of negatively worded items was reversed before the analysis in order to achieve correspondence between the ordering of the response categories following the original grouping of items into the wellbeing dimensions based on Jeffrey *et al.* (2015). The response categories range from 1 to 4 and are defined as follows: none or almost none of the time (1); some of the time (2); most of the time (3); and all or almost all of the time (4). Therefore, the items' level of measurement is ordinal, i.e. categorical.

2.3 Statistical analysis

MGCFA was carried out using Mplus Version 8.4 for testing measurement invariance of gender (men and women) and employment status (employed and unemployed) groups: configural, metric and scalar (Brown [3]; Millsap and Yun-Tein [28]).

In performing MGCFA, the following sequence of decisions was adopted as presented in our previous work (Charalampi *et al.* [10]):

1. Initially, separate Confirmatory Factor Analyses (CFAs) were performed for both demographic and social groups (Brown [3]) following the

4

sequence of decisions presented in our previous work (Charalampi [6]; Charalampi *et al.* [7, 8, 9]). As Brown [3: 246] pointed out, “if markedly disparate measurement models are obtained between groups, this outcome will contraindicate further invariance evaluation”. Model fit was considered adequate if $\chi^2/df < 3$, CFI and TLI values were greater than or close to .95 and RMSEA $\leq .06$ with the 90% CI upper limit $\leq .06$, or acceptable if $\chi^2/df < 3$, the CFI and TLI values were $> .90$ and RMSEA $< .08$ with the 90% CI upper limit $< .08$.

2. Identifying the marker indicator: if no item indicated by theory had meaningful metric to serve as marker indicator, then all items had to be tested and used as such to define the metric of the latent variable (Brown [3]; Han *et al.* [22]).
3. To test the measurement invariance of the emotional wellbeing scale across gender and employment status, a series of MGCFAs were performed with robust weighted least squares (WLSMV). Across groups, we sequentially constrained parameters in three models: the configural, metric, and scalar models. The configural model is a “baseline multiple-groups model whereby the factor loadings and thresholds [in the case of categorical items] are freely estimated in all groups” (Brown [3: 370]). In the metric model, the factor loadings are constrained to equality across groups. In the scalar model, the constraint of equal indicator intercepts across groups is added as well. In each case, model fit was considered adequate or acceptable as in step 1.
4. Measurement invariance was established by sequentially comparing and testing the decrease in model fit from one model to the subsequent one based on the change (Δ) in both CFI and RMSEA. In this case of ordinal items, Δ CFI and Δ RMSEA of 0.004 and 0.050, respectively, from the configural to the metric model and Δ CFI and Δ RMSEA of 0.004 and 0.010, respectively, from the metric to the scalar model indicated a lack of measurement invariance (Rutkowski and Svetina, [34]; see also, Raudenská, [32]).

3 Results

In Table 1, the standardized factor loadings for the gender (men and women) and employment status (employed and unemployed) groups of separate CFAs are presented. As shown, in all cases, CFA resulted in a one-factor model that was to be tested. As no item had a meaningful metric by theory to serve as marker indicator, then models using E1 and E7, i.e. the items with the highest factor loadings, could be tested if we were to follow Liu *et al.* [27]. However, all items (E1-E11) were tested as suggested by Brown [3] and Han *et al.* [22].

Table 1. Standardized factor loadings for the gender and employment status groups of separate CFAs performed with robust weighted least squares of the polychoric correlation matrix on a combined sample of eight European countries: European Social Survey, 2006 and 2012

	2006				2012			
	Female	Male	Unempl.	Employed	Female	Male	Unempl.	Employed
E1	.846	.825	.831	.838	.831	.824	.811	.820
E2	.692	.608	.648	.614	.657	.584	.626	.572
E3	.593	.586	.616	.558	.559	.529	.609	.529
E4	.631	.596	.574	.575	.626	.593	.622	.567
E5	.706	.697	.651	.690	.676	.664	.651	.660
E6	.576	.511	.505	.486	.563	.521	.531	.502
E7	.848	.817	.851	.823	.836	.807	.843	.811
E8	.726	.682	.763	.657	.717	.665	.724	.653
E9	.582	.526	.559	.516	.576	.478	.491	.484
E10	.649	.618	.652	.617	.641	.623	.657	.618
E11	.575	.523	.545	.557	.642	.590	.683	.616

The first and second largest factor loadings are in boldface.

In men and women in the eight countries data sets of 2006 (Table 2) and 2012 (Table 3), although overall fit statistics for the one-factor model were consistent with acceptable model fit for men and inadequate model fit for women, the performance of the models was not markedly different. In both groups, all freely estimated factor loadings were statistically significant ($p < .001$) and salient (standardized factor loadings ranged from .511 to .848 and .478 to .836 for the 2006 and 2012 data sets, respectively). However, both analyses for men resulted in four modification indices exceeding 100 and for women nine and twelve for 2006 and 2012, respectively, suggesting the presence of “salient focal areas of misspecification” (Brown [3: 372]).

In the 2006 dataset (Table 2), the configural model resulted in acceptable model fit. Almost identical results were obtained when using E1-E11 as marker indicators for testing metric and scalar invariance. In all cases, the metric and scalar models resulted in acceptable and adequate model fit, respectively. In all cases, the change in both CFI and RMSEA from the configural to the metric model ranged from -0.007 to -0.006 (i.e., < 0.004) and 0.012-0.013 (i.e., < 0.050), respectively, and the change in both CFI and RMSEA from the metric to the scalar model ranged from 0.000-0.001 (i.e., < 0.004) and 0.006-0.007 (i.e., < 0.010), respectively. Therefore, all the models using E1-E11 as marker indicators demonstrated metric and scalar measurement invariance. If one were

to decide among these models, the model with the best overall results was the one using E8 as marker indicator.

Table 2. Measurement invariance tests of the 2006 European Social Survey Emotional wellbeing scale for gender performed on a combined sample of eight European countries using different items (E1-E11) as marker indicator

Gender	χ^2/df	CFI (Δ CFI)	TLI	RMSEA (Δ RMSEA)	RMSEA 90% CI
Single group solutions					
Male ($n = 7,621$)	40.30	.966	.954	.072	.069-.075
Female ($n = 9,188$)	59.05	.972	.961	.079	.077-.082
Measurement invariance:					
Configural model	49.29	.970	.959	.076	.074-.078
Marker indicator E1					
Metric model	35.35	.976 (-.006)	.971	.064 (.012)	.062-.066
Scalar model	28.20	.977 (.001)	.977	.057 (.007)	.055-.059
Marker indicator E2					
Metric model	35.10	.977 (-.007)	.971	.064 (.012)	.062-.066
Scalar model	28.20	.977 (.000)	.977	.057 (.007)	.055-.059
Marker indicator E3					
Metric model	34.95	.977 (-.007)	.971	.064 (.012)	.062-.065
Scalar model	28.20	.977 (.000)	.977	.057 (.007)	.055-.059
Marker indicator E4					
Metric model	35.36	.976 (-.006)	.971	.064 (.012)	.062-.066
Scalar model	28.20	.977 (.001)	.977	.057 (.007)	.055-.059
Marker indicator E5					
Metric model	35.07	.977 (-.007)	.971	.064 (.012)	.062-.066
Scalar model	28.20	.977 (.000)	.977	.057 (.007)	.055-.059
Marker indicator E6					
Metric model	35.25	.976 (-.006)	.971	.064 (.012)	.062-.066
Scalar model	28.20	.977 (.001)	.977	.057 (.007)	.055-.059
Marker indicator E7					
Metric model	35.17	.976 (-.006)	.971	.064 (.012)	.062-.066
Scalar model	28.20	.977 (.001)	.977	.057 (.007)	.055-.059
Marker indicator E8					
Metric model	35.21	.976 (-.006)	.971	.064 (.012)	.062-.066
Scalar model	28.20	.977 (.001)	.977	.057 (.007)	.055-.059
Marker indicator E9					
Metric model	35.24	.976 (-.006)	.971	.064 (.012)	.062-.066
Scalar model	28.20	.977 (.001)	.977	.057 (.007)	.055-.059
Marker indicator E10					
Metric model	34.73	.977 (-.007)	.972	.063 (.013)	.061-.065
Scalar model	28.20	.977 (.000)	.977	.057 (.006)	.055-.059
Marker indicator E11					
Metric model	35.17	.976 (-.006)	.971	.064 (.012)	.062-.066
Scalar model	28.20	.977 (.001)	.977	.057 (.007)	.055-.059

Note. *df* = degrees of freedom; CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root-mean- square error of approximation; CI = confidence interval. Note, that the χ^2/df for the configural model is the mean value of the χ^2/df of the two groups.

In the 2012 dataset (Table 3), the configural model resulted in acceptable model fit. Almost identical results were obtained when using E1-E11 as marker indicators for testing metric and scalar invariance. In all cases, the metric and scalar models resulted in acceptable and adequate model fit, respectively. In all cases, the change in both CFI and RMSEA from the configural to the metric model was -0.006 (i.e., < 0.004) and 0.012 (i.e., < 0.050), respectively, and the change in both CFI and RMSEA from the metric to the scalar model was 0.000 (i.e., < 0.004) and 0.006 (i.e., < 0.010), respectively. Therefore, all the models using E1-E11 as marker indicators demonstrated metric and scalar measurement invariance. If one were to decide among these models, the models with the best overall results with slightly better TLI values were those using E2, E3, E5, E7, E8 and E10 as marker indicators. However, if one wanted comparable results of 2006 and 2012 for gender, then E8 should be used as marker indicator.

In both the unemployed and the employed groups in the eight countries data sets of 2006 (Table 4), overall fit statistics for the one-factor model resulted in inadequate model fit. Data sets of 2012 (Table 5) resulted in an acceptable model fit for both groups. However, in both cases, the performance of the models was not markedly different. In both groups, all freely estimated factor loadings were statistically significant ($p < .001$) and salient (standardized factor loadings ranged from .486 to .851 and .484 to .843 for the 2006 and 2012 data sets, respectively). The analyses for the unemployed resulted in relatively small modification indices ranging from 10-28 and 10-46 for 2006 and 2012, respectively. The analyses for the employed resulted in nine and 11 modification indices exceeding 80 for 2006 and 2012, respectively, suggesting the presence of “salient focal areas of misspecification” (Brown [3: 372]).

In the 2006 dataset (Table 4), the configural model resulted in acceptable model fit. Almost identical results were obtained when using E1-E11 as marker indicators for testing metric and scalar invariance. In all cases, the metric and scalar models resulted in acceptable and adequate model fit, respectively. In all cases, the change in both CFI and RMSEA from the configural to the metric model was -0.007 (i.e., < 0.004) and ranged from 0.010-0.011 (i.e., < 0.050), respectively, and the change in both CFI and RMSEA from the metric to the scalar model was -0.002 (i.e., < 0.004) and ranged from 0.008-0.009 (i.e., < 0.010), respectively. Therefore, all the models using E1-E11 as marker indicators demonstrated metric and scalar measurement invariance. In this case, there was no obvious choice among the models tested. However, if one wanted to present results of 2006 along with those for gender, then E8 should be used as marker indicator.

Table 3. Measurement invariance tests of the 2012 European Social Survey Emotional wellbeing scale for gender performed on a combined sample of eight European countries using different items (E1-E11) as marker indicator

Gender	χ^2/df	CFI (Δ CFI)	TLI	RMSEA (Δ RMSEA)	RMSEA 90% CI
Single group solutions					
Men ($n = 7,760$)	36.67	.968	.956	.068	(.065-.071)
Women ($n = 9,291$)	62.53	.968	.956	.081	(.079-.084)
Measurement invariance:					
Configural model	49.08	.969	.957	.075	.073-.077
Marker indicator: E1					
Metric model	35.07	.975 (-.006)	.969	.063 (.012)	.061-.065
Scalar model	28.32	.975 (.000)	.975	.057 (.006)	.055-.058
Marker indicator: E2					
Metric model	34.83	.975 (-.006)	.970	.063 (.012)	.061-.065
Scalar model	28.32	.975 (.000)	.975	.057 (.006)	.055-.058
Marker indicator: E3					
Metric model	34.44	.975 (-.006)	.970	.063 (.012)	.061-.065
Scalar model	28.32	.975 (.000)	.975	.057 (.006)	.055-.058
Marker indicator: E4					
Metric model	35.23	.975 (-.006)	.969	.063 (.012)	.061-.065
Scalar model	28.32	.975 (.000)	.975	.057 (.006)	.055-.058
Marker indicator: E5					
Metric model	34.57	.975 (-.006)	.970	.063 (.012)	.061-.065
Scalar model	28.32	.975 (.000)	.975	.057 (.006)	.055-.058
Marker indicator: E6					
Metric model	35.12	.975 (-.006)	.969	.063 (.012)	.061-.065
Scalar model	28.32	.975 (.000)	.975	.057 (.006)	.055-.058
Marker indicator: E7					
Metric model	34.98	.975 (-.006)	.970	.063 (.012)	.061-.065
Scalar model	28.32	.975 (.000)	.975	.057 (.006)	.055-.058
Marker indicator: E8					
Metric model	34.88	.975 (-.006)	.970	.063 (.012)	.061-.065
Scalar model	28.32	.975 (.000)	.975	.057 (.006)	.055-.058
Marker indicator: E9					
Metric model	35.04	.975 (-.006)	.969	.063 (.012)	.061-.065
Scalar model	28.32	.975 (.000)	.975	.057 (.006)	.055-.058
Marker indicator: E10					
Metric model	34.46	.975 (-.006)	.970	.063 (.012)	.061-.065
Scalar model	28.32	.975 (.000)	.975	.057 (.006)	.055-.058
Marker indicator: E11					
Metric model	35.00	.975 (-.006)	.969	.063 (.012)	.061-.065
Scalar model	28.32	.975 (.000)	.975	.057 (.006)	.055-.058

Note. df = degrees of freedom; CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root-mean-square error

of approximation; CI = confidence interval. Note, that the χ^2/df for the configural model is the mean value of the χ^2/df of the two groups.

Table 4. Measurement invariance tests of the 2006 European Social Survey Emotional wellbeing scale for employment status performed on a combined sample of eight European countries using different items (E1-E11) as marker indicator

Employment status	χ^2/df	CFI (Δ CFI)	TLI	RMSEA (Δ RMSEA)	RMSEA 90% CI
Single group solutions					
Unemployed ($n = 866$)	6.00	.971	.959	.076	.067-.085
Employed ($n = 8,240$)	52.32	.961	.946	.079	.076-.082
Measurement invariance:					
Configural model	28.96	.960	.945	.078	.076-.081
Marker indicator: E1					
Metric model	21.66	.967 (-.007)	.960	.067 (.011)	.065-.070
Scalar model	16.65	.969 (-.002)	.969	.059 (.008)	.056-.061
Marker indicator: E2					
Metric model	21.55	.967 (-.007)	.960	.067 (.011)	.065-.070
Scalar model	16.65	.969 (-.002)	.969	.059 (.008)	.056-.061
Marker indicator: E3					
Metric model	21.64	.967 (-.007)	.960	.067 (.011)	.065-.070
Scalar model	16.65	.969 (-.002)	.969	.059 (.008)	.056-.061
Marker indicator: E4					
Metric model	21.70	.967 (-.007)	.960	.067 (.011)	.065-.070
Scalar model	16.65	.969 (-.002)	.969	.059 (.008)	.056-.061
Marker indicator: E5					
Metric model	21.53	.967 (-.007)	.960	.067 (.011)	.065-.070
Scalar model	16.65	.969 (-.002)	.969	.059 (.008)	.056-.061
Marker indicator: E6					
Metric model	21.71	.967 (-.007)	.959	.067 (.011)	.065-.070
Scalar model	16.65	.969 (-.002)	.969	.059 (.008)	.056-.061
Marker indicator: E7					
Metric model	21.69	.967 (-.007)	.960	.067 (.011)	.065-.070
Scalar model	16.65	.969 (-.002)	.969	.059 (.008)	.056-.061
Marker indicator: E8					
Metric model	21.81	.967 (-.007)	.959	.068 (.010)	.065-.070
Scalar model	16.65	.969 (-.002)	.969	.059 (.009)	.056-.061
Marker indicator: E9					
Metric model	21.68	.967 (-.007)	.960	.067 (.011)	.065-.070
Scalar model	16.65	.969 (-.002)	.969	.059 (.008)	.056-.061
Marker indicator: E10					
Metric model	21.59	.967 (-.007)	.960	.067 (.011)	.065-.070
Scalar model	16.65	.969 (-.002)	.969	.059 (.008)	.056-.061
Marker indicator: E11					
Metric model	21.61	.967 (-.007)	.960	.067 (.011)	.065-.070
Scalar model	16.65	.969 (-.002)	.969	.059 (.008)	.056-.061

Note. df = degrees of freedom; CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root-mean-square error

of approximation; CI = confidence interval. Note, that the χ^2/df for the configural model is the mean value of the χ^2/df of the two groups.

In the 2012 dataset (Table 5), the configural model resulted in acceptable model fit. Almost identical results were obtained when using E1-E11 as marker indicators for testing metric and scalar invariance. In all cases, the metric and scalar models resulted in acceptable and adequate model fit, respectively. In all cases, the change in both CFI and RMSEA from the configural to the metric model ranged from -0.009 to -0.008 (i.e., < 0.004) and 0.012-0.013 (i.e., < 0.050), respectively, and the change in both CFI and RMSEA from the metric to the scalar model ranged from -0.001-0.000 (i.e., < 0.004) and 0.007-0.008 (i.e., < 0.010), respectively. Therefore, all the models using E1-E11 as marker indicators demonstrated metric and scalar measurement invariance. In this case, there was no obvious choice among the models tested. However, if one wanted to present results of 2006 and 2012 along with those for gender, then E8 should be used as marker indicator.

Conclusions

In this paper, we demonstrated empirically a methodology for identifying a marker indicator to define the metric of the latent variable when there is no obvious item with a meaningful metric to serve as such using the eleven-item unidimensional scale measuring emotional wellbeing from the European Social Survey of 2006 and 2012. Measurement invariance was tested for gender (men and women) and employment status (unemployed and employed) groups in a combined sample of eight European countries: Belgium, France, Germany, Netherlands, Poland, Portugal, Russian Federation and Spain. To decide between the eleven items of the emotional wellbeing scale and empirically explore their suitability, MGCFAs were performed using a different item each time as Brown [3] and Han *et al.* [22] proposed.

In all cases of testing for measurement invariance of the two gender (men and women) and employment status (unemployed, employed) groups on both datasets (2006 and 2012), the configural model resulted in acceptable model fit. Almost identical results were obtained when using E1-E11 as marker indicators for testing metric and scalar invariance. In all cases, the metric and scalar models resulted in acceptable and adequate model fit, respectively. Therefore, all the models using E1-E11 as marker indicators demonstrated metric and scalar measurement invariance.

Table 5. Measurement invariance tests of the 2012 European Social Survey Emotional wellbeing scale for employment status performed on a combined sample of eight European countries using different items (E1-E11) as marker indicator

Employment status	χ^2/df	CFI (Δ CFI)	TLI	RMSEA (Δ RMSEA)	RMSEA 90% CI
Single group solutions					
Unemployed ($n = 1,250$)	7.05	.976	.967	.070	.062-.077
Employed ($n = 8,154$)	49.97	.959	.943	.077	.075-.080
Measurement invariance:					
Configural model	28.89	.960	.945	.077	.074-.080
Marker indicator: E1					
Metric model	20.77	.968 (-.008)	.961	.065 (.012)	.062-.067
Scalar model	16.50	.969 (-.001)	.970	.057 (.008)	.055-.060
Marker indicator: E2					
Metric model	20.65	.968 (-.008)	.961	.065 (.012)	.062-.067
Scalar model	16.50	.969 (-.001)	.970	.057 (.008)	.055-.060
Marker indicator: E3					
Metric model	20.62	.968 (-.008)	.961	.065 (.012)	.062-.067
Scalar model	16.50	.969 (-.001)	.970	.057 (.008)	.055-.060
Marker indicator: E4					
Metric model	20.72	.968 (-.008)	.961	.065 (.012)	.062-.067
Scalar model	16.50	.969 (-.001)	.970	.057 (.008)	.055-.060
Marker indicator: E5					
Metric model	20.54	.969 (-.009)	.962	.064 (.013)	.062-.067
Scalar model	16.50	.969 (.000)	.970	.057 (.007)	.055-.060
Marker indicator: E6					
Metric model	20.69	.968 (-.008)	.961	.065 (.012)	.062-.067
Scalar model	16.50	.969 (-.001)	.970	.057 (.008)	.055-.060
Marker indicator: E7					
Metric model	20.76	.968 (-.008)	.961	.065 (.012)	.062-.067
Scalar model	16.50	.969 (-.001)	.970	.057 (.008)	.055-.060
Marker indicator: E8					
Metric model	20.75	.968 (-.008)	.961	.065 (.012)	.062-.067
Scalar model	16.50	.969 (-.001)	.970	.057 (.008)	.055-.060
Marker indicator: E9					
Metric model	20.74	.968 (-.008)	.961	.065 (.012)	.062-.067
Scalar model	16.50	.969 (-.001)	.970	.057 (.008)	.055-.060
Marker indicator: E10					
Metric model	20.69	.968 (-.008)	.961	.065 (.012)	.062-.067
Scalar model	16.50	.969 (-.001)	.970	.057 (.008)	.055-.060
Marker indicator: E11					
Metric model	20.72	.968 (-.008)	.961	.065 (.012)	.062-.067
Scalar model	16.50	.969 (-.001)	.970	.057 (.008)	.055-.060

Note. df = degrees of freedom; CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root-mean-square error

of approximation; CI = confidence interval. Note, that the χ^2/df for the configural model is the mean value of the χ^2/df of the two groups.

In the case of testing for measurement invariance of the two gender groups (men and women) for the 2006 dataset, one item (E8) gave slightly better results than all the others. In this respect, if one wished to present comparable results for the 2012 dataset and the two employment status (unemployed and employed) groups, the same item should be used as marker indicator. This item had a high factor loading (Table 1) confirming thus the results of Liu *et al.* [27: 18] for considering an “indicator of the latent common factor with a high factor loading” as marker indicator. These results suggest that researchers should first investigate and report on the suitability of items to serve as marker indicators instead of relying on the default software rule of the first item in the list (Brown [3]; Bowen and Masa [2]).

References

1. E.M. Andresen, J.A. Malmgren, W.B. Carter and D.L. Patrick. Screening for depression in well older adults: Evaluation of a short form of the CES-D, *American Journal of Preventive Medicine*, 10, 2, 77-84, 1994.
2. N.K. Bowen and R.D. Masa. Conducting measurement invariance tests with ordinal data: a guide for social work researchers, *Journal of the Society for Social Work Research*, 6, 2, 229-249, 2015.
3. T.A. Brown, *Confirmatory factor analysis for applied research* (2nd edition), The Guilford Press, New York, 2015.
4. S. Carey (Ed.), *Measuring adult literacy: The International Adult Literacy Survey (IALS) in the European context*, Office for National Statistics, London, 2000.
5. R.N. Carleton, M.A. Thibodeau, M.J.N. Teale, P.G. Welch, M.P. Abrams, T. Robinson and G.J.G. Asmundson. The Center for Epidemiologic Studies Depression scale: A review with a theoretical and empirical examination of item content and factor structure. *Plos One*, 8, 3, e58067, 2013. doi: 10.1371/journal.pone.0058067
6. A. Charalampi, *The importance of items' level of measurement in investigating the structure and assessing the psychometric properties of multidimensional constructs* (Doctoral dissertation). Retrieved from the National Archive of Ph.D. Theses, National Documentation Centre (ND 44012), 2018.
7. A. Charalampi, C. Michalopoulou and C. Richardson. Determining the structure and assessing the psychometric properties of multidimensional scales constructed from ordinal and pseudo-interval items. *Communications in Statistics: Case Studies, Data Analysis and Applications*, 5, 1, 26-38, 2019. doi: 10.1080/23737484.2019.1579683
8. A. Charalampi, C. Michalopoulou and C. Richardson. Psychometric validation of constructs defined by ordinal-valued items. In C. H. Skiadas & C. Skiadas (Eds), *Demography of Population Health, Aging and Health Expenditures* (Chapter 20). The Springer Series on Demographic Methods and Population Analysis 50, 2020. doi: 10.1007/978-3-030-44695-6_20
9. A. Charalampi, C. Michalopoulou and C. Richardson. Validation of the 2012 European Social Survey measurement of wellbeing in seventeen European countries,

- Applied Research in Quality of Life, 15, 1, 73-105, 2020. doi: 10.1007/s11482-018-9666-4
10. A. Charalampi, A.E. Paltoglou, C. Michalopoulou and C. Richardson. Laying the groundwork for testing measurement invariance in large-scale cross-national studies. Manuscript in preparation, 2020.
 11. G.W. Cheung and R.S. Lau. A direct comparison approach for testing measurement invariance, *Organizational Research Methods*, 15, 2, 167-198, 2012. doi: 10.1177/1094428111421987
 12. G.W. Cheung and R.B. Rensvold. Testing factorial invariance across groups: A reconceptualization and proposed new method, *Journal of Management*, 25, 1, 1-27, 1999.
 13. G.W. Cheung and R.B. Rensvold. Evaluating goodness-of-fit indexes for testing measurement invariance, *Structural Equation Modeling*, 9, 2, 233-255, 2002.
 14. S.R. Cole, I. Kawachi, S.J. Maller and L.F. Berkman. Test of item-response bias in the CES-D scale: Experience from the New Haven EPESE study, *Journal of Clinical Epidemiology*, 53, 285-289, 2000.
 15. E. Davidov, A cross-country and cross-time comparison of the human values measurements with the second round of the European Social Survey, *Survey Research Methods*, 2, 1, 33-46, 2008.
 16. E. Davidov, B. Meuleman, J. Cieciuch, P. Schmidt and J. Billiet. Measurement equivalence in cross-national research, *Annual Review of Sociology*, 40, 1, 55-75, 2014. doi: 10.1146/annurev-soc-071913-043137
 17. E. Davidov, B. Muthén and P. Schmidt. Measurement invariance in cross-national studies: Challenging traditional approaches and evaluating new ones, *Sociological Methods & Research*, 47, 4, 631-636, 2018.
 18. European Social Survey, Measuring and reporting on Europeans' wellbeing: Findings from the European Social Survey, ESS ERIC, London, 2015. Retrieved from: https://www.europeansocialsurvey.org/docs/findings/ESS1-6_measuring_and_reporting_europeans_wellbeing.pdf
 19. European Social Survey Round 3 Data. Data file edition 3.7. NSD-Norwegian Centre for Research Data, Norway-Data Archive and distributor of ESS data for ESS ERIC, 2006.
 20. European Social Survey Round 6 Data. Data file edition 2.3. NSD-Norwegian Centre for Research Data, Norway-Data Archive and distributor of ESS data for ESS ERIC, 2012.
 21. B.F. French and W.H. Finch. Multigroup confirmatory factor analysis: Locating the invariant referent sets, *Structural Equation Modeling*, 15, 96-113, 2008. doi: 10.1080/10705510701758349
 22. K. Han, S.M. Colarelli and N.C. Weed. Methodological and statistical advances in the consideration of cultural diversity in assessment: A cultural review of group classification and measurement invariance testing, *Psychological Assessment*, 31, 12, 1481-1496, 2019.
 23. K. Jeffrey, S. Abdallah and A. Quick. Europeans' personal and social wellbeing: Topline results from Round 6 of the European Social Survey, ESS Topline Result Series – Issue 5, 2015. http://www.europeansocialsurvey.org/docs/findings/ESS6_toplines_issue_5_personal_and_social_wellbeing.pdf. Accessed 2 June 2016.
 24. E. Jung and M. Yoon. Two-step approach to partial factorial invariance: Selecting a reference variable and identifying the source of noninvariance, *Structural Equation Modeling*, 24, 1, 65-79, 2017. doi: 10.1080/10705511.2016.1251845

25. J. Karim, R. Weisz, Z. Bibi and S. ur Rehman. Validation of the eight-item Center for Epidemiologic Studies Depression scale (CES-D) among older adults, *Current Psychology*, 34, 4, 681-692, 2015. doi: 10.1007/s12144-014-9281-y
26. L. Kish, Multi-population survey designs: Five types with seven shared aspects, *International Statistical Review*, 62, 2, 167-186, 1994.
27. Y. Liu, R.E. Millsap, S.G. West, J-Y. Tein, R. Tanaka and K.J. Grimm. Testing measurement invariance in longitudinal data with ordered-categorical measures, *Psychological Methods*, 22, 3, 486-506, 2017. doi: 10.1037/met0000075
28. R.E. Millsap and J. Yun-Tein. Assessing factorial invariance in ordered-categorical measures, *Multivariate Behavioral Research*, 39, 3, 479-515, 2004.
29. S. Missine, C. Vandeviver, S. Van de Velde and P. Bracke. Measurement equivalence of the CES-D 8 depression-scale among ageing population in eleven European countries, *Social Science Research*, 46, 38-47, 2014.
30. D.L. Putnick and M.H. Bornstein. Measurement invariance conventions and reporting: The state of the art and future directions for psychological research, *Developmental Review*, 41, 71-90, 2016.
31. L.S. Radloff, The CES-D scale: A self-report depression scale for research in the general population, *Applied Psychological Measurements*, 1, 3, 385-401, 1977.
32. P. Raudenská, The cross-country and cross-time measurement invariance of positive and negative affect scales: Evidence from European Social Survey, *Social Science Research*, 86, 1-12, 2020. doi: 10.1016/j.ssresearch.2019.102369
33. J.-B. E. M. Steenkamp and H. Baumgartner. Assessing measurement invariance in cross-national consumer research, *Journal of Consumer Research*, 25, 1, 78-107, 1998.
34. L. Rutkowski and D. Svetina. Measurement invariance in international surveys: categorical indicators and fit measure performance. *Applied Measurement in Education*, 30, 1, 39-51, 2017. doi:10.1080/08957347.2016.1243540.
35. The ESS Sampling Expert Panel. Sampling guidelines: Principles and implementation for the European Social Survey, ESS ERIC Headquarters, London, 2016. <http://www.europeansocialsurvey.org/>
36. S. Van de Velde, P. Bracke, K. Levecque and B. Meuleman. Gender differences in depression in 25 European countries after eliminating measurement bias in the CES-D 8, *Social Science Research*, 39, 396-404, 2010.
37. S. Van de Velde, K. Levecque and P. Bracke. Measurement equivalence of the CES-D 8 in the general population in Belgium: A gender perspective, *Archives of Public Health*, 67, 15-29, 2009. doi: 10.1186/0778-7367-67-1-15
38. R.J. Vandenberg. Toward a further understanding of and improvement in measurement invariance methods and procedures, *Organizational Research Methods*, 5, 2, 139-158, 2002.
39. R.J. Vandenberg and C.E. Lance. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research, *Organizational Research Methods*, 3, 1, 4-70, 2000.
40. H. Xu and T.J.G. Tracey. Use of multi-group confirmatory factor analysis in examining measurement invariance in counseling psychology research, *The European Journal of Counselling Psychology*, 6, 1, 75-82, 2017. doi: 10.5964/ejcop.v6i1.120

Application of Stochastic Process in Speech Recognition

Mary Chriselda A ¹

¹ Student, St. Joseph's College (Autonomous), Bangalore, India.
(E-mail: chriseldaoliver23@gmail.com)

Abstract. This paper deals with the usage of the specific branch of machine learning: Reinforcement Algorithm, to model the speech recognition pattern. This is so devised such that it aims to improve the reliability through trial and error using a rewards approach. which can further be converted to a stochastic differential equation to provide results such as the probabilities attached to taking a particular course of action on the basis of the rewards received. This paper deals with the problem arising due to selection of the best pattern(phoneme) from the competing ones leading to a desirable outcome. It is a known fact that some values of the output are preferred to the others, a detailed analysis of the approaches used to solve this premise has been corroborated. Therefore, this paper summarizes the efficacy in selection of the pattern that can efficiently model the speech recognition algorithm at a future time period.

Keywords: Machine Learning, Reinforcement Learning Algorithm, Stochastic Differential Equation, Martingales, Optimal control

1. INTRODUCTION

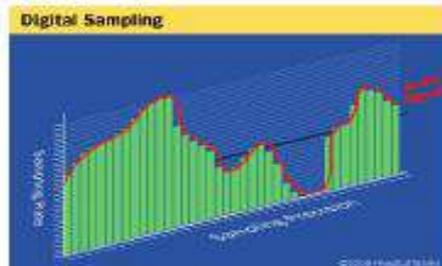


fig1: DIGITAL SAMPLING



Speech Recognition Software's proselytes analog signals into digital data by discretization of the sample data into computational content.

These software's playing a vital role in the proliferation of Artificial Intelligence is one of the potential vision for the future. Quoting Joel Walmsley - "The ethical use of artificial intelligence will rely upon the workings of that technology becoming more and more transparent" it is imperative to understand these models building techniques which could not only enhance but also improve the workings of these latest artificial intelligence systems [2] The technique used for the speech recognition software is through Reinforcement learning, a branch of machine learning which does not aim for the target output unlike the supervised learning, rather the input data is used to train the algorithm and then forecast how appropriate each of the output is to the desired result. As stated, that sometimes one would prefer a particular output to the other, the paper elaborates the approach used in achieving this optimal outcome using the stochastic process to model. It is highly essential to analysis these key elements while constructing the model:

- Existence of the pattern that could be nailed somewhere around scientific procedures. For instance, an feature vector for this situation scaling through time
- Availability of information applicable to the model
- Framing and testing of the speculation utilizing the example perceptions while preparing the calculation
- Sensitivity Analysis of the model

This paper has been organized into the following sections.

- Section 1 deals with the Introduction followed by the Literature Survey.
- Section 2 gives the mathematical proof for the construction of the Speech Recognition Model
- Section 3 is about the Analytical Method used in modelling
- section 4 deals with the Alternative Approach based on the pattern found from the analytical method.
- Section 5 deals with the Sensitivity Analysis
- Section 6 is Conclusion followed by Section 7 which is Acknowledgement and finally concluded with the References Section.

A. LITERATURE SURVEY

Recent years have witnessed tremendous progression in the field of Artificial Intelligence like Google Home, Alexa, Siri. This has in fact divulged into newer techniques being used ranging from the hidden markov models to neural networks, Gaussian models to MIXFIT[1].The first wave of interest spurred

around V.Mnih's first system game to play Atari 2600 which surpasses the human expert. This game takes into inclusion the past experiences through trial and error to whet its information processing ability[6]. The other example would be of the 2017 developed Pac-Man game using divide and rule strategy to classify each problem as a separate target to which a reinforcement learning was applied to[5]. The complicated locomotive behaviours illustrated in this paper[4] by Nicolas Hess and others show how robots in a simulated parkour like environment can learn through reward functions to a reinforcement algorithm. The other novelty was the use of this technique in the game of AlphaGo by Deep learning to play against strategic games by using the training dataset This algorithm has been widely applicable in the actuarial control cycles dealing with the financial asset pricing as well .An excellent example would be of a fund manager who is investing the assets in a pension fund the follow up of measuring the fund's solvency which can be affected by the unknown factors on the liability side might want a particular investment strategy[31] Another fabulous application is the genetic algorithm based on selective breeding where mutations are introduced to select the best surviving breeds that could continue its legacy to the next generation [32]

2. SPEECH RECOGNITION MODEL

Definition:

Computational approach to goal-directed learning from interaction with the environment using idealized learning situations.[1][23]

Reinforcement Learning is more like the 'carrot-stick' approach to allow for modifications using the user feedback, where most of the 'learning' or training of the dataset occurs when errors are committed allowing it to tweak its rules to improve the performance. In Reinforcement Learning the agent involved will learn to choose the action for each state u in S and at each time t in J . The possibilities of these actions A_t will give a chance that the agent would choose an action that would lead to a state s at time $t+1$. This is what we call by policy as to the function defined on a state to state on taking a particular action course .In this paper we are dealing with stochastic maps. For increasing the credibility of this probability, a rewards function is associated to ensure a more optimal value which would be closer to the accurate desired outcome. The construction of the model is divided into the following stages:

B. Stochastic Analysis

We define a stochastic process with continuous state space S and time set J such that:[25][26][27][28]

$$\{X_t: t \in J\} \tag{1}$$

The basic structures required for the stochastic process X_T are:

- sample path Ω indicates that every outcome ω in path Ω determines the sample path $X_T(\omega)$
- a set of events F which is applied on a subset X_S where $0 < s < t$, this F_T which belong to the family of natural filtrations describes the internal information gained by the process X_T up to t .
- state transition function

$$P(X_{t+1} = v | X_t = u, A_I) \quad (2)$$
- output function as

$$P(Y | X_t = u, A_I) \quad (3)$$

In this case we consider A_I to be a random variable belonging to $U(0,1)$ which are the set of actions to be taken by the agent.

For the filtration to be applicable on any martingale it is necessary that

- X_T is integrable for any $t \geq 0$
- X_T is measurable in F_T

Theorem:

For the simplification of the process we have considered the Markov assumption that any future development of the process depends on the current state alone and is independent of any past history of the process.[30][29][18]

$$P[X_t \in A | X_{s1} = x_1, X_{s2} = x_2, \dots, X_{sn} = x_n, X_s = x] \quad (4)$$

$$= P[X_t \in A : X_s = x] \quad (5)$$

for all times

$$s_1 < s_2 < s_3 < \dots < s < t$$

and the states

$$x_1, x_2, \dots, x_n, x \in S$$

and all the subsets A in S

C. MARTINGALES

A sequence of real-valued random variables

$$X_t : \omega \rightarrow R_t \geq 0$$

where

$$t \geq 0$$

on the probability space is called a martingale. On applying the filter to the stochastic process as follows when $E[|X_T| < \infty]$ for any $t \geq 0$ the obtained martingale is as follows

$$E[X_t | F_s] = X_s \quad (5)$$

for any

$$0 < s < t$$

Lemma:

Consider a random variable v such that

$$X_t = E[v | F_t] \quad (6)$$

for all t in J . [14][15][16][17][18]

By the tower property of conditional expectations, we can rewrite as

$$E[X_t | F_{t-1}] = E[E[v | F_t] | F_{t-1}] = E[v | F_{t-1}] = X_{t-1} \quad (7)$$

for all t in J .

This equation can be rewritten as following

$$E[X_{t+1} - X_t | X_t] = 0 \quad (8)$$

for any t in J .

This means that the process has a time invariant property in the mean as the mean of a martingale is constant

$$E[X_t] = E[X_0] \quad (9)$$

for all t in J

D. Hamilton-Jacobi Bellman Equation

Proof:

Definition:

The Bellman's principle of optimality states that irrespective of the initial state and action, the remaining action must constitute an optimal mapping with respect to the state resulting from the first action.[20][21][22]

Theorem:

For any $x \in \mathbb{R}^n$

The value function is given as

$$v(s, x) = \sup_{a \in U(s,t)} E \left[\int_s^t f(s, X_s, a_s) ds + v(t, X_t | X_s = x) \right] \quad (10)$$

$$0 < s < t < T$$

We have introduced this value function which is derived from the Reinforcement Learning given as

$$v = E[R_t | X_t = i, A_t] \quad (11)$$

in which the reward R_t depends on the probability that the action A_t leads to a desirable outcome. The agent then finds the supremum of the rewards. This introduction to the Hamilton-Jacobi Bellman equation (16) focusses on the concept that the agent would learn by the law of effect. The newer probabilities are then appended by taking a particular action on the basis of the rewards received.

Using the lemma (12,13) we can rewrite equation (16) as

$$E \left[\int_s^t f(s, X_s, a_s) ds + E \left[\int_t^T f(s, X_s, a_s) ds + kX_t \mid F_t \right] \mid X_s = x \right] \quad (12)$$

$$= E \left[\int_s^t f(s, X_s, a_s) ds + k(X_t) \mid X_s \right] \quad (13)$$

$$= k(s, X_s, a_s) \leq v(t, X_s) \quad (14)$$

Hence, we can rewrite it as

$$k(s, x, a) \leq \text{Sup}_{a \in U} E \left[\int_s^t f(s, X_s, a) ds + u(s, X_s) \mid X_s = x \right] \quad (16)$$

We consider $\epsilon \geq 0$ then there exists

$$a \in U(s, t) \quad (17)$$

Such that for every $\delta > 0, x_i$'s

$$v(s, x_i) \leq \epsilon + k(t, x_i, a) \quad (18)$$

Using the Lipschitz condition,[17][18][19]

$$v(s, X_s) - v(t, x_i) \leq C |X_s - x_i| < C \quad (19)$$

We fix $\delta > 0$ such that

$$2C\epsilon < \epsilon \quad (20)$$

$$\begin{aligned} v(s, x_i) &\leq 2\epsilon + k(s, x_i, a) \leq 3\epsilon + E \left[\int_s^t f(s, X_s, a_s) ds \right. \\ &\left. + E \left(\int_t^T f(s, X_s, a_s) ds + k(Y_s) \mid Y_s = S_s \right) \mid X_s = x \right] \quad (21) \end{aligned}$$

This implies that the controls $a \in U(s, t)$ for each i .

E. Existence of Solution

Lemma:

By using the Martingale Optimality Theorem

We fix the initial state x_0 at $t=0$. Let

$$M_t = \int_0^t f(s, X_s, a) ds + v(t, X_t) \quad (22)$$

Then for any control $a \in u[0, T]$ X_t defined on filtration F_t is a super martingale

This shows that a is optimal control action if the given Hamilton-Jacobi-Equation derived is a martingale[7][8][9]

Proof:

For any $a \in U$

$$v(t, X_t) \geq E \left[\int_s^t f(s, X_s, a) ds + v(s, X_s) \mid F_t \right] \quad (23)$$

$$\begin{aligned} X_t &\geq \int_0^t f(s, X_s, a) ds + E \left[\int_s^t f(s, X_s, a) ds + v(s, X_s) \mid F_t \right] \\ &= E[X_t \mid F_t] \quad (24) \end{aligned}$$

$$k(0, x, a) = E(X_t) = E(X_0) = v(0, x) \quad (25)$$

So, this satisfies the martingale property so a is indeed the optimal action needed to be taken.

F. Uniqueness of Solution

We express the Hamilton Jacobi Bellman Equation as a stochastic differential equation

Proof:

Let v be a random variable which is independent of the

σ -algebra generated by $W_s, s \geq 0$ and such that

$$E[|Z|^2] < \infty \quad (26)$$

This satisfies the martingale property. Then the stochastic differential equation

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t \quad (27)$$

$0 \leq T, X_0 = a$

has a unique t -continuous solution $X_t(w)$ with the property that $X_t(w)$ is adapted to the filtration F_t generated by W and

$$E\left[\int_0^T |X_T|^2 dt\right] < \infty \quad (28)$$

$$v(t) = E\left[|x_t - \widehat{X}_t|^2\right] + 3(1+t)^2 D^2 \int_0^T v(s) ds \quad (29)$$

This equation is similar to equation (15)

Put $v(s)=0$. So,

$$P[|X_t - \widehat{X}_t|] = 0 \quad (30)$$

This corresponds to the equation (34) which proves the uniqueness of the solution.

G. Completeness of Solution

We consider the Hamilton-Jacobi-bellman equation as a non-linear partial differential equation

Proof:

Theorem:

As we define U to be the open set of control actions in \mathbb{R}^n and $a_i \in U(s,t)$ such that $f \in (U \times \mathbb{R} \times \mathbb{X}^n)$ be a strictly positive function. Then a_i solves the Monge-Ampere's Equation if and only if it solves the Hamilton-Bellman-Jacobi equation.[3]

$$\det [D^2 v(x)] = f(x, v(x), D v(x))^n \quad (31)$$

which is a positively defined on U .

$$\min_{a \in U} Tr[AD^2 v(x) - t f(x, v(x), D(v(x)))] (det A)^{\frac{1}{t}} = 0 \quad (32)$$

$$\min[\sum_{j=1}^t \frac{\partial^2 v(x)}{\partial x_j^2} - t f(x, v(x), D v(x)) \sum_{j=1}^n A_j^{\frac{1}{t}} | A_i \geq 0, \sum_{j=1}^n A_{I_j} = 1] = 0$$

For Completeness we use the Geometric-Arithmetic inequality. Let $t \in J$ [15][16]

$a_i \geq 0$

Then,

$$\prod a_i^{\frac{1}{t}} \leq \frac{\sum a_i}{t} \quad (34)$$

Using Young's inequality putting $t=1$

$$\prod_{j=1}^{t+1} a_i^{\frac{1}{t+1}} = a_{t+1}^{\frac{1}{t+1}} \prod a_i^{\frac{1}{t+1}} \quad (35)$$

$$\leq \frac{1}{t+1} a_{t+1} + \left(1 - \frac{1}{t+1}\right) \prod a_i^{\frac{1(1-\frac{1}{t+1})-1}{t+1}} \quad (36)$$

$$\leq \frac{1}{t+1} a_{t+1} + \frac{t}{t+1} \prod a_i^{\frac{1}{t}} \quad (37)$$

and induction hypothesis

$$\prod_{j=1}^t a_i^{\frac{1}{t}} \leq \sum \frac{a_i}{t} \quad (38)$$

We can conclude that

$$= \prod_{j=1}^{t+1} a_i^{\frac{1}{t+1}} \leq \frac{a_{t+1}}{t+1} + \frac{a_i}{t+1} \quad (39)$$

$$\frac{\sum_{j=1}^{t+1} a_i}{t+1} \quad (40)$$

This proves the completeness of the equation.

Using the Bellman's principle of optimality the optimal control action for the agent advised to take implied by solving the HJB using Martingales to prove the existence and uniqueness of the value function or the output function. The completeness of this solution was further proven by using the Theorem[3]. The reason why these steps were undertaken was to negate the premise that there could have existed many solutions to the HJB equation, the restriction is supplied by adding boundary conditions in the form of the Markovian Property and the value function. The proof of the equivalence between the Monge-Ampere's and HJB equations also conveys the fact that the control action belonging to the set U is a compact set through the completeness proof. This invigorates the fact that one could use

Theorem: The Heine-Borel theorem states that a set is compact if any of its open cover has finite subcover.

It is also a known fact that a bounded sequence in R^n is always convergent. [38]

Theorem: The Bolzano Weistrass Theorem states that a set in R^n is sequentially compact if it closed and bounded. [35,36,37]

Definition: A real number l is said to be the limit of the sequence $a_i \in U$ as i tends to ∞ if for every $\epsilon > 0$, there exists a positive integer m such that

$$|U - l| < \epsilon \forall n > m$$

Lemma: From the theorem it can be proved that the limit of the convergent sequence is unique [35,36,37] and from this it can also imply the fact that a continuous mapping on this compact set U is uniformly continuous.

Therefore, it can be proved that the control action necessary for obtaining the value in the output function is unique, bounded and uniformly continuous. This verifies the premise stated in this paper that some actions could be preferred over others which are more profitable in obtaining the valid results. Hence, the method proves the optimal strategy for obtaining the accurate results when dealing with the phonemes in a Speech Recognition Software.

3. ANALYTICAL METHOD

Steps ensured for computing the results:

- Firstly, the 'raw' data obtained from the audio file has been converted to the text format and stored.
- The data is then stored in a spreadsheet and imported for use; the type of data used in this are ordinal, continuous
- The data is prepared for use by cleaning, checking for missing values and errors

H. Feature Scaling

It is important to introduce weights w_i which somewhat play the role of β 's in the regression model. Since we have employed techniques from Reinforcement learning it is quite apt to introduce scaling of the variable. In place of the rewards function we can introduce a loss function as well.

$$L(w_1, w_2, \dots, w_i) + \lambda \sum_{i=1}^J w_i^2 \quad (41)$$

It is quite obvious that the complexity of the model can be viewed through the weights.

I. Training, Validation and Testing

The data is divided into a 3:1:1 split. It is in this stage that the fitting of the data model can be estimated from the parameters. On modelling, it is imperative to find the hidden patterns on which further analysis can be done. This was identified to follow the Hidden Markov Model on further computation as seen in the Results section below. The training, testing and validation is done to ensure the performance of the dataset correlates with the objective under study.

J. Model Performance

For this model we use the Vapnik-Chervonenkis inequality[23] as a proposed measure to address the difference in the model in sample error to the model out-sample error by comparing with a threshold value.

$$P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq K \quad (42)$$

Here K is the threshold value. This model performance has been computed rather than manually solving for using the Scenario Weights for Importance Measurement (SWIM), the threshold value considered are the probabilities that are stressed. It is ensured that the value of the probabilities of a small reward is kept below a specific probability.

4. ALTERNATIVE APPROACH

K. Algorithm

The algorithm for simulations makes use of the hidden markov model given as the following. The HMM can be used as a generator to give the observation sequence.[11][12][13]

- T is the number of observations in sequence
- Choose the initial state $m_i=S_i$
- According to the initial state distribution π set $t=1$
- Choose $O_t=v_n$ according to the the probability distribution in state S_i
- Transit to a new state $m_{i+1}=S_i$
- Set $t=t+1$ return step 3
- else if $t < T$ otherwise terminate

Using the above algorithm, the output has been simulated using Python to observe the most probable sequence of the phoneme 'p' and 'd' and its probability. This is using the HMM (Hidden Markov Model) approach for the Speech Recognition model created. The probability of the transitions to each state was simulated using the input phoneme 'p' and 'd' for simplicity. This analytical method was specifically mentioned in this paper due to its feasible usage in economic and financial models, weather forecasting etc. But this approach lacked the credibility as the rewards or loss function for every wrong output was not taken into consideration.

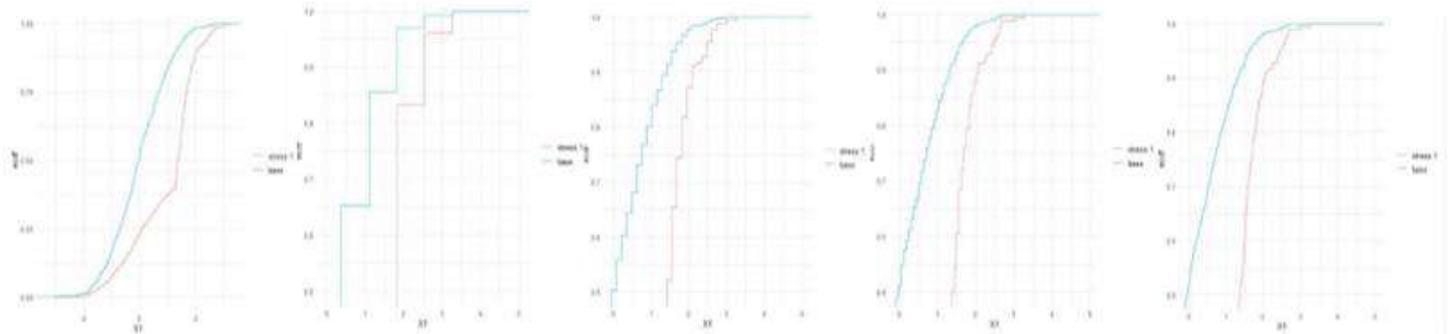
```
>>>
===== RESTART: C:\Python27\vib2.py =====
(3.3929083008000003e-08, ['p', 'd', 'd', 'p', 'd', 'd', 'd', 'd', 'd', 'p'])
>>> |
```

Fig 2: OUTPUT GENERATED FOR THE MOST PROBABLE SEQUENCE AND ITS PROBABILITY

5. SENSITIVITY ANALYSIS

In the above sections we have already proved that the rewards function in Section (II) can lead to an optimal result. This is further substantiated using the programming techniques to simulate these observations. For the sensitivity analysis we have made use of the Scenario Weights for Importance Measurement (SWIM) including the possibilities of the simulated scenarios by changing the input data to obtain the patterns. It was proven to be specially advantages as the methodology was useful for stochastic models like the one modelled in this paper. Adding to these benefits, the time span of cumbersome simulations was mitigated by working on a single set of simulated scenarios. This also particularly gives us a valuable judgement of the change in the distributions with the increase in the size of n as seen in the graphs plotted above. We have taken the stress which is the modification of a particular variable of interest in terms of the probabilities for different values of n ranging from 10,50,100,500,1000. It is noticeable that with the change in the value of n the curve is smoothed due to the graduation of the crude estimates by using the scenario weights. The cdf of the empirical distribution of the stressed probabilities under the scenario weights were computed and plotted for different values of n after extracting the realizations of the stochastic model the function generated act as a new object on which the Reinforcement algorithm was applied to find the rewards and the expected time taken for obtaining the results. Further the model performance is indicated by the different sensitivity measures which compare the stressed probability of the stochastic model with the baseline model.

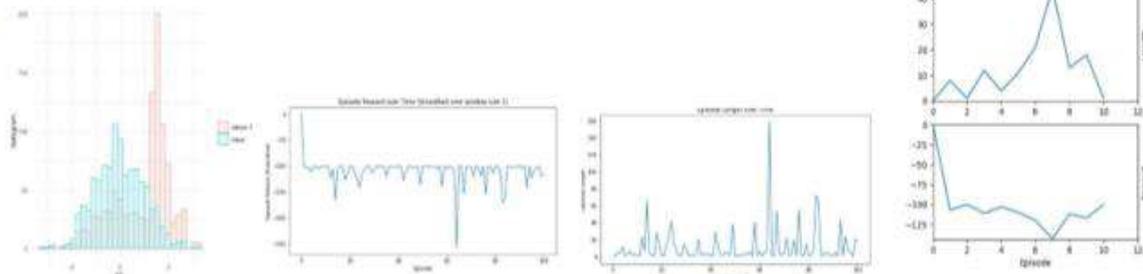
Hence this approach is far more reliable in terms of the empirical distribution and simulation of realizations with the Reinforcement algorithm by adding rewards function to provide a far more optimal outcome than the Hidden Markov Models.[24][30]



GRAPHS FOR THE SCENARIOS SIMULATED UNDER THE STRESSED PROBABILITIES

Chi-squared test for given probabilities
data: h
X-squared = 25.766, df = 49, p-value = 0.9974

stress	type	X1
stress 1	Gamma	1.00
stress 1	Kolmogorov	0.44
stress 1	Wasserstein	3.38



SENSITIVITY ANALYSIS

6. CONCLUSION

The Machine Learning has always been quotes as the " good old fashionable AI" by thespians across this field remarking the feasibility to reproduce the research, by tuning it with the time and trend. Most of the learning as proposed in this Speech Recognition Model takes inclusive learning through simulation of observations which could not have been curated otherwise. The amalgamation of Reinforcement Learning in the Hamilton-Jacobi-Bellman paved way to make use of the martingales and converting it into a stochastic differential equation proving the existence, uniqueness and completeness of the optimal solution. These results were further validated using the R-software and Python to help visualize the implications of the model. The proposed paper deals with the algorithm for pattern finding in a sequence of phoneme giving out its probability, mathematical proof for the construction and validation of the model. These findings also answer how to choose a particular action over the other with the help of the rewards function introduced in the model. As Further Clarke famously quipped "any sufficiently advanced technology is indistinguishable from magic", by the manner these modern AI systems have progressed definitely far outstrips the of human cognition.

7. ACKNOWLEDGMENT

The author would like to thank the reviewers for their constructive and insightful comments with respect to this paper. Also, a special note of gratitude is rendered to all the near and dear ones who supported immensely in the fruition of this paper.

8. REFERENCES

- [1] Hajime Masayuki Shigenobu , Reinforcement Learning by Stochastic Hill Climbing on Discounted Reward, .
- [2] Imran Masood , A two stage interval stochastic programming, . Pro-ceedings of the Twelfth International Conference on Machine Learning, Tahoe City, California, July 9–12, 1995, Pages 295-303
- [3] Mao .X, Stochastic Differential Equations and Applications ,Ele-viser,2007.
- [4] NicolasHeess DhruvaTB SrinivasanSriram JayLemmon JoshMerel Greg-Wayne YuvalTassa TomErez ZiyuWang S.M.AliEslami MartinRied-miller DavidSilver, EmergenceofLocomotionBehaviours inRichEnviron-ments, DeepMind
- [5] HarmvanSeijen MehdiFatemi JoshuaRomoff RomainLaroche Tavian-Barnes JeffreyTsang, Hybrid Reward Architecture for Reinforcement Learning, Deep Mind
- [6] vlad koray david alex ioannis daan martinriedmille, PlayingAtariwith-DeepReinforcementLearning, Deep Mind
- [7] Oksendal .B, Stochastic Differential Equations,
- [8] Papoulis and Pillai .U, Probability Random Processes and stochastic process,
- [9] Protter .PE, Stochastic Differential Equations, Stochastic Integration and

- Differential Equations Springer 249-361,2005.
- [10] Raisignhanian .MD, Ordinary and Partial Differential Equations, .
 - [11] Samuel .Karlin Howard ,M Taylor, A first course in stochastic processes,3rd edition Academic Press
 - [12] Shapiro Dentcheva D, Lectures On Stochastic Modelling,
 - [13] Shapiro .Phipott, A tutorial on stochastic programming,
 - [14] Oliver .Knill, Probability and Stochastic Processes with Applications,2002 OverSeas Press
 - [15] Alexandre J. Chorin ,and Ole H. Hald, Stochastic Tools for Mathematics and Science, Berkeley, California March, 2009.
 - [16] Gustav Ludvigsson, Kolmogorov Equations,June 2003Uppasala Universtat
 - [17] C.W. Gardiner, Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences Second Edition, 2nd edition Springer
 - [18] Rabi N Battacharya Edward. C Waymire , Stochastic Processes with Applications, SIAM
 - [19] Dr. Andreas Eberle , Stochastic Analysis,
 - [20] Ranu Dixit Navdeep Kaur2, Speech Recognition Using Stochastic Approach: A Review,International Journal of Innovative Research in Science, Engineering and Technology Vol. 2, Issue 2, February 2013
 - [21] William Feller , An Introduction to Probability Theory and its Applications, Wiley Second Edition
 - [22] John G Kemeny J Laurie Snell and A.W knapp: Finite Markov Chains, Springer
 - [22] Hamilton-Jacobi-Bellman Equations Analysis and Numerical Analysis
Iain Smears
 - [23] Statistical Learning Theory in Reinforcement Learning & Approximate Dynamic Programming A. Lazaric & M. Ghavamzadeh (INRIA Lille – Team SequeL ICML 2012
 - [24] K.L Chung:Elementary Probability Theory
 - [25] Sheldon M Ross: A First course in Probability
 - [26] R Ash : Basic Probability Theory
 - [27] P. G Hoel ,S.C Port and C.J Stone: Introduction to Probability Theory.
 - [28] B Prabhakar Rao,TSR Murthy :Probability Theory and Stochastic Processes
 - [29]Sheldon M Ross:Stochastic Processes,Wiley Series in probability and mathematical statistics
 - [30]G Casella and R L Berger:Statistical Inference.
 - [31] A.Chalk and C.McMurtrie :’ A practical introduction to Machine Learning concepts for actuaries”, Casualty Actuarial Society E-forum, Spring 2016
 - [32]Adarsh Sehgal,Hung Manh,La , Sushil J,Louis ,Hai Nguyen ,Deep Reinforcement Learning using Genetic Algorithm for Parameter Optimization,Neural and Evolutionary Computing 2019.
 - [33] Pesenti Silvana M and Bettini A,Miliosovich P, Tsankas (2020),”Scenario Weights for Importance Measurement (SWIM)”
 - [34] Pesenti SM,Miliosovich P, Tsankas A (2019): Reverse Sensitivity analysis testing : What does it take to break the model?” European Journal of Operations Research 274(2654-670)
 - [35] Tom M Apostol : “Mathematical Analysis”

Real estate pension schemes: modeling and perspectives

Valeria D'Amato, Emilia Di Lorenzo, Gabriella Piscopo, Marilena Sibillo and Roberto Tizzano

Abstract

The paper is focused on a contractual scheme where an immediate life annuity is obtained by paying a single premium in the form of real estate rights (RERs). Such contract is framed within the varied macrocosm of personal pension products, that proves to be an increasingly attractive and scientifically stimulating tool in the economic and social context in which we currently find ourselves. The basic structure requires the lender to pay the borrower/homeowner a life annuity against a part or the totality of the future liquidation value of his home at the time of his death. In this work we will explore some characteristics of the product, paying attention to the risk sources which impact on it.

Key words: Real Estate financial market; Demographic risk; Reverse mortgage;

Valeria D'Amato, Pharmacy Department, Campus Universitario, University of Salerno, Italy
e-mail: vdamato@unisa.it

Emilia Di Lorenzo, Gabriella Piscopo, Roberto Tizzano, Department of Economic and Statistical Sciences, University of Naples Federico II, Italy
e-mail: emilia.dilorenzo@unina.it, gabriella.piscopo@unina.it, roberto.tizzano@unina.it

Marilena Sibillo, Department of Economics and Statistics, Campus Universitario, University of Salerno, Italy
e-mail: msibillo@unisa.it

§1. A personal pension product on real estate rights.

The critical aspects of pension systems, viewed in relation to the problems of private savings, have long been the subject of scientific, legal, as well as political debate. The need to set up a reliable system able to support public and professional pension systems (with the latter often struggling, too) is a critical issue for governments and regulatory institutions. In the European context (cf. EIOPA (2014)) the relevant authorities have set up a definition and diffusion process of contractual lines, aiming at integrating existing pension systems and, above all, adapting to Europeans' mobility needs in the context of an ever-evolving job market (cf. EIOPA (2019)). In this context, some guidelines have been identified, aiming at homogenizing and standardizing, at least in a number of key areas, personal pension products' contractual profiles (which constitute the so-called "third pillar" in social security system). Their purpose is to ensure transparency, homogeneity and adequacy in terms of pricing and yield. This project responds to the EU's more general political need to contribute synergistically, via its many economic and financial sectors, to strengthening "the much needed, efficient and sustainable Capital Market Union", as explicitly maintained in a recent EIOPA Report of 29 / 11/2019 (cf. EIOPA (2019)).

Even though the governance, supervision and taxation issues of personal retirement products are often purely political in nature, scholars in this field remain puzzled by a number of questions. One can mention, in this sense, an extensive actuarial literature dedicated to issues pertaining to the

protection of potential buyers. Bovenberg and Nijman (2016), Van Binsbergen et al. (2014), Maurer et al. (2013), Denuit et al. (2011), D'Amato et al. (2018) and (2017), Chłoń-Domińczak (2016), for example, devise solutions based on risk sharing as well as on profit sharing in different pension schemes. Although the above list is inevitably limited, these works consistently feature a focus on the question of longevity risk: any contractual design must deal with European societies' aging populations. In particular, (cf. EUROSTAT (2018) and (2019), D'Amato et al. (2019), Loichinger et al. (2016)) the relationship, in various member states, between working life expectancy and life expectancy (LE) at the age of 50 it is less than a third, reaching 31.61% in Italy and 35.08% in Spain.

Longevity risk, however, is not a source of risk affecting only pension annuity providers, but also ordinary citizens, who "risk" outliving their savings. In consequence, actuarial literature has expanded the scope of its contributions, and provided proposals for retirement savings planning that take into account changing economic environments. In this context, particular emphasis is placed on those who, due to registry and/or income problems, are unable to access traditional credit lines to face needs they could not sustain with pension incomes alone. These products include Reverse Mortgages, by which elderly homeowners receive credit, by exchanging - entirely or partially - the value of their home for a sum of money (a lump sum or a periodic income). By virtue of the rights guaranteed by these contracts, homeowners will retain the right to live in his own home until they die. Several papers (cf. Alai et al. (2014), de la Fuente et al. (2018), Di Lorenzo et al. (2020a), Hanewald et al. (2016), Li et al. (2019), Nakajima et al. (2017), Shao et al. (2015)), examine these contracts' pricing and risk management. Additionally, some studies also focus on securitization proposals, viewed as risk management tools and, at the same time, as potential leverage for the development of the RMs market itself, which benefit from a sufficiently developed secondary market (cf. Di Lorenzo et al. (2020b), Huang et al. (2019), Merton et al. (2016), Wang et al. (2008)).

The proposal presented by D'Amato et al. (2019) of a custom-made private pension plan able to providing a flexible addition to other sources of retirement pension links up with this research context. This pension product consists of an immediate life annuity, triggered by the payment of a single-premium, in form of real estate rights (RERs), such as transferring the full or the bare property of a house or a similar realty to an insurer. The installments depend on the fair value of the transferred RER at the contract's issue, the life expectancy of the insured, the expected return of the real estate market and a risk-free financial rate; however, the definition of the valuation parameters is appropriately contextualized and expressed in relation to the insurer's position and financial "capabilities".

In this work in sect.2 we present the Real Estate Pension Scheme (REPS) in its basic actuarial connotations. In the third section, we focus in particular on the control of financial risk under this contract, detailing the components of the insurer's actuarial benefit assessment rate. In section 4, the source of financial risk is studied in its impact on the amount of the annuity payment due to the insured. Several tables show the sensitivity of the installment to changes in the significant parameters identified in our analysis. This study is extended to a number of European countries in order to make the comparison among various real estate markets. Some conclusions close the paper.

§2. The installments in the REPS pensions schemes.

The Real Estate Pension Scheme (REPS) we consider in this paper is based upon an underlying real estate right (RER), while the counterparties are the owner of such RER, who also serves as the insured, and the insurer. The insurer acquires from the insured, at the contract issue time, the single premium in the form of the real estate right (RER) (all or part of the property) that the insured transfers to him/her at the same time. The insurer's obligations consist of the payment of a periodic installment, which we assume to be constant, for the entire duration of the insured's life. The insured also maintains the right to reside in that property until the time of his/her death.

The exchange of cash flows at the issue time, which must achieve the contractual equilibrium at the issue time, is based on the value attributable to the transfer of all or part of the property by the homeowner to the insurer, in a form in which the insurer will only have full availability of the property at the time of the insured's death. This value shall be financially equivalent to the value of the life annuity that the insurer will pay to the policyholder throughout his/her life. This means that, from a strictly actuarial point of view, the transfer of all or part of the ownership rights from the insured to the insurer is equivalent to the single premium paid by him/her at the contract issue time. The amount of the single premium strictly depends on the RER fair value.

Let's indicate it by A.

In $t = 0$ the following relation holds:

$$A = R \ddot{a}_{T_x | i^*} \quad (1)$$

in which R is the installment the insurer will pay at the beginning of each period till the insured aged x is alive, T_x is the his/her residual lifetime at the issue time, and $\ddot{a}_{T_x | i}$ is the value, at the issue time, of an anticipated unitary life annuity. We are interested in the calculation of R, which is the financial commitment of the insurer: on its value the interest rate i^* at which the life annuity will be calculated, has a fundamental and relevant impact.

§3. Financial risk: considerations on the annuity valuation interest rate.

This section is dedicated to the analysis of the financial risk arising from the interest rate applied in the valuation of the insurer's obligations in formula (1).

The key risks involved in determining that rate are:

- (a) the risk related to the volatility of the RER value, which, in turn, depends upon real estate market price perspective.
- (b) the risk related to the volatility of capital markets. The insurer may need to resort to financial market, for example by issuing long term indexed rate bonds, to finance the payment of the annuity.

Equation (1) shows that the hedging action is carried out by controlling the interest rate i^* , that must be compatible with long-term interest rate perspectives. Should this rate turn out to be lower than the one he pays on the bonds he issues, the insurer would incur a financial loss and his solvency might be compromised. Of course, the longer the time contract horizon (that is the policyholder life expectancy), the higher the risk that financial rates unexpectedly creep up.

To address the above said risks, we propose a hedging strategy that is based on the interaction of the three relevant rates involved into a REPS valuation, according to the following relation which links the three relevant rates in the transaction we're dealing with here (cf. D'Amato et al., 2019):

$$i^* = \alpha i_{re} - \beta i_{rf} \quad (2)$$

in which:

i^* is the discount rate to be applied for the assessment of the insurer's future obligations

i_{re} is the expected growth rate of the real estate market

i_{rf} is the free risk financial rate for long-term investments

and α and β are two appropriate parameters through which to hedging strategy is developed.

The strategy consists of the two following actions:

Action 1: i_{re} is calibrated in order to hedge against the risk associated with the volatility of the real estate market and for this purpose it will be weighted using the parameter α , $0 < \alpha \leq 1$. α measures the impact of the uncertainty on the future evolution of the rates of return in the real estate market, increasing as the values attributed to this parameter decrease. The maximum value $\alpha = 1$,

represents the insurer's confidence in a future stable performance of the real estate market. As the value assigned to α decreases, the impact of that market's instability increases. All other things being equal, this is achieved in a decrease of i^* and consequently in an increase of R .

Action 2: i_{rf} is calibrated by assigning an appropriate value to the β parameter, $\beta \geq 1$, that weights the impact on i^* of the cost of capital incurred by the insurer that needs to resort to the financial market to fulfill the obligations. Such cost is based on the long-term free risk and increases at the worsening of the insurer financial position. The limit case of $\beta = 1$ represents the case of the highly liquid insurer, for which i_{rf} is simply an opportunity cost rate. Increasing values of β , due to a bad financial position of the insurer, determine increasing levels of i^* that is lower levels of R . Since at equal REPS premiums policyholders are attracted by higher levels of R , the competitiveness of the insurer depends upon the quality of their financial position: the better it is, the more competitive they are.

§4. Evidences on the sensitivity of the installments to the financial risk drivers.

The α and β parameters, calibrating respectively the impact on i^* of the uncertainty on the future evolution of the rates of return in the real estate market, and of the cost of capital incurred by the insurer if necessary, have an influence on the amount of the installment due to the insured. In this section we are going to provide some numerical evidences to the aim of pointing out the sensitivity of the installments when α and β assume different values. In order to give an overview of more than one geographical context, we have extended the analysis to two European countries: Italy and Germany.

We will calculate the level of the installments for the two countries on the basis of the house return forecasts reported in Table 1, considering the house price equal to 250,000.

We will describe three different scenarios assigning to the two parameters α and β to the following values:

$$\alpha = 0.5, 0.8, 1.0$$

$$\beta = 1.0, 1.5, 2.0$$

and the REPS contract we consider is issued on a male aged 55 and 60, with anticipated installments paid at the beginning of each month if the insured is alive. The survival probabilities have been forecasted by means of a standard Lee Carter model. The interest rate i_{re} has been estimated (cf. D'Amato et al. 2019) as the average value of the future house returns got by the best ARMA-GARCH model associated with the lowest information criteria for each country. In Table 1 as an example we report the real estate rates got for six European countries by means of this procedure:

Table 1. The real estate interest rates

Country	i_{re}
Italy	2.5%
France	5.0%
Germany	4.0%
Spain	2.0%
Austria	5.0%
Greece	2.0%

The following group of tables are referred to two values of the risk free interest rate: $i_{rf} = 0$, highlighting only the impact of α on the installment, and $i_{rf} = 1\%$. We will present the results as obtained for the two countries, if $\alpha = 0.50, 0.80, 1.00$ and $\beta = 1.00, 1.50, 2.00$, accordingly with eq. 2.

Italy

In Table 1 and Table 2 we can read the interest rates determined by equation 2 in two different risk free interest rate scenarios.

$i_{rf} = 0.00\%$	α		
	0.50	0.80	1.00
i^*	1.25%	2.00%	2.50%

Table 1. i^* values, $i_{rf} = 0.00\%$, for different values of α

$i_{rf} = 0.01$		α		
		0.50	0.80	1.00
β	1.00	0.0025	0.010	0.015
	1.50	-0.0025	0.005	0.010
	2.00	-0.0075	0.000	0.005

Table 2. i^* values, $i_{rf} = 0.01$ for different values of α and β

Tables 3 and 4 report the installments determined for three different ages at issue, $x = 55, 60, 65$ in the two cases of risk free interest rate. For $i_{rf} = 0.00\%$ the β parameter is irrelevant. We will consider not significant from the point of the numerical application the cases in which the values of the parameters α and β determine negative values for i^* .

$i_{rf} = 0.00\%$	α		
	0.50	0.80	1.00
$x = 55$	4713.25	4839.18	4918.74
$x = 60$	5603.96	5677.88	5724.32
$x = 65$	7389.46	7667.23	7857.00

Table 3. Installments for different values of α and different values of the age at issue; $i_{rf} = 0.00\%$.

$x = 55$		α		
$i_{rf} = 0.01$		0.50	0.80	1.00
β	1.00	4368	4479	4548
	1.50	-	4406	4479
	2.00	-	4328	4406

$x = 60$		α		
$i_{rf} = 0.01$		0.50	0.80	1.00
β	1.00	5261	5315	5347
	1.50	-	5280	5315
	2.00	-	5242	5280

Table 4. Installments for different values of α and different values of the age at issue; $i_{rf} =$

France

In Table 5 and Table 6 we report the interest rates determined by equation 2 in two different risk free interest rate scenarios.

$i_{rf} = 0.00\%$	α		
	0.50	0.80	1.00
i^*	2.50%	4.00%	5.00%

Table 5. i^* values, $i_{rf} = 0.00\%$, for different values of α and β

$i_{rf} = 0.01$		α		
		0.50	0.80	1.00
β	1.00	0.016	0.030	0.040
	1.50	0.010	0.025	0.035
	2.00	0.005	0.020	0.030

Table 6. i^* values, $i_{rf} = 0.01$ for different values of α and β

Tables 7 and 8 report the installments determined in three different ages at issue, $x = 55, 60, 65$ in the two cases of risk free interest rate. For $i_{rf} = 0.00\%$ the β parameter is irrelevant.

$i_{rf} = 0.00\%$	α		
	0.50	0.80	1.00
$x = 55$	4935.13	5146.05	5269.03
$x = 60$	5745.69	5869.12	5940.86
$x = 65$	7857.00	8448.98	8863.07

Table 7. Installments for different values of α and different values of the age at issue; $i_{rf} = 0.00\%$.

$x = 55$		α		
$i_{rf} = 0.01$		0.50	0.80	1.00
β	1.00	4579	4741	4836
	1.50	4499	4687	4791
	2.00	4427	4629	4741

$x = 60$		α		
$i_{rf} = 0.01$		0.50	0.80	1.00
β	1.00	5376	5446	5483
	1.50	5339	5423	5465
	2.00	5304	5398	5446

Table 8. Installments for different values of α and different values of the age at issue; $i_{rf} = 0.01$

Austria

In Table 5 and Table 6 we report the interest rates determined by equation 2 in two different risk free interest rate scenarios.

$i_{rf} = 0.00\%$	α		
	0.50	0.80	1.00
i^*	2.50%	4.00%	5.00%

Table 5. i^* values, $i_{rf} = 0.00\%$, for different values of α and β

$i_{rf} = 0.01$		α		
		0.50	0.80	1.00
β	1.00	0.016	0.030	0.040
	1.50	0.010	0.025	0.035
	2.00	0.005	0.020	0.030

Table 6. i^* values, $i_{rf} = 0.01$ for different values of α and β

Tables 7 and 8 report the installments determined in three different ages at issue, $x = 55, 60, 65$ in the two cases of risk free interest rate. For $i_{rf} = 0.00\%$ the β parameter is irrelevant.

$i_{rf} = 0.00\%$	α		
	0.50	0.80	1.00
$x = 55$	4935.13	5146.05	5269.03
$x = 60$	5745.69	5869.12	5940.86
$x = 65$	7857.00	8448.98	8863.07

Table 7. Installments for different values of α and different values of the age at issue; $i_{rf} = 0.00\%$.

$x = 55$ $i_{rf} = 0.01$		α		
		0.50	0.80	1.00
β	1.00	4497	4658	4753
	1.50	4418	4604	4707
	2.00	4347	4546	4658

$x = 60$ $i_{rf} = 0.01$		α		
		0.50	0.80	1.00
β	1.00	5353	5426	5465
	1.50	5315	5402	5447
	2.00	5280	5376	5426

Table 8. Installments for different values of α and different values of the age at issue; $i_{rf} = 0.01$

Germany

In Table 9 and Table 10 we report the interest rates determined by equation 2 in two different risk free interest rate scenarios.

$i_{rf} = 0.00\%$	α		
	0.50	0.80	1.00
i^*	2.00%	3.20%	4.00%

Table 9. i^* values, $i_{rf} = 0.00\%$, for different values of α and β

$i_{rf} = 0.01$		α		
		0.50	0.80	1.00
β	1.00	0.010	0.006	0.010
	1.50	-0.005	0.001	0.005
	2.00	-0.010	-0.004	0.000

Table 10. i^* values, $i_{rf} = 0.01$ for different values of α and β

Tables 11 and 12 report the installments determined in three different ages at issue, $x = 55, 60, 65$ in the two cases of risk free interest rate. For $i_{rf} = 0.00\%$ the β parameter is irrelevant.

$i_{rf} = 0.00\%$	α		
	0.50	0.80	1.00
$x = 55$	4733.08	4909.15	5015.88
$x = 60$	5677.88	5785.64	5850.64
$x = 65$	7667.23	8128.98	8448.98

Table 11. Installments for different values of α and different values of the age at issue; $i_{rf} = 0.00\%$.

$x = 55$		α		
$i_{rf} = 0.01$		0.50	0.80	1.00
β	1.00	4377	4392	4405
	1.50	-	4249	4308
	2.00	-	-	4235

$x = 60$		α		
$i_{rf} = 0.01$		0.50	0.80	1.00
β	1.00	5315	5387	5403
	1.50	-	5249	5280
	2.00	-	-	5242

Table 12. Installments for different values of α and different values of the age at issue; $i_{rf} = 0.01$

Spain

In Table 13 and Table 14 we report the interest rates determined by equation 2 in two different risk free interest rate scenarios.

$i_{rf} = 0.00\%$	α		
	0.50	0.80	1.00
i^*	1.00%	1.60%	2.00%

Table 13. i^* values, $i_{rf} = 0.00\%$, for different values of α and β

$i_{rf} = 0.01$		α		
		0.50	0.80	1.00
β	1.00	0.00	0.006	0.010
	1.50	-0.005	0.001	0.005
	2.00	-0.01	-0.004	0.000

Table 14. i^* values, $i_{rf} = 0.01$ for different values of α and β

Tables 15 and 16 report the installments determined in three different ages at issue, $x = 55, 60, 65$ in the two cases of risk free interest rate. For $i_{rf} = 0.00\%$ the β parameter is irrelevant.

$i_{rf} = 0.00\%$	α		
	0.50	0.80	1.00
$x = 55$	4691.33	4794.09	4859.68
$x = 60$	5578.16	5639.10	5677.88
$x = 65$	7298.68	7518.07	7667.23

Table 15. Installments for different values of α and different values of the age at issue; $i_{rf} = 0.00\%$.

$x = 55$		α		
$i_{rf} = 0.01$		0.50	0.80	1.00
β	1.00	4328	4421	4500
	1.50	-	4344	4428
	2.00	-	-	4351

$x = 60$		α		
$i_{rf} = 0.01$		0.50	0.80	1.00
β	1.00	5242	5287	5315
	1.50	-	5250	5280
	2.00	-	-	5242

Table 16. Installments for different values of α and different values of the age at issue; $i_{rf} = 0.01$

Greece

In Table 13 and Table 14 we report the interest rates determined by equation 2 in two different risk free interest rate scenarios.

$i_{rf} = 0.00\%$	α		
	0.50	0.80	1.00
i^*	1.00%	1.60%	2.00%

Table 13. i^* values, $i_{rf} = 0.00\%$, for different values of α and β

$i_{rf} = 0.01$		α		
		0.50	0.80	1.00
β	1.00	0.00	0.006	0.010
	1.50	-0.005	0.001	0.005
	2.00	-0.01	-0.004	0.000

Table 14. i^* values, $i_{rf} = 0.01$ for different values of α and β

Tables 15 and 16 report the installments determined in three different ages at issue, $x = 55, 60, 65$ in the two cases of risk free interest rate. For $i_{rf} = 0.00\%$ the β parameter is irrelevant.

$i_{rf} = 0.00\%$	α		
	0.50	0.80	1.00
$x = 55$	4691.33	4794.09	4859.68
$x = 60$	5578.16	5639.10	5677.88
$x = 65$	7298.68	7518.07	7667.23

Table 15. Installments for different values of α and different values of the age at issue; $i_{rf} = 0.00\%$.

$x = 55$		α		
$i_{rf} = 0.01$		0.50	0.80	1.00
β	1.00	4249	4336	4390
	1.50	-	4264	4322
	2.00	-	-	4249

$x = 60$		α		
$i_{rf} = 0.01$		0.50	0.80	1.00
β	1.00	5180	5232	5258
	1.50	-	5197	5225
	2.00	-	-	5190

Table 16. Installments for different values of α and different values of the age at issue; $i_{rf} = 0.01$

The numerical results, immediately evident, reflect the strategic meaning of the parameters α and β as explained in sect.3. In fact we can observe that the installments, increasing with the age, increase if α increases. This trend is less marked when the risk free interest rate is different from 0. The perception of high volatility in real estate markets impacts the more the i_{re} rate is affected. Increasing values for α produce a decrease in the rate to use for valuation of the installments to be paid to the policyholder: this behavior leads to a decreasing

installment. This can be understood as a form of hedging against the risk of high volatility in financial markets in the form of an implicit loading in the financial risk driver. The installments decrease when β increases, as it is intuitive because it is linked to the insurer's need to resort to the financial market to fulfill the obligations.

References

Alai D. H., Chen H., Cho D., Hanewald K, Sherris M.: Developing Equity Release Markets: Risk Analysis for Reverse Mortgage and Home Reversion. *North American Actuarial Journal*, 18:1, (2014) pp. 217-241.

Bovenberg, L., Nijman, T.: Personal pensions with risk sharing. *Journal of Pension Economics & Finance*, (2016), Published online: 08 August 2016: 1-17

Chłóń-Domińczak, A. (2016). Impact of changes in multi-pillar pension systems in CEE countries on individual pension wealth. *Journal of Pension Economics & Finance*, Published online: 05 December 2016, 1-12

de la Fuente Merencio I., Navarro E., Serna G.: Estimating the No-Negative-Equity Guarantee in Reverse Mortgages: International Sensitivity Analysis. in *New Methods in Fixed Income Modeling*. Springer International Publishing DOI: 10.1007/978-3-319-95285-7_13 (2018)

Denuit, M., Haberman, S., Renshaw, A.: Longevity-Indexed life annuities. *North American Actuarial Journal*, 15, Issue 1, 97-111 (2011)

D'Amato V., Di Lorenzo E., Haberman S., Sibillo M., Tizzano R.: Pension Schemes versus Real Estate. *Annals of Operations Research* <https://doi.org/10.1007/s10479-019-03241-y>; First Online 10 May 2019; DOI <https://doi.org/10.1007/s10479-019-03241-y>; Springer US; Print ISSN 0254-5330; Online ISSN 1572-9338 (2019)

Di Lorenzo E., Piscopo G., Sibillo M., Tizzano R.: Reverse mortgages through artificial intelligence: new opportunities for the actuaries, *Decisions in Economics and Finance* <http://link.springer.com/article/10.1007/s10203-020-00274-y>; First Online February 2020; DOI <https://doi.org/10.1007/s10203-020-00274-y>; Springer (2020a)

Di Lorenzo E., Piscopo G., Sibillo M., Tizzano R.: Risk Management strategies for Reverse Mortgage contracts: proposals for securitization. (2020b) Working paper

EIOPA: Call for advice from the European Insurance and Occupational Pension Authority (EIOPA). On the development of an EU single market for personal pension products (PPP) (2014)

https://eiopa.europa.eu/Publications/Requests%20for%20advice/Personal_pension_EIOPA_Anexx_-_CfA_EIOPA.pdf#search=reverse%20mortgage

EIOPA: Consultation paper concerning technical advice, implementing and regulatory technical standards for the Pan-European Personal Pension Product (2019) <https://eiopa.europa.eu/Publications/Consultations/Consultation%20Paper%20on%20PEPP.pdf>

EUROSTAT Statistical Books: Living conditions in Europe. 2018 edition

EUROSTAT: Duration of Working Life Statistics (2019) https://ec.europa.eu/eurostat/statistics-explained/index.php/Duration_of_working_life_statistics#Increase_in_expected_duration_of_working_life_in_the_EU

Hanewald K., Post T., Sherris M.: Portfolio Choice in Retirement: What is the Optimal Home Equity Release Product??. *Journal of Risk and Insurance*, Vol. 83, Issue 2 (2016) pp. 421-446.

Huang, H. C., Wang, C. W., Miao, Y. C.: Securitization of Crossover Risk in reverse Mortgage. *The Geneva Papers on Risk and Insurance. Issues and Practice*, Vol. 36, No. 4, SPECIAL ISSUE ON LONGEVITY (October 2011), pp. 622-647.

Li J., Kogure A., Liu J.: Multivariate Risk-Neutral Pricing of Reverse Mortgages under the Bayesian Framework. *Risks* (2019), 7, 11, pp. 1-12.

Loichinger E., Weber D.: Trends in Working Life Expectancy in Europe. *Journal of Aging and Health*, Vol 28(7), 2016.

Maurer, R., Rogalla, R. and Siegelin, I.: Participating Payout Life Annuities. Lessons from Germany. *ASTIN Bull*, 43(02), (2013), 159-187.

Merton R. C., Neng Lai R.: On an efficient design of the Reverse Mortgage: A Possible Solution for Aging Asian Population (2016), SSRN-id3075087.pdf

Nakajima M., Telywcava I.: Reverse Mortgage Loans: A quantitative Analysis??. *The Journal of Finance*, Vol. 72, Issue 2 (2017) pp. 911-950.

Shao A. W., Hanewal K, Sherris M: Reverse mortgage pricing and risk analysis allowing for idiosyncratic home price risk and longevity risk. *Insurance: mathematics and Economics*, 63 (2015), pp.76-90.

Wang L., Valdez E. A., Piggot J.: Securitization of Longevity Risk in Reverse Mortgage. *North American Actuarial Journal*, 12 (2008) pp. 345-371.

Van Binsbergen, J. H., Broeders, D., De Jong, M., Koijen, R. S. J.: Collective pension scheme and individual choice. *Journal of Pension Economics & Finance*, (2014), 13, Issue 2, 210-225

Applying Interval PCA and Clustering to Quantile Estimates: Empirical Distributions of Fertilizer Cost Estimates for Yearly Crops in European Countries ¹.

Dominique Desbois¹

¹ UMR Economie publique, INRAE-AgroParisTech, Université Paris-Saclay, 16 rue Claude Bernard, F-75231 Paris Cedex 05, France.
(E-mail: dominique.desbois@inrae.fr)

Abstract. The decision to adopt one or another of the sustainable land management alternatives should not be based solely on their respective benefits in terms of climate change mitigation but also based on the performances of the productive systems used by farm holdings, assessing their environmental impacts through the cost of fertilizer resources used. This communication uses the symbolic clustering tools in order to analyse the conditional quantile estimates of the fertilizer costs of yearly crop productions in agriculture, as a replacement proxy for internal soil erosion costs. After recalling the conceptual framework of the estimation of agricultural production costs, we present the empirical data model, the quantile regression approach and the interval principal component analysis clustering tools used to obtain typologies of European countries on the basis of the conditional quantile distributions of fertilizer cost empirical estimates. The comparative analysis of econometric results for main products between European countries illustrates the relevance of the typologies obtained for international comparisons to assess land management alternatives based on their impact on agricultural carbon sequestration in soils.

Keywords: principal component analysis, hierarchic clustering, interval estimates, quantile regression, input-output model, symbolic data analysis, agricultural production cost, fertilizer, yearly crops, micro-economics.

“Applied economists increasingly want to know what is happening to an entire distribution, to the relative winners and losers, as well as to averages.”
Angrist et Pischke [1]

1 Economics of agricultural carbon sequestration in soils

Signatory States to the 2015 Paris Agreement have set a common goal of achieving carbon neutrality. According to a logic of net emissions flow adopted by several European countries, France has adopted a Climate Plan in July 2017 with a target of zero net emissions (ZEN) of greenhouse gases, at the 2050 horizon (Quinet [29]).

Carbon sequestration in soils is one of the means proposed to achieve common goals of reducing greenhouse gas emissions while improving the productivity and sustainability of agricultural land in both developed and developing countries (SM CRSP, 2008). In addition to their soil carbon storage capacity, sustainable

⁶th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



land management technologies benefit farmers by increasing yields and reducing production costs.

For the European Union, a group of experts from the European Commission on agricultural markets also proposes to encourage farmers to store carbon on the basis of adapted agricultural practices (EC, 2016). However, on one hand, the evolution of the CAP's regulatory frameworks by 2020 shows that the proposed instruments alone cannot support large-scale projects on the agricultural soil carbon storage in Europe: in fact, there is very little likely that the future CAP budget is sufficient (Jevnaker and Wettestad, [19]). Hence, the decision to adopt one or another of the sustainable land management alternatives should not be based solely on their respective benefits in terms of climate change mitigation but also based on the consideration of the farmers, assessing comprehensively the productivity, resource utilization and environmental impact of the productive system.

In the framework of the ANR ASSES project, integrated into the OTE-Med Eranet, we propose to better assess the economic cost of erosion for farmers by estimating the costs of restoring soil fertility, conceived as an ecosystem service for the benefit of agriculture. The economic evaluation of erosion distinguishes between two types of costs: on-site and off-site costs: in this paper, we focus on the on-site costs and in particular the costs induced by the resulting loss of nutrients. A review of the literature shows that estimates of the cost of soil erosion due to nutrient loss are significant and vary greatly depending on the type of crops grown and the production regions. In order to evaluate erosion costs due to nutrient losses, we estimate the production costs of fertilizers using an input/output methodology.

The integration of agriculture in the 27 Member States of the European Union (EU) have raised both in the context of competitive markets as markets subject to regulation, recurring needs for estimating costs of production for major agricultural products, all along the successive reforms of the Common Agricultural Policy (CAP). The analysis of agricultural production costs is a tool for analyzing economic results of farmers: it allows to assess the price competitiveness of farmers, one of the major elements for development and sustainability of food chains in the European regions. To meet the needs of simulations and impact assessment in the various common market organizations, we must be able to provide information on the entire distribution of production costs for the assessment of public agricultural policy options. Based on the observation of asymmetry and heterogeneity within the empirical distribution of agricultural inputs, we propose a methodology adapted to the problem of estimating the empirical distributions of fertilizer costs of production for the main agricultural products in a European context where agricultural holdings remain mainly oriented towards multiple productions (Desbois *et al.* [11]).

We first present the empirical model for estimating the fertilizer costs of production, derived from an econometric cost allocation approach inspired by Divay and Meunier [14] using microeconomic data to build an input-output matrix. Then, we introduce the estimation methodology according to the conditional quantiles proposed by Koenker and Bassett [21]. Next, we present the

symbolic data analysis procedures used to explore the empirical estimates of conditional quantile distribution intervals based on the concepts and methods provided by the symbolic approach of Billard and Diday [2]. Then, we present the graphs of results from the analysis tools of the symbolic data applied to the estimation intervals of the conditional quantiles. Eventually, we conclude on the relevance of this approach applied to the production of pig, proposing an extension of this type of analysis at the regional level.

2 Conceptual framework and methodological aspects

First, we present the methodology for estimating input costs, among which the fertilizer costs. Secondly, we introduce the factorial analysis and the clustering procedure of the estimation intervals in the formalism of the symbolic data.

2.1 The empirical model for estimating the fertilizer costs of production

Inspired by Divay and Meunier [14], the allocation of the sum x_i of the input costs for farm holding i is made by linear decomposition along the gross products Y_i^j of farm holding i for each production j , where u_i is a random vector with a zero mathematical expectation:

$$(1) \quad x_i = \sum_{j=1}^p \beta_j Y_i^j + u_i$$

As Cameron and Trivedi [4], we assume that the data generator process is a linear model with multiplicative heteroscedasticity characterized in matrix form by:

$$(2) \quad x = Y'\beta + u \quad \text{with} \quad u = Y'\alpha \times \varepsilon \quad \text{and} \quad Y'\alpha > 0$$

where $\varepsilon \sim iid[0, \sigma]$ is a random-vector identically and independently distributed with zero mean and constant variance σ^2 . Under this assumption, $\mu_q(x|Y, \beta, \alpha)$, the q^{th} conditional quantile of the production cost x , conditioned by Y and the α and β parameters, is derived analytically as follows:

$$(3) \quad \mu_q(x|Y, \beta, \alpha) = Y'[\beta + \alpha \times F_\varepsilon^{-1}(q)] = Y'\gamma.$$

where F_ε is the cumulative distribution function (CDF) of the errors.

The technical coefficient for the j^{th} product of the q^{th} quantile of the fertilizer cost is defined by the j^{th} component of the multivariate slope vector:

$$(4) \quad \gamma^j(q) = [\beta + \alpha \times F_\varepsilon^{-1}(q)]^j.$$

Following D'Haultfoeuille and Givord [15], three models can be derived:

- i) $x = Y'\beta + u$ with $u = K\varepsilon$, homoscedastic errors $V(\varepsilon|Y) = \sigma^2$, denoted as the *location-shift* model, *i.e.* the linear model of conditional quantile with homogeneous slopes; while $Y'\alpha = K$ is constant, the conditional quantiles $\mu_q(x|Y, \beta, \alpha) = Y'\beta + KF_\varepsilon^{-1}(q)$ get all the same β slope, but differ only by a constant gap, growing as q , the quantile order, increases;
- ii) $x = Y'\beta + (Y'\alpha)\varepsilon$ and $Y'\alpha > 0$ with heteroscedastic residuals, referred as the *location-scale shift* model, *i.e.* the linear model of heterogeneous conditional quantile slopes.

- iii) $X = Y'\gamma_\xi$ with ξ random variable independent of Y following a uniform distribution over the interval $[0,1]$ such as $\xi \rightarrow Y'\gamma_\xi$ be strictly increasing whatever Y , designated as the random coefficient model. ξ corresponds to a random component determining the rank of the individual within the distribution of X . Under the strong distributional hypothesis of rank invariance, the random coefficient γ_q represent the effect of a marginal change in Y for agricultural holdings located at the q^{th} quantile of the ξ distribution. This distributional assumption of rank invariance means that median farms in terms of input productivity would maintain the $q = 0.5$ rank, regardless of the different levels of production Y_i registered for the i^{th} farm holding.

2.2 The procedures for estimating and testing conditional quantiles

The quantile regression is defined for each quantile of order q as the solution of a problem of minimization of the sum of the deviations in absolute value (L_1 norm):

$$(5) \quad \hat{\beta}(q) = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i \in \{i / x_i \geq y'_i \beta\}} q |x_i - y'_i \beta| + \sum_{i \in \{i / x_i \leq y'_i \beta\}} (1-q) |x_i - y'_i \beta| \right\}$$

can be written in matrix form (6):

$$\hat{\beta}(q) = \arg \min_{\beta \in \mathbb{R}^p} \left\{ q e'(X - Y'\beta \geq 0) \delta'[X - Y'\beta] + (1-q) e'(Y'\beta - X \geq 0) \delta'[Y'\beta - X] \right\}$$

with $e'(X - Y'\beta \geq 0)$, index of farms i such as $x_i - y'_i \beta \geq 0$, and δ^1 , vector of absolute deviations.

Thus, the linear optimization problem solving methods developed for the L_1 (absolute deviation) regression easily extend to quantile regression (Koenker and d'Orey, [22]). Although the simplex method (Dantzig [9]) has an algorithmic complexity in $O(n^6)$, the Karmarkar [20]'s method of the "interior-point" is in practice preferable as soon as the sample size becoming large, because of its reduced algorithmic complexity to $O(n^{3.5})$. For large samples, Portnoy and Koenker [28] have shown that a combination of the "interior-point" algorithm and a smoothing algorithm for the objective function by Madsen and Nielsen [26] makes quantile regression calculations competitive with those of least squares regression.

The weighted conditional quantiles have been proposed by Koenker and Zhao [23] as L-estimates¹ in linear heteroscedastic models. The $W = \{w_i, i = 1, \dots, n\}$ weighting of the observations leads to a quantile regression scheme solving the following minimization problem (7):

¹ An L-estimate is an estimate defined by a linear combination of ordinal statistics.

$$\hat{\beta}_\omega(q) = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i \in \{i / x_i \geq y'_i \beta\}} w_i q |x_i - y'_i \beta| + \sum_{i \in \{i / x_i \leq y'_i \beta\}} w_i (1-q) |x_i - y'_i \beta| \right\}$$

The weighted estimation procedure uses the "predictor-corrector" implementation of the primal-dual algorithm proposed by Lustig *et al.* [25]

Given the size of the Farm Accounting Data Network (FADN) sample, its non-random selection and the existence a priori of distinct sub-populations (e.g. specialized types of farming), we opted for the resampling method, based on the Markov Chain Marginal Bootstrap (MCMB) technique. Without distributional assumption, this method yields robust empirical confidence intervals in a reasonable computation time (He and Hu [18]).

For a given product j_0 such as yield crops and the l^{th} European country, the estimation interval of technical coefficients for q^{th} conditional quantile of the fertilizer costs

$$(8) \quad z_l^q = [Inf_{-}\hat{y}_l^{j_0}(q); Sup_{-}\hat{y}_l^{j_0}(q)] = \left[\underline{z}_l^q; \overline{z}_l^q \right]$$

is obtained by MCMB.

2.3 Symbolic PCA of the fertilizer cost distributions

The symbolic approach has been introduced by Diday [13] in order to take in account several values rather a single one attached to a variable into the framework of exploratory methods of data analysis. Within this conceptual framework of symbolic data analysis, the extension of principal component analysis (PCA) to interval data was initially proposed by Cazes *et al.* [5] and later improved by Chouakria *et al.* [7] with the Vertex and the Center methods using either the vertices or the center of the hyper-rectangle defined by interval values as a multidimensional support for the initial PCA. In this paper, we propose to assess different PCA variants around the Vertex or the Center Methods, proposed by Garro and Rodriguez [24] in order to maximize the variance of the projections or to minimize the distance between the vertices and the projections of the hyper-rectangle, on the basis of distributional data.

As symbolic objects, the L national distributions $\Omega = \{\omega_1, \dots, \omega_l, \dots, \omega_L\}$ are described by a set of $Q = 5$ descriptors², which are the estimation intervals of $\{z^{0.10}, z^{0.25}, z^{0.50}, z^{0.75}, z^{0.90}\}$, coding for the D1 and D9 deciles combined with the three quartiles Q1, Q2 and Q3.

Let define the set of $L \times Q$ "within interval"-value matrices,

$$\mathcal{M} = \left\{ Z \in M_{L \times Q} \mid z_l^q \in \left[\underline{z}_l^q; \overline{z}_l^q \right] \right\}.$$

² This choice of a small number of descriptors was made for comparative convenience with some more classical graphic approaches (Desbois *et al.* [11]); however, like this earlier work, it could be extended without disadvantage to sets of descriptors of cardinality $Q=9$ (deciles), or even $Q=99$ (percentiles) if the analysis objectives required it.

2.3.1 The center-PCA of the interval distribution for quantile estimates

Let us define $U \in \mathcal{M}$, the center-interval matrix of Z , by:

$$U = [U^1, \dots, U^q, \dots, U^Q] = \begin{bmatrix} u_1^1 & \dots & u_1^Q \\ \vdots & u_l^q & \vdots \\ u_L^1 & \dots & u_L^Q \end{bmatrix} \text{ with } u_l^q = \frac{\overline{z_l^q} + \underline{z_l^q}}{2};$$

$$V = \begin{bmatrix} v_1^1 & \dots & v_1^Q \\ \vdots & v_l^q & \vdots \\ v_L^1 & \dots & v_L^Q \end{bmatrix} \text{ with } v_l^q = \left[\frac{z_l^q - \hat{\mu}^q}{\sqrt{L}\hat{\sigma}^q}; \frac{\overline{z_l^q} - \hat{\mu}^q}{\sqrt{L}\hat{\sigma}^q} \right]$$

where $\hat{\mu}^q$ and $\hat{\sigma}^q$ are respectively the mean and the standard deviation of the q^{th} column vector U^q of the matrix U .

According to Cazes *et al.* [5], the interval principal components are defined by the following equations:

$$(9) \quad \underline{\varphi}_l^q = \sum_{k=1,K; \zeta_k^q < 0} (\overline{u}_l^k - \hat{\mu}^k) \zeta_k^q + \sum_{k=1,K; \zeta_k^q \geq 0} (\underline{u}_l^k - \hat{\mu}^k) \zeta_k^q$$

$$(10) \quad \overline{\varphi}_l^q = \sum_{k=1,K; \zeta_k^q < 0} (\underline{u}_l^k - \hat{\mu}^k) \zeta_k^q + \sum_{k=1,K; \zeta_k^q \geq 0} (\overline{u}_l^k - \hat{\mu}^k) \zeta_k^q$$

where ζ_k^q is the q^{th} coordinate of the k^{th} eigenvector of $U'U$, the variance-covariance matrix of U .

According to Rodriguez, Diday and Winsberg [31], the pattern of duality in the center-PCA implies the following relationships:

$$(11) \quad \underline{\varphi}_h^q = \max \left[\sum_{k=1, \dots, Q; \zeta_k^q < 0} \overline{v}_h^k \zeta_k^q + \sum_{k=1, K; \zeta_k^q \geq 0} \underline{v}_h^k \zeta_k^q; -1 \right]$$

$$(12) \quad \overline{\varphi}_h^q = \min \left[\sum_{k=1, \dots, Q; \zeta_k^q < 0} \underline{v}_h^k \zeta_k^q + \sum_{k=1, K; \zeta_k^q \geq 0} \overline{v}_h^k \zeta_k^q; 1 \right]$$

where ζ_k^q is the q^{th} coordinate of the h^{th} eigenvector of VV' the inertia matrix of

V , and $\overline{v}_h^k = \sup_{l_h \in L} \{v_{l_h}^k\}$ respectively $\underline{v}_h^k = \inf_{l_h \in L} \{v_{l_h}^k\}$. This duality pattern

determines the infimum and the supremum of the hyper-rectangle defined by the projection of the q^{th} vector of V in the direction of the h^{th} principal component of VV' .

2.3.2 The 'best point' PCA of the interval distribution for quantile estimates

In the bivariate case ($q = 2$) with the Q1 ($Z^{0.25}$) and Q3 ($Z^{0.75}$) quartiles, the vertex submatrix \bar{Z}_l associated with the l^{th} country, is defining the $n = 2^q = 4$ vertices of a Q1 by Q3 rectangle \mathcal{H}_l (cf. figure 1):

$$(13) \quad \bar{Z}_l = \begin{matrix} \bar{Z}^{0.25} & \bar{Z}^{0.75} \\ \left. \begin{matrix} z_l^{0.25} & z_l^{0.75} \\ z_l^{0.25} & z_l^{0.75} \\ \bar{z}_l^{0.25} & \bar{z}_l^{0.75} \\ z_l^{0.25} & z_l^{0.75} \end{matrix} \right\} \end{matrix}.$$

Via a similar process for $l = 1, \dots, L$, let us define $\bar{Z} = (\bar{Z}_1, \dots, \bar{Z}_l, \dots, \bar{Z}_L)'$, the vertex-interval matrix, by its submatrices \bar{Z}_l of the l^{th} country ω_l , represented by \mathcal{H}_l the hyper-rectangle build with $n_l = 2^{q_l}$ vertices of the q_l non-trivial intervals.

$$\bar{Z}_l = \begin{bmatrix} \bar{z}_{s_1}^1 & \dots & \bar{z}_{s_1}^q & \dots & \bar{z}_{s_1}^{q'} & \dots & \bar{z}_{s_1}^Q \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \bar{z}_{s_h}^1 & \dots & \bar{z}_{s_h}^q & \dots & \bar{z}_{s_h}^{q'} & \dots & \bar{z}_{s_h}^Q \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \bar{z}_{s_{n_l}}^1 & \dots & \bar{z}_{s_{n_l}}^q & \dots & \bar{z}_{s_{n_l}}^{q'} & \dots & \bar{z}_{s_{n_l}}^Q \end{bmatrix}$$

In this way, the vertices of hyper-rectangles \mathcal{H}_l are vectors of \mathbb{R}^Q , while the Q estimates of the conditional quantiles are elements of \mathbb{R}^N , with $N = \sum_{l=1}^L n_l$.

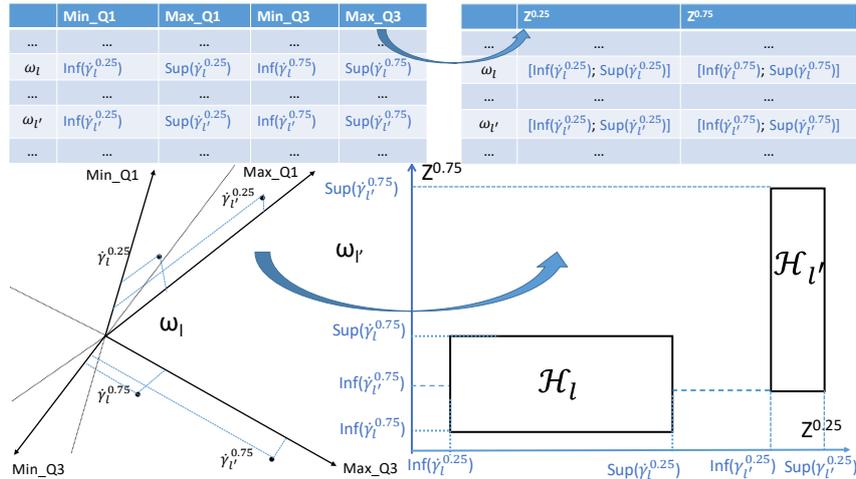


Fig. 1. The symbolic coding of the estimation intervals for the technical coefficients of the lower (Q1) and higher (Q3) quartiles of fertilizer costs

Let us apply PCA to $Z \in \mathcal{M}$, a within-interval value matrix. The k^{th} principal component of the l^{th} country is given by:

$$(14) \quad \psi_l^k = \sum_{q=1}^Q (z_l^q - \mu_q) w_q^k$$

where $\mu_q = \frac{1}{L} \sum_{l=1}^L z_l^q$ is the average of the q^{th} conditional quantile of cost estimates and w_q^k , the q^{th} coordinate of the k^{th} eigenvector of the variance-covariance matrix of Z .

Defining the supplementary normalised vertex $\tilde{Z} = (\tilde{Z}_1, \dots, \tilde{Z}_l, \dots, \tilde{Z}_L)'$ by its l^{th} submatrix, where σ_q is the standard deviation of Z^q

$$\tilde{Z}_l = \begin{bmatrix} \frac{\overline{z_l^1} - \mu_1}{\sqrt{L}\sigma_1} & \dots & \frac{\overline{z_l^q} - \mu_q}{\sqrt{L}\sigma_q} & \dots & \frac{\overline{z_l^Q} - \mu_Q}{\sqrt{L}\sigma_Q} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{z_l^1 - \mu_1}{\sqrt{L}\sigma_1} & \dots & \frac{z_l^q - \mu_q}{\sqrt{L}\sigma_q} & \dots & \frac{z_l^Q - \mu_Q}{\sqrt{L}\sigma_Q} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{z_l^1 - \mu_1}{\sqrt{L}\sigma_1} & \dots & \frac{z_l^q - \mu_q}{\sqrt{L}\sigma_q} & \dots & \frac{z_l^Q - \mu_Q}{\sqrt{L}\sigma_Q} \end{bmatrix}$$

Each s_h vertex of hyper-rectangle of the l^{th} national distribution of fertilizer cost estimate \tilde{Z}_l can be projected on the principal components of the Z-PCA, with the following k^{th} coordinates:

$$(15) \quad c_{s_h}^k = \sum_{q=1}^Q \tilde{z}_{s_h}^q w_q^k.$$

According to Rodriguez [30], the minimum and maximum of the k^{th} coordinate for each estimation interval for the l^{th} country can be computed as follows:

$$(16) \quad \underline{\psi}_l^k = \underset{s_h = 1, \dots, n_l}{\text{Inf}}\{c_{s_h}^k\} = \sum_{\{q|w_q^k < 0\}} (\bar{z}_l^q - \mu_q) w_q^k + \sum_{\{q|w_q^k \geq 0\}} (\underline{z}_l^q - \mu_q) w_q^k$$

$$(17) \quad \bar{\psi}_l^k = \underset{s_h = 1, \dots, n_l}{\text{Sup}}\{c_{s_h}^k\} = \sum_{\{q|w_q^k < 0\}} (\underline{z}_l^q - \mu_q) w_q^k + \sum_{\{q|w_q^k \geq 0\}} (\bar{z}_l^q - \mu_q) w_q^k$$

Let us denote t_h the eigenvectors of $\tilde{Z}\tilde{Z}'$ for $h = 1, \dots, H$, the coordinate of the q^{th} quantile estimates on the h^{th} principal component is given by:

$$(18) \quad r_h^q = \sum_{s=1}^N \tilde{Z}'_q^s t_s^h$$

According to Garro and Rodriguez [24], by projection of the q^{th} quantile estimate on the h^{th} principal component in the direction of t_h , the infimum and supremum values of the hyper-rectangle \mathcal{H}_l are computed as follows:

$$(19) \quad \underline{\chi}_l^q = \underset{s_l = 1, \dots, n_l}{\text{Inf}}\{r_{s_l}^q\} = \sum_{\{s|t_s^h < 0\}} \bar{z}_l^s t_s^h + \sum_{\{s|t_s^h \geq 0\}} \underline{z}_l^s t_s^h$$

$$(20) \quad \bar{\chi}_l^q = \underset{s_l = 1, \dots, n_l}{\text{Sup}}\{r_{s_l}^q\} = \sum_{\{s|t_s^h < 0\}} \underline{z}_l^s t_s^h + \sum_{\{s|t_s^h \geq 0\}} \bar{z}_l^s t_s^h$$

Thus, the Z-PCA provides a dual representation of the fertilizer empirical cost distributions represented by their estimation intervals, which are the symbolic objects, and conditional quantiles which are the descriptors of these symbolic objects.

Let us define $\mathcal{U}(Z) = \{w_1^Z, \dots, w_s^Z, \dots, w_S^Z\}$, the orthonormal basis of eigenvectors issued from the variance-covariance matrix of Z , and the function

$\Psi(Z): \mathcal{M} \rightarrow \mathbb{R}^+ \cup \{0\}$ based on the Euclidean norm $\|\cdot\|$,

such as $\Psi(Z) = \sum_{l=1}^L \|\tilde{Z}_l - Pr_{\mathcal{U}(Z)}(\tilde{Z}_l)\|^2$

and where $Pr_{\mathcal{U}(Z)}(\tilde{Z}_l)$ is the projection of the sub-matrix \tilde{Z}_l , coding the vertices of the hyper-rectangle \mathcal{H}_l , on $\mathcal{U}(Z)$, as an appropriate orthonormal basis.

The interval-valued matrix Z^* that solves the optimization problem

$$(21) \quad \underset{Z \in \mathcal{M}}{\text{Min}} \Psi(Z)$$

is estimated through Procedure (below), using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm, in order to find the minimal distance to \tilde{Z} , the vertex matrix.

Procedure 1 Minimizing the squared distance

Input:

$Z \in \mathcal{M}$, a $L \times Q$ matrix with s principal components;

TOL , a numerical threshold of tolerance;

$ITER$, a maximum number of iterations.

i) $Z \leftarrow U$, the center matrix as the initial value;

ii) $Z^* \leftarrow lbfgs(Z, \text{objective} = \Psi(Z), TOL, ITER)$;

iii) Compute the $[\underline{\psi}_L^*; \overline{\psi}_L^*]$ coordinates,
applying (16) & (17) duality relationships;

iv) **Return** $[\underline{\psi}_L^*; \overline{\psi}_L^*]$

Source: adapted from Garro and Rodriguez [24];

Nota bene: *lbfgs* is a function implementing BFGS algorithm, from the *nloptr* package by Ypma *et al.* [32].

Let us define the function $\Lambda(Z, s): \mathcal{M} \times N \rightarrow \mathbb{R}^+$ such as $\Lambda(Z, s) = \sum_{h=1}^s \lambda_h$, the variance of the first s components issued from the PCA of Z , where λ_h is the h^{th} eigenvalue associated to the h^{th} eigenvector of $U(Z)$.

The interval-valued matrix Z^s that solves the optimization problem

$$(22) \quad \underset{Z \in \mathcal{M}}{\text{Max}} \Lambda(Z, s)$$

is estimated through Procedure 2 (below) using the BFGS algorithm, in order to maximize the variance of the first s components.

Procedure 2 Maximizing the variance of the first components

Input: $Z \in \mathcal{M}$, a $L \times Q$ matrix, with s principal components;

TOL , a numerical threshold of tolerance;

$ITER$, a maximum number of iterations.

i) $Z \leftarrow U$, the center matrix as an initial value;

ii) $Z^s \leftarrow lbfgs(Z, \text{objective} = \Lambda(Z, s), TOL, ITER)$;

iii) Compute the $[\underline{\chi}_L^*; \overline{\chi}_L^*]$ coordinates,
applying (19) & (20) duality relationships;

iv) **Return** $[\underline{\chi}_L^*; \overline{\chi}_L^*]$.

Source: adapted from Arce and Rodriguez.

Nota bene: *lbfgs* is a function implementing BFGS algorithm, from the *nloptr* package by Ypma and *al.*[31].

2.4 Symbolic clustering analysis of the fertilizer cost distributions

The local dissimilarities between country l and country l' , associated with these estimation intervals of technical coefficients for the q^{th} conditional quantile, are computed according to the Euclidean distance metric:

$$(23) \delta_M(z_l^q, z_{l'}^q) = \sqrt{\left(\text{Inf}_{\hat{\gamma}_l^{j_0}}(q) - \text{Inf}_{\hat{\gamma}_{l'}^{j_0}}(q)\right)^2 + \left(\text{Sup}_{\hat{\gamma}_l^{j_0}}(q) - \text{Sup}_{\hat{\gamma}_{l'}^{j_0}}(q)\right)^2}.$$

For this metric M , a global dissimilarity between country l and country l' based on the differences over the national distributions of estimation intervals for the technical coefficients is computed according to the following quadratic criterion:

$$(24) d(\omega_l, \omega_{l'}) = \left(\sum_{q=1}^Q \delta_M^2(z_l^q, z_{l'}^q)\right)^{1/2}.$$

Given a matrix of dissimilarities between national empirical distributions of fertilizer costs issued from the previous computations, we can use the methods of unsupervised clustering. In a way similar to the Ward's method, Chavent *et al.* [6] proposes a divisive hierarchical clustering algorithm on symbolic data (DIVCLUS-T), valid for both interval data and categorical data. Subsequently, we detail for interval data the principles on which the operations of this unsupervised clustering procedure are based.

The divisive hierarchical clustering algorithm recursively splits each cluster into two sub-clusters, starting from the whole set of countries as symbolic objects

$$\Omega = \{\omega_1, \dots, \omega_l, \dots, \omega_L\}.$$

At each partition in k symbolic clusters $P_K = \{C_1, \dots, C_k, \dots, C_K\}$, a cluster has to be divided in order to get a partition P_{K+1} , with $K + 1$ clusters, optimizing the selected adequacy criterion based on the inertia.

The inertia of the k^{th} cluster is defined by $I(C_k) = \sum_{l \in C_k} \mu_l d_M^2(z_l, g(C_k))$ where μ_l is the weight of the l^{th} country and $g(C_l)$ is the cluster centroid defined as:

$$(25) g(C_k) = \frac{1}{\sum_{l \in C_k} \mu_l} \sum_{l \in C_k} \mu_l z_l.$$

The intra inertia is defined by the sum of the inertias of the clusters to their centroids:

$$(26) W(P_K) = \sum_{k=1, \dots, K} I(C_k).$$

The inter inertia is defined by the inertia of the centroids with regards to the g overall centroid of Ω , as follows:

$$(27) B(P_K) = \sum_{k=1, \dots, K} \mu_k d_M^2(g(C_k), g) \text{ where } \mu_k = \sum_{l=1, \dots, k} \mu_l.$$

For a partition P_K , the total inertia sums the intra inertia with the inter inertia:

$$(28) I(\Omega) = W(P_K) + B(P_K).$$

Hence, minimizing the heterogeneity (measured by W) is equivalent to maximizing the homogeneity (measured by B).

Generated by the logical binary choice (*yes/no*) to a numerical binary question $\Psi = [Is z^q \leq c?]$, let us denote $\{A_k, \bar{A}_k\}$ the induced bipartition of a cluster C_k formed of n_k objects. In order to choose among the $n_k - 1$ possible bipartitions of the C_k cluster, a discriminating criterion can be defined by the following ratio:

$$(29) D(\Psi) = \frac{B^q(A_k, \bar{A}_k)}{I^q(C_k)} = 1 - \frac{W^j(A_k, \bar{A}_k)}{I^q(C_k)},$$

where the inter inertia $B^q(A_k, \bar{A}_k)$ and the inertia $I^q(C_k)$ are computed with regards to the q^{th} conditional quantile. Hence, minimizing the intra inertia $W\{A_k, \bar{A}_k\}$ is equivalent to maximizing the inter inertia $B\{A_k, \bar{A}_k\}$ and, as a result, to the $D(\Psi)$ discriminating criterion.

As in Ward method, the “upper hierarchy” (Mirkin [27]) of partition P_K is indexed by the height h of a cluster C_K , defined by its inter inertia as follows:

$$(30) \quad h(C_k) = B(A_k, \bar{A}_k) = \frac{\mu(A_k)\mu(\bar{A}_k)}{\mu(A_k)+\mu(\bar{A}_k)} d^2(g(A_k), g(\bar{A}_k))$$

The DIVCLUS-T algorithm splits the cluster C_K^* that maximises $h(C_K)$, ensuring that the next partition $P_{K+1} = P_K \cup \{A_K, \bar{A}_K\} - C_K^*$ has the minimum intra inertia value, with respect to the rule

$$(31) \quad W(P_{K+1}) = W(P_K) - h(C_K^*).$$

In order to determine an optimal clustering, we use as the internal quality index for each partition P_K , the log of the determinant ratio computed as follows:

$$(32) \quad \varkappa_K = N \log \left(\frac{\det(T)}{\det(WG^{(K)})} \right)$$

where $T = Z'Z$ is the total scatter matrix (N times the total variance-covariance matrix) and $WG^{(K)} = \sum_{k=1}^K W^{(k)}$ the sum of the within-group scatter matrices, $W^{(k)}$ for each group C_k of the partition P_K in K groups.

The optimal score for the quality index is given by the *min_diff* decision rule:

$$K^* = \arg_min_K \{\partial_K - \partial_{K-1}\}$$

with $\partial_K = \varkappa_{K+1} - \varkappa_K$, using procedure *ClusterCrit* proposed by Desgraupes [12] for needed computations.

3 Results

Based on the gross product, the estimation according to the quantiles provides a conditional allocation of the fertilizer costs by main products, within the framework of a multi-product exploitation. In the framework of the Farm Accountancy Cost Estimation and Policy Analysis project (FACEPA) research project, the managers in charge of the Knowledge Based Bio-Economy project of the 7th EU Framework Program of Research has chosen to focus on the main agricultural commodities produced at a level sufficiently broad at the European level to allow meaningful cross-country comparisons for the twelve European Member States which are the main producers (EU12), choosing 2006 as a baseline for comparison convenience.

We analyse the results obtained in particular for the yield crops about fertilizer inputs. The figures are estimated from a quantile regression of the fertilizer inputs on a decomposition of the gross product into five product aggregates (yearly crops, permanent crops, pasture livestock, off-ground livestock, others) for the set of twelve European countries (UE12) selected on 2006.

Table 1 presents for yield crops the estimation intervals of conditional quantiles (lower decile D1, lower quartile Q1, median Q2, upper quartile Q3, upper decile D9) of the fertilizer inputs of agricultural production.

Country	D1	Q1	Q2	Q3	D9
Austria	[0.000 ; 0.029]	[0.043 ; 0.057]	[0.068 ; 0.086]	[0.106 ; 0.127]	[0.155 ; 0.179]
Belgium	[0.009 ; 0.019]	[0.023 ; 0.030]	[0.038 ; 0.047]	[0.056 ; 0.080]	[0.082 ; 0.110]
Denmark	[0.018 ; 0.024]	[0.035 ; 0.035]	[0.056 ; 0.056]	[0.094 ; 0.094]	[0.140 ; 0.140]
France	[0.023 ; 0.028]	[0.053 ; 0.065]	[0.125 ; 0.125]	[0.182 ; 0.182]	[0.232 ; 0.232]
Germany	[0.004 ; 0.009]	[0.025 ; 0.033]	[0.082 ; 0.082]	[0.140 ; 0.140]	[0.181 ; 0.181]
Hungary	[0.020 ; 0.038]	[0.056 ; 0.071]	[0.093 ; 0.110]	[0.138 ; 0.164]	[0.197 ; 0.197]
Italy	[0.007 ; 0.011]	[0.019 ; 0.022]	[0.041 ; 0.041]	[0.078 ; 0.078]	[0.121 ; 0.121]
Netherlands	[0.001 ; 0.004]	[0.004 ; 0.006]	[0.009 ; 0.012]	[0.017 ; 0.022]	[0.026 ; 0.029]
Poland	[0.024 ; 0.032]	[0.052 ; 0.059]	[0.088 ; 0.099]	[0.146 ; 0.165]	[0.215 ; 0.228]
Spain	[0.013 ; 0.017]	[0.025 ; 0.033]	[0.058 ; 0.058]	[0.103 ; 0.103]	[0.169 ; 0.169]
Sweden	[-0.007 ; 0.016]	[0.003 ; 0.038]	[0.100 ; 0.100]	[0.215 ; 0.215]	[0.293 ; 0.293]
United-Kingdom	[0.006 ; 0.029]	[0.036 ; 0.047]	[0.088 ; 0.088]	[0.137 ; 0.137]	[0.171 ; 0.171]

Tab. 1. Yield Crops, estimation intervals for technical coefficients of quantile fertilizer costs for € 1 of gross product, EU12. Source: author's processing, from EU-FADN 2006.

The pre-visualization of the fertilizer cost estimates is done according to the graph in Fig. 2, showing the conditional quantile point estimates in ascending order for each country. This graph of point estimates of conditional quantiles of fertilizer costs for yield crop by country highlights some distributional facts. Below 3%, the overall level of the Netherlands distribution curbs (*iNED* and *sNED* on Figure 1) is the lowest of the twelve European countries studied, with the exception of the lower bound (*i*) of the first decile (*D1*) in Sweden (*SVE*) which is negative. The Netherlands distribution is also the flattest of the twelve European distributions analysed, followed by the distributions for Italy and Belgium, which have fairly moderate slopes and overall estimation levels below 13%. The Netherlands distribution illustrates the *location shift* linear model of conditional quantile with homogeneous slopes.

Conversely, the maximum and minimum curves of the Swedish distribution (*iSVE* and *sSVE*) are the steepest (from 1,6% to near 30%), immediately followed by those of France (*iFRA* and *sFRA*) and Poland (*iPOL* and *sPOL*). These three countries illustrate the *location-scale shift* linear model of conditional quantile with heterogeneous slopes.

Next, Hungary (*iHUN* and *sHUN*), Germany (*iDEU* and *sDEU*), Austria (*iOST* and *sOST*), the United Kingdom (*iUKI* and *sUKI*) and Spain (*iESP* and *sESP*) form an intermediate group where, on the basis of this first graph, it becomes difficult to distinguish clear differences between these national distributions.

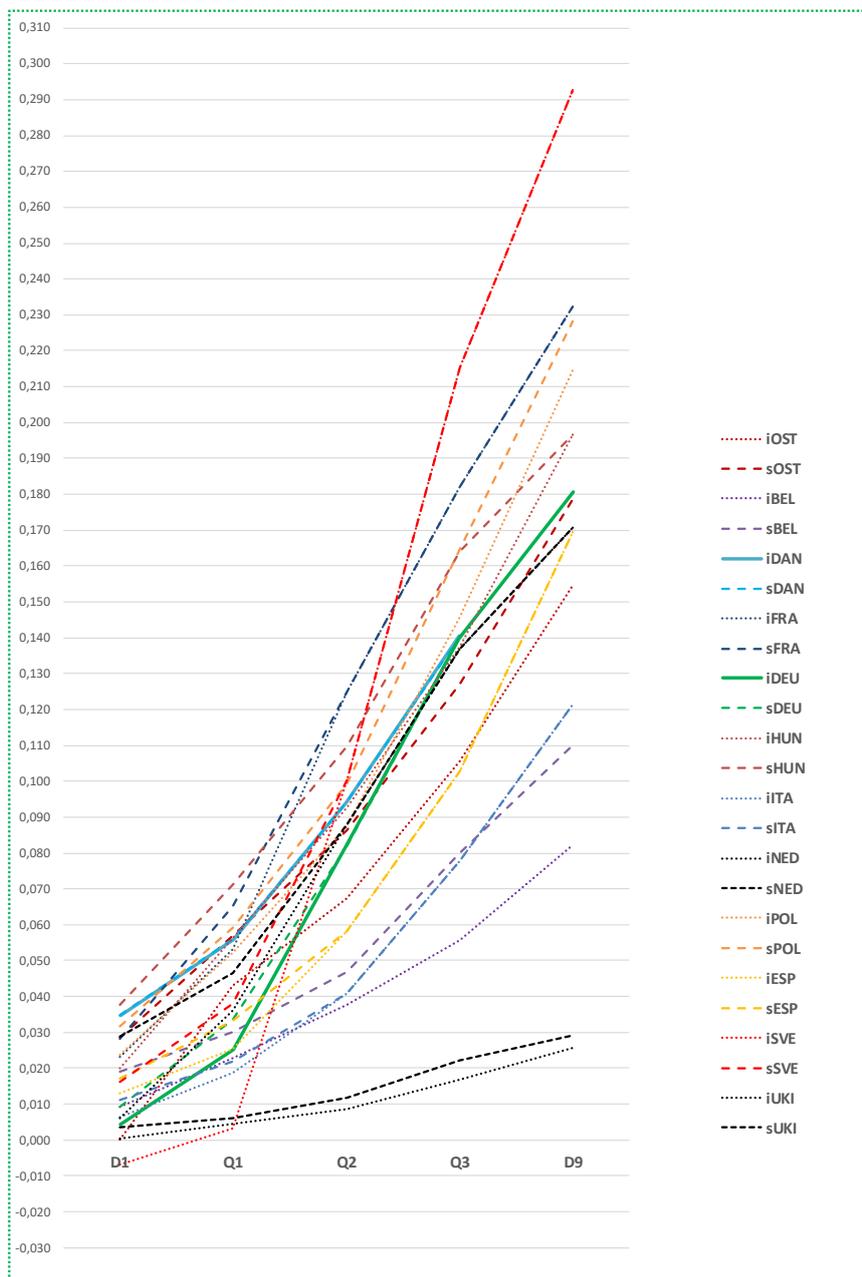


Fig. 2. Yield Crops, interval estimation for fertilizer coefficients of conditional quantiles for 12 EU member States; *iOST* stands for Austria infimum, respectively, *sOST* for Austria supremum. Source: author's processing, from EU-FADN 2006.

3.1 The interval PCAs of fertilizer cost estimates

Applying equations (9) and (10), the “centers” option of the interval PCA shows a correlation circle displaying the estimate quantile coordinates on the first two principal components with the highest negatives correlation for D1, Q1 and Q2 quantiles. The larger fans which indicate the greater infimum-supremum intervals of estimation are found for D1 and Q1 quantiles meanwhile Q2, Q3 and D9 quantiles displays the smallest which indicate the lower interval ranges of estimates.

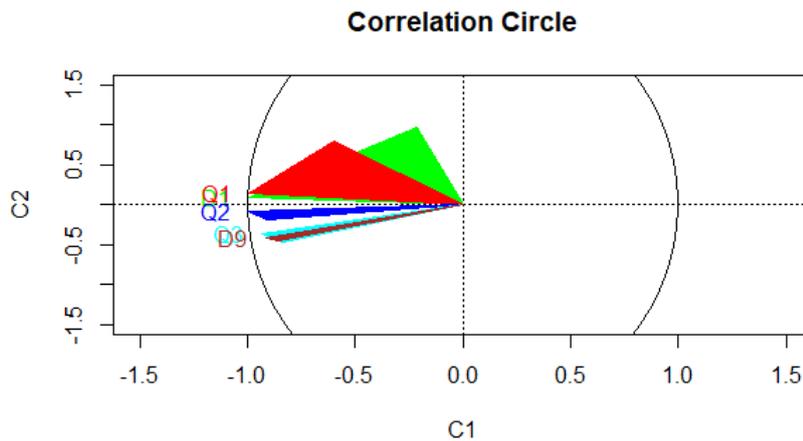


Fig. 3. Symbolic PCA (‘centers’ option) for Quantile Estimates, factorial plane F1xF2 of EU12 countries. Source: author’s processing, from EU-FADN 2006.

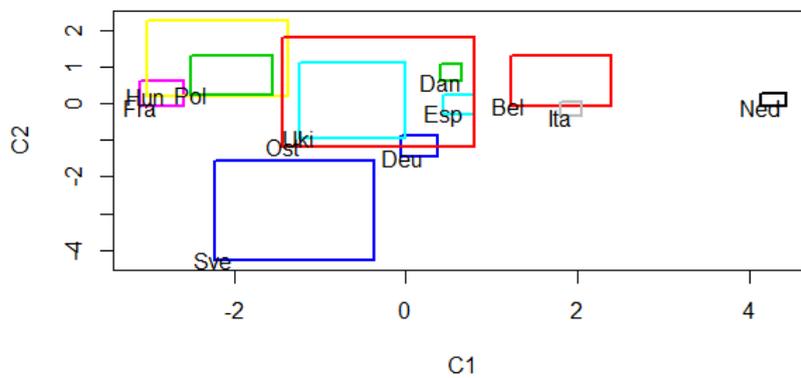


Fig. 4. Symbolic PCA (‘centers’ option) for Quantile Estimates, factorial plane C1xC2 of EU12 countries. Source: author’s processing, from EU-FADN 2006.

In the factorial plane of the first two components C1xC2 (fig.4), the Netherlands are plotted with Belgium and Italy in the first quadrant ($C1C2>0$), indicating a lower general level of fertilizer estimates. In the opposed half-plan ($C1<0$), France, Hungary, Poland are plotted with Sweden, indicating much higher general levels of fertilizer estimates. Along the C1 component negative side, Austria and United Kingdom are nearer from Sweden while Germany (Deu), Denmark (Dan) and Spain (Esp) along the positive side of C1 component are plotted nearer Belgium.

Along the C2 component, Sweden is clearly opposed to the other countries, taking in account its extreme D9 estimates.

Countries symbolised by a larger rectangle are Austria, Sweden, Hungary, United-Kingdom, Belgium, and Poland, which correspond to those with greater interval range. Conversely, countries symbolised by a smaller rectangle are Denmark, Italy, Netherlands, Spain, Germany, and France, which are characterised by a narrower range of estimate intervals.

For individuals, alternate projections are provided by the “best point” PCA options, the optimised distance option on one hand, and on the other hand the optimized variance option.

As shown in table 2, the optimised variance option of the PCA maximizes the variance of the first components since the cumulative percentage of variance of the first factorial plan is the highest (98.7%) compared to the optimal distance option (94.9 %) and to the classical PCA (97.4%). So, the optimised variance option provides a more complete summary in two dimensions.

%_cum_var	Classic	Optdist	Optvar
C1	75,2	69,5	65,4
C2	97,4	94,9	98,7
C3	99,3	99,2	100,0
C4	99,9	99,9	100,0
C5	100,0	100,0	100,0

Tab. 2. Comparison of the percentage of cumulative variance between the principal components of the three following PCA options: classical PCA (clC_{ssic}), optimised distance (Optdist), and optimised variance (Optvar). Source: author’s processing, from EU-FADN 2006.

Except for the third principal component (C3), the optimised distance option of the interval PCA displays the minimum absolute deviation (MAD) between supremum and infimum vertices over the principal components, compared to the centers option and the optimised variance options (cf. table 3). So the optimised distance option provides a narrower display of interval estimates for quantile.

Method	MAD_C1	MAD_C2	MAD_C3	MAD_C4	MAD_C5
Centers	0,93	1,27	1,21	0,89	0,33
Optdist	0,61	0,83	0,88	0,32	0,27
Optvar	0,67	1,17	0,73	0,39	1,06

Tab. 3. Comparison of the mean absolute deviation (MAD) between the principal components of three PCA options: centers PCA (Centers), optimised distance (Optdist), and optimised variance (Optvar). Source: author's processing, from EU-FADN 2006.

In the first factorial plane, the optimised distance and the optimised variance options display a pattern of correlations between quantile estimates and principal components very similar to those of the classic PCA on the two first principal components. As shown by their contributions to inertia (table 5), the first two principal components have roughly the same definition in terms of quantile. The correlations between quantile estimates and the other principal components (C3, C4 and C5) are different from the classical PCA for the optimised variance option, however without few practical implications due to the very small level of inertia (below 5%) expressed by these components.

Contr. (%)	C1			C2			C3			C4			C5		
	Classic	Optdist	Optvar												
D1	13	11	5	43	41	49	35	48	16	9	0	3	0	0	27
Q1	19	15	16	22	29	28	26	49	5	30	5	0	3	2	50
Q2	25	28	30	2	1	1	20	1	29	31	45	23	22	25	18
Q3	22	24	25	16	12	11	0	1	1	3	2	63	58	60	0
D9	21	22	24	17	17	11	19	1	49	26	47	12	17	13	4

Tab. 4. Comparison of the relative contribution to inertia (Contr.) between the principal components of the three PCA options: classic PCA (Classic), optimised distance (Optdist), and optimised variance (Optvar). Source: author's processing, from EU-FADN 2006.

The contributions to inertia for the national distributions of fertilizer estimates (table 5) show similar patterns on the two first components between the optimised options and the classic PCA, with the exception of Poland opposed to Sweden in the optimised variance option, instead of Hungary in classic and optimised distance options for the C2 component.

Contr. (%)	C1			C2			C3			C4			C5		
	Classic	Optdist	Optvar												
Bel	7	11	9	3	1	3	0	0	0	2	3	0	2	2	0
Dan	1	1	1	5	4	1	16	8	5	0	9	58	0	0	1
Deu	0	0	0	10	7	4	16	1	6	1	15	10	1	1	0
Esp	1	1	1	0	0	0	12	3	25	5	17	28	45	31	0
Fra	18	17	20	1	1	3	5	2	6	21	13	0	23	31	46
Hun	11	13	8	11	11	1	0	0	16	0	3	0	15	17	48
Ita	8	7	7	0	0	0	1	0	3	2	3	0	1	0	0
Ned	41	35	40	0	1	0	2	1	11	6	9	0	2	4	4
Ost	0	0	1	1	0	15	26	77	3	48	7	0	0	2	0
Po1	9	10	8	5	2	1	13	2	12	3	17	0	3	3	0
Sve	4	3	3	64	69	70	7	2	2	0	0	3	7	6	0
Uki	1	2	1	0	3	0	3	5	11	11	4	0	2	3	0

Tab. 5. Comparison of the relative contribution to inertia (Contr.) between the principal components of the three PCA options: classic PCA (Classic), optimised distance (Optdist), and optimised variance (Optvar). Source: author's processing, from EU-FADN 2006.

As summarized by the mean absolute deviation in table 3, the display of all country rectangle projections is the largest into the centers option (figure 4) and the smallest into the optimised distance option (figure 6) while the display of the optimised variance option (figure 5) is of medium range between the two previous options, both in the lengths (dimension 1 of the first principal component) and the widths (dimension 2 of the second principal component).

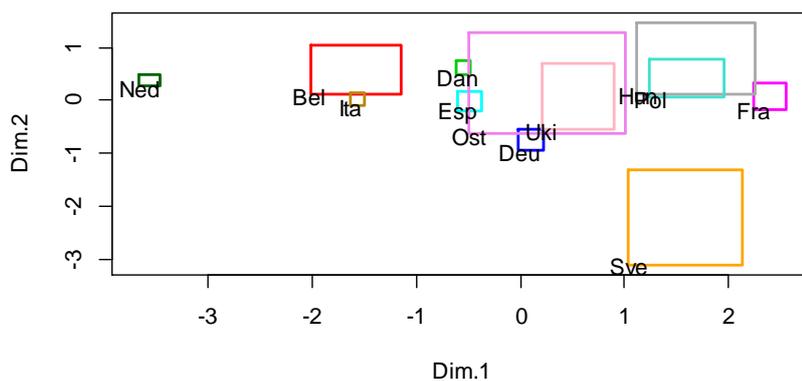


Fig. 5. Symbolic PCA ('optimized.distance' option) for Quantile Estimates, factorial plane F1x2 of EU12 countries. Source: author's processing, from EU-FADN 2006.

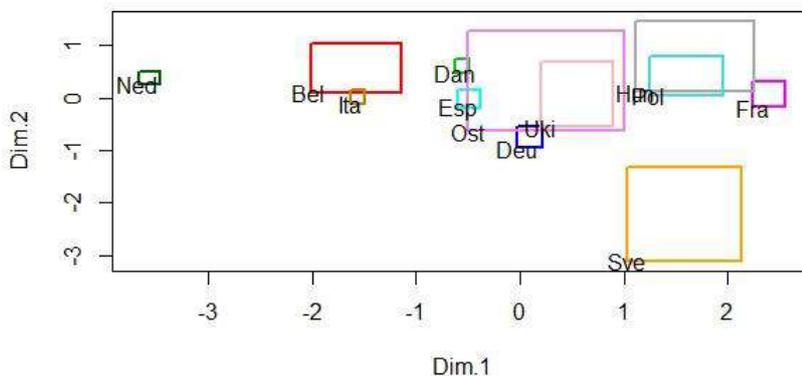


Fig. 6. Symbolic PCA ('optimized.variance' option) for Quantile Estimates, factorial plane F1x2 of EU12 countries. Source: author's processing, from EU-FADN 2006.

By the relative sizes and locations of their hyper-rectangle projections, these three factorial representations (figures 4, 5, and 6) distinguish clearly Netherlands on the first principal component, as the archetype of the *location shift* model, and

Sweden, on the second principal component, as the archetype of the *location scale shift* model.

3.2 The divisive hierarchy of fertilizer cost estimates

The divisive hierarchy obtained with Euclidean distance option shows that the set of D1, Q1, Q2 and Q3 quantile estimates is used by the discriminant values, which implies keeping these parameters to describe the distribution, and possibly extending it by a finer quantile scale allowing some of the national distributions to be better distinguished.

The first partition in two clusters corresponds to the supremum level of the median estimate (Q2S).

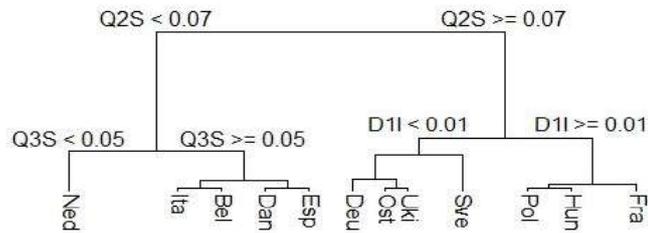


Fig. 7. Symbolic Divisive Clustering ('Euclidean distance' option) for Quantile Estimates, EU12 countries. Source: author's processing, from EU-FADN 2006.

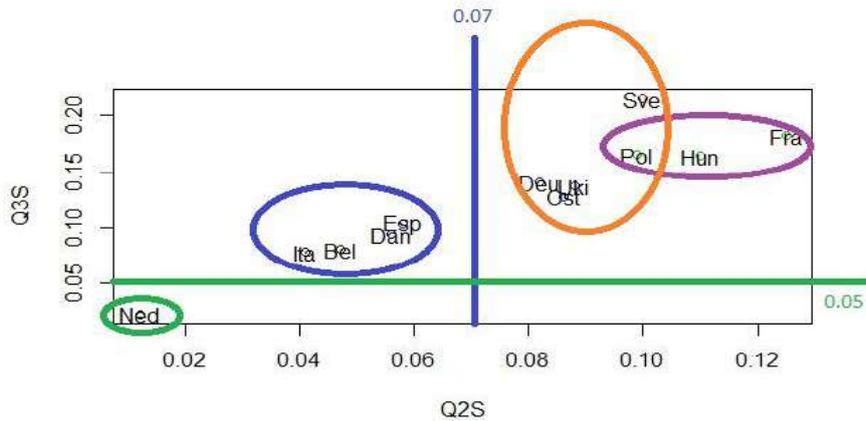


Fig. 8. Symbolic Divisive Clustering (C4 optimal partition for Determinant Ratio Criterion) for Quantile Estimates, factorial plane F1xF2 of EU12 countries. Source: author's processing, from EU-FADN 2006.

At the top of the divisive hierarchy, the clustering procedure allows to identify two contrasted models for empirical distributions of the fertilizers technical coefficients for yearly crops production costs used to € 1,000 of gross product. As the first cluster, Netherlands (*Ned*) and the group of Italy (*Ita*), Belgium, (*Bel*), Denmark (*Dan*) and Spain (*Esp*) grouped by their supremum median ($Q2S$) levels which are lower than € 7, are split in the following divisive step by the supremum higher quartile ($Q3S$) level of € 5 which identifies Netherlands as the less intensive in fertilizer input. Netherlands is the archetype of the *location-shift* model formalizing the assumption of homogeneous producers in their fertilizer costs. As the second cluster, for which their supremum median ($Q2S$) of fertilizer cost is greater than 7 €, is split into two groups: first, the group for which the fertilizer first decile input is greater than € 1, i.e. the subgroup formed by Poland (*Pol*), and Hungary (*Hun*) aggregated with France (*Fra*); second, the group formed by Sweden (*Sve*) aggregated with the subgroup formed by Germany (*Deu*), Austria (*Ost*) and United-Kingdom (*Uki*), on the basis of their fertilizer first decile lesser than € 1 input level. This latter group illustrates the *location-scale shift* model, formalizing the assumption of heterogeneous producers in their fertilizer costs. The partition into four groups displays by figure 8 is the optimal partition for the minimum difference in the logarithm of the ratio of determinants (package *ClusterCrit*), which is a consistent rule with the criterion of the DIVCLUS-T algorithm.

Conclusions

Based on quantile regression and symbolic data analysis, this paper presents a global methodology which aims to keep as much as possible relevant information for the policy design, all along the econometric process of estimating and analyzing agricultural fertilizer costs for yearly crops production. The different properties of three options of interval PCA (centers, optimized distance and optimized variance) are described allowing to identify different models of distributional scale, notably that of the *location shift* model opposite that of the *location-scale shift* one. Differences and similarities between interval estimates are exploited using divisive hierarchical clustering to produce two country clusters identifying through quantile cost thresholds the archetypes of the *location shift* model and the *location-scale shift* one. The differences between four groups of countries are delimited by optimal thresholds expressed according to the conditional quantiles in unitary terms of the gross product. These thresholds can be used for segmenting farm populations to later analyze the differential impacts of agricultural policy measures. We will apply this methodology at the second level of the European Nomenclature of Territorial Units for Statistics (NUTS 2, 281 regions).

References

1. J. Angrist and J. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton Univ Press, Princeton, 2009.

2. L. Billard and E. Diday. *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, Wiley Interscience, Chichester, 2006.
3. C. G. Broyden, « The Convergence of a Class of Double-rank Minimization Algorithms », *Journal of the Institute of Mathematics and Its Applications*, vol. 6, 1970, p. 76-90.
4. A. C. Cameron and P. K. Trivedi. *Microeconometrics. Methods and Applications*. Cambridge University Press, 2005.
5. P. Cazes, A. Chouakria, E. Diday and Y. Schektman. Extension de l'analyse en composantes principales à des données de type intervalle, *Revue de statistique appliquée*, 45, 3, 5-24, 1997.
6. M. Chavent, Y. Lechevalier, O. Briant. DIVCLUS-T: A monothetic divisive hierarchical clustering method. *Computational Statistics & Data Analysis*, 52, 2, 687-701, 2007.
7. A. Chouakria, E. Diday, and P. Cazes. An improved factorial representation of symbolic objects. In *Studies and Research, Proceedings of the Conference on Knowledge Extraction and Symbolic Data Analysis (KESDA'98)*, Luxembourg: Office for Official Publications of the European Communities, 276–289, 1998.
9. G.B. Dantzig. Programming in a linear structure, *Econometrica*, 17, 73–74, 1949.
10. D. Desbois. Estimation des coûts de production agricoles : approches économétriques. PhD dissertation directed by J.C. Bureau and Y. Surry, ABIES-AgroParisTech, Paris, 2015.
11. D. Desbois, J.-P. Butault, and Y. Surry. Distribution des coûts spécifiques de production dans l'agriculture de l'Union européenne : une approche reposant sur la méthode de régression quantile, *Économie rurale*, 361, 3-22, 2017.
12. Desgraupes B. Clustering Indices. Lab Modal'X, Paris-Ouest University, 22 pages, November 2017.
13. E. Diday. Thinking by Classes in Data Science: the Symbolic Data Analysis Paradigm, *WIREs Computational Statistics*, 8, 171-205, 2006.
14. J.F. Divay and F. Meunier. Deux méthodes de confection du tableau entrées-sorties. *Annales de l'INSEE*, 37, 59-109, 1980.
15. X. D'Haultfoeuille and P. Givord. La régression quantile en pratique, *Economie et statistique*, 71, 85-111, 2014.
16. R. Fletcher, « A New Approach to Variable Metric Algorithms », *Computer Journal*, vol. 13, 1970, p. 317-322.
17. D. Goldfarb, « A Family of Variable Metric Updates Derived by Variational Means », *Mathematics of Computation*, vol. 24, 1970, p. 23-26.
18. X. He and F. Hu. Markov Chain Marginal Bootstrap, *Journal of the American Statistical Association*, 97, 783–795, 2002.
19. T. Jevnaker and J. Wettstad. Ratcheting Up Carbon Trade: The Politics of Reforming EU Emissions Trading, *Global Environmental Politics*, 17(2), pp. 105–124, 2017.
20. R. Karmarkar. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4, 373-395, 1984.
21. R. Koenker, and G. Bassett. Regression quantiles. *Econometrica*, 46, 33-50, 1978.
22. R. Koenker and V. d'Orey. Remark AS R92: A remark on algorithm AS 229: Computing dual regression quantiles and regression rank scores. *Applied Statistics*, 43, 410-414, 1994.
23. R. Koenker and Q. Zhao. L-estimation for linear heteroscedastic models. *Journal of Nonparametric Statistics*, 3, 223-235, 1994.
24. J. Garro and O. Rodriguez. Optimized Dimensionality Reduction Methods for Interval-Valued Variables and Their Application to Facial Recognition Entropy 2019, 21, 1016; doi:10.3390/e21101016

25. I. J. Lustig, R. E. Marsden and D. F. Shanno. On implementing Mehrotra's predictor-corrector interior-point method for Linear Programming. *SIAM Journal on Optimization*, 2, 435-449, 1992.
26. K. Madsen and H. B. Nielsen. A Finite Smoothing Algorithm for Linear Estimation, *SIAM Journal on Optimization*, 3, 223-235, 1993.
27. B. Mirkin. *Clustering for Data Mining. A Data Recovery Approach*. Chapman & Hall, London, 2005.
28. S. Portnoy and R. Koenker. The gaussian hare and the laplacian tortoise: Computation of squared-errors vs. absolute-errors estimators. *Statistical Science*, 1, 279-300, 1977.
29. A. Quinet. *La valeur de l'action pour le climat. Une valeur tutélaire du carbone pour évaluer les investissements et les politiques publiques*, France Stratégie, 2019.
30. Rodriguez, O. *Classification et Modèles Linéaires en Analyse des Données Symboliques*. PhD Thesis, Paris IX-Dauphine University, France, 2000.
31. O. Rodriguez, E. Diday and S. Winsberg *Generalizations of Principal Component Analysis*
31. D. F. Shanno, « Conditioning of Quasi-Newton Methods for Function Minimization », *Mathematics of Computation*, vol. 24, 1970, p. 647-656.
32. J. Ypma *et alii*. Package 'nloptr', CRAN repository, 2020

ⁱ This communication is the continuation of some of author's works done during the preparation of the PhD dissertation (Desbois [10]), co-directed by Y. Surry and J.C. Bureau, supported by the *Farm Accountancy Cost Estimation and Policy Analysis* project (FACEPA) of the 7th Framework Program of the European Community (FP7 / 2007-2013, Approval No. 212292). This mention does not imply any approval by the persons and organizations mentioned, the author assuming full responsibility for the text.

Simulation studies for a special mixture regression model with multivariate responses on the simplex

Agnese Maria Di Brisco¹, Roberto Ascari², Sonia Migliorati², and Andrea Ongaro²

¹ Department of Studies in Economics and Business (DISEI), Università del Piemonte Orientale, via E. Perrone 18, 28100, Novara, Italy
(E-mail: agnese.dibrisco@uniupo.it)

² Department of Economics, Management and Statistics (DEMS), Università di Milano-Bicocca, P.zza dell'Ateneo Nuovo 1, 20126, Milano, Italy
(E-mail: roberto.ascari@unimib.it, sonia.migliorati@unimib.it, andrea.ongaro@unimib.it)

■

Abstract. Compositional data are defined as vectors whose elements are strictly positive and subject to a unit-sum constraint. When the multivariate response is of compositional type, a proper regression model that takes account of the unit-sum constraint is required. This contribution illustrates a new multivariate regression model for compositional data that is based on a mixture of Dirichlet distributed components. Its complex structure is offset by good theoretical properties (among which identifiability) and a greater flexibility than the standard Dirichlet regression model. We perform intensive simulation studies to evaluate the fit of the proposed regression model and its robustness in the presence of multivariate outliers. The (Bayesian) estimation procedure is performed via the efficient Hamiltonian Monte Carlo algorithm.

Keywords: mixture distribution, multivariate regression, compositional data, Hamiltonian Monte Carlo.

1 Introduction

Compositional data, namely proportions of some whole, are encountered in several fields of science and require proper statistical tools of analysis [1]. Indeed, compositional data have the peculiarity of being vector of proportions lying on the simplex

space: $\mathcal{S}^D = \{\mathbf{Y} : Y_j > 0, j = 1, \dots, D, \sum_{j=1}^D Y_j = 1\}$. The analysis of compositional

data is challenging since it cannot make use of standard techniques that might lead to distorted results due to ignoring the unit-sum constraint. A fruitful strategy in the analysis of compositional data takes advantage of statistical distributions defined on the simplex. Among them, the Dirichlet distribution is a widespread one for a D -dimensional vector $\mathbf{y} \in \mathcal{S}^D$. A regression model based on the Dirichlet distribution is straightforward for compositional data and proves to behave satisfactorily [3][6][7]. As a counterpart, it has some limitations among which its inability of modeling multimodality, heavy tails, and the eventual presence of outliers. A convenient approach to induce multimodality and an overall increase in flexibility is to consider a mixture

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain



distribution. In this regard, we propose to resort to a special finite mixture of Dirichlet components referred to as the Extended Flexible Dirichlet (EFD) [9]. Moreover, we illustrate a regression model based on the EFD distribution [11]. Aim of this work is to intensively study the behavior of the EFD regression model in many simulated scenarios covering some relevant statistical issues such as the presence of outliers, heavy tails, and latent groups. We compare the EFD regression model with the Dirichlet one in terms of fit and estimates of the regression parameters.

The rest of the paper is organized as follows. Section 2 introduces the Dirichlet and the EFD distributions and it shows convenient parameterizations for regression purposes. Section 3 outlines details on the EFD regression model. Section 3.1 provides an overview on the HMC algorithm, a Bayesian approach to inference especially suited for mixture models. Section 4 illustrates several simulation studies that have been performed to evaluate the behavior and the fit to data of the EFD regression model in comparison to the Dirichlet one.

2 Dirichlet and EFD distributions

A Dirichlet distributed D -dimensional vector $\mathbf{y} \in \mathcal{S}^D$ has a probability density function (p.d.f.) as follows:

$$f_D(\mathbf{y}; \boldsymbol{\alpha}) = \frac{\Gamma(\boldsymbol{\alpha}^+)}{\prod_{j=1}^D \Gamma(\alpha_j)} \prod_{j=1}^D y_j^{\alpha_j-1}, \quad (1)$$

where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_D)^\top$, $\alpha_j > 0$, and $\boldsymbol{\alpha}^+ = \sum_{j=1}^D \alpha_j$. With the aim of regressing a compositional vector onto a set of covariates, it is convenient to work with an alternative parameterization based on the mean vector $\bar{\boldsymbol{\alpha}} = (\bar{\alpha}_1, \dots, \bar{\alpha}_D)^\top \in \mathcal{S}^D$, where $\bar{\alpha}_j = \mathbb{E}[Y_j] = \frac{\alpha_j}{\boldsymbol{\alpha}^+}$ for $j = 1, \dots, D$, and $\boldsymbol{\alpha}^+ = \sum_{j=1}^D \alpha_j > 0$ that represents the precision parameter of the Dirichlet distribution.

An EFD distributed D -dimensional vector $\mathbf{y} \in \mathcal{S}^D$ has the following p.d.f.:

$$f_{EFD}(\mathbf{y}; \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}) = \left(\prod_{r=1}^D \frac{y_r^{\alpha_r-1}}{\Gamma(\alpha_r)} \right) \sum_{h=1}^D p_h \frac{\Gamma(\alpha_h) \Gamma(\boldsymbol{\alpha}^+ + \tau_h)}{\Gamma(\alpha_h + \tau_h)} y_h^{\tau_h}, \quad (2)$$

where $\mathbf{p} \in \mathcal{S}^D$ and vectors $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_D)^\top$ and $\boldsymbol{\tau} = (\tau_1, \dots, \tau_D)^\top$ have positive elements. The EFD distribution function admits the following mixture representation:

$$\text{EFD}(\mathbf{y}; \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{p}) = \sum_{r=1}^D p_r \text{Dir}(\mathbf{y}; \boldsymbol{\alpha} + \tau_r \mathbf{e}_r), \quad (3)$$

where $\text{Dir}(\cdot; \cdot)$ denotes the Dirichlet distribution, and \mathbf{e}_r is a vector of zeros except for the r -th element which is equal to one. It is worth noting that the EFD distribution contains the Dirichlet as an inner point when $\tau_r = 1$ and $p_r = \bar{\alpha}_r$ for every $r = 1, \dots, D$. The p.d.f. of the EFD admits a variety of shapes including, but not limited to, uni- and multi-modal ones. Moreover, the richer parameterization of the EFD with respect to the Dirichlet allows for a more flexible modelization of the dependence structure of the composition. Last, the EFD distribution shows several theoretical properties, i.e.

some simplicial forms of dependence/independence and identifiability [9], that make it tractable from computational and inferential points of view.

To define a regression model based on the EFD, it is convenient to adopt an alternative parameterization that explicitly includes the mean vector. To this end, note that the r -th Dirichlet component in (3) has a mean vector $\boldsymbol{\lambda}_r = (1 - w_r)\bar{\boldsymbol{\alpha}} + w_r\mathbf{e}_r$, (where $\bar{\boldsymbol{\alpha}} = \frac{\boldsymbol{\alpha}}{\alpha^+}$ and $w_r = \frac{\tau_r}{\alpha^+ + \tau_r}$), which can be interpreted as a weighted average of a common barycenter, $\bar{\boldsymbol{\alpha}}$, and the r -th simplex vertex \mathbf{e}_r . The first order moment of the EFD easily follows from its mixture structure:

$$\boldsymbol{\mu}_j = \mathbb{E}[Y_j] = \sum_{r=1}^D p_r \boldsymbol{\lambda}_{rj} = \bar{\alpha}_j \sum_r p_r (1 - w_r) + p_j w_j. \quad (4)$$

However, the parameterization of the EFD based on $\boldsymbol{\mu}_j$, p_j and w_j ($j = 1, \dots, D$) is not variation independent since some constraints hold between parameters and thus the following inequalities referred to w_j , $j = 1, \dots, D$, can be derived:

$$(i) w_j < \frac{\boldsymbol{\mu}_j}{p_j}, \quad (ii) w_j > \frac{\boldsymbol{\mu}_j}{p_j} - \frac{1 - \sum_r p_r w_r}{p_j}.$$

Variation independence, whose lack is a potential issue for Bayesian inference through Monte Carlo (MC) methods, can be achieved by normalizing w_j as follows:

$$\tilde{w}_j = \frac{w_j}{\min\left\{\frac{\boldsymbol{\mu}_j}{p_j}, 1\right\}}, \quad j = 1, \dots, D. \quad (5)$$

The parameterization of the EFD distribution depending on $\boldsymbol{\mu} \in \mathcal{S}^D$, $\mathbf{p} \in \mathcal{S}^D$, $\tilde{w}_j \in (0, 1)$ for every j , and $\alpha^+ > 0$ has the double benefit of being variation independent—useful for Bayesian inference—and explicitly including the mean vector—useful for regression purposes.

3 Dirichlet and EFD regression models

Since both parameterizations of the Dirichlet and of the EFD illustrated in Section 2 explicitly include the mean vector $\boldsymbol{\mu}$, it is possible to derive a regression model for compositional data. Let $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_n)^\top$ be the response matrix such that \mathbf{Y}_i , for $i = 1, \dots, n$, is a D -dimensional vector on the simplex, and let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$ be the design matrix such that \mathbf{x}_i are $(K + 1)$ -dimensional vectors. The mean vector \mathbf{v}_i of \mathbf{Y}_i can be regressed onto a set of covariates in accordance to a GLM strategy [8]. Indeed, since \mathbf{v}_i lies on the simplex, a multinomial logit link function can be adopted:

$$g(v_{ij}) = \log\left(\frac{v_{ij}}{v_{iD}}\right) = \mathbf{x}_i^\top \boldsymbol{\beta}_j, \quad (6)$$

where $v_{ij} = \mathbb{E}[Y_{ij}]$, $\mathbf{x}_i = (1, x_{i1}, \dots, x_{iK})^\top$ is the vector of covariates, and $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, \dots, \beta_{jK})^\top$ is a vector of regression coefficients. Please note that the D -th category is conventionally fixed as baseline, so that $\beta_{Dk} = 0$ for $k = 0, 1, \dots, K$, and thus:

$$v_{ij} = g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta}_j) = \begin{cases} \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta}_j)}{1 + \sum_{r=1}^{D-1} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_r)}, & \text{for } j = 1, \dots, D-1 \\ \frac{1}{1 + \sum_{r=1}^{D-1} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_r)}, & \text{for } j = D. \end{cases} \quad (7)$$

If \mathbf{Y}_i are Dirichlet distributed, we recover the Dirichlet regression (DirReg) model [67] by substituting v_{ij} with $\bar{\alpha}_{ij}$ in (6). Similarly, if \mathbf{Y}_i are EFD distributed, we get the EFD regression (EFDReg) model [11] by replacing v_{ij} with μ_{ij} in (6).

3.1 Inference and fit

To obtain estimates of the unknown parameters of EFDReg and DirReg models we favor a Bayesian approach. This choice is mainly motivated by the difficulty, both computational and analytical, of likelihood-based inferential approaches in dealing with complex models such as mixtures. Conversely, the finite mixture structure of the EFD distribution can be advantageously treated as an incomplete data problem in a Bayesian paradigm [4]. Among the MC methods, a recent solution is the Hamiltonian Monte Carlo (HMC) [10] algorithm, a generalization of the Metropolis algorithm which combines Markov Chain Monte Carlo (MCMC) and deterministic simulation methods. The (simulated) posterior distributions of the unknown parameters are simulated on the basis of the full likelihood and prior distributions. With regard to priors' choice, we adopt non- or weakly informative priors to induce the minimum impact on the posteriors [2], and suppose prior independence. We select a multivariate normal with zero mean vector and diagonal covariance matrix with “large” values of the variances as non-informative prior for the regression parameters $\boldsymbol{\beta}_j$. Moreover, we adopt a Uniform(0, 1) prior for \tilde{w}_j , $j = 1, \dots, D$, and a Dirichlet prior with hyperparameter $\mathbf{1}$ for the vector \mathbf{p} . Last, we use a Gamma(g, g) prior, with g “small” and equal to 0.001, for the precision parameter α^+ . Among the variety of fitting criteria, we favor the Watanabe-Akaike information criterion (WAIC) [13] because it is fully Bayesian and well-defined for non-regular models such as mixture ones as well. As a general rule, the smaller the criterion is, the better the model fit.

4 Simulation studies

To compare the performances of the EFDReg and DirReg models, we simulated a variety of scenarios that cover many potentially tricky problems among which multimodality, as well as presence of heavy tails, outliers, and latent groups. In what follows we illustrate the samples' simulation schemes and the main inferential results for each scenario. We took advantage of the HMC algorithm for estimating the vector of unknown parameters $\boldsymbol{\eta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_{D-1}, \alpha^+, \mathbf{p}, \tilde{\mathbf{w}})^\top$. The algorithm is easily implemented by the Stan modeling language [12]. We run chains of length 10,000 and we discarded the first half. Moreover, we checked the convergence to the target distribution through graphical tools, such as trace-plots, density-plots, and autocorrelation-plots, as well as diagnostic measures such as the potential scale reduction factor, the effective sample size, and Raftery-Lewis test [5]. Each scenario is replicated 500 times to evaluate MC estimates. Computational times are of approximately 60 seconds for DirReg and 300 seconds for EFDReg.

Fitting study First, we evaluated some fitting studies by simulating from Dirichlet (scenario (i)) and EFD (scenario (ii)) regression models. The objective of these studies is to analyze the goodness of fit and estimates of regression coefficients. The sample size is $n = 150$, and the multivariate response lies on the 3-part simplex. In both

scenarios, a quantitative covariate x , uniformly distributed in $(-0.5, 0.5)$, is included in the regression model for the mean (see equation (6)), with regression coefficients set equal to $\beta_{10} = 1$, $\beta_{11} = 2$, $\beta_{20} = 0.5$, $\beta_{21} = -3$. In scenario (i) the response is Dirichlet distributed and the precision parameter is $\alpha^+ = 50$. In scenario (ii) the response is EFD distributed, and the remaining parameters are fixed equal to $\alpha^+ = 50$, $\mathbf{p} = (1/3, 1/3, 1/3)^\top$, and $\tilde{\mathbf{w}} = (0.6, 0.2, 0.7)^\top$. Ternary plots and scatterplots of each element of the composition y_{ij} , $j = 1, 2, 3$, versus the quantitative covariate x are shown in Figures 1 (scenario (i)) and 2 (scenario (ii)). It is worth noting the absence of whatever cluster in scenario (i), while in scenario (ii) it is clear the presence of clusters and multiple modes.

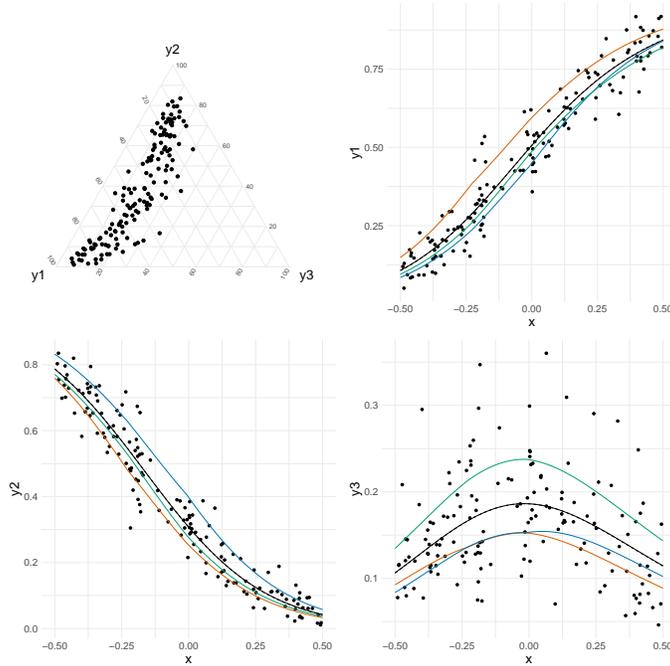


Fig. 1: Representations of one replication from scenario (i).

Presence of outliers To evaluate the behavior of the DirReg and EFDReg models in the presence of outliers we perturbed scenario (i) according to the following perturbation scheme. We randomly selected 15 observations (10% of the sample size) and we applied the perturbation operation defined as $\mathbf{y} \oplus \boldsymbol{\delta} = \mathcal{C}\{y_1 \cdot \delta_1, \dots, y_D \cdot \delta_D\} \in \mathcal{S}^D$, where \mathbf{y} and $\boldsymbol{\delta}$ are vectors on the simplex playing the roles of perturbed and perturbing element, respectively. Moreover, the closure operation $\mathcal{C}\{\cdot\}$ is defined as $\mathcal{C}\{\mathbf{q}\} = \{q_1/q^+, \dots, q_D/q^+\}$ with $q^+ = \sum_{j=1}^D q_j$ and $q_j > 0, \forall j = 1, \dots, D$. The neutral element of the perturbation operation is $\boldsymbol{\delta} = (1/D, \dots, 1/D)^\top$, so that if element y_j is perturbed by δ_j greater (lower) than $1/D$, the perturbation is upward (downward). We set three scenarios of perturbation by fixing the perturbing element $\boldsymbol{\delta}$

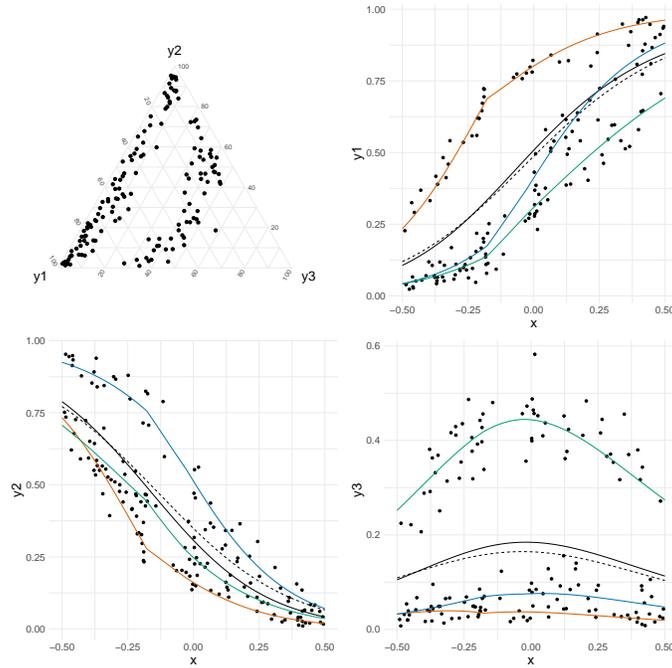


Fig. 2: Representations of one replication from scenario (ii).

equal to $(0.86, 0.07, 0.07)^\top$ in scenario (I), $(0.07, 0.86, 0.07)^\top$ in scenario (II), and $(0.07, 0.07, 0.86)^\top$ in scenario (III). Figures 3, 4, and 5 show the effect of perturbation on the Dirichlet-distributed responses. In all plots the perturbed points are in light blue while unperturbed points are in black. Looking at the scatterplots, we can observe that scenario (I) assumes some outlying observations upward for the first element and downward for the second and third elements of the composition; this is coherent with the chosen vector δ which has the first element greater than 0.5 and the second and third elements lower than 0.5. Instead, in scenarios (II) and (III) the second and third elements of the composition respectively are perturbed upward while the remaining elements are perturbed downward. Focusing on the ternary plots, it is worth noting that the effect of perturbation in scenario (III) is clearly visible in that the group of perturbed values, in blue, is well-separated from the remaining points, in black. The overall effect of perturbing vector $\delta = (0.07, 0.07, 0.86)^\top$ is thus to shift the cloud of points towards the bottom-right vertex of the plot. Conversely, in scenarios (I) and (II) the perturbed points are overall shifted towards the bottom-left and top vertex of the ternary plot, respectively, i.e. in a region with a higher presence of unperturbed points.

Presence of latent groups The following simulation study explores the case of presence of a latent (unobserved) covariate that induces the occurrence of clusters. So data are simulated by including an additional covariate in the regression model that is assumed unknown, not accounted for by the estimates of the Dirichlet and EFDReg models. In particular, we replicated the generating mechanism of fitting study (i) by

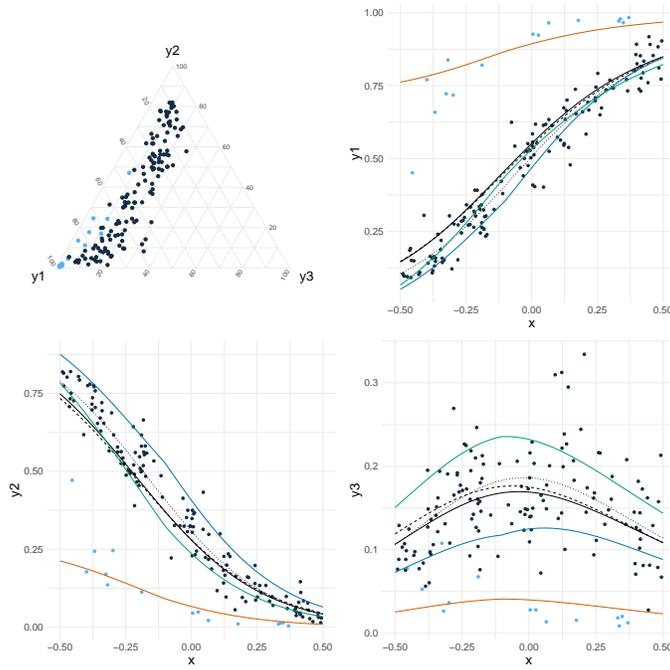


Fig. 3: Scenario (I). Perturbed points are in light-blue and unperturbed points are in black.

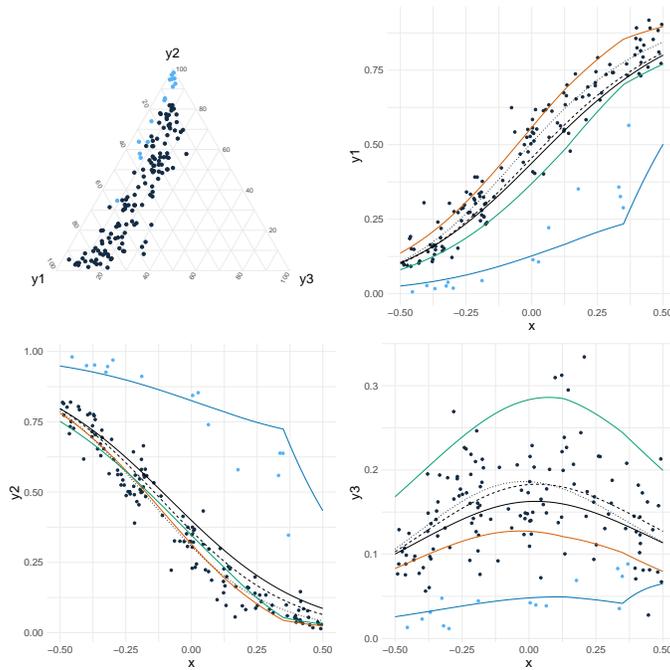


Fig. 4: Scenario (II). Perturbed points are in light-blue and unperturbed points are in black.

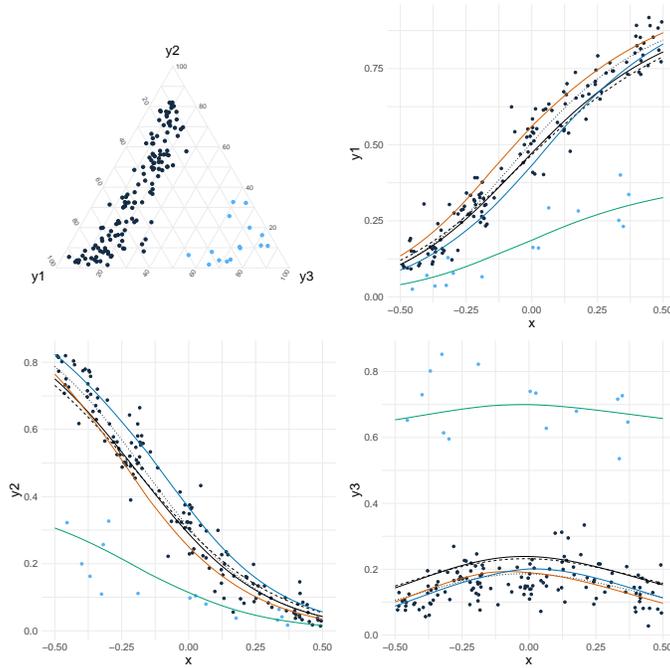


Fig. 5: Scenario (III). Perturbed points are in light-blue and unperturbed points are in black.

adding a latent dichotomous covariate (scenario (a)) and a latent covariate with three categories (scenario (b)). In scenario (a), the additional regression coefficients are $\beta_{12} = -1$ and $\beta_{22} = 2$, and in scenario (b) they include also $\beta_{13} = 0.5$ and $\beta_{23} = -3$. With respect to the dichotomous covariate of scenario (a), the categories have probabilities of 0.3 and 0.7. In scenario (b), the three categories of the latent covariate have probabilities of 0.3, 0.15, and 0.55. Figures 6 and 7 show one random replication from scenarios (a) and (b) with latent groups, respectively. In the ternary plots, points are colored and shaped according to their belonging to the latent groups. The existence of two and three clusters respectively is particularly visible from the scatterplots referred to the first and second elements of the composition.

Generic mixture of Dirichlet Last, we evaluate the case of a generic mixture of two Dirichlet distributions: $\pi \text{Dir}(\mathbf{y}_i; \bar{\boldsymbol{\alpha}}_i, \alpha_1^+) + (1 - \pi) \text{Dir}(\mathbf{y}_i; \bar{\boldsymbol{\alpha}}_i, \alpha_2^+)$. Please note that the mixture structure is not of EFD type. Both Dirichlet distributions have the same regression model equal to that of scenario (i), but they differ in their precision parameters that are equal to $\alpha_1^+ = 2$ and $\alpha_2^+ = 50$, respectively. The mixing proportion parameter π is equal to 0.3.

The generic mixture of two Dirichlet distributions has been chosen to induce heavier tails than the one of the Dirichlet. The ternary plot on top left panel of Figure 8 illustrates one random replication from the generic mixture where green points belong to the first component of the mixture and orange triangles belong to the second component. It is possible to observe that the majority of points (belonging to the second component of the mixture) are placed on the ternary plot and on the scatterplots sim-

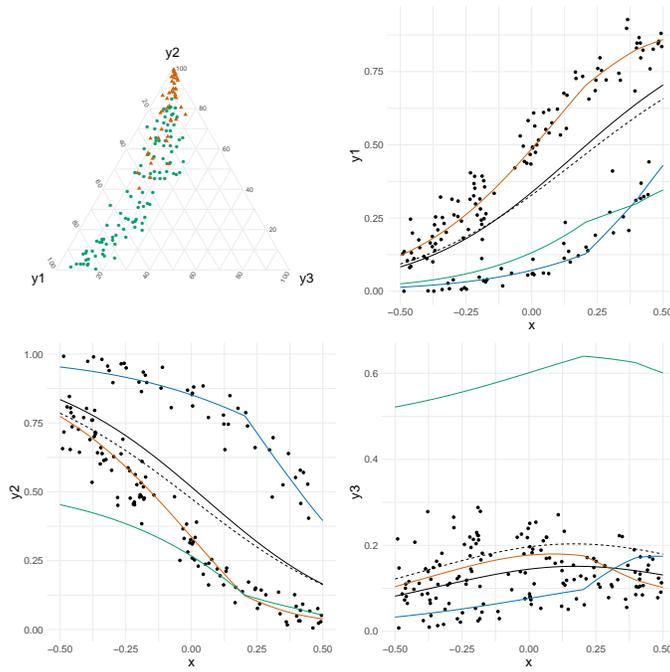


Fig. 6: Representations of one replication in the presence of latent groups, scenario (a).

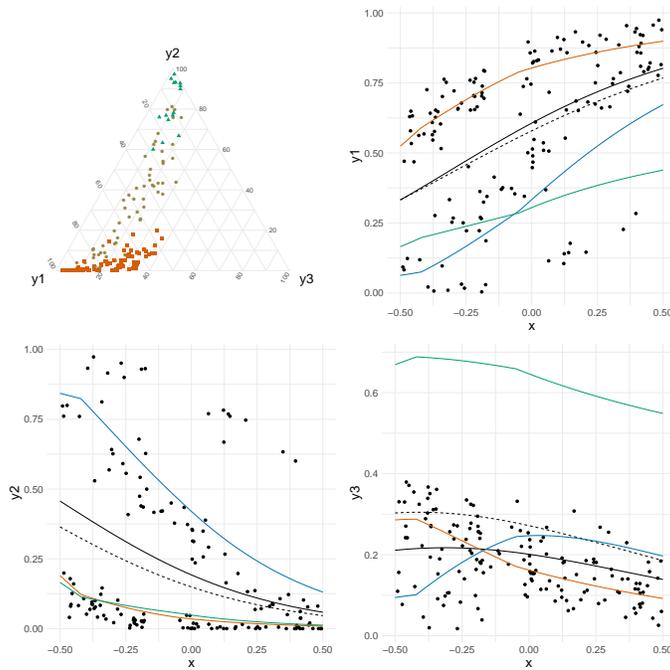


Fig. 7: Representations of one replication in the presence of latent groups, scenario (b).

ilarly to scenario (i). At the same time the group of data coming from the Dirichlet with the smaller precision parameter is far from the remaining points. Focusing on the scatterplots referred to the first and second elements of the composition (top-right and bottom-left panels of Figure 8), it is worth noting that the responses belonging to the first component of the mixture, i.e., the one with the smaller precision parameter, depart from the data cloud both upward and downward.

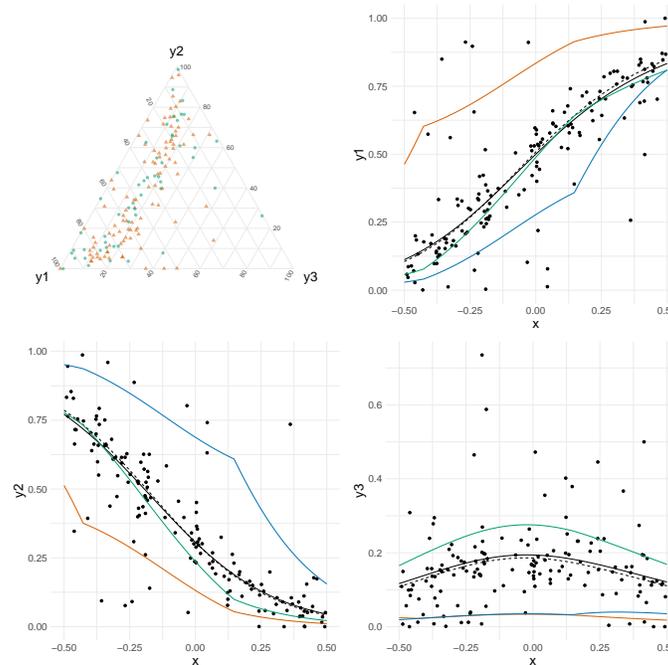


Fig. 8: Representations of one replication of a generic mixture of two Dirichlet distributions.

4.1 Comments

Table 1 shows the WAIC values in all simulation studies. In fitting study (i), where the data generating mechanism is Dirichlet, the WAIC of both models is comparable, while in all remaining scenarios the EFDReg model is far better than the DirReg one. The superiority in fit of the EFDReg model is particularly noticeable in fitting study (ii), in all scenarios with outliers, and in the presence of a latent group induced by a dichotomous covariate (scenario (a)). Scenario (b) (i.e. three latent groups) and the scenario from a generic mixture of two Dirichlet distributions are particularly challenging and result in a difficulty in fit for both models. Nevertheless, the EFDReg model is capable to provide a better adaptation to data (lower WAIC) than the DirReg.

Let us now analyze and comment the posterior means and MSEs for the Dirichlet and EFD regression models in all scenarios. All results can be found in Tables 2, 3, and 4. Moreover, we deepen the analysis of the two models by inspecting the regression

curves that are superimposed on the scatterplots in Figures 1-8. In all Figures, black solid lines refer to the EFD model and black dashed lines refer to the Dirichlet one. In some scenarios only the solid line appears meaning that the regression curves of both models are almost coincident. Colored lines are referred to the component means λ_1 (orange), λ_2 (blue), and λ_3 (green) of the EFDReg model.

Scenario	Fitting Studies		Presence of outliers			Latent groups		Generic Mixture
	(i)	(ii)	(I)	(II)	(III)	(a)	(b)	
Dir	-948.029	-512.614	-633.131	-623.990	-565.643	-430.142	-611.100	-557.735
EFD	-946.140	-883.605	-849.742	-814.106	-833.923	-770.366	-764.415	-593.995

Table 1: WAIC values in the simulation studies.

Scenario	Fitting study (i)		Latent groups (a)		Latent groups (b)	
	Dir	EFD	Dir	EFD	Dir	EFD
$\beta_{10} = 1$	1.001 (0.001)	1.001 (0.001)	0.514 (0.231)	0.850 (0.024)	0.756 (0.060)	1.104 (0.012)
$\beta_{11} = 2$	1.998 (0.018)	1.992 (0.018)	1.567 (0.203)	1.665 (0.131)	1.328 (0.461)	1.299 (0.510)
$\beta_{20} = 0.5$	0.501 (0.001)	0.502 (0.001)	0.883 (0.148)	1.275 (0.605)	-0.600 (1.213)	-0.044 (0.230)
$\beta_{21} = -3$	-3.006 (0.021)	-2.998 (0.021)	-1.963 (1.086)	-2.096 (0.835)	-1.568 (2.096)	-1.634 (1.945)
$\alpha^+ = 50$	50.052 (4.338)	53.636 (5.473)	5.894 (0.258)	21.241 (2.023)	3.033 (0.121)	5.389 (0.600)
p_1	—	0.290 (0.111)	—	0.640 (0.030)	—	0.582 (0.071)
p_2	—	0.315 (0.119)	—	0.353 (0.030)	—	0.409 (0.071)
p_3	—	0.395 (0.116)	—	0.007 (0.002)	—	0.009 (0.001)
\tilde{w}_1	—	0.149 (0.031)	—	0.608 (0.032)	—	0.584 (0.037)
\tilde{w}_2	—	0.146 (0.041)	—	0.707 (0.021)	—	0.780 (0.019)
\tilde{w}_3	—	0.151 (0.032)	—	0.461 (0.056)	—	0.421 (0.012)

Table 2: Posterior means for the Dirichlet and EFD regression models in fitting study (i) and in scenarios (a) and (b) with latent groups. MSEs for the regression coefficients and SEs for remaining parameters are in parenthesis.

Results about the fitting study with Dirichlet distributed data (scenario (i)), can be found in the second and third columns of Table 2. It is worth noting that both models provide precise estimates for the regression parameters and similar MSEs. This is confirmed by almost identical regression curves for the Dirichlet (black dashed line) and EFD (black solid line) models (see scatterplots in Figure 1). The DirReg model provides a precise estimate also for the precision parameter α^+ , while the EFDReg model slightly overestimates it. Looking at the additional parameters of the EFDReg model, we can observe that the adaptation to Dirichlet distributed data is achieved thanks to equally weighted (estimated p_j equal to approx 0.3 for $j = 1, 2, 3$) and close component means (small estimated distances \tilde{w}_j between components). Graphically, the regression curves referred to component means of the EFDReg model (colored solid lines) are close together and with similar distances.

In fitting study (ii) the EFDReg model well adapts to data and provides precise estimates with low MSEs and SEs for all the parameters (see Table 4). On the con-

trary, the DirReg model, in trying to adapt to data, estimates a considerably lower precision than the true one, and it fails also to correctly estimates some of the regression parameters. From scatterplots in Figure 2 it emerges that the regression curves for the EFDReg model adapt very well to data (both for the overall mean and for the component means), while they are systematically more flat for the DirReg model.

The estimates of the unknown parameters in the three scenarios with outliers are shown in Table 3. Moreover, the regression curves of the Dirichlet and EFD models are plotted on the scatterplots in Figures 3, 4, and 5 referred to scenarios (I), (II), and (III), respectively. The estimates of the regression parameters of the Dirichlet and EFD models are affected by the presence of outliers. The element of flexibility used by the DirReg model in order to adapt to data that depart from the Dirichlet distribution is given by the precision parameter, that is systematically underestimated in all scenarios with outliers. Conversely, the EFDReg model can take advantage of its special mixture structure to better adapt to data. It is worth noting that in all scenarios with outliers, one component of the mixture is dedicated to the group of perturbed values as indicated by the corresponding p_j estimate which is around 0.1. The remaining two components equally describe the remaining majority of unperturbed data with estimates of p_j 's between 0.3 and 0.5. The analysis of the regression curves allows to better understand the different behavior of the DirReg and EFDReg models. The regression curves of the DirReg model are slightly shifted with respect to the regression curves of the DirReg in the scenario without perturbation (dotted lines in Figures 3-5) in the direction of the perturbed values. Instead, looking at the component means of the EFD we note that the first, second, and third components of the mixture are entirely dedicated to model the subgroup of outliers in scenarios (I), (II), and (III).

Scenario	Outliers (I)		Outliers (II)		Outliers (III)	
	Dir	EFD	Dir	EFD	Dir	EFD
$\beta_{10} = 1$	1.134 (0.019)	1.183 (0.036)	0.914 (0.008)	0.992 (0.001)	0.699 (0.092)	0.682 (0.103)
$\beta_{11} = 2$	1.840 (0.065)	1.748 (0.081)	1.871 (0.030)	1.924 (0.022)	1.830 (0.075)	1.962 (0.014)
$\beta_{20} = 0.5$	0.472 (0.002)	0.502 (0.002)	0.683 (0.035)	0.898 (0.180)	0.262 (0.058)	0.192 (0.097)
$\beta_{21} = -3$	-2.693 (0.112)	-2.895 (0.035)	-2.728 (0.123)	-2.349 (0.462)	-2.658 (0.165)	-2.929 (0.020)
$\alpha^+ = 50$	14.930 (1.177)	37.754 (7.304)	15.012 (1.205)	35.590 (4.526)	13.328 (0.940)	41.561 (4.424)
p_1	—	0.121 (0.170)	—	0.553 (0.282)	—	0.479 (0.228)
p_2	—	0.372 (0.300)	—	0.147 (0.086)	—	0.434 (0.229)
p_3	—	0.507 (0.309)	—	0.300 (0.281)	—	0.087 (0.008)
\tilde{w}_1	—	0.744 (0.054)	—	0.281 (0.084)	—	0.155 (0.050)
\tilde{w}_2	—	0.265 (0.090)	—	0.706 (0.057)	—	0.153 (0.052)
\tilde{w}_3	—	0.272 (0.101)	—	0.265 (0.101)	—	0.613 (0.030)

Table 3: Posterior means for the Dirichlet and EFD regression models in scenarios (I), (II), and (III) with outliers. MSEs for the regression coefficients and SEs for remaining parameters are in parenthesis.

Results concerning the presence of some latent groups in data are shown in the last four columns of Table 2. The estimates of regression parameters are biased for both models. Once again, the DirReg model tries to adapt to data by estimating a very low value for the precision parameter, nevertheless this results in a very poor fit. The

regression curves of the DirReg model, reported in Figures 6 and 7, severely miss the data cloud, particularly in scenario (a). The EFDReg model has a satisfactory behavior in scenario (a) where the latent covariate has two categories with probabilities of 0.3 and 0.7. These latent clusters are grasped by the EFDReg model with an estimate equal to 0.64 and 0.353 for the mixing proportions p_1 and p_2 of the first and second component, and an estimate close to zero for p_3 . This is clearly reflected by the regression curves of the component means of the EFD model plotted in Figures 6 and 7. It is worth noting that the orange and blue lines λ_1 and λ_2 perfectly fit the two data clouds. On the contrary, the green line λ_3 has a very poor fit, but this does not affect the overall fit of the model since the third component of the mixture has a probability of occurrence around zero. Scenario (b) is more challenging for the EFDReg model. Please remind that this scenario assumes the existence of a latent covariate having three categories with probabilities of 0.3, 0.15, and 0.55. Nevertheless, the EFD model is able to capture only two out of the three latent clusters, as witnessed by the estimate of the third mixing proportion p_3 which is close to zero. A look at the regression curves of the component means of the EFD model (Figure 7) better explains this behavior. The first scatterplot, referred to the first element of the response, shows a good fit of the orange curve λ_1 . The remaining two curves λ_2 and λ_3 are unable to describe the two visible clusters of data since they are placed in the middle. In the second scatterplot, referred to second elements of the response, the blue curve well adapts to one cluster, the green and orange ones are almost overlapping and well fit a second cluster while a third cluster of data is missed by all curves. In the third scatterplot, referred to the third element of the response, the blue and orange curves cross the data cloud, but the green one completely misses it. Overall, the EFDReg model has an excessively rigid mixture structure to well adapt to this scenario, whilst remaining a far better model than the Dirichlet one.

Scenario Model	Fitting study (ii)		Scenario Model	Generic Mixture	
	Dir	EFD		Dir	EFD
$\beta_{10} = 1$	1.087 (0.015)	1.014 (0.010)	$\beta_{10} = 1$	1.006 (0.010)	0.947 (0.010)
$\beta_{11} = 2$	1.990 (0.069)	1.999 (0.012)	$\beta_{11} = 2$	2.063 (0.149)	1.967 (0.083)
$\beta_{20} = 0.5$	0.752 (0.068)	0.511 (0.010)	$\beta_{20} = 0.5$	0.501 (0.014)	0.457 (0.014)
$\beta_{21} = -3$	-2.409 (0.395)	-3.009 (0.014)	$\beta_{21} = -3$	-2.967 (0.189)	-2.866 (0.159)
$\alpha^+ = 50$	6.444 (0.306)	50.153 (4.253)	α^+	5.619 (0.892)	6.877 (1.684)
$p_1 = 1/3$	—	0.335 (0.024)	p_1	—	0.149 (0.239)
$p_2 = 1/3$	—	0.335 (0.034)	p_2	—	0.189 (0.293)
$p_3 = 1/3$	—	0.331 (0.035)	p_3	—	0.662 (0.353)
$\tilde{w}_1 = 0.6$	—	0.601 (0.016)	\tilde{w}_1	—	0.563 (0.230)
$\tilde{w}_2 = 0.2$	—	0.199 (0.032)	\tilde{w}_2	—	0.553 (0.229)
$\tilde{w}_3 = 0.7$	—	0.694 (0.029)	\tilde{w}_3	—	0.732 (0.153)

Table 4: Posterior means for the Dirichlet and EFD regression models in fitting study (ii) and in case of a generic mixture of Dirichlet. MSEs for the regression coefficients and SEs for remaining parameters are in parenthesis.

The last two columns of Table 4 show the estimates in case observations come from a generic mixture of two Dirichlet distributions. It is worth recalling that this

scenario assumes that the second mixture component follows the same Dirichlet distribution as in scenario (i), and the first component differs from the second one because of the presence of a lower precision parameter. Both the DirReg and EFDReg models provide reasonably unbiased estimates of the regression parameters, despite the MSEs being greater than the ones in scenario (i). To confirm this, the regression curves (dashed and solid lines in Figure 8) well adapt to the majority of observations and are almost overlapping. The presence of a group of data, around 30%, coming from the Dirichlet distribution with a lower precision parameter forces the DirReg model to provide a low estimate of the precision parameter in trying to adapt to data. The EFDReg performs better than the DirReg model since it is capable to recognize the presence of some clusters in data. In particular, it dedicates the third component to describe the majority of data, indeed the estimate of p_3 is approximately equal to 0.7. Instead, the first and second components are dedicated to data coming from the second component of the generic mixture, and they show similar estimates of all parameters (p_j and \tilde{w}_j). In this regard, the green curve is near the solid one, particularly in the scatterplots referred to the first and second elements of the composition. Differently, the blue and orange curves fit the values from the second component of the generic mixture, and they are placed either upward or downward with respect to the majority of points in the scatterplots.

References

1. Aitchison, J.: *The Statistical Analysis of Compositional Data*. The Blackburn Press, London (2003)
2. Albert, J.: Bayesian computation with R. Springer Science. In *ASA Proceedings of Section on Statistical Graphics* (1987)
3. Campbell, G., Mosimann, J. E.: *Multivariate analysis of size and shape: modelling with the Dirichlet distribution & Business Media* (2009)
4. Frühwirth-Schnatter, S.: *Finite mixture and Markov switching models*. Springer Science & Business Media (2006)
5. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B.: *Bayesian Data Analysis*, 3rd edn. CRC Press, London (2013)
6. Hijazi, R.H., Jernigan, R.W.: Modelling compositional data using Dirichlet regression models. *J. Appl. Probab. Statist.* **4**, 77-91 (2009)
7. Maier, M.J.: *Dirichletreg: Dirichlet regression for compositional data in R*. Research Report Series, Department of Statistics and Mathematics, University of Economics and Business, Vienna (2014)
8. McCullagh, P., Nelder, J.: *Generalized linear models*. Chapman & Hall, London (1989)
9. Ongaro, A., Migliorati, S., Ascari, R.: A New Mixture Model On The Simplex. *Statistics and Computing* <https://doi.org/10.1007/s11222-019-09920-x>
10. Neal, R.M.: An Improved Acceptance Procedure for the Hybrid Monte Carlo Algorithm. *J. Comput. Phys.*, **111**(1), 194-203 (1994)
11. Di Brisco, A. M., Ascari, R., Migliorati, S., Ongaro, A.: A new regression model for bounded multivariate responses. In *Smart Statistics for Smart Applications - Book of Short Papers SIS*, 817-822 (2019)
12. Stan Development Team: *Stan Modeling Language Users Guide and Reference Manual* (2016) <http://mc-stan.org/>
13. Vehtari, A., Gelman, A., Gabry, J.: Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, **27**(5), 1413-1432 (2017)

Properties of American-type Options under a Markovian Regime Switching Model

Marko Dimitrov¹, Lu Jin², and Ying Ni¹

¹ Division of Applied Mathematics, Mälardalen University, Västerås, Sweden
(E-mails: marko.dimitrov@mdh.se; ying.ni@mdh.se)

² Department of Informatics, University of Electro-communications, Tokyo, Japan
(E-mail: jinlu@inf.uec.ac.jp)

Abstract. In this paper, a model under which the underlying asset follows a Markov regime-switching process is considered. The underlying economy is partially observable in a form of a signal stochastically related to the actual state of the economy. The American option pricing problem is formulated using a Partially Observable Markov Decision Process (POMDP). We review our previous research on the analytical structural properties of American option prices and optimal strategies. We present also results of numerical experimental studies under some deviations from our sufficient conditions.

Keywords: Hidden Markov Chain, Optimal Strategy, Partially Observable Markov Decision Process, Totally Positive of Order 2.

1 Introduction

An American call (put) option is a financial contract that gives its holder the right but not obligation to buy (sell) a so-called underlying asset for a predetermined strike price K on or before a maturity time T . Note that American options can be exercised any time before or at the maturity, which makes the pricing problem of an American option difficult. Knowing some analytical properties of the American option price and the structural properties of the optimal exercising strategies is very useful for the pricing problem.

Many researchers studied the analytical properties of the optimal early exercise and hold regions, as well as the optimal exercise boundary for models with a one-dimensional underlying process. Kim and Byun [3] considered the simplest binomial tree model with constant volatility and derived the properties of the optimal exercise boundary. In their later paper, Kim *et al.* [4] authors considered deterministic volatility. Broadie and Detemple [5] has derived properties of the optimal exercising boundary when the underlying process follows a geometric Brownian motion. Sato and Sawaki [6] assumed that the economic situation is observable, and later extended the model in Sato and Sawaki [7] where the economic situation is partially observable.

This present paper is a continuation of the authors' earlier research in Jin *et al.* [1]. Refer to Jin *et al.* [1] and references therein for more related work.

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



In Jin *et al.* [1], the underlying asset price process follows a discrete-time partially observable Markovian regime-switching process. Some parameter of the asset price relative is governed by a hidden Markov chain which describes the evolution of economic conditions. In Jin *et al.* [1] the American option pricing problem is formulated using a partially observable Markov decision process. Sufficient conditions for a monotonic one-threshold exercising policy with respect to the asset prices and probability matrices for the economic conditions are derived.

In this work, numerical experimental studies on the structural properties of early exercising strategies for American options are conducted. In particular, the interest is to see how the structural properties are affected under some violation of sufficient conditions. Results obtained here are useful for future practical uses of this pricing model, as model parameters calibrated from the market data do not always satisfy ideal conditions.

This paper proceeds as follows. The Markovian regime-switching model is presented in Section 2 and the American option pricing problem is formulated in Section 3. In Section 4, a summary without proofs of some analytical results from our previous research Jin *et al.* [1] is given. In the end, in Section 5, results of numerical experimental studies are presented.

2 Partially Observable Markovian Regime-Switching Model

Consider an American option with a strike price K and a maturity time T . The payoff function for call (put) is given by $v^e(s) = \max\{s - K, 0\}$ ($v^e(s) = \max\{K - s, 0\}$) with domain $s \in [0, \infty)$.

Make a grid on $[0, T]$ using $M + 1$ grid points $\{0, 1h, \dots, th, \dots, Mh = T\}$. Here, M is an integer and h is the length of period. Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \in \mathbb{M}}, Q)$, $\mathbb{M} = \{0, 1, \dots, M\}$ be a complete filtered probability space where the probability measure Q is the market-chosen risk-neutral probability measure which we take as given. All expectations are taken with respect to Q hereafter, all stochastic processes and probability distributions are defined in the above probability space.

In the formulated model, the asset price process with a deterministic initial value S_0 , defined on times epochs $1, 2, \dots, t, \dots, M$, is given by

$$S_t = S_{t-1}X_t, \quad t = 1, \dots, M,$$

where the *price relative* X_t depends on a variable economic situation Z at time t .

Next, suppose that economy Z_t takes a value from a finite state space $\mathbb{Z} = \{1, 2, \dots, n\}$. The economy states are ordered in ascending order with 1 being the worst economic situation and n being the best. Let the changing of economic situation Z be defined by known a Transition Probability Matrix (TPM) $\mathbf{P} = [p_{ij}]_{i,j \in \mathbb{Z}}$ where p_{ij} refers to the probability that the economic situation transits from level i to level j .

At each period, consider a random variable Y that provides incomplete information related to the real economic situation Z . An observation of Y comes from a finite set $\mathbb{Y} = \{1, 2, \dots, m\}$. Let $\mathbf{\Gamma} = [\gamma_{j\theta}]_{j \in \mathbb{Z}, \theta \in \mathbb{Y}}$ be a Conditional Probability Matrix (CPM) that describes the relationships between the economic situation and the observations. Here, $\gamma_{j\theta} = P(Y = \theta | Z = j)$ is the element of $\mathbf{\Gamma}$ in j -th row and θ -th column.

Let $\boldsymbol{\pi} = (\pi_1, \dots, \pi_n)$ be a probability vector, called the *economy information vector*, that expresses the information about the economic situation. Here,

$$\pi_j = P(Z = j), \quad j = 1, 2, \dots, n, \quad \sum_{j=1}^n \pi_j = 1.$$

At any period, the pair $(s, \boldsymbol{\pi})$ is called a process state, meaning that the current asset price is s and the economy information vector is $\boldsymbol{\pi}$.

3 American Option pricing

It is well known that it is never optimal to early exercise an American call option on a non-dividend-paying underlying asset. Therefore, consider dividend-paying underlying asset with a continuous compounded dividend yield, denoted by δ , throughout the paper.

At each time epoch t , the option holder can choose to early exercise or hold the option. If the holder decides to early exercise, a payoff of $v^e(s)$ is received where s is the underlying asset price at time t .

Assume that a holder decides to hold the option, then the information vector at the beginning of the next period is updated to $\mathbf{T}(\boldsymbol{\pi}, \theta)$, given the observation θ with probability $\psi(\theta | \boldsymbol{\pi})$. The probability $\psi(\theta | \boldsymbol{\pi})$ is given by

$$\psi(\theta | \boldsymbol{\pi}) = \sum_{j=1}^n \sum_{i=1}^n \pi_i p_{ij} \gamma_{j\theta},$$

and the j -th element of the updated information vector $\mathbf{T}(\boldsymbol{\pi}, \theta)$ is

$$T_j(\boldsymbol{\pi}, \theta) = \frac{\sum_{i=1}^n \pi_i p_{ij} \gamma_{j\theta}}{\psi(\theta | \boldsymbol{\pi})}.$$

In the next step, formulate the optimal stopping problem using a partially observable Markov decision process.

Denote N be the remaining periods to maturity, for example $N = M = T/h$ at the beginning of option transaction and $N = 0$ at maturity.

Let $r > 0$ be the continuous compounded risk-free interest rate. Assume that X_j is distributed as the well-known risk-neutral distribution in a binomial tree,

$$P(X_j = x_j) = \begin{cases} q_j, & x_j = u_j \\ 1 - q_j, & x_j = d_j \end{cases}$$

where

$$q_j = \frac{e^{(r-\delta)h} - d_j}{u_j - d_j}, \quad u_j = e^{\sigma_j \sqrt{h}}, \quad d_j = \frac{1}{u_j},$$

and $\sigma_j > \sigma_{j'}$ for $j < j'$. The arbitrage opportunities are excluded when

$$d_j < e^{(r-\delta)\sqrt{h}} < u_j, \quad j \in \mathbb{Z}.$$

Note that the economy is ordered according to the volatility σ_i , where higher volatility is linked to a worse economy and lower volatility to a better economy. Indeed, low volatility indicates usually a stable market, on the other hand, most of the assets show very high volatility under a bad economy.

Consider an American option with the current process state $(s, \boldsymbol{\pi})$ and remaining periods to maturity N . The option price $v_N(s, \boldsymbol{\pi})$, is given by

$$v_N(s, \boldsymbol{\pi}) = \max \left\{ \begin{array}{l} \max\{K - s, 0\} = v_N^e(s) \\ \beta \sum_{\theta=1}^m \psi(\theta|\boldsymbol{\pi}) \sum_{k=1}^2 v_{N-1}[sx_j^k, \mathbf{T}(\boldsymbol{\pi}, \theta)] P(x_j^k) = v_N^h(s, \boldsymbol{\pi}) \end{array} \right.$$

where $x_j^1 = u_j$, $x_j^2 = d_j$, and $\beta = e^{-rh}$ ($0 < \beta < 1$) is the discount factor.

The quantity denoted by $v_N^e(s, \boldsymbol{\pi})$ is the payoff when the holder exercises the option at the beginning of the current period, and denoted by $v_N^h(s, \boldsymbol{\pi})$ is the value when the holder decides to hold and follow the optimal strategy in the remaining periods. As the payoff of early exercise does not depend on the remaining periods, notation $v^e(s)$ is used instead of $v_N^e(s)$.

The hold value $v_0^h(s, \boldsymbol{\pi})$ is zero at the maturity. So the option value at the maturity is simply the payoff function, i.e.

$$v_0(s, \boldsymbol{\pi}) = \max\{v^e(s), v_0^h(s, \boldsymbol{\pi})\} = v^e(s).$$

4 Analytical Structural Properties of American Options

First review the following definitions of totally positive property of order 2 (Karlin [2]).

Definition 1. If for two vectors $\mathbf{x} = (x_1, x_2, \dots, x_n)$, and $\mathbf{y} = (y_1, y_2, \dots, y_n)$

$$\begin{vmatrix} x_i & x_j \\ y_i & y_j \end{vmatrix} \geq 0, \quad 1 \leq i < j \leq n,$$

holds, it is said that \mathbf{y} dominates \mathbf{x} in the sense of totally positive ordering of order 2, denoted by $\mathbf{x} \stackrel{\text{TP}_2}{\prec} \mathbf{y}$.

Definition 2. Let $\mathbf{X} = [x_{ij}]_{ij}$ be an $n \times m$ matrix for which

$$\det(\mathbf{B}) \geq 0$$

for every submatrix $\mathbf{B} = [x_{i_k j_l}]_{kl}$ of dimensions 2×2 where $1 \leq i_1 < i_2 \leq n$, $1 \leq j_1 < j_2 \leq m$. Matrix \mathbf{X} is said to have a property of totally positive of order two, denoted by $\mathbf{X} \in \text{TP}_2$.

To obtain structural properties of American options, impose the following assumptions.

- (A-1) Transition probability matrix \mathbf{P} , corresponding to economic situation, has a TP_2 property.
- (A-2) Conditional probability matrix $\mathbf{\Gamma}$, corresponding to signal, has a TP_2 property.

The TP_2 property of TPM implies that a better economy in a period moves to a more progressive situation in the next period. The TP_2 property of CPM implies that a better economic situation gives rise to higher output levels for the observations probabilistically.

In the authors' previous research Jin *et al.* [1], analytical structural properties of American options were derived using the sufficient conditions given by assumptions (A-1) and (A-2). Summary of some results that are important for the experimental studies in this paper follow without proof.

Proposition 1. *For a put (call) American option, if assumptions (A-1) and (A-2) hold, then hold value of the option $v_N^h(s, \boldsymbol{\pi})$ is*

1. decreasing (increasing) in asset price s for every N and $\boldsymbol{\pi}$;
2. increasing in remaining period N for every s and $\boldsymbol{\pi}$;
3. decreasing in information vector $\boldsymbol{\pi}$ in the sense of TP_2 for every s and N .

Proposition 1 establish the monotonicity of $v_N^h(s, \boldsymbol{\pi})$ in N, s , and $\boldsymbol{\pi}$, that is, to be monotone in remaining time to maturity, asset price and the economic situation.

Proposition 2. *For an American put (call) option, (i) $v_N^h(s, \boldsymbol{\pi})$ is a convex function of s , (ii) the decreasing (increasing) rate of $v_N^h(s, \boldsymbol{\pi})$ in s is less than 1 for any $\boldsymbol{\pi}$ under the assumptions (A-1) and (A-2).*

Proposition 3. *For an American put or call option, the difference between $v_N^h(s, \boldsymbol{\pi})$ and $v^e(s)$ is increasing in N for any s and $\boldsymbol{\pi}$ under the assumptions (A-1) and (A-2).*

Proposition 4. *For an American put or call option, the difference between $v_N^h(s, \boldsymbol{\pi})$ and $v^e(s)$ is decreasing in $\boldsymbol{\pi}$ in the sense of TP_2 ordering for any N and s under the assumptions (A-1) and (A-2).*

Propositions 2, 3 and 4 provide some properties of the relationship between $v_N^h(s, \boldsymbol{\pi})$ and $v^e(s)$, and these properties are important for the following discussion on the thresholds for the following two regions.

The early exercising region and holding region for every remaining periods to maturity N are defined by

- Exercise region

$$\begin{aligned} D_N^e &= \{(s, \boldsymbol{\pi}) \mid v_N^h(s, \boldsymbol{\pi}) < v^e(s)\} \\ &= \{(s, \boldsymbol{\pi}) \mid v_N(s, \boldsymbol{\pi}) = v^e(s)\} \end{aligned}$$

- Hold region

$$\begin{aligned} D_N^h &= \{(s, \boldsymbol{\pi}) \mid v_N^h(s, \boldsymbol{\pi}) > v^e(s)\} \\ &= \{(s, \boldsymbol{\pi}) \mid v_N(s, \boldsymbol{\pi}) = v_N^h(s, \boldsymbol{\pi})\} \end{aligned}$$

Next, investigate the thresholds in both $\boldsymbol{\pi}$ and s for the above two regions. To discuss the threshold in $\boldsymbol{\pi}$, one needs to define the set of all TP_2 -ordered sets of economy information vectors, denoted by Π , as follows

$$\Pi = \bigcup \{\boldsymbol{\pi}^i, i \in \mathcal{J} : \boldsymbol{\pi}^k \stackrel{TP_2}{\prec} \boldsymbol{\pi}^l \text{ for } k \leq l, k, l \in \mathcal{J}\},$$

where \mathcal{J} refers to the enumeration of information vectors in each TP_2 -ordered set. There are infinitely many elements, i.e. TP_2 -ordered sets, in Π . The set Π , which is the union of all such ordered-sets, matches the space of all economy information vectors $\{(\pi_1, \dots, \pi_n) : \sum_{i=1}^n \pi_i = 1\}$.

Notation TP_2^* is used to denote an arbitrary TP_2 -ordered set in Π .

Proposition 5. *For an American put or call option, there exists at most one threshold $\boldsymbol{\pi}_N(s)$ for any s and N which separates an ordered set TP_2^* into two regions: an (early) exercise for any $\boldsymbol{\pi}$ ($\in TP_2^*$) less than $\boldsymbol{\pi}_N(s)$, and a hold region otherwise. Moreover, $\boldsymbol{\pi}_N(s') \stackrel{TP_2}{\prec} \boldsymbol{\pi}_N(s)$ for $s < s'$, and $\boldsymbol{\pi}_{N^1}(s) \stackrel{TP_2}{\prec} \boldsymbol{\pi}_{N^2}(s)$ for $N^1 < N^2$.*

Proposition 5 focuses on economy situation and presents a property of the threshold in $\boldsymbol{\pi}$ ($\in TP_2^*$). Figures 1 and 2 are used to illustrate the thresholds for different s and N .

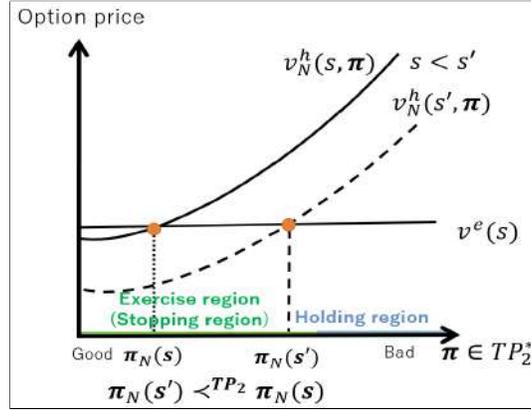


Fig. 1. Threshold $\boldsymbol{\pi}_N(s)$ with different asset price s for the case of an American put option

Next, focus on the asset price s and obtain a similar property of the threshold in s as given in Proposition 6.

Proposition 6. *For an American put or call option, there exists at most one threshold $s_N(\boldsymbol{\pi})$ for any $\boldsymbol{\pi}$ and N which separates the space of s into two regions:*

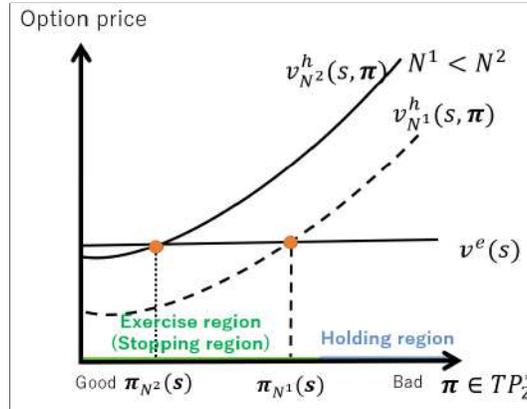


Fig. 2. Threshold $\pi_N(s)$ with different remaining periods N for the case of an American put option

(early) exercise region ($s < s_N(\pi)$) and holding region ($s > s_N(\pi)$). Moreover, $s_N(\pi^1) \leq s_N(\pi^2)$ for $\pi^1 \prec^{TP_2^*} \pi^2$, and $s_{N^1}(\pi) \geq s_{N^2}(\pi)$ for $N^1 < N^2$.

From the above propositions 5 and 6, we know that the information space (s, π) for any $\pi (\in TP_2^*)$ is divided into at most two regions, D_N^e and D_N^h , and the area of holding region D_N^h increases with the remaining periods N as shown in Figure 3. This means it is preferable to hold the option if more time periods remain.

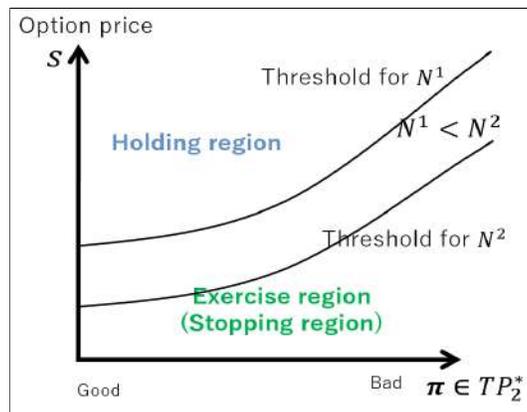


Fig. 3. Threshold on space $\pi_N(s)$ with different remaining periods N for the case of an American put option

5 Numerical Experimental Studies

In this section, results of experimental studies are presented. The object of interest is the behavior of the analytical properties of optimal exercising strategies under some deviations from the strict assumptions (A-1) and (A-2).

5.1 Three-State Model

Assume that the information vector consists of three pieces of information. Hence, consider a three-state model. For the details of the model implementation refer to the previous research Jin *et al.* [1].

The thresholds in the numerical examples are obtained as follows. Let S be a set of initial asset prices S_0 . First, take a finite subset $S^* \subseteq S$, and an order-set $\text{TP}_2^* \in \Pi$ for $n = 3$. Then, for a fixed economy information vector $\boldsymbol{\pi} \in \text{TP}_2^*$ compute the option price for every $S_0 \in S^*$. Initial asset price $s^* \in S^*$ is the one-threshold that splits the set S^* into a (early) exercise region D_N^e and a hold region D_N^h .

5.2 TP_2^* -ordered set of economy information vectors

To determine a TP_2 ordered set of economy information vectors, that is a set $\text{TP}_2^* \in \Pi$ for $n = 3$, the following proposition was used.

Proposition 7. *Let $\boldsymbol{\pi}_1 = (p_1, p_2, 1 - p_1 - p_2)$ and $\boldsymbol{\pi}_2 = (q_1, q_2, 1 - q_1 - q_2)$ such that $\boldsymbol{\pi}_1 \neq \boldsymbol{\pi}_2$. If $p_1 \neq 0$ and $p_1 = q_1$, then $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ are not TP_2 comparable. If $p_1 \geq q_1$ and $p_2 = q_2$, then $\boldsymbol{\pi}_1 \stackrel{\text{TP}_2}{\prec} \boldsymbol{\pi}_2$.*

Proof. Let $p_1 = q_1$,

$$\begin{aligned}\boldsymbol{\pi}_1 &= (p_1, p_2, 1 - p_1 - p_2), \\ \boldsymbol{\pi}_2 &= (p_1, q_2, 1 - p_1 - q_2).\end{aligned}$$

Assume opposite that $\boldsymbol{\pi}_1 \stackrel{\text{TP}_2}{\prec} \boldsymbol{\pi}_2$. Then $q_2 \geq p_2$ from

$$\begin{vmatrix} p_1 & p_2 \\ p_1 & q_2 \end{vmatrix} \geq 0.$$

However $q_2 \geq p_2$ leads to

$$\begin{vmatrix} p_1 & 1 - p_1 - p_2 \\ p_1 & 1 - p_1 - q_2 \end{vmatrix} \leq 0,$$

Therefore, $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ are not TP_2 comparable if $p_1 = q_1$. Assume that $p_2 = q_2$, by the definition of totally positive of order two, it follows that $\boldsymbol{\pi}_1 \stackrel{\text{TP}_2}{\prec} \boldsymbol{\pi}_2$ because $p_1 \geq q_1$.

Name	Notation	Parameters
Maturity time	T	8/252
Number of steps	M	4
Time duration of a step	h	2/252
Volatility vector	$\boldsymbol{\sigma}$	(0.5, 0.3, 0.1)
Strike price	K	100
Interest rate	r	0.02
Dividend yield (American call)	δ	0.1
TPM	\mathbf{P}	$[p_{ij}]_{i,j=1,2,3}$
CPM	$\mathbf{\Gamma}$	$[\gamma_{ij}]_{i,j=1,2,3}$

Table 1. General model test parameters

5.3 Numerical Results

In this subsection, unless it is said otherwise, the parameters used for computation are given in Table 1.

The choice of parameters has to satisfy assumptions (A-1) and (A-2). Hence, both the TPM and CPM matrices should have the property of TP_2 .

The experimental study in this paper will be concentrated on the threshold, that is the early exercise and hold regions. Other results may be found in Jin *et al.* [1]. To show the early exercise and hold regions, as well as the monotonicity of threshold in the information vectors $\boldsymbol{\pi}$, a set $\text{TP}_2^* \in \mathcal{I}$ and a set S^* of initial asset prices are required. Denote the following set with TP_2^* ,

$$\begin{aligned}\pi_2 &= 0.05, \\ \pi_1 &= \pi_2 + 0.03 \times i, \quad i = 0, 1, \dots, 30, \\ \pi_3 &= 1 - \pi_1 - \pi_2.\end{aligned}$$

As explained in Section 5.2, this is a set of TP_2 ordered information vectors. Thus, $\text{TP}_2^* \in \mathcal{I}$. Next, the set of initial asset prices used for the experimental studies is

$$S_p^* = \left\{ \left(0.7 + \frac{0.3}{5000} \times i \right) \times K : i \in \{0, 1, \dots, 5000\} \right\}$$

for an American put, and

$$S_c^* = \left\{ \left(1 + \frac{0.3}{5000} \times i \right) \times K : i \in \{0, 1, \dots, 5000\} \right\}$$

for an American call option with dividend yield $\delta = 0.1$ were used. Let TPM and CPM be

$$\mathbf{P} = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.1 & 0.4 & 0.5 \\ 0.05 & 0.25 & 0.7 \end{bmatrix}, \quad \mathbf{\Gamma} = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.1 & 0.4 & 0.5 \\ 0.05 & 0.4 & 0.55 \end{bmatrix}.$$

respectively. This choice of matrices satisfies TP_2 property, $\mathbf{P}, \mathbf{\Gamma} \in \text{TP}_2$.

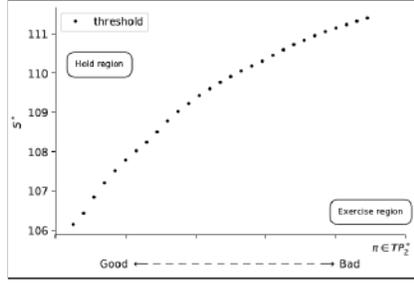


Fig. 4. An example of the optimal stopping regions for an American call option with dividend yield and the monotonicity of the threshold in π with parameters π given in Table 1, TP_2^* and S_c^*

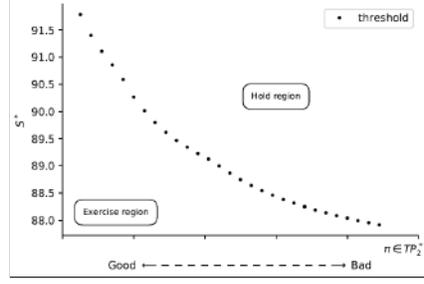


Fig. 5. An example of the optimal stopping regions for an American put option with dividend yield and the monotonicity of the threshold in π with parameters π given in Table 1, TP_2^* and S_p^*

Figures 4 and 5 show that the threshold is decreasing and increasing in π for $N = 4$, as well as the exercise and hold regions for both American call option with dividend yield and American put, respectively.

Consider an American put option in the following until the end. To see the model behavior when \mathbf{P} , \mathbf{I} or both do not have TP_2 property, a various sets $TP_2^i \in \Pi, i = 1, 2, \dots, 10000$ were used. To do such Monte Carlo simulation, sets $TP_2^i \in \Pi, i = 1, 2, \dots, 10000$ were randomly generated, as well as \mathbf{P} , \mathbf{I} or both. The following three cases were considered

1. both TPM and CPM are randomly generated;
2. CPM is randomly generated;
3. TPM is randomly generated.

In all three cases sets of TP_2 ordered vectors were randomly generated according to the property explained in Section 5.2.

Case 1: TPM and CPM are randomly generated matrices that satisfy conditions of the transition probability matrix of a Markov chain. As expected, varying both matrices, TPM and CPM, the claim that the threshold is monotonically increasing does not stand. Figure 6 shows that the threshold is decreasing in information vector π under assumptions of Case 1, on the other hand, Figure 7 shows that the threshold is increasing in information vector π under assumptions of Case 1.

Case 2: In this case, TPM \mathbf{P} is fixed and given by

$$\mathbf{P} = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.1 & 0.4 & 0.5 \\ 0.05 & 0.25 & 0.7 \end{bmatrix}$$

and CPM \mathbf{I} is a randomly generated matrix that satisfies conditions of the transition probability matrix of a Markov chain. Based on the simulation, deviation from the sufficient assumption $\mathbf{I} \in TP_2$ does not affect the monotonicity of the threshold as much as the deviation from the sufficient conditions presented in Case 1. Figures 8 and 9 show that the threshold is increasing in the

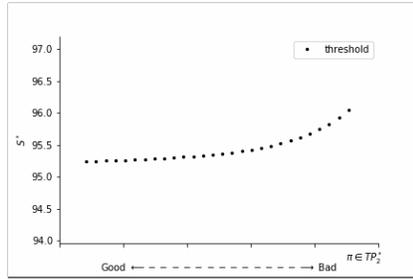


Fig. 6. An example of Case 1, where matrices TPM and CPM were randomly generated and the threshold is decreasing in information vector π

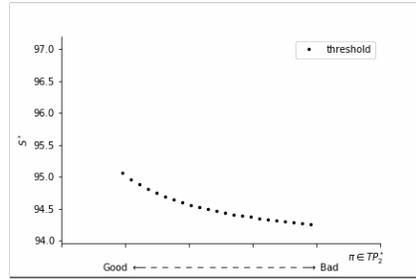


Fig. 7. An example of Case 1, where matrices TPM and CPM were randomly generated and the threshold is increasing in information vector π

information vector π under assumptions of Case 2, that is, CPM is a randomly generated matrix.

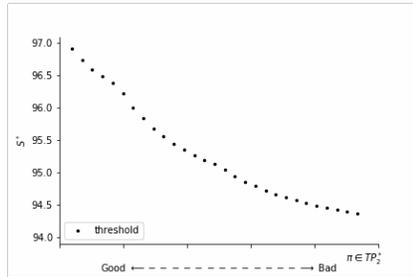


Fig. 8. An example of Case 2, where matrix CPM was randomly generated and the threshold is increasing in information vector π

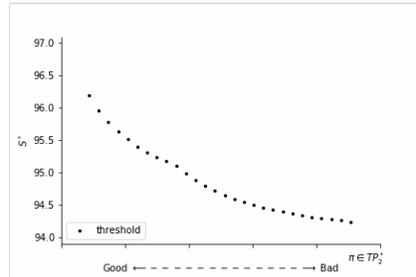


Fig. 9. Another example of Case 2, where matrix CPM was randomly generated and the threshold is increasing in information vector π

Case 3: Assume now that CPM Γ is fixed and given by

$$\Gamma = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.1 & 0.4 & 0.5 \\ 0.05 & 0.4 & 0.55 \end{bmatrix}$$

and TPM \mathbf{P} is randomly generated matrix that satisfies conditions of transition probability matrix of a Markov chain. Violation of assumption $\mathbf{P} \in TP_2$ influence the results obtained for the threshold after a couple of simulations. It is clear that the condition $\mathbf{P} \in TP_2$ affects the monotonicity of threshold more than the condition $\Gamma \in TP_2$. Figure 6 shows that the threshold is increasing in information vector π under assumptions given in Case 3, and Figure 7 shows that the threshold is decreasing in information vector π under assumptions of Case 3.

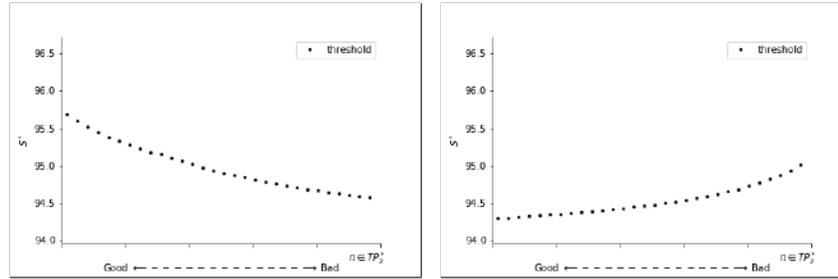


Fig. 10. An example of Case 3, where matrix TPM was randomly generated and the threshold is increasing in information vector π

Fig. 11. An example of Case 3, where matrix TPM was randomly generated and the threshold is decreasing in information vector π

6 Conclusion and future research

Review of previous research and new experimental results of the American option pricing and the corresponding optimal exercising strategies under a novel model were presented. Under this model, the asset price follows an extended binomial tree with the volatility parameter governed by a discrete-time hidden Markov chain.

The numerical results indicate that violation of sufficient conditions affect the structural properties of American put and call options with dividend yield. That is, the TP_2 property of the transition probability matrix for the economic situations is important for having the monotonically decreasing (increasing) property of the exercising threshold. One of the assumptions may be relaxed in practice as it does not affect the monotonicity to great extent. Specifically, the TP_2 property of CPM may be ignored as it has little effect on the monotonicity.

For future research, the model shall be generalized by permitting a more general probability distribution for the asset price dynamics. As the results of this research are limited to the pricing of short-maturity options future research may contain extensions of the model for options with longer maturities.

References

1. L. Jin, M. Dimitrov and Y. Ni. *Valuation and Optimal Strategies for American Options under a Markovian Regime-Switching Model*, In: Rancic M., Malyarenko A. Silvestrov S. (Eds.), SPAS conference proceedings 2019, Volume I, Västerås, Sweden, To be published by Springer (submitted).
2. S. Karlin *Total Positivity*, Volume I. Stanford University Press; 1968.
3. I. J. Kim, S. J. Byun. Optimal Exercise Boundary in a Binomial Option Pricing Model, *The Journal of Financial Engineering*, **3**, 2, 137–158, 1994
4. I. J. Kim, S. J. Byun, and S. Lim. Valuing and Hedging American Options under TimeVarying Volatility, *Journal of Derivatives Accounting* **1**, 195-204, 2004
5. M. Broadie, J. Detemple. American options on dividend-paying assets, *Fields Institute Communications* **22**, 69–97, 1999

6. K.i Sato, K. Sawaki. The dynamic pricing for callable securities with Markov-modulated prices, *Journal of the Operations Research Society of Japan*, 57, 87-103, 2014
7. K. Sato, K. Sawaki. The Dynamic Valuation of Callable Contingent Claims With a Partially Observable Regime Switch, *RIMS Kokyuroku*, 1933, 157-167, 2015 (In Japanese)

Age exaggeration ruses: infrequent age overstatement distorts the mortality curve at old age

Dalkhat M. Ediev

(dalkhat@hotmail.com; ediev@iiasa.ac.at)

North-Caucasian State Academy (Russia),

Lomonosov Moscow State University (Russia),

International Institute for Applied Systems Analysis (Austria)

A mathematical model of age exaggeration is suggested that enables studying the statistical manifestations of the age overstatement formally, in simulations, and empirically. Our findings indicate that even an infrequent age exaggeration may substantially distort the mortality curve at an advanced old age. Old-age mortality distortions that were previously attributed to the effects of mortality selection in the heterogeneous population are shown to be produced by the age exaggeration too. The model proposed here may be used in both reconstructing the actual mortality profile and assessing the mode and extent of age misreporting.

Acknowledgements

The work was supported by the Russian Foundation for Basic Research (grant № 18-01-00289 «Mathematical models and methods of correcting the distortions of the age structure and mortality rates of elderly population»).

Introduction.

Age exaggeration is a tough problem creating obstacles when studying longevity and mortality at advanced old age. The models and methods suggested in dealing with the age misreporting problem address the issue by pooling the advanced old age together into an extended open age interval to avoid considering the mortality rate explicitly at advanced old ages subject to the age exaggeration (Horiuchi and Coale, 1982; Mitra, 1984; Ediev, 2018, 2019). That approach, however, does not provide a tool for assessing the extent of age exaggeration. Neither does it offer means to reconstructing the actual age pattern of mortality at advanced old age (see, however, the constrained extrapolation model that deals with the latter issue (Ediev, 2017)).

Here, we attempt an alternative approach by explicitly modeling the age exaggeration and examining the extended age profiles of mortality that may result from the age exaggeration under different scenarios. We consider a model where a given fraction of the elderly exaggerates their age by a given amount. Albeit simple, the model offers valuable insights into possible effects of age exaggeration.



First, we compensate for the lack of empirical data and develop a better intuition for consequences of age exaggeration by using a mathematical model and running simulations based on our model. After examining the patterns of age exaggeration and resulting mortality distortions in simulations, we examine possible age exaggeration empirically. We conclude by discussion.

1. Age exaggeration model and formal results

To examine which sorts of distortions the age exaggeration may bring to the mortality curve, we introduce a simple schematic mathematical model. We assume that a fraction α of people above the age x_0 , randomly chosen, exaggerate their true age by the same amount, namely, by δ years each. Although this assumption simplifies the real-life age misreporting in many ways, the model does reflect the most important aspect of the age exaggeration and has good practical potential. More complex patterns of age exaggeration may be thought of as a mix of the elementary exaggerations we consider here. For brevity, we will call the introduced age exaggeration model as (x_0, α, δ) model.

Under the (x_0, α, δ) model, death rates at ages $x \geq x_0 + \delta$ will be distorted because younger, in case $\delta > 0$, or older, in case $\delta < 0$, people will be erroneously added to the observed population thereby replacing the true death rates by the weighted average of the rates at ages x and $x - \delta$:

$$\tilde{M}(x) = \frac{\tilde{D}(x)}{\tilde{P}(x)} = \frac{(1-\alpha)P(x)M(x) + \alpha P(x-\delta)M(x-\delta)}{(1-\alpha)P(x) + \alpha P(x-\delta)} = M(x) \left(1 - \frac{1 - \frac{M(x-\delta)}{M(x)}}{1 + \frac{1-\alpha}{\alpha} \frac{P(x)}{P(x-\delta)}} \right), \quad x \geq x_0 + \delta, \quad (1)$$

here, $\tilde{M}(x)$ is the observed death rate at age x , $M(x)$ is the true death rate at age x , $P(x)$ is the population of age x , and $D(x) \stackrel{\text{def}}{=} M(x)P(x)$ is the deaths' intensity at age x . Hereinafter, the tilde mark ($\tilde{}$) marks observed (possibly biased due to the age exaggeration effects) quantities as opposed to the true, non-observed, ones that are not marked. If, as typical at old age, mortality increases by age and $\alpha, \delta > 0$, then the observed death rate is an underestimation of the true rate, $\tilde{M}(x) < M(x), x \geq x_0 + \delta$. If, on the contrary, the elderly were to reduce their true age ($\alpha > 0, \delta < 0$) the observed death rates would be overestimated as compared to the true rates: $\tilde{M}(x) > M(x), x \geq x_0 + \delta$.

To examine the patterns of distortions to the mortality curve qualitatively, we use Gompertz (Gompertz, 1825; Thatcher, V. Kannisto and Vaupel, 1998) mortality model that implies, for any two ages x and y :

$$\frac{M(y)}{M(x)} = e^{b(y-x)}, \quad (2)$$

and the stable population model (Preston, Heuveline and Guillot, 2001) where:

$$\frac{P(y)}{P(x)} = e^{-r(y-x)} \frac{l(y)}{l(x)} = e^{-r(y-x)} e^{-\int_x^y M(z) dz} = e^{-r(y-x)} e^{-\frac{1}{b} M(x) (e^{b(y-x)} - 1)}. \quad (3)$$

Here, $l(x)$ is the survival, i.e., the proportion surviving from birth to age 0.

Combining (1)-(3), we get the analytical expression for distortion of the observed death rates:

$$\frac{\tilde{M}(x)}{M(x)} = 1 - \frac{1-e^{-b\delta}}{1+\frac{1-\alpha}{\alpha}e^{-r\delta-\frac{1}{b}M(x)(1-e^{-b\delta})}}, \quad x \geq x_0 + \delta. \quad (4)$$

As it follows from (4), the higher the death rate (i.e., the older the age x), the stronger the proportionate bias of the death rate caused by age exaggeration. Two limit cases of (4) deserve closer attention. First, consider younger ages where the death rate is low as compared to the population growth rate so as $M(x) \ll |r|b\delta$. In that case, (4) reduces to:

$$\frac{\tilde{M}(x)}{M(x)} \approx 1 - \frac{1-e^{-b\delta}}{1+\frac{1-\alpha}{\alpha}e^{-r\delta}}, \quad x \geq x_0 + \delta. \quad (5)$$

That means, *at ages where the growth rate dominates the death rate in shaping the population age structure, proportionate distortions to the observed death rates are age-independent*. If, additionally, $|r\delta| \ll 1$ and $|b\delta| \ll 1$, (5) reduces to simply $\frac{\tilde{M}(x)}{M(x)} \approx 1 - \alpha\delta b \approx e^{-\alpha\delta b}$ and does not depend on the growth rate r . In other words, at young ages and with moderate age exaggeration, the proportionate distortion of the death rates is determined by the amount of increase of the death rate over the mean magnitude ($\alpha\delta$) of the age exaggeration and not on the growth rate.

In the opposite limit case of advanced old ages where $M(x) \gg |r|b\delta$,

$$\frac{\tilde{M}(x)}{M(x)} \approx 1 - \frac{1-e^{-b\delta}}{1+\frac{1-\alpha}{\alpha}e^{-\frac{1}{b}M(x)(1-e^{-b\delta})}}. \quad (6)$$

That is, at an advanced old age with high mortality and with moderate age exaggeration, the observed death rates are also distorted in a growth-parameter-independent way.

The two limit cases suggest that the population growth rate (and therefore, the departure of the population age structure from stationary) might be less important a factor for the distortions of the mortality curve as compared to other factors such as the age exaggeration parameters and the level and the slope of the mortality curve.

Patterns of the distortions in the observed population composition by age fit the following relation:

$$\frac{\tilde{P}(x)}{P(x)} = 1 + \alpha \left(e^{r\delta + \frac{1}{b}M(x)(1-e^{-b\delta})} - 1 \right), \quad x \geq x_0 + \delta. \quad (7)$$

At a young age, population distortions (7) are sensitive to the growth rate, $\frac{\tilde{P}(x)}{P(x)} \approx 1 + \alpha r\delta$, unlike in the death rates' distortions. At advanced old age, the population distortions depend on the death rate rather than the growth parameter but remain α -specific: $\frac{\tilde{P}(x)}{P(x)} \sim 1 + \alpha(e^{M(x)\delta} - 1)$.

Because the distortions to the death rates due to the age exaggeration vary with age, they also affect the pace at which the death rates change. In particular, from (4), the Life-table Ageing Rate (LAR) (Horiuchi and Wilmoth, 1997) is subject to the following alterations:

$$\begin{aligned}
\frac{d \ln \tilde{M}(x)}{dx} &= \frac{d \ln M(x)}{dx} + \frac{d}{dx} \ln \left(1 - \frac{1-e^{-b\delta}}{1 + \frac{1-\alpha}{\alpha} e^{-r\delta - \frac{1}{b} M(x)(1-e^{-b\delta})}} \right) = b + \\
&\frac{1}{1 - \frac{1-e^{-b\delta}}{1 + \frac{1-\alpha}{\alpha} e^{-r\delta - \frac{1}{b} M(x)(1-e^{-b\delta})}}} \frac{1-e^{-b\delta}}{\left(1 + \frac{1-\alpha}{\alpha} e^{-r\delta - \frac{1}{b} M(x)(1-e^{-b\delta})}\right)^2} \frac{1-\alpha}{\alpha} e^{-r\delta - \frac{1}{b} M(x)(1-e^{-b\delta})} \left(-\frac{1}{b}\right) \left(1 - \right. \\
e^{-b\delta}) b M(x) &= b - \frac{(1-e^{-b\delta})^2 \frac{1-\alpha}{\alpha} e^{-r\delta - \frac{1}{b} M(x)(1-e^{-b\delta})}}{\left(e^{-b\delta} + \frac{1-\alpha}{\alpha} e^{-r\delta - \frac{1}{b} M(x)(1-e^{-b\delta})}\right) \left(1 + \frac{1-\alpha}{\alpha} e^{-r\delta - \frac{1}{b} M(x)(1-e^{-b\delta})}\right)} M(x). \quad (8)
\end{aligned}$$

Next, we infer analytically, if the classical relation (3) between the stable population growth, mortality, and age composition may still hold if the population numbers and the death rates are subject to distortions due to the age exaggeration. In the stable population, as it follows from (3), there must be a relation between the age structure, mortality, and growth rate:

$$\frac{d \ln P(x)}{dx} = -r - M(x). \quad (9)$$

When observing the biased age structure (7) and the death rates (4) at ages $x > x_0 + \delta$,

$$\begin{aligned}
\frac{d \ln \tilde{P}(x)}{dx} &= \frac{d \ln P(x)}{dx} + \frac{d \ln \left(1 + \alpha \left(e^{r\delta + \frac{1}{b} M(x)(1-e^{-b\delta})} - 1 \right) \right)}{dx} = -r - M(x) + \\
&\frac{\alpha e^{r\delta + \frac{1}{b} M(x)(1-e^{-b\delta})} \frac{1}{b} \frac{dM(x)}{dx} (1-e^{-b\delta})}{1 + \alpha \left(e^{r\delta + \frac{1}{b} M(x)(1-e^{-b\delta})} - 1 \right)} = -r - M(x) \left(1 - \frac{\alpha e^{r\delta + \frac{1}{b} M(x)(1-e^{-b\delta})} (1-e^{-b\delta})}{1 + \alpha \left(e^{r\delta + \frac{1}{b} M(x)(1-e^{-b\delta})} - 1 \right)} \right) = -r - \\
M(x) \left(1 - \frac{1-e^{-b\delta}}{1 + \frac{1-\alpha}{\alpha} e^{-r\delta - \frac{1}{b} M(x)(1-e^{-b\delta})}} \right) &= -r - \tilde{M}(x). \quad (10)
\end{aligned}$$

In other words, *at ages $x > x_0 + \delta$, the stable-population relation between the observed age composition, death rates, and the growth rate holds despite the distortions caused by the age misstatement.* This, unfortunately, implies it will typically be hard to detect the age exaggeration from observed population numbers and death rates. Note, however, that the stable-population relation between the population growth, age composition, and mortality will be broken at ages at ages x_0 to $x_0 + \delta$ where the death rates will not be affected by age exaggeration while the population size will.

More generally, the usual population balance (Keyfitz and Caswell, 2005):

$$\frac{\partial}{\partial t} P(x, t) + \frac{\partial}{\partial x} P(x, t) = -D(x, t), \quad (11)$$

where the second variable in all functions indicates time, will not be violated by age- and time-independent age misstatements in any (including non-stable) migration-closed population *at ages $x > x_0 + \delta$:*

$$\begin{aligned}
\frac{\partial}{\partial t} \tilde{P}(x, t) + \frac{\partial}{\partial x} \tilde{P}(x, t) &= \frac{\partial}{\partial t} \left((1-\alpha)P(x, t) + \alpha P(x-\delta, t) \right) + \frac{\partial}{\partial x} \left((1-\alpha)P(x, t) + \alpha P(x-\delta, t) \right) \\
&= (1-\alpha) \left(\frac{\partial}{\partial t} P(x, t) + \frac{\partial}{\partial x} P(x, t) \right) + \alpha \left(\frac{\partial}{\partial t} P(x-\delta, t) + \frac{\partial}{\partial x} P(x-\delta, t) \right) = -(1-\alpha)D(x, t) - \alpha D(x-\delta, t) = -\tilde{D}(x, t). \quad (12)
\end{aligned}$$

The population balance will be broken, however, at ages x_0 and $x_0 + \delta$ where population numbers will also be subject to jumps caused by the onset of age exaggeration process.

The last two formal results show that detecting the age exaggeration from the data alone, without specific assumptions as regards, for example, the shape of the mortality curve might be a very tough task.

We conclude this section by studying how the age exaggeration may affect the life expectancy estimates in the case of a stationary population. The stationary population case is convenient to study because our earlier results suggest the age structure and distribution of deaths at ages $x > x_0 + \delta$ will be consistent with each other despite the distortions caused by the age exaggeration. In particular, the remaining life expectancy at age $x_0 + \delta$ in stationary population is the mean age of the observed distribution of deaths. At ages older than $x_0 + \delta$, the observed distribution of deaths is a mix of the correct distribution taken with weight $(1 - \alpha)l(x_0 + \delta)$, where $l(x)$ is the survival curve at age x , and shifted distribution of deaths at ages $x > x_0$ with weight $\alpha l(x_0)$. Therefore, the observed remaining life expectancy at age $x_0 + \delta$ in the stationary population equals:

$$\tilde{e}(x_0 + \delta) = \frac{(1-\alpha)l(x_0+\delta)e(x_0+\delta)+\alpha l(x_0)e(x_0)}{(1-\alpha)l(x_0+\delta)+\alpha l(x_0)} \quad (13)$$

here, $e(x)$ is the remaining life expectancy at age x . When mortality at ages x_0 to $x_0 + \delta$ falls to levels where $l(x_0 + \delta) \approx l(x_0)$ and $e(x_0) \approx \delta + e(x_0 + \delta)$, Eq. (13) leads to the limit

$$\tilde{e}(x_0) - e(x_0) \approx \tilde{e}(x_0 + \delta) - e(x_0 + \delta) \approx \alpha\delta. \quad (14)$$

That would be an upper limit to the remaining life expectancy bias at all ages up to $x \leq x_0$, because, at those ages, using (13) and inequalities $e(x_0) \leq \delta + e(x_0 + \delta)$, $l(x_0 + \delta) \leq l(x)$, it follows:

$$\begin{aligned} \tilde{e}(x) - e(x) &= \frac{l(x_0+\delta)}{l(x)} [\tilde{e}(x_0 + \delta) - e(x_0 + \delta)] = \frac{l(x_0+\delta)}{l(x)} \left[\frac{(1-\alpha)l(x_0+\delta)e(x_0+\delta)+\alpha l(x_0)e(x_0)}{(1-\alpha)l(x_0+\delta)+\alpha l(x_0)} - \right. \\ &e(x_0 + \delta) \left. \right] = \frac{l(x_0+\delta)}{l(x)} \frac{\alpha l(x_0)[e(x_0)-e(x_0+\delta)]}{(1-\alpha)l(x_0+\delta)+\alpha l(x_0)} \leq \frac{l(x_0+\delta)}{l(x)} \frac{\alpha l(x_0)\delta}{(1-\alpha)l(x_0+\delta)+\alpha l(x_0)} = \\ &\frac{l(x_0)}{l(x)} \frac{l(x_0+\delta)}{(1-\alpha)l(x_0+\delta)+\alpha l(x_0)} \alpha\delta \leq \alpha\delta. \end{aligned} \quad (15)$$

We found that, in the stationary population, at ages younger than the first age where the age exaggeration begins, distortion to the remaining life expectancy will be bounded by the upper limit that is equal to the mean magnitude of the age exaggeration $\alpha\delta$. That upper limit will only be reached when mortality below age $x_0 + \delta$ falls to zero. For a non-stationary population, however, the stationary upper limit may be exceeded, as, for example, in stable growing populations.

2. Numerical simulations

To have a better idea of distortions due to age miss-statements in the (x_0, α, δ) model, we run numerical simulations using (4), (7) and (8) (Figures 1-4).

Distortions to population age structure depend substantially on the population growth rate and, by implication, deviations from the stationary age composition (Figure 1: in this and following figures, the horizontal axis is the correct death rate in log-scale that, in Gompertzian mortality assumed here, represents age). At large age exaggerations δ , distortion patterns for various combinations of the growth rate and α overlap with each other. That, unfortunately, makes it problematic to reconstruct the age exaggeration parameters from the observed population age composition.

Confirming the results for the limit cases (5) and (6), distortions of the death rates (Figure 2) are not much sensitive to the population growth rate as compared to the age exaggeration parameters. This suggests it might be possible to reconstruct the exaggeration model parameters as well as the real death rates from the observed deviations of the death rates' patterns from assumed model mortality curves. The underestimation of the death rates at an advanced old age is much more substantial than at a younger age. In extreme cases, the observed death rates may go down to as low as only dozens of percent of the correct rate.

Strong biases at old age, both absolute and relative, in comparison to younger ages, make it possible for the observed death rates to follow unrealistic non-monotone trajectories at old age (Figure 3)¹. Consequently, the observed LAR may profoundly mislead because of the age exaggeration (Figure 4). Somewhat counterintuitively, the age at observed maximum mortality (at which LAR equals zero) is younger for less-prevalent (but strong) age exaggerations than for the more-prevalent ones. This is an important observation pointing to the possibility to empirically distinguish the cases of age exaggeration by a small percent of the population from cases of a heterogeneous population. As shown by Vaupel and Yashin (1985), population heterogeneity in frailty may well produce sorts of distorted age patterns of the death rates that we demonstrate here for the age exaggeration. Indeed, the two cases are, generally speaking, identical mathematically, as a proportion of people who misplace themselves into a wrong age category may formally be interpreted as a sub-population of the reported age with different frailty. Old-age mortality peaking at some age seems to be a sign of few people strongly exaggerating their age rather than more substantial parts of the population differing in their frailty. An idea about differential frailty may be developed based on educational-levels life expectancies. Eurostat data suggest that educational categories differ by up to a few years in terms of life expectancy at age 60 (Eurostat, 2017) (Table 1). That implies selection processes in a heterogeneous population might typically manifest themselves in reducing the steepness (LAR) of the mortality curve rather than producing non-monotone mortality at old age. The latter may, therefore, be a hallmark of age exaggeration.

¹ Notably, such non-monotone patterns of old-age mortality exist, although not common, even in the presumably high-quality data collection of the Human Mortality Database (University of California and Max Planck Institute for Demographic Research (Rostock), 2020).

Unless α is large, biases are small at young ages (with small M_x). On the other hand, the distortions at advanced ages may be substantial, even at low age exaggeration levels (both in terms of α and δ).

Table 1. Remaining life expectancy at age 60 by educational category, years 2007-2013, selected countries

	Less than primary, primary and lower secondary education	Upper secondary and post-secondary non- tertiary education	Tertiary education
Bulgaria	17.3	20.4	21.5
Czech Republic	20.3	21.0	23.5
Denmark	21.8	22.6	23.6
Estonia	18.0	21.2	23.4
Greece	23.8	25.3	25.5
Croatia	20.5	19.9	22.5
Italy	24.2	26.0	25.8
Hungary	18.1	21.4	21.5
Malta	23.2	23.9	24.6
Poland	20.7	21.1	23.3
Portugal	23.7	24.2	25.2
Romania	19.3	20.7	20.3
Slovenia	22.0	23.1	24.4
Slovakia	18.9	20.8	22.2
Finland	23.1	23.9	24.7
Sweden	23.3	24.1	25.0
Norway	22.8	24.0	24.9
Macedonia, FYR	18.1	19.1	20.8
Turkey	21.4	21.8	22.6
AVERAGE	21.1	22.3	23.4
Difference to the Tertiary level	2.4	1.1	-

Source: Own elaboration on (Eurostat, 2017).

Figure 1. Distortions to the observed population numbers in relevant age groups as compared to the true population by age: by δ (“Delta”, in years, columns), b (rows), α (“Alpha”, in percent, colors), r (line types). Horizontal axis: the true death rate $M(x)$, log-scale.

Note: values below 1 indicate underestimation; values above 1 indicate overestimation.

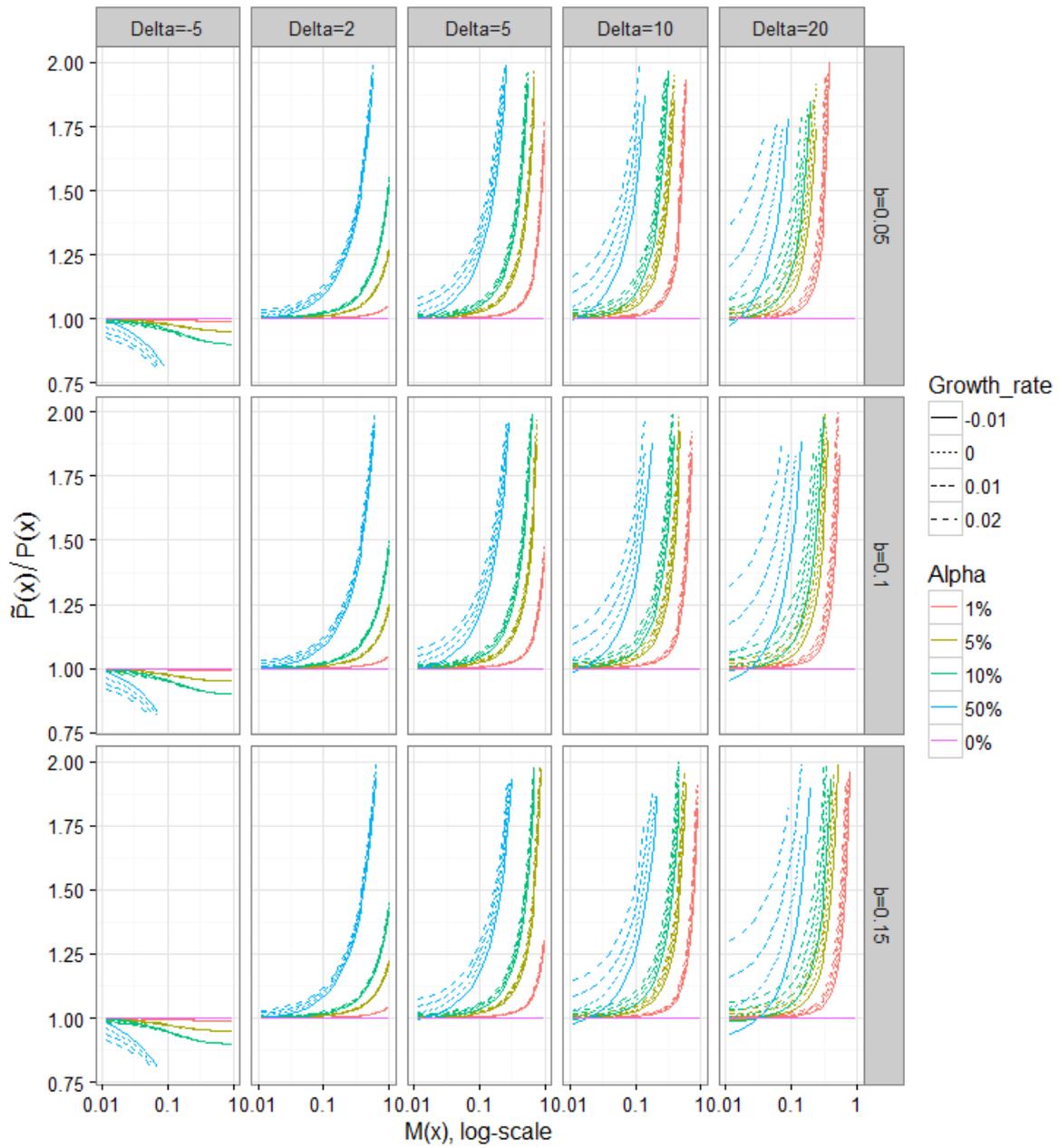


Figure 2. Distortions to the observed death rates as compared to the true rates: by δ (“Delta”, in years, columns), b (rows), α (“Alpha”, in percent, colors), r (line types). Horizontal axis: the true death rate $M(x)$, log-scale.

Note: values below 1 indicate underestimation; values above 1 indicate overestimation.

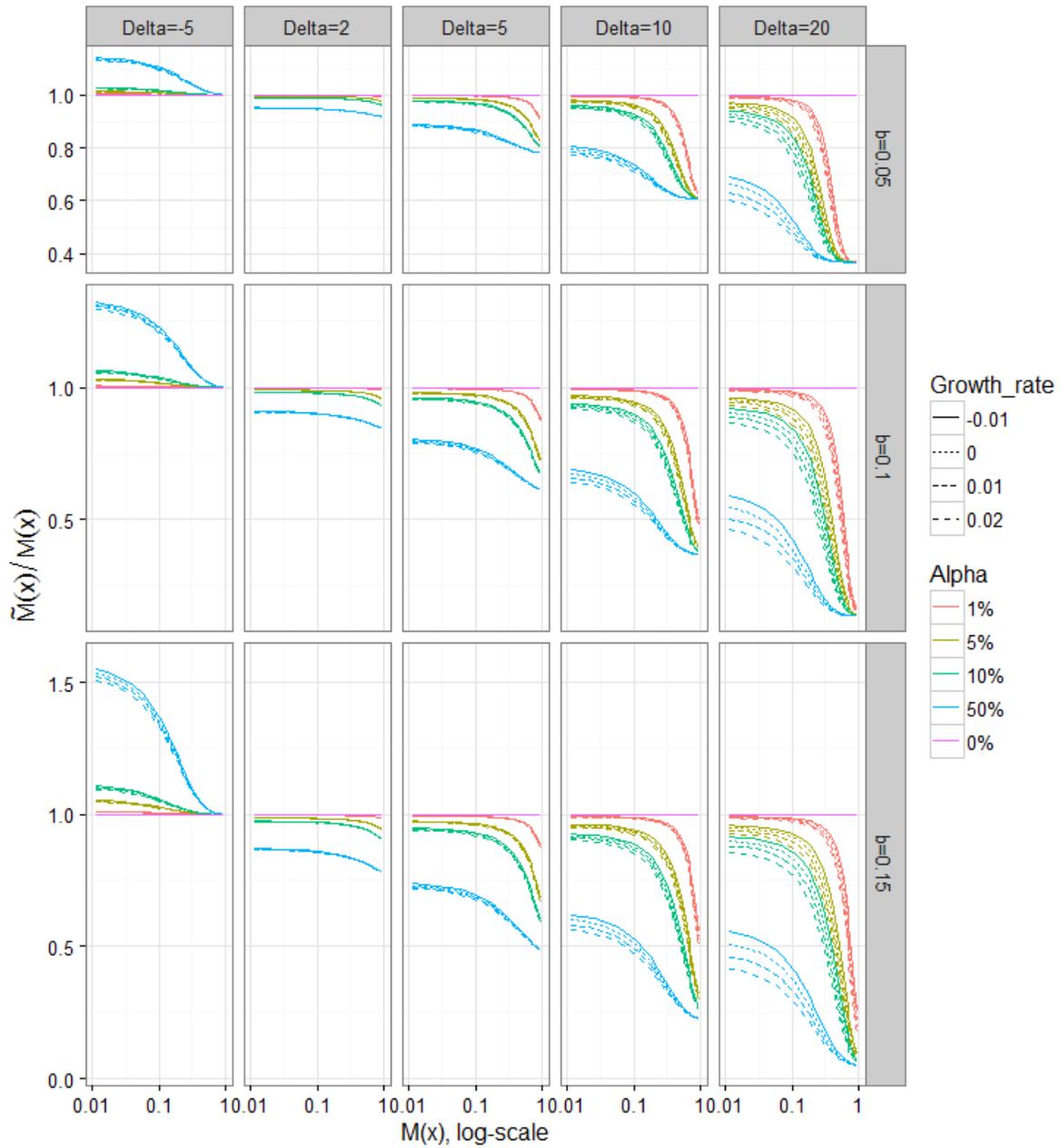


Figure 3. Observed death rates (vertical axis) vs true death rates (horizontal axis, log-scale): by δ (“Delta”, in years, columns), b (rows), α (“Alpha”, in percent, colors), r (line types).

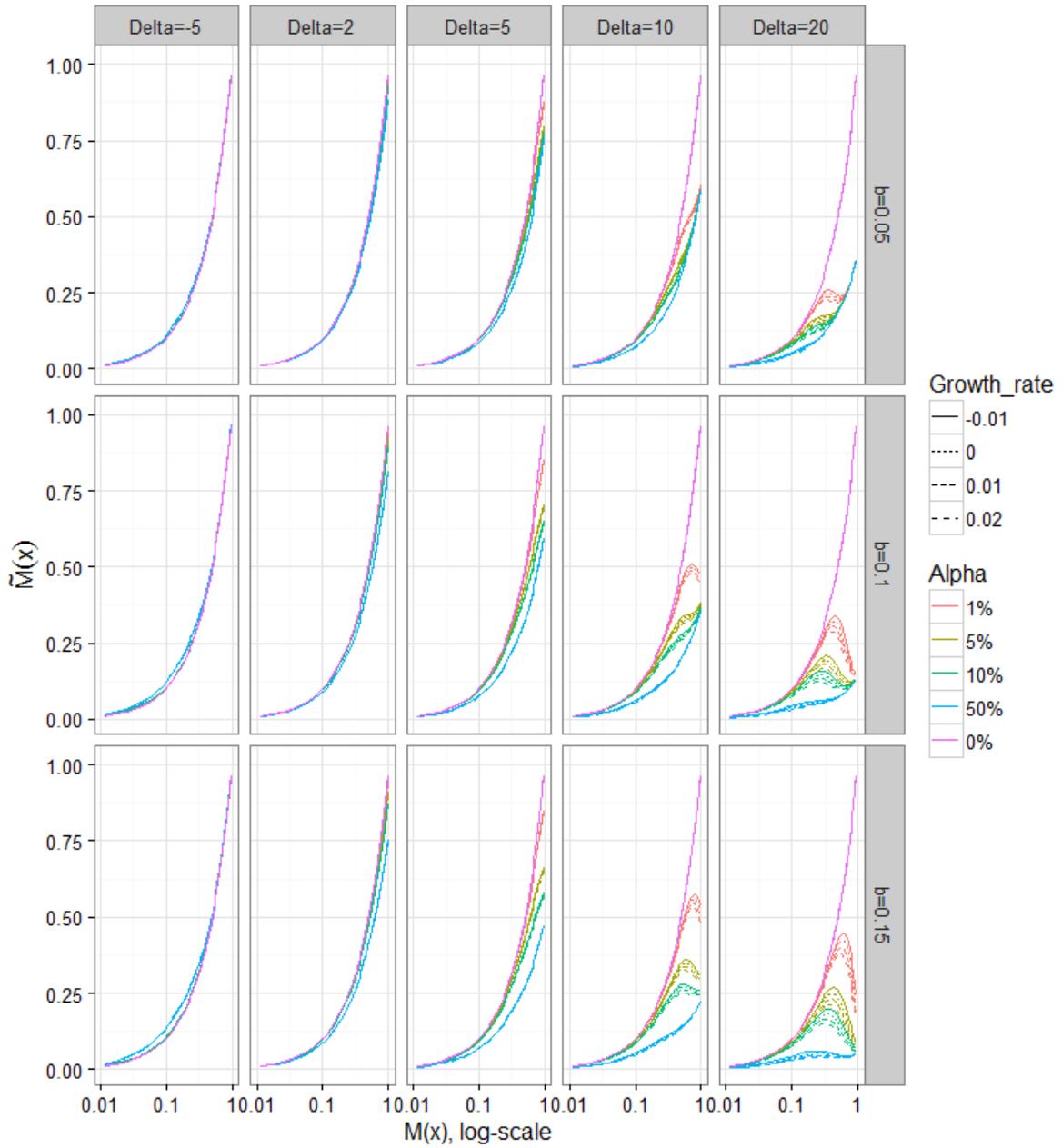
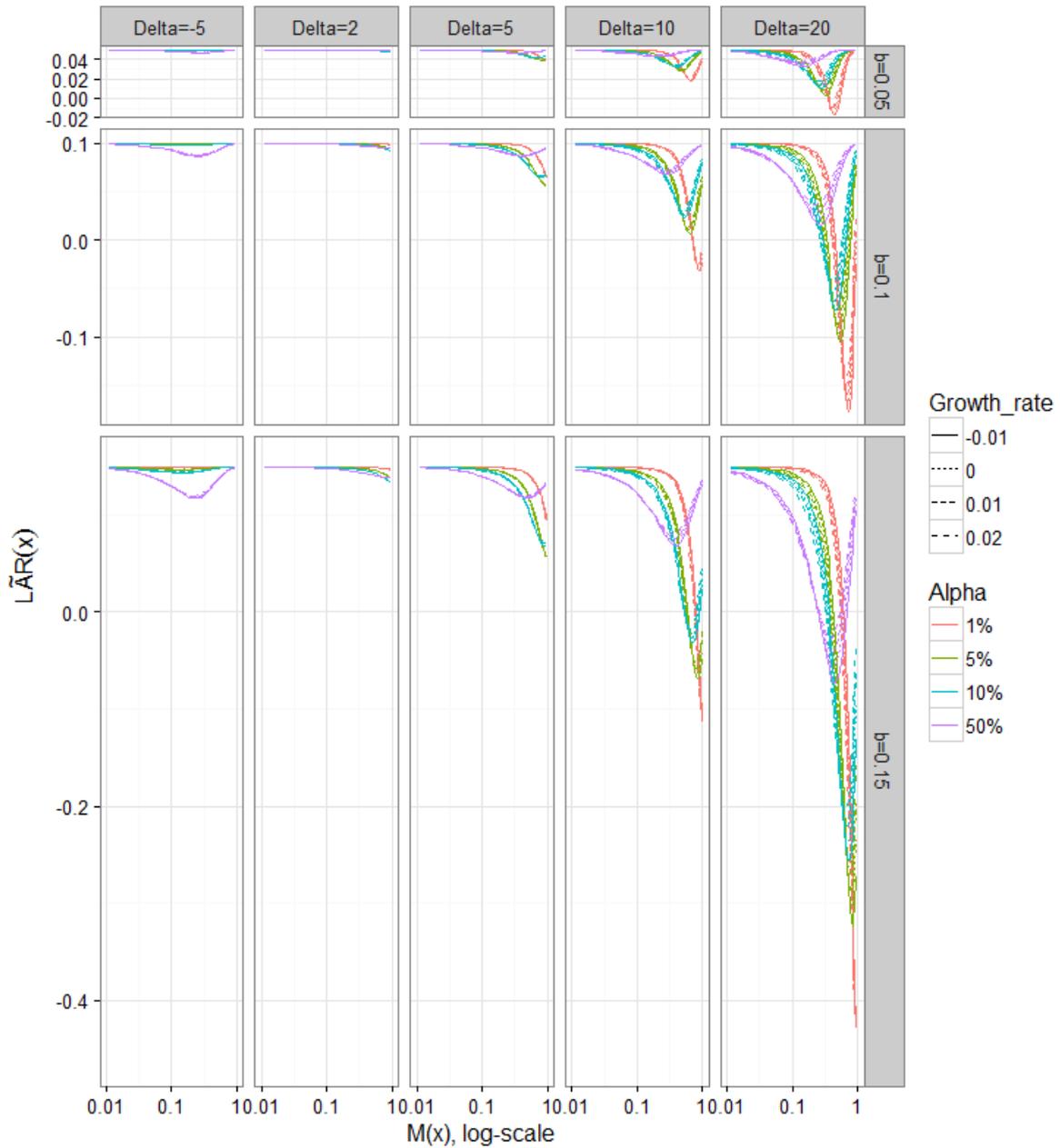


Figure 4. Observed life table ageing rate (LAR) at relevant age: by δ (“Delta”, in years, columns), b (rows), α (“Alpha”, in percent, colors), r (line types). Horizontal axis: the true death rate $M(x)$, log-scale.



3. Empirical assessment: Calibrating the age exaggeration model

We illustrate the practical relevance of the model presented by applying it to empirical data. The latter comes from the Human Life Table Database (2020) where we have selected some mortality cases that show non-monotone patterns at old age. That is because our numerical investigation suggests non-monotone mortality at old age might be a hallmark of age

exaggeration that distinguishes effects of age exaggeration from effects of population heterogeneity. The list of selected populations is available in Table 2.

For each of the selected populations, we found the model parameters that fit best the empirical data while assuming an underlying model of real mortality rates. For the underlying mortality model, we tested the two common demographic models: the exponential Gompertz (Gompertz, 1825) and the logistic Kannisto (Thatcher, Väinö Kannisto and Vaupel, 1998) models. Estimation results for the Gompertz model are presented in Table 2. Graphical results for both models are shown in Figures 5 and 6, respectively. Parameters estimated suggest that, indeed, we may typically deal with only small fractions of the population exaggerating their age by a substantial amount. As a result, distortions of the life expectancy (at birth or at the age when people start exaggerating their age) are usually small. Nevertheless, the mortality curve at advanced old age may be substantially biased even in the case of moderate age exaggeration prevalence.

Table 2. Reconstructed age exaggeration model parameters and observed and corrected life table indicators. Selected populations of the Human Life Table Database (2020).

Country	Sex	Period	b	x_0	α	δ	$e(0)$	$e(0)$ cor- rected	$e(0)$ bias	$e(x_0)$	$e(x_0)$ cor- rected	$e(x_0)$ bias
China rural	M	1981	0.087	73	4.1%	17.0	65.9	65.9	-0.02	8.1	8.2	-0.05
China rural	F	1981	0.098	56	1.0%	23.0	68.7	68.6	0.11	21.2	21.1	0.14
Cuba	M	2005_07	0.091	75	6.5%	13.9	76.1	75.9	0.11	10.9	10.7	0.18
Cuba	F	2005_07	0.103	77	9.5%	12.6	80.1	79.9	0.20	11.2	10.9	0.30
Mexico	M	1983_85	0.072	76	3.0%	17.0	66.2	66.1	0.02	10.1	10.0	0.06
Mexico	F	1983_85	0.084	76	4.4%	17.0	72.2	72.2	0.02	10.8	10.8	0.03
Turkey	M	2013_14	0.096	63	0.9%	21.0	75.3	75.3	0.03	17.6	17.6	0.03
Turkey	F	2013_14	0.117	71	6.7%	13.0	80.7	80.6	0.18	14.7	14.4	0.21
Uruguay	M	2004	0.099	80	4.5%	8.0	71.7	71.6	0.04	6.6	6.4	0.12
Uruguay	F	2004	0.124	68	21.3%	5.4	79.0	78.3	0.76	16.8	15.9	0.92

Source: Own elaboration on (HLTDB, 2020).

Figure 5. Selected empirical death rates (circles), reconstructed underlying unbiased death rates (Gompertz model; the upper green solid lines) and the fitted (x_0, α, δ) model (the lower red solid lines). The horizontal axes: age; the vertical axes: the death rate.

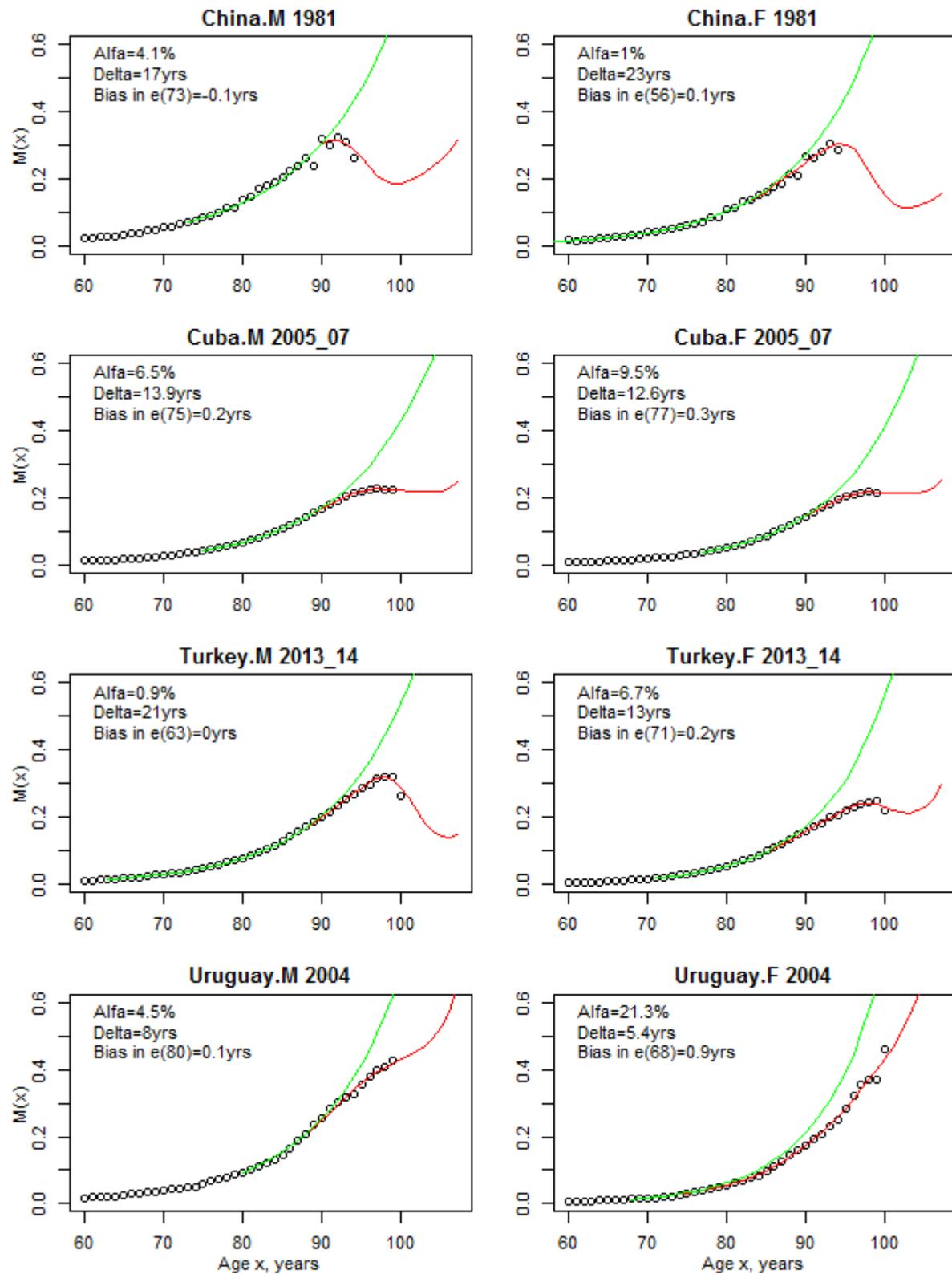
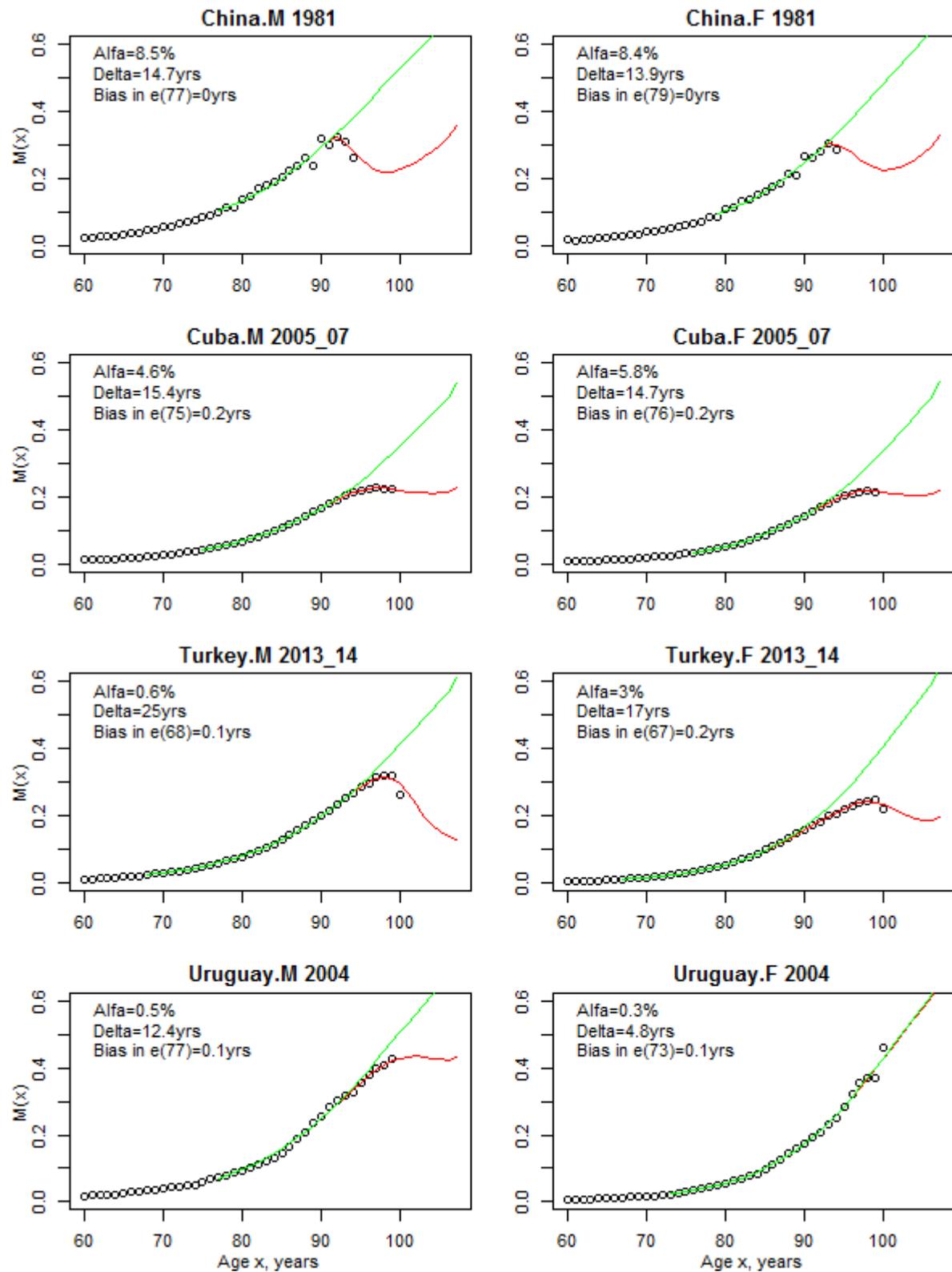


Figure 6. Selected empirical death rates (circles) by age, reconstructed underlying unbiased death rates (Kannisto model; the upper green solid lines) and the fitted (x_0, α, δ) model (the lower red solid lines). The horizontal axes: age; the vertical axes: the death rate.



4. Conclusions

Our theoretical, numerical, and empirical findings suggest that age exaggeration may have profound effects on mortality and population development over the age scale.

It is often presumed that the age exaggeration is practiced by a small fraction of the population only. That, nonetheless, may lead to significant distortions of the death rates at advanced old age and distort the shape of the mortality curve. We found that infrequent, yet, large, age overstatements may lead to deceleration and even non-monotone behavior of the mortality curve at advanced old age. While non-monotone mortality and mortality plateau were suggested to be possible outcomes of selection effect in heterogeneous populations (Vaupel and Yashin, 1985; Missov, Németh and Daňko, 2016), our findings suggest the age exaggeration might be another, if not the primary, possible source of such effect. Indeed, a more frequent age exaggeration may be even worse in terms of biasedness of the mortality curve. Such patterns, indeed, do not exclude more prevalent but weaker age exaggerations. Even more, stronger (albeit few) exaggerations practiced in the population may, perhaps, be taken as indications of wider problems. When few people overstate their true age by large amounts, more people will likely exaggerate their age more moderately.

Given the difficulty of distinguishing the age exaggeration from the mortality selection, it is hard to suggest the model proposed here as the main tool of correcting the biasedness of the remaining life expectancy caused by the age exaggeration. For that purpose, one may resort to other methods found efficient in the literature (Ediev, 2018, 2019). Yet the behavioural age exaggeration model presented here may be both a useful supplement to other methods. In combination with such alternatives, the model may also be used to assess the extent of the age exaggeration prevalence in the population.

References

- Ediev, D. M. (2017) 'Constrained Mortality Extrapolation to Old Age: An Empirical Assessment', *European Journal of Population*. doi: 10.1007/s10680-017-9434-4.
- Ediev, D. M. (2018) 'Expectation of life at old age: revisiting Horiuchi-Coale and reconciling with Mitra', *Genus*, 74(1). doi: 10.1186/s41118-018-0029-7.
- Ediev, D. M. (2019) 'On the sources of instability of the Mitra model for years of life at old-age', *Communications in Statistics: Case Studies, Data Analysis and Applications*, forthcoming.
- Eurostat (2017) *Database - Eurostat, 2017*. Available at: <http://ec.europa.eu/eurostat/data/database> (Accessed: 10 January 2017).
- Gompertz, B. (1825) 'On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies', *Philosophical Transactions of the Royal Society of London*, 115, pp. 513–583.
- HLTDB (2020) *The Human Life-Table Database jointly developed by Max Planck Institute for Demographic Research (Rostock, Germany), Department of Demography at the University of California (Berkeley, USA), and Institut national d'études démographiques (Paris, France)*. Available

at: <http://www.lifetable.de/> (Accessed: 10 January 2017).

Horiuchi, S. and Coale, A. J. (1982) 'A Simple Equation for Estimating the Expectation of Life at Old Ages', *Population Studies*, 36(2), pp. 317–326. doi: 10.2307/2174203.

Horiuchi, S. and Wilmoth, J. R. (1997) 'Age patterns of the life table aging rate for major causes of death in Japan, 1951-1990.', *The journals of gerontology. Series A, Biological sciences and medical sciences*, 52(1), pp. B67-77. Available at:

<http://www.ncbi.nlm.nih.gov/pubmed/9008660> <http://biomedgerontology.oxfordjournals.org/cgi/content/abstract/52A/1/B67>.

Keyfitz, N. and Caswell, H. (2005) *Applied mathematical demography*. Springer.

Missov, T. I., Németh, L. and Daňko, M. J. (2016) 'How much can we trust life tables? Sensitivity of mortality measures to right-censoring treatment', *Palgrave Communications*. Nature Publishing Group, 2, p. 15049. doi: 10.1057/palcomms.2015.49.

Mitra, S. (1984) 'Estimating the Expectation of Life at Older Ages', *Population Studies*, 38(2), pp. 313–319. doi: 10.2307/2174079.

Preston, S. H., Heuveline, P. and Guillot, M. (2001) *Demography: Measuring and Modeling Population Processes*. Oxford: Blackwell Publishers. doi: 10.2307/1535065.

Thatcher, A. R., Kannisto, V. and Vaupel, J. W. (1998) *Odense Monographs on Population Aging 5: The force of mortality at ages 80 to 120*. Odense: Odense University Press.

Thatcher, A. R., Kannisto, V. and Vaupel, J. W. (1998) *The Force of Mortality at Ages 80-120. Monographs on Population Aging*. Odense, Denmark: Odense University Press. Available at: <http://www.demogr.mpg.de/Papers/Books/Monograph5/ForMort.htm>.

University of California, B. and Max Planck Institute for Demographic Research (Rostock) (2020) *Human Mortality Database. Online database sponsored by University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany)*. Available at: www.mortality.org (Accessed: 27 January 2020).

Vaupel, J. W. and Yashin, A. (1985) 'Heterogeneity's ruses: some surprising effects of selection on population dynamics.', *The American statistician*, 39(3), pp. 176–85. Available at:

<http://www.ncbi.nlm.nih.gov/pubmed/12267300> (Accessed: 11 January 2017).

Outlier Detection in Consumer Price Index by Using Fuzzy Clustering: A Case Study of Iranian Inflation Data

Mostafa Farrokhfal¹, Mohsen Saadatpour², Reza Hadizadeh³

¹ Price Indices Office, Statistical Centre of Iran No. 1, Corner of Rahi Moayeri St., Dr. Fatemi Ave., Tehran 1414663111, I. R. Iran
(E-mail: m.farokhfal@gmail.com)

² Price Indices Office, Statistical Centre of Iran No. 1, Corner of Rahi Moayeri St., Dr. Fatemi Ave., Tehran 1414663111, I. R. Iran
(E-mail: m.saadatpour@aut.ac.ir)

³ Price Indices Office, Statistical Centre of Iran No. 1, Corner of Rahi Moayeri St., Dr. Fatemi Ave., Tehran 1414663111, I. R. Iran
(E-mail: reza.h.fuzzy@gmail.com)

Abstract. High volume of pricing data and time limitation are two main constraints of publishing Consumer Price Index (CPI) monthly reports in Statistical Centre of Iran (SCI). To examine accuracy of data collected from a variety of sources, Observation Officers of SCI have been used experimental and traditional methods. In this paper, a new approach for examining the accuracy of data is proposed. Prices of goods and services are grouped in two types: low-volatile and high-volatile data. Using clustering data by fuzzy K-means algorithm, outlier detection is performed and clean data is obtained, Next data is used to calculate clean relative price by filtering method. Results of the proposed approach indicate that quality of relative price is boosted and CPI is calculated in less time compared to traditional methods, thereby demonstrating high performance of the algorithm.

Keywords: Consumer Price Index, Data mining, Fuzzy Clustering, Outlier Data.

1 Introduction

The sharp rise in prices over the past year and policymaker's attention to this rate have forced the Iran's NSO to minimize the time required to collect data to release inflation results. Traditional methods to resolve this necessity were ineffective, and we met this need with process correction by modern methods on data mining on prices.

The previous process involved collecting prices with a paper questionnaire and prices were checked once in the provinces and then re-examined by experts at the Iran's NSO. The new process of collecting prices is done electronically with a tablet. Prices are also not monitored in the provinces, and the modifications which have been made at this stage have been mainly to design an electronic

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



questionnaire. Checking prices in traditional way is known as a time-consuming method and we replaced it with data mining method.

We review and validate over 350,000 prices per month (240,000 in urban and 110,000 in rural areas). Checking this number is not only to calculate relative prices and indices (total index in urban, rural and whole country and decile total index), but also to calculate the average price of foods and the monthly household purchasing power change.

Our new approach improved this process by two important ways: change the traditional-based method to scientific-based and practical method in terms of relative prices. Second, ability to check prices with more detail and high accuracy due to limited time.

The paper is organized as follow: in section 2, we introduce former related work on Data mining, section 3 debates our clustering methods contain theoretical and proposed algorithm. In section 4 show empirical result with figures and tables of clustering in low and high price volatility of items. Finally, conclusion gather in section 5.

1.1 The literature review

Irrespective of how the 'representative outlet' is determined, it is clear that households may face different inflation due to both different spending patterns and differences in the price changes of various goods and services [1]. It could weaken the ability of the price index to proxy the experiences of all households. This is, to some extent, borne out by the claims of individuals across the income distribution that the rates of inflation which they experience are significantly different to the official consumer inflation statistics [2]. Thiprungsri and Vasarhelyi in [3] examine the application of cluster analysis in the accounting domain, particularly discrepancy detection in audit. This study, clustering accounting data, has been shown as a good method for anomaly detection and searching the use of clustering technology to automate fraud filtering during an audit. They employ cluster analysis to help auditors focus their efforts when evaluating group life insurance claims.

Zazzaro and CIRA in [4] focused on how to mine large datasets by applying the ECF-means algorithm, in order to detect potential outliers. They utilized Fuzzy Logic for this matter. Through the three exposed case studies, the experimental outcomes on real world datasets, and the comparison with the results of other outlier detection methods, they proposed algorithm to provide other types of deeper detections; the first case study related to the famous Wine dataset from the UCI Machine Learning Repository; the second one involved the analysis and exploration of data in meteorological domain, where various results were explained. Finally, the third case study explored the well-known Iris dataset which traditionally has no outliers, while new information is discovered by the ECF-means algorithm and exposed here with many results.

The processing system and checking prices for the Price Indices such as Consumer Price Index (CPI) is always being updated. Therefore, the current methods used by the system for methodological tasks such as outlier detection are being reviewed to determine if changes can be made to enhance the

efficiency of the system and the quality of the estimates. The paper [5] studied different non-parametric outlier detection methods that could be implemented in the redesigned CPI system. Mehrotra and Joshi in [6] proposed an algorithm using K-means clustering and C5.0 decision tree, where Euclidean distance was used to find the closest cluster for the data set and then decision tree was built for each cluster using C5.0 decision tree technique and the rules of decision tree is used to classify each anomalous and normal instances in the dataset. Their method gives classification with high accuracy in the results. Loureiro and others in [7] described a methodology for the application of hierarchical clustering methods to the task of outlier detection. This methodology was tested on the problem of cleaning official Statistics data. They worked to detecting erroneous foreign trade transactions in data collected by the Portuguese Institute of Statistics (INE). The results of the application to this problem clearly met the performance criteria outlined by the human experts. Thiprungsri and Vasarhelyi examined the application of cluster analysis in the accounting domain, particularly discrepancy detection in audit. Cluster analysis groups data so that points within a single group or cluster are similar to one another and distinct from points in other clusters. They used cluster analysis to help auditors focus their efforts when evaluating group life insurance claims [3].

2 Proposed Methodology

Cluster analysis is a way of “slicing and dicing” data to allow the grouping together of similar entities and the separation of dissimilar ones. Issues arise due to the existence of a diverse number of clustering algorithms that each one has different techniques and inputs, and with no universally optimal methodology. Thus, a framework for cluster analysis and validation methods are needed. Our approach is to use cluster ensembles from a diverse set of algorithms.

Fuzzy C-Means (FCM) is a soft clustering algorithm proposed by Bezdek [8]. Unlike K-means algorithm in which each data object is the member of only one cluster, a data object is the member of all clusters with varying degrees of fuzzy membership between 0 and 1 in FCM. Hence, the data objects closer to the centers of clusters have higher degrees of membership than objects scattered in the borders of clusters.

It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m \|x_i - c_j\|^2, \quad 1 \leq m \leq \infty$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j , x_i is the i^{th} of d -dimensional measured data, c_j is the d -dimension center of the cluster, and $\|*\|$ is any norm expressing the similarity between any measured data and the center.

Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{i=1}^c \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}, \quad c_j = \frac{\sum_{i=1}^N u_{ij}^m x_i}{\sum_{i=1}^N u_{ij}^m}$$

This iteration will stop in the time of where is a termination criterion between 0 and 1, whereas k are the iteration steps. This procedure converges to a local minimum or a saddle point of J_m . the algorithm is composed of the following steps:

First, it needs to preparation and pre-processing of data in order to use fuzzy method. This section consists of 2 steps:

- Step 1: Removing Incomplete, poor quality and confused data. This step removed some records had missing information or were not consistent with other information.
- Step 2: Extract data and create a database. The purpose of this step is to create an integrated database of urban price of consumer goods and services. A database is a repository of collected data from various sources that have been stored and structured in different shapes.

Second, after collecting urban prices of goods and services of consumption and preparing them, Fuzzy clustering is used to assign a price to one of the 7 clusters (The number of clusters is obtained from the assessment of available data, this assessment contains types of imported or domestic, fresh and frozen Items, packed or unpacked.). For this purpose, R software has been used. R software is a powerful software in the field of data mining and modern statistical methods. Clustering in R software is performed by various packages. In the present study, the package "ppclust" and the "fcm" function for data clustering have been used. Then, the following steps are suggested for finding outliers.

In this proposed algorithm, the most focus on the prices in the first and last clusters, Because of the possibility of outliers in these clusters is greater. These clusters are sorted and averaged, and then identified incorrect prices excluded from future calculations in eight stages, using the algorithm 1.

Algorithm 1:

- 1- Calculate the average of the clusters, sort cluster in ascending order so that the cluster with the lowest and highest mean is rated 1 and 7, respectively.
- 2- In cluster i, if $\max_i/\min_i > 5$ then prices that $p < \max_i/5$ remove in cluster with minimum mean (first cluster) also prices that $p > \min_i*5$ remove in cluster with maximum mean (7th cluster).
- 3- If $\text{Avg}_i/\text{Avg}_{i-1} > 2$ then the cluster with lowest mean (first cluster), prices with two standard deviations from the mean are omitted only from the lower bound, and the cluster with the highest mean (seventh cluster), prices with the difference of two standard deviations from the mean is only eliminated from the upper bound.

- 4- If a cluster has a standard deviation of zero (or has only one price or has several equal prices) and the cluster has only one member or less than one percent of the total quotes, then the whole cluster must be removed.
- 5- After eliminating possible quotes by steps 1 through 4, means are updated and re-ranked.
- 6- Calculate the standard deviation of the clusters, then the clusters are sorted in ascending order, forming a Std_i/Std_{i-1} ratio. After calculating the ratio of standard deviations, rank based on the mean. If the first or 7th clusters had a standard deviation ratio of more than 2, then prices with two standard deviations from the mean are only eliminated from the lower bound of the first cluster, and in the 7th cluster, prices with two standard deviations from the mean are only removed from upper bound.
- 7- In step 6, if the first cluster has less than one percent of all quotes then the lowest price should be eliminated and also if the 7th cluster has less than one percent of all quotes then the highest price should be eliminated.
- 8- Calculate the standard deviation of the clusters, then the clusters are sorted in ascending order, forming a Std_i/Std_{i-1} ratio, if the ratio was greater than 2 for the first or last cluster, then prices with two standard deviations from the mean only eliminated from the lower bounds for the first cluster, and in the 7th cluster, prices with two standard deviations from the mean are only removed from the upper bounds.

3 Experimental Result

This section will present the performance and accuracy of proposed algorithm and for this CPI dataset have been used. We considered two categories of items in this paper. The first group is prices with high volatility and the second is prices with low volatility. The following tables show the representative items for each of these groups and their statistics:

Table 1 – Three items with High volatility

Item Name	Average Price(IRR)	Min Price (IRR)	Max Price (IRR)	Standard deviation
Shrimp	559,491	120,000	1,200,000	35,830
Gas heaters	3,538,826	1,750,000	10,400,000	194,435
Mobile	18,968,860	780,000	115,000,000	2,276,294

Table 2 – Three items with low volatility

Item Name	Average Price (IRR)	Min Price (IRR)	Max Price (IRR)	Standard deviation
Potato	16,718	10,000	46,000	496
Onion	17,635	9,800	35,000	438
Liquid oil	58,087	36,500	180,000	1,730

Usually, Items with high volatility in price are that mainly imported or domestic, have various brands or various properties, are fresh or frozen Items, while bulk foods or items with limited properties have prices with low volatility. Figure 1 shows a cluster of shrimp item using the K-means algorithm in R software and display outlier data.

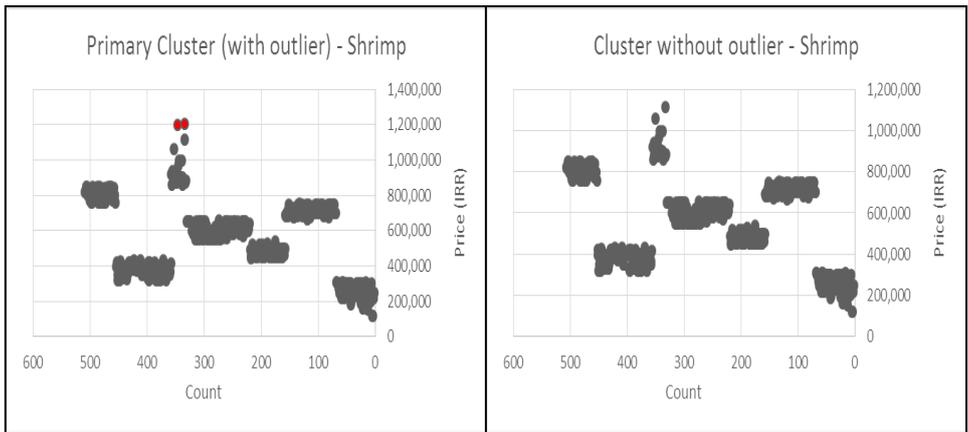


Figure 1. Shrimp cluster before and after outlier detection

In this case, the prices of 1200000 and 1195000 Rials are eliminated by the algorithm due to the large standard deviation created in their cluster. Of course, we only remove these high prices that are part of the cluster (not all clusters) and are identified as outlier by the algorithm from the current period calculations and keep it for next period calculations. Other prices are expected to tend the level of observed high prices; as high prices may be collected later this month. Figure 2 shows the clustering and identification of outliers for an item in low-volatility group.

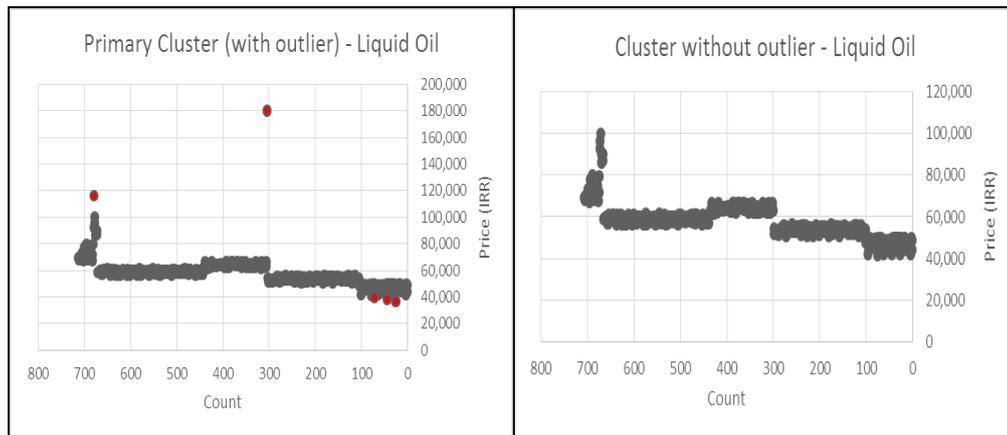


Figure 2. Liquid Oil cluster before and after outlier detection

In this example, Cluster 3 is identified as outliers by proposed algorithm and we remove the entire cluster from the computation. Removed cluster will not be included in the calculations of the following month. (Unlike the high prices that are part of a cluster) also the 4 prices smaller than the first cluster (prices 36500 to 41400 Rials) and the maximum prices of 6th clusters (116000 Rials) are identified by the algorithm and eliminated as outliers from the current period calculations. Minimum prices which are excluded from the clusters are not returned to the calculations just like the outlier clusters.

The following table (Table 3) show the status of the clusters in the liquid oil before the algorithm is implemented:

Table 3 – Initial clustering status for liquid oil samples

Cluster	Count of Prices	Average Price	Min Price	Max Price	Standard Dev. of Price
1	105	47,214	36,500	50,000	2,959
2	198	53,769	50,625	56,250	1,335
5	228	58,893	56,399	61,333	1,207
4	138	64,064	61,500	66,700	1,294
7	33	70,890	67,000	80,000	3,617
6	9	93,833	86,000	116,000	9,487
3	2	180,000	180,000	180,000	-
Grand Total	713	58,087	36,500	180,000	1,730

In this example, Cluster 3 is identified as outliers by proposed algorithm and the entire cluster is removed from the computation. Removed cluster will not be included in the calculations the following month. (Unlike the high prices that are part of a cluster) also the 4 prices smaller than the first cluster (prices 36500 to 41400 Rials) and the maximum prices of 6th clusters (116000 Rials) are

identified by the algorithm and eliminated as outliers from the current period calculations. Minimum prices which are excluded from the clusters are not returned to the calculations just like the outlier clusters.

Table 4 show the status of the clusters in the liquid oil after the algorithm is implemented:

Table 4 – clustering status after remove outlier for liquid oil samples

Cluster	Count of Prices	Average Price	Min Price	Max Price	Standard Dev. of Price
1	101	47,595	41,400	50,000	2,286
2	198	53,769	50,625	56,250	1,335
5	228	58,893	56,399	61,333	1,207
4	138	64,064	61,500	66,700	1,294
7	33	70,890	67,000	80,000	3,617
6	8	91,063	86,000	100,000	4,888
Grand Total	706	57,776	41,400	100,000	7,141

As table 4 show, with eliminating 7 quotes, we were able to get better results in comparison with primary clustering. (Reduce of standard deviation in cluster 1 from 2,959 to 2,286 and sharp decline in standard deviation from 9,487 to 4,888 in cluster 6)

4 Conclusion and Future Work

In this paper, the prices we identified in the traditional way as outlier data, automatically took the algorithm out of the calculations. Now we don't need to investigate 350,000 quotes with traditional method in 3 days.

After more, we constructed relative price with confirmed prices in step one and then using fuzzy K-means method for clustering and identify outlier relative price. Results of the purposed approach indicate that quality of relative price is boosted and CPI is calculated in less time compared to traditional methods, thereby demonstrating high performance of the algorithm.

References

1. Hait, Pavel, and Petr Jansky. "Inflation differentials among Czech households." CERGE-EI Working Paper Series 508, 2014.
2. Oosthuizen, Morne. "Consumer Price inflation across the income distribution in South Africa.", Development Policy Research Unit Working Paper 07/129. 2007.
3. Thiprungsri, Sutapat, and Miklos A. Vasarhelyi. "Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach." International Journal of Digital Accounting Research 11, 2011.
4. Zazzaro, Gaetano, and Angelo Martone. "Fuzzy Outlier Detection by Applying the ECF-Means Algorithm.", International Journal on Advances in Software, vol 12 no 1 & 2, 2019.

5. Rais, Saad. "Outlier detection for the consumer price index." In Proceedings of the Survey Methods Section : Statistical Society of Canada Annual Meeting, vol. 110. 2008.
6. Mehrotra, Mani, and Nakul Joshi. "Anomaly Detection in Temporal data Using Kmeans Clustering with C5. 0.", The International Journal of Engineering and Science (IJES), 2017.
7. Loureiro, Antonio, Luis Torgo, and Carlos Soares. "Outlier detection using clustering methods : a data cleaning application." In Proceedings of KNet Symposium on Knowledge-based systems for the Public Sector. Bonn: Springer, 2004.
8. Bezdek, James C., Robert Ehrlich, and William Full. "FCM: The fuzzy c-means clustering algorithm." *Computers & Geosciences* 10, no. 2-3, 1984.

An empirical study of the emergence of Taylor's power law in random graph models and real networks

István Fazekas¹, Csaba Noszály², and Noémi Uzonyi³

¹ University of Debrecen, Debrecen, Hungary

(E-mail: fazekas.istvan@inf.unideb.hu)

² University of Debrecen, Debrecen, Hungary

(E-mail: noszaly.csaba@inf.unideb.hu)

³ University of Debrecen, Debrecen, Hungary

(E-mail: unomi95@gmail.com)

Abstract. In this paper we study the emergence of Taylor's power law in random graph models and real life networks. We analyse numerically the networks, and we consider the degrees of the nodes inside certain communities of the networks. We calculate the empirical mean values and empirical variances of the degrees and then we look for a functional relationship between these two quantities. We visualize these pairs on double logarithmic scale. Since Taylor's law states that the variance is a power function of the mean, therefore on double logarithmic scale a straight line should appear. Besides visualization, we calculate the residual sum of squares and the Pearson correlation coefficient. The results of our study will show that the fit to a straight line is perfect in certain cases, it is fair in several other cases and even in the worst case, the straight line shows a vague tendency of the points.

Keywords: Taylor's power law, random graph, real network, empirical study.

1 Introduction

Our world consists of relationships. Think of the biological network that builds our bodies, social networks, or infrastructure networks. The abstraction of these complex systems is the same: they are made up of individuals with similar properties and the relationships between them.

However, they may differ in many characteristics. Consider one of the elementary properties of networks, the size. For instance, the network of the Zachary karate club had 34 members, see [1]. However, the Facebook's social network consisted of 2.41 trillion active members in June 2019. There was a change in the life of the Zachary karate club and the club split into two groups. Think of how this split may appear on Facebook. It is obvious that there are classes inside almost every network. E.g. in the case of social network, we can think of a family, a group of friends, or the population of a city or a country. What are the differences or similarities between these classes?

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



In 1961, ecologist Lionel Roy Taylor studied groups of different species, see [2]. The result of his research became famous as Taylor's power law, see [3]. The law was originally defined for ecosystems, with particular reference to the assessment of the spatial grouping of organisms. This is because Taylor studied the number and standard deviation of individuals of different species per unit area. The general Taylor's power law is the following.

Consider a population C and divide it into classes C_1, C_2, \dots, C_n . Consider also a quantity ξ that can be measured for each member of C . Inside class C_i , let μ_i and s_i be the quantity's mean value and variance, respectively. Then Taylor's power law states that

$$s_i \sim a\mu_i^b, \quad i = 1, \dots, n,$$

where a and b are (positive) constants. Does this law prevail in real networks and in random graph models? We deal with the empirical study of this problem.

For a network we can imagine the following version of Taylor's power law. Let C denote the nodes of a network, divide it into classes C_1, C_2, \dots, C_n so that inside each class the connections are strong, but between different classes the connections are relatively weak. Let the quantity ξ be the degree of the given node. We emphasize that we count connections only inside the given class. As above, let μ_i and s_i be the mean value and variance of the degrees inside class C_i . Then we want to show that $s_i \sim a\mu_i^b, i = 1, \dots, n$. To visualize this law, let $x_i = \log \mu_i, y_i = \log s_i, i = 1, \dots, n$. Then $s_i \sim a\mu_i^b$ means that

$$y_i = \log s_i \sim \log a + b \log \mu_i = A + Bx_i, \quad i = 1, \dots, n.$$

Therefore we should check linear relation. Equivalently, using log scales on both axes, the points $(\mu_i, s_i), i = 1, \dots, n$, should be around a straight line. This phenomenon will be discovered on our figures.

The results of our study will show that the fit to a straight line is perfect in certain cases, it is fair in several other cases and even in the worst case the straight line shows a vague tendency of the points. To show the goodness of the linear approximation in each case we calculate the correlation coefficient and the residual sum of squares, too.

2 The classification algorithm

Let us consider a network. The first task is to classify the nodes of the network. To accomplish this task, we use a classification algorithm called Label Propagation, see [4] and [5]. Here we give a short description of the algorithm.

In the initial state, we consider that each vertex of the network forms a group on its own. Going through the vertices one by one, the classification of each node will be determined by the following steps of the procedure, [5]. Let $C_x(t)$ denote the label of node x at time t .

- (1) Initialize the labels at all nodes in the network. For a given node x , let $C_x(0) = x$.
- (2) Set $t = 1$.

- (3) Arrange the nodes in the network in a random order and set it to X .
- (4) For each $x \in X$ chosen in that specific order, let

$$C_x(t) = f(C_{x_{i_1}}(t), \dots, C_{x_{i_m}}(t), C_{x_{i_{(m+1)}}}(t-1), \dots, C_{x_{i_k}}(t-1)),$$

where x_{i_1}, \dots, x_{i_m} are neighbours of x that have already been updated in the current iteration, while $x_{i_{(m+1)}}, \dots, x_{i_k}$ are neighbours that are not yet updated in the current iteration. The function f here returns the label occurring with the highest frequency among the neighbours. Select a label at random if there are multiple highest frequency labels.

- (5) If every node has a label that the maximum number of their neighbours have, then stop the algorithm. Else, set $t = t + 1$ and go to (3).

The code was implemented in Python.

```
for vertex in Graph.nodes():
    neighclasses=[classes[neighbor] for neighbor in G.neighbors(vertex)]
    freq=Counter(neighclasses)
    mx=max(freq.values())
    maxs=[k for k in freq.keys() if freq[k]==mx ]
    if freq[classes[vertex]]<mx:
        classes[vertex]=maxs[rn.randint(0,len(maxs)-1)]
```

The essence of the Label Propagation algorithm is the following. We take all the neighbours of any given vertex in turn and consider the classes of the neighbours. We count how many neighbours of the given vertex have in each class. Our given vertex is classified into the class in which it has the most neighbours. If the maximum number of the neighbours occurs in more than one class, then we choose uniformly at random one of those classes and assign our given vertex to that class.

```
result=[[1,1,1] for _ in range(numvertex)]
for v in range(numvertex):
    classv=classes[v]
    if v in G:
        degv=sum([classes[neighbor]==classv for neighbor in G.neighbors(v)])
        result[classv][0]+=degv
        result[classv][1]+=degv*degv
        result[classv][2]+=1
    ...
f=lambda x: [math.log(x[0]/(math.log(x[2])+1)),
             math.log(x[1]/(math.log(x[2])+1))]
```

Using this classification, we get strong relationship within a class and weak relationship outside the class. After completing the classification, each class is studied separately. Using function f , we calculate the internal degree of the members of the group, that is, we consider only the edges that run to another vertex of the same class. Then we calculate the empirical mean and empirical variance of the internal degrees within all groups. After performing the calculations, we plot the result using logarithmic scales on both axes.

Besides the graphical overview, some statistical indicators are also calculated. The Pearson correlation between the mean and standard deviation of the degrees is considered to characterize the linear relationship between the two quantities. By applying linear regression and using log scales, a straight line is fitted to the mean and standard deviation points, and then the RSS value (Residual Sum of Squares) is calculated, i.e., the mean of the squares of the distances of the points from the fitted straight line. If this value is relatively large, the points do not fit well on the line. We also observe Pearson's correlation coefficient mentioned above, which shows the goodness of fitting the points to the line. If the correlation is close to one, we experience a functional relationship, so the fit is good.

3 The experiments

First, the algorithm was tested for the well-known, classical case of the Zachary karate club, [1]. The club consisted of 34 members, the edges of the network represented the relationships that existed in the life outside the club. However, a conflict emerged in the life of the club, which led to the split of the club, known ways. Using the Label Propagation algorithm, the classification was done according to our reality-based expectation. We get two large groups, just as in reality the club split into two groups, and a small group, which can show a click, see Figure 1. The comparison of the mean degree and standard deviation within the resulting groups was also in line with our expectations, so although the graphical representation is not spectacular due to the small number of subclasses, we can still state that in the Zachary karate club network, Taylor's law is fulfilled. Plotting the mean degrees and standard deviations of the three groups on logarithmic scale, the three points fit to the regression line $y = -0.48 + 1.74x$. The Pearson correlation between the two criteria is 0.9889 so there is a strong positive linear relationship. The RSS value is 0.0004 so the points fit well to the regression line.

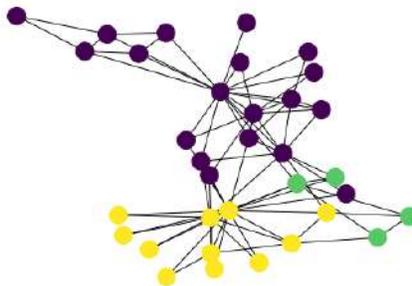


Fig. 1. Zachary karate club classification

3.1 Networks generated by classical models and random graph models

In this subsection we verify Taylor's law in networks generated by classical and random graph models. The built-in graph generators of the Python NetworkX package ([6], [7]) was used to generate the networks.

Firstly, we examined the case of the **balanced trees**. For simplicity, the generation was done with the parameters *Branchingfactor* = 10 and *Height* = 2. After the classification, only a few groups were detected, so the graphical representation is not spectacular. The mean degree - standard deviation points on the double logarithmic scale fit to the regression line $y = -1.39 + 1.2.46x$. The Pearson correlation of the two variables is 0.9886 so there is a strong positive linear relationship. The RSS value is 0.0022.

A **barbell graph** is a non-directed graph consisting of two non-overlapping n -vertex cliques connected by a single edge. The classification was done according to our expectations, the obtained mean degree - standard deviation points fit to the regression line $y = -1.9 + 1.84x$. The correlation shows a strong linear relationship in this case as well, its value is 0.9905, the RSS is 0.0055.

In the case of the **Dorogovtsev-Goltsev-Mendes hierarchically constructed graph**, see [8], we start from a triangle formed by three nodes. When a new node connects to the graph, an existing edge is selected randomly, and the new vertex links to both endpoints on that chosen edge. The classification formed a sufficient number of groups to obtain a graphically evaluable result. Thus, on the x axis we indicate the mean of the degrees inside the classes, and on the y axis we represent the corresponding standard deviations, on both axes using logarithmic scale. A straight line is fitted to the points by linear regression. The obtained lines, the corresponding r correlation coefficients, and the RSS values are the following.

Figure 2: $y = -0.56 + 1.79x$, $r = 0.9934$, RSS=0.0012

Figure 3: $y = -0.59 + 1.81x$, $r = 0.9933$, RSS=0.0025

Figure 4: $y = -0.61 + 1.81x$, $r = 0.9958$, RSS=0.0046

Figure 5: $y = -0.66 + 1.85x$, $r = 0.9968$, RSS=0.0043

Then, we studied the well-known **Erdős-Rényi model**, see [9]. In this case, we choose one of all the graphs with vertices n and edges m according to uniform distribution. The generation of the networks was performed with different parameters. The study was started with small graphs with 50 nodes and 150 edges, then we increased the network size to 100 nodes and 250 edges, 500 nodes and 1250 edges, 5000 nodes and 15000 edges, and finally 10000 nodes and 30000 edges as parameters. The result of the graphical overview is shown in Figure 6. Regarding the statistical metrics, similar values were obtained in all cases, the correlation for the largest graph examined is 0.9924, which shows a strong linear relationship. However, the RSS values are high, the value for the largest graph examined is 0.0045, indicating that the points do not fit well on the line, showing a larger scatter around the regression line, as the figure also shows. Next, the well-known **preferential attachment model** was examined, see [10]. We generated the networks with vertices of 1000, 5000 and 10000 using parameter $m = 2$. After the classification, the graphical survey of

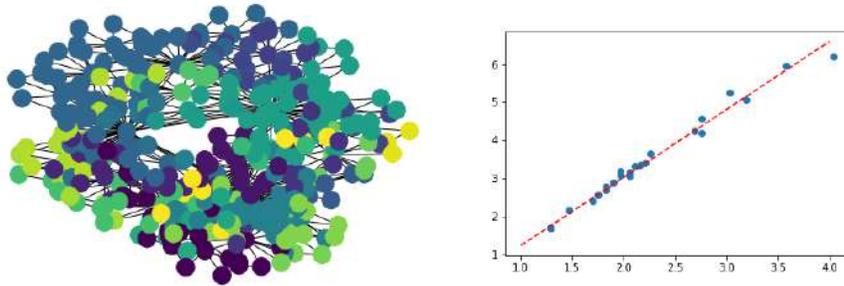


Fig. 2. Classification of the 500-node Dorogovtsev-Goltsev-Mendes network and graphical study of Taylor's law

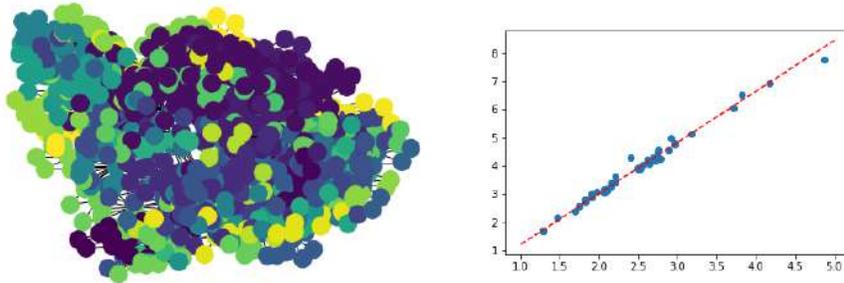


Fig. 3. Classification of the 1000-node Dorogovtsev-Goltsev-Mendes network and graphical study of Taylor's law

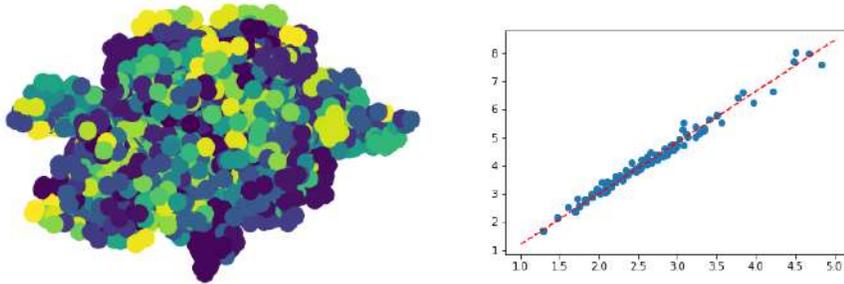


Fig. 4. Classification of the 2000-node Dorogovtsev-Goltsev-Mendes network and graphical study of Taylor's law

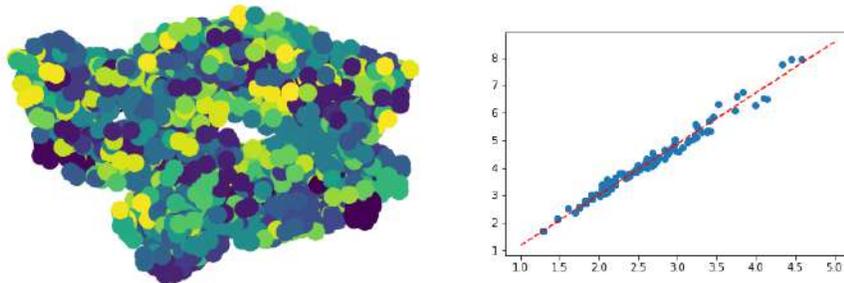


Fig. 5. Classification of the 3000-node Dorogovtsev-Goltsev-Mendes network and graphical study of Taylor's law

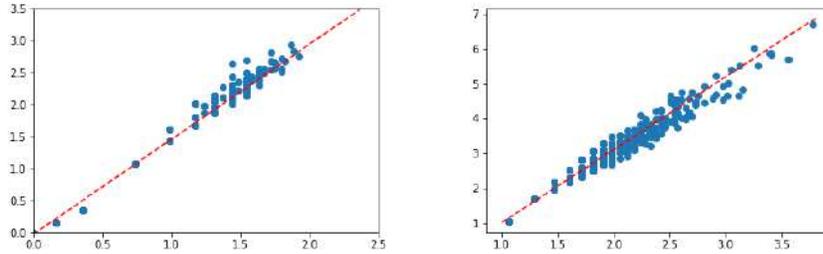


Fig. 6. Study of Taylor’s law in the Erdős-Rényi graph with 10000 vertices and 30000 edges (on the left) and in the Barabási-Albert graph with 10000 vertices (on the right)

Taylor’s law showed the result which can be seen on the Figure 6. It indicates that the points are more scattered around the straight line. The line fitted by linear regression is $y = -1.1 + 2.1x$. The correlation coefficient is above 0.99 in all three cases, so the linear relationship is strong. However, the RSS values are relatively high, 0.0087, 0.0058, and 0.0063, suggesting that the points do not fit well on the line, showing a larger scatter around the regression line, see Figure 6 on the left.

Next, we generate networks with the **Holme-Kim algorithm**, see [11], which is similar to the model of Barabási. This algorithm differs from the Barabási-Albert model in that every step after the new vertex connects to the network, it also associates to a neighbour of the selected vertex with probability p , thus creating a triangle. For the generated 20000-node network, the regression line is $y = -0.81 + 0.88x$, see Figure 7. The correlation coefficient is 0.9966. The RSS is 0.0039, thus the scatter around the line can be considered medium. We studied also the **Newman-Watts-Strogatz** and **Watts-Strogatz small-**

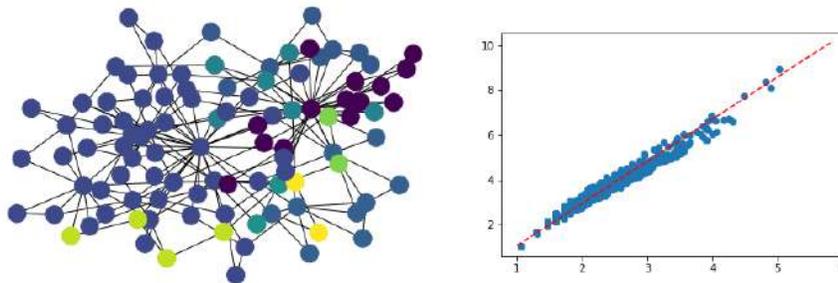


Fig. 7. Classification of the Holme-Kim network and the graphical study of Taylor’s law in a graph with 20000 vertices

world models, see [12]. For the Newman-Watts-Strogatz graph, we create a ring with n vertices, in which each vertex is connected to its k nearest neighbours. Then, for every existing (u, v) edge, we randomly select a vertex w and create a new (u, w) edge with probability p . For the Watts-Strogatz model, the

selected (u, v) edge is deleted from the graph. In both cases our classification divided the networks into just a few groups, so the graphical representation is not spectacular, but the trend according to Taylor’s law can still be observed. The linear relationship is strong, the correlation is about 0.99. The value of RSS is 0.003.

For the **duplication-divergence graph** model, see [13], a graph with n vertices is created by copying the initial nodes and keeping the edges associated with the original nodes with probability p . We examined networks with 100, 1000, and then 5000 vertices, and the graphical overview shows that Taylor’s law is fulfilled. The correlation is strong in all cases, being above 0.99 and the RSS value is moderately low, being around 0.003. The fitted line for the largest graph is $y = -0.01 + 1.54x$, see Figure 8.

In the case of the **random-lobster graph**, the network is a tree that reduces to a caterpillar when pruning all leaf nodes. A caterpillar is a tree that reduces to a path graph when pruning all leaf nodes. Networks with 200, 400, 2000, 4000, 7000 nodes were analysed. The graphical examinations in these cases are also in accordance with Taylor’s law. The correlation is strong in all cases, being above 0.99, the RSS value is low, being 0.0021 for the largest graph examined. The regression line for the largest graph is $y = -0.02 + 1.44x$.

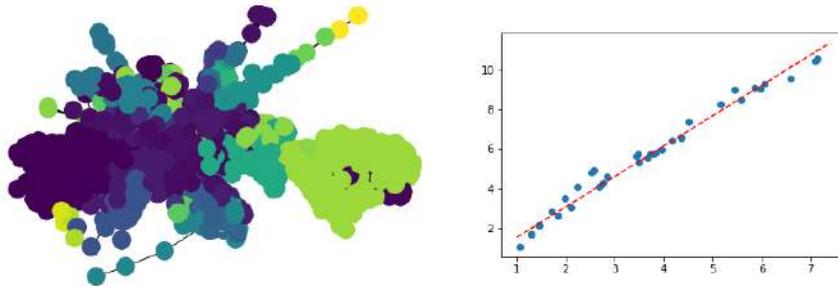


Fig. 8. Classification of the Duplication-Divergence network and the study of Taylor’s law in a 5000-node graph

3.2 Real-life networks

The main objective of our work is to analyse real-life networks from the perspective of Taylor’s law. The networks were downloaded from the Network Repository, see [14], public database at *www.networkrepository.com*. The networks are stored in edge list format, which can be imported to our Python program for analysis using the NetworkX package.

Social networks Firstly, animal social networks were analyzed, then we continued our research with human social networks. As Lionel Roy Taylor did, we first examined networks and groups of animals. In the first case, the nodes of the network from [14], [15] are **birds**. The edges of the network extend

between individuals who, as members of a colony, used the same nest chamber during a given period at the same time, together, or sequentially. The network consists of 18 vertices and 28 edges. Despite its small size, Taylor's law can still be observed. The correlation between the mean degree and the standard deviation of the degree is 0.9959, which shows a strong linear relationship. The regression line is $y = -0.87 + 1.92x$, see Figure 9. The value of the RSS is 0.0002, so the points fit the regression line very well.

The social network of **bottlenose dolphins** from [14], [16] consists of 62 individuals, among which 159 regularly observed associations were detected. The line fitted to the points is $y = -0.65 + 1.77x$, see Figure 9. The correlation is 0.9956, which shows a strong linear relationship. The RSS value is 0.0015, which is very low, so the fit is good.

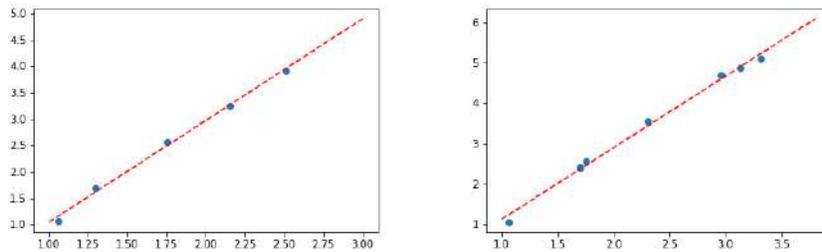


Fig. 9. Graphical study of Taylor's law in the network of nesting birds (on the left) and bottlenose dolphins (on the right)

We examined the '**who voted who**' network of the Wikipedia from [14], [17], [18]. The database contains all votes from the beginning of the Wikipedia to January 2008. The vertices of the network are Wikipedia users, and the edges represent the votes. We have 2900 votes of 889 users for analysis. The correlation is 0.9901. The regression line is $y = -0.45 + 1.58x$, see Figure 10. The value of RSS is 0.0015, so we can say that the points fit well on the regression line.

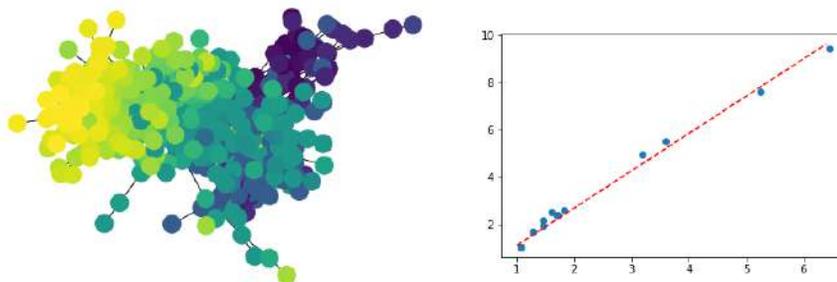


Fig. 10. Classification of the Wikipedia voting network and examination of Taylor's law

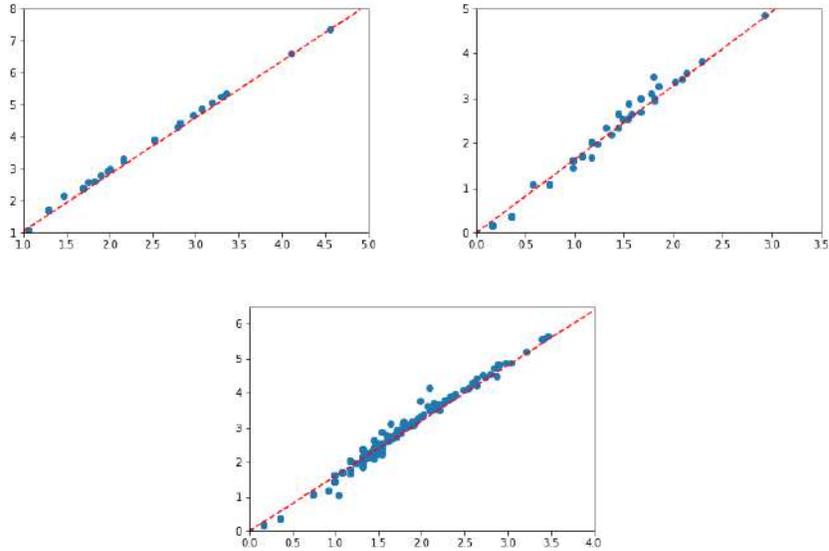


Fig. 11. The study of Taylor’s law in the social networks of Hamsterster (on the top left), Anybeat (on the top right) and Brightkite (at the bottom)

Next, we analysed the **Hamsterster** social network from [14], [19], the network of the website [20]. The nodes in the network are people, and the edges represent family and friendships. The network contains 2400 nodes and 16600 edges. The correlation is strong, 0.9926. The regression line is $y = -0.03 + 1.58x$, see Figure 11. RSS shows a moderately good fit, with a value of 0.0037.

Then we took a review of **Anybeat’s online community** from [14], [21], where users have the opportunity to interact with people near them, or anyone, anywhere in the world, see [22]. The network consists of 12600 nodes and 67100 edges representing interactions. The correlation is 0.9865. The fitted line is $y = -0.9 + 1.93x$, see Figure 11. The RSS value is low, 0.0011.

The **Brightkite network** from [14], [23] is a location-based social network with 58200 nodes and 214100 edges. After the analysis, the obtained regression line is: $y = -0.01 + 1.52x$, see Figure 11. The correlation is 0.9902. The RSS value is very low, 0.0014, so the fit is good.

Then we examined collaboration networks. The **Erdős network** of scientific collaborations from [14], [24] represents 7500 collaborative relationships between 6100 authors, see [25]. The regression line is $y = -0.02 + 1.84x$ (see bottom left on Figure 12). The correlation is 0.9932 and the RSS value is 0.007, which is very high, so the points do not fit to the line well. We also experience scattering in the graphical representation.

Arxiv High Energy Physics collaboration network from [14], [26] consists of 12000 nodes and 118500 edges. The regression line is $y = -0.01 +$

$1.61x$ (see top left on Figure 12). The correlation is 0.9968, which shows a good fit. RSS is 0.0028.

Arxiv Condensed Matter's collaborative network from [14], [27] consists of 25100 vertices and 93400 edges. The regression line is $y = -0.01 + 1.68x$ (see top right on Figure 12). The correlation is 0.9920. The RSS value is 0.0142, which shows a high mean squared deviation from the regression line. Last in

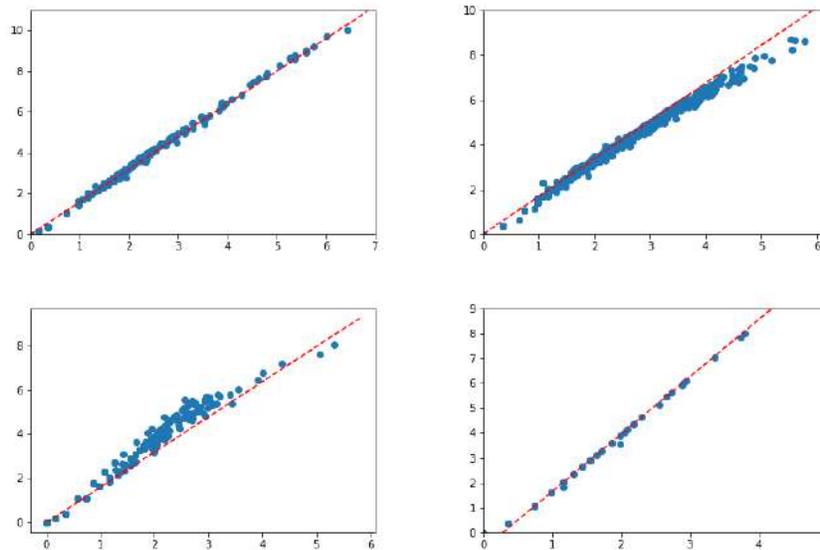


Fig. 12. The study of Taylor's law in the Arxiv High Energy Physics (on the top left), Arxiv Condensed Matter's (on the top right), Erdős (at the bottom left) collaboration networks and the e-mail network (at the bottom right)

the category of social networks, we studied the **network of electronic mails** from [14], [28]. 54400 e-mails were sent among 32400 users. The graphical representation shows the fulfilment of Taylor's law very nicely. The correlation is 0.9902. The regression line is $y = -0.65 + 2.3x$ (see bottom right on Figure 12). The RSS value is 0.0002, which shows a very good fit.

Biological networks In this section, different biological networks are studied.

The networks represent gene-functional associations. All the considered cases show compliance with Taylor's law, however, the figures are not always spectacular due to the small number of groups.

In the graphical study, the clearest result was obtained for the **human gene regulatory network** from [14], [29], see [30]. The human gene regulatory network from the analysis of gene expression profiles consists of 14000 vertices and 9 million edges. The correlation coefficient is 0.9995. The regression line is $y = 0.002 + 2.12x$, see Figure 13. The RSS value is 0.0012, which shows a very good fit.

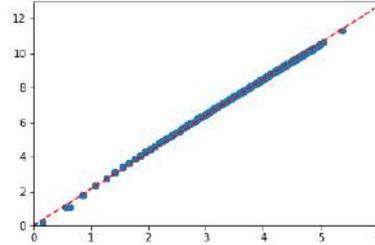


Fig. 13. The study of Taylor’s law in the human gene regulatory network

Infrastructure networks In this section, we performed an empirical study of different infrastructure networks.

In the first case, the graph represents the topology of the **electricity grid** in the western states of the United States from [14], [31], see [32]. It consists of 4900 vertices and 6600 edges. The regression line is $y = -0.3 + 1.54x$ (see top left of Figure 14). The correlation is 0.9945 and the RSS value is 0.0046, so we experience scattering around the regression line.

Aviation network from [14], [33] was downloaded from openflights.org, contains links between two non-US-based airports, see [34]. It consists of 2900 vertices and 30500 edges. The fitted line is $y = -0.03 + 1.51x$ (see top right of Figure 14). The correlation is 0.99 and the RSS value is 0.026.

Then we examined the **road network in Luxembourg** from [14], [35], which is made up from 114600 junctions and 119700 roads. The regression line is $y = -0.03 + 1.48x$ (see bottom left of Figure 14). The correlation coefficient is 0.9909 and the RSS value is 0.0056, which is high, predicting that we experience a scatter around the regression line, the points do not fit well.

Finally, we study a subnet of Amazon from [14], [36], that includes users and products, and the relationship between them embodies evaluation, see [37]. This is a bipartite graph. The network consists of 2.1 million nodes and 5.8 million edges. The fitted regression line is $y = -0.004 + 2.01x$ (see bottom right of Figure 14). The correlation is 0.9987, the RSS value is 0.017, which is high. We find that the points are scattered below the fitted line.

4 Summary

The aim of this empirical study was to observe the fulfillment of Taylor’s power law in artificially generated and real networks. To implement the study, we first had to discover the community structure of the networks by applying the Label Propagation algorithm. Then we considered the empirical mean and empirical standard deviation of the degrees within the classes. These value pairs were interpreted by visualizing on double logarithmic scale. The linear relationship between the values was characterized by determining Pearson’s correlation coefficient. By linear regression, a straight line was fitted to the observed points. The accuracy of the fit to the regression line was also certified by calculating the Residual Sum of Squares index.

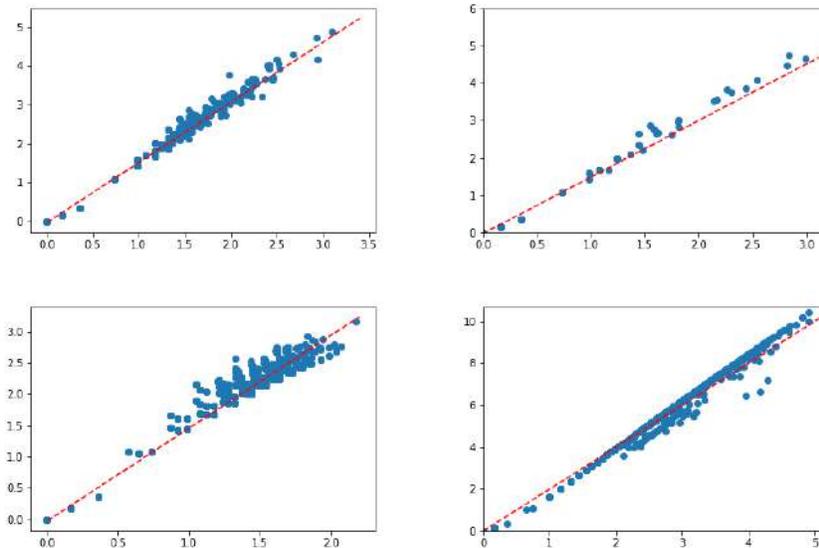


Fig. 14. The study of Taylor’s law in the US electricity grid (on the top left), the aviation grid (on the top right), the Luxembourg road network (at the bottom left) and the Amazon evaluation network (at the bottom right)

We found that in each of the considered networks there is a strong linear relationship between the empirical mean degree and the empirical standard deviation. The smallest Pearson correlation value is 0.9818. This relationship is often functional, in some cases more of a stochastic relationship. The slopes of the regression lines are between 1.5 and 2.5, usually around 2. The RSS value ranged from 0.0002 to 0.0142. Considering the graphical representation as well, we found that we have very good matches up to roughly 0.004 RSS. This occurred in 20 cases out of 35. In the other cases, we experienced a larger scatter around the straight line. In only a few cases were there greater deviation from the straight line, in which case Taylor’s law is not exactly fulfilled.

This work was supported by the EFOP-3.6.3-VEKOP-16-2017-00002.

References

1. W. W. Zachary. An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research*, 33, 452–473, 1977.
2. L. R. Taylor. Aggregation, variance and the mean. *Nature*, 189, 732–735, 1961.
3. J. N. Perry. Taylor’s Power Law for Dependence of Variance on Mean in Animal Populations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 30, 3, 254–263, 1981.
4. U. N. Raghavan, R. Albert and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76, 036106, 2007.
5. en.wikipedia.org/wiki/Label_propagation_algorithm

6. A. A. Hagberg, D. A. Schult and P. J. Swart. Exploring Network Structure, Dynamics, and Function using NetworkX. *Proceedings of the 7th Python in Science conference (SciPy 2008)*, G. Varoquaux, T. Vaught, J. Millman (Eds.), 11–16, 2008.
7. <https://networkx.github.io>
8. S.N. Dorogovtsev, A.V. Goltsev and J.F.F. Mendes. Pseudofractal Scale-free Web. *Phys. Rev. E* 65, 066122, 2002.
9. P. Erdős and A. Rényi. On Random Graphs. *Publ. Math.*, 6, 290 1959.
10. A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286, 509–512, 1999.
11. P. Holme and B. J. Kim Growing. Scale-free networks with tunable clustering. *Phys. Rev. E*, 65, 026107, 2002.
12. D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393, 440–442, 1998.
13. I. Ispolatov, P. L. Krapivsky and A. Yuryev. Duplication-divergence model of protein interaction network, *Phys. Rev. E*, 71, 061911, 2005.
14. Ryan A. Rossi and Nesreen K. Ahmed. The Network Data Repository with Interactive Graph Analytics and Visualization. AAAI, <http://networkrepository.com>, 2015.
15. <http://networkrepository.com/aves-weaver-social-00.php>
16. <http://networkrepository.com/mammalia-dolphin-social.php>
17. J. Leskovec, D. Huttenlocher and J. Kleinberg. Signed networks in social media, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1361–1370, 2010.
18. <http://networkrepository.com/soc-wiki-Vote.php>
19. <http://networkrepository.com/soc-hamsterster.php>
20. Hamsterster, Hamsterster social network, <http://www.hamsterster.com>
21. <http://networkrepository.com/soc-anybeat.php>
22. M. Fire, R. Puzis, and Y. Elovici. Link Prediction in Highly Fractional Data Sets, *Handbook of Computational Approaches to Counterterrorism*, Springer, 2012.
23. <http://networkrepository.com/soc-loc-brightkite.php>
24. <http://networkrepository.com/ca-Erdos992.php>
25. V. Batagelj and A. Mrvar. Some analyses of Erdos collaboration graph, *Social Networks*, 22, 2, 173–186, 2000.
26. <http://networkrepository.com/ca-HepPh.php>
27. <http://networkrepository.com/ca-CondMat.php>
28. <http://networkrepository.com/email-EU.php>
29. <http://networkrepository.com/bio-human-gene2.php>
30. M. Bansal, V. Belcastro, A. Ambesi-Impiombato and D. Di Bernardo. How to infer gene networks from expression profiles. *Molecular systems biology*, 3, 1, 2007.
31. <http://networkrepository.com/inf-power.php>
32. D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393, 6684, 440–442, 1998
33. <http://networkrepository.com/inf-openflights.php>
34. T. Opsahl. Why anchorage is not (that) important: Binary ties and sample selection. <https://toreopsahl.com/2011/08/12/why-anchorage-is-not-that-important-binary-ties-and-sample-selection/>, 2011
35. <http://networkrepository.com/road-luxembourg-osm.php>
36. <http://networkrepository.com/rec-amazon-ratings.php>
37. A. Mukherjee, B. Liu and N. Glance. Spotting fake reviewer groups in consumer reviews. *Proceedings of the 21st international conference on World Wide Web*, 191–200, 2012.

Health Care Need Adjusted Prospective Old-age Dependency Ratio in Selected European Countries

Tomáš Fiala¹, Jitka Langhamrová², and Jana Vrabcová³

¹ Department of Demography, Faculty of Informatics and Statistics, Prague University of Economics and Business, nám. W. Churchilla 4, 130 67 Praha 3, Czech Republic
(E-mail: fiala@vse.cz)

² Department of Demography, Faculty of Informatics and Statistics, Prague University of Economics and Business, nám. W. Churchilla 4, 130 67 Praha 3, Czech Republic
(E-mail: langhamj@vse.cz)

³ Department of Demography, Faculty of Informatics and Statistics, Prague University of Economics and Business, nám. W. Churchilla 4, 130 67 Praha 3, Czech Republic
(E-mail: jana.langhamrova@vse.cz)

Abstract. One of the not so often mentioned consequence of the population ageing is the expected increase of financial burden of health care systems. Standard and commonly used simple indicators of this burden are usually based on the relation of the number of old (or the oldest old) persons and the number of persons in productive age. The threshold of old age is very often determined as the age of 65 years. But many latest studies propose that due to the permanent increase of life expectancy the definition of the old age threshold should be more likely based on the expected remaining length of life, on the expected time to death. According to some analyses, especially during the last five years of the life of old-age persons the need of health care is relatively high. Possible indicator of the health care financial burden of this type can be e.g. the Health care need adjusted prospective old-age dependency ratio. It is determined as the proportion of the expected number of old age persons having time to death lower than 5 years and the number of persons in productive age where the upper threshold of productive age is defined using not the chronological but the prospective age.

The paper contains the development of the values of the Health care need adjusted prospective old-age dependency ratio and other alternative indicator of ageing in selected European countries in the period since 1950 until 2100. The estimates of future values are based on the latest Eurostat population projection.

Keywords: Population ageing, Prospective age, Time to death, Old-age dependency ratio, Health care need adjusted prospective old-age dependency ratio.

1 Introduction

Population ageing is one of the most important phenomena of the latest decades of the previous and of course of this century. This topic is very often discussed not only by demographers and economists, but also by general public and media. All demographically advanced populations face and will face the ageing process, which is, in many aspects, unprecedented and dynamically deepening. Despite the continuing changes in demographic development, especially the

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



continuing increase of the life span and also of the healthy life span, most of the analytical approaches to the population ageing phenomena use conventional tools and standard indicators and measures, based usually on the assumption that the threshold of old age is in each population fixed, unchanged in time, equal usually to 65 years of age, which is the retirement age in many countries. But in fact, already in 1975 Ryder has published the idea of re-examination of the concept of fixed threshold of old age. "*We measure age in terms of the number of years elapsed since birth. This seems to be a useful and meaningful index of the stages of development from birth to maturity. Beyond maturity, however, such an index becomes progressively less useful as a clue to other important characteristics. To the extent that our concern with age is what it signifies about the degree of deterioration and dependence, it would seem sensible to consider the measurement of age not in terms of years elapsed since birth but rather in terms of the number of years remaining until death.*" (Ryder [7], p. 16.) He suggested the idea to consider the point of entry into old age as the value of age at which the life expectancy is equal to a given, relatively low, value, say, e.g. 10 years.

The idea of flexible old age threshold depending on the remaining life expectancy was suggested later also e.g. by Fuchs [3]. Siegel [11] proposed the idea of the old-age threshold defined as the age when the life expectancy equals not to 10 but to 15 years. Wachter used for the time until deaths the term thanatological¹ age (Riffe [6]).

This idea of alternative concept of definition of human's age and new indicators based of this forward-looking conception was minutely treated by Sanderson and Scherbov in several papers. They introduced a new forward-looking definition of age (so called prospective age) and argued that, along with the traditional backward-looking concept of age, this definition provides a more informative basis to discuss population aging (Sanderson and Scherbov [8]).

The indicators of ageing based on prospective age instead of biological age threshold of old age show that the increase of many indicators of ageing in such a case will not be so dramatic in comparison with indicators defined by standard way (Sanderson and Scherbov [9], [10]). Calculations of some prospective indicators for Czech population were published by Klapková *et al.* [5]. An overview of prospective indicators presents e.g. Šprocha [13].

The aim of the paper is to present some new approaches to the analysis of population ageing using indicators based on the concept of prospective old-age threshold and to compare the values of new indicators with the values of corresponding standard ones. The alternative threshold of old-age has been chosen as the age at which the remaining life expectancy reaches the value 20 years which is close to the value of the remaining life expectancy at the age of 65 years in many European countries at present time. The paper brings results not only for the past period (since 1950 until 2017) but also prospects of the future development until 2100 according to the latest population projection of Eurostat.

¹Thanatos was the Greek god of death

2 Methodology and Data

While the standard old-age threshold (or the upper bound of productive age) usually used is 65 years of age, the “alternative” threshold has been chosen as the age at which the remaining life expectancy reaches the value 20 years. As in the standard case, the threshold should be the same for both males and females. Because of this the unisex life expectancy has been used for calculations and the old-age threshold $x(t)$ in the year t has been defined as the value fulfilling the equation

$$e_{t,x(t)} = 20, \quad (1)$$

where $e_{t,x}$ denotes the unisex life expectancy in the year t at the age x (average value of the life expectancy of males and females). Usually there exists no integer value $x(t)$ fulfilling (1). Linear interpolation method has been used for more precise determination of the value of the prospective old-age threshold

$$x(t) = x_0(t) + \frac{e_{x_0(t)} - 20}{e_{x_0(t)} - e_{x_0(t)+1}}, \quad (2)$$

where $x_0(t)$ denotes highest integer value of age at which the life expectancy is higher than or equal to 20 years while in the age of $x_0(t)+1$ the value of life expectancy is lower than 20 years. If the life expectancy is rising in time the value of the prospective age is increasing, too.

While the standard proportion of old-age persons is defined by the usual way

$$s_{65+} = \frac{S_{65+}}{S}, \quad (3)$$

where S_{65+} is the number of persons at the age 65 or older, S is the total population size, the prospective proportion means the proportion of persons at the age equal to $x(t)$ or older (i.e. proportion of persons having life expectancy equal or lower than 20 years). Calculation formula makes use of linear interpolation method again

$$s_{x(t)+} = \frac{S_{x(t)+}}{S} = \frac{(1 - (x(t) - x_0(t))) \cdot S_{x_0(t)} + \sum_{x=x_0(t)+1}^{\omega-1} S_x}{S}. \quad (4)$$

where $\omega-1$ denotes the highest value of complete age in the population.

Assuming that the lower threshold of productive age is 20 years (which is at present more realistic value than 15 years used in the past), the standard definition of the old age dependency ratio is rather simple

$$OADR = \frac{S_{65+}}{S_{20-64}}. \quad (5)$$

while the prospective type is calculated by the following way

$$POADR = \frac{S_{x(t)+}}{S_{20-x(t)}} = \frac{(1 - (x(t) - x_0(t))) \cdot S_{x_0(t)} + \sum_{x=x_0(t)+1}^{\omega-1} S_x}{\sum_{x=20}^{x_0(t)-1} S_x + [x(t) - x_0(t)] \cdot S_{x_0(t)}} . \quad (6)$$

A very simple indicator of population ageing may be the average age of the population

$$\bar{x} = \frac{\sum_{x=0}^{\omega-1} S_x \cdot (x+1/2)}{S} . \quad (7)$$

The natural corresponding prospective counterpart indicator is the average age of remaining life of the population called PARYL – population average remaining years of life

$$PARYL = \frac{1}{2} \cdot \sum_{x=0}^{\omega-1} S_x \cdot (e_x + e_{x+1}) . \quad (8)$$

The formula reflects the fact that the life expectancy e_x means the average remaining lengths of life of a person of exact age x , not completed age.

One of the most often mentioned consequences of the population ageing is the increasing financial burden of pension system. Not so often mentioned but very important challenge is the expected increasing financial burden of health care systems.

The old-age dependency ratio and prospective old-age dependency ratio, respectively, may be a very rough indicators mainly of the financial burden of the pension system. Unlike the pension system where the height of the annuity is not changing (or is slightly increasing by inflation adjustment), health expenditures increase dramatically on average at the end of life. At the same time, these expenditures depend more on remaining lifetime than on calendar age, at least beyond the age of 65+ (Zweifel *et al.* [14]).

Spijker *et al.* [12] propose the elderly population having a time to deaths (thanatological age) less than 5 years to be the population of acute health care needs. The ratio of this population size to size of population in productive age – called Health care need old-age dependency ratio (OADR5TTD) and Health care need adjusted prospective old-age dependency ratio (POADR5TTD), respectively are more appropriate indicators of health care expenditures.

Number of persons at the (complete) age x having time to deaths (thanatological age) lower than 5 years can be calculated by the formula

$$S_{x,TTD<5} = \frac{1}{2} \cdot S_x \cdot \left[\left(1 - \prod_{i=x}^{x+4} p_i\right) + \left(1 - \prod_{i=x+1}^{x+5} p_i\right) \right] , \quad (9)$$

where p_i is the life table survival probability at the exact age i .

The standard health care need old-age dependency ratio is then

$$OADR5TTD = \frac{S_{65+,TTD<5}}{S_{20-64}} = \frac{\sum_{x=65}^{\omega-1} S_{x,TTD<5}}{S_{20-64}}, \quad (10)$$

while the alternative form using prospective threshold of old-age has a little bit complicated formula

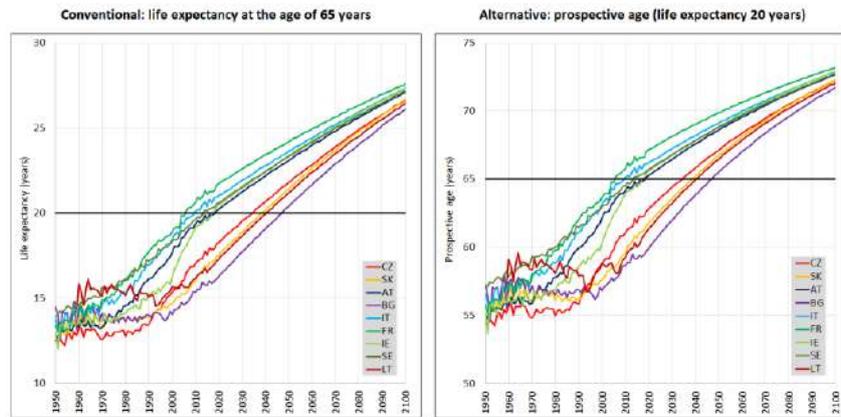
$$POADR5TTD5 = \frac{S_{x(t)+,TTD<5}}{S_{20-x(t)}} = \frac{(1-(x(t)-x_0(t))) \cdot S_{x_0(t),TTD<5} + \sum_{x=x_0(t)+1}^{\omega-1} S_{x,TTD<5}}{\sum_{x=20}^{x_0(t)-1} S_x + [x(t)-x_0(t)] \cdot S_{x_0(t)}}. \quad (11)$$

The analysis has been carried out for the following 9 European countries: Czechia, Slovakia, Austria, Bulgaria, Italy, France, Ireland, Sweden and Lithuania. The analysis covers the period since 1950 until 2100. Data for the previous period (1950–2017) came from Human Mortality Database – HMD [4]: population by age and sex as of 1st January, age-specific mortality rates and life expectancies by sex and age form period life tables. (For Lithuania there were available data only since 1959.) Analogical data for the future period 2018–2100 were taken over from the baseline variant of the latest population projection created by Eurostat [1].

3 Main results

The value of the life expectancy at the age of 65 years was about 12.5–14.5 years in 1950. Due to long-term mortality stagnation in post-communist countries the value in 1990 in these countries was only several months higher, in Bulgaria even lower than in 1990 while in other countries the increase reached several years. This gap remained until present times. In 2017 the value of life expectancy at 65 years in France was higher than 21 years while in Bulgaria it was lower than 16 years. The mortality scenario of the Eurostat population projection is of convergent type assuming that until 2100 the difference in life expectancy at 65 years between analysed countries should diminish to 1.5 years, nevertheless the value in post-communist countries is supposed to remain lower than in other countries (Fig. 1; right graph).

The value of prospective threshold of the old age (the age at which the remaining life expectancy equals to 20 years) was about 55–57 years in 1950. At present time there is apparent a remarkable gap between values for post-communist countries and other countries. While at present times the value of prospective threshold of old age is for Austria, Italy, France, Ireland and Sweden about 65–67 years, for Czechia, Slovakia, Bulgaria and Lithuania reaches only about 60–63 years. The level of 65 years will be in post-communist countries reached probably as late as in the second half of 30th, in Bulgaria even in late 40th. Until the end of this century, the value should in all countries reach about 72–73 years (Fig. 1; right graph).



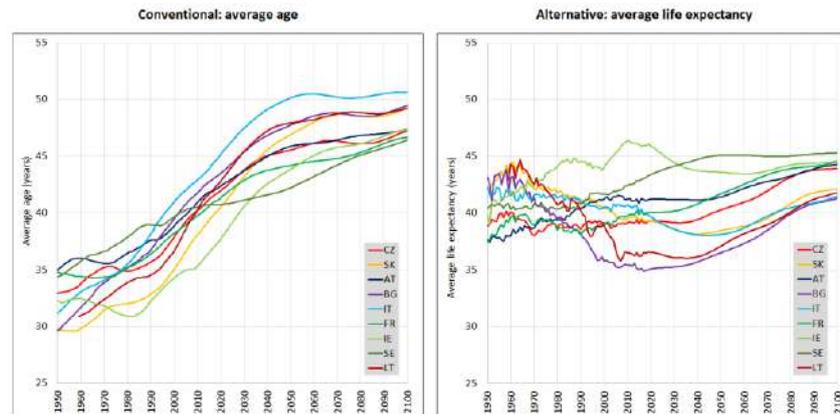
Source of data: HMD (until 2017), Eurostat population projection (since 2018), authors' calculations

Fig. 1. Life expectancy at 65 years and prospective age

One of the simplest indicators of population ageing is the average age of the population. In 1950 its values were between 29–35 years. During next decades, there was a relatively rapid increase of average age in some countries. Until 1990, its value in Bulgaria and Italy grew by about 7 years, on the other hand due to high fertility the value in Ireland was almost the same like in 1950. Until present times the average age continued to grow, its values in 2017 were approximately between 40–43 years except Ireland (37 years) and Italy (44.5 years). Because of assumed stable (or even slowly growing) fertility and slowing increase of life expectancy, the average age is supposed to be relatively stable in the second half of this century. In 2100 the average age in Sweden, France, Ireland, Czechia and Austria is expected to be about 47 years while in Lithuania, Slovakia and Bulgaria a little bit over 49 years and in Italy almost 51 years of age (Fig. 2; left graph).

Despite relatively rapid increase of the average age of the population the average value of life expectancy, i.e. the value of average remaining life span of the population is not decreasing so much and, in some countries, or periods, its values have even growing tendency. In 1950, its value was about 37–43 years, so it was a little bit higher than values of average age. Until 1990, the value decreased in all post-communist countries as well as in Italy, on the other hand there was a growth in the remaining countries, in Ireland even almost by 5 years. Until 2017 the trends were similar, there was a little bit increase also in Czechia. In the period projected the average life expectancy is supposed to continue to grow slowly than so far except Ireland. In this country, the average life expectancy would drop a little bit because of decrease of fertility in the 80th and 90th of the previous century. Due to decrease of mortality, the values of average remaining life expectancy in France, Austria, Czechia, Sweden and Ireland would be in 2100 about 4-7 years higher than in 1950 and would reach

4445 years. On the contrary values in Bulgaria, Latvia² and Italy would be 1–2 years lower, Slovakia will have in 2100 approximately the same value as in 1950, i.e. about 4142 years (Fig. 2; right graph).



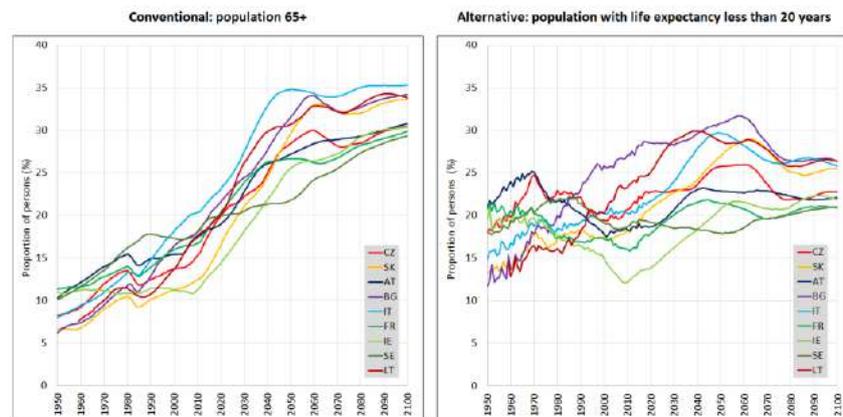
Source of data: HMD (until 2017), Eurostat population projection (since 2018), authors' calculations
 Fig. 2. Average age and average remaining life expectancy

Indicator of population ageing very often used is the proportion of elderly population. Standard threshold of old age is usually 65 years. In 1950, the proportion of persons at the age of 65 and over was especially in post-communist countries very low: about 6–7% in Bulgaria and Slovakia, about 8% in Czechia and also in Italy, while in other countries it reached 10–11.5%. Until 1990, the proportion grew in all countries by several percentage points to 10–15%. In 2017 the lowest proportion of seniors was in Ireland (only about 13.5%), in Slovakia (15%) while the values in other countries have reached about 18–22%. The growths should continue relatively rapidly during the first half of the present century, it would slow down in the second part. Until the end of this century the proportion of persons 65 years and older should reach in Sweden, France, Ireland, Czechia, and Austria about 29–31%, in Slovakia, Latvia and Bulgaria about 34% and in Latvia more than 35% of the total population (Fig. 3; left graph).

Using alternative definition of old-age threshold (age at which the life expectancy equals 20 years) the values and trends of development of the proportion of elderly population are quite different. While in 1950 the prospective old-age threshold is in all countries about 55 years, until 2100 it grows to about 72 years and the threshold differs among analysed countries (Fig. 1, right graph). In 1950 the lowest proportion of elderly population was in Bulgaria (under 12%), the highest value – 21% – had Austria. Until 1990 due to increase of average age and stagnation of mortality the proportion in Bulgaria grew to 23%, which was the highest value of all countries. The lowest value – a little bit over 16%, was reached in Ireland. Because of decrease of mortality,

² Comparing values in 1959 and 2100

the proportion of prospective elderly population in France, Ireland and Austria was in 1990 lower than in 1950. The drop of proportion of elderly population continued until 2017 in Ireland and Austria, the value for Sweden was also a little bit lower than in comparison with 1990. The proportion of elderly population remains until 2017 the highest in Bulgaria (almost 29%), the lowest in Ireland (only about 13.5%). Values in other countries reached about 18–22%. Until 2100 the values should slightly grow. While in France, Sweden, Ireland, Austria and Czechia they would be about 21–23%, in the remaining countries they should reach about 25–26% (Fig. 3; right graph).

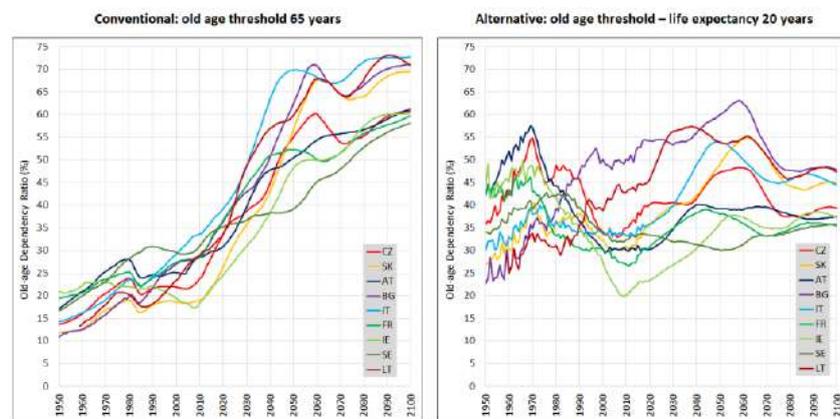


Source of data: HMD (until 2017), Eurostat population projection (since 2018), authors' calculations
 Fig. 3. Proportion of elderly population

The old-age dependency ratio is a very rough indicator of the financial burden of the pension system based on the PAYG principle. The standard limits of productive age are at present times usually 20 and 65 years, respectively. The values of the ratio in analysed countries differed in 1950 relatively high. While in Bulgaria the value of the ratio was only a little bit less than 11%, the value for Ireland reached almost 21%. During following decades, the ratio grew in all countries. In 1990 the lowest value (about 18%) had Slovakia while the highest (almost 31%) was reached in Sweden. The growth continued also in next years and in 2017 the lowest value (almost 23%) was in Ireland and the highest (almost 38%) in Italy. After continual growths during the period projected it is expected that the values of the ratio should reach about 58–61% in Sweden, France, Ireland, Czechia and Austria and about 69–73% in Slovakia, Lithuania, Bulgaria and Italy in the end of this century (Fig. 4; left graph).

Using the prospective concept of the upper productive age threshold (as the age when the remaining life expectancy equals to 20 years) the values and trends of development of the prospective old-age dependency ratio are quite different. In 1950 (when the upper limit of productive age was lower than 65 years in all countries analysed) the lowest value of the ratio (less than 23%) was in Bulgaria while the value in Ireland (over 44%) was the highest. Unlike the standard ratio, the values of prospective one diminished until 1990 in Austria, France and

Ireland. The value for France was only a little bit higher than 30% while for Bulgaria almost 47%. During next years when the mortality in post-communist countries started to decrease more rapidly again, the value of ratio drops also in Czechia and Slovakia. There was a remarkable difference between countries in 2017: while the ratio for Ireland was lower than 23%, for Bulgaria reached more than 54%. In the period projected the differences among analysed countries should diminish. The values of the ratio in 2100 should be in France, Sweden, Ireland, Austria and Czechia about 35–40% while in Italy and Slovakia about 44–45%, in Bulgaria and Lithuania about 47%. The upper value is almost the same like in 1950 (Fig. 4; right graph).

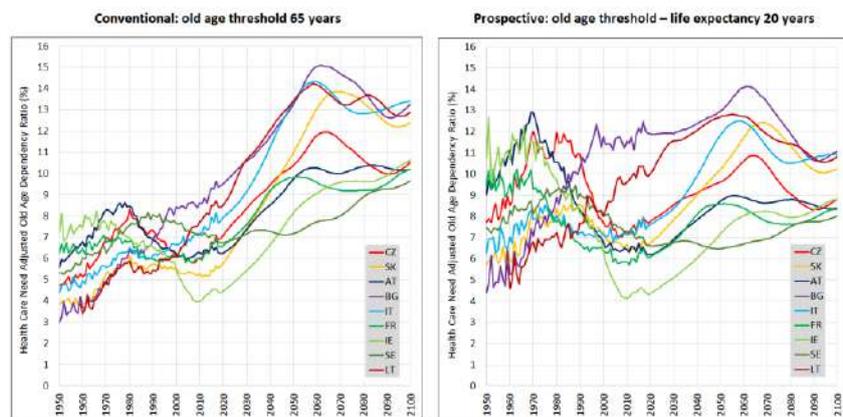


Source of data: HMD (until 2017), Eurostat population projection (since 2018), authors' calculations
 Fig. 4. Old-age dependency ratio

The Health care need adjusted old-age dependency ratio can be used as a very rough indicator of the financial burden of the health care system. The values of this indicator are lower than those of the old-age-dependency ratio but the trends of development will be very similar. In 1950 the lowest value (3%) was in Bulgaria, the highest value (Ireland) reached more than 7%. Until 1990 the ratio grew in all countries except France and Ireland and the differences between countries diminished. In 1990 the lowest value (about 5.5%) had (surprisingly) Lithuania while the highest value (over 8%) was reached in Sweden. During next more than 15 years the values of the ratio rose only in Lithuania, Bulgaria, Italy and France while in other countries there was a small drop. In 2017 the lowest value (4.5%) was in Ireland and the highest value (almost 10%) in Bulgaria. During the remaining decades of this century permanent growths (especially in the first half of this century) is expected. The values of the ratio should in 2100 reach about 10% in Sweden, Austria, France, Ireland and Czechia and about 13% in Slovakia, Lithuania, Bulgaria and Italy (Fig. 5; left graph).

Using the prospective concept of the old age threshold, the values and trends of development of the ratio are a little bit different. The differences are not so remarkable as in the case of old-age dependency ratio because in the case of the

health care need old-age dependency ratio the change of the determination of the old age threshold affects mainly the value of the denominator of the formula while the value of the numerator is not so much different. In 1950 the lowest value of the ratio (about 4.5%) was in Bulgaria while the highest value (in Ireland) reached almost 11%. Until 1990 the values in France, Ireland and Austria diminished. The lowest value (France) was in 1990 only a little bit lower than 7% while for Czechia (highest value) almost 11%. During next years due to rapid decrease of mortality, the value of ratio drops also in Czechia, Slovakia, and also in Sweden. In 2017 the lowest value (in Ireland) was only 4.5% while in Bulgaria (highest value) reached almost 12.5%. In the period projected the ratio should drop a little bit in Bulgaria, in other countries the growths is expected and the differences among analysed countries should diminish. The values of the ratio in 2100 should be in Sweden, France, Austria, Ireland and Czechia about 8-9% while in Slovakia, Lithuania, Italy and Bulgaria about 10–11%. Let us remark that in Ireland, France and Austria the values in 2100 should be a little bit lower than in 1950 (Fig. 5; right graph).



Source of data: HMD (until 2017), Eurostat population projection (since 2018), authors' calculations
 Fig. 5. Health care need old-age dependency ratio

Conclusions

The population ageing is compensated by the increase in life expectancy. Because of this fact the values of indicators of ageing based on prospective age threshold of old age do not grow so dramatically in time as the values of conventional indicators, some of them show relatively stable development in time, especially in some countries. E.g. while the proportion of persons aged 65 and older should be in 2100 in all countries several time higher than in 1950, the proportion of persons having remaining life expectancy 20 years or less in France, Austria, Ireland and Sweden would be in 2100 only by several percentage points higher than in 1950. The average age of population is all analysed countries is growing but the average remaining age remains relatively stable. According the Eurostat projection the ageing of population in Czechia,

Austria, Ireland, France and Sweden should be a little bit slower than in Slovakia, Bulgaria. Italy and Lithuania.

From the economic point of view, the most important issue is, if the retirement age in the future would be determined by the prospective age concept (or by relative prospective age, e.g. Fiala, Langhamrová [2]). The second important condition is that there should be enough appropriate job opportunities for older persons. It of course depends not only on the development of economy but also on the fact if the expected growth of the life expectancy would be accompanied by the simultaneous improvement of the health state of seniors, by increase in the healthy life expectancy.

References

1. Eurostat database. EUROPOP2019 – Population projection at national level. (2019–2100) (proj_19n). Available at <https://ec.europa.eu/eurostat/data/database> (data downloaded on [15-05-2020]).
2. Fiala, T., Langhamrová, J. Pension age based on relative prospective age concept. In Applications of Mathematics and Statistics in Economics (AMSE2018) [CD-ROM]. Kutná Hora, 29.08.2018 – 02.09.2018. Prague : Oeconomica Publishing House, 2018. 9 p. ISBN 978-80-245-2277-7.
3. Fuchs, V. R. Though Much is Taken: Reflections on Aging, Health, and Medical Care. *Milbank Memorial Fund Quarterly. Health and Society*. Vol. 62, No. 2, Special Issue: Financing Medicare: Explorations in Controlling Costs and Raising Revenues (Springer, 1984), pp. 142-166. Published by: Wiley on behalf of Milbank Memorial Fund. [cit. 21-03-2019] <http://www.jstor.org/stable/3349821>
4. Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (data downloaded on [15-05-2020]).
5. Klapková, M., Šídlo, L., Šprocha, B. Koncept prospektivního věku a jeho aplikace na vybrané ukazatele demografického stárnutí (The Concept of Prospective Age and its Application to Selected Indicators of Demographic Ageing). *Demografie*, Vol. 58 (2016), pp. 126-141. Český statistický úřad, Praha.
6. Riffe, Timothy. The force of mortality by life lived is the force of increment by life left in stationary populations. *Demographic Research*. Vol. 32 (2015), pp. 27-834. 10.4054/DemRes.2015.32.29.
7. Ryder, N. B. Notes on Stationary Populations. *Population Index*, Vol. 41, No. 1, pp. 3-28. Published by: Office of Population Research, 1975. [cit. 19-03-2019] <http://www.jstor.org/stable/2734140>.
8. Sanderson, W. C., Scherbov, S. Average Remaining Lifetimes Can Increase as Human Populations Age. *Nature*, 2005, Vol. 435, No. 7 043, p. 811–813. DOI: 10.1038/nature03593. <http://www.nature.com/doifinder/10.1038/nature03593>.
9. Sanderson, W. C., Scherbov, S. Remeasuring Aging. *Science*, Vol. 329 (2010), No. 5 997, pp. 1287–1288. DOI: 10.1126/science.1193647. [cit. 16-03-2019] <http://science.sciencemag.org/content/sci/suppl/2010/09/07/329.5997.1287.DC1/pfSandersonSOM.pdf>.

10. Sanderson, W. C., Scherbov, S. The Characteristics Approach to the Measurement of Population Aging. *Population and Development Review*, Vol. 39 (2013), No. 4, pp. 673–685. DOI: 10.1111/j.1728-4457.2013.00633.x. [cit. 12-03-2019] <http://onlinelibrary.wiley.com/doi/10.1111/j.1728-4457.2013.00633.x/full>.
11. Siegel, J. S. *A Generation of Change: a Profile of America's Older Population*. New York: Russell Sage Foundation, 1993.
12. Spijker, J. – Riffe, T.L.M – MacInnes, J.: Incorporating time-to-death (TTD) in health-based population ageing measurements. Presented at the New Measures of Age and Ageing, Vienna, 3 – 5 December, 2014.
13. Šprocha, B. Niektoré nové prístupy k analýze populačného starnutia (Some New Approaches to the Analysis of the Population Ageing). *Slovenská štatistika a demografia (Slovak Statistics and Demography)*. Vol. 29 (2019). No. 4, pp. 2335.
14. Zweifel P, Felder S, Meiers M. Aging of population and health care expenditure: a red herring. *Health Econ* 1999; 8: 485–496.

Acknowledgement

This article was supported by the Czech Science Foundation No. GA ČR 19-03984S under the title *Economy of Successful Ageing*.

The Kelly bet and stock market investment

James Freeman and Zhan Xu

Alliance Manchester Business School, University of Manchester, Booth Street West,
Manchester, M15 6PB, UK

Email: jmacfreeman@gmail.com

Abstract: The Kelly strategy for bet-sizing has long dominated gambling applications. More recently, the Kelly criterion has come to the fore in stock market investment – a development underpinned by extensive analytic theory. Complementing the latter, this new study examines the efficacy of the Kelly method from an empirical perspective. In particular, using the S&P 500 dataset for illustration, the return growth rates realised by Kelly style leveraging are found to be vastly superior to a straight non-leveraging alternative over a 12 year cycle.

Key words: Kelly bet, leveraging, trading strategy, S&P 500, Relative Strength Index (RSI) Terminal Wealth Relative (TWR) index

1 Introduction

As for gamblers, the twin cardinal concerns of investors or traders are:

- Finding a set of profitable opportunities
- Sizing investments [1]

The latter, regarded by most professionals as the more challenging of the two, was first explored by Kelly [2] - his seminal analysis being adapted for stock market application by Thorp & Kassouf [3].

Building on this work, Maclean and Ziemba [4] developed a Kelly-style approach for use in asset management in 2006 [5], [6], [7].

More recently, Nekrasov [8] reported on the value of deploying so-called ‘fractional’ Kelly strategies for managing arbitrarily large investment portfolios.

Complementing the huge amount of analytical modelling for the Kelly criterion in the literature - see for example: Poundstone [9] Rotando [10], Ziemba [11] - the paper reports on the effectiveness of Kelly for stock market application from an empirical standpoint.

A background to the study – including details of the Terminal Wealth Relative and the Relative Strength Indices used in data preparation - is provided in Section 2. Section 3 delineates the experimental design employed and Section 4, the relative performance of selected Kelly and non-Kelly investment strategies. Conclusions for the study appear in Section 5.

2. Background

For a bet with two outcomes, the Kelly bet formula can be written:

$$f^* = \frac{bw - l}{b} = \frac{p(b+1) - 1}{b} = \frac{\text{Edge}}{\text{Odds}}$$

f^* = proportion of the current capital to wager;

b = net odds gained on the bet (" b to 1"); namely, one would receive \$ b , in addition to the \$1 invested, for a successful \$1 bet (otherwise the \$1 investment is lost)

w = probability of winning

$l = 1 - w$ = probability of losing

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain



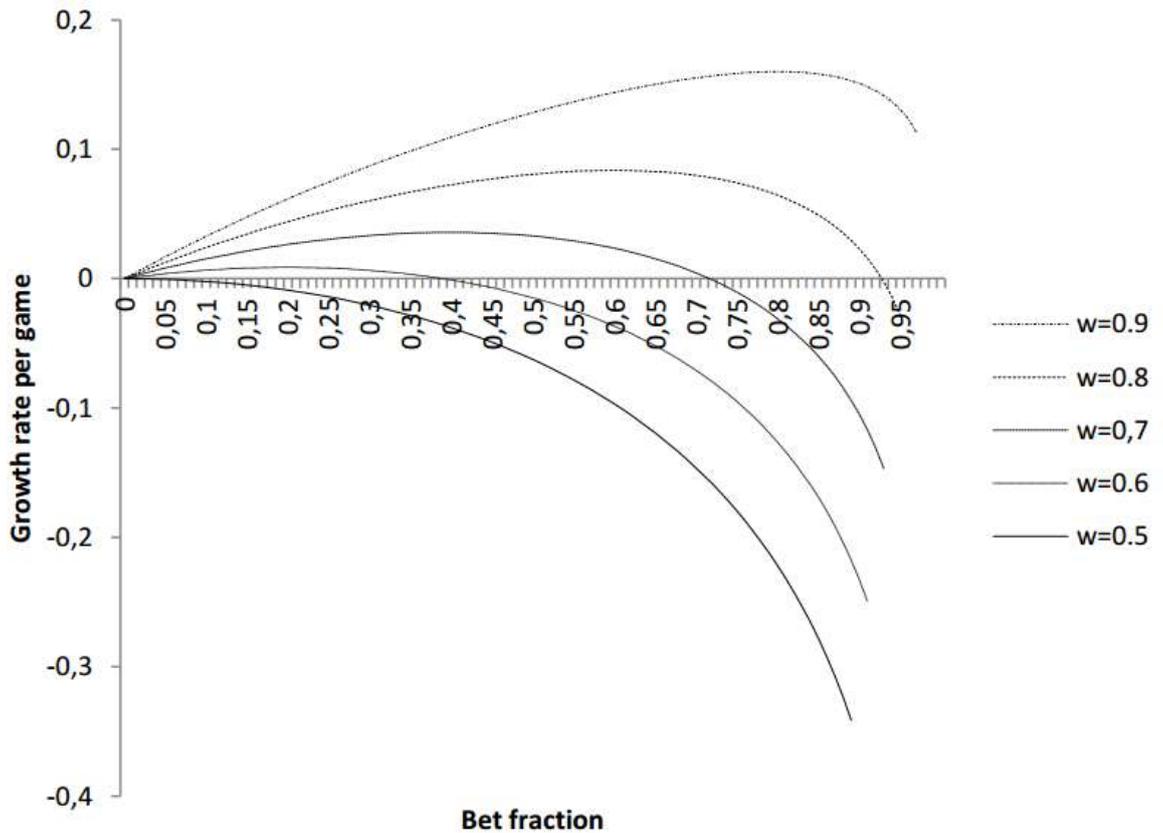


Fig.1 Capital growth rate by bet fraction, f

From the preceding graph, it is evident that:

- a positive growth rate is only achievable when $w > 0.5$ in other words when we have ‘a positive expectation bet’
- the relationship between the betting proportion, f and growth rate per bet is concave.
- The highest growth rate occurs when

$$f = f^* = w - l = 2w - 1$$

e.g. when $w = 0.7$, growth rate is maximised when

$$f = f^* = 0.4 (= 0.7 - 0.3)$$

2.1 Terminal Wealth Relative

Assume our initial bankroll takes the value W_0 and that this transforms into W_n after a sequence of n investments. Then the Terminal Wealth Relative

$$TWR = \frac{W_n}{W_0} = \prod_{i=1}^n (1 + \theta r_i)$$

Where

θ = leverage factor

r_i = payoff rate of investment ($i = 1, 2, \dots, n$)

and the average growth rate of the bankroll is estimated by:

$$G_n(\theta) = \frac{1}{n} \ln \left(\frac{W_n}{W_0} \right) = \frac{1}{n} \sum_{i=1}^n \ln(1 + \theta r_i)$$

In broad terms, the leverage factor θ is analogous to the betting fraction f in the original Kelly formulation but for the particular experimental design, here, is assumed to be constant

In practice, θ - unlike f - has to be estimated by numerical methods. (See for example the ‘historical-returns’ method described by Balsara [12].

2.2 Relative Strength Index

For each trading period an upward change U or downward change D is calculated. Up periods are characterized by the close being higher than the previous close:

$$U = \text{close}_{\text{now}} - \text{close}_{\text{previous}}$$

$$D = 0$$

Conversely, a down period is characterized by the close being lower than the previous period's close:

$$U = 0$$

$$D = \text{close}_{\text{previous}} - \text{close}_{\text{now}}$$

The average U and D are calculated using an n -period smoothed moving average (SMMA) which is an exponentially smoothed moving average with $\alpha = 1/\text{period} = 1/n$ [13]. Historically, Wilder [14] recommended the use of $n = 14$ days.

Denoting

- the average price increase by SMMA (U, n) and
- the average price drop by SMMA (D, n)

then the ‘relative strength’ (RS) is given by

$$RS = \frac{\text{SMMA}(U, n)}{\text{SMMA}(D, n)}$$

RS gives a measure as to how well an investment is performing against the market or benchmark. In practice RS is converted to the Relative Strength Index (RSI):

$$RSI = 100 - \frac{100}{1 + RS}$$

which ranges in value from 0 – 100.

Typically, a graph of the ‘momentum indicator’, RSI is positioned above or below the price chart e.g



Fig.2 Relative Strength Index Daily chart, Wal-Mart (Tradegroup)

As the latter graph shows, the indicator has an upper line set at 70 and a lower line at 30. When the index rises above 70, the asset is considered ‘overbought’ or over-valued. Conversely, when the index decreases below the 30 level, the stock is considered ‘oversold’ or under-valued. [15].

3. Empirical study

The strategy tested in this study combines elements of both TMR and RSI methodologies.

Following Wilder [14], the focus is on long - not short – trades.

The data used for the analysis is the S&P 500 stock market index covering the period December 2000 to December 2012:



Fig.3 S&P 500 stock market index (2000 – 2012)

Notation

t = day of trade

P = Price at time τ on day t

P^{τ} = Highest price on day t

$P^{H,t}$ = Lowest price on day t

$P^{L,t}$ = Closing price on day t

Summary statistics for the Daily Return:

$$r_t = \ln \left(\frac{P_{C,t}}{P_{C,t-1}} \right)$$

are as follows:

Observations	Mean	Highest Value	Lowest Value	Kurtosis	Skewness
3017	0.00438%	10.957%	-9.4689%	7.90333	-0.17429

In line with RSI theory we filter those observations from the full dataset that satisfy the following conditions:

$$RSI_{t-1} < 30 \quad (1)$$

$$P_t > P_{H,t-1} \quad (2)$$

The resulting dataset of 106 observations we then split into contiguous training and test samples of 53 observations each. Details are as follows:

Period	Observations	Mean	Highest Value	Lowest value	Kurtosis	Skewness
20/12/2000-20/12/2012	106	0.4523%	8.124%	-11.5031%	11.852	-1.298
20/12/2000-4/6/2005	53	0.6102%	8.124%	-11.5031%	12.408	-2.135
4/6/2005-20/12/2012	53	0.2893%	6.509%	-7.856	7.619	-0.626

To offset the heavy tails that asset returns are typically affected by, the data are then processed using the HCSE (heteroscedasticity corrected standard errors - or White-Huber standard errors) procedure [16] – yielding the summary:

Period	Observations	Mean	Robust HCSE	p-value
20/12/2000-20/12/2012	106	0.4142%	0.212%	0.031
20/12/2000-4/6/2005	53	0.6102%	0.246%	0.112
4/6/2005-20/12/2012	53	0.2893%	0.179%	0.139

According to the significant p-value of 0.031 (< 5%) shown for the full sample here, we deduce the positive expected payoff condition needed for the Kelly procedure to apply, is satisfied for the time period in question.

4. Analysis

In the analysis of the training set that follows, we assume that the transaction fee is 0.04%. To allow for this and to offset problems of gaps occurring between the closing prices in one period and the opening prices in the following period, the simulation has incorporated the additional rules:

$$r_t | (P_{L,t} < P_{H,t-1}) = \ln \left(\frac{P_{C,t}}{P_{H,t-1}} \right) - 0.0004$$

$$r_t | (P_{L,t} > P_{H,t-1}) = \ln \left(\frac{P_{C,t}}{P_{L,t}} \right) - 0.0004$$

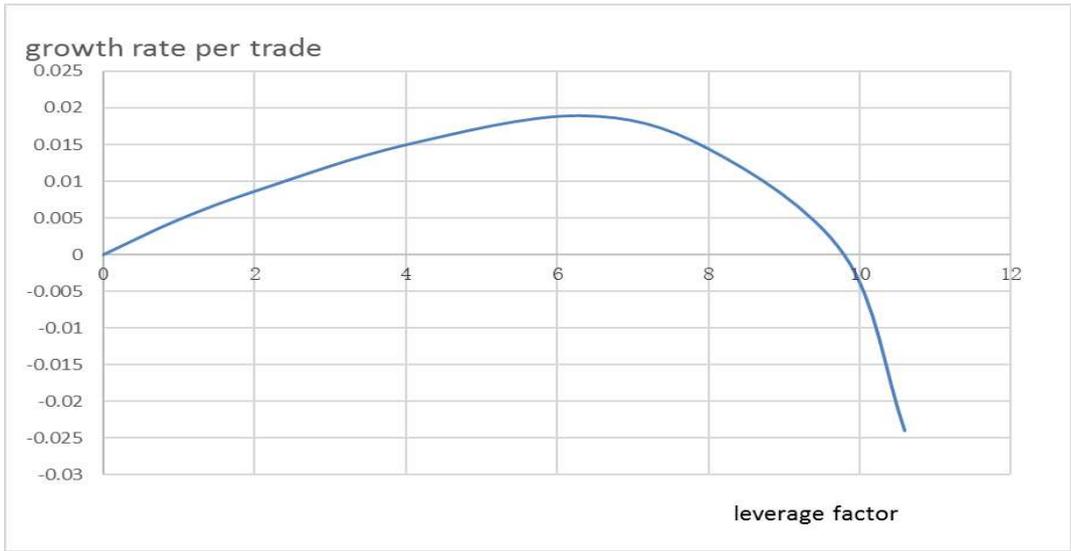


Fig.4 Growth rate by trade versus leverage factor

The preceding graph – which summarises the principal results from the analysis – shows that the pattern linking leverage factor to growth rate per trade identifies very closely with Capital growth rate by bet fraction f plots shown earlier in Figure 1.

Here we see that the growth rate is maximized at $G = 1.936\%$ when $\theta = 6.3$. This is the optimal value (θ^*) we use for the leverage factor in the TWR analysis of the test data that follows.

By way of contrast, we also highlight corresponding capital growth results for the case where $\theta = 1$ i.e. when no leveraging is in effect as such.



Fig.5 Growth of TWR for $\theta = \theta^* = 6.3$ and $\theta = 1$

Clearly from Figure 5, TWR growth is vastly superior for Kelly style trading compared to the chosen non-Kelly alternative: indeed, the capital growth rate realised with Kelly – see Table 1 below - can be shown to be some five times higher

Leverage factor, θ	Final TMR	% Growth Rate
1	1.156	0.274
$\theta^* = 6.3$	2.653	1.809

Table 1: Final TMR and % Growth Rate for Kelly versus non-Kelly style trading

5. Conclusions

The study provides new empirical evidence in favour of the use of the Kelly criterion for stock market investment.

Though the modelling undertaken for this prototype project - of necessity - is subject to a number of simplifying assumptions – in particular that investors are risk-neutral, a constant leverage factor applies, finding matching trades is not a problem [17] the results obtained so far have been very encouraging. So much so that a range of real-life adaptations to it has already been completed as part of the next phase of the research.

References

1. N. Yoder. The Kelly Criterion: How to apply the famous betting formula in investment, trading and professional gambling. 2019. [Online], Available: <https://medium.com/@nickyoder/the-kelly-criterion-cd986d037d87> (Accessed 2020-2-22).
2. J. L. Kelly. A New Interpretation of Information Rate. *Bell System Technical Journal*. 35, 4, 917–926. 1956.
3. E.O. Thorp and S.T. Kassouf, *Beat the Market: A Scientific Stock Market System*, Random House, 1967.
4. L.C. Maclean, and W.T. Ziemba, Capital Growth Theory and Practice. In S. A. Zenios and W. T. Ziemba (Eds.), *Handbook of asset and liability management*, Handbooks in Finance, pp. 430–469. North Holland. 2006.
5. M. Pabrai. *The Dhandho Investor: The Low-Risk Value Method to High Returns*, Wiley, 2007.
6. E.O. Thorp. The Kelly Criterion: Part I. *Wilmott Magazine*, May 2008.
7. E.O. Thorp. The Kelly Criterion: Part II. *Wilmott Magazine*, September 2008.
8. V. Nekrasov. Kelly Criterion for Multivariate Portfolios: A Model-Free Approach (September 30, 2014). Available at SSRN: <https://ssrn.com/abstract=2259133> or <http://dx.doi.org/10.2139/ssrn.2259133>
9. W. Poundstone. *Fortune's Formula: The Untold Story of the Scientific Betting System That Beat the Casinos and Wall Street*. Macmillan, 2005.
10. L.M. Rotando and E.O. Thorp. The Kelly criterion and the stock market. *The American Mathematical Monthly*, 99, 922–931. 1992.
11. W.T. Ziemba. *Great investment ideas*, World Scientific Publishing Co. Ltd., 2017.
12. N. J. Balsara. *Money Management Strategies for Futures Traders*. Wiley, 1992.
13. <https://user42.tuxfamily.org/chart/manual/Exponential-Moving-Average.html> (Accessed 2020-2-22).
14. J.W. Wilder Jr. *New concepts in technical trading systems*. Greensboro, NC: Hunter Publishing Company, 1978.
15. <https://www.investopedia.com/articles/active-trading/042114/overbought-or-oversold-use-relative-strength-index-find-out.asp> (Accessed 2020-2-22).
16. http://www3.grips.ac.jp/~yamanota/Lecture_Note_9_Heteroskedasticity (Accessed 2020-2-22).
17. C. Lundström. Day Trading Profitability across Volatility States: Evidence of Intraday Momentum and Mean-Reversion. *Umeå Economic Studies*, 861. Umeå University, 2013.

Spatial Triplet Markov Trees for Auxiliary Variational Inference in Spatial Bayes Networks

Hugo Gangloff^{1, 2}, Jean-Baptiste Courbot³, Emmanuel Monfrini⁴, and
Christophe Collet¹

¹ ICube, Université de Strasbourg - CNRS UMR 7357, Illkirch, France

(E-mail: hugogangloff@unistra.fr, c.collet@unistra.fr)

² GEPROVAS, Strasbourg, France

³ IRIMAS UR 7499 Université de Haute-Alsace, Mulhouse, France

(E-mail: jean-baptiste.courbot@uha.fr)

⁴ SAMOVAR - CNRS UMR 5157, Évry, France

(E-mail: emmanuel.monfrini@telecom-sudparis.eu)

Abstract. In this article, we develop a Triplet Markov Tree model with auxiliary random variables for an approximate inference in an intractable probabilistic model. It is based on recent advances on probabilistic modeling and variational inference with auxiliary random variables. The new Triplet Markov Tree model performs better than the classical Mean-Field variational inference and than a tree-structured variational inference. Our study provides insights and motivations for the developing work around models involving auxiliary random variables.

Keywords: Variational Inference, Auxiliary Random Variables, Triplet Markov Models.

1 Introduction

1.1 Position of the problem

Modeling the relations between random variables of probabilistic models requires to find a compromise between tractability and richness of the correlations.

Markov tree-structured probabilistic models are therefore a popular choice because they offer the ability to introduce correlations while preserving tractability, in particular, the deterministic retrieval of the marginals [14]. Highly structured trees, known as dyadic- or quadtree-Markov Trees (DMT) [6, 8, 11, 19] (Figure 1), have been used for the probabilistic treatment of mono-dimensional or bi-dimensional data series.

The modelization with Markov process has been generalized by using auxiliary variables in [4, 5, 12]. In these models, a carefully chosen auxiliary process enables the introduction of richer correlations while, preserving the Markov property and the model tractability.

Based on the successful structured and hierarchical dyadic Markov trees, we present the Spatial Bayes Network (SBN) model which introduces, with

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



respect to the DMT model, additional correlations between neighboring nodes (Figure 2), which might be an appealing feature for the practitioner. However, as soon as those additional correlations are established, the network contains many loops, inference becomes intractable and it must be approximated.

1.2 Variational Inference

Variational inference (VI) [9] [14] [3] [18] is an approach to perform inference in a complex probability distribution using a simpler one, called the *variational distribution*. The inference problem is then recast as an optimization problem. There exists several designs of VI. The Mean-Field (MF) VI is the most popular approach to VI: it consists in using a variational distribution with a fully factorized, unstructured, form [9]. While computation is easy in this context, the result of the optimization problem can fail to reflect key correlations between random variables, because of the lack of structure. Therefore, over the years, *structured* VI has been developed. It consists in using a structured variational distribution, in which computation is still tractable, in order to better approximate the correlations in the original intricate distribution. Structured VI can lead to dramatic increase in performances thanks to the improved modelling of the correlations [2] [7] [13]. Generally speaking, VI leads to a non-convex optimization problem. However, adding more structure makes some local optima disappear [18], hence the enhanced inference results. A recent trend in structured VI is to propose variational distribution using auxiliary variables, which leads to further improvements in the inference [1] [16] [17].

1.3 Outline of the paper

The main goal of this article is to develop a Spatial Triplet Markov Tree (STMT) model of [5] to exhibit the richness of the correlations it induces. We do so by using the STMT distribution in a VI procedure with auxiliary variables to approximate the intractable SBN distribution.

We first present the definitions of the probabilistic models mentioned in the introduction: DMT, SBN and STMT. We then develop the variational inference framework required to perform approximate inference in the intractable SBN. To the best of our knowledge, structured VI has never been studied with neither DMT nor STMT as the variational distribution. Finally we numerically show on an example the interest of STMT and auxiliary variable modeling to enhance the modelization.

2 Model definitions

2.1 Dyadic Markov Tree (DMT)

Let $\mathbf{X} = (X_1, \dots, X_N)$ be a random process which can be real or discrete-valued. We will refer to the realizations of the random variable with the notation $p(\mathbf{X} = \mathbf{x}) = p(\mathbf{x})$.

We define a tree graph over a set of nodes $\mathcal{S} = \{\mathcal{S}^0, \dots, \mathcal{S}^{n-1}\}$, where each \mathcal{S}^k is a resolution. \mathcal{S}^0 has a unique node r , the root node. Let $\bar{\mathcal{S}} = \mathcal{S} \setminus \mathcal{S}^0$, then $\forall s \in \bar{\mathcal{S}}$, a parent node of s is denoted s^- . In *dyadic* trees each node has 2 sons. Note that Markov trees have no cycles; each node has exactly one parent.

A DMT (illustrated in Figure 1) has the following joint distribution [11]:

$$p(\mathbf{x}) = p(x_r) \prod_{s \in \bar{\mathcal{S}}} p(x_s | x_{s^-}). \quad (1)$$

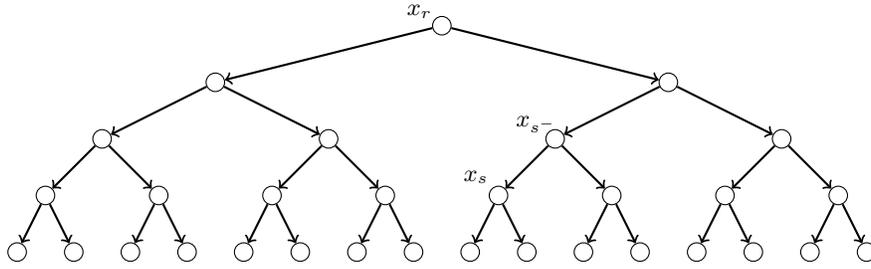


Fig. 1: Graphical model corresponding to a DMT. As an illustration, a node x_s , its father x_{s^-} and the root node x_r have been annotated.

In all tree structured graphs inference can be carried out exactly by a message passing technique [14]. In DMT, message passing is often stated in the form of the the upward-downward approach as given in [6].

2.2 Spatial Bayes Network (SBN)

We introduce a new model which we call Spatial Bayes Network (SBN). The model is a hierarchical Bayes network which contains cycles, in order to better capture local correlations between random variables.

In order to distinguish between the two sons of a father node s^- in a dyadic tree: let s_L and s_R be respectively the *left* and *right* son of s^- . We also define s^{\leftarrow} (resp. s^{\rightarrow}) as the left (resp. right) neighbouring node of s . Let $v: \bar{\mathcal{S}} \rightarrow \bar{\mathcal{S}}$ be a mapping from a node to a neighbouring node of its father, such that:

$$v: s \mapsto \begin{cases} (s^-)^{\leftarrow} & \text{if } s \text{ is a } \textit{left} \text{ node,} \\ (s^-)^{\rightarrow} & \text{if } s \text{ is a } \textit{right} \text{ node,} \end{cases} \quad (2)$$

The SBN model (illustrated in Figure 2) has the following distribution:

$$p(\mathbf{x}) = p(x_{s_r}) \prod_{s \in \bar{\mathcal{S}}} p(x_s | x_{s^-}, x_{v(s)}). \quad (3)$$

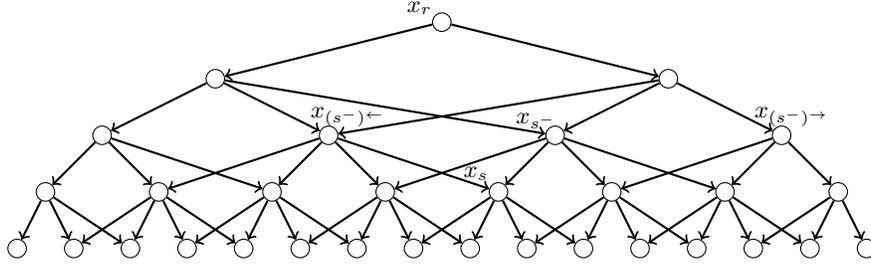


Fig. 2: Graphical model corresponding to a dyadic SBN. As an illustration, a node x_s , its father x_{s^-} , the left neighbour of its father $x_{(s^-)\leftarrow}$, the right neighbour of its father $x_{(s^-)\rightarrow}$ and the root node x_r have been annotated.

2.3 Spatial Triplet Markov Tree (STMT)

Along with the previously introduced process \mathbf{X} , let $\mathbf{V} = (V_1, \dots, V_N)$ be a (real- or discrete-valued) random process which is an *auxiliary* process. Let \mathbf{T} be a process such that $\mathbf{T} = (\mathbf{X}, \mathbf{V})$ and having the joint distribution:

$$p(\mathbf{t}) = p(\mathbf{t}_r) \prod_{s \in \bar{\mathcal{S}}} p(\mathbf{t}_s | \mathbf{t}_{s^-}). \quad (4)$$

We now describe the hypothesis and the transitions to build a dyadic *Spatial Triplet Markov Tree* model as introduced in [5] (illustrated in Figure 4)¹. A node at site s is a 3-tuple $(X_s, V_s^\leftarrow, V_s^\rightarrow)$; we then assume that we have the factorization, $\forall s \in \bar{\mathcal{S}}$:

$$\begin{aligned} p(\mathbf{t}_s | \mathbf{t}_{s^-}) &= p(x_s, \mathbf{v}_s | x_{s^-}, \mathbf{v}_{s^-}), \\ &= p(x_s | x_{s^-}, \mathbf{v}_{s^-}) p(v_s^\leftarrow | x_{s^-}, \mathbf{v}_{s^-}) p(v_s^\rightarrow | x_{s^-}, \mathbf{v}_{s^-}). \end{aligned} \quad (5)$$

To further define the transition laws of each of the variable we define the notion of *inner* and *outer* variables for the V variables. We define that, within *Left* (resp. *Right*) sons, V_{sL}^\leftarrow (resp. V_{sR}^\rightarrow) is an outer variable and V_{sL}^\rightarrow (resp. V_{sL}^\leftarrow) is an inner variable. Figure 3 illustrates these concepts for a particular node.

We now detail Equation 5, with special care on the variable type (*left*, *right*, *inner* or *outer*). For X_s sons:

$$\begin{cases} p(x_s^L | x_{s^-}, \mathbf{v}_{s^-}) &= p(x_s^L | x_{s^-}, v_{s^-}^\leftarrow) \\ p(x_s^R | x_{s^-}, \mathbf{v}_{s^-}) &= p(x_s^R | x_{s^-}, v_{s^-}^\rightarrow), \end{cases} \quad (6)$$

for inner V_s sons:

$$\begin{cases} p(v_{sL}^\rightarrow | x_{s^-}, \mathbf{v}_{s^-}) &= p(v_{sL}^\rightarrow | x_{s^-}, v_{s^-}^\rightarrow) \\ p(v_{sR}^\leftarrow | x_{s^-}, \mathbf{v}_{s^-}) &= p(v_{sR}^\leftarrow | x_{s^-}, v_{s^-}^\leftarrow), \end{cases} \quad (7)$$

¹We keep referring to our model as *Spatial Triplet* as introduced in [5] because of its origin and the close definition: the original model is linked with image processing hence *Spatial* and it uses a triplet Markov tree. While in this article we only have two processes, this does not change the construction because the missing *observed* process only plays a marginal role in the original model.

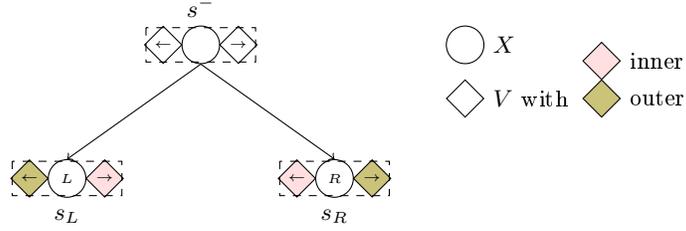
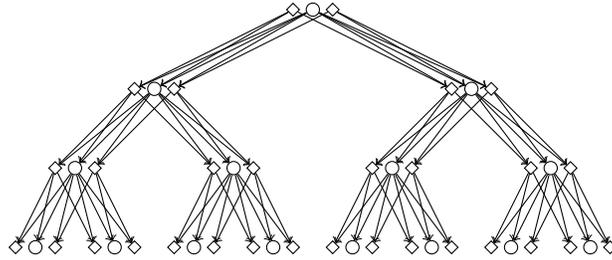


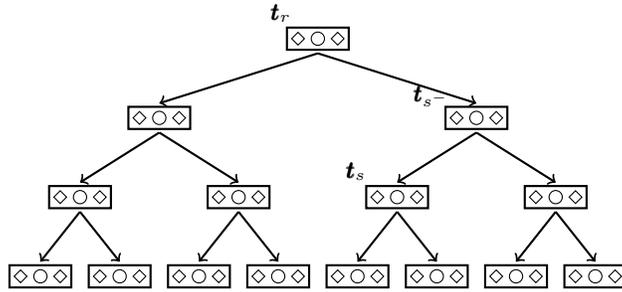
Fig. 3: Details of the STMT construction: a father node s^- linked to its 2 sons (s^L, s^R). The directionality of the V_s is specified as well as their type (inner or outer).

and for outer V_s sons:

$$\begin{cases} p(v_{s^L}^{\leftarrow} | x_{s^-}, \mathbf{v}_{s^-}) = p(v_{s^L}^{\leftarrow} | v_{s^-}^{\leftarrow}, x_{s^-}) \\ p(v_{s^R}^{\rightarrow} | x_{s^-}, \mathbf{v}_{s^-}) = p(v_{s^R}^{\rightarrow} | v_{s^-}^{\rightarrow}, x_{s^-}). \end{cases} \quad (8)$$



(a)



(b)

Fig. 4: Graphical model corresponding to a dyadic STMT. (a) depicts all the links, (b) gives a condensed view highlighting the preserved Markov tree structure. On (b), a node t_s , its father t_{s^-} and the root node t_r have been annotated.

Remark: The conditioning of the V_s variables can be seen as being the same conditioning as the X variable it spatially "refers" to. The Triplet tree

framework then provides a way to simulate X variables conditionally to the realizations of V variables which behave similarly to the neighboring X variables, hence the closeness with SBN we emphasize in this article.

A generalized version of the upward-downward algorithm of [6] enables us to compute deterministically the marginals in the STMT model [5].

3 Variational inference in SBNs

Within the Variational Inference (VI) framework [14], we aim at finding a distribution $q(\mathbf{x})$ which approximates well $p(\mathbf{x})$. We commonly seek to minimize the *reverse* Kullback-Leibler $\mathbb{KL}(q(\mathbf{x})||p(\mathbf{x}))$. By definition:

$$\mathbb{KL}(q(\mathbf{x})||p(\mathbf{x})) = \mathbb{E}_q[\log p(\mathbf{x})] - \mathbb{E}_q[\log q(\mathbf{x})]. \quad (9)$$

In this article, $p(\mathbf{x})$ is the joint distribution of a SBN (Section 2.2):

$$p(\mathbf{x}) = \prod_{s \in \mathcal{S}} p(x_s | x_{s-}, x_{v(s-)}) \quad (10)$$

In the context of structured variational inference, we will reparametrize and work with another representation of Equation 10. Indeed, we need to use a concise notation embedding the notion of clusters of variables for each of the terms of the product. The notations are adapted from [2]. We can write:

$$p(\mathbf{x}, \mathbf{y}) = \prod_{d_s} p_{d_s} \quad (11)$$

where d_s represents the cluster of variables $(x_s, x_{s-}, x_{v(s-)})$ (note that these clusters overlap).

In the following we develop the necessary material to approximate Equation 10 with DMT and STMT, which benefit from their tree structure and deterministic marginalization. Therefore the approximations obtained by each of the models of interest will be caused by the enhanced correlations that some models have with respect to others.

Algorithmically, the steps of the VI procedure for a variational distribution q can be summarized:

1. Initialize all the factors of q .
2. $\forall j \in \mathcal{S}$, update q_{c_j} with the expression that minimizes the Kullback-Leibler divergence.
3. Check convergence and repeat step 2 if needed.

The convergence can be assessed, *e.g.*, by monitoring the values of the cost function or monitoring stationarity in the estimated variational parameters.

3.1 Structured variational inference with DMTs

The variational distribution r is here defined with the structure of a Markov Tree (DMT) (Section 2.1):

$$r(\mathbf{x}) = \prod_{s \in \mathcal{S}} r(x_s | x_{s-}) = \prod_{c_s} r_{c_s}, \quad (12)$$

where c_s represents the cluster of variables (x_s, x_{s-}) . Using a MT structure as a posterior to approximate the more complex posterior p is a compromise between additional structure and tractability. Indeed, once the posterior transitions will be learnt through the VI optimization, the posterior marginals are easily computable with the upwards-downwards algorithm [6].

The Kullback-Leibler divergence to minimize (Equation 9) becomes in this section:

$$\mathbb{KL}(r||p) = \sum_{\mathbf{x}} \prod_{c_s} r_{c_s} \left(\log p(\mathbf{x}) - \sum_{c_k} \log r_{c_k} \right). \quad (13)$$

Let us now isolate one of the factors, r_{c_j} of the variational distribution r . This leads to splitting sums and product according to the different clusters c_s . It follows that:

$$\mathbb{KL}(r||p) = \sum_{x_j, x_{j-}} r_{c_j} \sum_{\mathbf{x} \setminus \{x_j, x_{j-}\}} \prod_{c_i \neq c_j} r_{c_i} \left(\log p(\mathbf{x}) - \log r_{c_j} - \sum_{c_k \neq c_j} \log r_{c_k} \right). \quad (14)$$

We want to minimize this quantity with the constraint $\prod_{c_s} r_{c_s} = 1$. Therefore, we introduce a Lagrangian multiplier $\lambda > 0$ and the quantity to optimize becomes:

$$\widetilde{\mathbb{KL}}(r||p) = \mathbb{KL}(r||p) + \lambda \left(\prod_{c_s} r_{c_s} - 1 \right). \quad (15)$$

Taking the functional derivative of Equation 15 and setting this derivative to 0, we have the expression of the variational factor $r_{c_j}^*$ minimizing the divergence:

$$\log r_{c_j}^* = \mathbb{E}_{\prod_{c_i \neq c_j} r_{c_i}} \left[\sum_{d_s} \log p_{d_s} - \sum_{c_k \neq c_j} \log r_{c_k} \right] + \text{const.}, \quad (16)$$

where const. denotes the group of constant terms with respect to c_j . This expression can be simplified because additional terms in the three sums do not depend on c_j (and these terms can merge with the const. term). Let \mathcal{D}_j be a set whose elements are clusters of variables d_j containing x_j , \mathcal{B}_j a set of clusters of variables b_j also containing x_j but with the condition $b_j \neq c_j$. With such reparametrization, it follows that Equation 16 leads to:

$$r_{c_j}^* = \frac{1}{Z_j} \exp \left(\mathbb{E}_{\prod_{c_i \neq c_j} r_{c_i}} \left[\sum_{d_j \in \mathcal{D}_j} \log p_{d_j} - \sum_{b_j \in \mathcal{B}_j} \log r_{b_j} \right] \right), \quad (17)$$

with Z_j a normalization constant.

Remark: The expectation appearing in Equation 17 requires to sample from the joint law $\mathbf{x} \setminus \{x_j, x_{j-}\}$ given x_j and x_{j-} . Such a sampling can be done when r is an MT because of its straightforward sparse decomposition. Note also that, in fact, we only need to sample the variables in each cluster (upon which the expectation is computed, *i.e.* d_j or b_j). Then \mathbf{X} never needs to be fully sampled.

3.2 Auxiliary variable variational inference with STMTs

We now develop VI over STMT using auxiliary variables as initiated in 1. Let \mathbf{V} be the auxiliary process which will be used to augment both p and t . The new *auxiliary* variational lower bound to minimize is:

$$\begin{aligned} \mathbb{KL}(t(\mathbf{x}, \mathbf{v})||p(\mathbf{x}, \mathbf{v})) &= \mathbb{E}_t[\log p(\mathbf{x}, \mathbf{v})] - \mathbb{E}_t[\log t(\mathbf{x}, \mathbf{v})], \\ &= \mathbb{E}_t[\log p(\mathbf{x})] + \mathbb{E}_t[\log p(\mathbf{v}|\mathbf{x})] - \mathbb{E}_t[\log t(\mathbf{x}, \mathbf{v})]. \end{aligned} \quad (18)$$

Following 10, we have that the Kullback-Leibler divergence with auxiliary variables is lower-bounded by the Kullback-Leibler divergence without auxiliary variables:

$$\mathbb{KL}(t(\mathbf{x}, \mathbf{v})||p(\mathbf{x}, \mathbf{v})) \geq \mathbb{KL}(t(\mathbf{x})||p(\mathbf{x})). \quad (19)$$

So the cost function with auxiliary variables is at best equal to the cost function without auxiliary variables.

However, the use of auxiliary variables offers much more flexibility in the modelization and enables the modeling of richer correlations between the variables of interest.

We focus now on STMT as the variational distribution to approximate the SBN. The auxiliary variables have been introduced to better reflect the correlations in SBNs, while keeping the Markov-tree property. Hence, at the end of the VI procedure, exact marginal computation can be done, again with a generalized upward-downward algorithm.

Let t be the STMT distribution, we have from Section 2.3:

$$\begin{aligned} t(\mathbf{x}, \mathbf{v}) &= \prod_{s \in \mathcal{S}} t(x_s, \mathbf{v}_s | x_{s-}, \mathbf{v}_{s-}), \\ &= \prod_{s \in \mathcal{S}} t(x_s | x_{s-}, v_{n(s-)}) t(v_s^{\leftarrow} | x_s, v_{n'(s-)}) t(v_s^{\rightarrow} | x_s, v_{n''(s-)}), \\ &= \prod_{c_s} t_{c_s} \prod_{c'_s} t_{c'_s} \prod_{c''_s} t_{c''_s}, \end{aligned} \quad (20)$$

with $c_s = (x_s, x_{s-}, v_{n(s-)}), c'_s = (v_s^{\leftarrow} | x_s, v_{n'(s-)})$ and $c''_s = (v_s^{\rightarrow} | x_s, v_{n''(s-)}).$

We want to minimize Equation [18](#), which becomes for STMT:

$$\begin{aligned} \mathbb{KL}(t||p) = \sum_{\mathbf{x}, \mathbf{v}} \prod_{c_s} t_{c_s} \prod_{c'_s} t_{c'_s} \prod_{c''_s} t_{c''_s} & \left(\log p(\mathbf{x}) - \right. \\ & \left. \left(\sum_{c_k} \log t_{c_k} + \sum_{c'_k} \log t_{c'_k} + \sum_{c''_k} \log t_{c''_k} \right) \right). \end{aligned} \quad (21)$$

Obtaining the update equations for each variational transition follows the same steps as for DMT-VI (which are omitted here for brevity). We have $\forall t_{c_j}$:

$$\begin{aligned} \log t_{c_j}^* = \mathbb{E}_{\prod_{c_i \neq c_j} t_{c_i} \prod_{c'_s} t_{c'_s} \prod_{c''_s} t_{c''_s}} & \left[\log p(\mathbf{x}) - \right. \\ & \left. \sum_{c_k \neq c_j} \log t_{c_k} - \sum_{c'_k} \log t_{c'_k} - \sum_{c''_k} \log t_{c''_k} \right] + \text{const.}, \end{aligned} \quad (22)$$

where the last term regroups constant terms with respect to c_j .

Update equations for $t_{c'_j}$ and $t_{c''_j}$ are similar but the term $\log p(\mathbf{x})$ is replaced by $\log p(\mathbf{v}|\mathbf{x})$ in both cases.

Now STMT VI is still a relatively sparse and highly structure network, hence, the updates equations can be simplified as in Equation [17](#). We need to carefully select the subsets of the variables involved in the expectation while the other join the constant term that is unimportant.

4 Experiments and Results

4.1 Experimental set-up

We now consider variational inference on the small SBN network given in Figure [5a](#). Similar experiments have been conducted in the same context, to evaluate a VI approximation, for example in [13](#). Due to its small size the SBN of Figure [5a](#) represents a slightly modified probability distribution from that the SBN used up to now. Indeed, we needed to treat in a specific fashion the root node to induce SBN-like correlations on a 3-layered network only. It follows that p has the following distribution:

$$\begin{aligned} p(a, a^\leftarrow, a^\rightarrow, b, c, d, e, f, g) = & p(a)p(a^\leftarrow)p(a^\rightarrow)p(b|a, a^\leftarrow)p(c|a, a^\rightarrow) \\ & p(d|b)p(e|b, c)p(f|c, b)p(g|c). \end{aligned} \quad (23)$$

We are interested in retrieving the marginals in the SBN using VI. . We successively consider 3 VI techniques: MF VI (Figure [6a](#))¹, MT VI (Figure [6b](#)) and STMT VI (with auxiliary nodes) (Figure [6c](#)).

The developments of the previous section can be straightforwardly used to conduct VI over DMT and STMT. Note that for the STMT VI, we also

¹This approach is the most popular and the associated equations have not been developed in this article for brevity. Many resources cover the topic, *e.g.* [14](#).

need to provide the SBN with auxiliary nodes [1]. We need to keep the property that $p(\mathbf{x}, \mathbf{v}) = p(\mathbf{x})p(\mathbf{v}|\mathbf{x})$, where $\mathbf{x} = \{a, a^\leftarrow, a^\rightarrow, b, c, d, e, f, g\}$ and $\mathbf{v} = \{b^\leftarrow, b^\rightarrow, c^\leftarrow, c^\rightarrow, d^\leftarrow, d^\rightarrow, e^\leftarrow, e^\rightarrow, f^\leftarrow, f^\rightarrow, g^\leftarrow, g^\rightarrow\}$; in order to ensure that $p(\mathbf{x})$ (Equation 23) is the same between the three VI procedures. In STMT VI, p is then:

$$\begin{aligned}
 p(\mathbf{x}, \mathbf{v}) = & p(\mathbf{x})p(b^\leftarrow | b, a^\rightarrow)p(b^\rightarrow | a, a^\rightarrow)p(c^\leftarrow | a, a^\leftarrow)p(c^\rightarrow | a, a^\leftarrow) \\
 & p(d^\leftarrow | c^\rightarrow, b^\leftarrow)p(d^\rightarrow | b^\leftarrow, c^\rightarrow)p(e^\leftarrow | b^\rightarrow, c^\leftarrow)p(e^\rightarrow | c^\leftarrow, b^\rightarrow) \\
 & p(g^\leftarrow | c^\rightarrow, b^\leftarrow)p(g^\rightarrow | b^\leftarrow, c^\rightarrow)p(f^\leftarrow | b^\rightarrow, c^\leftarrow)p(f^\rightarrow | c^\leftarrow, b^\rightarrow),
 \end{aligned} \tag{24}$$

with

$$\begin{aligned}
 p(b|a, a^\leftarrow) &= p(b^\leftarrow | a, a^\leftarrow) = p(c^\leftarrow | a, a^\leftarrow), \\
 p(c|a, a^\rightarrow) &= p(b^\rightarrow | a, a^\rightarrow) = p(c^\rightarrow | a, a^\rightarrow), \\
 p(d|b) &= p(d^\leftarrow | b^\leftarrow) = p(e^\leftarrow | b^\leftarrow), \\
 p(e|b, c) &= p(d^\rightarrow | b^\leftarrow, b^\rightarrow) = p(f^\leftarrow | c^\leftarrow, b^\rightarrow), \\
 p(f|c, b) &= p(g^\leftarrow | c^\rightarrow, c^\leftarrow) = p(e^\rightarrow | b^\rightarrow, c^\leftarrow), \\
 p(g|c) &= p(f^\rightarrow | c^\rightarrow) = p(g^\rightarrow | c^\rightarrow).
 \end{aligned}$$

The model p with auxiliary nodes is described in Figure 5b

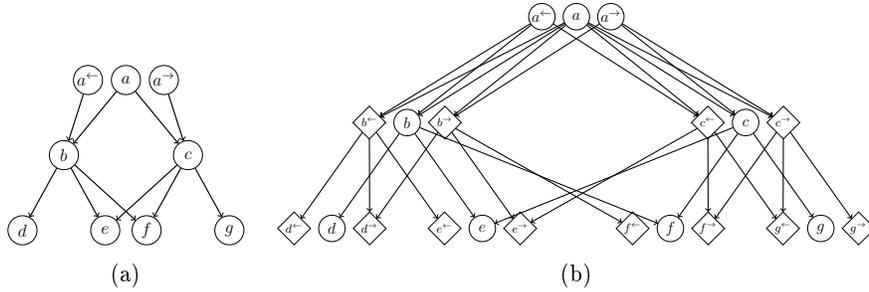


Fig. 5: (a) Adapted SBN, (b) Adapted SBN with auxiliary nodes.

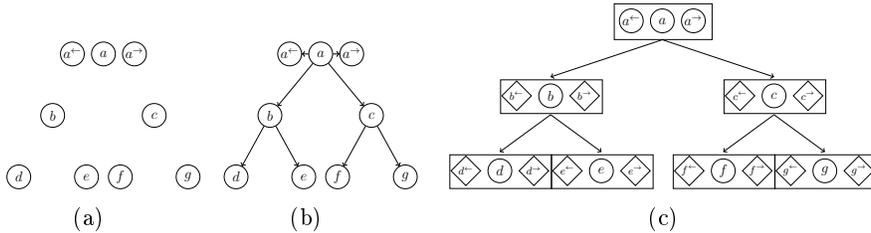


Fig. 6: Variational distributions for the VI procedure. (a) MF VI, (b) DMT VI and (c) STMT VI.

The random variables are chosen with values in $\{0, 1\}$. Our goal is to estimate the *true* marginals $p(x = 1), \forall x \in \mathbf{x}$ of the SBN of Figure 5a. In this synthetic example those true marginals are given as well as the transitions in p (we do not cover the parameter estimation problem): the transitions of Equation 23 which also totally define Equation 24 are taken randomly.

4.2 Results

We define a marginal *error* for a variational distribution q and a random variable x by $e_q(x) = p(x = 1) - q(x = 1)$. These errors are computed and stored over 1000 different SBNs p whose transitions are randomly chosen. Figure 7 depicts the values of the Kullback-Leibler divergence of the three VI procedures. We see a rapid convergence which is related to the small size of the considered SBN. We then choose to stop the VI process after 30 iterations. The value for MF VI and DMT VI are comparable and are shown on the same graph. The minimization is better in the case of DMT VI, the structured VI; this is reflected in Figure 8 which illustrates the goodness of the estimated marginals.

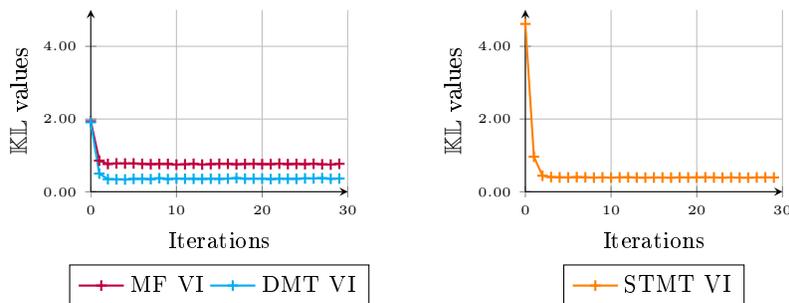


Fig. 7: The values of the cost function (Kullback-Leibler divergence values) of the minimization problem for the three VI procedures. Note that the STMT VI cost function integrates auxiliary variables and is not comparable with the others, hence it is plotted in another graph.

The analysis of Figure 8 illustrates the interest of the STMT structure to approximate the marginals of SBN: this VI procedure exhibits in all cases the smallest error with respect to the true marginals. We also observe that progressively adding structure increases the quality of the approximation since MF VI performs worse than DMT VI which performs in turn worse than STMT VI. Note that the left/right symmetry of the SBN can be found in the behaviour of the error rates: d is similar to g , e is similar to f , and so on. Moreover, we notice that the DMT VI performs worse for random variables b and c . This can be explained by the fact that those variables are central in the SBN and are more correlated with others variables. We note also that the errors computed at nodes b and c in the STMT VI remain stable.

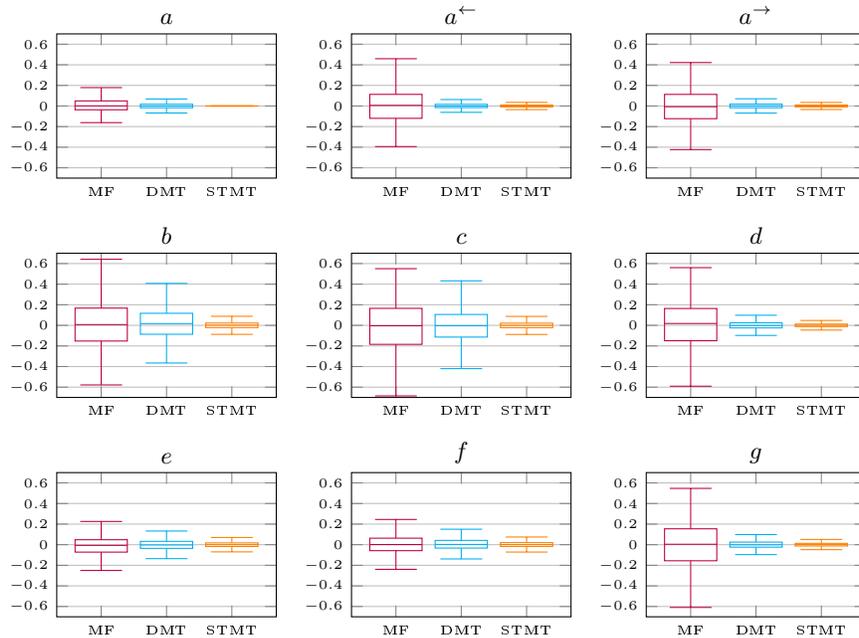


Fig. 8: Boxplots to illustrate the error with respect to the true marginal, for each node, for the three VI procedures. The series of errors spans over 1000 different experiments (different p transitions randomly chosen).

5 Conclusion

In this article we explored the potential of adding auxiliary random variables in order to develop more complex and rich correlations in probabilistic models. We showed that the triplet Markov framework enables to handle the additional auxiliary variables while preserving tractability of computations and analytical solutions.

We developed the STMT model and illustrated its potential as a variational distribution to approximate a much more complex Bayesian network: the SBN model. STMT performed better than the classical MF variational inference but also than the DMT variational inference methods.

Further work might consider studying theoretical results to quantify and qualify the relation of the SBN and STMT models. We might also extend the developed algorithms from dyadic trees to quadrees [11] [5] in order to treat bi-dimensional data such as images.

References

- [1] F. V. Agakov and D. Barber. “An auxiliary variational method”. In: *International Conference on Neural Information Processing*. Springer, 2004, pp. 561–566.

- [2] C. Bishop and J. Winn. “Structured variational distributions in VIBES”. In: (2003).
- [3] D. M. Blei et al. “Variational inference: A review for statisticians”. In: *Journal of the American statistical Association* 112.518 (2017), pp. 859–877.
- [4] J.-B. Courbot et al. “Oriented triplet Markov fields”. In: *Pattern Recognition Letters* 103 (2018), pp. 16–22.
- [5] J.-B. Courbot et al. “Triplet Markov Trees for Image Segmentation”. In: *2018 IEEE Statistical Signal Processing Workshop (SSP)*. 2018, pp. 233–237.
- [6] J.-B. Durand et al. “Computational methods for hidden Markov tree models—An application to wavelet trees”. In: *IEEE Transactions on Signal Processing* 52.9 (2004), pp. 2551–2560.
- [7] Z. Ghahramani and M. I. Jordan. “Factorial hidden Markov models”. In: *Advances in Neural Information Processing Systems*. 1996, pp. 472–478.
- [8] H. Hanzouli-Ben Salah et al. “A framework based on hidden Markov trees for multimodal PET/CT image co-segmentation”. In: *Medical physics* 44.11 (2017), pp. 5835–5848.
- [9] M. I. Jordan et al. “An introduction to variational methods for graphical models”. In: *Machine learning* 37.2 (1999), pp. 183–233.
- [10] D. P. Kingma. “Variational inference & deep learning: A new synthesis”. PhD thesis. 2017.
- [11] J.-M. Laferté et al. “Discrete Markov image modeling and inference on the quadtree”. In: *IEEE Transactions on image processing* 9.3 (2000), pp. 390–404.
- [12] P. Lanchantin et al. “Unsupervised segmentation of randomly switching data hidden with non-Gaussian correlated noise”. In: *Signal Processing* 91.2 (2011), pp. 163–175.
- [13] S. L. Lauritzen and D. J. Spiegelhalter. “Local computations with probabilities on graphical structures and their application to expert systems”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 50.2 (1988), pp. 157–194.
- [14] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [15] V. Olariu et al. “Modified variational Bayes EM estimation of hidden Markov tree model of cell lineages”. In: *Bioinformatics* 25.21 (2009), pp. 2824–2830.
- [16] R. Ranganath et al. “Hierarchical variational models”. In: *International Conference on Machine Learning*. 2016, pp. 324–333.
- [17] T. Salimans, D. A. Knowles, et al. “Fixed-form variational posterior approximation through stochastic linear regression”. In: *Bayesian Analysis* 8.4 (2013), pp. 837–882.
- [18] C. Zhang et al. “Advances in variational inference”. In: *IEEE transactions on pattern analysis and machine intelligence* 41.8 (2018), pp. 2008–2026.
- [19] P. Zwiernik. *Semialgebraic statistics and latent tree models*. Chapman and Hall/CRC, 2015.

Invariant description for regret of UCB strategy for Gaussian multi-arm bandit

Sergey Garbar* and Alexander Kolnogorov*

* Department of applied mathematics and information science, Novgorod State University, ul. B. S.-Peterburgskaya, d. 41, 173003 Velikiy Novgorod, Russia
(e-mails: Sergey.Garbar@novsu.ru, Alexander.Kolnogorov@novsu.ru)

Abstract. A variation of upper confidence bound strategy for Gaussian multi-armed bandit is considered. Invariant descriptions with the control horizon equal to 1 are obtained for upper bounds and for regret. A set of simulations are performed for different settings of MABs to determine with the control horizon that is equal to 1 and maximum regrets.

Keywords: multi-armed bandit problem, UCB rule, invariant description, gaussian multi-armed bandit.

1 Introduction

We consider multi-armed bandit (MAB) problem, which can be imagined as a slot machine that has two or more arms (levers). Choosing an arm yields some random income associated with it (Berry and Fristedt[1]). The gambler (decision-making agent) begins with no initial knowledge about rewards (incomes) associated with the arms. The goal of the gambler is to maximize the total expected reward. To this end, during the game he or she should determine the arm that has the largest corresponding one-step reward to use it predominantly. The problem is also known as the problem of expedient behavior (Tsetlin[2]) and of adaptive control in a random environment (Sragovich[3]). This is a classic reinforcement learning problem that exemplifies the exploration–exploitation tradeoff dilemma, so it is also faced in machine learning (Auer[4], Lugosi[5]). MABs have also been used to model problems such as managing research projects in a large organization like a science foundation or a pharmaceutical company (Berry and Fristedt[1], Gittins[6]).

Further we consider Gaussian MAB with J arms. Formally it is a controlled random process $\xi(n), n = 1, 2, \dots, N$. Value $\xi(n)$ at time n only depends on the currently chosen arm y_n and is interpreted as reward and has a normal distribution with probability density function

$$f_D(x|m_l) = (2\pi D)^{-1/2} e^{-\frac{(x-m_l)^2}{2D}}$$

if $y_n = l$, and $l = 1, \dots, J$. Variance D is assumed to be known and equal for all the arms and expected values m_1, \dots, m_J are assumed to be unknown. The

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



requirement for a prior knowledge of variance can be omitted as the algorithm that we consider further changes only slightly when the variance is changed moderately (e.g. 5–10% change). Hence the variance can be estimated during the initial control stage. Gaussian MAB arises when batch data processing is optimized and there are two or more processing methods available with different a priori unknown efficiencies Kolmogorov[7].

So, each arm has an associated mean reward, which is unknown and remains fixed for the duration of control. Reward can be represented as this given mean plus Gaussian random variable with zero mean and variance D .

A control strategy σ determines a choice of action y_n depending of currently available information about process history. Below we only consider with the strategy proposed in Bather[8]. Suppose that at the step n the l -th arm was chosen n_l times and let $X_l(n)$ denote corresponding cumulative reward (for $l = 1, \dots, J$). In this case $X_l(n)/n_l$ is an estimator of the mean reward m_l for this arm. Since the goal is to maximize the total expected reward, it might seem reasonable always to apply the action corresponding to currently largest value $X_l(n)/n_l$. However, such a rule can result in a significant losses due to the fact that initial estimate $X_l(n)/n_l$, corresponding to the largest m_l , can by chance take a lower value and consequently this action will be never applied.

Instead of estimates of $\{m_l\}$ themselves it is proposed to consider the upper bounds of their confidence intervals

$$U_l(n) = \frac{X_l(n)}{n_l} + \frac{aD^{1/2}}{n_l^{1/2}} (2 + \zeta_l(n)),$$

where $a > 0$ and $\zeta_l(n)$ are i.i.d. random variables with exponential distribution; $l = 1, 2, \dots, J$; $n = 1, 2, \dots, N$. This choice of an arm negotiates the exploration—exploitation trade-off: it greedily selects an arm with the highest estimated mean, but on the other hand there is an exploration term $(2 + \zeta_l(n))$ that can be derived from an experimental design approach with the notion of information gain.

Discussed strategy prescribes to initially apply once every action. Then at each point of time n it is necessary to choose the action corresponding to the highest value $\{U_l(n)\}$. Strategies of that kind are called UCB (upper confidence bound) rules.

Strategy that is reviewed in this paper is equivalent to the strategy described in Bather[8] up to summands of order n_l^{-1} when $a = 2/15$, but there it is applied to Bernoulli MAB problem, therefore one should use $D = 0.25$ as it is the maximum value for the variance of Bernoulli one-step income. It is stated in Bather[8] that at $J = 2$ (i.e. when 2-armed bandit is considered) the maximal expected normalized regret (scaled to the value $(DN)^{1/2}$) does not exceed 0.72 for large N . However, explanation of this result is not presented in Bather[8] and, to the best of our knowledge, was not published later.

To explain this estimation of the regret for the rule we aim to build an invariant description of the control strategy on the unit horizon in the domain of “close” distributions, as in case of “close” distributions the maximum values of expected regret are attained. We consider the batch version of the strategy described in Bather[8] and show that expected regret only depends on number of processed

batches and some invariant characteristics of the parameter. Note that batch (parallel) strategies are important when processing time of the datum is significant, because in this case the total processing time depends on the number of batches rather than on the total number of data. The found value of maximum normalized regret for the batch UCB rule is 0.76 for 2-armed bandit, i.e. is almost the same as in Bather[8].

2 Obtaining an invariant description of the considered control algorithm

Considered MAB can be described with a vector parameter $\theta = (m_1, \dots, m_J)$. For applied strategy σ the total expected reward is less than maximally possible by the value which is called the regret (the loss function). The regret after N rounds is defined as the expected difference between the reward sum associated with an optimal strategy and the sum of the collected rewards and is equal to

$$L_N(\sigma, \theta) = E_{\sigma, \theta} \left(\sum_{n=1}^N (\max(m_1, \dots, m_J) - \xi_n) \right).$$

Here $E_{\sigma, \theta}$ denotes the expected value calculated with respect to measure generated by strategy σ and parameter θ . We aim to estimate the upper bound of the maximum regret calculated over the set of acceptable values of parameter which is chosen as follows

$$\Theta = \{m_l = m + d_l(D/N)^{1/2}; m \in (-\infty, +\infty), |d_l| \leq C < \infty, l = 1, \dots, J\}.$$

This is set of parameters describes “close” distributions. Their definitive feature is the difference between expected values of the order $N^{-1/2}$.

Maximal normalized regrets are observed on that domain and have the order $N^{1/2}$ (see Vogel[9]). For “distant” distributions the normalized regrets have smaller values. For example, they have order $\log N$ if $\max(m_1, \dots, m_J)$ exceeds all other $\{m_l\}$ by some $\delta > 0$ (see Lai *et al* [10]).

Further we consider strategies that can change the arm only after using it M times in a row. These strategies allow batch (and also parallel) processing. We assume for simplicity that $N = MK$ where K is the number of batches. For batch strategy the upper bounds take form

$$U_l(k) = \frac{X_l(k)}{k_l} + \frac{a(MD)^{1/2}}{k_l^{1/2}} (2 + \zeta_l(n)),$$

where k is the number of processed batches, k_l is the number of batches for which l -th arm was chosen and $X_l(k)$ is the corresponding cumulative reward after processing k batches ($k = 1, 2, \dots, K$).

We denote by

$$I_l(k) = \begin{cases} 1, & \text{if } U_l(k) = \max(U_1(k), \dots, U_J(k)), \\ 0, & \text{otherwise} \end{cases}$$

the indicator of chosen action for processing the $(k + 1)$ -th batch according to considered rule at $k > J$ (recall that at $k \leq J$ every arm is chosen once for a batch). Note that with probability 1 only one of values of $\{I_l(k)\}$ equals to 1.

Using this notation, we can write out cumulative reward for each arm as

$$X_l(k) = k_l M \cdot (m + d_l (D/N)^{1/2}) + \sum_{i=1}^k I_l(i) \eta_l(MD; i)$$

where $\eta_l(MD; i) \sim N(0, \sqrt{MD})$ are i.i.d. normally distributed random variables with zero means and variances equal to MD .

Note that for arm l indicator equals 1 exactly k_l times, so $\sum_{i=1}^k I_l(i) \eta_l(MD; i)$ is the sum of k_l Gaussian random variables, which can be presented a scaled by standard deviation standard normal random variable. In that case

$$X_l(k) = k_l M \cdot (m + d_l (D/N)^{1/2}) + \sqrt{k_l} (MD)^{1/2} \eta,$$

where $\eta \sim N(0, 1)$ is a standard normal random variable.

If we introduce the following notation:

$$t = kK^{-1}, t_l = k_l K^{-1}, \varepsilon = K^{-1},$$

we can write the upper bounds for batch strategy as

$$U_l(k) = Mm + d_l \left(\frac{MD}{K}\right)^{1/2} + \frac{\sqrt{k_l MD} \eta}{t_l K^{1/2}} + \frac{a \sqrt{MD}}{t_l^{1/2} K^{1/2}} (2 + \zeta_l(n)), l = 1, 2, \dots, J.$$

We can apply the following linear transformation which does not change the arrangement of bounds:

$$u_l(t) = (U_l(k) - Mm) \left(\frac{K}{MD}\right)^{1/2}.$$

This way we obtain an expression for choosing the arm in the UCB strategy with the control horizon that is equal to 1, i.e. in the invariant form:

$$u_l(t) = d_l + \frac{\eta}{t_l^{1/2}} + \frac{a}{t_l^{1/2}} (2 + \zeta_l(t)), l = 1, 2, \dots, J.$$

Next task is to find the expression for regret. Let us assume without loss of generality, that $d_1 = \max(d_1, \dots, d_J)$. In that case

$$\begin{aligned} L_N(\sigma, \theta) &= \sqrt{\frac{D}{N}} \sum_{l=2}^J (d_1 - d_l) E_{\sigma, \theta} \left(\sum_{n=1}^K M I_l(k) \right) = \sqrt{\frac{D}{N}} \sum_{l=1}^J (d_1 - d_l) M E_{\sigma, \theta}(k_l) \\ &= \sqrt{\frac{D}{N}} \sum_{l=1}^J (d_1 - d_l) E_{\sigma, \theta}(t_l). \end{aligned}$$

In normalized form (scaled by $(DN)^{1/2}$) we get the following expression for regret:

$$(DN)^{-1/2} L_N(\sigma, \theta) = \sum_{l=2}^J (d_1 - d_l) E_{\sigma, \theta} \left(\sum_{n=1}^K \varepsilon I_l(k) \right) = \sum_{l=2}^J (d_1 - d_l) E_{\sigma, \theta}(t_l).$$

We can present the described results in form of the following theorem:

Theorem 2.1. For Gaussian multi-armed bandit with J arms, fixed known variance D and unknown expected values m_1, \dots, m_J the usage of the batch UCB rule with bounds

$$U_l(k) = \frac{X_l(k)}{k_l} + \frac{a(MD)^{1/2}}{k_l^{1/2}}(2 + \zeta_l(n)),$$

results in invariant description on the unit control horizon which is described by

$$u_l(t) = d_l + \frac{\eta}{t_l^{1/2}} + \frac{a}{t_l^{1/2}}(2 + \zeta_l(t)).$$

For normalized (scaled by $(DN)^{1/2}$) regret the expression

$$(DN)^{-1/2}L_N(\sigma, \theta) = \sum_{l=2}^J (d_1 - d_l)E_{\sigma, \theta} t_l$$

holds.

As we obtained an invariant description for losses in UCB strategy for the Gaussian MAB, we can justify using Monte-Carlo method for studying the corresponding model.

3 Simulation results

To study the normalized regret values for the UCB rule for Gaussian MAB we perform the following tasks: first we determine the optimal value for the parameter a of the UCB rule. Then we find the maximum regrets for cases of 2-armed and 3-armed bandits and conditions where they occur. Finally, we try to find worst-case regrets for MAB. As there is an invariant description for the strategy, a Monte-Carlo simulation method can be used to carry out these tasks. In all the following simulations we take $d_1 = 0$ and d_j is always shown on the horizontal axis of the figures. In this case we can also regard d_j as a difference between mean rewards of bandit's best (most lucrative) and worst arms.

Figure 1 shows how normalized regret depends on parameter a of the model in interval $[1.0, 6.0]$ where losses are the highest. Horizon size of 400 was considered with regret averaged over 4000 simulations.

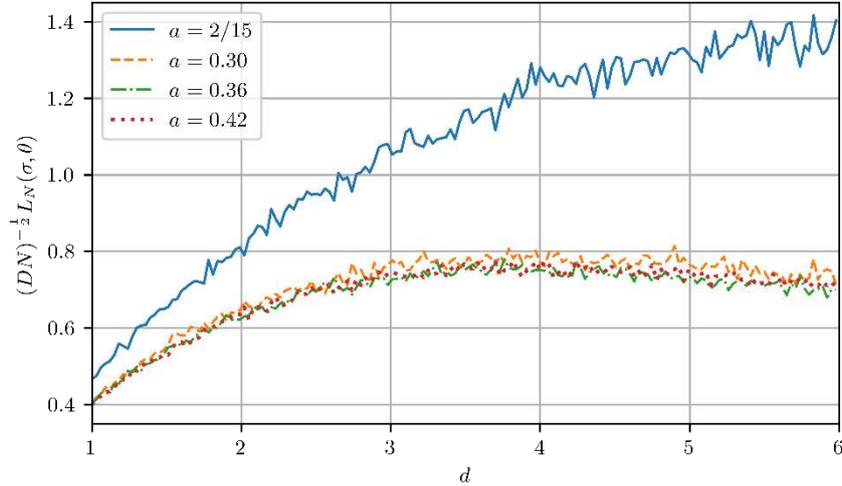


Fig. 1. Relation between difference between mean rewards and regrets for different values of parameter a for considered UCB rule for arms for 2-arm bandit

Optimal value (that is the one that yields the lowest maximum regret) for a can be found in the interval $[0.33, 0.39]$, which is rather different from $2/15$ that was found in Bather[8]. That might be due to a fact that in Bather[8] rather small control horizons ($N = 50$) were considered. In this case added terms of order n_l^{-1} which are present in the proposed rule can considerably affect values $\{U_l(n)\}$. While figure 1 shows the relation between regret and parameter a for the 2-armed bandit, we checked that the value is approximately the same in case of 3-armed bandit also. d_2 is shown on the horizontal axis.

Hence, for the following simulations value $a = 0.36$ is used.

Figure 2 shows relation between difference between mean rewards of arms d_2 for 2-armed bandit and normalized regret averaged over 10000 simulations for 2-armed bandit ($J = 2$). We consider horizons $N = 100, 400, 1500$ shown here by different lines. We indeed see that regret does not differ much between $N = 400$ and $N = 1500$. Maximum normalized regret is approximately equal to 0.76 and is reached when $d_2 = d_2 - d_1 \approx 3.9$.

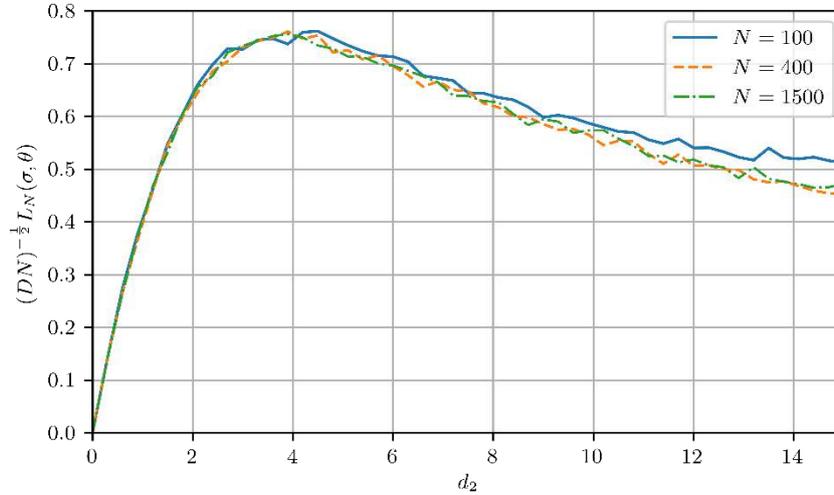


Fig. 2. Relation between difference between mean rewards d_2 for arms of 2-armed bandit and normalized regret

Next, we study 3-armed bandit to determine what is the worst case for the regret and how it depends on relation between means of rewards of different arms. We set $d_1 = 0$ and then consider different values for d_2 : figure 3 shows cases where $d_2 = 0, 1, 2, 3, 4$ as different lines, and d_3 is shown on the horizontal axis. Each point is calculated as an average regret over 10000 simulations, control horizon was chosen $N = 400$. Worst case for normalized regret 1.29 was reached when $d_1 = d_2 = 0$ and $d_3 \approx 4.5$. An explanation for this can be a fact that more lucrative strategy has not one but two competitor strategies with close distributions.

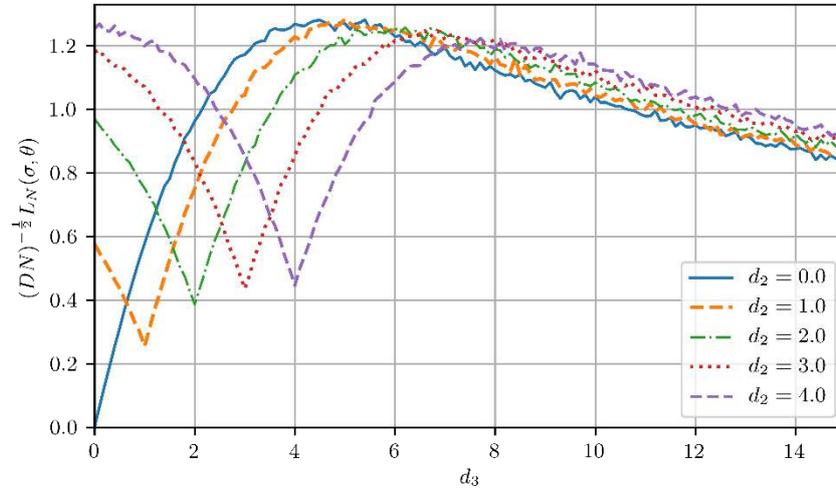


Fig. 3. Relation between mean rewards of 3-armed bandit and normalized regret: $d_1 = 0$, d_2 is shown with different lines, d_3 is on the horizontal axis

Next assume that the J -armed bandit with $(J - 1)$ arms having the same mean reward (i.e. $d_1 = \dots = d_{J-1} = 0$) and one arm having a reward with a bigger mean will have high regret. Figure 4 shows how normalized regret is related to difference between mean reward of the best arm d_j compared to other arms. Different lines show this relation for MABs with different number of arms J . Black dots mark maximum regret for each MAB of such configuration. The data for the figure was averaged over 5000 simulations.

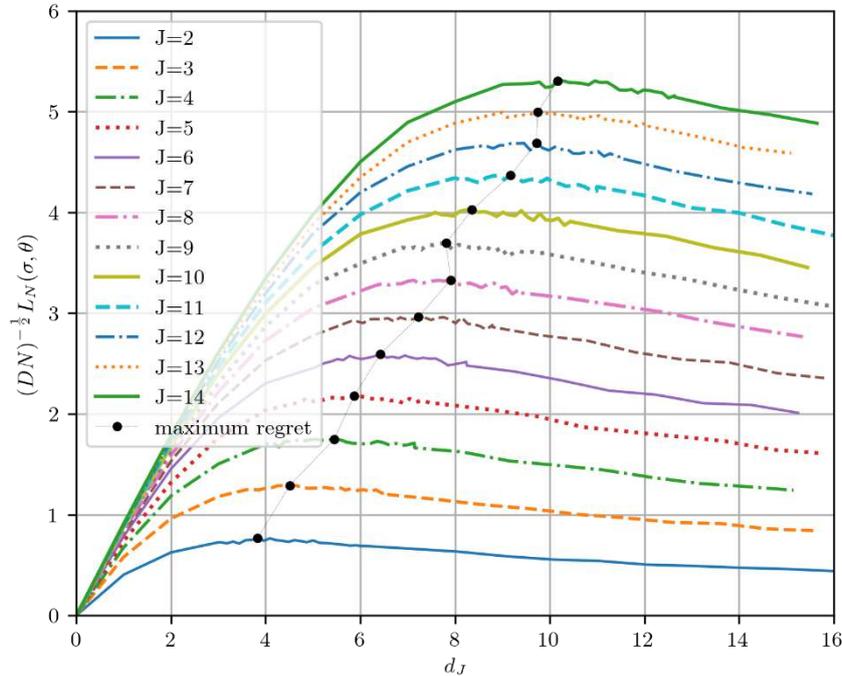


Fig. 4. Relation between mean rewards of J -armed bandit and normalized regret: $d_1 = \dots = d_{J-1} = 0, d_J$ if on the horizontal axis. Black dots mark maximum regret for each MAB of such configuration

We can see that relation between number of bandit's arms J and maximum regret is close to linear for considered values of J (coefficient of determination is $R^2 = 0.994$ for linear regression model).

Conclusions

We reviewed a variant of UCB rule proposed in Bather[8] and applied it for Gaussian MAB.

An invariant description for strategy and worst-case normalized regrets were found.

Values for worst-case normalized regrets were found with a set of Monte-Carlo simulations with fairly large control horizons.

The reported study was funded by RFBR, project number 20-01-00062.

References

1. Berry, D. A. and Fristedt, B. Bandit Problems: Sequential Allocation of Experiments, Chapman and Hall, London, New York, 1985
2. Tsetlin, M. L. Automaton Theory and Modeling of Biological Systems, Academic Press, New York, 1973

3. Sragovich, V. G. *Mathematical Theory of Adaptive Control*, World Sci., Singapore, 2006
4. Auer, P. Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine Learning Research*. 3 397–422., 2002
5. Lugosi, G. and Cesa-Bianchi, N. *Prediction, Learning and Games*. University Press, New York, Cambridge, 2006
6. Gittins, J. C. *Multi-armed bandit allocation indices*, Wiley-Interscience Series in Systems and Optimization., Chichester: John Wiley & Sons, Ltd., 1989
7. Kolmogorov, A.V. Parallel design of robust control in the stochastic environment (the two-armed bandit problem). *Automation and Remote Control* 73 689–701, 2012
8. Bather, J.A. The Minimax Risk for the Two-Armed Bandit Problem. *Mathematical Learning Models Theory and Algorithms*. Lecture Notes in Statistics. SpringerVerlag, New York. 20 1–11, 1983
9. Vogel, W. An Asymptotic Minimax Theorem for the Two-Armed Bandit Problem. *Ann. Math. Statist.* 31 444–451, 1960
10. Lai, T.L., Levin, B., Robbins, H. and Siegmund, D. Sequential Medical Trials (Stopping Rules/Asymptotic Optimality). *Proc. Nati. Acad. Sci. USA*. 77 3135–3138, 1980

America's Zika virus and its Similarities with African and Asian Lineages

Jesús E. García * V.A. González-López †

July 5, 2020

Abstract

According to the literature, see [1] there are two Zika lineages, the African (East and West) and the Asian. Phylogenetic studies expose the similarity between the Americas virus and the African genotypes [2], and also those studies report a strong correspondence between the Americas virus and the Asian strains that circulated in French Polynesia during the 2013–2014 outbreak [1]-[2]-[3]. It has been postulated that the virus originated in East Africa and then spread into both West Africa and Asia around 80-100 years ago. And that the Asian genotype has been gradually evolving and spreading geographically throughout Asia and the Pacific Islands. Thus, a conjecture of closer proximity is established between America's Zika and Asia's Zika compared to the proximity between America's Zika and Africa's Zika. In this paper, we compare genomic sequences coming from America with collections from Africa and Asia respectively, in order to define if the sequences of America are more similar to the Asian or to the African ones. To proceed with the comparison, we apply a stochastic metric between sequences (see [4]). We also classify the sequences coming from each set (i) the American collection, (ii) the Asian collection, (iii) the African collection defining for each set a sequence that represents the collection (see [4], [5]). With these representative sequences, we show that there is a greater similarity between the sequence coming from America and the Asian one. Based on the Partition Markov Model (see [6]), we have revealed the stochastic organization of these sequences that justifies our discovery.

Keywords: Partition Markov Models; Metric between Processes; Bayesian Information Criterion

1 Introduction

In this paper, we look for evidence that allows us to discard or support the following geographical assumption about the Zika genetic structure, the sequences coming from America are closer to the sequences coming from Asia than those coming from Africa. There is a range of studies in the field of the evolution of Zika that reveals a possibility of transformation of the genetic structure. It has been postulated that the virus originated in East Africa and then spread into both West Africa and Asia around 80-100 years ago. And that the Asian genotype has been gradually evolving and spreading geographically throughout Asia and the Pacific Islands. Here the problem is placed since this last place is considered a possible entrance door of the genetic version that prevails in the sequences of America.

The theoretical elements that allow giving an answer to these questions come from the area of stochastic processes. Among such notions, which we will address later, we highlight three. The first of them is a metric that allows deciding whether two samples, coming from stochastic processes, follow the same stochastic law or not (see [4]). The second is an indicator that, based on a collection of samples from independent stochastic processes, it allows selecting the sample that best represents the collection (see [5]). In relation to the third notion, this is constituted by a model that is built using a collection of samples from stochastic processes and which allows us to reveal what the

*Department of Statistics, University of Campinas, Sergio Buarque de Holanda, 651, Campinas, S.P., CEP: 13083-859, Brazil. E-mail: jg@ime.unicamp.br

†Department of Statistics, University of Campinas, Sergio Buarque de Holanda, 651, Campinas, S.P., CEP: 13083-859, Brazil. E-mail: veronica@ime.unicamp.br

samples have in common and what they do not have in common, in terms of units of the state space (see [6]). In this article, we consider each genomic and complete sequence of Zika (in FASTA format) as a sample coming from a stochastic process of finite order and arranged in a finite alphabet, which is the genetic alphabet. Such abstraction has already been used in the case of sequences coming from America, but with the purpose of modeling the profile of Zika in America (see [6]).

The organization of this paper is as follows; section 2 introduces the notions we will use as well as the notation. Section 3 describes the data and its source. Section 4 shows the results and section 5 shows the conclusion.

2 The Markovian Model

The section gives the theoretical framework of the paper. The notions of proximity between processes represented by samples and the criterion of classification of samples are introduced. In the latter case, a sample that best represents the process is identified from a collection of samples. Such a representative sample, by Zika's lineage, can be used to compare Zika's lineages. The section concludes by showing the model that allows us to identify, from a collection of samples (a sample by provenance), the states that operate stochastically equivalently (in terms of the transition probabilities).

Let (X_t) be a discrete time Markov chain on a finite alphabet A with finite order o . Let us call $\mathcal{S} = A^o$ the state space, denote $a_m^n = a_m a_{m+1} \dots a_n$ where $a_i \in A$, $m \leq i \leq n$. For each $a \in A$ and $s \in \mathcal{S}$ define the transition probability $P(a|s) = \text{Prob}(X_t = a | X_{t-o}^{t-1} = s)$. If x_1^n is a sample coming from the stochastic process (X_t) , the number of occurrences of s in the sample x_1^n is denoted by $N_n(s)$ and the number of occurrences of s followed by a in the sample x_1^n is denoted by $N_n(s, a)$. Then, $\frac{N_n(s, a)}{N_n(s)}$ is the estimator of the transition probability $P(a|s)$. Consider now, two Markov chains $(X_{1,t})$ and $(X_{2,t})$, of order o , arranged on the finite alphabet A with state space \mathcal{S} . Given $s \in \mathcal{S}$ denote by $\{P(a|s)\}_{a \in A}$ and $\{Q(a|s)\}_{a \in A}$ the sets of transition probabilities of $(X_{1,t})$ and $(X_{2,t})$ respectively. The local metric d_s , introduced by [4], is given now, and it allows us to establish how far or near the samples are.

Definition 2.1. Consider two Markov chains $(X_{1,t})$ and $(X_{2,t})$, of order o , with finite alphabet A , state space $\mathcal{S} = A^o$ and independent samples $x_{1,1}^{n_1}, x_{2,1}^{n_2}$ respectively.

i. For a string $s \in \mathcal{S}$,

$$d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) = c_{\alpha, n_1, n_2} \sum_{a \in A} \left\{ \sum_{j=1,2} N_{n_j}(s, a) \ln \left(\frac{N_{n_j}(s, a)}{N_{n_j}(s)} \right) - N_{n_1+n_2}(s, a) \ln \left(\frac{N_{n_1+n_2}(s, a)}{N_{n_1+n_2}(s)} \right) \right\},$$

ii.

$$d_{\max}(x_{1,1}^{n_1}, x_{2,1}^{n_2}) = \max_{s \in \mathcal{S}} \{d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2})\},$$

with $c_{\alpha, n_1, n_2} = \frac{\alpha}{(|A|-1) \ln(n_1+n_2)}$, $N_{n_1+n_2}(s, a) = N_{n_1}(s, a) + N_{n_2}(s, a)$, $N_{n_1+n_2}(s) = N_{n_1}(s) + N_{n_2}(s)$, where N_{n_1} and N_{n_2} are given as usual, computed from the samples $x_{1,1}^{n_1}$ and $x_{2,1}^{n_2}$ respectively. With α a real and positive value.

In [4] is proved that d_s is a metric, this is d_s is such that a) $d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) \geq 0$ with equality $\Leftrightarrow \frac{N_{n_1}(s, a)}{N_{n_1}(s)} = \frac{N_{n_2}(s, a)}{N_{n_2}(s)} \quad \forall a \in A$; b) $d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) = d_s(x_{2,1}^{n_2}, x_{1,1}^{n_1})$ and c) $d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) \leq d_s(x_{1,1}^{n_1}, x_{3,1}^{n_3}) + d_s(x_{3,1}^{n_3}, x_{2,1}^{n_2})$. The two notions introduced by definition 2.1 are statistically consistent, then, by increasing the $\min\{n_1, n_2\}$ grows their ability to detect *discrepancies* and *similarities*.

In the application we use $\alpha = 2$ (see definition 2.1-i.), with this value ($\alpha = 2$), to decide that the sequences follow the same law when $d_s < 1$, is equivalent to use the *Bayesian Information Criterion*, see [8] and [4].

To follow is introduced a notion that makes possible the classification of sequences that belong to a group of sequences.

Definition 2.2. Given a finite collection $\{x_{j,1}^{n_j}\}_{j=1}^m$ of independent samples from independent processes $\{(X_{j,t})\}_{j=1}^m$ with probabilities $\{P_j\}_{j=1}^m$, over the finite alphabet A , with state space $\mathcal{S} = A^o$ ($o < \infty$). For a fixed $i \in \{1, 2, \dots, m\}$ define

$$V(x_{i,1}^{n_i}) = \text{median}\{\text{dmax}(x_{i,1}^{n_i}, x_{j,1}^{n_j}) : j \neq i, 1 \leq j \leq m\}.$$

Where, given a sequence $\{z_j\}_{j=1}^l$, $\text{median}\{z_j, 1 \leq j \leq l\} = z_{(k+1)}$ if $l = 2k + 1$ and $\text{median}\{z_j, 1 \leq j \leq l\} = \frac{z_{(k)} + z_{(k+1)}}{2}$ if $l = 2k$, for k an integer and $z_{(j)}$ denoting the j th order statistic of the collection $\{z_j\}_{j=1}^l$.

With the V values attributed to each sample, we can proceed to order the samples, from lowest to highest value of V , in order to identify their classification. As we can perceive from the definition 2.2, low values of V indicate that these samples represent the whole group better, while high values of V indicate little representativeness. The next result, proved in [5], gives an adequate tool to classify sequences, according to their underlying laws. According to Theorem 1 of [5], under the assumptions of definition 2.2, for each i , $1 \leq i \leq m$, set $\xi_i = |\{j : 1 \leq j \leq m, P_j = P_i\}|$,

i.

$$V(x_{i,1}^{n_i}) \xrightarrow{\min\{n_1, \dots, n_m\} \rightarrow \infty} \infty \text{ if, and only if, } \xi_i \leq \lceil \frac{m}{2} \rceil.$$

ii.

$$V(x_{i,1}^{n_i}) \xrightarrow{\min\{n_1, \dots, n_m\} \rightarrow \infty} 0, \text{ if, and only if, } \xi_i > \lceil \frac{m}{2} \rceil.$$

Where $\lceil x \rceil$ is the smallest integer greater than or equal to x . The result guarantees that if at least 50% of the samples of the set follow the same law; each of them receives a value of V close to zero. And, if this does not happen, V takes arbitrarily large values identifying discrepancies in the generating laws of the sequences.

We now show an extension of the previous notions, specifically developed for a collection of independent stochastic processes, see [6]. Consider $\mathcal{F} = \{(X_t^j)\}_{j=1}^p$, a collection of p independent, discrete time, Markov chains on the same finite alphabet A . To simplify the notation, we will assume that all the processes have the same memory o . $\mathcal{S} = A^o$ is the state space of each Markov chain in the collection. For each $j \in J = \{1, 2, \dots, p\}$, $a \in A$ and $s \in \mathcal{S}$, $P^j(s) = \text{Prob}(X_{t-o}^{j,t-1} = s)$ and $P^j(a|s) = \text{Prob}(X_t^j = a | X_{t-o}^{j,t-1} = s)$. Now we define a space, on which the model is created, this space considers all the p independent processes and the state space, which is unique to all of them. Define $M = J \times \mathcal{S}$. To continue, we introduce the notion that defines the model.

Definition 2.3. Consider a collection of p independent processes $\mathcal{F} = \{(X_t^j)\}_{j=1}^p$ of p independent, discrete time, Markov chains on the same finite alphabet A , with memory o . $M = J \times \mathcal{S}$, $J = \{1, \dots, p\}$, $\mathcal{S} = A^o$,

i. the elements $(i, s), (j, r) \in M$, are equivalent if $P^i(a|s) = P^j(a|r)$ for all $a \in A$;

ii. the collection \mathcal{F} has Markov partition $\mathcal{L} = \{L_1, L_2, \dots, L_{|\mathcal{L}|}\}$ if \mathcal{L} is the partition of M defined by the relationship introduced in i.

The partition \mathcal{L} is minimal, that is, it has the smallest cardinal $|\mathcal{L}|$. Moreover, there are no two parts in the partition that share all the transition probabilities.

The following notation, applies the principle that for each part of the partition (definition 2.3), all the elements share the same probability. Then, if $\mathcal{F} = \{(X_t^j)\}_{j=1}^p$ has Markov partition $\mathcal{L} = \{L_1, L_2, \dots, L_{|\mathcal{L}|}\}$, for any $L \in \mathcal{L}$, we will denote for all $a \in A$,

$$P_L(a) = P^i(a|s) \text{ for any } (i, s) \in L. \quad (1)$$

Then, the model is completely specified once \mathcal{L} is estimated and also once it is estimated the set of probabilities conditioned to the structure \mathcal{L} . The estimation procedure is developed in [6] consisting of a metric on the space M that consistently identifies the partition given by definition 2.3.

The model allows us to reveal what a specific group of sequences have in common, thus revealing what is different about them.

In the next section, we describe the data and its source. We also detail the alphabet, order of the processes necessary for the computation of the notions 2.1, 2.2 and for building the model given by definition 2.3.

3 Data Set

All the sequences are in format *FASTA* obtained from the source of the National Center for Biotechnology Information Support Center (NCBI), available in <https://www.ncbi.nlm.nih.gov/nucleotide/>. We consider each genomic sequence as a sample, that is to say, that x_1^n is composed by the concatenation of elements of the alphabet A . As each sequence takes values in the genomic alphabet we define $A = \{a,c,g,t\}$, composed by the four bases: adenine (a), cytosine (c), guanine (g) and thymine (t). So, A has cardinal $|A| = 4$. In table 1, we give the identifier codes of the complete genomic sequences of Asia and Africa.

Lineage	Sequence	Year	Origin
Asian	KU312312.1	2015	Suriname
	KJ776791.2	2013	French Polynesia
	EU545988.1	2007	Yap Island
African	KF268948.1	2013	Central African Republic
	KF268949.1	2013	Central African Republic
	KF268950.1	2013	Central African Republic
	AY632535.2	2009	Uganda
	LC002520.1	2014	Uganda
	DQ859059.1	2006	Uganda

Table 1: Complete genomic sequences of Zika by lineage.

The list of complete sequences coming from America is given in table 2, separated by country (a total of $m = 153$ samples). The list of cases recorded in table 2 has been used to determine a unique model for the sequences of America, see [6]. And, as a precedent for research on the distance between the sequences of America, see [7].

The memory o allowed is such that $o < \log_{|A|}(n) - 1$, where n is the sample size coming from the sequence, in this case, $n \geq 10000$ for all the sequences. In the modeling problem of genomic sequences, the elements of A are organized in triples, so $o = 3, 6$ are the recommended orders. In the present study we use $o = 3$, then $S = A^o = \{a, c, g, t\}^3$.

The next section shows the results, bearing in mind that we seek an answer to the conjecture of greater proximity between the sequences of America (all of table 2) and Asia to that found between the sequences of America and Africa.

4 Results

Since most of the sequences are from Brazil, we first compared the 44 sequences from Brazil with the African and Asian lineages. The subsection 4.1 is intended for such a comparison. Subsection 4.2 the comparison is made between the American sequences and the Asian and African sequences. Finally, in section 4.2 and through definition 2.3 we expose the differences and similarities between the sequences properly identified by their origins: America, Asia, and Africa.

4.1 Comparison between Brazilian, Asian and African Sequences.

In figure 1, we show the *dm_{ax}* values organized in a dendrogram, see definition 2.1-ii. This graphic includes only the sequences of Brazil and those of Asian and African lineages. We see that the sequences from Asia are shown as the closer. And, the closest to the Brazilian ones are KU312312.1 (Suriname) and KJ776791.2 (French Polynesia).

Such a conclusion is verified by table 3, where we record the values of V , according to definition 2.2 and considering as the whole set of sequences the 44 sequences from Brazil, the 3 sequences from Asia, and the 6 sequences from Africa. V offers us an excellent notion of how Asian sequences can be seen as much closer to Brazilian in comparison with the African ones (with the highest values of V). Furthermore, it is possible to point out that sequence KU312312.1 (Suriname) could be considered the closest to the Brazilian set. The dendrogram (figure 1) reveals that there are clusters of Brazilian sequences that are far from the majority, for example (i) composed by KY785439.1, KY559004.1

Country	Sequences
Brazil (44)	KX197192.1, KY014296.2, KY014297.2, KY014301.2, KY014307.2 KY014308.2, KY014313.2, KY014317.2, KY014320.2, KY558999.1 KY559001.1, KY559003.1, KY559004.1, KY559005.1, KY559006.1 KY559007.1, KY559009.1, KY559010.1, KY559011.1, KY559012.1 KY559013.1, KY559014.1, KY559015.1, KY559017.1, KY559018.1 KY559019.1, KY559021.1, KY559023.1, KY559024.1, KY559027.1 KY559031.1, KY559032.1, KY785410.1, KY785426.1, KY785427.1 KY785429.1, KY785433.1, KY785437.1, KY785439.1, KY785450.1 KY785455.1, KY785456.1, KY785479.1, KY817930.1
US (34)	KX832731.1, KX842449.2, KX922703.1, KX922704.1, KX922705.1 KX922706.1, KX922707.1, KY014295.2, KY014298.1, KY014316.2 KY014325.2, KY014326.1, KY075932.1, KY075933.1, KY075934.1 KY075935.1, KY075936.1, KY325464.1, KY325465.1, KY325466.1 KY325467.1, KY325468.1, KY325469.1, KY325471.1, KY325472.1 KY325473.1, KY325476.1, KY325477.1, KY325479.1, KY785412.1 KY785445.1, KY785457.1, KY785459.1, KY785474.1
DOM (23)	KY014300.2, KY014302.3, KY014303.2, KY014304.2, KY014305.2 KY014314.2, KY014318.3, KY014321.2, KY785413.1, KY785415.1 KY785420.1, KY785423.1, KY785435.1, KY785441.1, KY785447.1 KY785449.1, KY785453.1, KY785463.1, KY785465.1, KY785470.1 KY785475.1, KY785476.1, KY785484.1
MEX (19)	MF801391.1, MF801395.1, MF801396.1, MF801398.1, MF801402.1 MF801403.1, MF801404.1, MF801406.1, MF801407.1, MF801408.1 MF801409.1, MF801410.1, MF801412.1, MF801413.1, MF801414.1 MF801417.1, MF801418.1, MF801420.1, MF801423.1
HND (13)	KY014306.2, KY014310.2, KY014312.2, KY014315.2, KY014319.2 KY014327.2, KY785414.1, KY785418.1, KY785442.1, KY785444.1 KY785448.1, KY785452.1, KY785461.1
NIC (7)	MF434516.1, MF434517.1, MF434518.1, MF434520.1, MF434521.1 MF434522.1, MF801426.1
JAM (4)	KY785419.1, KY785424.1, KY785430.1, KY785432.1
COL (3)	KY785417.1, KY785466.1, KY785469.1
PRI (2)	KY785462.1, KY785464.1
VEN (2)	KX702400.1, KX893855.1
CUB (1)	MF438286.1
MTQ (1)	KY785451.1

Table 2: Complete genomic sequences of Zika by country: Brazil (BRA), Colombia (COL), Cuba (CUB), Dominican Republic (DOM), Honduras (HND), Jamaica (JAM), Martinique (MTQ), Mexico (MEX), Nicaragua (NIC), Puerto Rico (PRI), United States (USA), Venezuela (VEN).

and KY817930.1 and (ii) composed by KY559001.1, KY559009.1, KY559010.1, KY014308.2 and KY559014.1. Consistently, we also note that the 8 sequences previously mentioned are those that receive the highest values of V (among the Brazilian ones), according to the records in table 3. KY785439.1 stands out with the highest value $V = 0.19626$.

4.2 Comparison between sequences coming from America, Africa and Asia.

The general comparison can be visualized using the dendrogram of figure 2 constructed from the d_{max} values (see definition 2.1-ii). We see (figure 2) that among the clusters furthest from the majority are two that contain the African sequences (i) composed of DQ859059.1 (Uganda), AY632535.2 (Uganda) and LC002520.1 (Uganda) and (ii) KF268949.1 (Central African Republic), KF268948.1 (Central African Republic), KF268950.1 (Central African Republic), KY785449.1

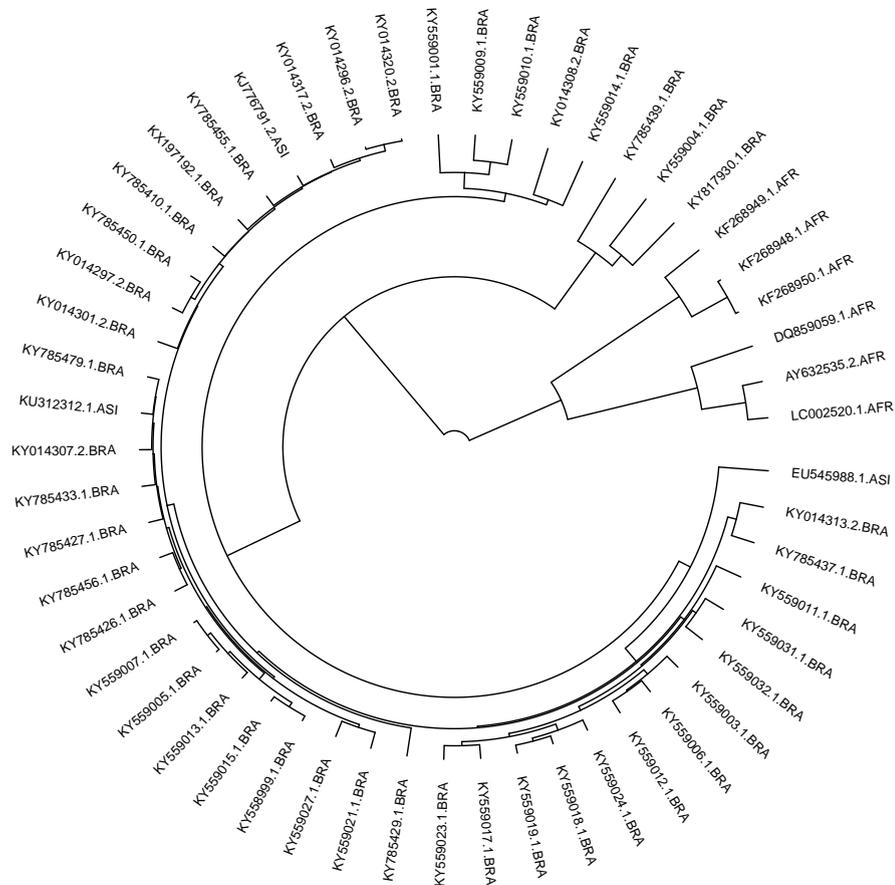


Figure 1: Dendrogram built from $dmax$ values, see definition 2.1-ii, for the sequences of Brazil (table 2), African and Asian lineages (table 1). Brazil (BRA), Asia (ASI), Africa (AFR).

(Dominican Republic), KY785439.1 (Brazil), KY785444.1 (Honduras). Once again, the Brazilian sequence KY785439.1 is shown to be particularly distant from the Brazilian ones and, also, from the American ones. In relation to the Asian sequences, we see that they are confused with the American ones, mainly KU312312.1 (Suriname) and KJ776791.2 (French Polynesia).

We apply the notion V - definition 2.2 - to 3 scenarios (a) sequences from America, (b) sequences from Asia, and (c) sequences from Africa, separately, in order to identify that sequence with the lowest V in each case. The assumption that governs this computation is the following, each group (a), (b), or (c) shows a certain homogeneity in its composition, so we proceed to select that sequence by a group that can be considered a good representative of the group. Given the definition of V , the most representative sequence of the group is the one with the lowest V . Table 4 shows the classifications in cases (b) and (c), and for (a) the indicated sequence is KY014318.3 (Dominican Republic), see also [7]. Thus, among the Asian sequences, the most representative is KU312312.1 (Suriname), and among the African ones, the most representative is KF268948.1 (Central African Republic). Table 5 exposes the $dmax$ values between the most representative sequences.

With these elements given by the procedure of choosing the most representative sequence by a group and the value obtained by $dmax$ between such sequences, we can conclude that there is greater proximity between the America's set and the Asian's set, see table 5.

Now we look at the constitution of each of those 12 parts. We see in table 7, their compositions. Let's think about the configurations, that is to say in the structures xyz_i , where i registers to which genomic sequence the configuration belongs. $i = 1$ refers to sequence KF268948.1 (Central African Republic), $i = 2$ refers to sequence KU312312.1 (Suriname), and $i = 3$ refers to sequence

Sequence	Median of Dmax	Sequence	Median of Dmax
KY559005.1.BRA	0.02006	KY559003.1.BRA	0.03332
KY559027.1.BRA	0.02058	KX197192.1.BRA	0.03631
KY559007.1.BRA	0.02107	KY559006.1.BRA	0.03671
KY558999.1.BRA	0.02151	KY559019.1.BRA	0.03693
KY559015.1.BRA	0.02224	KY559018.1.BRA	0.03694
KU312312.1.ASI	0.02237	KY785429.1.BRA	0.03754
KY785410.1.BRA	0.02317	KY014313.2.BRA	0.04061
KY014307.2.BRA	0.02347	KY559031.1.BRA	0.04452
KY559013.1.BRA	0.02411	KY559032.1.BRA	0.04885
KY785479.1.BRA	0.02588	KY559011.1.BRA	0.05050
KY559012.1.BRA	0.02666	KY785437.1.BRA	0.05109
KY785426.1.BRA	0.02702	KY559010.1.BRA	0.05490
KY785450.1.BRA	0.02759	EU545988.1.ASI	0.05673
KY785433.1.BRA	0.02795	KY559009.1.BRA	0.07528
KY014296.2.BRA	0.02796	KY014308.2.BRA	0.07719
KY014320.2.BRA	0.02796	KY559001.1.BRA	0.08371
KY785455.1.BRA	0.02824	KY559014.1.BRA	0.10132
KY785427.1.BRA	0.02856	KY817930.1.BRA	0.14787
KY559021.1.BRA	0.02865	KY559004.1.BRA	0.15016
KY014301.2.BRA	0.02866	KY785439.1.BRA	0.19626
KY559023.1.BRA	0.02994	KF268948.1.AFR	0.20933
KY559024.1.BRA	0.03034	KF268950.1.AFR	0.20936
KY014317.2.BRA	0.03069	KF268949.1.AFR	0.23211
KY559017.1.BRA	0.03152	LC002520.1.AFR	0.36991
KY785456.1.BRA	0.03184	AY632535.2.AFR	0.38538
KJ776791.2.ASI	0.03204	DQ859059.1.AFR	0.41088
KY014297.2.BRA	0.03206		

Table 3: V values, see definition 2.2, for the sequences of Brazil (table 2), African and Asian lineages (table 1). Brazil (BRA), Asia (ASI), Africa (AFR).

Asian sequence	V	African sequence	V
KU312312.1	0.02777	KF268948.1	0.14306
KJ776791.2	0.03778	AY632535.2	0.14309
EU545988.1	0.04613	KF268950.1	0.14309
		KF268949.1	0.14450
		DQ859059.1	0.15173
		LC002520.1	0.16114

Table 4: V values of each set of Asian sequences (left), African sequences (right). In bold letter the most representative sequence by set.

	KU312312.1 (Suriname)	KF268948.1 (Central African Republic)
KY014318.3 (Dominican Republic)	0.01364	0.17449
KU312312.1 (Suriname)	0	0.19697

Table 5: $dmax$ values between the most representative sequences of America, Africa and Asia. In bold letter the least value.

KY014318.3 (Dominican Republic). If the law is, in fact, the same for the 3 sequences ($i = 1, 2, 3$) we would expect that all the configurations xyz and independent of i are in the same part, which in fact we verify from table 7 that does not occur. In table 8, we show the states that, having the

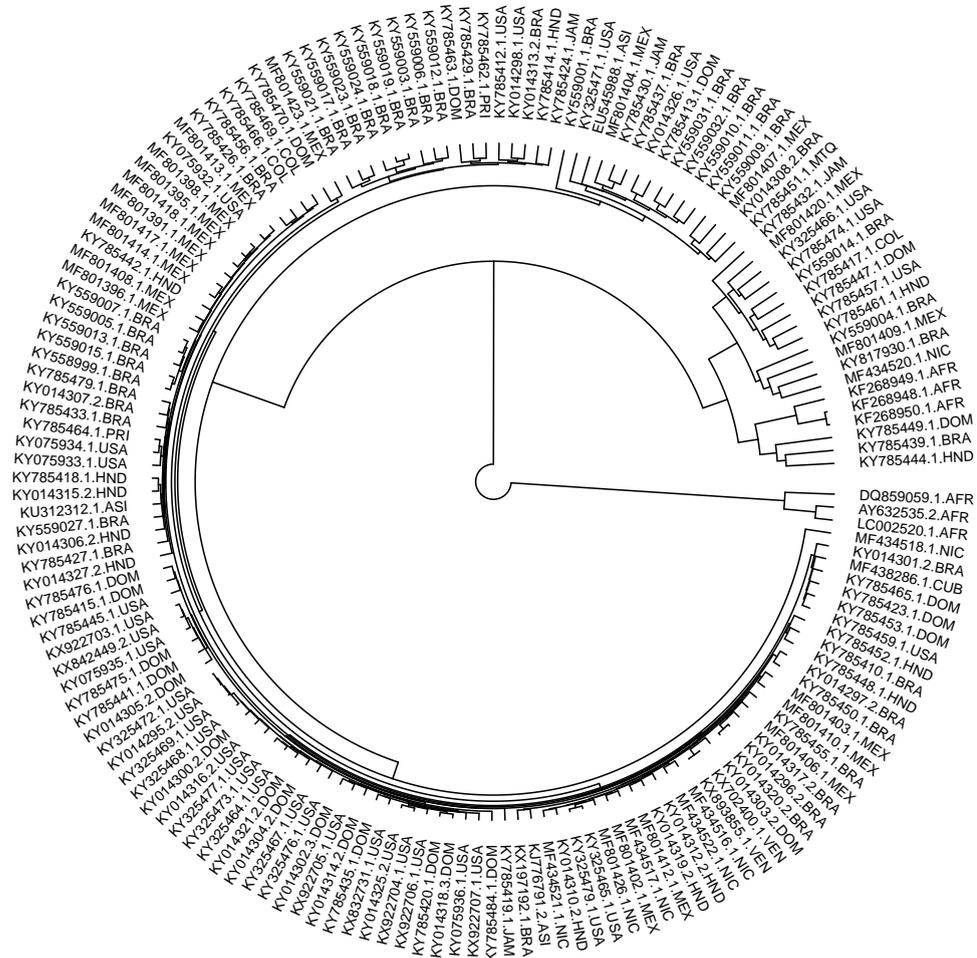


Figure 2: Dendrogram built from d_{max} values, see equation (2.1)-ii, for the sequences of America (table 2), African and Asian lineages (table 1).

i of L_i	a	c	g	t
1	0.29821	0.17092	0.35854	0.17233
2	0.39141	0.23853	0.14075	0.22931
3	0.37231	0.19614	0.25889	0.17266
4	0.22066	0.19150	0.42656	0.16128
5	0.28877	0.21828	0.22811	0.26484
6	0.19964	0.27293	0.25944	0.26799
7	0.25857	0.26475	0.29012	0.18656
8	0.12574	0.20433	0.41320	0.25672
9	0.30381	0.27060	0.11439	0.31119
10	0.17818	0.25769	0.36680	0.19733
11	0.42098	0.20546	0.07040	0.30316
12	0.13222	0.22882	0.46249	0.17647

Table 6: Transition Probabilities - equation (1) $\hat{P}_{L_i}(\cdot), \cdot \in A = \{a, c, g, t\}$. In bold letter the highest values.

same configuration, xyz , have been allocated in different parts. The states that are not mentioned in table 8 have been allocated in the same part regardless of the sequence, that is, for example,

i of L_i	Elements
1	$aaa_1, aaa_2, aaa_3, gaa_2, gaa_3, aga_1, gca_1, gca_2, gca_3, gga_1, gga_2, gga_3, tag_2, tag_3, aga_2, aga_3, cga_1, ttg_2, ttg_3, gaa_1, ttg_1$
2	$aac_1, acc_1, acc_2, gtc_1, acc_3, agg_1, gtc_2, gtc_3, aac_2, aac_3, gcc_2, gcc_1, gcc_3, ttc_3, ttc_2, cac_2, cac_3, gac_2, gac_3, cac_1, ccc_1, ccc_3, ccc_2, cgc_1, ctc_1$
3	$aag_1, tgg_1, aag_2, ggg_3, ggg_2, aag_3, cag_1, cgg_1, cgg_2, cgg_3, agg_2, agg_3, tgg_2, tgg_3, ggg_1, gag_1, gag_2, gag_3$
4	$aat_1, aat_2, aat_3, atg_1, ctg_2, ctg_3, atg_2, atg_3, tat_2, tat_3, ctg_1, gcg_1, gtg_1, gtg_2, gtg_3$
5	$aca_1, tag_1, aca_2, aca_3, tga_1, tga_3, tga_2, cgc_2, cgc_3, ttc_1, tca_1, tca_2, tca_3$
6	$acg_1, gta_2, gta_3, taa_1, ccg_2, ccg_3, cta_1, tct_1, tta_2, tta_3, cca_1, cta_2, tcg_1, tct_2, tct_3, gta_1, tta_1, cca_2, cca_3, taa_2, tcg_2, tcg_3, taa_3$
7	$acg_2, acg_3, ata_2, ata_3, caa_1, caa_2, caa_3, cag_2, cag_3$
8	$act_1, ggt_2, ggt_3, act_2, act_3, ggt_1, agt_1, ttt_2, ttt_3, cgt_1, agt_2, agt_3, cct_1, cgt_2, att_2, cgt_3, att_3, ctt_2, ctt_3, att_1, cct_2, cct_3$
9	$agc_1, agc_2, agc_3, tcc_1, ggc_2, ggc_3, tac_2, tac_3, tcc_2, tcc_3, ctc_2, ctc_3, ggc_1, tgc_1, tgc_2, tgc_3$
10	$ata_1, ccg_1, cga_2, cga_3, gcg_3, gcg_2, gtt_2, gtt_3, cat_1, ctt_1, cat_2, cat_3, cta_3, gtt_1$
11	$atc_1, gac_1, atc_2, atc_3, tac_1$
12	$gat_1, tgt_1, tat_1, gct_1, gct_2, gct_3, ttt_1, gat_2, gat_3, tgt_2, tgt_3$

Table 7: The subscript i indicates the provenance of sequence i , where $i = 1$ refers to sequence KF268948.1 (Central African Republic), $i = 2$ refers to sequence KU312312.1 (Suriname), and $i = 3$ refers to sequence KY014318.3 (Dominican Republic).

State	Sequences 2 & 3	Sequence 1	State	Sequences 2 & 3	Sequence 1
agg	L_3	L_2	ctt	L_8	L_{10}
ata	L_7	L_{10}	gac	L_2	L_{11}
acg	L_7	L_6	gcg	L_{10}	L_4
cag	L_7	L_3	tac	L_9	L_{11}
ccg	L_6	L_{10}	tag	L_1	L_5
cga	L_{10}	L_1	tat	L_4	L_{12}
cgc	L_5	L_2	ttc	L_2	L_5
ctc	L_9	L_2	ttt	L_8	L_{12}
	State	Sequences 1 & 2	Sequence 3		
	cta	L_6	L_{10}		

Table 8: To the right of each state, we report the part where it was included (see table 7), according to the sub-index related to the sequence of origin. KF268948.1 (Central African Republic) is the sequence 1, KU312312.1 (Suriname) is the sequence 2 and KY014318.3 (Dominican Republic) is the sequence 3.

the state aac for the 3 sequences is found in part L_2 . We can clearly see how the xyz states from sequences 2 (KU312312.1 (Suriname)) and 3 (KY014318.3 (Dominican Republic)) stay together, except in the case of cta . This fact strengthens the evidence given by table 5, reporting the most pronounced proximity between sequences from America and Asia.

5 Concluding Remarks

In this paper we use three powerful tools, originating from stochastic processes to verify an epidemiological conjecture, we give a magnitude to the validity of such conjecture, and we show the reasons that allow it to be verified. We use a metric between processes (see [4]), an indicator of representativeness between samples of processes (see [5]) and a model of communality (see [6]) to identify the greatest proximity between the genetic structure of sequences originating in America and those from Asia when comparing them with the proximity between the sequences from America and those from Africa. The ordering between the sequences of Brazil, Africa, and Asia, see figure 1 and table 3, already shows that the last ones positioned are the African ones, exposing the genetic

diversity announced by the literature, see [1], [2], [3]. These pieces of evidence are confirmed in the general comparison, which positions the African sequences as the most distant (see figure 2). With the help of the metric d_s , after applying the index V , (definitions 2.1, 2.2), we can concretely quantify this distance (see table 5), and confirm the conjecture. Furthermore, through the model (definition 2.3) we give meaning to it, since we find the states where the sequences of America and Asia are shown to be equivalent and we see that this occurs in most of the states of the state space, explaining the proximity identified (see tables 7, 8 and 6).

References

- [1] Enfissi, A., Codrington, J., Roosblad, J., Kazanji, M., Rousset, D. (2016). Zika virus genome from the Americas. *The Lancet*, 387(10015), 227-228.
- [2] Lanciotti, R. S., Lambert, A. J., Holodniy, M., Saavedra, S., Signor, L. D. C. C. (2016). Phylogeny of Zika virus in western hemisphere, 2015. *Emerging infectious diseases*, 22(5), 933.
- [3] Zanoluca, C., Melo, V. C. A. D., Mosimann, A. L. P., Santos, G. I. V. D., Santos, C. N. D. D., Luz, K. (2015). First report of autochthonous transmission of Zika virus in Brazil. *Memórias do Instituto Oswaldo Cruz*, 110(4), 569-572.
- [4] García, Jesús E., Gholizadeh R. , González-López, V.A. (2018). A BIC - based consistent metric between Markovian processes. *Applied Stochastic Models in Business and Industry*, 34(6), 868-878.
- [5] Fernández M., García Jesús E., Gholizadeh R., González-López V.A. Sample Selection Procedure in Daily Trading Volume Processes. *Mathematical Methods in the Applied Sciences* DOI: 10.1002/mma.5705 .
- [6] Cordeiro, M.T.A., García, Jesús E., González-López V.A. & Londoño, S.L.M. Partition Markov Model for Multiple Processes. *Mathematical Methods in the Applied Sciences* DOI: 10.1002/mma.6079 .
- [7] García, Jesús E., González-López V.A., Londoño, S.L.M. and Cordeiro, M.T.A., Similarity between Strains of Zika from Tropical and Subtropical Regions, Chapter in *Data Analysis and Applications* Iste Science Publishing ltd. (forthcoming).
- [8] Schwarz G (1978). Estimating the dimension of a model. *The annals of statistics* **6** (2), 461-464.

Tail estimation via linear combinations of generalized means

M. Ivette Gomes¹ and Fernanda Figueiredo²

¹ FCUL, DEIO and CEAUL, Universidade de Lisboa, Portugal
(E-mail: ivette.gomes@fc.ul.pt)

² Faculdade de Economia da Universidade do Porto and CEAUL, Portugal
(E-mail: otilia@fep.up.pt)

Abstract. Heavy-tailed data, from an underlying model with a Pareto-type right tail, are quite common in financial and telecommunication traffic time series, among many other areas of application. When analyzing such a type of data one never knows how much the underlying model differs from a strict Pareto model. And this is the unique situation where the Hill estimator of a positive *extreme value index* (EVI) is “perfect”. Together with the use of generalized means in the estimation of parameters of extreme events, like the EVI, bias-corrected and asymptotically best linear unbiased EVI-estimators are proposed and discussed. The finite-sample behaviour is assessed through Monte-Carlo simulation.

Keywords: Best linear unbiased estimation, Pareto-type right-tails, Statistics of extremes, Semi-parametric tail index estimation.

1 Introduction and motivation for the need of statistical extreme value theory

Extreme value theory (EVT) helps the control of potential disastrous and catastrophic events, high relevant to our society and with a high social impact. Domains of application of EVT include the fields of biostatistics, environment, finance, hydrology, insurance, meteorology and structural engineering, among many others.

Statistics of univariate extremes, as well as multivariate and spatial extremes, have recently faced a big development. In the eighties there has been a shift from the area of parametric statistics of extremes, based on probabilistic asymptotic results in EVT, towards semi-parametric or even non-parametric approaches. But parametric modeling is becoming again quite popular.

To motivate the interest for this area, we merely refer the North Sea Historical Floods, February 1, 1953 (for several other examples, see, among others, the books by Beirlant *et al.*[4], by Castillo *et al.*[14], and by Gomes *et al.*[28]). Indeed, at the early morning of the aforementioned day, the level of the water exceeded 5.6 metres above the sea level, several maritime defenses were destroyed, provoking the flooding of areas in Holland, England, Belgium, Denmark and France, with the death of around 2500 people. As a consequence

⁶*th* SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain



of the North Sea floods, the Dutch government formed the so-called ‘*Delta Committee*’. It was decided that dams should be built in such a way that the probability of a flood in a certain year should be 1 out of 10 000. *But data have been observed during a much shorter period!* . . . It is thus necessary to extrapolate beyond observed data! . . . either univariate or multivariate, dependent (or not) of time and/or space. Indeed, EVT is needed because the world is not always ‘normal’, either multivariate or univariate, a simple symmetric model with quadratic exponential right and left tails.

Risky events are in the tails of the underlying distribution, with most of the data in the center, being observations in the tail quite scarce, even when we are dealing with ‘big data’. Estimates much beyond the observed maximum/minimum are often required, and we thus need to consider models for the tails, usually based on asymptotic results. Indeed, the answer to the question whether there is a hidden pattern underlying this type of extreme events is positive, and we shall briefly sketch it in the following.

In an univariate set-up, let us assume that, after a possible adequate transformation, the available transformed sample, $\mathbf{X}_n = (X_1, \dots, X_n)$, can be regarded as a sample of size n of independent, identically distributed (IID) or more generally stationary weakly dependent random variables (RVs) from a cumulative distribution function (CDF) F . Let us use the notation $X_{1:n} \leq \dots \leq X_{n:n}$ for the associated ascending *order statistics* (OSs). Further assume that there exist sequences of real constants $\{a_n > 0\}$ and $\{b_n \in \mathbb{R}\}$ such that $(X_{n:n} - b_n)/a_n$ converges in distribution to a non-degenerate RV. Then (Gnedenko[18]), the limiting CDF is necessarily of the type of the *general extreme value* (GEV) CDF, given by

$$\text{GEV}_\xi(x) \equiv G_\xi(x) = \begin{cases} e^{-(1+\xi x)^{-1/\xi}}, & 1 + \xi x > 0, \text{ if } \xi \neq 0, \\ e^{-e^{-x}}, & x \in \mathbb{R}, \text{ if } \xi = 0. \end{cases} \quad (1)$$

The CDF F is then said to belong to the *max-domain of attraction* of GEV_ξ , and the notation $F \in \mathcal{D}_\mathcal{M}(\text{GEV}_\xi)$ is used. The parameter ξ is the *extreme value index* (EVI), the primary parameter of extreme events. The GEV_ξ model, in (1), is perhaps the most relevant univariate asymptotic model in statistical EVT. For other relevant asymptotic models, and different approaches to statistics of univariate extremes, see the recent overviews by Beirlant *et al.*[5], Scarrott and MacDonald[43] and Gomes and Guillou[20].

The EVI in (1) measures the heaviness of the right-tail function $\bar{F}(x) := 1 - F(x)$, and the heavier the right-tail, the larger ξ is (see Figure 1, for an illustration of the right-tail heaviness). We shall here consider *heavy-tailed* models, i.e. *Pareto-type* underlying CDFs, with a positive EVI, often called tail index, working thus in

$$\mathcal{D}_\mathcal{M}^+ := \mathcal{D}_\mathcal{M}(\text{GEV}_{\xi>0}),$$

with $\text{GEV}_\xi \equiv G_\xi$ defined in (1). Note that, in an univariate framework, and with \mathcal{R}_a denoting the class of regularly varying functions with an index of regular variation a (see Bingham *et al.*[7], for details on regular variation), and with the notation $U(t) := F^{\leftarrow}(1 - 1/t)$, $t \geq 1$, $F^{\leftarrow}(y) := \inf\{x : F(x) \geq y\}$ the

generalized inverse function of the underlying model F ,

$$F \in \mathcal{D}_{\mathcal{M}}^+ \iff \bar{F} = 1 - F \in \mathcal{R}_{-1/\xi} \iff U \in \mathcal{R}_{\xi}. \quad (2)$$

As an example of a CDF in $\mathcal{D}_{\mathcal{M}}^+$, we merely mention the Fréchet CDF, $F(x) = \exp(-x^{-\alpha})$, $x \geq 0$, $\alpha > 0$ ($\xi = 1/\alpha$), among many others.

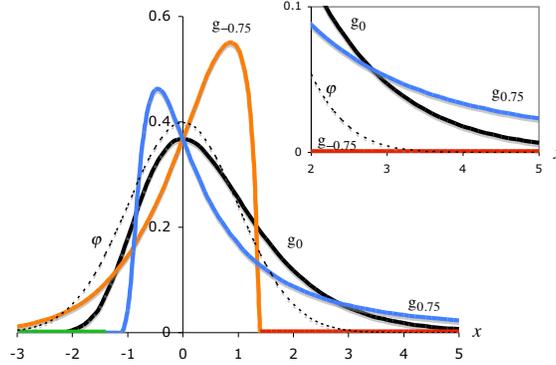


Fig. 1. Probability density function (PDF), $g_{\xi} = \text{GEV}'_{\xi} \equiv G'_{\xi}$, for $\xi = -0.75$, $\xi = 0$ and $\xi = 0.75$, together with the normal PDF, φ .

Under the first-order condition, in (2), the EVI-estimators are consistent for intermediate k , i.e. levels k such that

$$k = k_n \rightarrow \infty \quad \text{and} \quad k/n \rightarrow 0, \quad \text{as } n \rightarrow \infty. \quad (3)$$

To achieve asymptotic normality for the EVI-estimators, under the general first order semi-parametric framework in (2), we need often to know the rate of convergence in (2), i.e. we need to assume an additional second-order condition, such as

$$\lim_{t \rightarrow \infty} \frac{\ln U(tx) - \ln U(t) - \xi \ln x}{A(t)} = \begin{cases} \frac{x^{\rho} - 1}{\rho}, & \text{if } \rho < 0, \\ \ln x, & \text{if } \rho = 0, \end{cases} \quad (4)$$

for all $x > 0$, where A is a suitably chosen function of constant sign near infinity (positive or negative), and $\rho \leq 0$ is the ‘shape’ second-order parameter. The limit function in (4), whenever non-null, is necessarily of this given form, and $|A| \in RV_{\rho}$ (Geluk and de Haan[17]). We shall further assume throughout the paper that the shape second-order parameter is negative, i.e. $\rho < 0$.

Remark 1. For the strict Pareto model, i.e. if $F(x) = 1 - \alpha x^{-1/\xi}$, $x \geq \alpha^{\xi} > 0$, with $\xi > 0$, the numerator of the right hand-side of (4) is null, i.e. $\ln U(tx) - \ln U(t) - \xi \ln x \equiv 0$.

Remark 2. For Hall-Welsh class of Pareto-type models (Hall and Welsh[35]), with a right tail function

$$\bar{F}(x) = \left(\frac{x}{C}\right)^{-1/\xi} \left(1 + \frac{\beta}{\rho} \left(\frac{x}{C}\right)^{\rho/\xi} + o(x^{\rho/\xi})\right), \quad \text{as } x \rightarrow \infty, \quad (5)$$

$\alpha > 0$, $\beta \neq 0$, a ‘scale’ second-order parameter, $\rho < 0$, (4) holds and we may choose $A(t) = \xi\beta t^\rho$.

We shall address the estimation of the EVI, giving emphasis to the use of *generalized means* (GMs), *reduced-bias* (RB) EVI-estimators and *best linear unbiased estimators* (BLUEs). In Section 2, and for a heavy right-tail, i.e. $\xi > 0$, in (1), we discuss a few estimators of the EVI, the primary parameter in statistical EVT. Like most estimators in the general area of Statistics, the most common estimator of a positive ξ , the Hill (H) EVI-estimator is an average of adequate statistics, based on the k upper observations. Trying to improve the performance of the classical H EVI-estimators, recent classes of reliable EVI-estimators based on adequate GMs are put forward. Again, a high variance for small k and a high bias for large k can appear, and thus the need for bias-reduction and/or an adequate choice of the tuning parameters under play. In Section 3, RB EVI-estimators are discussed, and Section 4 is dedicated to the introduction of different possible classes of *best linear* (BL) EVI-estimators. In Section 5, using Monte Carlo simulation techniques, we exhibit the finite sample behaviour of some of the EVI-estimators under discussion. Finally, in Section 6, overall conclusions are drawn. One of the main points of the article is that, as even asymptotically equivalent estimators may exhibit very diversified finite sample properties, it is always sensible to work, in practice, with a few EVI-estimators, possibly dependent on tuning parameters, which make them more flexible. This enables us to choose a reliable estimate of ξ , the primary parameter of extreme events, and the basis for the estimation of other parameters of rare events.

2 Classes of Hill and generalized means’ EVI-Estimators

2.1 The Hill EVI-estimators

For heavy-tailed models, the classical EVI-estimators are the Hill (H) estimators (Hill[36]), which can be written as the average of the log-excesses or of the scaled log-spacings, i.e.

$$\begin{aligned} H(k) = H(k; \underline{X}_n) &:= \frac{1}{k} \sum_{i=1}^k V_{ik} \equiv \frac{1}{k} \sum_{i=1}^k V_i, \\ V_{ik} &:= \ln \frac{X_{n-i+1:n}}{X_{n-k:n}}, \quad V_i := i \ln \frac{X_{n-i+1:n}}{X_{n-i:n}}, \quad 1 \leq i \leq k < n. \end{aligned} \quad (6)$$

Remark 3. The Hill EVI-estimators in (6) are unbiased if the underlying model is a strict Pareto model. Aban and Meerschaert[1] have shown that for this strict Pareto model, the Hill estimator in (6) is indeed the BLUE, and also the uniformly minimum variance unbiased estimator of the tail index ξ , based on the $k + 1$ upper OSs of the sample.

2.2 Power mean-of-order- p (H_p) EVI-estimators

We can write

$$H(k) = \sum_{i=1}^k \ln \left(\frac{X_{n-i+1:n}}{X_{n-k:n}} \right)^{1/k} = \ln \left(\prod_{i=1}^k \frac{X_{n-i+1:n}}{X_{n-k:n}} \right)^{1/k} =: \ln \left(\prod_{i=1}^k U_{ik} \right)^{1/k} \quad (7)$$

and also,

$$H(k) = \ln \left(\prod_{i=1}^k \left(\frac{X_{n-i+1:n}}{X_{n-i:n}} \right)^i \right)^{1/k} =: \ln \left(\prod_{i=1}^k U_i \right)^{1/k}. \quad (8)$$

The H EVI-estimators are thus the logarithm of the *geometric mean* (or *mean-of-order-0*) of U_{ik} , $1 \leq i \leq k$, in (7) or of U_i , $1 \leq i \leq k$, in (8). Brillhante *et al.*[8], and almost simultaneously Paulauskas and Vaičiulis[38] and Beran *et al.*[6] considered as basic statistics, the *power mean-of-order- p* of U_{ik} , $1 \leq i \leq k$, in (7), for $p \geq 0$. More generally, Gomes and Caeiro[19], and also Caeiro *et al.*[13], considered those same statistics for any $p \in \mathbb{R}$, i.e.

$$M_p(k) = \begin{cases} \left(\frac{1}{k} \sum_{i=1}^k U_{ik}^p \right)^{1/p}, & \text{if } p \neq 0, \\ \left(\prod_{i=1}^k U_{ik} \right)^{1/k}, & \text{if } p = 0, \end{cases} \quad (9)$$

and the associated class of EVI-estimators:

$$H_p(k) = H_p(k; \underline{X}_n) := \begin{cases} \left(1 - M_p^{-p}(k) \right) / p, & \text{if } p < 1/\xi, p \neq 0, \\ \ln M_0(k) = H(k), & \text{if } p = 0. \end{cases} \quad (10)$$

If in (9), we replace U_{ik} by U_i , using the notation \tilde{M}_p for those statistics, we then trivially get the following associated class of EVI-estimators:

$$\tilde{H}_p(k) = \tilde{H}_p(k; \underline{X}_n) := \begin{cases} \left(1 - \tilde{M}_p^{-p}(k) \right) / p, & \text{if } p < 1/\xi, p \neq 0, \\ \ln \tilde{M}_0(k) = H(k), & \text{if } p = 0, \end{cases} \quad (11)$$

surely asymptotically equivalent to $H_p(k)$, in (10). As shown in Brillhante *et al.*[9], and in the sense of minimal asymptotic *mean square error* (MSE), there is, for $H_p(k)$ in (10), an optimal positive p , an explicit function of ρ . The same happens for $\tilde{H}_p(k)$, in (11).

2.3 Lehmer's mean-of-order- p (L_p) EVI-estimators

Beyond the average, the p -moments of log-excesses, i.e.

$$M_{k,n}^{(p)} := \frac{1}{k} \sum_{i=1}^k V_{ik}^p, \quad p \geq 1 \quad [M_{k,n}^{(1)} \equiv H(k)], \quad (12)$$

introduced in Dekkers *et al.*[15], have also played a relevant role in the EVI-estimation, and can more generally be parameterized in $p \in \mathbb{R} \setminus \{0\}$. And a simple generalization of the average is Lehmer's mean-of-order- p : Given a set of positive numbers $\underline{a} = (a_1, \dots, a_k)$, such a mean generalizes both the arithmetic mean ($p = 1$) and the harmonic mean ($p = 0$), being defined as

$$L_p(\underline{a}) := \sum_{i=1}^k a_i^p / \sum_{i=1}^k a_i^{p-1}, \quad p \in \mathbb{R}.$$

The H EVI-estimators can thus be considered as the Lehmer mean-of-order-1 of the k log-excesses $\underline{V} := (V_{ik}, 1 \leq i \leq k)$, defined in (6). Following Penalva *et al.*[40] (see also, Gomes *et al.*[34]; Penalva[39]), and with E denoting a standard exponential RV, $V_{ik} = \xi E_{k-i+1:k} + o_p(1)$. Since $\mathbb{E}(E^p) = \Gamma(p+1)$, for all $p > -1$, with $\Gamma(\cdot)$ denoting the Gamma function, the law of large numbers enables us to say that

$$\frac{1}{k} \sum_{i=1}^k V_{ik}^p \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \Gamma(p+1)\xi^p.$$

Hence the reason for the class of L_p EVI-estimators,

$$L_p(k) = L_p(k; \underline{X}_n) := \frac{L_p(\underline{V})}{p} = \frac{1}{p} \frac{\sum_{i=1}^k V_{ik}^p}{\sum_{i=1}^k V_{ik}^{p-1}} = \frac{M_{k,n}^{(p)}}{pM_{k,n}^{(p-1)}} \quad \left[L_1(k) \equiv H(k) \right], \quad (13)$$

consistent for all $\xi > 0$ and real $p > 0$, provided (3) holds, and with $M_{k,n}^{(p)}$ defined in (12). The p -moments of the log-excesses can obviously be replaced by the p moments of the scaled log-spacings, and the consideration of an \tilde{L}_p -class of EVI-estimators, in the lines of what we have suggested for the building of the \tilde{H}_p -class in (11).

3 Classes of reduced-bias EVI-estimators

Despite of the interesting behaviour of GMs' EVI-estimators, either in (10) or in (13), for an adequate choice of (p, k) , these estimators, just like the H EVI-estimators in (6), usually exhibit a high variance for small values of k and a high bias for large k . Consequently, the adequate accommodation of such asymptotic bias has recently been extensively addressed in the literature on statistical EVT. We mention the pioneering papers by Peng[42], Beirlant *et al.*[3], Feuerverger and Hall[16] and Gomes *et al.*[22], among others, based either on log-excesses or on scaled log-spacings between subsequent extreme OSs from a Pareto-type distribution, both defined in (6). In these papers, authors are led to propose *second-order reduced-bias* (SORB) EVI-estimators, with asymptotic variances larger than or equal to $(\xi(1-\rho)/\rho)^2$, where $\rho(< 0)$ is the aforementioned 'shape' second-order parameter, in (5). More details on this type of EVI-estimators can be found in Gomes and Guillou[20].

For Pareto-type models, Caeiro *et al.* ([11], [12]) and Gomes *et al.* ([24], [25]), among others, have been able to reduce the bias without increasing the asymptotic variance, kept at ξ^2 , just as happens with the Hill EVI-estimators. Those estimators, called *minimum-variance reduced-bias* (MVRB) EVI-estimators, are all based on an adequate ‘external’ and a bit more than consistent estimation of the pair of second-order parameters, $(\beta, \rho) \in (\mathbb{R} \setminus \{0\}, \mathbb{R}^-)$ in $A(t) = \xi \beta t^\rho$, done through consistent estimators, denoted by $(\hat{\beta}, \hat{\rho})$, such that $\hat{\rho} - \rho = o_p(1/\ln n)$, and outperform the classical estimators for all k . Different algorithms for the estimation of (β, ρ) can be found in Gomes and Pestana [21], among others. Among the most common MVRB EVI-estimators, we just mention the simplest class in Caeiro *et al.* [11]. Such a class has the functional form

$$\bar{H}(k) \equiv \bar{H}_{\hat{\beta}, \hat{\rho}}(k) := H(k) \left(1 - \frac{\hat{\beta}}{1-\hat{\rho}} \left(\frac{n}{k} \right)^{\hat{\rho}} \right), \quad (14)$$

with $(\hat{\beta}, \hat{\rho})$ an adequate estimator of the vector (β, ρ) of second-order parameters, in (5), and can also be regarded as a linear combination of either the log-excesses or of the scaled log-spacings, in (6).

More generally than the class in (14), we can consider the class introduced in Gomes *et al.* [32], given by

$$\bar{H}_p(k) \equiv \bar{H}_p(k; \hat{\beta}, \hat{\rho}) := H_p(k) \left(1 - \frac{\hat{\beta}(1-pH_p(k))}{1-\hat{\rho}-pH_p(k)} \left(\frac{n}{k} \right)^{\hat{\rho}} \right) \quad [\bar{H}_0 \equiv \bar{H} \text{ in (14)}],$$

with $H_p(k)$ given in (10). We further mention the partially RB class of EVI-estimators in Gomes *et al.* [30], which also reveals some interesting properties. It is also obviously sensible to refer the class of RB EVI-estimators introduced in Gomes *et al.* [34], given by

$$\bar{L}_p(k) \equiv \bar{L}_p(k; \hat{\beta}, \hat{\rho}) := \left(1 - \frac{\hat{\beta}(n/k)^{\hat{\rho}}}{(1-\hat{\rho})^p} \right) L_p(k) \quad [\bar{L}_1 \equiv \bar{H} \text{ in (14)}],$$

with $L_p(k)$ given in (13).

4 Classes of BL EVI-estimators

In Statistics, we often put the question whether the combination of information can improve the performance of an estimator. We can then be led to think on BLUEs, which are unbiased linear combinations of an adequate set of statistics, with minimum variance among the class of such linear combinations. The basic theorem underlying this theory is due to Aitken [2]: *If \mathbf{X} is a vector of observations with mean values $\mathbb{E}\mathbf{X} = \mathbf{A}\theta$ depending linearly on the unknown vector of parameters θ , with a known coefficient matrix \mathbf{A} , and with a covariance matrix $\delta^2 \Sigma$, known up to a scale factor δ^2 , the least-squares estimator of θ is the vector θ^* that minimizes the quadratic form $(\mathbf{X} - \mathbf{A}\theta)' \Sigma^{-1} (\mathbf{X} - \mathbf{A}\theta)$. Such a vector is the vector of solutions of the “normal equations”, $\mathbf{A}' \Sigma^{-1} \mathbf{A} \theta^* = \mathbf{A}' \Sigma^{-1} \mathbf{X}$. This solution is given by $\theta^* = (\mathbf{A}' \Sigma^{-1} \mathbf{A})^{-1} \mathbf{A}' \Sigma^{-1} \mathbf{X}$, and $\text{Var}(\theta^*) = \delta^2 (\mathbf{A}' \Sigma^{-1} \mathbf{A})^{-1}$.*

Given a vector of m statistics directly related to the tail index ξ , let us say

$$\mathbf{T} \equiv \mathbf{T}_m \equiv (T_{ik}, \quad i = k - m + 1, \dots, k), \quad 1 \leq m \leq k,$$

where k is intermediate, i.e. (3) holds, let us assume that, asymptotically, the covariance matrix of \mathbf{T} is well approximated by $\xi^2 \Sigma$, i.e. it is known up to the scale factor ξ^2 , and that its mean value is asymptotically well approximated by $\xi \mathbf{s} + \varphi(n, k) \mathbf{b}$, with \mathbf{s} and \mathbf{b} known. Just as mentioned in Gomes *et al.*[23], it is thus sensible to investigate the following question: “*Is it possible to find a linear combination of this set of statistics, asymptotically unbiased and of minimum variance*”? Such a linear combination will be called an ‘*asymptotically BLUE*’ (ABLUE) based on \mathbf{T} , and will be denoted $\text{BL}_{\mathbf{T}}$. Just as discussed in the above mentioned paper, our goal is then to find a vector $\mathbf{a}' = (a_1, \dots, a_m)$ such that $\mathbf{a}' \Sigma \mathbf{a}$ is minimum, subject to the conditions $\mathbf{a}' \mathbf{s} = 1$ and $\mathbf{a}' \mathbf{b} = 0$. As already mention in Gomes *et al.*[23], among others, the solution of such a problem is easily obtained if we consider the function,

$$H(\mathbf{a}; \alpha, \beta) = \mathbf{a}' \Sigma \mathbf{a} - \alpha (\mathbf{a}' \mathbf{s} - 1) - \beta \mathbf{a}' \mathbf{b},$$

and solve the equations:

$$\begin{cases} 2 \Sigma \mathbf{a} - \alpha \mathbf{s} - \beta \mathbf{b} = 0 \\ \mathbf{a}' \mathbf{s} = 1 \\ \mathbf{a}' \mathbf{b} = 0. \end{cases} \quad (15)$$

From the first equation in (15) we get $2 \Sigma \mathbf{a} = \mathbf{P} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$, with $\mathbf{P} = [\mathbf{s} \quad \mathbf{b}]$. Hence,

$$\mathbf{a} = \frac{1}{2} \Sigma^{-1} \mathbf{P} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \quad (16)$$

$\mathbf{P}' \mathbf{a} = \frac{1}{2} (\mathbf{P}' \Sigma^{-1} \mathbf{P}) \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$, and $\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = 2 (\mathbf{P}' \Sigma^{-1} \mathbf{P})^{-1} \mathbf{P}' \mathbf{a}$, if there exists $(\mathbf{P}' \Sigma^{-1} \mathbf{P})^{-1}$. But from the last two equations in (15) we get $\mathbf{a}' \mathbf{P} = [1 \quad 0]$, i.e., $\mathbf{P}' \mathbf{a} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, and consequently,

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = 2 (\mathbf{P}' \Sigma^{-1} \mathbf{P})^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \quad (17)$$

If we incorporate (17) in (16), we get

$$\mathbf{a} = \Sigma^{-1} \mathbf{P} (\mathbf{P}' \Sigma^{-1} \mathbf{P})^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = -\frac{1}{\Delta} \mathbf{b}' \Sigma^{-1} (\mathbf{s} \mathbf{b}' - \mathbf{b} \mathbf{s}') \Sigma^{-1}, \quad (18)$$

where $\Delta = \|\mathbf{P}' \Sigma^{-1} \mathbf{P}\|$. Since we have denoted $\mathbf{T} \equiv \mathbf{T}_m$ the vector of the m statistics we use to construct the estimator, we get

$$\text{BL}_{\mathbf{T}}(k; m) := \mathbf{a}' \mathbf{T}, \quad \mathbf{a} \text{ given in (18)}. \quad (19)$$

With exact results we can further derive that

$$\text{Var}(\text{BL}_{\mathbf{T}}(k; m)) = \xi^2 \mathbf{b}' \Sigma^{-1} \mathbf{b} / \Delta.$$

Gomes *et al.*[23] have first considered ‘asymptotically best linear combinations’ of the H EVI-estimators in (6) and of log-excesses, $V_{ik}, 1 \leq i \leq k$, also in (6), being led to a computationally time-consuming EVI-estimator. They have then decided to consider similar computations, but based on the scaled log-spacings, $V_i, 1 \leq i \leq k$, defined also in (6). Such a derivation led them to much simpler linear combinations, with almost the same exact behaviour and obviously equivalent asymptotic properties. We now suggest a more general framework, considering for \mathbf{T} , in (19), any of the aforementioned EVI-estimators, in (10) and in (13), or the statistics U_{ik} and $U_i, 1 \leq i \leq k$, respectively given in (7) and (8). And we may even try putting the not yet answered question: “*Is it possible to find a linear combination of any of the aforementioned sets of statistics or other similar fulcral statistics, asymptotically unbiased and that suffers no increase in the asymptotic variance?*”

5 A short illustration with simulated data

Apart from the H_p EVI-estimators, in (10), we also consider the simplest class of second-order MVRB EVI-estimators in Caeiro *et al.*[11], given in (14). Just as an illustration of large-scale Monte-Carlo simulation studies undertaken in several articles on the topic (see Gomes *et al.*[32], among others), we picture Figure 2, where, for a GEV underlying model, with $\xi = 0.1$, the patterns of simulated expected values (E) and *root mean square errors* (RMSE) of the H EVI-estimators, the estimated optimal H_p EVI-estimator, denoted by \hat{H}^* , the associated RB EVI-estimator, denoted by \hat{H} , and the BL EVI-estimator based on scaled log-spacings are presented.

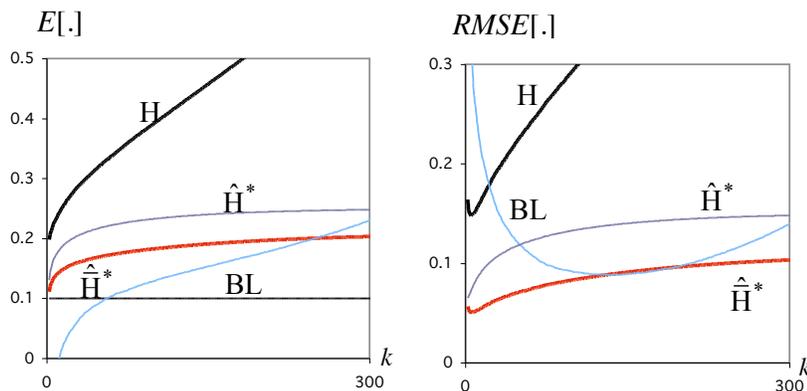


Fig. 2. IID set-up and $\text{GEV}_{0.1}$ underlying model

Similar patterns were obtained for other simulated models and for other BL EVI-estimators.

6 A few overall comments

For all k , there is a reduction in RMSE, as well as in bias, and \widehat{H}_p^* , the estimated optimal H_p beats the H EVI-estimators for all simulated parents. Even better results have been obtained for \widehat{L}^* , which is known to beat asymptotically \widehat{H}^* in the whole (ξ, ρ) -plane (see Penalva *et al.*[41]). As expected, \widehat{H}^* beats \widehat{H}^* for all k . The BL EVI-estimators based on the scaled log-spacings can beat \widehat{H}^* , regarding RMSE, but have never beaten \widehat{H}^* , for any of the simulated models, and other BL EVI-estimators are thus desired. The choice of the tuning parameters (k, p) can be done through heuristic sample-path stability algorithms, like the ones in Gomes *et al.*[29], and Neves *et al.*[37]. But alternatively, it is also sensible to use a bootstrap algorithm of the type of the ones in Gomes *et al.* ([26], [27]), and Brilhante *et al.*[8], among others. See also Gomes *et al.*[31] for a reliable EVI-estimation based on the bootstrap methodology. For further algorithmic details on an adaptive EVI-estimation, see also Caeiro and Gomes[10], and Gomes *et al.*[33], where R-scripts are provided.

Acknowledgements. Research partially supported by FCT—Fundação para a Ciência e a Tecnologia, project UIDB/00006/2020 (CEAUL).

References

1. I. Aban and M. Meerschaert. Generalized least squares estimators for the thickness of heavy tails. *J. Statistical Planning and Inference*, 119, 2, 341–352, 2004.
2. A.C. Aitken. On least squares and linear combinations of observations. *Proc. Roy. Soc. Edin.*, 55, 42-48, 1935.
3. J. Beirlant, G. Dierckx, Y. Goegebeur and G. Matthys. Tail index estimation and an exponential regression model. *Extremes*, 2, 2, 177-200, 1999.
4. J. Beirlant, Y. Goegebeur, J. Segers and J. Teugels. *Statistics of Extremes. Theory and Applications*. Wiley, 2004.
5. J. Beirlant, F. Caeiro and M.I. Gomes. An overview and open research topics in statistics of univariate extremes. *Revstat—Statist. J.*, 10, 1, 1–31, 2012.
6. J. Beran, D. Schell and M. Stehlík. The harmonic moment tail index estimator: asymptotic distribution and robustness. *Ann. Inst. Statist. Math.*, 66, 193–220, 2014.
7. N. Bingham, C.M. Goldie and J. Teugels. *Regular Variation*. Cambridge Univ. Press, Cambridge, 1987.
8. M.F. Brilhante, M.I. Gomes and D. Pestana. A simple generalization of the Hill estimator. *Computat. Statistics and Data Analysis*, 57, 1, 518–535, 2013.
9. M.F. Brilhante, M.I. Gomes and D. Pestana. The mean-of-order p extreme value index estimator revisited. In Pacheco, A. *et al.* (eds.), *New Advances in Statistical Modeling and Application*. Springer-Verlag, Berlin and Heidelberg, 163–175, 2014.

10. F. Caeiro and M.I. Gomes. Threshold Selection in Extreme Value Analysis. Chapter in: Dipak Dey and Jun Yan (eds.), *Extreme Value Modeling and Risk Analysis: Methods and Applications*, Chapman-Hall/CRC, 69–87, 2015.
11. F. Caeiro, M.I. Gomes and D. Pestana. Direct reduction of bias of the classical Hill estimator. *Revstat—Statist. J.*, 3, 2, 111–136, 2005.
12. F. Caeiro, M.I. Gomes and L. Henriques-Rodrigues. Reduced-bias tail index estimators under a third order framework. *Comm. Statist. Theory Methods*, 38, 7, 1019–1040, 2009.
13. F. Caeiro, M.I. Gomes, J. Beirlant and T. de Wet. Mean-of-order p reduced-bias extreme value index estimation under a third-order framework. *Extremes*, 19, 4, 561–589, 2016.
14. E. Castillo, A. Hadi, N. Balakrishnan and J.M. Sarabia. *Extreme Value and Related Models with Applications in Engineering and Science*. Wiley, Hoboken, New Jersey, 2005.
15. A.L.M. Dekkers, J.H.J. Einmahl and L. de Haan. A moment estimator for the index of an extreme-value distribution. *Ann. Statist.*, 17, 1833–1855, 1989.
16. A. Feuerverger and P. Hall. Estimating a tail exponent by modelling departure from a Pareto distribution. *Ann. Statist.*, 27, 760–781, 1999.
17. J. Geluk and L. de Haan. *Regular Variation, Extensions and Tauberian Theorems*. CWI Tract 40, Center for Mathematics and Computer Science, Amsterdam, Netherlands, 1987.
18. B.V. Gnedenko. Sur la distribution limite du terme maximum d’une série aléatoire. *Ann. Math.*, 44, 423–453, 1943.
19. M.I. Gomes and F. Caeiro. Efficiency of partially reduced-bias mean-of-order- p versus minimum-variance reduced-bias extreme value index estimation. In M. Gilli *et al.* (eds.), *Proceedings of COMPSTAT 2014*, ISI/IASC 289–298, 2014.
20. M.I. Gomes and A. Guillou. Extreme value theory and statistics of univariate extremes: a review. *International Statistical Review*, 83, 2, 263–292, 2015.
21. M.I. Gomes and D. Pestana. A sturdy reduced-bias extreme quantile (VaR) estimator. *J. American Statistical Association*, 102, 477, 280–292, 2007.
22. M.I. Gomes, M.J. Martins and M.M. Neves. Alternatives to a semi-parametric estimator of parameters of rare events—the Jackknife methodology. *Extremes*, 3, 3, 207–229, 2000.
23. M.I. Gomes, F. Figueiredo and S. Mendonça. Asymptotically best linear unbiased tail estimators under a second order regular variation. *J. Statist. Planning and Inference*, 134, 2, 409–433, 2005.
24. M.I. Gomes, M.J. Martins and M.M. Neves. Improving second order reduced bias extreme value index estimation. *Revstat—Statist. J.*, 5, 2, 177–207, 2007.
25. M.I. Gomes, L. de Haan and L. Henriques-Rodrigues. Tail index estimation for heavy-tailed models: accommodation of bias in weighted log-excesses. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 70, 1, 31–52, 2008.
26. M.I. Gomes, S. Mendonça and D. Pestana. Adaptive reduced-bias tail index and VaR estimation via the bootstrap methodology. *Comm. in Statistics—Theory and Methods*, 40, 16, 2946–2968, 2011.
27. M.I. Gomes, F. Figueiredo and M.M. Neves. Adaptive estimation of heavy right tails: the bootstrap methodology in action. *Extremes*, 15, 463–489, 2012.
28. M.I. Gomes, M.I. Fraga Alves and C. Neves. *Análise de Valores Extremos: uma Introdução*. Edições SPE & INE (in Portuguese), 2013.
29. M.I. Gomes, L. Henriques-Rodrigues, M.I. Fraga Alves and B.G. Manjunath. Adaptive PORT-MVRB estimation: an empirical comparison of two heuristic algorithms. *J. Statist. Comput. and Simul.*, 83, 6, 1129–1144, 2013.

30. M.I. Gomes, M.F. Brillhante, F. Caeiro and D. Pestana. A new partially reduced-bias mean-of-order-p class of extreme value index estimators. *Computat. Statistics and Data Analysis*, 82, 223–237, 2015.
31. M.I. Gomes, F. Figueiredo, M.J. Martins and M.M. Neves. Resampling methodologies and reliable tail estimation. *South African Statistical J.*, 49, 1–20, 2015.
32. M.I. Gomes, M.F. Brillhante and D. Pestana. New reduced-bias estimators of a positive extreme value index. *Comm. in Statistics—Simulation and Computation*, 45, 1–30, 2016.
33. M.I. Gomes, F. Caeiro, L. Henriques-Rodrigues and B.G. Manjunath. Bootstrap methods in statistics of extremes. In Longin, F. (ed.), *Handbook of Extreme Value Theory and Its Applications to Finance and Insurance*. Handbook Series in Financial Engineering and Econometrics (Ruey Tsay Adv.Ed.). John Wiley and Sons, Hoboken, New Jersey, Chapter 6, 117–138, 2016.
34. M.I. Gomes, H. Penalva, F. Caeiro and M.M. Neves. Non-reduced versus reduced-bias estimators of the extreme value index—efficiency and robustness. In A. Colubi *et al.* (eds.), *COMPSTAT 2016—22nd International Conference on Computational Statistics*, ISI/IASC, 279–290, 2016.
35. P. Hall and A.W. Welsh. Adaptive estimates of parameters of regular variation. *Ann. Statist.*, 13, 331–341, 1985.
36. B.M. Hill. A simple general approach to inference about the tail of a distribution. *Ann. Statist.*, 3, 1163–1174, 1975.
37. M.M. Neves, M.I. Gomes, F. Figueiredo and D. Prata-Gomes. Modeling extreme events: sample fraction adaptive choice in parameter estimation. *J. Statistical Theory and Practice*, 9, 1, 184–199, 2015.
38. V. Paulauskas and M. Vaičiulis. On the improvement of Hill and some others estimators. *Lith. Math. J.*, 53, 336–355, 2013.
39. H. Penalva. *Contributos Computacionais e Metodológicos na Estimação do Índice de Valores Extremos*. PhD Thesis, Instituto Superior de Agronomia, Universidade de Lisboa, 2017.
40. H. Penalva, F. Caeiro, M.I. Gomes and M.M. Neves. *An Efficient Naive Generalization of the Hill Estimator—Discrepancy between Asymptotic and Finite Sample Behaviour*. Notas e Comunicações CEAUL 02/2016. Available at: <http://www.ceaul.fc.ul.pt/notas.html?ano=2016>, 2016.
41. H. Penalva, M.I. Gomes, F. Caeiro and M.M. Neves. A couple of non reduced bias generalized means in extreme value theory: an asymptotic comparison. *Revstat*, accepted. Available at: <https://www.ine.pt/revstat/pdf/Acoupleofnonreducedbiasgeneralizedmeansinextremevaluetheory.pdf>, 2018.
42. L. Peng. Asymptotically unbiased estimators for the extreme-value index. *Statist. Probab. Lett.*, 38, 2, 107–115, 1998.
43. C. Scarrot and A. MacDonald. A review of extreme value threshold estimation and uncertainty quantification. *Revstat—Statist. J.*, 10, 1, 33–60, 2012.

Estimating parameters for systems of ordinary differential equations using the principle of stochastic Runge-Kutta solvers

Flavius Guias¹

Department of Mechanical Engineering, Dortmund University of Applied Sciences and Arts, Sonnenstr.96, 44139 Dortmund, Germany
(E-mail: flavius.guias@fh-dortmund.de)

Abstract. The values of the parameter vector driving a system of ordinary differential equations are determined by an optimization algorithm which minimizes an error function constructed in order to match given statistical data. For this we employ the principle of the stochastic Runge-Kutta type scheme for autonomous ordinary differential equations introduced by the author in [1] and [2]. The method is based on a predictor computed as a path of a Markov jump process whose values are improved by using steps of Runge-Kutta type, which have a strong effect in reducing the stochastic fluctuations. The result of this scheme is a smooth numerical approximation of high order of the solution of the differential equation. In the present paper, instead of using the data produced by the Markov jump process, we use the variations of the statistical data for computing the predictor and obtain next an improved approximation by the Runge-Kutta steps on a short time interval, where the parameters are assumed to be constant. We use an initial guess for the parameter vector and minimize then the error functional by standard optimization methods in MATLAB in order to obtain an optimal set of parameters which delivers a solution as close as possible to the real data points. We illustrate this approach by considering an application to a modified SIR-system which models the time evolution of an epidemy and use statistical data available for COVID-19.

Keywords: ordinary differential equations, SIR model, Markov jump processes, COVID-19.

1 Introduction

The general problem which motivates the present paper can be described as calibrating the parameters of a system of ordinary differential equations in order to match given statistical data. Here we can distinguish between constant parameters and parameters which may vary in time. The main impulse of this research was given by the attempt to set up a mathematical model which can give insights regarding the time evolution of the COVID-19 pandemy in several countries, by matching the statistical data provided for example at <https://worldometers.info/coronavirus/> .

¹*6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain*



The SIR model is a standard approach used for the modeling of epidemics, see [3], [4]. It belongs to the so-called *compartmental models*, since the population is divided into several compartments: S - *susceptible*, I - *infected* and R - *recovered*. The dynamics occurs in several stages, by transitions from one stage to the next. Susceptible individuals can get the disease by contact with infected ones. The corresponding rate is directly proportional to the numbers S and I, while the transition from I to R occurs at a rate proportional to the number I of infected individuals. A person in this last stage is already immune and cannot contribute anymore to the spread of the disease. The system of ordinary equations which corresponds to the basic SIR model is the following:

$$\begin{aligned}\frac{dS}{dt} &= -\beta \cdot I \cdot \frac{S}{N} \\ \frac{dI}{dt} &= \beta \cdot I \cdot \frac{S}{N} - \alpha \cdot I \\ \frac{dR}{dt} &= \alpha \cdot I\end{aligned}\tag{1}$$

The number N denotes the size of the whole population and the presence of the factor N^{-1} is necessary for the correct scaling of the bilinear terms. The coefficients of the linear terms, i.e. the rates of passing from one state to another, are inversely proportional to the average time spent in the corresponding state. With T_{inf} the average time spent in the infectious state, we can therefore assume that $\alpha = T_{inf}^{-1}$, while for β we can consider the form $\beta = R(t) \cdot T_{inf}^{-1}$. The term $R(t)$ denotes the time dependent *effective reproduction number*, i.e. the average number of further infections produced by the contacts with one infectious individual. At the beginning of the epidemic its value is equal to R_0 , the so called *basic reproduction number*, but after that it may vary due to restrictions, social distancing, increased frequency of testing, etc.

In this paper we will modify the above SIR model in order to describe adequately the features of the COVID-19 disease and, on the other hand, to account for the available statistical data, which records the numbers of new infections, the current infections, the recoveries and the deaths.

First, we have to add a new component D , which counts the death cases and to consider transitions into D from the state I and sometimes directly from S , if the infection was discovered and recorded only after the death of the patient.

Another reason to build in additional features is the fact that the standard SIR model doesn't match the structure of the statistical data. In the simple model every person in state I is considered to be a potential propagator of the infection, which in this concrete case is not true. As we will see from the calibrated models, the average time spent in state I until an individual is officially reported as recovered is at least 20 days, but it may be also much larger, depending on the reporting policy of each particular country. However, the actual average time in which an individual is infectious and can spread the disease is much smaller. In the case of COVID-19 this value is considered in [6] to be of about only 4 days: a person is highly infectious 1-2 days before

the onset of the symptoms and 2-3 days after that, since once the disease is suspected or confirmed, the ill person is in most cases isolated and/or tries consciously to avoid the further spreading of the disease. In fact only a small part of the group of 'active cases', as recorded by the statistics and modeled in first instance by the variable I , is contributing to the spread of the disease. Therefore it makes sense to split this category as $I = I_1 + I_2$, where the variables I_1 and I_2 denote respectively the infectious individuals and the ill individuals which are not infectious any more, including also the ones which have effectively recovered, but not reported as such by the statistics.

By the above considerations, the system of differential equations which we will use in this paper is the following:

$$\begin{aligned}
\frac{dS}{dt} &= -\frac{R(t)}{T_{inf}} \cdot I_1 \cdot \frac{S}{N} - \tilde{\mu}_d \cdot S \\
\frac{dI_1}{dt} &= \frac{R(t)}{T_{inf}} \cdot I_1 \cdot \frac{S}{N} - T_{inf}^{-1} \cdot I_1 \\
\frac{dI_2}{dt} &= T_{inf}^{-1} \cdot I_1 - T_{conv}^{-1} \cdot I_2 - \mu_d \cdot I_2 \\
\frac{dR}{dt} &= T_{conv}^{-1} \cdot I_2 \\
\frac{dD}{dt} &= \mu_d \cdot I_2 + \tilde{\mu}_d \cdot S
\end{aligned} \tag{2}$$

T_{conv} is the average convalescence time, i.e. the time spent in the state I_2 . We assume that deaths can not occur from state I_1 (that is not within the first few days after onset of the disease), but only at later stages, from the state I_2 with daily death rate μ_d . For countries where the reporting policy counts COVID-19 cases also if the positive testing occurred only after the death of the person, we will assume a positive death rate $\tilde{\mu}_d$ which describes the transition directly from S to D , otherwise this parameter is considered to be 0.

Our goal is to develop a method to estimate the parameters $R(t)$, T_{inf} , T_{conv} , μ_d (and possibly $\tilde{\mu}_d$) by minimizing the mean-square error of the solution of the system relative to a given data set. The statistical data consists in daily time series recording the number of new infections, the active cases, recoveries and deaths. Since we have $S + I + R + D = N$, it is sufficient to match only three of the four components, namely I , R , D . The initial conditions $I(0)$, $R(0)$, $D(0)$ are taken from the data set, basically at the time point where we have about 100 cases.

At this stage we have to make certain additional assumptions. Firstly, since we know only $I(0)$, we consider that $I_1(0) = f \cdot I(0)$ and $I_2(0) = (1 - f) \cdot I(0)$, introducing therefore a new parameter f which has also to be estimated. Secondly, we will assume that $R(t)$ takes constant values over d consecutive days. As current value we take $d = 4$.

We next give a brief overview on the structure and the contents of the present paper.

In section 2 we present the optimization algorithm which we use for estimating the parameters and which can be applied to a general class of problems.

For a given vector of parameters we basically need a numerical approximation of the solution of the system in the time points corresponding to those of the data set. For this purpose we use a variant of the stochastic Runge-Kutta solver developed in [1] and [2]. The original scheme uses as a predictor the path of a Markov jump process and by applying corrections steps of Runge-Kutta type one obtains an improved approximation. In this case however, instead of the simulated jump process we use the daily variations of the given data set. Since we don't need to perform a stochastic simulation, the method is fast and, if we are close to the optimum, it shows a good agreement with the results of a standard ODE-solver in MATLAB. The optimization step is performed also by a standard MATLAB function, which searches for the set of parameters which minimizes the mean square error between the computed and given data. By using a splitting algorithm and alternative computations we can obtain better results. Namely, we alternatively fix either the vector of the time dependent parameters and optimize with respect to the time independent ones, or we change the roles, and can thus perform several iterations with different starting values in the search for the minimum.

In section 3 we apply this method to the system of equations (2) using the COVID-19 data from countries like Germany, Italy, Spain, France, Sweden and Romania. Comparing the dynamics of the epidemy, we can note that the curves of the estimated parameter $R(t)$ have for all countries a similar profile, which is consistent also to the official data. For Germany the values of the reproduction number are estimated and provided by the Robert-Koch Institute, see [6] and [7], by using the data set of new infections.

While in the case of the parameters $R(t)$ and T_{inf} , which both influence the dynamics of new infections, we have basically a similar development in all countries (for T_{inf} values between 3.16-3.86), the values of the other time-independent estimated parameters show significant differences in handling the reporting of the recoveries and of the death cases. This is reflected also in the different speeds of decay of the curves of the active cases, after passing over the peak of the pandemy in the respective countries.

In countries like Italy, Spain or France, if we take first $\tilde{\mu}_d = 0$ (no direct transitions from S to D), the death curve computed by our model deviates from the real data and the death rates μ_d are about double than in Germany. However, by introducing also the parameter $\tilde{\mu}_d$, the model fits much better with the reported data and the death rate μ_d becomes closer to that of Germany. We can also note that the estimated values of the average time T_{conv} spent in the state I_2 show large variations among different countries, revealing very different policies of reporting. The values vary from about 12 in the case of Germany, to 18 in the case of Spain, 30 for Romania, 43-44 for Italy/France and 120 for Sweden. This exhibits also an influence on the curves of active cases, since the conclusion drawn from these results is that in many countries we have substantial delays in reporting the recoveries. Which in turn implies that the curves of the active cases remain at relatively high levels, as long as the effective recoveries are not officially reported.

In the case of France, even looking only at the plots of the data sets, we note that they exhibit locally a chaotic profile, especially the curve of active cases,

and the discrepancy between the estimated curve and original data suggests an irregular rhythm of reporting the recoveries, indicating that T_{conv} is probably also variable in time. However, with the exception of France, for all other considered countries the results produced by the model with the estimated parameters shows a good agreement with the given statistical data.

Finally, we summarize the results of the present paper in the concluding section 4.

2 The algorithm for parameter estimation

The general problem which we are attempting to solve can be formulated as follows. Consider a system of ordinary differential equations of the form $\mathbf{y}'(t) = \mathbf{F}(\mathbf{R}(t), \mathbf{p}; \mathbf{y}(t))$. The equations depend therefore on a set $\mathbf{R}(t)$ of time-dependent parameters and on another set \mathbf{p} of time independent parameters. The goal is to find a set of parameters such that the error in mean-square sense between the solution $\mathbf{y}(t_i)$ and given data points \mathbf{Y}_i , with $i = 1, \dots, n$, is minimal, considering that the initial condition is $\mathbf{y}(0) = \mathbf{Y}_0$. Typically we consider $t_i = i$, if the time unit is for example one day.

For this we set up a MATLAB-function which returns the above mentioned error depending on the current set of parameters and perform the search for a minimizer. Each function call needs to calculate numerical approximations \mathbf{y}_i^* for $\mathbf{y}(i)$. Our choice for the solver which will be presented below is based on the requirements that it should be cheap and accurate, since the optimization algorithm involves a large number of function evaluations. The modified version of a stochastic algorithm of Runge-Kutta types presented in [1] and [2] fulfills these criteria.

2.1 The numerical scheme driven by the variations of the statistical data

Assuming that we have an approximation \mathbf{y}_i^* , the steps for computing \mathbf{y}_{i+1}^* , with a time interval h between them, are the following.

1. Compute predictors $\bar{\mathbf{y}}_{i+1/2}$ at time distance $h/2$ and $\bar{\mathbf{y}}_{i+1}$ at time distance h from \mathbf{y}_i^* . The original method uses the direct simulation of a Markov jump process (see also [5]), while here we use the increments of the data set. If the data set provides only the daily variations, at the mid point we can consider as increment a normally distributed random variable with mean equal to the half of the daily variation and a small variance, while for the second point we consider the remaining difference to the daily variation. This step can be performed very fast, since we don't simulate each jump of the Markov process, involving sampling from a large probability table and updating after every change. Moreover, in this latter case, in order to cover an interval of length h , we would need to perform a large number of jumps if we aim at a good precision. Alternatively, applying the so called *tau-leap* method from [5] to simulate several jumps at once, won't make

use of the data set, but only of the actual values of the parameters, so we don't pursue this approach.

We have to add a remark concerning the particular model considered in this paper. From the data set we can read only the variations of the component I , but not of the parts I_1 and I_2 . For their initial values we made the assumption that they are proportions of f and $1 - f$ from I , assuming for their variations the same proportions from the variations of I . The parameter f , which is also part of the parameter set to be estimated, is in fact a hidden one (since for the error we compare only the sum $I = I_1 + I_2$ to the values of the data set), entering in the initial condition but also in the approximating dynamics for the model as well.

2. The first correction step computes improved values at the endpoints of the two subintervals of length $h/2$ by applying a similar step to the Runge-Kutta method of order 2. Namely we take

$$\begin{aligned}\tilde{\mathbf{y}}_{i+1/2} &= \mathbf{y}_i^* + (h/4) \cdot (\mathbf{F}(\mathbf{y}_i^*) + \mathbf{F}(\tilde{\mathbf{y}}_{i+1/2})) \\ \tilde{\mathbf{y}}_{i+1} &= \tilde{\mathbf{y}}_{i+1/2} + (h/4) \cdot (\mathbf{F}(\tilde{\mathbf{y}}_{i+1/2}) + \mathbf{F}(\tilde{\mathbf{y}}_{i+1}))\end{aligned}$$

For simplicity, in the above and in the next formulas the dependence of \mathbf{F} on the parameters is suppressed.

3. Finally, we apply another correction step on the whole time interval of length h by a step similar to the Runge-Kutta method of order 3 and obtain thus the desired approximation:

$$\mathbf{y}_{i+1}^* = \mathbf{y}_i^* + (h/6) \cdot (\mathbf{F}(\mathbf{y}_i^*) + 4\mathbf{F}(\tilde{\mathbf{y}}_{i+1/2}) + \mathbf{F}(\tilde{\mathbf{y}}_{i+1}))$$

If we take $h = 1$, the above method computes \mathbf{y}_{i+1}^* starting from \mathbf{y}_i^* by considering also intermediate values at the midpoint. In practice, in order to improve precision, we take $h = 0.5$, so we compute the desired final value at the end of the time interval of length 1 in two steps, by computing first an intermediate value $\mathbf{y}_{i+1/2}^*$ at the midpoint, and starting from this the value we get \mathbf{y}_{i+1}^* . For the considered model this turns out to be enough, but of course it is possible to apply the method with arbitrarily small time steps $h = 0.25, 0.125, \dots$. Since in the computations of the values \mathbf{y}_i^* we use for the predictors directly the dynamics drawn from the data set, if the parameters are far from the optimum, these values are not close to the solution of the system computed by a classical solver, being in principle more closer to the data set, especially concerning the profile of the curves. However, if we have a good set of parameters, the two approximations are close to each other. In all the figures below we plot the curves computed by the above method but also (with thin dotted lines, not mentioned in the legend) the solutions curves computed by a MATLAB ODE-solver for the same parameter set. For the most part of the curves they are almost indistinguishable, but this good agreement is reached only for $h = 0.5$. If we take $h = 1$, the difference would be slightly larger.

2.2 The optimization algorithm

The first step of the estimation procedure consists in finding a good first guess for the values $R(t)$ of the time dependent parameter. Recall that we consider

them to be constant over consecutive intervals of d days, usually taking $d = 4$, unless otherwise specified. We can perform this by hand, by trial and error, aiming that the polygonal chain of the computed values to be optically as close as possible to the data points of the state I . For the other time-independent parameters we use first a rough guess, for example $T_{inf} = 4$, $\mu_d = 0.005$, $f = 0.5$, while for the values for T_{conv} , which vary strongly from case to case, we find a proper starting value also by trial and error, even if the match is not very precise.

After this preliminary step for determining the initial value of the parameter set, we apply several runs of a splitting algorithm, by optimizing alternatively with respect to \mathbf{p} by keeping \mathbf{R} fixed, and then with respect to \mathbf{R} by keeping \mathbf{p} fixed. The parameter vector \mathbf{p} has 4 to 5 elements, while the length of \mathbf{R} depends on the number of data points and on d . Typically we have $d = 4$ and 72 data points, so \mathbf{R} has 18 elements, but in one case we will consider also $d = 1$. We finally note that all computations were performed in MATLAB.

3 Applications to the COVID-19 model

We present first the summary of the results for the considered countries.

country	start	days	T_{inf}	T_{conv}	$\mu_d(\tilde{\mu}_d = 0)$	$\mu_d/\tilde{\mu}_d$	imprv.err.
Germany	01.03.20	72	3.86	12.23	0.0043	0.0034 / 2.50e-7	0.67%
Italy	22.02.20	80	3.40	43.01	0.0090	0.0041 / 3.58e-6	28.32%
Spain	01.03.20	72	3.16	18.27	0.0105	0.0047 / 5.00e-6	14.83%
France	01.03.20	72	3.79	45.04	0.0112	0.0067 / 2.29e-6	6.23%
Sweden	06.03.20	68	3.67	120.49	0.0083	0.0065 / 9.18e-7	1.44%
Romania	14.03.20	60	3.64	29.92	0.0054	0.0032 / 3.32e-7	3.60%

We estimated the parameters by considering first $\tilde{\mu}_d = 0$ and after that by including also this parameter in the set of those to be estimated. Recall that a positive value of $\tilde{\mu}_d$ would mean that we have direct transitions $S \rightarrow D$. That is, we speak mostly of asymptomatic persons at which the infection was discovered only after their death, which is likely to have a different cause, since otherwise, if typical symptoms would have been recorded during his lifetime, the patient would pass through the states $S \rightarrow I_1 \rightarrow I_2 \rightarrow D$.

The last column of the table records the relative improvement of the error with respect to the data points if we assume $\tilde{\mu}_d \neq 0$ compared to the case when we disregard this parameter. We can note three different types of behavior. For countries like Germany, Sweden or Romania, the improvement of the error and the effective value of $\tilde{\mu}_d$ are small (on average 1% , respectively of order 10^{-7}). For these cases we assume therefore that keeping $\tilde{\mu}_d = 0$ is a reasonable assumption. For countries like Italy and Spain however, the improvements in the error are substantial, as well as the values of $\tilde{\mu}_d$, which are of order 10^{-6} . Here it seems reasonable to consider that the effect $S \rightarrow D$ is significant and must be included in the model. The third case is that of France, which is somehow inbetween, but closer to Italy and Spain, so we finally include it in the same category, but with the mention that, as we will see below, the data of

France is more problematic from the point of view of our model, since probably the rate of reporting the recoveries is not constant in time.

The table contains the values of T_{inf} and T_{conv} computed in the choice for $\tilde{\mu}_d$ which is *not* marked in gray, but the results show to be stable and very similar also in the other case. Especially the parameter T_{inf} shows similar values for all considered countries as well as $R(t)$, as can be seen in Figure 1.

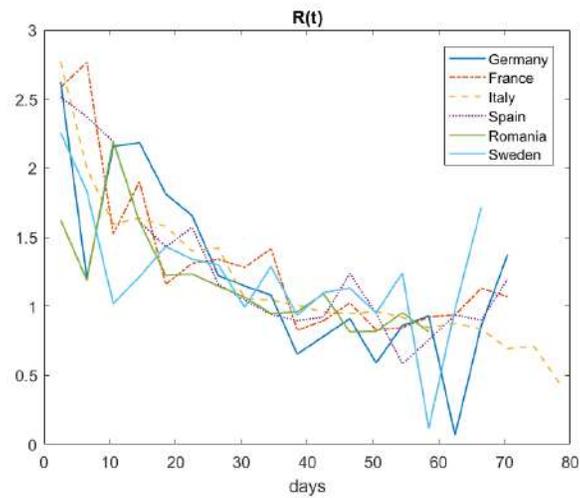


Fig. 1. The evolution of the effective reproduction number $R(t)$ in different countries

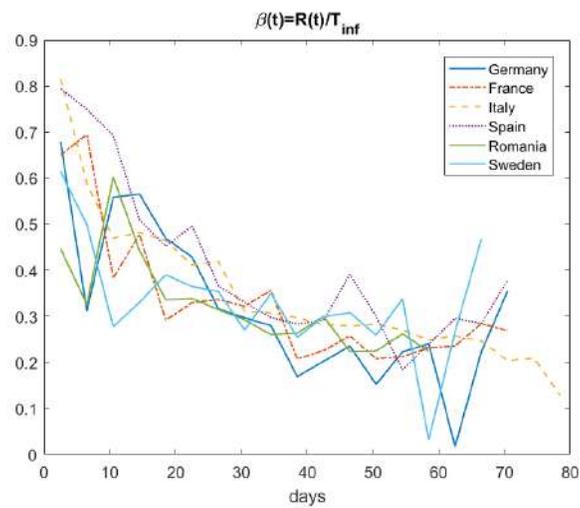


Fig. 2. The evolution of the factor $\beta(t) = R(t)/T_{inf}$ in different countries

The infection dynamics shows to be very similar in all countries, since the values of $R(t)$ and T_{inf} are within the same range. Note that if the differences between the values of T_{inf} would be substantial, comparing the different $R(t)$ wouldn't be very relevant, since in fact the rate of new infections is driven by the factor $\beta(t) = R(t)/T_{inf}$, which is plotted in Figure 2. Since however $R(t)$ and T_{inf} are similar, we conclude therefore that the values of the other parameters, namely T_{conv} , μ_d and $\tilde{\mu}_d$, are the factors that explain the differences in the plots of the data sets of each country.

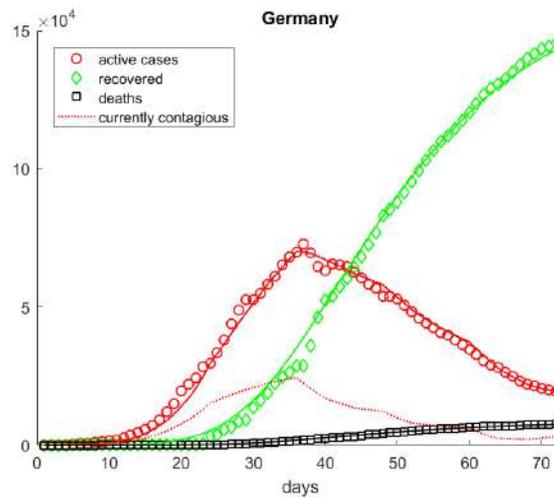


Fig. 3. Data and approximation for Germany

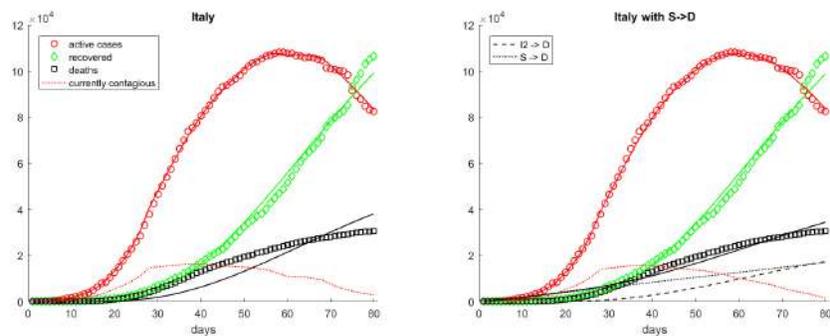


Fig. 4. Data and approximation for Italy

Figure 3 shows the plot for Germany with $\tilde{\mu}_d = 0$, where the dotted line represents (here and in the plots for the other countries) the curve for I_1 , i.e. the currently contagious individuals. Figures 4 and 5 compare the results for

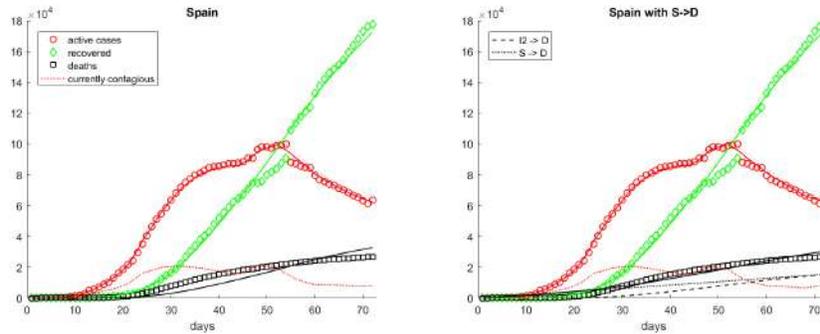


Fig. 5. Data and approximation for Spain

Italy and Spain with $\tilde{\mu}_d = 0$ (left picture) and $\tilde{\mu}_d \neq 0$ (right picture). We note the the approximation of the death curve is much better in the second case, where the black dashed line depicts the deaths $I_2 \rightarrow D$, whereas the black dotted line the deaths $S \rightarrow D$. The continuous black line approximates the total deaths as recorded in the statistics.

Figure 6 shows the plots for Romania and Sweden. We note that for Sweden we computed a very large value for T_{conv} , and this extreme behaviour can be also seen in the corresponding graphic. The statistical data for the total number recoveries shows constant values over large time intervals, followed by sudden jumps of large magnitude. Both elements lead to the evident conclusion that the data for Sweden concerning the recoveries exhibits significant delays in reporting and updating. We also note that for Sweden the estimated curves for the deaths and recoveries are almost identical and can hardly be distinguished.

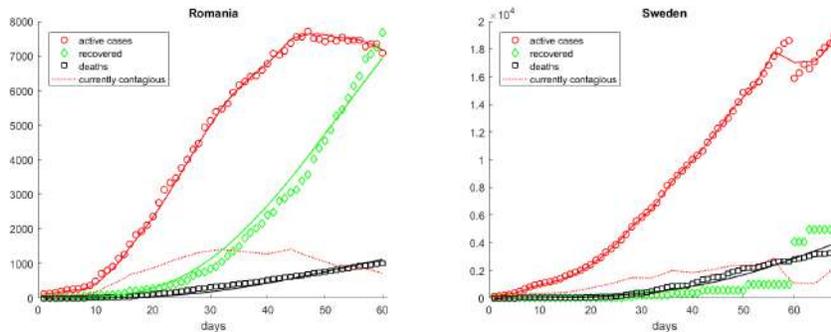


Fig. 6. Data and approximation for Romania (left) and Sweden (right)

The results for France are shown in Figure 7 and show that by considering $\tilde{\mu}_d > 0$, the approximation of the death curve is indeed improving, but the discrepancy of the model compared to the data of recovered patients point to the fact that T_{conv} probably cannot be viewed as a global constant parameter, i.e. for France the reportings do not occur with a regular frequency.

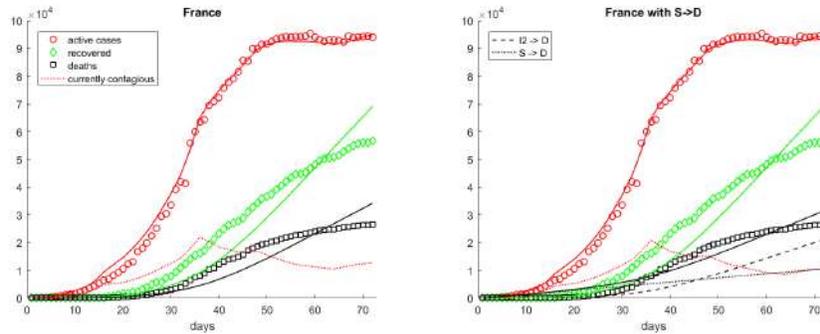


Fig. 7. Data and approximation for France

Using the data from Germany, we will discuss next a comparison of the values of $R(t)$ computed by our method and the official values reported by the Robert-Koch Institute (RKI), see [6], [7]. We used for comparison the data provided there between March 6 and May 6. The procedure used in the mentioned references, called *nowcasting*, delivers a smoothed value by a 4-day moving average based on the data of the previous four days. Moreover, it attempts to estimate the real value of $R(t)$ in the corresponding day, by taking also into account the delays in the reportings. Therefore, the recorded data reflects basically the situation at about four days before. For this reason, in order to be able to compare with our data, we have to shift our curve corresponding to the R -values with four days to the left, in order to match the above described delay. Furthermore, the plot takes into account that the RKI-data starts in 6.03, whereas our data in 1.03.

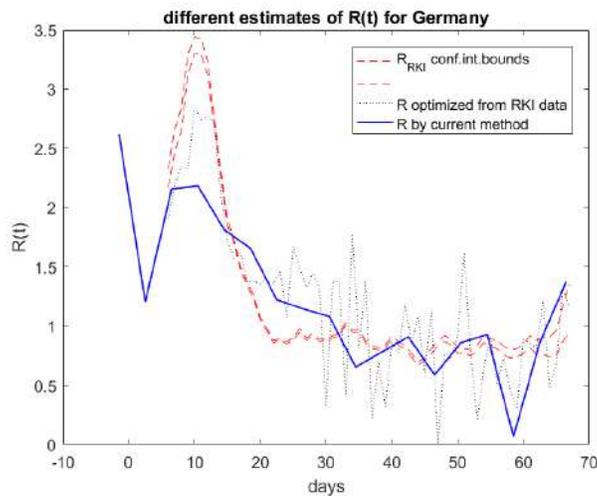


Fig. 8. Approximating the reproduction number $R(t)$ for Germany

The results are plotted in Figure 8, where we can see three curves. The dashed lines represent the boundaries of the confidence intervals of $R(t)$ provided by RKI, while the continuous line connects the values computed by the present method, located in the middle of the 4-day intervals where we assume that $R(t)$ is constant. In the same figure we plot with the dotted line the 'optimized' valued $R(t)$ using as starting point the RKI-data. More precisely, if we try to fix the $R(t)$'s as given by RKI and try to compute optimized values only for the time-independent parameters, we obtain unplausible values and a larger approximation error as in our method. The reason for this is that the RKI data is a smoothed version of the raw data which is not explicitly available, but this smoothing alters the dynamics (since we compare our results to unsmoothed data sets). Consequently, we performed several optimization steps according to our method, using as initial guess for $R(t)$ the official smooth RKI data and considering $d = 1$. The obtained result is represented in Figure 8 by the dotted line and shows a good agreement to the curve computed by the method of this paper, which in some sense is also a smoothed one.

4 Conclusion

In this paper we presented a general method of estimating the parameters of a system of ordinary differential equations in order to match given statistical data. As application we chose a modified SIR model which is designed to describe the time evolution of the COVID-19 pandemy. The results turned out to be promising and the estimated values of the parameters are able to give insights regarding the evolution of the pandemy in several countries, but also to reveal different policies and approaches in reporting the official data.

References

1. F. Guiaş and P. Eremeev. Improving the stochastic direct simulation method with applications to evolution partial differential equations. *Appl. Math. Comput.*, **289** : 353–370, (2016)
2. F. Guiaş. High precision stochastic solvers for large autonomous systems of differential equations. *Int. J. Math. Models and Meth. in Appl. Sci.* **13** : 60-63, (2019)
3. F. Brauer and C. Castillo-Chávez. *Mathematical Models in Population Biology and Epidemiology*. Springer, New-York, 2001.
4. F. Brauer. Compartmental Models in Epidemiology in F. Brauer, P. van den Driessche and J. Wu (Eds.) *Mathematical Epidemiology*, Springer, Berlin, Heidelberg 2008, Chapter 2 : 19–79
5. D. Gillespie. Stochastic Simulation of Chemical Kinetics. *Annu. Rev. Phys. Chem.*, **58** : 35–55, 2007.
6. M. an der Heiden, O. Hamouda. Schätzung der aktuellen Entwicklung der SARS-CoV-2-Epidemie in Deutschland – Nowcasting. *Epid Bull* 2020;17 : 10 – 16 — DOI 10.25646/6692.4
7. https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Projekte_RKI/Nowcasting.html

Modelling Nigerian Female Mortality: An Application of Four Stochastic Mortality Models

Adegbilero-Iwari Oluwaseun Eniola.^{1,2*}, Chukwu Angela Unna.²

¹Department of Community Medicine, Afe-Babalola University, Ado-Ekiti, Nigeria

²Department of Statistics, University of Ibadan, Ibadan, Nigeria.

*E-mail of the corresponding author: seuneniola01@gmail.com

Abstract

Nigeria, with an estimated population of over 200 million people is depicted as the most populous black Nation in the World with the female gender constituting about 49.4% of its total Population. However, from infancy to adulthood, the female folk in Nigeria is faced with challenges; from low-socio cultural values to poor socio-economic growth and a frail health care system. In the midst of the afore-mentioned challenges, survival for her is not a *privilege* but a *fight*.

With the aid of four stochastic mortality models and mortality data obtained from the Global Health Observatory, this study assesses the peculiarities surrounding the Nigerian age-specific female mortality data. It also illustrates the potentiality of the Gamma-Normal version of the Lee-Carter model over three existing variants of the Lee and Carter (1992) model.

Keywords: Age, Female mortality, Population, Mortality models, Gamma-normal



1. Introduction

1.1 *The Role of mortality models*

Whether in developed or developing countries, mortality models play a very significant role in demographic projections as they help to provide a sufficient understanding of mortality trends and the uncertainty of future mortality rates.

Lee and Carter (1992)^[1] published a stochastic model for the long-run forecasts of the level and age pattern of mortality. Though designed for the United States, the model has been applied by many developed countries to model their mortality pattern with satisfactory results.^[2,3,4,5,6,7,8] The model has not gained much prominence in Africa. However, it has been applied on Nigerian and Zimbabwean mortality data.^[9,10,11]

1.2 *Study Background*

Nigeria is the most populous black nation in the world with an estimated population of about 201 million in 2019.^[12] The female gender constitutes about 49.4% of its total Population. Unlike her counterpart in the developed regions of the world, the female folk in Nigeria is faced with challenges ranging from low-socio cultural values to poor socio-economic growth and a frail health care system. In 2015, the average life expectancy at birth for a female in the developed and less developed region of the world stood at 82 and 72 years, respectively. For a Nigerian female, it was estimated as 54 years ranking the country as the second lowest in average life expectancy in West Africa and the third lowest in Africa.^[13]

1.3 *Child and Maternal Mortality in Nigeria*

Child or under-five mortality is the probability of dying between birth and age 5 years expressed per 1,000 live births. Of the six regions of the world, only Europe achieved the Millennium Development Goal (MDG) target of reducing child mortality by two-thirds between 1990 and 2015.^[13] A further reduction of child mortality rate (CMR) to less than 25 deaths per 1,000 live births by 2030 is target 3.2 of the Sustainable Development Goals (SDGs).^[14] The Nigerian Demographic and Health Survey (NDHS) in its 2018 report highlighted a CMR of 132 deaths per 1,000 live births with higher mortality rates in the rural parts of the Country.^[15]

Maternal mortality affects women in the child-bearing age-group (15-49) years. In 2015, Nigeria alone accounted for 19% of all estimated global maternal deaths, with an estimated maternal mortality ratio (MMR) of over 800 maternal deaths per 100,000 live births.^[16] Retrospective reviews carried out in the Northern part of the Country had reported MMR of about 1,271 to 1,625 maternal deaths per 100,000 live birth.^[17,18] On the contrary, in the 46 most developed countries of the world, the total number of maternal deaths in 2015 was only 12 deaths per 100,000 live births.^[19]

1.4 Objectives of the study

- i. To examine past time trends in the pattern of mortality across the age-groups.
- ii. To illustrate the potentiality of the Gamma-Normal Lee-Carter over three other variants of the Lee-Carter model

2 Review of Methods

2.1 The Lee-Carter model

The Lee and Carter (1992)^[1] method is based on a blend of statistical time series techniques and a simple approach to dealing with the age distribution of mortality. The methodology defines the logarithm of a time series of age-specific death rates as the addition of an age-specific component that is time independent and another component which is the product of a time-varying parameter and an age-specific component that represents how rapidly or slowly mortality at each age varies when the general level of mortality changes. The resulting estimate of the time-varying parameter is then modeled and forecast as a stochastic time series using standard methods. The Lee-Carter model is specified as:

$$\ln m_{xt} = a_x + b_x k_t + \varepsilon_{xt} \quad (1)$$

The practical use of the model in (1) assumes that the error term ε_{xt} are normally distributed i.e. $\varepsilon_{xt} \sim N(0, \sigma^2)$.

2.2 The Brouhns model

Brouhns *et al.* (2002)^[20] implemented an alternative procedure for a log-bilinear formulation of the Lee-Carter model based on Poisson error structures. Specifically, they switched from a classical linear model to a generalized linear model, substituting Poisson random variation for the number of deaths for an additive error term on the logarithm of mortality rates. The researchers pointed out that the main drawback of the OLS estimation through the Singular value decomposition (SVD) procedure is that the errors are assumed to be homoscedastic. Instead of resorting to SVD for estimating a_x , β_x and k_t they determine these parameters by maximizing the log-likelihood based on model:

$$D_{xt} = e^{a_x + b_x k_t} E_{xt} + e^{\varepsilon_{xt}} \quad (2)$$

$$D_{xt} \sim \text{poisson}(E_{xt} \mu_x(t)), \quad \mu_x(t) = \exp(a_x + \beta_x k_t)$$

Where, D_{xt} refers to death counts at age x and time t and E_{xt} is the population exposed to risk at age x and time t .

2.3 The Renshaw-Haberman model

Renshaw and Haberman (2006)^[21] extension of the Lee-carter method bears some resemblance with [2]. However, the former proposed the incorporation of a cohort effect in their model. This

was motivated by the lack of goodness of fit of [1] to the death rates of England and Wales and the fact that its application led to a cohort effect of decreasing death rates. Furthermore they investigated the feasibility of constructing mortality forecasts on the basis of the first two sets of Singular Value Decomposition vectors, rather than just on the first set of such vectors, as obtained in the Lee–Carter approach. Under the Gaussian framework the model is expressed as:

$$\ln [m(x, t)] = a(x) + b_1(x)k_1(t) + b_2(x) l_{t-x} + \varepsilon_{xt} \quad (3)$$

2.4 The Gamma-Normal Lee-Carter model

Adegbilero-Iwari and Chukwu (2018)^[22] proposed the gamma-normal version of the Lee-Carter model. Their work was motivated by a prior experience of the Lee-Carter model with the Nigerian female mortality data in 2012.^[9] Their proposed methodology utilized the contributions of Zografos and Balakrishnan (2009)^[23] who suggested the novel gamma generating distribution and Lima *et al.* (2015)^[24] who proposed the gamma-normal distribution. The gamma-normal distribution is particularly useful for heterogeneous data allowing for greater flexibility of its tails. The model is given as

$$\ln m_{xt} = a_x + b_x k_t + \varepsilon_{xt}^* \quad (4)$$

$$\text{Where } \varepsilon_{xt}^* \sim GN(\alpha, \mu, \sigma^2)$$

Parameter α in (4) gives an idea of the degree of heterogeneity present in a given data set

2.4.1 Parameters of the Gamma-normal Lee-Carter model

The gamma-generated distribution by parent F as introduced by [23] has its probability density function given as:

$$g(y) = \frac{1}{\Gamma\alpha} [-\ln[1 - F(y)]]^{\alpha-1} f(y), \quad y \in R, \quad \alpha > 0. \quad (5)$$

The Lee-carter approach assumes that its error term follows a normal distribution whose probability density function must satisfy (6) and (7) below.

$$f(y) = \frac{1}{\sigma} \Phi \left[\frac{y - (a_x + b_x k_t)}{\sigma} \right] \quad (6)$$

$$\Phi \left[\frac{y - (a_x + b_x k_t)}{\sigma} \right] = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \left[\frac{y - (a_x + b_x k_t)}{\sigma} \right]^2}$$

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2} \left[\frac{y - (a_x + b_x k_t)}{\sigma} \right]^2}, \quad \sigma > 0, \quad b_x > 0, \quad k_t > 0 \quad (7)$$

Its cumulative distribution function is given by;

$$F(y) = \Phi \left(\frac{y - (a_x + b_x k_t)}{\sigma} \right) \quad (8)$$

The probability density function of the Gamma-normal Lee-Carter is derived by substituting equations (7) and (8) into (5) to give;

$$g(y) = \frac{1}{\Gamma\alpha_x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{y-(a_x+b_xk_t)}{\sigma_i}\right]^2} \left[-\ln\left[1 - \Phi\left(\frac{y-(a_x+b_xk_t)}{\sigma_i}\right)\right]\right]^{\alpha_x-1} \quad (9)$$

its likelihood function is given by:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n \frac{1}{\Gamma\alpha_x} \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i-(a_x+b_xk_t)}{\sigma_i}\right)^2} \left\{-\ln\left[1 - \Phi\left(\frac{y_i-(a_x+b_xk_t)}{\sigma_i}\right)\right]\right\}^{\alpha_x-1} \\ &= \prod_{i=1}^n e^{-\frac{1}{2\sigma_i^2}[y_i-(a_x+b_xk_t)]^2} \left\{-\ln\left[1 - \Phi\left(\frac{y_i-(a_x+b_xk_t)}{\sigma_i}\right)\right]\right\}^{\alpha_x-1} (\sigma_i\Gamma\alpha_x)^{-n} (2\pi)^{-\frac{n}{2}} \end{aligned} \quad (10)$$

The log-likelihood function is expressed as:

$$\begin{aligned} \ln L(\theta) &= -n(\ln \sigma_i + \ln \Gamma\alpha_x) - \frac{n}{2}(\ln 2\pi) - \frac{1}{2\sigma_i^2} \sum_{i=1}^n [y_i - (a_x + b_xk_t)]^2 + \\ &(\alpha_x - 1) \sum_{i=1}^n \ln \left\{-\ln \left[1 - \Phi\left(\frac{y_i-(a_x+b_xk_t)}{\sigma_i}\right)\right]\right\} \end{aligned} \quad (11)$$

Therefore with respect to (10), the components of the score vector $U(\theta)$ of the Gamma-normal Lee-Carter are finally of the forms:

$$U_{\alpha_x}(\theta) = \frac{1}{\sigma_i} \sum_t \left[\frac{y_i-(a_x+b_xk_t)}{\sigma_i}\right] + \frac{(\alpha-1)}{\sigma_i} \sum_t \frac{\phi\left[\frac{y_i-(a_x+b_xk_t)}{\sigma_i}\right]}{\left[1-\Phi\left(\frac{y_i-(a_x+b_xk_t)}{\sigma_i}\right)\right] \ln\left[1-\Phi\left(\frac{y_i-(a_x+b_xk_t)}{\sigma_i}\right)\right]} \quad (12)$$

$$U_{\alpha}(\theta) = \sum_t \ln \left\{-\ln \left[1 - \Phi\left(\frac{y_i-(a_x+b_xk_t)}{\sigma_i}\right)\right]\right\} - n\psi(\alpha) \quad (13)$$

$$U_{\sigma_i}(\theta) = \frac{n}{\sigma_i} + \frac{1}{\sigma_i} \sum_{i=1}^n \left[\frac{y_i-(a_x+b_xk_t)}{\sigma_i}\right]^2 + \frac{(\alpha-1)}{\sigma_i} \sum_{i=1}^n \frac{\left[\frac{y_i-(a_x+b_xk_t)}{\sigma_i}\right] \phi\left(\frac{y_i-(a_x+b_xk_t)}{\sigma_i}\right)}{\left[1-\Phi\left(\frac{y_i-(a_x+b_xk_t)}{\sigma_i}\right)\right] \ln\left[1-\Phi\left(\frac{y_i-(a_x+b_xk_t)}{\sigma_i}\right)\right]} \quad (14)$$

$$U_{b_x}(\theta) = \frac{k_t}{\sigma_i} \sum_t \left[\frac{y_i-(a_x+b_xk_t)}{\sigma_i}\right] + \frac{k_t(\alpha-1)}{\sigma_i} \sum_t \frac{\phi\left[\frac{y_i-(a_x+b_xk_t)}{\sigma_i}\right]}{1-\Phi\left(\frac{y_i-(a_x+b_xk_t)}{\sigma_i}\right) \ln\left[1-\Phi\left(\frac{y_i-(a_x+b_xk_t)}{\sigma_i}\right)\right]} \quad (15)$$

$$U_{k_t}(\theta) = \frac{b_x}{\sigma_i} \sum_x \left[\frac{y_i-(a_x+b_xk_t)}{\sigma_i}\right] + \frac{b_x(\alpha-1)}{\sigma_i} \sum_x \frac{\phi\left[\frac{y_i-(a_x+b_xk_t)}{\sigma_i}\right]}{1-\Phi\left(\frac{y_i-(a_x+b_xk_t)}{\sigma_i}\right) \ln\left[1-\Phi\left(\frac{y_i-(a_x+b_xk_t)}{\sigma_i}\right)\right]} \quad (16)$$

Setting these expressions to zero and solving them simultaneously yields the maximum likelihood estimates (MLEs) of the parameters.

3 Results

3.1 Data Source and structure

The data set used for the study is the age-specific mortality data of Nigerian females. It was obtained from the Global Health Observatory, an arm of the WHO Indicator and Measurement Registry (IMR).^[25] For the mortality data set, the age distribution of the population ranges between less than one year old to 85 years and above for the time period 2000 to 2015. Throughout the study, the number of deaths (d_{xt}), the central exposures or the population exposed to risk (E_{xt}) and the mortality rates (m_{xt}) are arranged in a rectangular array format comprising ages (on the row) $x = x_1, x_2, \dots, x_k$ and calendar years (on the columns) $t = t_1, t_2, \dots, t_n$.

3.2 Normality Test

To test for normality, two non-parametric procedures were used; the Kolmogorov-Smirnov and Shapiro-Wilks normality tests. While the Kolmogorov-Smirnov procedure was significant at 1%, confirming non-normality with P-values lesser than or equal to 0.000645 across the 19 age-groups, the Shapiro-Wilks procedure confirmed non-normality with P-values lesser than or equal to 0.07518 in 15 age-groups at 10% significance level. Moreover, across the time periods, the Kolmogorov-Smirnov and Shapiro-Wilks procedures were significant at 1%, confirming non-normality with P-values lesser than or equal to 4.18×10^{-13} and 1.58×10^{-5} . This confirms the need for a model allowing for a non-Gaussian distribution structure.

3.3 Estimate of Parameters

The Gamma-Normal Lee-Carter (GNLC) is compared with three other variants; the Lee and Carter model (LC), the Brouhns model (BR) and the Renshaw-Haberman model (RH). The packages *ilc*, *StMoMo* and *bbmle* in R software were used in obtaining the maximum likelihood estimate of the parameters.^[26,27,28] Results are displayed in Tables 1-4

Table 1: Comparison of the MLEs of Parameter a_x under the four models

	LC	BR	RH	GNLC
<1	-2.461892	-2.461851	-2.6237619	-2.461892
1-4	-4.164983	-4.165188	-4.4210464	-4.164983
5-9	-5.069584	-5.070319	-5.2007739	-5.069584
10-14	-5.683282	-5.685167	-5.6382799	-5.683282
15-19	-5.521461	-5.521461	-5.4084323	-5.521461
20-24	-5.278079	-5.280118	-5.1915108	-5.278079
25-29	-4.977923	-4.979045	-4.9542377	-4.977923
30-34	-4.738466	-4.738359	-4.7842194	-4.738466
35-39	-4.521071	-4.519965	-4.5915937	-4.521071
40-44	-4.479415	-4.478644	-4.5137096	-4.479415
45-49	-4.435550	-4.435540	-4.4027834	-4.435550
50-54	-4.248468	-4.248570	-4.1516312	-4.248468
55-59	-3.974881	-3.974918	-3.8380744	-3.974881

60-64	-3.522756	-3.522780	-3.3607945	-3.522756
65-69	-3.052064	-3.052062	-2.8875846	-3.052064
70-74	-2.525927	-2.525934	-2.3698408	-2.525927
75-79	-2.013919	-2.013953	-1.8768102	-2.013919
80-84	-1.512042	-1.512117	-1.4006350	-1.512042
85+	-1.078266	-1.078415	-0.9954274	-1.078266

3.3.1 Comparison of MLEs of Parameter a_x across all the Models

Table 1 shows the maximum likelihood estimates of Parameter \hat{a}_x across the models. Parameter \hat{a}_x represents the general pattern of mortality by age. A similarity is noticed in the general pattern of mortality by age for all the models under consideration. Generally, the results show that \hat{a}_x values decreased with age from age-groups 0-1 to age-groups 10-14, but increased with age for age-groups 10-14 to 85⁺. This implies that the younger ages (from age-group 10-14) have a lower mortality rate than the older ages. There is a clear indication that Infant mortality rate is higher than child mortality rate which is in turn greater than mortality rate at ages 5-14 years. Across the models, mortality is also increasing as age increases within the child-bearing age-groups.

Table 2: Comparison of the MLEs of Parameter b_x under the four models

	LC	BR	RH	GNLC
<1	1.104649e-01	1.121297e-01	0.013588631	0.1105305631
1-4	1.575868e-01	1.601981e-01	0.056164504	0.1576193323
5-9	5.882967e-02	5.888257e-02	0.233382645	0.0587414376
10-14	8.805851e-02	8.803642e-02	0.147707367	0.0880098908
15-19	-1.394130e-17	1.638890e-16	0.002829039	0.0000267991
20-24	4.742123e-02	4.790192e-02	0.025341864	0.0474709506
25-29	7.796972e-02	7.943062e-02	0.018282374	0.0779606398
30-34	7.702937e-02	7.574334e-02	0.030265294	0.0770945235
35-39	5.927341e-02	5.686279e-02	0.068614830	0.0591113584
40-44	4.805914e-02	4.730043e-02	0.092077470	0.0479492269
45-49	3.783256e-02	3.706698e-02	0.078183912	0.0378279402
50-54	3.807446e-02	3.802905e-02	0.069666548	0.0380193337
55-59	3.910521e-02	3.916809e-02	0.060073108	0.0391321893
60-64	3.730551e-02	3.732741e-02	0.042542319	0.0373319652
65-69	3.213565e-02	3.211966e-02	0.030853607	0.0321785613
70-74	3.041920e-02	3.031044e-02	0.021081115	0.0304437756
75-79	2.631621e-02	2.608956e-02	0.011018519	0.0263550986
80-84	2.089943e-02	2.060288e-02	0.003354608	0.0209355434

85+	1.321902e-02	1.279999e-02	-0.005027751	0.0132608706
-----	--------------	--------------	--------------	--------------

3.3.2 Comparison of MLEs of Parameter b_x across the Models

The MLEs of the Parameter \hat{b}_x for the models are presented in Table 2. Parameter \hat{b}_x is the relative pace of change in mortality by age. It describes the tendency of mortality at age x to change as the general level of mortality changes. \hat{b}_x values exhibited a fluctuating increase and decrease across all the models. The Renshaw-Haberman model differs a bit from other models, however closer ties exist between the Lee-Carter and the proposed model. Across the models, age-group 1-4 has the most fluctuant mortality pattern since it has the highest \hat{b}_x value. Negative \hat{b}_x value is observed in both the Lee-carter and Renshaw-Haberman models but absent in the Brouhns and Gamma-normal Lee-Carter models.

Table 3: Comparison of the MLEs of Parameter k_t under the four models

Year	LC	BR	RH	GNLC
2000	1.81327	1.94073	1.64128	1.8157
2001	1.78974	1.81796	1.45225	1.78923
2002	1.76584	1.70236	1.59232	1.76397
2003	1.61195	1.59096	1.63979	1.61029
2004	1.43708	1.35376	1.40494	1.43581
2005	1.0702	0.9866	1.44142	1.07015
2006	0.61382	0.54789	0.82363	0.6138
2007	0.02544	0.14287	0.11706	0.02643
2008	-0.2173	-0.1596	0.03022	-0.2171
2009	-0.5093	-0.5122	-0.5552	-0.5094
2010	-0.8826	-0.8051	-0.5975	-0.8834
2011	-1.0727	-1.0677	-0.7544	-1.0743
2012	-1.2852	-1.3192	-1.3155	-1.2858
2013	-1.4702	-1.5306	-1.4612	-1.4718
2014	-2.0633	-2.1295	-2.5458	-2.0634
2015	-2.6268	-2.5594	-2.9133	-2.6202

3.3.3 Comparison of MLEs of Parameter k_t across all the Models

The MLEs of the mortality trend of all the models are shown in Table 3. Parameter \hat{k}_t is known as the mortality index. According to [1], when \hat{k}_t is linear in time, mortality at each age changes at its own constant exponential rate. All the models exhibited a similarity in the mortality trend except for the Renshaw-Haberman model. The mortality time trend or index shows a gradual

downward trend from 2000 to 2015. A closer look at the curve reveals that the pattern is not a straight linear trend but curve-linear in nature.

3.3.4 Estimate of parameter α under the Gamma-normal Lee-Carter model

Parameter $\hat{\alpha}$ in the Gamma-normal Lee-Carter model denotes the shape parameter. It gives us an idea of the degree of asymmetry of the distribution. Its demographic representation implies additional effects that might exist and could act constantly across age and time in human mortality experience especially in developing Countries. $\hat{\alpha}$ was approximately 0.9722. The higher the value of $\hat{\alpha}$ for a mortality data, the higher the level of heterogeneity in the data. This parameter is not present in the Lee-Carter model and the variants considered in this study.

Table 5: Measures of Goodness of fit results for the simulated data

years = 5, ages = 15					
	$\log[L(\hat{\theta})]$	AIC	BIC	CAIC	HQC
GNLC	-71.5810	207.162	281.3216	257.4477	236.7731
LC	-71.5817	209.1634	285.6405	263.8951	239.6998
BR	-59782.7	119631.3	119707.8	119653.3	119746.6
years = 20, ages = 20					
	$\log[L(\hat{\theta})]$	AIC	BIC	CAIC	HQC
GNLC	-198.322	510.6444	738.1579	529.9777	600.7427
LC	-198.338	512.6758	744.1807	532.7462	604.3548
BR	-208805	417725.8	417957.3	417745.9	417817.5
years = 30, ages = 20					
	$\log[L(\hat{\theta})]$	AIC	BIC	CAIC	HQC
GNLC	-281.065	696.1302	990.7245	715.8727	810.8098
LC	-281.100	698.2004	997.1916	726.5509	814.5917
BR	-296181	592498	592797	592515.7	592614.4

3.4 Measures of Goodness of fit results for the simulated data

For the simulated data, the Akaike information criterion (AIC), the Bayesian Information Criterion (BIC) and the Corrected Akaike Information Criterion (CAIC) were lowest for the Gamma-normal Lee-Carter than other models. The Brouhns and the Renshaw-Haberman models had convergence problems. Furthermore, the results presented in table 5 show that the proposed Gamma-Normal Lee-Carter model could not work beyond 30 years. This brings to mind the central limit theorem.

4 Conclusion

In this study, the Nigerian female mortality was modelled using four stochastic mortality models. It was observed that generally, across the models, the younger ages (from age-group 10-14) have a lower mortality rate than the older ages. However, there is a clear indication that Infant mortality rate is higher than child mortality rate which is in turn greater than mortality rate at ages 5-14 years. Mortality was also increasing with age in the child-bearing age-groups (15-49). Across the models, age-group 1-4 (under 5) had the most fluctuant mortality pattern.

Furthermore, a Gamma-normal version of the Lee-Carter model was compared to some existing variants of the Lee-Carter model which include: Lee and Carter (1992) and Brouhns (2002) and Renshaw and Haberman (2006). The proposed Gamma-normal version was able to accommodate variability at different ages better than the other existing variants used. The Gamma-normal Lee-Carter model should be considered a better alternative to the referenced classical Lee-Carter models especially when it comes to modelling heterogeneous mortality data which is usually the case in some developing countries.

References

- ¹Lee, R. & Carter, L. 1992. Modeling and forecasting the time series of U.S. mortality. *Journal of the American Statistical Association* 87.419: 659-671.
- ²Booth, H., Maindonald, J. & Smith, L. 2002. Applying Lee-Carter under Conditions of Variable Mortality Decline. *Population Studies* 56.3: 325-336.
- ³Booth, H., Tickle, L. & Smith, L. 2005. Evaluation of the variants of the Lee-Carter method of forecasting mortality: A multi-Country comparison. *New Zealand Population Review* 31.1:13-34.
- ⁴Hanna, S. 2007. Applying the Lee-Carter model to countries in Eastern Europe and the former Soviet Union. *Journal of Applied Statistics*, 33.2: 255-272
- ⁵Koissi, M., & Shapiro, A. 2008. The Lee-carter model under the condition of variables age-specific parameters. *Presented at the 43rd Actuarial Research Conference, Regina, Canada August 2008.*
- ⁶Koissi M., Shapiro A. & Högnäs G. 2004. Fitting and Forecasting Mortality Rates for Nordic Countries Using the Lee-Carter method. *Actuarial Research Clearing House* 1:21
- ⁷Wang, H. & Preston, S. H., 2009. Forecasting United States mortality using cohort smoking histories. *Proceedings of the National Academy of Sciences*, 106.2: 393-398.
- ⁸Wang J. 2007. "Fitting and Forecasting Mortality for Sweden: Applying the Lee-Carter Model. *Mathematical Statistics, Stockholm University.*
<https://www2.math.su.se/matstat/reports/serieb/2007/rep1/report.pdf>.
- ⁹Chukwu, A. U. & Oladipupo, O. E. 2012. Modelling adult mortality in Nigeria: An Analysis based on the Lee-carter model. *Studies in Mathematical Sciences* 5.2: 1-11.
- ¹⁰Chukwu, A. U. & Adegbilero-Iwari, O. E. 2015. The performance of the Lee-Carter model on heterogeneous adult mortality data. *Mathematical Theory and Modeling* 5.4: 198-207
- ¹¹Taruvinga, R., Gachira, W., Chiwanza, W. & Nkomo, D. 2017. Comparison of the Lee-Carter and arch in modelling and forecasting mortality in Zimbabwe. *Asian Journal of Economic Modelling* 5.1: 11-22

- ¹² United Nations, Department of Economic and Social Affairs, Population Division 2019. *World Population Prospects 2019: Data Booklet* (ST/ESA/SER.A/424).
https://population.un.org/wpp/Publications/Files/WPP2019_DataBooklet.pdf
- ¹³ United Nations, Department of Economic and Social Affairs, Population Division 2017. *World Mortality 2017 – Data Booklet* (ST/ESA/SER.A/412).
<https://www.un.org/en/development/desa/population/publications/pdf/mortality/World-Mortality-2017-Data-Booklet.pdf>
- ¹⁴ SDG Indicators: Global indicator framework for the Sustainable Development Goals and targets of the 2030 Agenda for Sustainable Development
<https://unstats.un.org/sdgs/indicators/indicators-list/>
- ¹⁵ National Population Commission (NPC) [Nigeria] and ICF. 2019. 2018 Nigeria DHS Key Findings. Abuja, Nigeria and Rockville, Maryland, USA: NPC and ICF.
<https://dhsprogram.com/pubs/pdf/FR359/FR359.pdf>
- ¹⁶ World Health Organization and Unicef. 2015. Trends in maternal Mortality: 1990-2015: Estimates from WHO, UNICEF, UNFPA, World Bank Group and the United Nations Population Division.
<https://www.unfpa.org/publications/trends-maternal-mortality-1990-2015>
- ¹⁷ WHO. Sexual & Reproductive Health 2020.
<https://www.who.int/reproductivehealth/maternal-health-nigeria/en/>
- ¹⁸ Audu L.R. & Ekele B.A. 2002. A ten year review of maternal mortality in Sokoto Northern Nigeria. *West Afr. J med.* 2002. 21: 74-76.
- ¹⁹ Yar’zever S.I., 2014. Temporal Analysis of Maternal Mortality in Kano State, Northern Nigeria: A Six-Year Review. *American Journal of Public Health Research*, 2.2: 62-67.
- ²⁰ Brouhns, N., Denuit, M. & Vermunt J. 2002. A Poisson log-bilinear regression approach to the construction of projected life-tables. *Insurance: Mathematics and Economics* 31.3: 373–393.
- ²¹ Renshaw, A. E. & Haberman, S. 2006. A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insurance: Mathematics and Economics* 38.3: 556–570
- ²² Adegbilero-Iwari, O. E. and Chukwu, A. U. 2018. Gamma-normal version of the Lee- Carter model for forecasting mortality. *Mathematical Theory and Modelling*. 8.8:41-50.
- ²³ Zografos, K. & Balakrishnan, N. 2009. On families of Beta- and generalized Gamma-Generated distributions and associated inference. *Statistical Methodology* 6.4: 344-362.
- ²⁴ Lima, M., Cordeiro, G., & Ortega, E. 2015. A new extension of the normal distribution. *Journal of Data Science* 13.2: 385-408
- ²⁵ World Health Organisation. 2017. Global Health observatory.
<http://apps.who.int/gho/data/view.main.61200?lang=en>
- ²⁶ Haberman, S. & Butt, Z., 2010. Ilc: A Collection of R Functions for Fitting a Class of Lee-Carter Mortality Models using Iterative Fitting Algorithm. *Cass Actuarial Research Paper* 190. https://www.cass.city.ac.uk/data/assets/pdf_file/0019/37180/190ARP.pdf
- ²⁷ Villegas, A., Kaishev, V. K. & Millossovich, P. 2015. StMoMo: An R package for Stochastic mortality modelling. *In 7th Australasian Actuarial Education and Research Symposium*.
<https://cran.rproject.org/web/packages/StMoMo/vignettes/StMoMoVignette.pdf>.

²⁸Bolker, B. 2017. Maximum likelihood estimation and analysis with the bbmle package.
<https://cran.r-project.org/web/packages/bbmle/bbmle.pdf>

Appendix



Figure 1: Boxplot showing presence of outliers

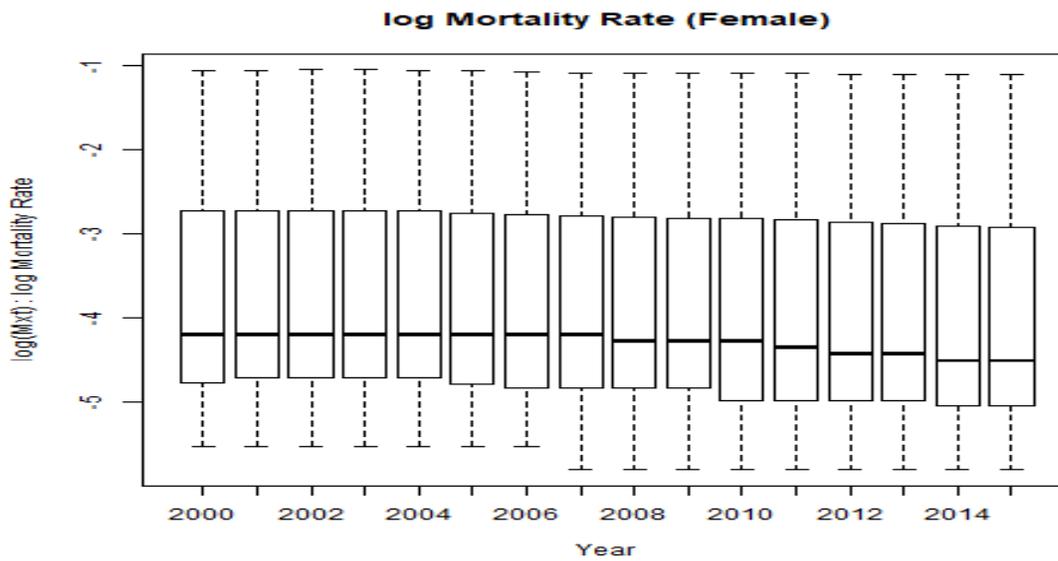


Figure 2: Boxplot showing absence of outliers

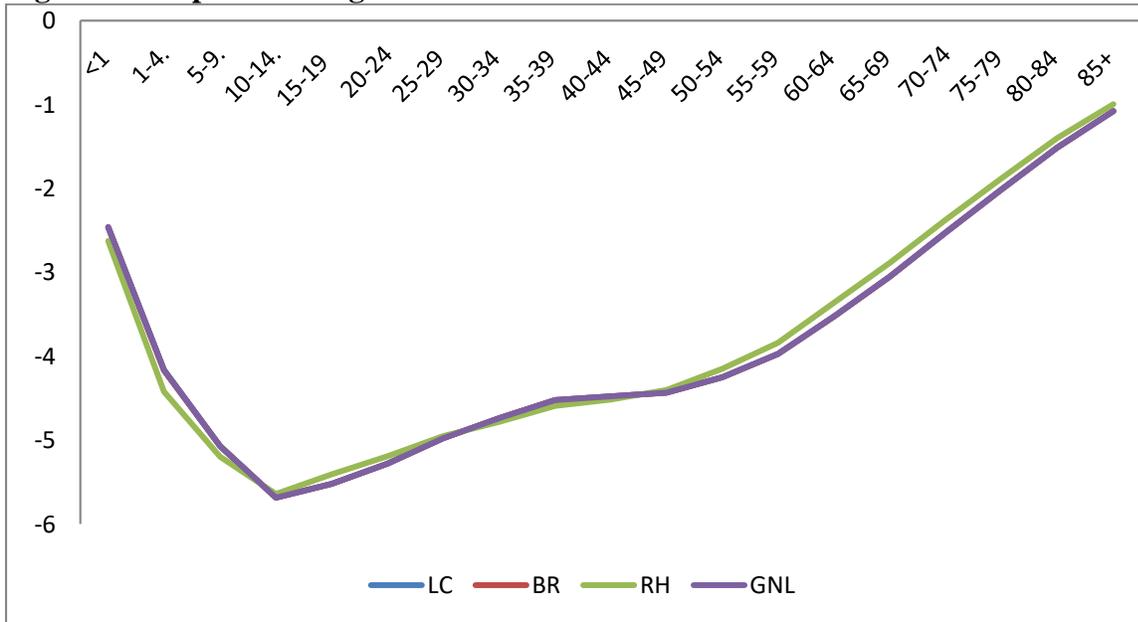


Figure 3: Parameter α_x across the models

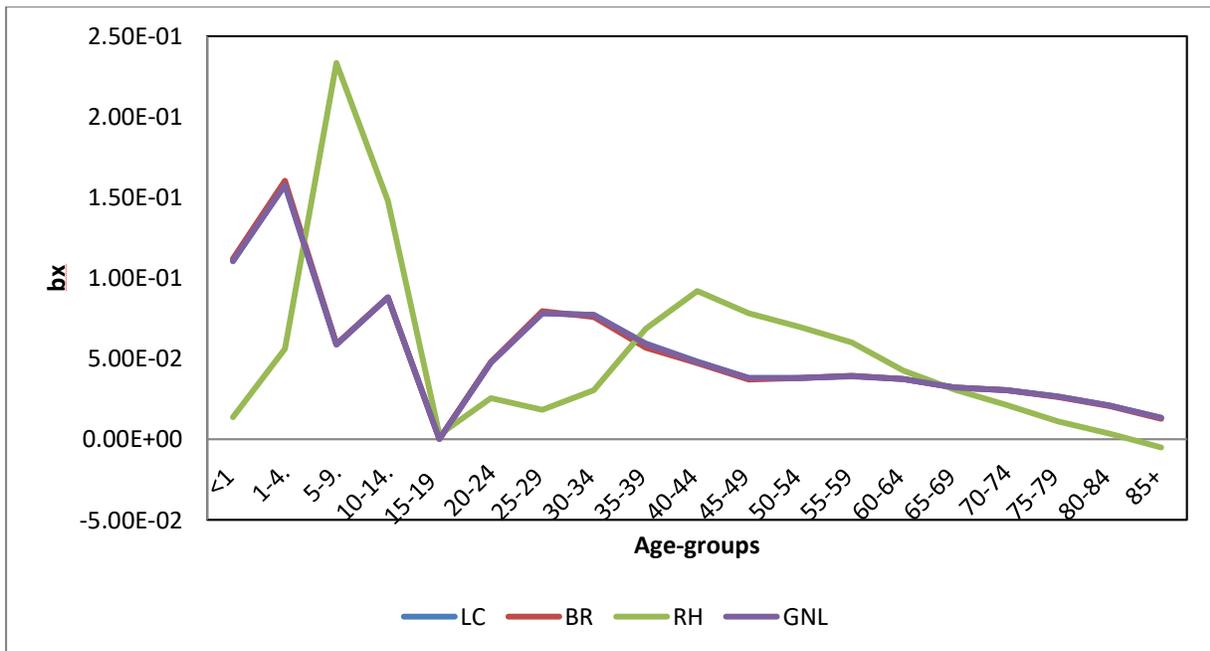


Figure 4: Parameter b_x across the models

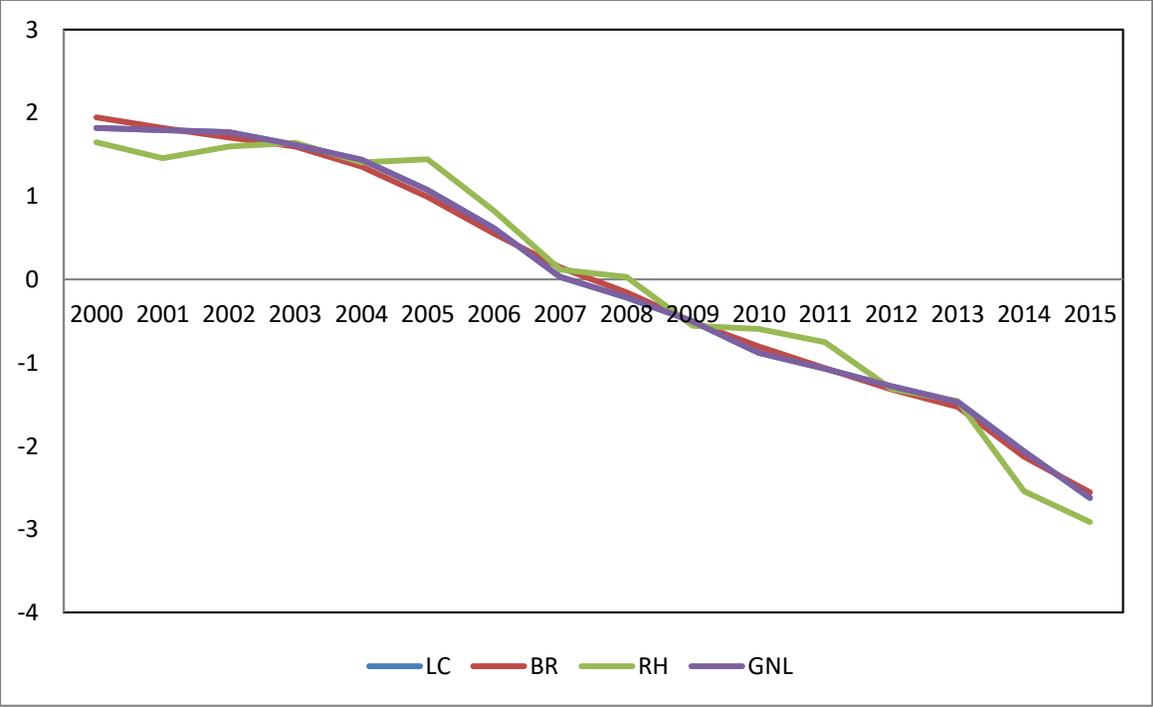


Figure 5: Parameter k_t across the models

Minimax Strategies for Bernoulli Two-Armed Bandit on a Moderate Control Horizon

Alexander Kolnogorov¹ and Denis Grunev²

¹ Yaroslav-the-Wise Novgorod State University, 41 B-St. Petersburgskaya Str, Velikiy Novgorod, 173003, Russia

(E-mail: Alexander.Kolnogorov@novsu.ru)

² Yaroslav-the-Wise Novgorod State University, 41 B-St. Petersburgskaya Str, Velikiy Novgorod, 173003, Russia

(E-mail: leonard_max@mail.ru)

Abstract. We consider a Bernoulli two-armed bandit problem on the moderate control horizon as applied to optimization of processing moderate amounts of data if there are two processing methods available with different a priori unknown efficiencies. One has to determine the most effective method and provide its predominant application. Unlike the big data processing for which several approaches, including batch processing, have been developed the optimization of moderate data processing is not good investigated now. We consider minimax approach and search for minimax strategy and minimax risk as Bayesian ones corresponding to the worst-case prior distribution for which Bayesian risk attains its maximal value. Close to the worst-case prior distribution and corresponding Bayesian risk are obtained by numerical methods. Calculations show that determined strategy provides the value of maximal regret close to determined Bayesian risk and, hence, is approximately minimax one. Results can be applied to big data processing if the data arise by batches of moderate size with approximately uniform properties.

Keywords: two-armed bandit problem, minimax and Bayesian approaches, main theorem of the game theory, moderate data processing.

1 Introduction

We consider Bernoulli two-armed bandit as applied to optimization of data processing. Bernoulli two-armed bandit (see, e.g., Berry and Fristedt[1]) is a controlled random process ξ_n , $n = 1, 2, \dots, N$, which values depend only on currently chosen actions y_n as follows

$$\Pr(\xi_n = 1|y_n = \ell) = p_\ell, \quad \Pr(\xi_n = 0|y_n = \ell) = q_\ell,$$

where $p_\ell + q_\ell = 1$, $\ell = 1, 2$. So, Bernoulli two armed bandit is described by a vector parameter $\theta = (p_1, p_2)$. Probabilities p_1, p_2 are assumed to be fixed but a priori unknown. However, the set Θ of possible values of parameter is assumed to be known. The values $\xi_n = 1$ and $\xi_n = 0$ of the process are interpreted as successful and unsuccessful processing of data item number n and are often considered as currently obtained income. We assume that the total number of

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST

processed data N has a priori known moderate value. It is usually interpreted as control horizon.

A control strategy σ determines the choice of the action y_{n+1} depending on the available history of the process, i.e. cumulative numbers n_1, n_2 of both actions applications ($n_1 + n_2 = n$) and corresponding cumulative incomes X_1, X_2 . Hence, strategy σ can be described by a finite array of probabilities $\sigma_\ell(X_1, n_1, X_2, n_2) = \Pr(y_{n+1} = \ell | X_1, n_1, X_2, n_2)$ where $X_1 = 0, \dots, n_1$, $X_2 = 0, \dots, n_2$, $n_1 + n_2 = n$, $n = 1, \dots, N$, $\ell = 1, 2$. Let's denote by $K(N)$ the total number of independently assigned probabilities $\{\sigma_\ell(X_1, n_1, X_2, n_2)\}$.

Let's describe the goal of the control. If one knew the parameter $\theta = (p_1, p_2)$ he should always choose the action corresponding to the maximum of p_1, p_2 , his total expected income is thus equal to $N \max(p_1, p_2)$. The application of the strategy σ provides the expected income which is less than maximal one by the value

$$L_N(\sigma, \theta) = N \max(p_1, p_2) - \mathbb{E}_{\sigma, \theta} \left(\sum_{n=1}^N \xi_n \right), \quad (1)$$

called the regret. Here $\mathbb{E}_{\sigma, \theta}$ denotes mathematical expectation calculated over the measure generated by strategy σ and parameter θ . The minimax risk is defined as

$$R_N^M(\Theta) = \inf_{\{\sigma\}} \sup_{\Theta} L_N(\sigma, \theta). \quad (2)$$

Minimax approach to the two-armed bandit problem was proposed in Robbins[2]. It was shown in Fabius and van Zwet[3] that it is virtually impossible to find explicit minimax strategies already for $N \geq 5$. On the other hand, the asymptotic minimax theorem was proved in Vogel[4] which states that $R_N(\Theta)$ has the order $N^{1/2}$ as $N \rightarrow \infty$. Then some additional approaches were proposed which provide the order $N^{1/2}$ of maximal regret, e.g. upper confidence bound rule (see Auer[5]) and mirror descent algorithm (see Juditsky *et al.*[6]). These results, as applied to data processing, can be interpreted as describing big data processing. In this connection, we also note Perchet *et al.*[7] and Kolmogorov[8], Kolmogorov[10] where batch data processing is considered and asymptotically optimal estimates of the minimax risk are established.

The structure of the paper is the following. In section 2 we consider approximation of minimax risk and minimax strategy by their determination on the finite subset of parameters. According to the main theorem of the game theory they are searched for as Bayesian ones corresponding to the worst-case prior distribution on which Bayesian risk attains its maximum value. In section 3 recursive Bellman-type equation is given for finding Bayesian risk and Bayesian strategy. In section 4 we present recursive equation for finding the regret. Section 5 contains numerical results. Discussion is presented in section 6.

2 Approximation on the finite set of parameters

Let's consider a finite set of parameters $\{\theta_1, \dots, \theta_K\}$. Denote by

$$R_N^M(\theta_1, \dots, \theta_K) = \inf_{\{\sigma\}} \max_{k=1, \dots, K} L_N(\sigma, \theta_k)$$

minimax risk on the finite set $\{\theta_1, \dots, \theta_K\}$. Assume that Θ is a closed set. It was proved in Kolmogorov[9] that the following equality holds

$$R_N(\Theta) = \max_{\theta_1, \dots, \theta_K \in \Theta} R_N^M(\theta_1, \dots, \theta_K), \quad (3)$$

i.e. minimax risk (2) on the entire set Θ is equal to minimax risk on some its finite subset. For K the upper bound holds $K \leq K(N) + 1$, where $K(N)$ is the number of independently assigned probabilities of the strategy σ .

Determination of the minimax risk $R_N^M(\theta_1, \dots, \theta_K)$ can be done with the use of the main theorem of the game theory. Let

$$\lambda_k = \Pr(\theta = \theta_k), \quad k = 1, \dots, K,$$

denote a prior distribution on the set of parameters $\theta_1, \dots, \theta_K$ where $\theta_k = (p_{1,k}, p_{2,k})$. Obviously, $\lambda_k \geq 0$, $k = 1, \dots, K$, $\sum_{k=1}^K \lambda_k = 1$. Denote by

$$R_N^B(\lambda_1, \dots, \lambda_K) = \min_{\{\sigma\}} \sum_{k=1}^K L_N(\sigma, \theta_k) \lambda_k \quad (4)$$

Bayesian risk calculated with respect to the prior distribution λ_k , $k = 1, \dots, K$. Since regret (1) is a continuous function of σ , θ and $\{\sigma\}$, $\{\theta_1, \dots, \theta_K\}$ are compact sets then, according to the main theorem of the game theory, the equality holds

$$R_N^M(\theta_1, \dots, \theta_K) = R_N^B(\lambda_1^0, \dots, \lambda_K^0) = \max_{\lambda_1, \dots, \lambda_K} R_N^B(\lambda_1, \dots, \lambda_K), \quad (5)$$

where $\lambda_1^0, \dots, \lambda_K^0$ is the worst-case prior distribution. Note that $R_N^B(\lambda_1, \dots, \lambda_K)$ is a concave function of $\lambda_1, \dots, \lambda_K$. Hence, determination of $R_N^M(\theta_1, \dots, \theta_K)$ by formula (5) is not laborious.

3 Calculation of Bayesian risk on the finite set of parameters

Calculation of the Bayesian risk $R_N^B(\lambda_1, \dots, \lambda_K)$ can be done with the use of the standard recursive Bellman-type dynamic programming equation. Given a history of the process (X_1, n_1, X_2, n_2) , let's denote by $\lambda_k(X_1, n_1, X_2, n_2) = \Pr(\theta = \theta_k | X_1, n_1, X_2, n_2)$, $k = 1, \dots, K$ the posterior probability distribution. It can be determined as

$$\lambda_k(X_1, n_1, X_2, n_2) = \frac{B(X_1, n_1 | p_{1,k}) B(X_2, n_2 | p_{2,k}) \lambda_k}{P(X_1, n_1, X_2, n_2)}, \quad (6)$$

$k = 1, \dots, K$, where

$$P(X_1, n_1, X_2, n_2) = \sum_{k=1}^K B(X_1, n_1 | p_{1,k}) B(X_2, n_2 | p_{2,k}) \lambda_k$$

and

$$B(X, n|p) = \binom{n}{X} p^X (1-p)^{n-X}.$$

If additionally to apply $B(X, n|p) = 1$ if $n = 0, X = 0$ then (6) holds if $n_1 = 0$ and/or $n_2 = 0$ as well.

Denote by $R_{N-n}(X_1, n_1, X_2, n_2)$ Bayesian risk on the latter $(N-n)$ steps calculated with respect to the posterior distribution (6). Denote $x^+ = \max(x, 0)$. To determine Bayesian risk (4) one has to solve the following standard recursive dynamic programming equation

$$\begin{aligned} & R_{N-n}(X_1, n_1, X_2, n_2) \\ &= \min(R_{N-n}^{(1)}(X_1, n_1, X_2, n_2), R_{N-n}^{(2)}(X_1, n_1, X_2, n_2)), \end{aligned} \quad (7)$$

where $R_{N-n}^{(1)}(X_1, n_1, X_2, n_2) = R_{N-n}^{(2)}(X_1, n_1, X_2, n_2) = 0$ if $n = N$ and

$$\begin{aligned} R_{N-n}^{(1)}(X_1, n_1, X_2, n_2) &= \sum_{k=1}^K \lambda_k(X_1, n_1, X_2, n_2) \\ &\times \left((p_{2,k} - p_{1,k})^+ + E_x^{(1,k)} R_{N-n-1}(X_1 + x, n_1 + 1, X_2, n_2) \right), \\ R_{N-n}^{(2)}(X_1, n_1, X_2, n_2) &= \sum_{k=1}^K \lambda_k(X_1, n_1, X_2, n_2) \\ &\times \left((p_{1,k} - p_{2,k})^+ + E_x^{(2,k)} R_{N-n-1}(X_1, n_1, X_2 + x, n_2 + 1) \right), \end{aligned} \quad (8)$$

if $n < N$. Here $E_x^{(\ell,k)} R(x) = q_{\ell,k} R(0) + p_{\ell,k} R(1)$, $q_{\ell,k} = 1 - p_{\ell,k}$.

In equations (6)–(7) risk $R_{N-n}^{(\ell)}(\cdot)$ is equal to expected cumulative regret on the residual control horizon $N-n$ if at first the ℓ -th action was chosen and then the control was optimally implemented ($\ell = 1, 2$). Bayesian strategy prescribes to choose the action corresponding to currently smaller value $R_{N-n}^{B(\ell)}(\cdot)$, $\ell = 1, 2$; in case of the draw the choice is arbitrary. Bayesian risk (4) is equal to

$$R_N^B(\lambda_1, \dots, \lambda_K) = R_N(0, 0, 0, 0). \quad (9)$$

Lets put $r(X_1, n_1, X_2, n_2) = R_{N-n}(X_1, n_1, X_2, n_2)P(X_1, n_1, X_2, n_2)$ and $r^{(\ell)}(X_1, n_1, X_2, n_2) = R_{N-n}^{(\ell)}(X_1, n_1, X_2, n_2)P(X_1, n_1, X_2, n_2)$, $\ell = 1, 2$. Then equation (7)–(8) takes the form

$$\begin{aligned} & r(X_1, n_1, X_2, n_2) \\ &= \min(r^{(1)}(X_1, n_1, X_2, n_2), r^{(2)}(X_1, n_1, X_2, n_2)), \end{aligned} \quad (10)$$

where $r^{(1)}(X_1, n_1, X_2, n_2) = r^{(2)}(X_1, n_1, X_2, n_2) = 0$ if $n = N$ and

$$\begin{aligned} & r^{(1)}(X_1, n_1, X_2, n_2) \\ &= g^{(1)}(X_1, n_1, X_2, n_2) + r(X_1, n_1 + 1, X_2, n_2)h_0(X_1, n_1) \\ &\quad + r(X_1 + 1, n_1 + 1, X_2, n_2)h_1(X_1, n_1), \\ & r^{(2)}(X_1, n_1, X_2, n_2) \\ &= g^{(2)}(X_1, n_1, X_2, n_2) + r(X_1, n_1, X_2, n_2 + 1)h_0(X_2, n_2) \\ &\quad + r(X_1, n_1, X_2 + 1, n_2 + 1)h_1(X_2, n_2) \end{aligned} \quad (11)$$

if $n < N$. Here $h_0(X_\ell, n_\ell) = (n_\ell + 1 - X_\ell)/(n_\ell + 1)$, $h_1(X_\ell, n_\ell) = (X_\ell + 1)/(n_\ell + 1)$, $\ell = 1, 2$,

$$g^{(1)}(X_1, n_1, X_2, n_2) = \sum_{k=1}^K (p_{2,k} - p_{1,k})^+ B(X_1, n_1 | p_{1,k}) B(X_2, n_2 | p_{2,k}) \lambda_k,$$

$$g^{(2)}(X_1, n_1, X_2, n_2) = \sum_{k=1}^K (p_{1,k} - p_{2,k})^+ B(X_1, n_1 | p_{1,k}) B(X_2, n_2 | p_{2,k}) \lambda_k.$$

Bayesian strategy prescribes to choose the action corresponding to currently smaller value $r^{(\ell)}(X_1, n_1, X_2, n_2)$, $\ell = 1, 2$; in case of the draw the choice is arbitrary. Bayesian risk (4) is equal to

$$R_N^B(\lambda_1, \dots, \lambda_K) = r(0, 0, 0, 0). \quad (12)$$

Equation (10)–(11) follows from (7)–(8) and formula (12) follows from (9). We need only verify expressions for $h_x(X_\ell, n_\ell)$, ($x \in \{0, 1\}$). It is sufficient to consider $h_x(X_1, n_1)$. In this case

$$\begin{aligned} h(X_1, n_1, x) &= \frac{\iint_{\Theta} p_1^x q_1^{1-x} B(X_1, n_1 | p_1) B(X_2, n_2 | p_2) \lambda(p_1, p_2) dp_1 dp_2}{P(X_1 + x, n_1 + 1, X_2, n_2)} \\ &= \frac{\iint_{\Theta} p_1^x q_1^{1-x} B(X_1, n_1 | p_1) B(X_2, n_2 | p_2) \lambda(p_1, p_2) dp_1 dp_2}{\iint_{\Theta} B(X_1 + x, n_1 + 1 | p_1) B(X_2, n_2 | p_2) \lambda(p_1, p_2) dp_1 dp_2} \\ &= \frac{p_1^x q_1^{1-x} B(X_1, n_1 | p_1)}{B(X_1 + x, n_1 + 1 | p_1)} = \frac{\binom{n_1}{X_1}}{\binom{n_1 + 1}{X_1 + x}}. \end{aligned}$$

One can verify straightforwardly that this corresponds to expressions for $h_x(X_\ell, n_\ell)$, ($x \in \{0, 1\}$).

Formulas (10)–(12) turned out to be more convenient for numerical experiments.

4 Calculation of the regret

Similarly one can calculate a regret corresponding to the strategy σ and prior distribution λ_k , $k = 1, \dots, K$ which is defined as

$$L_N(\sigma, \lambda) = \sum_{k=1}^K L_N(\sigma, \theta_k) \lambda_k, \quad (13)$$

where $L_N(\sigma, \theta)$ is defined in (1). Given strategy $\sigma_\ell(X_1, n_1, X_2, n_2) = \Pr(y_{n+1} = \ell | X_1, n_1, X_2, n_2)$, one has to solve the following recursive equation

$$\begin{aligned} l(X_1, n_1, X_2, n_2) &= \sigma_1(X_1, n_1, X_2, n_2) \times l^{(1)}(X_1, n_1, X_2, n_2) \\ &\quad + \sigma_2(X_1, n_1, X_2, n_2) \times l^{(2)}(X_1, n_1, X_2, n_2), \end{aligned} \quad (14)$$

where $l^{(1)}(X_1, n_1, X_2, n_2) = l^{(2)}(X_1, n_1, X_2, n_2) = 0$ if $n = N$ and

$$\begin{aligned}
& l^{(1)}(X_1, n_1, X_2, n_2) \\
&= g^{(1)}(X_1, n_1, X_2, n_2) + l(X_1, n_1 + 1, X_2, n_2)h_0(X_1, n_1) \\
&\quad + l(X_1 + 1, n_1 + 1, X_2, n_2)h_1(X_1, n_1), \\
& l^{(2)}(X_1, n_1, X_2, n_2) \\
&= g^{(2)}(X_1, n_1, X_2, n_2) + l(X_1, n_1, X_2, n_2 + 1)h_0(X_2, n_2) \\
&\quad + l(X_1, n_1, X_2 + 1, n_2 + 1)h_1(X_2, n_2)
\end{aligned} \tag{15}$$

if $n < N$. Regret (13) is then equal to

$$L_N^B(\sigma, \lambda) = l(0, 0, 0, 0). \tag{16}$$

Notice that if one has to calculate regret (1) he should choose the prior distribution concentrated at the single parameter θ .

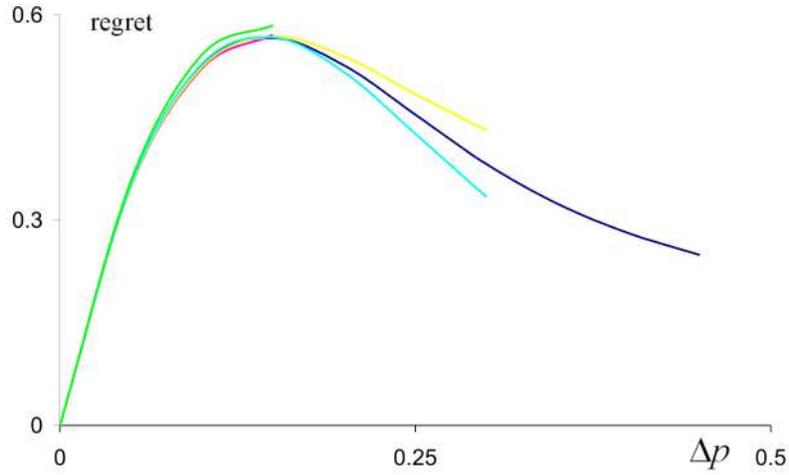


Fig. 1. Regret as a function of Δp for $N = 20$.

5 Numerical results

In what follows we consider the set of parameters $\Theta = \{(p_1, p_2) : 0 \leq p_\ell \leq 1, \ell = 1, 2\}$ which describes all Bernoulli two-armed bandits. Since this set is symmetric, i.e. contains both parameters (p_1, p_2) and (p_2, p_1) for all p_1, p_2 , the worst-case prior distribution is symmetric as well (see Fabius and van Zwet[3], Kolmogorov[9]). Let's put $K = 2M$ and take the prior distribution as follows:

$$p_{1,i} = p_i + \Delta p_i, \quad p_{2,i} = p_i - \Delta p_i, \quad p_{1,M+i} = p_i - \Delta p_i, \quad p_{2,M+i} = p_i + \Delta p_i,$$

$$\Pr\{(p_{1,i}, p_{2,i})\} = \Pr\{(p_{1,M+i}, p_{2,M+i})\} = \lambda_i, \quad i = 1, \dots, M,$$

where $0 < p_1 < \dots < p_M < 1$, $0 < \Delta p_i \leq \min(p_i, 1 - p_i)$, $\lambda_i \geq 0$, $i = 1, \dots, M$, $\lambda_1 + \dots + \lambda_M = 0.5$. Obviously, this prior distribution is symmetric. It is completely described by the values p_1, \dots, p_M , $\Delta p_1, \dots, \Delta p_M$, $\lambda_1, \dots, \lambda_M$ and Bayesian risk (12) calculated with respect to this prior can be considered as a function of finite number variables p_1, \dots, p_M , $\Delta p_1, \dots, \Delta p_M$, $\lambda_1, \dots, \lambda_M$. To minimize this function, standard numerical techniques can be used. We used gradient method. Numerical calculations were implemented for

i	1	2	3	4	5	6	7
p_i	0.214	0.312	0.373	0.436	0.503	0.560	0.640
Δp_i	0.154	0.149	0.147	0.145	0.144	0.143	0.143
λ_i	0.115	0.165	0.164	0.147	0.134	0.125	0.150

Table 1. Prior distribution for $N = 20$

$N = 20, 30, 40, 50$. Approximately the worst-case prior distributions that were determined using gradient method are presented in Tables 1–4. We also calculate corresponding normalized Bayesian risks $r_N^B(\lambda) = (DN)^{-1/2} R_N^B(\lambda)$ where $D = 0.25$ is the maximum variance of one-step reward and $R_N^B(\lambda)$ are calculated according to (10)–(12). Normalized Bayesian risks are as follows $r_{20}^B(\lambda) = 0.568$, $r_{30}^B(\lambda) = 0.572$, $r_{40}^B(\lambda) = 0.576$ and $r_{50}^B(\lambda) = 0.580$. Then determined strat-

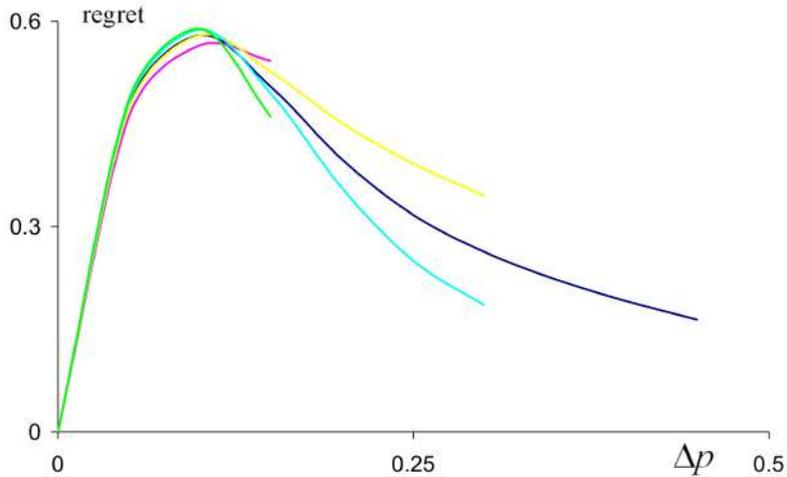


Fig. 2. Regret as a function of Δp for $N = 50$.

egy was used to calculate the normalized regret $l_N^B(\sigma, \lambda) = (DN)^{-1/2} L_N^B(\sigma, \lambda)$, where $L_N^B(\sigma, \lambda)$ is calculated according to (14)–(16) as a function of $(p, \Delta p)$. On Figure 1 normalized regrets are presented as functions of Δp for fixed p by the following lines: for $p = 0.2$ – Magenta, for $p = 0.35$ – Yellow, for $p = 0.5$

– Blue, for $p = 0.65$ – Cyan and for $p = 0.8$ – green. One can see that all the lines have approximately equal maxima. These maxima are as follows: 0.585 for $N = 20$, 0.559 for $N = 30$, 0.577 for $N = 40$ and 0.591 for $N = 50$. One can see that these maxima are close to determined normalized Bayesian risks, therefore determined strategies are close to minimax ones.

i	1	2	3	4	5	6	7
p_i	0.215	0.326	0.395	0.460	0.537	0.632	0.761
Δp_i	0.128	0.124	0.122	0.121	0.119	0.118	0.114
λ_i	0.143	0.173	0.192	0.190	0.161	0.106	0.034

Table 2. Prior distribution for $N = 30$

i	1	2	3	4	5	6	7
p_i	0.215	0.325	0.388	0.450	0.525	0.615	0.737
Δp_i	0.115	0.109	0.107	0.106	0.105	0.103	0.101
λ_i	0.110	0.163	0.179	0.181	0.173	0.145	0.049

Table 3. Prior distribution for $N = 40$

Notice that the shapes of the lines corresponding to regrets for $N = 30, 40, 50$ are very similar to presented on Figure 1. So, we include here additionally only Figure 2 which demonstrates the regrets corresponding to $N = 50$.

i	1	2	3	4	5	6	7
p_i	0.215	0.319	0.391	0.466	0.543	0.647	0.784
Δp_i	0.103	0.098	0.097	0.095	0.094	0.093	0.090
λ_i	0.123	0.174	0.205	0.204	0.171	0.103	0.020

Table 4. Prior distribution for $N = 50$

6 Conclusion

The proposed approach can be used for storage of the parameters of the minimax strategy for different moderate control horizons. One possible method is to store all independently assigned probabilities $\{\sigma_\ell(X_1, n_1, X_2, n_2)\}$. The number $K(N)$ of these probabilities is equal to the number of histories $\{(X_1, n_1, X_2, n_2)\}$. Given n_1, n_2 , the number of histories is $(n_1 + 1) \times (n_2 + 1)$. Hence, given n ($n = n_1 + n_2$), the number of histories is

$$\sum_{n_1=0}^n (n_1 + 1) \times (n - n_1 + 1).$$

Finally, given control horizon N , the total number of histories is equal to

$$K(N) = \sum_{n=0}^{N-1} \sum_{n_1=0}^n (n_1 + 1) \times (n - n_1 + 1).$$

For example, $K(20) = 8855$, $K(50) = 292825$. So, $K(N)$ is a significant value and the storage of independently assigned probabilities $\{\sigma_\ell(X_1, n_1, X_2, n_2)\}$ for multiple control horizons N requires significant computer memory. But more important is that possible errors in some probabilities $\{\sigma_\ell(X_1, n_1, X_2, n_2)\}$ may cause large regrets.

Another approach is to keep in computer memory the worst-case prior distributions similar to presented in Tables 1–4. Determination of corresponding Bayesian strategy is very fast. Once the batch of data of the size N arises, computer provides close to optimal strategy for their processing.

This approach may be developed for the case when the number N of arising data is not known exactly a priori. Corresponding strategies should provide uniform minimization of the normalized regret $(DN)^{-1/2}L_N(\sigma, \theta)$ on the range of control horizons $N_1 \leq N \leq N_2$.

7 Acknowledgments

The reported study was funded by RFBR, project number 20-01-00062.

References

1. D. A. Berry and B. Fristedt. *Bandit Problems: Sequential Allocation of Experiments*, Chapman and Hall, London, New York, 1985.
2. H. Robbins. Some Aspects of the Sequential Design of Experiments. *Bull. Amer. Math. Soc.*, 58, **5**, 527-535, 1952.
3. J. Fabius and W. R. van Zwet. Some remarks on the two-armed bandit. *Ann. Math. Statist.*, 41, 1906-1916, 1970.
4. W. Vogel. An asymptotic minimax theorem for the two-armed bandit problem. *Ann. Math. Stat.*, 31, 444-451, 1960.
5. P. Auer. Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine Learning Research*. 3, 397–422, 2002.
6. A. Juditsky, A. V. Nazin, A. B. Tsybakov and N. Vayatis. Gap-Free Bounds for Stochastic Multi-Armed Bandit. *Proc. 17th World Congress IFAC*. (Seoul, Korea, July 6–11), 11560–11563, 2008.
7. V. Perchet, P. Rigollet, S. Chassang and S. Snowberg. Batched Bandit Problems. *Annals of Statistics*. 44, **2**, 660–681, 2016.
8. A. V. Kolmogorov. Two-armed bandit problem for parallel data processing systems. *Problems of Information Transmission*. 48, **1**, 72–84, 2012.
9. A. V. Kolmogorov. Finding minimax strategy and minimax risk for Bernoulli multi-armed bandit. *Proceedings of the 2014 International Conference on Mathematical Models and Methods in Applied Sciences (MMAS '14)*. (Saint Petersburg, Russia, September 23–25), 59–66, 2014.
10. A. V. Kolmogorov. Gaussian two-armed bandit and optimization of batch data processing. *Problems of Information Transmission* 54, **1**, 84–100, 2018.

A model of random walk with varying transition probabilities

Tomáš Kouřim¹ and Petr Volf²

¹ Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Czech Republic

(E-mail: kourim@outlook.com)

² Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Prague

(E-mail: volf@utia.cas.cz)

Abstract This paper considers a model of one-dimensional discrete time random walk in which the position of the walker is controlled by varying transition probabilities. These probabilities depend explicitly on the previous move of the walker and implicitly on the entire walk history. Hence, transition probabilities evolve in time making the walk a non-Markovian stochastic process. The paper follows on the recent work of the authors. Two basic versions of the model are introduced, some of their properties are recalled and new theoretical results derived. Then, more complex variants of models are presented. Development of walks themselves as well as the properties of connected sequences of transition probabilities are illustrated also with the aid of simulations. Possible applications of the model in real life situations are discussed and briefly described, too.

Keywords: Random walk, history dependent transition probabilities, success punishing/rewarding walk.

1 Introduction

Stochastic processes and the corresponding mathematical theory represent a significant part of mathematics. One of the most prominent of such processes is the random walk, introduced by Pearson over hundred years ago [6]. This concept has been then further elaborated by many authors creating a number of different versions of a random walk [7] and there are still new possibilities and options how the classical random walk can be altered and adapted to specific application field. The model discussed in this paper follows on the work of Turban [8] and represents yet another version of a random walk, walk with varying transition probabilities. The model falls within a rather broad class of processes presented in recent work of Davis and Liu [1], but not all assumptions from [1] are met.

The original inspiration for the model comes from one of its applications – modeling of sports events. Many types of sport, such as tennis or volleyball, are played in a strictly discrete manner with steps divided by individual *points*,

6th SMTDA Conference Proceedings, 2–5 June 2020, Barcelona, Spain

© 2020 ISAST



games or *sets*. One sport match can be thus viewed as a random walk with individual parts of the match representing the steps of the walk. Success, i.e. scoring a point or winning a set, then significantly affects further development of the entire walk by changing the transition probability. Other real life situations with similar properties can be found everywhere in areas where both “successes” or “failures” occur. In fact, also discrete time recurrent counts data occurrence can be often modeled in a similar way, when the event probability is affected by the recent history of the process. Such cases include the recurrence of diseases, recidivism in crime or repeated defects and maintenance of a technical device.

The present contribution continues on the recent authors’ exploration of the model of a random walk with varying probabilities. Selected properties of the model are presented and possible real life implementations of the model are discussed. The rest of the paper is organized as follows. Next section introduces theoretical properties of the model and describes in detail the two main variants of the model. Section 3 presents possible applications of the model and the last section concludes this work.

2 Theoretical properties

As already mentioned, modeling sport events served as a motivation for the presented model. The probability of a success (i.e. scoring a goal, achieving a point etc.) is at the center of interest in such modeling. After each occurrence of such success its probability either decreases or increases, and thus two basic model alternatives exist – *success punishing* and *success rewarding*. The basic version of the model operates with starting success probability p_0 and a memory coefficient λ affecting the severity of probability change after a success as input parameters. Formally the walk is defined as follows:

Definition 1. Let $p_0 \in (0, 1)$, $\lambda \in (0, 1)$ be constant parameters, $\{P_n\}_{n=0}^\infty$ and $\{X_n\}_{n=1}^\infty$ sequences of discrete random variables, $X_t \in \{-1, 1\}$ and $P_t \in (0, 1)$ for each t , and $P_0 = p_0$. For $t \geq 1$ let

$$P(X_t = 1|P_{t-1} = p_{t-1}) = p_{t-1}, \quad P(X_t = -1|P_{t-1} = p_{t-1}) = 1 - p_{t-1},$$

and (*success punishing*)

$$P_t = \lambda P_{t-1} + \frac{1}{2}(1 - \lambda)(1 - X_t) \tag{1}$$

or (*success rewarding*)

$$P_t = \lambda P_{t-1} + \frac{1}{2}(1 - \lambda)(1 + X_t). \tag{2}$$

The sequence $\{S_n\}_{n=0}^\infty$, $S_n = S_0 + \sum_{i=1}^n X_i$ for $n \in \mathbb{N}$, with $S_0 \in \mathbb{R}$ some given starting position, is called a *random walk with varying probabilities*, with $\{X_n\}_{n=1}^\infty$ being the steps of the walker and $\{P_n\}_{n=0}^\infty$ transition probabilities. Depending on the chosen formula to calculate P_t the walk type is either *success punishing* (1) or *success rewarding* (2).

The model was first introduced in [2] and a more thorough description was provided in [4]. Selected properties were then presented in [5] and a practical implementation of the model in modeling tennis matches was presented in [3]. The basic results are recalled in the following sections and then variance of S_t is derived and described in more detail.

2.1 Success punishing model

The basic properties of the *success punishing* version of the walk are presented in this section. Previous results are presented as a set of expressions only, the reader is kindly asked to see referred papers for full proves of those expressions. Newly described properties are then provided with full proves and all necessary details.

For the expected value and variance of the step size for the $t \geq 1$ iteration of the walk X_t it holds [5]

$$EX_t = (2\lambda - 1)^{t-1}(2p_0 - 1), \quad (3)$$

$$Var X_t = 1 - (2\lambda - 1)^{2(t-1)}(2p_0 - 1)^2. \quad (4)$$

For the expected value and variance of the transition probability for the $t \geq 1$ iteration of the walk P_t it holds [2, 5]

$$EP_t = (2\lambda - 1)^t p_0 + \frac{1 - (2\lambda - 1)^t}{2}, \quad (5)$$

$$Var P_t = (3\lambda^2 - 2\lambda)^t p_0^2 + \sum_{i=0}^{t-1} K(i; p_0, \lambda)(3\lambda^2 - 2\lambda)^{t-1-i} - k(t; p_0, \lambda)^2, \quad (6)$$

where

$$k(t; p_0, \lambda)^2 = EP_t = (2\lambda - 1)^t p_0 + \frac{1 - (2\lambda - 1)^t}{2}$$

and

$$K(t; p_0, \lambda)^2 = k(t; p_0, \lambda)^2 \cdot (-3\lambda^2 + 4\lambda - 1) + (1 - \lambda)^2.$$

Finally, the expected position of the walker S_t after $t \geq 1$ iterations can be expressed as [2]

$$ES_t = S_0 + (2p_0 - 1) \frac{1 - (2\lambda - 1)^t}{2(1 - \lambda)}. \quad (7)$$

The last formula missing is the one expressing the variance of the position of the walker $Var S_t$. Before it is presented, let us first prove a support proposition.

Proposition 1. *For all $t \geq 1$*

$$E(P_t S_t) = (2\lambda - 1)^t p_0 S_0 + \sum_{i=0}^{t-1} (2\lambda - 1)^i q(t - 1 - i; p_0, S_0, \lambda), \quad (8)$$

where

$$q(j; p_0, S_0, \lambda) = (1 - \lambda)s(j; p_0, S_0, \lambda) + 2\lambda\pi(j; p_0, \lambda) + (1 - 2\lambda)p(j; p_0, \lambda) + \lambda - 1$$

and $p(j; p_0, \lambda) = EP_j$ is given by (5), $s(j; p_0, S_0, \lambda) = ES_j$ by (7) and $\pi(j; p_0, \lambda) = EP_j^2$ is given as

$$EP_t^2 = (3\lambda^2 - 2\lambda)^t p_0^2 + \frac{1 - (3\lambda^2 - 2\lambda)^t (\lambda + 1)}{3\lambda + 1} - (p_0 - \frac{1}{2})(3\lambda^2 - 4\lambda + 1)M(t-1; \lambda), \quad (9)$$

where

$$M(t; \lambda) = \sum_{i=0}^{t-1} (3\lambda^2 - 2\lambda)^{t-1-i} (2\lambda - 1)^i.$$

Proof. The formula for EP_t^2 follows from the proof of Proposition 2.5 in [5]. To prove (8) let us start with expressing the value of $E(P_t S_t)$ from the knowledge of past steps as

$$\begin{aligned} E(P_t S_t) &= E[E(P_t S_t | P_{t-1})] = \\ &= E[E((\lambda P_{t-1} + \frac{1}{2}(1 - \lambda)(1 - X_t))(S_{t-1} + X_t) | P_{t-1})] = \\ &= E[E(\lambda P_{t-1} S_{t-1} + \frac{1 - \lambda}{2} S_{t-1} - \frac{1 - \lambda}{2} X_t S_{t-1} + \\ &\quad + \lambda X_t P_{t-1} + \frac{1 - \lambda}{2} X_t - \frac{1 - \lambda}{2} X_t^2 | P_{t-1})]. \end{aligned}$$

Using $E(X_t | P_{t-1}) = 2P_{t-1} - 1$ and $EX_t^2 = 1$ we get

$$E(P_t S_t) = (2\lambda - 1)E(P_{t-1} S_{t-1}) + (1 - \lambda)ES_{t-1} + 2\lambda EP_{t-1}^2 + (2\lambda - 1)EP_{t-1} + \lambda - 1. \quad (10)$$

Further we will continue using mathematical induction. For $t = 1$ using the definition of the walk it holds that

$$\begin{aligned} E(P_1 S_1) &= p_0(p_0 \lambda (S_0 + 1)) + (1 - p_0)[(1 - (1 - p_0)\lambda)(S_0 - 1)] = \\ &= (2\lambda - 1)p_0 S_0 + (1 - \lambda)S_0 + 2\lambda p_0^2 - (2\lambda - 1)p_0 + \lambda - 1. \end{aligned}$$

When substituting $t = 1$ into (8) we obtain

$$\begin{aligned} E(P_1 S_1) &= (2\lambda - 1)p_0 S_0 + \sum_{i=0}^0 (2\lambda - 1)^i q(0 - i; p_0, S_0, \lambda) = \\ &= (2\lambda - 1)p_0 S_0 + (1 - \lambda)s(0; p_0, \lambda) + \\ &\quad + 2\lambda\pi(0; p_0, \lambda) + (1 - 2\lambda)p(0; p_0, \lambda) + \lambda - 1 \end{aligned}$$

and finally using (5), (7) and (9)

$$E(P_1 S_1) = (2\lambda - 1)p_0 S_0 + (1 - \lambda)S_0 + 2\lambda p_0^2 + (1 - 2\lambda)p_0 + \lambda - 1.$$

Equation (8) thus holds for $t = 1$. Now for the induction step $t \rightarrow t + 1$, after substituting (8) into (10) we get $E(P_{t+1} S_{t+1}) =$

$$\begin{aligned} &= (2\lambda - 1)E(P_t S_t) + (1 - \lambda)E S_t + 2\lambda E P_t^2 + (2\lambda - 1)E P_t + \lambda - 1 = \\ &= (2\lambda - 1)[(2\lambda - 1)^t p_0 S_0 + \sum_{i=0}^{t-1} (2\lambda - 1)^i q(t - 1 - i)] + (1 - \lambda)s(t) + \\ &\quad + 2\lambda \pi(t) + (2\lambda - 1)p(t) + \lambda - 1 = \\ &= (2\lambda - 1)^{t+1} p_0 S_0 + \sum_{i=0}^t (2\lambda - 1)^i q(t - i). \end{aligned}$$

Theorem 1. For all $t \geq 1$

$$Var S_t = t + 4 \sum_{i=0}^{t-1} \sigma(i; p_0, 0, \lambda) - a(t; p_0, \lambda),$$

with $\sigma(i; p_0, S_0, \lambda) = E(P_t S_t)$ given by Proposition 1 and

$$a(t; p_0, \lambda) = (2p_0 - 1) \sum_{i=0}^{t-1} \frac{1 - (2\lambda - 1)^i}{1 - \lambda} + (2p_0 - 1)^2 \frac{(1 - (2\lambda - 1)^t)^2}{4(1 - \lambda)^2}.$$

Proof. As clearly the value S_0 does not affect the variance, let us from now assume $S_0 = 0$. From the definition of variance

$$Var S_t = E S_t^2 - E^2 S_t \tag{11}$$

and (7) follows that to prove the proposition it is enough to prove that

$$E S_t^2 = t + 4 \sum_{i=0}^{t-1} \sigma(i; p_0, 0, \lambda) - (2p_0 - 1) \sum_{i=0}^{t-1} \frac{1 - (2\lambda - 1)^i}{1 - \lambda}. \tag{12}$$

First of all, let us express $E S_t^2$ from the knowledge of past walk development. From the definition of the expected value and the definition of the walk it follows

$$\begin{aligned} E S_t^2 &= E[E(S_t^2 | P_{t-1})] = E[E((S_{t-1} + X_t)^2 | P_{t-1})] = \\ &= E(S_{t-1}^2 + 2(2P_{t-1} - 1)S_{t-1} + 1) = \\ &= E S_{t-1}^2 + 4E(P_{t-1} S_{t-1}) - 2E S_{t-1} + 1, \end{aligned} \tag{13}$$

where the fact that $E X_t^2 = 1$ for all t was used. The theorem will be proved using mathematical induction again. For $t = 1$ we get from the definition of the walk

$$E S_1^2 = p_0(S_0 + 1)^2 + (1 - p_0)(S_0 - 1)^2 = 1.$$

Substituting $t = 1$ into (12) we obtain

$$ES_1^2 = 1 + 4\sigma(0; p_0, 0, \lambda) - (2p_0 - 1) \frac{1 - (2\lambda - 1)^0}{1 - \lambda} = 1$$

and (12) thus holds for $t = 1$. Now for the induction step $t \rightarrow t + 1$ we get by substituting (12) into (13)

$$\begin{aligned} ES_{t+1}^2 &= ES_t^2 + 4E(P_t S_t) - 2ES_t + 1 = \\ &= t + 4 \sum_{i=0}^{t-1} \sigma(i; p_0, 0, \lambda) - (2p_0 - 1) \sum_{i=0}^{t-1} \frac{1 - (2\lambda - 1)^i}{1 - \lambda} + \\ &\quad + 4\sigma(t; p_0, 0, \lambda) - 2(2p_0 - 1) \frac{1 - (2\lambda - 1)^t}{2(1 - \lambda)} + 1 = \\ &= (t + 1) + 4 \sum_{i=0}^t \sigma(i; p_0, 0, \lambda) - (2p_0 - 1) \sum_{i=0}^t \frac{1 - (2\lambda - 1)^i}{1 - \lambda}, \end{aligned}$$

which proves (12). Substituting (12) and (7) into (11) then proves the theorem.

Corollary 1. For $t \rightarrow +\infty$

$$Var S_t = +\infty$$

and

$$\lim_{t \rightarrow +\infty} (Var S_t - (c_1(p_0, \lambda)t + c_2(p_0, \lambda))) = 0,$$

where $c_i(p_0, \lambda)$ are some t -independent constants.

Corollary 1 shows that with $t \rightarrow +\infty$ $Var S_t$ behaves as a linear function with respect to t . This can be seen also on Figure 1 together with a comparison of observed and theoretical values of $Var S_t$.

2.2 Success rewarding model

Similar formulas can be derived for the *success rewarding* model. Once again for previous results only the formulas are presented with proofs in the referred literature, new properties are derived with full complexity. For the sake of clarity the set of expressions is presented in the same manner as in the previous section.

For the expected value and variance of the step size for the $t \geq 1$ iteration of the walk X_t it holds [5]

$$EX_t = 2p_0 - 1, \tag{14}$$

$$Var X_t = 4p_0(1 - p_0). \tag{15}$$

For the expected value and variance of the transition probability for the $t \geq 1$ iteration of the walk P_t it holds [5]

$$EP_t = p_0, \tag{16}$$

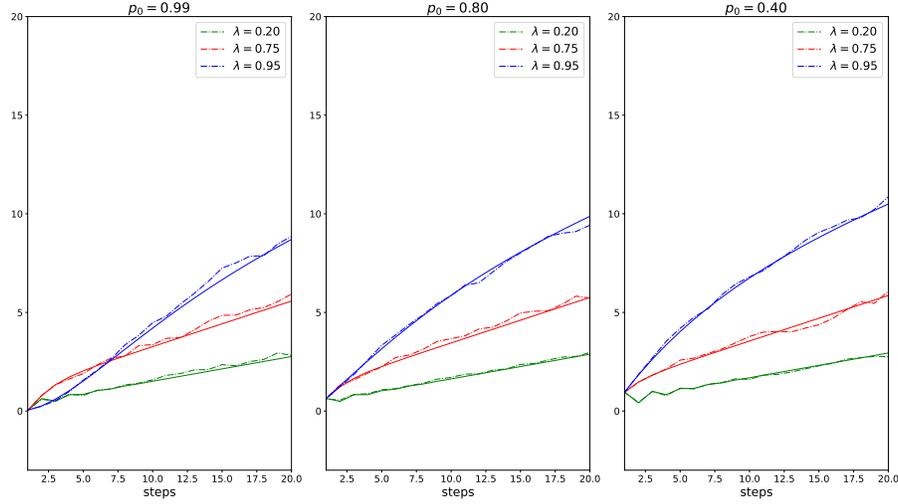


Figure 1. Observed (dash-dotted) and theoretical (solid lines) values of $Var S_t$ – *success punishing* model. The data were obtained from 1000 walks generated with given parameters.

$$Var P_t = (2\lambda - \lambda^2)^t p_0^2 + p_0(1 - \lambda)^2 \sum_{i=0}^{t-1} (2\lambda - \lambda^2)^i - p_0^2. \quad (17)$$

As the sum in the formula equals $\frac{1 - (2\lambda - \lambda^2)^t}{1 - (2\lambda - \lambda^2)}$, it can be further simplified as

$$\begin{aligned} Var P_t &= (2\lambda - \lambda^2)^t p_0^2 + p_0(1 - \lambda)^2 \frac{1 - (2\lambda - \lambda^2)^t}{(1 - \lambda)^2} - p_0^2 = \\ &= p_0[(2\lambda - \lambda^2)^t (p_0 - 1) + 1] - p_0^2, \\ Var P_t &= p_0(1 - p_0)(1 - (2\lambda - \lambda^2)^t). \end{aligned}$$

Finally, the expected position of the walker S_t after $t \geq 1$ iterations can be expressed as [5]

$$ES_t = S_0 + t(2p_0 - 1). \quad (18)$$

To prove a formula allowing to compute the variance of the position of the walker, let us again start with a support proposition.

Proposition 2. For all $t \geq 1$

$$E(P_t S_t) = p_0 S_0 + p_0 t + 2\lambda p_0 (p_0 - 1) \frac{1 - (2\lambda - \lambda^2)^t}{(1 - \lambda)^2}. \quad (19)$$

Proof. We will once again start with expressing $E(P_t S_t)$ from the knowledge of the past step.

$$\begin{aligned} E(P_t S_t) &= E(E(P_{t-1} S_{t-1} | P_{t-1})) = \\ &= E[E((\lambda P_{t-1} + \frac{1}{2}(1-\lambda)(1+X_t))(S_{t-1} + X_t) | P_{t-1})] = \\ &= E[E(\lambda P_{t-1} S_{t-1} + \frac{1-\lambda}{2} S_{t-1} + \frac{1-\lambda}{2} X_t S_{t-1} + \\ &\quad + \lambda X_t P_{t-1} + \frac{1-\lambda}{2} X_t + \frac{1-\lambda}{2} X_t^2 | P_{t-1})] \end{aligned}$$

and using $E(X_t | P_{t-1}) = 2P_{t-1} - 1$ and $E X_t^2 = 1$ finally

$$E(P_t S_t) = E(P_{t-1} S_{t-1}) + 2\lambda E P_{t-1}^2 - (2\lambda - 1) E P_{t-1}. \quad (20)$$

Further we will continue using mathematical induction. For $t = 1$ using the definition of the walk it holds that

$$\begin{aligned} E(P_1 S_1) &= p_0(1 - (1 - p_0)\lambda)(S_0 + 1) + (1 - p_0)\lambda p_0(S_0 - 1) = \\ &= p_0 S_0 + 2\lambda p_0^2 - (2\lambda - 1)p_0. \end{aligned}$$

When substituting $t = 1$ into (19) we obtain

$$\begin{aligned} E(P_1 S_1) &= p_0 S_0 + p_0 + 2\lambda p_0(p_0 - 1) \frac{1 - (2\lambda - \lambda^2)^0}{(1 - \lambda)^2} = \\ &= p_0 S_0 + p_0 + 2\lambda p_0(p_0 - 1) \end{aligned}$$

and finally

$$E(P_1 S_1) = p_0 S_0 + 2\lambda p_0^2 - (2\lambda - 1)p_0.$$

Equation (19) thus holds for $t = 1$. Now for the induction step $t \rightarrow t + 1$ we get by substituting (19) into (20)

$$E(P_{t+1} S_{t+1}) = E(P_t S_t) + 2\lambda E P_t^2 - (2\lambda - 1) E P_t$$

and further using

$$E P_t^2 = p_0((2\lambda - \lambda^2)^t(p_0 - 1) + 1),$$

which follows from the proof of Proposition 3.7 in [5], and (16)

$$\begin{aligned} E(P_{t+1} S_{t+1}) &= p_0 S_0 + p_0 t + 2\lambda p_0(p_0 - 1) \frac{1 - (2\lambda - \lambda^2)^t}{(1 - \lambda)^2} + \\ &\quad + 2\lambda p_0((2\lambda - \lambda^2)^t(p_0 - 1) + 1) - (2\lambda - 1)p_0 = \\ &= p_0 S_0 + p_0(t + 1) + 2\lambda p_0(p_0 - 1) \left[\frac{1 - (2\lambda - \lambda^2)^t}{(1 - \lambda)^2} + (2\lambda - \lambda^2)^t \right] = \\ &= p_0 S_0 + p_0(t + 1) + 2\lambda p_0(p_0 - 1) \frac{1 - (2\lambda - \lambda^2)^{t+1}}{(1 - \lambda)^2}. \end{aligned}$$

Theorem 2. For all $t \geq 1$ holds

$$\text{Var } S_t = 4p_0(1-p_0)t^2 + a(p_0, \lambda)t - a(p_0, \lambda) \frac{1 - (2\lambda - \lambda^2)^t}{(1-\lambda)^2},$$

where

$$a(p_0, \lambda) = \frac{8p_0(1-p_0)}{(1-\lambda)^2}.$$

Proof. As clearly the value S_0 does not affect the variance, let us from now assume $S_0 = 0$. From the definition of variance and (18) follows that to prove the theorem it is enough to prove that

$$ES_t^2 = t^2 + a(p_0, \lambda)t - a(p_0, \lambda) \frac{1 - (2\lambda - \lambda^2)^t}{(1-\lambda)^2}. \quad (21)$$

First of all let us recall that formula (13) holds for the *success rewarding* type of the model as well. The theorem will be once again proved using mathematical induction. For $t = 1$ the definition of the walk yields the same result as in the proof of Theorem 1. By substituting $t = 1$ into (21) we obtain

$$ES_1^2 = 1 + a(p_0, \lambda) - a(p_0, \lambda) = 1$$

and (21) thus holds for $t = 1$. Now for the induction step $t \rightarrow t + 1$ we get by substituting (21), (19) and (18) into (13)

$$\begin{aligned} ES_{t+1}^2 &= ES_t^2 + 4E(P_t S_t) - 2ES_t + 1 = \\ &= t^2 + a(p_0, \lambda)t - a(p_0, \lambda) \frac{1 - (2\lambda - \lambda^2)^t}{(1-\lambda)^2} + \\ &+ 4(p_0 t + 2\lambda p_0(p_0 - 1) \frac{1 - (2\lambda - \lambda^2)^t}{(1-\lambda)^2}) - 2t(2p_0 - 1) + 1 = \\ &= (t+1)^2 + a(p_0, \lambda)(t+1) - a(p_0, \lambda) \frac{1 - (2\lambda - \lambda^2)^{t+1}}{(1-\lambda)^2}. \end{aligned}$$

Substituting (21) and (18) into the definition of variance then proves the theorem.

Corollary 2. For $t \rightarrow +\infty$

$$\lim_{t \rightarrow +\infty} \text{Var } S_t = +\infty$$

and

$$\lim_{t \rightarrow +\infty} \left(\text{Var } S_t - \left(4p_0(1-p_0)t^2 + a(p_0, \lambda)t - \frac{a(p_0, \lambda)}{(1-\lambda)^2} \right) \right) = 0,$$

with $a(p_0, \lambda)$ as in Theorem 2.

Corollary 2 shows that with $t \rightarrow +\infty$ $\text{Var } S_t$ behaves as a quadratic function with respect to t . Similarly as with the *success punishing* model, such behavior is illustrated on Figure 2, which also shows the comparison of the theoretical value of position variance and an empirical one obtained using simulated data.

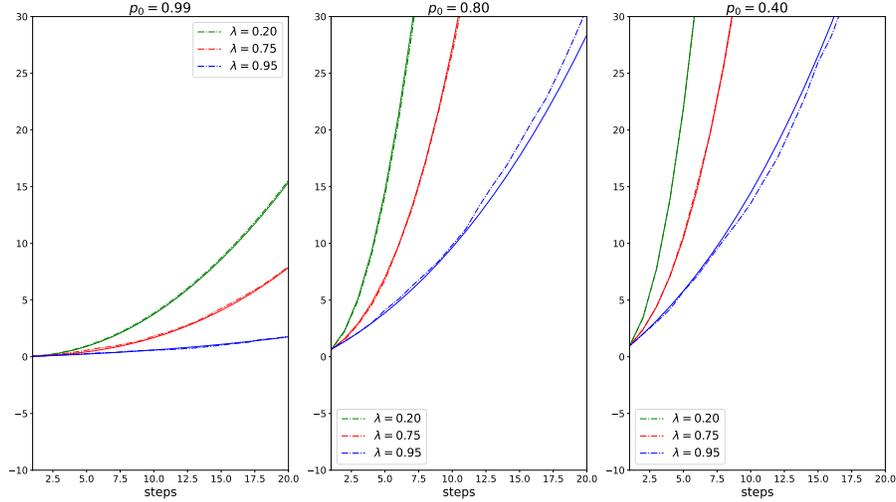


Figure 2. Observed (dash-dotted) and theoretical (solid lines) values of $Var S_t$ – *success rewarding* model. The data were obtained from 1000 walks generated with given parameters.

2.3 Two-parameter models

The presented model can be further extended by adding additional levels of complexity. The first option is to use two separate λ parameters for each direction of the walk. Maintaining the two basic options – *success punishing* and *success rewarding* models – this level of complexity can be defined as follows [5].

Definition 2. Let $p_0, \lambda_0, \lambda_1 \in (0, 1)$ be constant parameters, $\{P_n\}_{n=0}^{\infty}$ and $\{X_n\}_{n=1}^{\infty}$ sequences of discrete random variables, $X_t \in \{-1, 1\}$ and $P_t \in (0, 1)$ for each t , and $P_0 = p_0$. For $t \geq 1$ let

$$P(X_t = 1 | P_{t-1} = p_{t-1}) = p_{t-1}, \quad P(X_t = -1 | P_{t-1} = p_{t-1}) = 1 - p_{t-1},$$

and (*success punishing*)

$$P_t = \frac{1}{2}[(1 + X_t)\lambda_0 P_{t-1} + (1 - X_t)(1 - \lambda_1(1 - P_{t-1}))] \quad (22)$$

or (*success rewarding*)

$$P_t = \frac{1}{2}[(1 - X_t)\lambda_0 P_{t-1} + (1 + X_t)(1 - \lambda_1(1 - P_{t-1}))]. \quad (23)$$

The sequence $\{S_n\}_{n=0}^{\infty}$, $S_n = S_0 + \sum_{i=1}^n X_i$ for $n \in \mathbb{N}$, with $S_0 \in \mathbb{R}$ some given starting position, is called a *random walk with varying probabilities – two-parameter model*, with $\{X_n\}_{n=1}^{\infty}$ being the steps of the walker and $\{P_n\}_{n=0}^{\infty}$ transition probabilities. Depending on the chosen formula to calculate P_t the walk type is either *success punishing* (22) or *success rewarding* (23).

The derivation of exact model properties is not so straightforward as in the case with single lambda. The properties were thus studied with the help of

simulations. Figures 3 and 4 present again the variance of S_t . It seems that the position variance of the *success punishing* model goes to infinity linearly and of the *success rewarding* model quadratically, similarly as in the corresponding single parameter scenarios.

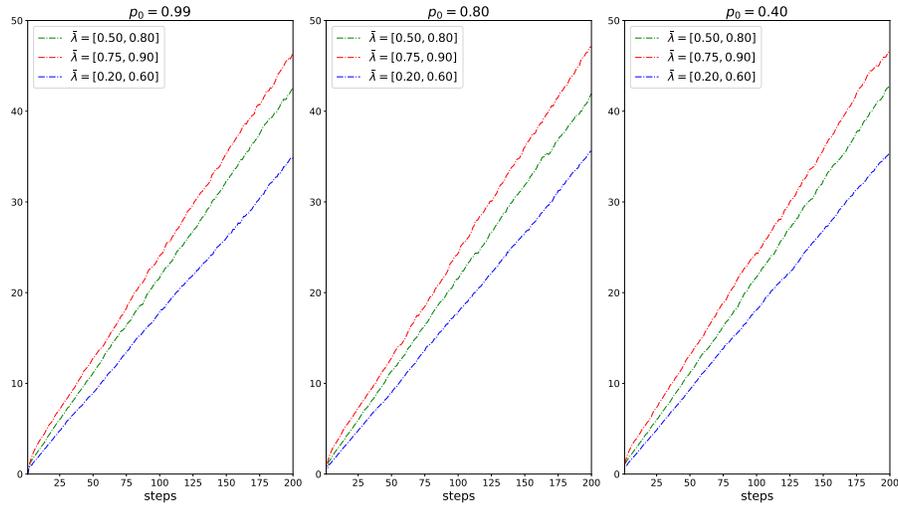


Figure3. The observed position variance of the *two-parameter success punishing* model. The data were gathered from 10000 simulations with given parameters.

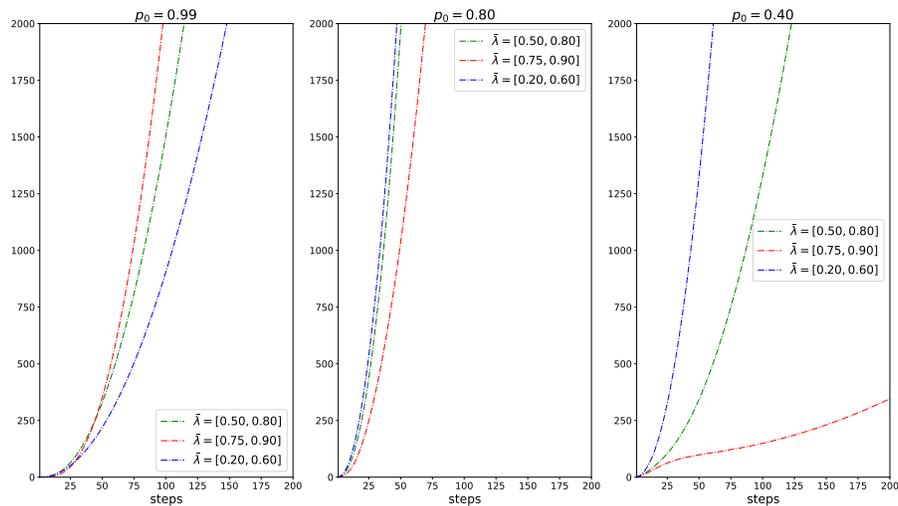


Figure4. The observed position variance of the *two-parameter success rewarding* model. The data were gathered from 10000 simulations with given parameters.

2.4 Other model alternatives

There are many possibilities how to further enhance the model. The model can be combined with a varying step size model from [8], the λ parameter can be handled as a function of time and position, or a combination of varying probability model with regression part (e.g. logistic) can be considered, which seems especially promising from the application point of view. These variations of the model will be subject of further study.

3 Model application

The model is especially well suited for simulation of random processes where a single or just a few events significantly affect the process's future development. An example such a process can be found in sports modeling. In such applications rather short walks occur, but they can be observed repetitively. For example in modeling tennis sets, the longest walk has only 5 steps (occurring in men Grand Slam or Davis Cup matches), but there are many matches played each year, which can be (under some assumptions) considered as multiple observations of the same walk. The authors recently presented a study where the model was used for modeling the men tennis Grand Slam matches with results suggesting the model might provide precious insights when modeling tennis. Here is a brief summary of the modeling approach.

The *success rewarding* version of the model was selected as the historical results show that the development of a tennis match follows such pattern. The p_0 parameter was obtained using input bookmaker odds from *Pinnacle Sports*, an industry leading bookmaker. The appropriate λ parameter was then found from historical data using the maximal likelihood estimate. The model was tested on a database consisting of 4255 tennis matches that took place between 2009 and 2018 and the results suggest that such a model could be used for *in-play* odds prediction. For full details of the model derivation and testing, see the original paper [3].

The quality of such *in-play* predictions was tested on a small study in real life setting with active betting against a bookmaker. The model from [3] was implemented into an automatic odds scraping and betting tool. Whenever the odds provided by bookmaker a_i were higher than the model implied odds, i.e. $a_i > \frac{1}{p}$, a bet was made. This test was carried over the 2019 men tennis US Open and resulted in 128 placed bets with the total amount 59.85 units bet. As the bets were not placed simultaneously, but rather consecutively, the theoretical total bankroll needed for the betting was only 0.52 units – the minimal actual account balance over the entire US Open. The actual number of wins was 57, slightly below the number of expected wins, but thanks to the average winning odds of 2.3 the final balance was plus 2.24 units, creating a theoretical win $\frac{2.24}{0.52} = 4.3$ times the investment, which is an outstanding performance.

This study just briefly shows the possibilities of the presented model and presents rather encouraging results. The testing dataset, consisting of only 128 bets, is however too small to provide a strong evidence favoring the model over

bookmaker's odds. A more conclusive test on a bigger dataset will be subject of further study.

4 Conclusion

The present paper continues in the research on one specific set of models of Bernoulli-like random walks, the models where the transition probabilities depend on the walk history. After reviewing basic models types and the results of previous studies, new properties of the model characterizing the variability of the walk were proved. These properties were explored additionally with the aid of simulations in order to compare derived theoretical results with empirical ones based on simulated data. The problem of parameter estimation was not addressed here, the properties of the maximum likelihood estimate of both λ and p_0 were studied in depth in [5] and utilized in [3], a study devoted to one of possible model applications, namely the modeling of tennis matches development. This kind of application was briefly recalled also in the present contribution and aggregated results of model testing on a new small dataset were reported.

Acknowledgement. The research was supported by the grant No. 18–02739S of the Grant Agency of the Czech Republic.

References

1. Richard A. Davis and Heng Liu. Theory and inference for a class of observation-driven models with application to time series of counts. *Statistica Sinica*, 26:1673–1707, 2016. MR3586234.
2. Tomáš Kouřim. Random walks with varying transition probabilities. In *Proceedings of Doktorandské dny FJFI*, pages 141–149. Czech Technical University in Prague, 2017. Available at <https://km.fjfi.cvut.cz/ddny/historie/historie/17-sbornik.pdf>.
3. Tomáš Kouřim. Random walks with memory applied to grand slam tennis matches modeling. In *Proceedings of MathSport International 2019 Conference (e-book)*, pages 220–227. Propobos Publications, 2019.
4. Tomáš Kouřim. Statistical analysis, modeling and applications of random processes with memory. *PhD Thesis Study, ČVUT FJFI*, 2019. Available at https://tomaskourim.com/publications/2019_rand_proc_w_memory.pdf.
5. Tomáš Kouřim and Petr Volf. Discrete random processes with memory: Models and applications. *Applications of Mathematics, accepted for publishing*, 2020.
6. Karl Pearson. The problem of the random walk. *Nature*, 72(1865):294, 1905.
7. Frank Spitzer. *Principles of random walk*, volume 34. Springer Science & Business Media, 2013.
8. Loïc Turban. On a random walk with memory and its relation with markovian processes. *Journal of Physics A: Mathematical and Theoretical*, 43(28):285006, 2010. MR2658904.

A priori estimation methodology on observation errors of a state space model with linear observation equation using Particle Filtering

Rodi Lykou¹ and George Tsaklidis¹

Department of Statistics and Operational Research, School of Mathematics,
Aristotle University of Thessaloniki, GR54124, Thessaloniki, Greece
(E-mail: lykourodi@math.auth.gr, tsaklidi@math.auth.gr)

Abstract. Observation errors of Particle Filter are studied over the case of a state-space model with a linear observation equation. The observation errors are estimated prior to an upcoming observation. This action is added to the basic algorithm of the filter as a new step for the acquisition of the state estimates. A simulating example is quoted to demonstrate the effectiveness of the proposed method. This intervention is useful in the presence of missing data problems mainly, as well as sample tracking for deprivation and impoverishment issues.

Keywords: Particle filter, Missing data, Single imputation.

1 Introduction

Particle Filter (PF) methodology conducts statistical inference on latent variables of stochastic processes taking into consideration noisy observations related to the latent variables (Gordon et al. [4]). This technique is founded both on the Monte-Carlo (MC) simulation of the hidden variables, as it has been proposed by Metropolis and Ulam [9], and the weight assignment to the random values produced during simulation, the particles. This procedure is repeated sequentially, at every time step of a stochastic process. The involvement of sequential MC simulation in the method adds a heavy computational burden to it. However, the nature of MC simulation renders PF suitable for a wide variety of state-space models, among which non-linear models with non-Gaussian noise are included. The weights are defined according to observations, which are also received at every time step. The weight assignment step constitutes an evaluation process of the existing particles, which are created at the simulation step.

As weight assignment according to an observation dataset is a substantial part of PF, missing observations hinder the function of the filter. Techniques that face the problem of missing data focus mainly on substitution of the missing data. In recent decades, Expectation-Maximization algorithm (Dempster et al. [2]) and Markov-Chain Monte Carlo methods, originally introduced by Metropolis et al. [8], have been popular within missing data problem handling.

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



These algorithms have been constructed independently of PF. Housfater et al. [5] devised Multiple Imputations Particle Filter (MIPF), wherein missing data are substituted by multiple imputations from a proposal distribution and these imputations are evaluated with an additional weight assignment according to their proposal distribution. Xu et al. [13] involved uncertainty on data availability in the observations with the form of additional random variables in the subject state-space model. All the aforementioned approaches are powerful, though computationally costly.

This study focuses on state-space models with linear observation equations and proposes estimation of missing observation errors, in cases of missing data, aiming at the approximation of weights, under the Missing At Random (MAR) assumption (Rubin [11],[12], pg. 53). Linearity in an observation equation permits sequential replacements of missing values with equal quantities of known distributions. Even if this method is applicable to a smaller set of models than the former ones, it is much faster though. In section 2, PF algorithm is presented analytically, in section 3 the new weight estimation step is introduced, in section 4 a simulating example is provided, where the results of the current method are compared with those of MIPF, while, in section 5 discussion and concluding remarks are quoted.

2 Particle Filter algorithm

Let $\{x_t\}_{t \in \mathbb{N}}$ be a stochastic process of m -dimensional latent variables, $x_t \in \mathbb{R}^m$, and $\{y_t\}_{t \in \mathbb{N}}$ be the k -dimensional process of noisy observations, $y_t \in \mathbb{R}^k$. The states and observations are inter-related according to the following state-space model

$$x_t = f(x_{t-1}) + v_t \quad (1)$$

$$y_t = c + Ax_t + u_t \Leftrightarrow y_t - c - Ax_t = u_t. \quad (2)$$

In this system of equations, f is a known deterministic function of x_t , v_t stands for the process noise and u_t symbolises the observation noise. Each sequence $\{v_t\}_{t \in \mathbb{N}}$ and $\{u_t\}_{t \in \mathbb{N}}$ consists of independent and identically distributed (iid) random variables following known distributions. Additionally, $c \in \mathbb{R}^k$ is a constant vector, while $A \in \mathbb{R}^{k \times m}$ is a constant matrix.

PF employs Bayesian inference for state estimation. The Bayesian approach aims at the construction of the posterior probability distribution function $p(x_t|y_{1:t})$, where $y_{1:t} = (y_1, y_2, \dots, y_t)$, resorting to the following recursive equations

$$p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})dx_{t-1} \text{ (prediction)}$$

$$p(x_t|y_{1:t}) = \frac{p(y_t|x_t)p(x_t|y_{1:t-1})}{\int p(y_t|x'_t)p(x'_t|y_{1:t-1})dx'_t} \text{ (update).}$$

These equations are analytically solvable in cases of linear state-space models of Gaussian noises. However, for more general models, analytical solutions are

usually infeasible. For this reason, PF turns to the concept of MC simulation and integration to represent the state posterior $p(x_t|y_{1:t})$ with a set of $N \in \mathbb{N}$ particles $x_t^i, i = 1, 2, \dots, N$ with corresponding weights $w_t^i, i = 1, 2, \dots, N$. Then, the state posterior distribution $p(x_t|y_{1:t})$ can be approximated by the discrete mass probability distribution of the weighted particles $\{x_t^i\}_{i=1}^N$ as

$$\hat{p}(x_t|y_{1:t}) \approx \sum_{i=1}^N w_t^i \delta(x_t - x_t^i),$$

where δ is the Dirac delta function and weights are normalized, so that $\sum_i w_k^i = 1$. As $p(x_t|y_{1:t})$ is unknown, this MC simulation is based on importance sampling, namely a known probability density $q(x_t|y_{1:t})$ is chosen and called importance density for the set of particles to be produced. Thus, the state posterior distribution is approximated as

$$\hat{p}(x_t|y_{1:t}) \approx \sum_{i=1}^N \tilde{w}_t^i \delta(x_t - \tilde{x}_t^i),$$

with $\tilde{x}_t^i \sim q(x_t|y_{1:t})$, while

$$\tilde{w}_t^i \propto \tilde{w}_{t-1}^i \frac{p(y_t|\tilde{x}_t^i)p(\tilde{x}_t^i|\tilde{x}_{t-1}^i)}{q(\tilde{x}_t^i|\tilde{x}_{t-1}^i, y_t)} \quad (3)$$

are the normalized particle weights for $i = 1, 2, \dots, N$.

In this framework, as PF continues to function for several time steps, the most probable particles are assigned bigger and bigger weights, while the weights of the other particles become negligible progressively. Thus, only a very small proportion of particles is finally used for the state estimation. This phenomenon is known as PF degeneracy. For its avoidance, a resampling with replacement step, according to the calculated weights, has been incorporated to the initial algorithm, having as a result the Sampling Importance Resampling (SIR) algorithm. Nevertheless, sequential resampling leads the particles to take values from a very small domain and exclude many other less probable values. This problem is called impoverishment. A criterion over the weight variance has been introduced for a decision to be taken at every time step, whether existing particles should be resampled or not, reaching the middle ground between degeneracy and impoverishment. Liu [6] (pg. 35-36) has quoted an analytical formula to measure the particle degeneracy with another quantity

$$N_{eff}(t) = \frac{N}{1 + Var_{p(\bullet|y_{1:t})}(w(x_t))}$$

named effective sample size. As this quantity cannot be calculated directly, it can be estimated as

$$\hat{N}_{eff}(t) = \frac{1}{\sum_{i=1}^N (\tilde{w}_t^i)^2}.$$

The decision on resampling is positive whenever $\hat{N}_{eff}(t) \leq N_T$, where $N_T = c_1 N, c_1 \in \mathbb{R}$ is a fixed threshold. A usually selected value for N_T is 75% N .

Algorithm 1 SIR Particle Filter

Require: N, q, N_{eff}, T
 Initialize $\{x_0^i, w_0^i\}$
for $t = 1, 2, \dots, T$ **do**
 1.Importance Sampling
 Sample $\tilde{x}_t^i \sim q(x_t|x_{0:t-1}^i, y_t)$
 Set $\tilde{x}_{0:t}^i = (x_{1:t-1}^i, \tilde{x}_t^i)$,
 Calculate importance weights
 $\tilde{w}_t^i \propto w_{t-1}^i \frac{p(y_t|\tilde{x}_t^i)p(\tilde{x}_t^i|x_{t-1}^i)}{q(\tilde{x}_t^i|x_{t-1}^i, y_t)}$.
 end for
 for $i = 1, 2, \dots, N$ **do**
 Nomalize weights $w_t^i = \frac{\tilde{w}_t^i}{\sum_{i=1}^N \tilde{w}_t^i}$
 2. Resampling
 if $\hat{N}_{eff}(t) \geq N_T$ **then**
 for $i = 1, 2, \dots, N$ **do**
 $x_{0:t}^i = \tilde{x}_{0:t}^i$
 end for
 else
 for $i = 1, 2, \dots, N$ **do**
 Sample with replacement index
 $j(i)$ according to the discrete
 weight distribution $P(j(i) =$
 $d) = w_t^d, d = 1, \dots, N$
 Set $x_{0:t}^i = \tilde{x}_{0:t}^{j(i)}$ and $w_t^i = \frac{1}{N}$
 end for
 end if
 end for

Algorithm 1 summarizes PF steps. The sampling part corresponds to the prior (prediction) step of Bayesian inference, while weight assignment and possible resampling constitute the posterior (update) step.

3 The missing data case - Estimation of weights

Concerning the missing data case, some new definitions need be quoted. Let a random indicator variable $R_{t,j}$, $j = 1, \dots, k$, indicate whether the k^{th} component of the t^{th} observation is available or not. That is,

$$R_{t,j} = \begin{cases} 0 & \text{the } j^{th} \text{ component is available at time } t \\ 1 & \text{otherwise} \end{cases}.$$

Additionally, sets Z_t and W_t are defined as the collections of available and missing components of the observation variables Y_t respectively for every time step $t \in \mathbb{N}$.

The proposed method of this paper is based on the MAR assumption, according to which the missing data mechanism does not depend on the missing observations, given the available ones:

$$P(R_{t,j}|Z_t, W_t) = P(R_{t,j}|W_t), t \in \mathbb{N}, j = 1, \dots, k.$$

Now, let x_{t-1}^i be a particle for the a posteriori estimation of the hidden state x_{t-1} of the state-space model (1)-(2). Then, according to Algorithm 1 and equation (1), the i^{th} prior estimate of the hidden state x_t is produced by equation

$$\tilde{x}_t^i = f(x_{t-1}^i) + v_t^i. \quad (4)$$

According to equation (2) the error of the estimate is calculated as

$$u_t^i = y_t - c - A\tilde{x}_t^i \quad (5)$$

The weight assigned to \tilde{x}_t^i depends on u_t^i , because, according to equations (3) and (5),

$$p(y_t|\tilde{x}_t^i) = p(u_t^i).$$

Then, as the two variables (\tilde{w}_t^i and u_t^i) are intimately connected, knowledge on the distribution of u_t^i leads to the distribution of \tilde{w}_t^i . Even if the distribution of \tilde{w}_t^i may not be exactly calculated, in cases where \tilde{w}_t^i are complicated functions of u_t^i , knowledge on the distribution of u_t^i will suffice to provide information on the weight distribution. Thereby, calculation of $p(u_t^i \in D), D \subset \mathbb{R}^k$, is of interest for the concomitant estimation of weights. On the supposition that the whole observation y_t is unavailable, sequential replacements of u_t^i and y_t from equations (5) and (2), respectively, contribute to the creation of the following formula

$$\begin{aligned} u_t^i &= c + Ax_t + u_t - c - A\tilde{x}_t^i \\ &= A(x_t - \tilde{x}_t^i) + u_t. \end{aligned}$$

As the whole observation is considered missing, particles \tilde{x}_t^i cannot be evaluated. Thus, both x_t and \tilde{x}_t^i are replaced according to equations (1) and (4)

$$\begin{aligned} u_t^i &= Af(x_{t-1}) + Av_t - Af(x_{t-1}^i) + Av_t^i + u_t \\ &= A(f(x_{t-1}) - f(x_{t-1}^i)) + Av_t - Av_t^i + u_t. \end{aligned}$$

However, the replacement of \tilde{x}_t^i can be avoided, as MC simulation has been implemented at this time point. The hidden state x_{t-1} is unknown, but its posterior distribution is available, so that a point estimate of it \hat{x}_{t-1} can be calculated. Therefore,

$$u_t^i \approx A(f(\hat{x}_{t-1}) - f(x_{t-1}^i)) + Av_t - Av_t^i + u_t. \quad (6)$$

As the distributions of the random variables v_t, v_t^i and u_t are known, the distribution of $Av_t - Av_t^i + u_t$ is also known, as it is the convolution of linear functions of the initial components v_t, v_t^i and u_t . Calculation of such convolutions is not always an easy task, as analytical solutions may not be feasible, leading to numerical approximation options (Lykou and Tsaklidis [7]). However, given that each noise process consists of iid variables and matrix A is constant, the distribution of this sum needs to be calculated only once. Therefore, since the quantity $A(f(\hat{x}_{t-1}) - f(x_{t-1}^i))$ is a constant at time t , the distribution of u_t^i can be approximated as

$$p(u_t^i) \approx p(A(f(\hat{x}_{t-1}) - f(x_{t-1}^i)) + Av_t - Av_t^i + u_t).$$

Estimation of u_t^i implies the estimation of y_t , according to equation (5). If observation y_t is partially available, its available components, say Z_t collection, can be placed in the above equations. Thus, some components of the observation error will also be available, while the rest of the components, say W_t collection, possibly dependent on the available ones, can be estimated with the same method. In any case, point estimators of u_t^i can be estimated along with their weights. Consequently, the initial PF algorithm sustains a slight change, as it is presented in Algorithm 2.

Algorithm 2 SIR Particle Filter for missing data with observation error estimation

Require: N, q, N_{eff}, T
Initialize $\{x_0^i, w_0^i\}$
for $t = 1, 2, \dots, T$ **do**
 1.Importance Sampling
 Sample $\tilde{x}_t^i \sim q(x_t|x_{0:t-1}^i, y_t)$
 Set $\tilde{x}_{0:t}^i = (x_{1:t-1}^i, \tilde{x}_t^i)$,
 Produce observation error estimates \hat{u}_t^i for the missing components Z_t and calculate importance weights $\tilde{w}_t^i \propto w_{t-1}^i \frac{p(y_t|\tilde{x}_t^i)p(\tilde{x}_t^i|x_{t-1}^i)}{q(\tilde{x}_t^i|x_{t-1}^i, y_t)}$.
end for
for $i = 1, 2, \dots, N$ **do**
 Nomalize weights $w_t^i = \frac{\tilde{w}_t^i}{\sum_{i=1}^N \tilde{w}_t^i}$
 2. Resampling
 if $\hat{N}_{eff}(t) \geq N_T$ **then**
 for $i = 1, 2, \dots, N$ **do**
 $x_{0:t}^i = \tilde{x}_{0:t}^i$
 end for
 else
 for $i = 1, 2, \dots, N$ **do**
 Sample with replacement index $j(i)$ according to the discrete weight distribution $P(j(i) = d) = w_t^d, d = 1, \dots, N$
 Set $x_{0:t}^i = \tilde{x}_{0:t}^{j(i)}$ and $w_t^i = \frac{1}{N}$
 end for
 end if
end for

4 Simulating example

A simulating example is presented in this section for the comparison of the basic PF algorithm, the proposed method and multiple imputation particle filter (MIPF) for $n = 5$ imputations (Housfater et al.[5]). The reduction step proposed by Acunã et al [1] is incorporated in the initial MIPF algorithm for the best possible results to be achieved. The data simulation is based on the state-space model of equations (1)-(2), with 2-dimensional variables

$$x_t = \begin{bmatrix} x_{1,t} \\ x_{2,t} \end{bmatrix} = \begin{bmatrix} \cos(x_{1,t-1} - x_{1,t-1}/x_{2,t-1}) \\ \cos(x_{2,t-1} - x_{2,t-1}/x_{1,t-1}) \end{bmatrix} + \begin{bmatrix} v_{1,t} \\ v_{2,t} \end{bmatrix}$$

$$y_t = x_t + u_t,$$

where $v_t = \begin{bmatrix} v_{1,t} \\ v_{2,t} \end{bmatrix} \sim N(\mu, S_1)$, $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $S_1 = \begin{pmatrix} 0.05 & 0 \\ 0 & 0.05 \end{pmatrix}$, $u_t \sim N(\mu, S_2)$, $S_2 = \begin{pmatrix} 0.03 & 0 \\ 0 & 0.03 \end{pmatrix}$ and N symbolizes Gaussian distribution. Initial condition $x_0 = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$ is considered known. Concerning missing data, $R_{t,j} \sim Bernoulli(0.15)$, $j = 1, 2$. $N = 100$ particles have been used for every filter. The distributions of noises are also considered known. The weighted mean is used as a point estimator of a hidden state. All the filters have been repeated for 100 times and their performance concerning their precision and consumed time has been recorded.

The results of the three methods are shown in Table 1. The mean (over the simulations) of Root-Mean-Square-Error (RMSE) of estimators for each component of the hidden states is presented. The mean of the two aforementioned columns is also calculated, as well as the mean time consumed in each

approach. (The code has been written in R project (R core team [10]). Package *mvnorm* (Genz et al. [3]) has also been used. Simulations have been performed on an AMD A8-7600 3.10GHz processor with 8 GB of RAM.) In the table, it is shown that the weight estimation with the suggested method outperforms MIPF concerning both precision and time elapsed. The proposed method is also compared with the results of the basic PF algorithm, at which all observations are available, and it seems that, even if the precision is inevitably reduced in the case of missing data, the computational time remains nearly the same. The small differentiation in the mean elapsed time is probably connected with the resampling decision. In Figures 1 and 2 the performances of the proposed method and MIPF are depicted for the two components of the state process respectively. The estimates of the two approaches are close to one another, tracking the hidden variables satisfactorily. Therefore, in this example, the suggested method appears to be the best option between the available ones in the missing data case.

Method	Mean RMSE for $(x_{1,t})$	Mean RMSE for $(x_{2,t})$	Overall mean precision	Mean time elapsed (sec)
Basic PF	0.1610253	0.1566881	0.1588567	2.5570
Weight est.	0.2065578	0.2102287	0.2083933	2.5491
MIPF	0.2267527	0.2173670	0.2220598	4.9137

Table 1. Comparison of the results over three methods: The basic PF algorithm, when all observations are available, weight estimation method, which was proposed in this study, and MIPF for $n = 5$ imputations. These methods are compared over the mean of RMSEs and time consumed over the 100 repeated implementations.

5 Discussion and conclusions

In this study, point estimation of observation errors is proposed for missing data cases, when PF is implemented and MAR assumption is adopted. This method is a single imputation procedure. Acuña et al. [1] have argued against single imputation, as it is rather simplistic, as it cannot attribute to a single value the distributional characteristics that can be approached and described by a sample of multiple imputation. Nevertheless, the primary target of the proposed technique is the minimization of the computational cost that is added to the initial PF algorithm, when missing data need be handled. For this purpose, interventions in the PF algorithm are slight. Moreover, in the provided simulation, the suggested method outperforms the multiple imputation approach even for a considerable number of imputations, whereas Acuña et al. [1] have noticed that MIPF with $n = 3, 4, 5$ imputations result in more than satisfactory results, according to the approximation of multiple imputation estimate efficiency provided by Rubin [12]. As a result, in this example, estimation of observation errors seems to perform well both in respect of the computational time it requires and the precision it achieves.

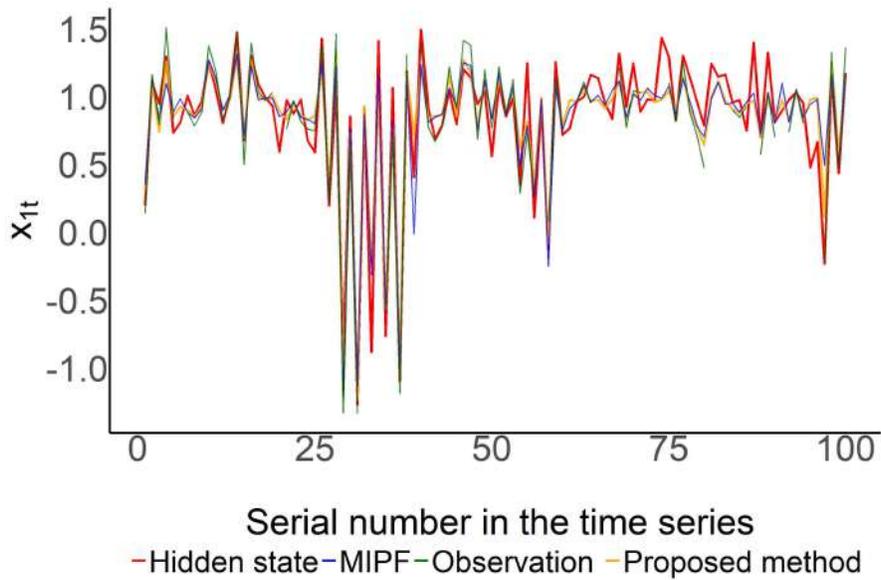


Fig. 1. Time series of the hidden values, the observations, the corresponding point estimates of the proposed method and MIPF imputations respectively, for the first component $x_{1,t}$ of the state process.

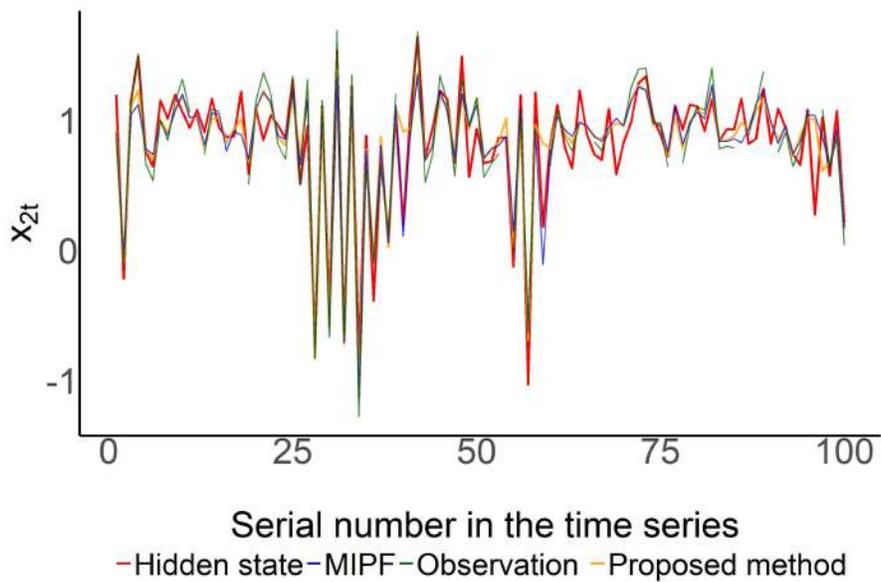


Fig. 2. Same as Figure 1 for the second component $x_{2,t}$ of the state process.

The contribution of such a method to deprivation and impoverishment problems also worths mentioning. This method permits the estimation of observation errors and their corresponding weights one time step after the current. The assessment of weight distribution for the forthcoming time steps could be very interesting, as far as it is connected with the two aforementioned problems. The evolution of weight distribution cannot be a priori estimated for multiple time steps, to the most of the authors' knowledge, but this is feasible for one step forward. As the weights of the next step can be estimated, the probabilities that a particle will be chosen at the resampling step can also be estimated.

References

1. D. E. Acuña, M. E. Orchard, J. F. Silva and A. Prez. Multiple-imputation-particle-filtering for uncertainty characterization in battery state-of-charge estimation problems with missing measurement data: Performance analysis and impact on prognostic algorithms. *International Journal of Prognostics and Health Management*, 6, 2015.
2. A. P. Dempster, N. M. Laird and D. B. Rubin. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, 39, 1-38, 1977.
3. A. Genz, F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl and T. Hothorn. mvtnorm: Multivariate Normal and t Distributions. *R package version 1.1-0*, 2014. <https://CRAN.R-project.org/package=mvtnorm>.
4. N. J. Gordon, D. J. Salmond and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F Radar and Signal Processing*, 140, 2, 107, 1993.
5. A. S. Housfater, X. P. Zhang and Y. Zhou. Nonlinear fusion of multiple sensors with missing data. *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings,IV*, 961-964, 2006.
6. J. S. Liu. Monte Carlo Strategies in Scientific Computing. *Springer-Verlag*, New York, 2001.
7. R. Lykou and G. Tsaklidis. Prior estimation of observation erros of Particle Filter. *32nd Panhellenic Statistics Conference*, 2019. (In Greek, under review)
8. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21, 1087-1092, 1953.
9. N. Metropolis and S. Ulam. The Monte Carlo Method, *Journal of the American Statistical Association*, 1949.
10. R Core Team. R: A Language and Environment for Statistical Computing, 2014, <http://www.r-project.org/>.
11. D. B. Rubin. Inference and missing data. *Biometrika*, 63, 581-592, 1976.
12. D. B. Rubin. Multiple Imputation for Nonresponse in Surveys. *Wiley*, 1987.
13. L. Xu, K. Ma, W. Li, Y. Liu and F.E. Alsaadi. Particle filtering for networked nonlinear systems subject to random one-step sensor delay and missing measurements. *Neurocomputing*, 275, 2162-2169, 2018.

Social vulnerability analysis to multi hazards at the inter-municipal scale: The case in southern Italy, Calabria

Roberta Maletta¹, Giuseppe Mendicino¹, and Luigi Maria Mollica²

¹ Department of Environmental Engineering (DIAM), University of Calabria, UNICAL, Via Pietro Bucci, 87036, Arcavacata di Rende (CS), Italy
(E-mail: robertamaletta@gmail.com, giuseppe.mendicino@unical.it)

² Civil Protection of Calabria Region, Viale Europa - Cittadella Regionale - Località Germaneto 88100 Catanzaro, Italy

Abstract. Vulnerability is an important component of risk assessment and it represents the main element in the perception of risk. In fact, the characteristics related to physical, socio-economic and cultural factors increase the susceptibility of an individual or a community, to the impacts of hazards. This study presents an application of the modified Social Vulnerability Index (SoVI) method in the territorial context of Marina di Gioiosa Ionica, located in southern Italy. In particular, principal components analysis (PCA) with varimax rotation and Kaiser criterion for component selection is adopted. The investigation utilizes 23 variables that, are reduced to four principal components, to represent 79% of variance of the data. Factor scores are calculated to get the final SoVI scores and to identify and map social vulnerable categories that affect an inter-municipal territorial context. A dedicated Geographic Information System (GIS) is used to capture, geo-process and display spatial data recorded at different scales. Results show significant differences in the spatial distribution of the social vulnerability, highlighting the multidimensionality and heterogeneity of the municipal characteristics. The highest social vulnerability index is concentrated in the southern portion of the study area. Social vulnerability map can be a useful tool for public decision-makers to implement effective planning and prevention and security policies. Moreover, a proper allocation of resources aims to reduce population health risks in the field of civil protection.

Keywords: Social vulnerability assessment, Multi hazards, Principal component analysis, vulnerability index and map, municipalities.

1 Introduction

In recent years, an increasing number of studies has been developed to give a proper definition of vulnerability. In general, vulnerability can be defined as an internal risk factor of the community or system, that is exposed to a natural or hand-made hazard, and it represents physical and socio-economic susceptibility or predisposition to damage phenomena (Cardona[1], Ciurean *et al.*[2]). However the concept of vulnerability has been employed by a large number of authors, from different knowledge areas (Birkmann[3], Adger and Kelly[4], Adger[5], Fekete *et al.*[6], Paul[7]). In each work, the definition of vulnerability

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



is correlated to the purpose of scientific study. As a matter of fact, researchers arising from different disciplines, state several methods or measurements of vulnerability (Hufschmidt[8]).

In this paper, a measure of social vulnerability is evaluated on the basis of methodology of Cutter *et al.*[9], which is based on the concept that disasters are mainly caused by social systems and human related characteristics, that make people vulnerable (Borden *et al.*[10]). This approach, in particular, uses a large number of measurable variables and a factor analysis to identify influential components, which make the territory socially vulnerable in multi-hazard scenarios. The present paper implements the Social Vulnerability Index (SoVI) method, but modifies variables used in the original construction to adapt them to the territorial context under investigation. The work provides several important contributions. In particular, social vulnerability method was calculated for an unusual reference inter-municipal scale, using census areas as spatial units, produced by the Italian National Institute of Statistics (ISTAT). In this framework, it is shown that the influence of adjoining several municipalities, is very important in planning activities. Moreover, the analysis takes into account also variables related to territorial emergency planning. In particular, the distances from strategic buildings and civil protection operational structures are assessed. Finally, the principal component analysis (PCA) results show driver variables and significant variations for several vulnerability levels. A set of geoprocessing activities was implemented by a dedicated Geographical Information Systems (GIS). The outline of the paper is as follows. In Section 2, a brief description of study area is provided, whereas in Section 3, the methodology and the dataset used for developing the statistical model are shown. Section 4 contains results from the factor analysis and maps, that display the resulting scores. Finally, Section 5 presents the conclusions of the study and potential developments.

2 Study area

This research was carried out to investigate social vulnerability of the inter-municipal territorial context of Marina di Gioiosa Jonica, in the Calabria region, in southern Italy (Fig. 1). The study area consists of six municipalities: Gioiosa Jonica, Grotteria, Mammola, Marina di Gioiosa Jonica, Martone e San Giovanni di Gerace. It is highlighted that the Calabria region has defined from 2016 new inter-municipal administrative boundaries called “Territorial Contexts”, for civil protection purposes. In fact, the geographical definition of these areas is aimed at reducing risk and improving the management of emergencies. The area under study is defined as the Territorial Context of Marina di Gioiosa Ionica. The total population in 2019 is 20’391 inhabitants and the total area is about 191 kmq. The perimeter of the union of the municipalities is included in the catchment area of the Torbido river. The physical environment of the territorial context is exposed to different types of risks: the mountainous area is subject to landslides, floods, and forest fires; and the valley areas are prone to floods and storms.

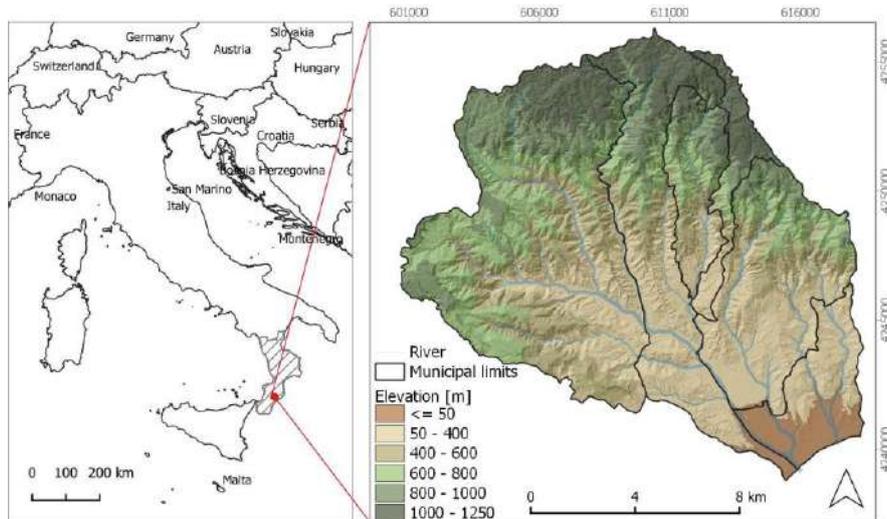


Fig. 1. Location of the study area in Italy (SR:WGS84/UTM zone 33N)

3 Data-set and methodology

Most of the data come from existing 2011 datasets of the Italian National Institute of Statistics (ISTAT), other ones are provided by local or regional authorities. The base unit in the statistical method is the population of census area, i.e. equal to 195, for which data are collected or derived. All variables are elaborated through GIS and the accuracy of the dataset is verified.

Based on the approach proposed by Cutter *et al.*[9], the present study uses the variables reported in Tab. 1. Some of the variables used in this study, refer to those reported in the literature (Mavhura *et al.*[11], Birkmann[3]). Other selected indicators differ from the original SoVI model, because they depend on both quality of available variables and territorial's situation and reality.

The variables are normalized and standardized before the statistical analysis. No values were missing in the dataset. During the processing some variables are discarded because they had little relationship with the components. Due to space limitations, only the final 23 variables, used in the factor analysis technique, are summarized in the Tab. 1.

The multivariate statistical technique is performed by the SPSS software. The PCA approach, condenses an original set of variables into a smaller number of principal components, that better describe variations in the data through identification and clustering of variables that measure the same theme. The use of this method is allowed for a consistent and reliable set of variables, that can be monitored over the time to assess any changes in social vulnerability in the territorial context. Varimax rotation is used to simplify the structure of the underlying dimensions and to produce more independence among components. The Kaiser criterion (eigenvalues > 1) is applied for the component selection.

Components are examined and added together to produce the final vulnerability index. In order to check the robustness of the model, the Kaiser-Meyer-Olkin (KMO) of sampling adequacy and the Barlett's Test of Sphericity are used.

Concept	N.	Variables	Data source
Special needs populations	1	percentage of residents with disabilities	local institutions
Demographic trend	2	population density	ISTAT
	3	percentage growth rate 2001-2011	ISTAT
Age	4	percentage of residents aged 5 and less	ISTAT
	5	percentage of residents aged 5 to 14 years	ISTAT
	6	percentage of residents aged 65 years and over	ISTAT
Family status and structure	7	percentage of separated and divorced residents	ISTAT
	8	percentage of widowed residents	ISTAT
	9	percentage of households with 6 and over	ISTAT
	10	percentage of households in rental housing	ISTAT
Gender	11	percentage of female population	ISTAT
Education	12	percentage of residents with lower secondary education	ISTAT
	13	percentage of residents with primary education	ISTAT
	14	percentage of residents with no education	ISTAT
Employment loss	15	percentage of residents aged 15 and over unemployed and looking for new jobs	ISTAT
	16	percentage of residents aged 15 and over housewife	ISTAT
Ethnicity	17	percentage of foreigner/stateless residents between 0 and 29 years	ISTAT
	18	percentage of foreigner/stateless residents aged 54 years and over	ISTAT
Development	19	percentage of employees who temporarily occupy the properties	ISTAT
	20	percentage of worker volunteers who temporarily occupy the properties	ISTAT
Tourists	21	percentage of beds of tourist accommodation facilities	web data
Emergency management and rescue organization	22	distance to nearest hospital (km)	calculated on the basis of transport modeling
	23	average distance from the civil protection operational structures (km)	

Tab. 1. Vulnerability concepts and variables used for the territorial context

Note that the latest variables reported in Tab. 1, take into account the elements of emergency planning. Such data are obtained through a transport modeling, based on the All or Nothing assignment method, which involves the concept of traffic distribution, planning and management (Thomas[12]). In this method the trips from any origin zone to any destination zone are calculated like the shortest path between them. In order to simulate demand and supply in the transportation

system, road network, origin points (centroids of the ISTAT census areas) and destination points (key elements of the emergency planning) are defined (Fig.2). The latter points refer to: strategic buildings (hospital, operating centers, etc.), external point access and civil protection operational structures (firefighters, voluntary civil protection associations, etc.). The analysis is carried out by using the AequilibraE plugin in the QGIS environment (Camargo[13]). Through this modeling, the shortest paths between origin and destination points, the travel times and flows affecting link road are obtained.

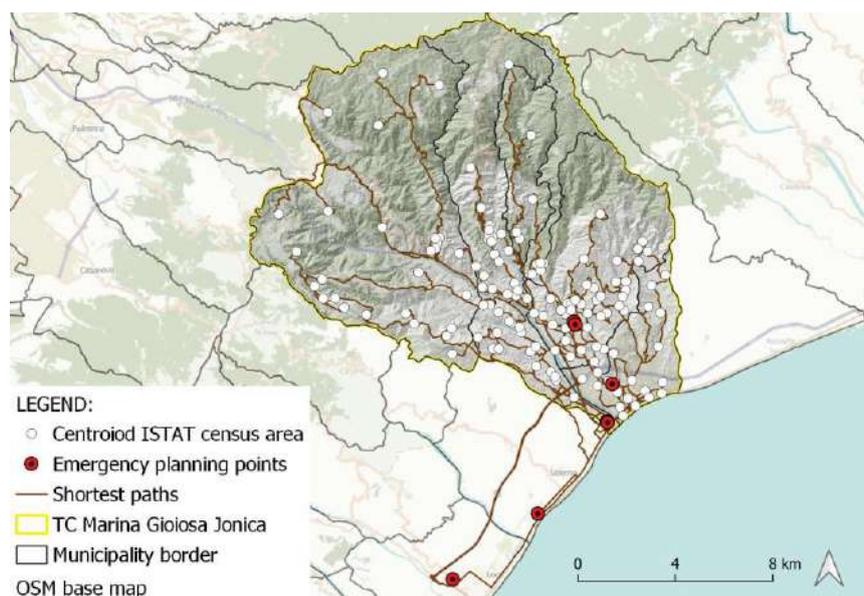


Fig. 2. Map of shortest paths between the origin and destination points in the transport modeling

4 Results

The adequacy of sampling is tested through Kaiser-Meyer-Olkin (KMO), while the strength of the relationship among variables is assessed through Bartlett's test of sphericity. These tests are conducted in principal component analysis to identify whether the data are appropriate for factor analysis. The KMO value calculated is 0.89 and shows that the sampling is quite adequate. As a general rule, the Bartlett test is supposed to be significant ($p < .001$). It tested the null hypothesis that the correlation matrix is an identity matrix. In this study it is highly significant (Sig.=0.000), implying that the data are appropriate for component analysis. After having passed the KMO and Bartlett's test of sphericity, the factor analysis technique is then conducted.

At the inter-municipal level, the PCA approach has generated four components, with eigenvalues greater than 1, accounting for 79% of the variance. Note that before reaching the final solution, several analyses based on the method

extraction, rotation and number of components are carried out (Osborne[14]). The components and communalities are shown in Tab. 2. Variables with the largest absolute loadings (greater than 0.5) are identified, subsequently the components are named in terms of their driver variables and their representation of social vulnerability. In PCA, the initial communalities are always 1.0 for all variables. Otherwise, extracted communalities are calculated as the variance accounted for in each variable by all extracted components. They are ranged from 0.412 to 0.98.

The four components of social vulnerability at the inter-municipal level are: 1) Population structure and socioeconomic status; 2) Emergency management and rescue organization; 3) Demographic trend and 4) Temporary occupants (tourists and volunteer workers).

Variables		Component Loading				Communality
Concept	n.	1	2	3	4	
Special needs populations	1	.614				.412
Demographic trend	2			.914		.861
	3			.900		.812
Age	4	.923				.880
	5	.949				.940
	6	.873				.907
Family status and structure	7	.878				.851
	8	.859				.872
	9	.874				.774
	10	.808				.786
Gender	11	.955				.980
Education	12	.947				.944
	13	.918				.928
	14	.695				.616
Employment loss	15	.895				.815
	16	.939				.939
Ethnicity	17	.750				.589
	18	.730				.591
Development	19	.690				.765
	20				.641	.432
Tourists	21				.774	.638
Emergency management and rescue organization	22		.957			.933
	23		.968			.937

Tab. 2. Rotated Component Matrix and Communality

The first main component, “Population structure and socioeconomic status”, contributes with 53.5 percent of the total variance. This component includes most of the important variables. It is highlighted that in risk scenarios, such as pandemics, fires, etc., the territorial demographic characteristics are necessary to manage the medical response. The diagram reported in Fig. 3, shows, in the study area, the distribution of age group in terms of municipality. In the inter-municipal territorial of Marina di Gioiosa Jonica, people aged 65 and older

represent a quarter of its total population and it contributes to the definition of high vulnerable areas (Fig.4a).

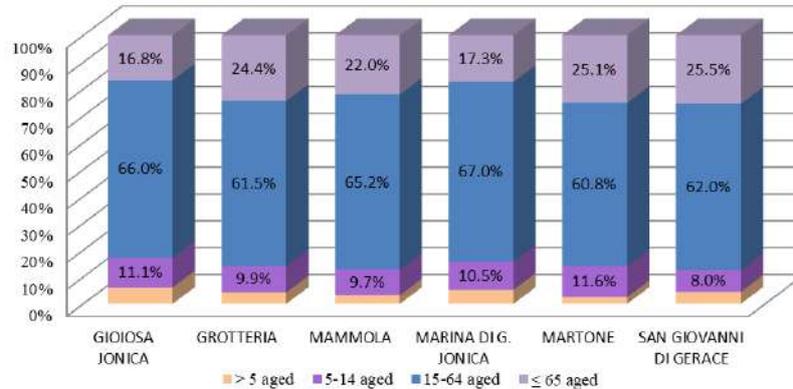


Fig. 3. Demographic variables by municipalities (data source: ISTAT, 2011)

The second component, “Emergency management and rescue organization”, contributes with 9 percent of the total variance. This component has strong loadings on the last two items and it is useful to identify hinterland areas with a high relative vulnerability due to the greater distances from the strategic emergency buildings (Fig. 4b). Third and fourth components, i.e. “Demographic trend” and “Temporary occupants” (tourists and volunteer workers), contribute with percentages equal to 8.5 and 8 of the total variance, respectively.

In order to identify the spatial patterns of social vulnerability, the component score weights created in a factor analysis process, are used. They are similar to coefficients in multiple regression analysis. For each component a cardinal direction is assigned: positive (+) if majority of the variables in the component increases vulnerability, negative (-) if majority of the variables decreases vulnerability and the absolute value if the component variables have a mixed impact on vulnerability. After adjustment, following components scores are added to obtain the value of the total vulnerability score. Each component score thus has an equal contribution to the overall vulnerability score and it is mapped at the census ISTAT unit. Once computed, the component scores and the SoVI scores are mapped using standard deviations (SD) from the mean to show the spatial variability across the municipalities (Fig 4, 5). On the study area at the census area level, SoVI scores greater than 1 SD are labeled as most vulnerable. On the contrary, values less than -1 SD are labeled as low vulnerability. The 0.5 SD indicates the breakpoints.

The relative scores displayed in Fig 4 a-d, represent how relative vulnerability varies across ISTAT census space. The Fig. 4e shows the municipal boundaries of the territorial context of Marina di Gioiosa Jonica.

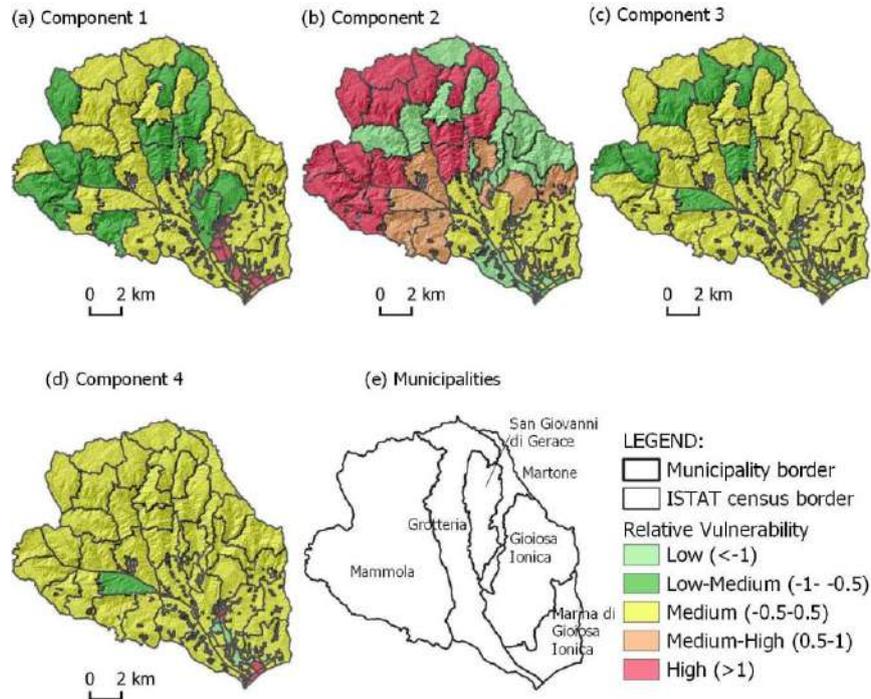


Fig. 4. Spatial distribution of relative Vulnerability index in Marina di Gioiosa Ionica context: a) Population structure and socioeconomic status; b) Emergency management and rescue organization; c) Demographic trend; d) Temporary occupants (tourists and volunteer workers); e) Reference map.

Final SoVI map shows that the most socially vulnerable zone is that referring to the lower part of the inter-municipal study area (Fig.5). It is worth noting that such municipality is the one that registers over the years a large population density as well as a low demographic decrease. Then such criticality is basically correlated to urbanization and most dominant economic activities. There are also other high vulnerability values in the historical centers of some municipalities. The northern part of the study area, presents a medium-high vulnerability level, due to Component 2, whereas the green areas, with the lowest vulnerability index, indicate close to zero or zero population.

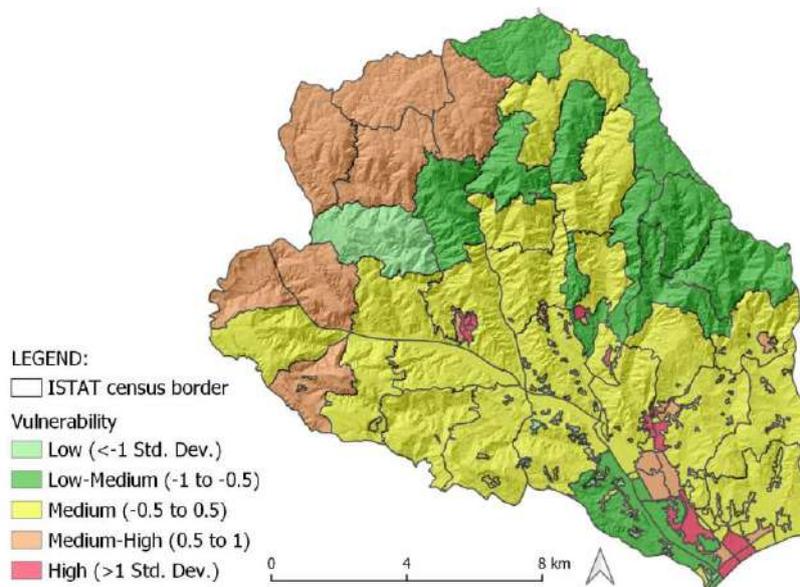


Fig. 5. Spatial distribution of final social vulnerability index the territorial context of Marina di Gioiosa Ionica.

Fig 6 shows the data of five categories of social vulnerability related to the number of affected ISTAT census zones. It should be noted that Grotteria municipality is the one with the largest number of census zones, with a total area of about 38 kmq, whereas the largest municipality is that of Mammola with a total area of about 81 kmq. It is highlighted that the number of Grotteria census zones is quite influenced by the narrow and elongated shape of the municipality, that has the greatest longitudinal extension. The spatial distribution of housing settlements in the territory makes it a low vulnerable municipality with the exception of the historical center area, located in the hinterland. The municipality of Grotteria is characterized by the highest number of ISTAT census areas with a low vulnerability index. The most vulnerable municipality is that of Marina di Gioiosa Ionica, which has a higher population density. The municipality of Mammola is the one with the largest number and extension of census areas that falls into the category of medium-high vulnerability. In fact, many areas are vulnerable due to the lack of services and their distance from strategic civil protection buildings. The municipality of Gioiosa Ionica is characterized by the highest number of ISTAT census areas with a medium vulnerability. In general by analyzing the values of the resulted vulnerability scores it must be noted that 28.7% of census areas are characterized by low and low-medium, while 21 % fall into categories labelled with high and medium-high vulnerability.

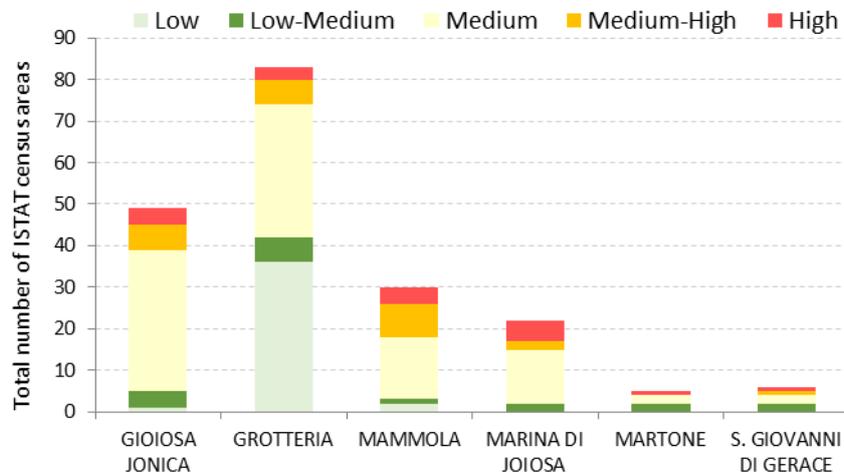


Fig. 6. Distribution of the social vulnerability values (from low to high) based on the number of ISTAT census zones, on the basis of the six municipalities belonging to the territorial context of Marina di Gioiosa Jonica.

Conclusions

The aim of this article is to evaluate the Social Vulnerability Index (SoVI), adapted at the inter-municipal level, in the southern part of Italy. The study area includes 6 municipalities and 195 ISTAT census zones. This area is called “Territorial Context of Marina di Gioiosa Ionica” and assumes a certain regional importance in emergency planning. Over the past ten years, in this area several hydrogeological/hydraulic and forest fire events were recorded. Then the research studies emphasize the importance to quantify a multi-risk vulnerability index.

The methodology to assess vulnerability is based on factorial analysis technique, using the extraction method of the Principal Component Analysis with Varimax rotation. The variables set is constructed on the basis of literature study and it reflects the multiple dimensions of social vulnerability.

Results show that historic centers and areas south of the territorial context have the high levels of social vulnerability. The study also identifies the driving components contributing to the overall vulnerability. The objective of this study is also to establish criteria for the assessment of vulnerability index and evaluate the spatial variation of the same index according to the different variables adopted. Nevertheless, vulnerability is a complex concept and for a better definition and quantification it’s necessary to evaluate also physical and climatic components of the territory.

The vulnerability maps provide useful territorial information that can support policy-makers for planning and emergency management. A good vulnerability assessment, can provide an indication of the housing areas that need

development and humanitarian aids and can provide guidance for better preparedness, response and mitigation strategies.

Acknowledgements

The authors express their gratefulness to the Regional Civil Protection, for availability of data and for technical support for this study.

References

1. Cardona, O.D., The need for rethinking the concepts of vulnerability and risk from a holistic perspective: A necessary review and criticism for effective risk management, in *Mapping Vulnerability: Disasters, Development and People*. 2013. p. 37-51.
2. Ciurean, R.L., D. Schroter, and G. T., *Conceptual Frameworks of Vulnerability Assessments for Natural Disasters Reduction*. In: Tiefenbacher, J., Ed., *Approaches to Disaster Management—Examining the Implications of Hazards, Emergencies and Disasters*, InTech., 2013.
3. Birkmann, J., Risk and vulnerability indicators at different scales: Applicability, usefulness and policy implications. *Environmental Hazards*, 2007. 7(1): p. 20-31.
4. Adger, W.N. and P.M. Kelly, Social vulnerability and resilience, in *Living with Environmental Change: Social Vulnerability, Adaptation and Resilience in Vietnam*. 2012. p. 19-34.
5. Adger, W.N., Vulnerability. *Global Environmental Change*, 2006. 16(3): p. 268-281.
6. Fekete, A., M. Damm, and J. Birkmann, Scales as a challenge for vulnerability assessment. *Natural Hazards*, 2010. 55(3): p. 729-747.
7. Paul, S., Vulnerability Concepts and its Application in Various Fields: A Review on Geographical Perspective. *Journal of Life and Earth Science (J. Life Earth Sci.)*, 2013. 8: p. 63-81.
8. Hufschmidt, G., A comparative analysis of several vulnerability concepts. *Natural Hazards*, 2011. 58(2): p. 621-643.
9. Cutter, S.L., B.J. Boruff, and W.L. Shirley, Social vulnerability to environmental hazards. *Social Science Quarterly*, 2003. 84(2): p. 242-261.
10. Borden, K., et al., Vulnerability of U.S. Cities to Environmental Hazards. *Journal of Homeland Security and Emergency Management*, 2007. 4.
11. Mavhura, E., B. Manyena, and A.E. Collins, An approach for measuring social vulnerability in context: The case of flood hazards in Muzarabani district, Zimbabwe. *Geoforum*, 2017. 86: p. 103-117.
12. Thomas, R., *Traffic Assignment Techniques*. 1991: Avebury Technical, England.
13. Camargo, P. *Aequilibræ*. 2018; Available from: <http://aequilibrae.com/>
14. Osborne, J.W., *Best Practices in Exploratory Factor Analysis*. 2014: CreateSpace Independent Publishing Platform.

Evolution of Forward Curves in the Heath–Jarrow–Morton Framework by Cubature Method on Wiener Space

Anatoliy Malyarenko¹ and Hossein Nohrouzian¹

Division of Applied Mathematics, Mälardalen University,
Box 883, SE-721 23, Västerås, Sweden
(E-mails: anatoliy.malyarenko@mdh.se, hossein.nohrouzian@mdh.se)

Abstract. The multi-curve extension of Heath–Jarrow–Morton framework is a popular method for pricing interest rate derivatives and overnight indexed swaps in the post-crisis financial market. That is, the set of forward curves is represented as a solution to a boundary value problem for an infinite-dimensional stochastic differential equation. In this paper, we review the post-crisis market proxies for interest rate models. Then, we consider a simple model that belongs to the above framework. This model is driven by a single Wiener process, and we discretize the space of trajectories of its driver by cubature method on Wiener space. After that, we discuss possible methods for numerical solution of the resulting deterministic boundary value problem in the finite dimensional case. Finally, we compare and contrast the obtained numerical solutions of cubature method and classical Monte Carlo simulation.

Keywords: Heath–Jarrow–Morton framework, , forward curves, interest rate derivatives, overnight indexed swaps, cubature method, Monte Carlo simulation.

1 Introduction and outline of the paper

An important class of financial instruments appears when the underlying asset is an *interest rate*. Such financial instruments are called *fixed-income instruments*, and their values depend on the random fluctuation of the underlying interest rates. Fixed-income instruments form the *fixed-income market*.

Moreover, the mathematical models which describe the stochastic dynamic of underlying interest rates in the fixed-income market are called *interest rate (term-structure) models*. Term-structure models are used to price *default-free zero-coupon bonds (bond options)* as well as pricing *interest rate derivatives*. An interest rate derivative's payoff depends on future interest rates and the most important interest rate derivatives are *swap option (swaption)*, *cap* and *floor* options. Future interest rates are called *forward rates* and implied by today's market zero (coupon bonds) rates.

Black's , Heath–Jarrow–Morton (HJM) and LIBOR market model (LMM) are some of the famous term structure models to evaluate forward rates. Knowing forward rates, one can calculate the price of bond options and interest rate derivatives. In our setting, we use the HJM framework and objectives are to:

¹6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain



1. include post-crisis multiplicative spread to evaluate forward (spread) curves,
2. include infinitely many bonds in evolution of forward (spread) curves,
3. use cubature method on Wiener space to calculate stochastic integrals.

The first objective is due to the fact that in the prior-crisis market, LIBOR rates were assumed to be risk-free rates while in the post-crisis market, overnight indexed swaps (OIS) rates are considered to be risk-free rates. The second objective is due to the variety of tradable bonds, e.g., governmental and cooperate bonds in different currencies and with different ratings. The third objective is due to the complexity of calculating high-dimensional stochastic integrals.

The outline of the paper is constructed as follows. In Section 2, we briefly review term-structure models, namely, spot-rate, continuous forward-rate and simple forward-rate models. In Section 3, we compare and contrast prior-crisis and post-crisis interest rate models. This also includes reviewing the meaning of multiplicative spreads, Musiela parameterization and explaining a system of stochastic partial differential equations which models a market with infinitely many assets (bonds). In Section 4, we explain the procedure of solving the system presented in the section before. In Section 5, we use the results of our previous work and evaluate forward curves using both cubature method on (one-dimensional) Wiener space and Monte Carlo simulation. Finally, we close this work by a discussion and future work section.

2 Interest rate (term-structure) models

The term structure models can be divided into three main categories. Namely, *spot-rate models*, *(continuous) forward rate models* and *market (or simple) forward rate models*. Following [12], [16], [8] and [11], let us briefly review the classical term structure models.

2.1 Spot-rate models

In the spot-rate models, the underlying interest rate is the time- t instantaneous spot rate $r(t)$ under the physical probability measure \mathbb{P} , and the *money-market account* $B(t)$ is given in terms of $r(t)$. That is,

$$B(t) = \exp\left(\int_0^t r(u)du\right), \quad 0 \leq t \leq T.$$

For a one-dimensional standard Brownian motion $W = \{W(t)\}_{0 \leq t \leq T}$, the instantaneous spot rate $r(t) = \{r(t)\}_{0 \leq t \leq T}$ is a solution to the following SDE

$$dr(t) = \alpha(t, r(t))dt + \beta(t, r(t))dW(t), \quad (1)$$

where α and β are drift and diffusion functions.

Also, given a filtered probability space $(\Omega, \mathfrak{F}, \mathbb{P}, (\mathfrak{F})_{t \geq 0})$, the fair price p (value) of a financial derivative $X = \{X(t)\}_{0 \leq t \leq T}$ with the payoff function f and under the equivalent martingale probability measure $\mathbb{Q} \sim \mathbb{P}$ is given by

$$p(t) = \mathbb{E}^{\mathbb{Q}} \left[B(t) \frac{f(X(T))}{B(T)} \mid \mathfrak{F}_t \right], \quad 0 \leq t \leq T. \quad (2)$$

The default-free zero-coupon bond P whose time- t fair price is $P(t, T)$ pays one unit of cash at time T . In other words, $f(P(T, T)) = P(T, T) = 1$. Thus, if the underlying asset in Equation (2) is a default-free zero-coupon bond, then the fair price of it becomes

$$p(t) = P(t, T) = \mathbb{E}^{\mathbb{Q}} \left[\exp \left(- \int_t^T r(u) du, \right) \middle| \mathfrak{F}_t \right], \quad 0 \leq t \leq T. \quad (3)$$

Finally, if we let $V(t, r) = P(t, T)$, then using Equation (3) and Feynman–Kač representation formula, the \mathbb{Q} -dynamics of Equation (1) for the short rate r yields to the following so called *term-structure equation*

$$\begin{cases} \frac{\partial}{\partial t} V(t, r) + \alpha(t, r) \frac{\partial}{\partial r} V(t, r) + \frac{1}{2} \beta(t, r) \frac{\partial^2}{\partial r^2} V(t, r) = rV(t, r), \\ V(T, r) = 1. \end{cases}$$

Remark 1. In the spot rate models, the amount $1/B(T)$ is called the *discounting factor* or *numéraire*. Furthermore, for $t \leq T$, the following relations between the default-free zero-coupon bonds and instantaneous spot-rates holds

$$r(t) = - \frac{\partial}{\partial T} \ln P(t, T) \Big|_{T=t}, \quad P(t, T) = \exp \left(- \int_t^T r(u) du \right) = \frac{B(t)}{B(T)}.$$

Affine models According to [16], estimating α and β in Equation (1) can be seen as a *calibration* of the model. However, α represents the drift coefficient in the risk-neutral world, and therefore cannot be estimated from historical data under physical probability measure \mathbb{P} . To solve this problem, one can calibrate the model to the set of today's price of default-free zero-coupon bonds and other liquid interest rate derivatives. In other words, we assume the so called *empirical term structure* which states that the set of initial prices of default-free zero-coupon bond is observable.

If we in Equation (1), let $\alpha(t, r) = \alpha_1(t) + \alpha_2(t)r$ and $\beta^2(t, r) = \beta_1(t) + \beta_2(t)r$ where $\alpha_1, \alpha_2, \beta_1$ and β_2 are deterministic functions, then we obtain *affine models*. Affine models are effective under empirical term structure. The term structure equation can be solved semi-analytically in the terms of (Riccati types) first order differential equations (see also [12, Theorem 15.1]).

2.2 Forward-rate models: continuous rates

Forward rate models can be used to explicitly describe the evolution of the full structure of the interest rates. The Black model proposed by [2] is one of the popular forward rate models. Let us focus however on the classical *HJM framework* developed in [9] and [10]. The state variable of the HJM framework is an infinite-dimensional object: the so called *forward curve* that cannot be described by a finite number of rates and factors. The HJM model (more generally the HJM framework) describes the no-arbitrage conditions which must

be satisfied by a model of forward rate curve. To begin with, let us denote the instantaneous forward rate by $f(t, T) = \{f(t, T) : 0 \leq t \leq T \leq T^*\}$, where T^* is *ultimate maturity* and is for example 20 or 30 years from today. Then, for $t \leq T$ the following relations holds

$$P(t, T) = \exp \left(- \int_t^T f(t, u) du \right), \quad (4)$$

$$f(t, T) = -\partial_T \ln P(t, T), \quad (5)$$

$$r(t) = f(t, t).$$

In the HJM framework, the evolution of forward rate curve satisfies

$$df(t, T) = \alpha(t, T)dt + \boldsymbol{\sigma}^\top(t, T)d\mathbf{W}(t), \quad (6)$$

where \mathbf{W} is a d -dimensional Wiener process, $\alpha(t, T)$ and $\boldsymbol{\sigma}(t, T)$ are respectively drift and volatility structures. The arbitrage-free dynamics of the forward rate curve under risk-neutral probability measure \mathbf{Q} satisfies the following SDE

$$df(t, T) = \left(\boldsymbol{\sigma}^\top(t, T) \int_t^T \boldsymbol{\sigma}(t, u) du \right) dt + \boldsymbol{\sigma}^\top(t, T)d\mathbf{W}(t). \quad (7)$$

In other words, under risk-neutral probability measure \mathbf{Q} , Equation (3) and Equation (4) are *compatible* if and only if the *HJM drift condition* holds, i.e.,

$$\alpha(t, T) = \boldsymbol{\sigma}^\top(t, T) \int_t^T \boldsymbol{\sigma}(t, u) du. \quad (8)$$

2.3 Forward-rate models: simple rates

The simple rate forward-rate models are closely related to the HJM framework with the difference that these models are constructed on the bases of *simple* rather than continuously compounded forward rates. Perhaps the most popular model in this category is the so called “*X interbank offered rate (XIBOR)*” market models, where X usually stands for a capital city name, for example LIBOR stands for London interbank offered rate.

The XIBOR market models are based on observable market rates. Let $0 \leq T_0 < \dots < T_m < T_{m+1}$ be a finite set of maturity or *tenor* (frequency of payments) dates and $\delta_i = T_{i+1} - T_i$ for $i = 0, \dots, m$ be the length of interval between tenor dates (i.e., m is the number of tenors). Denote by $L_i(t)$ the T_i -forward LIBOR rate at time t . Then,

$$L_i(t) = \frac{P(t, T_i) - P(t, T_{i+1})}{\delta_i P(t, T_{i+1})} \equiv \frac{P_i(t) - P_{i+1}(t)}{\delta_i P_{i+1}(t)}.$$

It can be shown that (see [12])

$$1 + \delta_i L_i(t) = \frac{P(t, T_i)}{P(t, T_{i+1})} = \exp \left(\int_{T_i}^{T_{i+1}} f(t, s) ds \right).$$

Thus, we can conclude that the XIBOR rate at time T_i is given by $L_i(T_i) = L(T_i, T_{i+1}) = L(T_i, T_i + \delta_i)$. To derive the SDE of XIBOR rates, we do not use the term “risk-neutral” but we say “spot measure” since in this approach, the bond price is a martingale when it is deflated (rather discounted) by the numéraire asset. That is, the stochastic process describing the forward rates becomes a martingale when we do not discount bond price by the continuously compounded rate, but instead divide it by the numéraire asset (see [8, p. 169]). Suppose the evaluation of the forward-XIBOR rate satisfies the following SDE

$$\frac{dL_i(t)}{L_i(t)} = \alpha_i(t)dt + \boldsymbol{\sigma}_i^\top(t)d\mathbf{W}(t), \quad 0 \leq t \leq T_i, \quad i = 1, \dots, m,$$

where the drift and diffusion structures can depend on time t and the vector of XIBOR rates, i.e., $(L_1(t), \dots, L_m(t))$. Define a function $N(t)$ as a unique integer satisfying $T_{N(t)-1} \leq t < T_{N(t)}$ such that $N(t)$ modifies the index of the next tenor date at time t . Then, the value of the drift term α_i can be determined and the last equation takes the following form

$$\frac{dL_i(t)}{L_i(t)} = \sum_{j=N(t)}^i \frac{\delta_j L_j(t) \boldsymbol{\sigma}_i^\top(t) \boldsymbol{\sigma}_j(t)}{1 + \delta_j L_j(t)} dt + \boldsymbol{\sigma}_i^\top(t) d\mathbf{W}(t),$$

for $0 \leq t \leq T_i$ and $i = 1, \dots, m$.

3 Prior-crisis vs post-crisis interest rate models

To begin with, following [11] let us review the meaning of following terms:

Swap is an exchange by cash flows in the future via a pre-agreed formula.

Interest rate swap is to exchange of a fixed interest rate (e.g., 5%) for a floating interest rate (e.g., 6-month XIBOR) on the same notional principal.

Overnight rates are the rates of overnight lending and borrowing to keep a certain amount of cash called *reserve* in an authorized institute. Overnight rates in the US, UK and euro zone are called federal fund rates, sterling overnight index average (SONIA) and euro overnight index average (EONIA).

Overnight indexed swap (OIS) is a swap where a fixed rate for a specific period is exchanged for the geometric average of overnight rates during the same period. Mathematically, in an OIS with the rate $K^{\text{OIS}}(t, T_m)$ and ℓ_i business days between $[T_{i-1}, T_i]$, i.e.,

$$T_{i-1} = t_0 < t_1 < \dots < t_{\ell_i} = T_i, \quad i = 1, \dots, m,$$

one party pays $(T_i - T_{i-1})K$ and the other party pays $(T_i - T_{i-1}) \cdot \bar{L}_{T_{i-1}}(T_{i-1}, T_i)$, where

$$\bar{L}_{T_{i-1}}(T_{i-1}, T_i) = \frac{1}{T_i - T_{i-1}} \left(\prod_{j=1}^{\ell_i} [1 + (t_j - t_{j-1})L_{t_j}(t_j - t_{j-1})] - 1 \right).$$

In fact, $K^{\text{OIS}}(t, T_m)$, is the value of K that makes the OIS value equal to zero. To estimate the rate K one can construct the OIS rate curve using the bootstrap techniques.

Forward rate agreement (FRA) is an agreement which specifies a particular underlying interest rate will be applied to a specific principal amount for a pre-agreed period of time in the future. Mathematically, a FRA starting at T_i , with maturity $T_i + \delta_i$, fixed rate K_i and notional amount N_i , is a contract which pays

$$N_i \delta_i (L_{T_i}(T_i, T_i + \delta_i) - K_i).$$

FRA rate $L_t(T_i, T_i + \delta_i)$, is the rate K fixed at time t such that the value of the contract becomes 0.

Prior to the financial crisis started in 2007, XIBOR rates were the market proxies for considering the term *risk-free* interest rate. In the post-crisis financial market however, the OIS rates are the market proxies for the risk-free interest rate. The reason of this switch is due to the fact that XIBOR rate are not completely risk-free.

Existence of the post-crisis spreads is due to the fact that FRA rates are typically greater than simply compounded OIS forward rates since XIBOR panel is periodically updated to include only creditworthy banks. Mathematically, in the post-crisis financial market, we have

$$L_t^{\text{OIS}}(T_i, T_{i+1}) := \frac{P(t, T_i) - P(t, T_{i+1})}{\delta_i P(t, T_{i+1})} < L_t(T_i, T_{i+1}), \quad (9)$$

where $\delta_i = T_{i+1} - T_i$ and $i = 1, \dots, m$.

Given $P(0, T_i)$ and $L_0(T_i, T_{i+1})$ for $T_i \geq 0$ and $i = 1, \dots, m$, we have constructed an arbitrage-free large market model consisting of infinitely many OIS zero coupon bonds and FRA contracts with all possible maturities (see [15]).

Now, our objective is to price the interest rate derivatives in the post-crisis market using modern Monte Carlo simulation by cubature method on Wiener space (see [13] and [14]).

3.1 Multiplicative spreads

Inequality (9) helps us to define the *multiplicative spread*. Following [5], on the one hand the *spot multiplicative spread* $S^{\delta_i}(t, t)$ is given by

$$S^{\delta_i}(t, t) := \frac{1 + \delta_i L_t(t, t + \delta_i)}{1 + \delta_i L_t^{\text{OIS}}(t, t + \delta_i)}.$$

The numerator of the above equation can be determined by the quoted XIBOR rate in the market, where as the denominator can be determined using the quoted OIS rates in the market and bootstrapping techniques. Thus, the multiplicative spread can be inferred from the market data.

On the other hand, the *forward multiplicative spread* $S^{\delta_i}(t, T_m)$ is given by

$$S^{\delta_i}(t, T) := \frac{1 + \delta_i L_t(T, T + \delta_i)}{1 + \delta_i L_t^{\text{OIS}}(T, T + \delta_i)} = (1 + \delta_i L_t(T, T + \delta_i)) \frac{P(t, T + \delta_i)}{P(t, T)},$$

for $0 \leq t \leq T \leq T^*$ and $i = 1, \dots, m$ (T is maturity and T^* is called ultimate maturity). Now, we can use Equation (4) in the HJM framework to calculate the given fraction, i.e., default-free coupon bonds, in the above equation.

3.2 Musiela parameterization

To include the forward multiplicative spread in Equation (7), we need to use the so called Musiela parameterization, since for different times t , the function $f(t, T)$ have different domains of definition. Let $s = T - t$ be the time to maturity, then the Musiela parameterization is given by

$$\theta_t(s) = f(t, t + s).$$

Now, Equation (7) can be re-written in the form of θ -dynamics and is called *Musiela equation*. That is, the following boundary value problem

$$\begin{cases} d\theta_t(s) = \left(\frac{\partial}{\partial s} \theta_t(s) + \tilde{\sigma}^\top(t, s) \int_0^s \tilde{\sigma}(t, u) du \right) dt + \tilde{\sigma}^\top(t, s) d\mathbf{W}(t), \\ \theta_0(s) = \theta^o(s), \end{cases} \quad (10)$$

where $\tilde{\sigma}(t, s) = \sigma(t, t + s)$.

3.3 A market model with infinitely many assets (bonds)

To include infinitely many assets (bonds), and multiplicative spread in the post-crisis forward rates models discussed up to Equation (10), we will use a framework which yield to the following system of stochastic partial differential equations (see [15])

$$\begin{cases} d\theta_t^i = \left(\frac{d}{ds} \theta_t^i + \kappa^i(\theta_t) \right) dt + \zeta_i(\theta_t) dW_t, \\ \theta_0^i = \eta_0^i. \end{cases} \quad (11)$$

Here we assume that for each t , the function $\theta_t(s) = (\theta_t^0(s), \dots, \theta_t^m(s))^\top$ belongs to the *Filipović space* H_{m+1}^λ , $\lambda \in \mathbb{R}^+$, the completion of the set of all differentiable functions $\mathbf{r}: [0, \infty) \rightarrow \mathbb{R}^{m+1}$ satisfying the condition

$$\|r\|^2 = \int_0^\infty \|\mathbf{r}(s)\|_{\mathbb{R}^{m+1}}^2 e^{-\lambda s} ds + \int_0^\infty \|\mathbf{r}'(s)\|_{\mathbb{R}^{m+1}}^2 e^{-\lambda s} ds < \infty$$

with respect to the norm $\|r\|$. Moreover, we assume that the drift functions κ^i map H_{m+1}^λ to H_1^λ , the diffusion functions ζ^i map H_{m+1}^λ to $L(\mathbb{R}^d, H_1^\lambda)$, where $L(\cdot, \cdot)$ is the space of linear operators. d/ds is the infinitesimal operator, $\eta_t^0(T) = \partial_T \ln P(t, T)$ and $\eta_t^i(T) = \partial_T \ln S^{\delta_i}(t, T)$ are the forward spread curves, and $\theta_t(s) = \boldsymbol{\eta}_t(t + s)$ is Musiela parameterization.

4 A framework in the post-crisis market

Consider the following “building blocks” for a market model.

- A stochastic basis $(\Omega, \mathfrak{F}, (\mathfrak{F}_t)_{t \geq 0}, \mathbf{P}^*)$;
- the real numbers $0 = u_0 < u_1 < u_2 < \dots < u_m$, where m is the number of tenors;

- the functions $\eta_0^0(t), \eta_0^1(t), \dots, \eta_0^m(t)$, called the *initial forward rates*;
- the random fields $\sigma_t^0(T), \sigma_t^1(T), \dots, \sigma_t^m(T)$, defined on the triangle

$$\{(t, T) \in \mathbb{R}^2: 0 \leq t \leq T < \infty\},$$

called the *forward rate volatilities*.

Our blocks constitute a subset of the set of building blocks given by [3, Definition 4.1]. The block \hat{Y} is missing in our list, the reason for that will be explained below.

Let $(\beta_t)_{t \geq 0}$ be a predictable process, integrable with respect to the Wiener process W_t . According to [3, Definition 3.2], a predictable process $(\Psi_t^W(\beta_t))_{t \geq 0}$ is called the *local exponent* of W at β if the stochastic process

$$\left(\exp \left(\int_0^t \beta_s dW_s - \int_0^t \Psi_s^W(\beta_s) ds \right) \right)_{t \geq 0}$$

is a local martingale. By [3, Proposition 3.3], we can calculate the local exponent if we know the differential characteristics of the Wiener process W_t . We refer to [6] for definition of differential characteristics, and use Proposition 4.6 of the above cited book instead. That is, if a semi-martingale X is a Lévy process, then the differential characteristics of X equal to its Lévy–Khintchine triplet. In particular, the triplet of the Brownian motion is $(0, 1, 0)$. By [3, Equation (3.4)], we obtain

$$\Psi_t^W(\beta_t) = \frac{1}{2}(\beta_t)^2.$$

Now, we put the initial forward rates $\eta_0^0(t), \eta_0^1(t), \dots, \eta_0^m(t)$ “in motion”. The time evolution of forward rates is described by the following system of infinite-dimensional SDEs:

$$\eta_t^i(T) = \eta_0^i(T) + \int_0^t \alpha_s^i(T) ds + \int_0^t \sigma_s^i(T) dW_s, \quad 0 \leq i \leq m, \quad (12)$$

where $\alpha_t^0(T), \dots, \alpha_t^m(T)$ are random fields satisfying the *drift condition*

$$\int_t^T \alpha_t^i(u) du = -\Psi_t^W(\Sigma_t^i(T) - \Sigma_t^0(T)) + \Psi_t^W(-\Sigma_t^0(T)), \quad 0 \leq i \leq m,$$

with

$$\Sigma_t^i(T) = \int_t^T \sigma_t^i(u) du, \quad 0 \leq i \leq m. \quad (13)$$

With our convention $X_t = W_t$, the drift condition becomes

$$\int_t^T \alpha_t^i(u) du = -\frac{1}{2}(\Sigma_t^i(T))^2 + \Sigma_t^i(T)\Sigma_t^0(T).$$

Differentiating both hand sides with respect to T and taking into account Equation (13), we obtain

$$\alpha_t^i(T) = -\sigma_t^i(T)(\Sigma_t^i(T) - \Sigma_t^0(T)) + \sigma_t^0(T)\Sigma_t^i(T). \quad (14)$$

Let us introduce the following notations

$$\boldsymbol{\eta}_t(T) = (\eta_t^0(T), \dots, \eta_t^m(T))^\top, \quad \boldsymbol{\theta}_t(s) = \boldsymbol{\eta}_t(t+s),$$

and assume that the volatility structures have the form

$$\sigma_t^i(T) = \zeta_i(\boldsymbol{\theta}_t)(T-t), \quad 0 \leq t \leq T < \infty, \quad 0 \leq i \leq m,$$

where the Musiela-parametrised stochastic process $(\boldsymbol{\theta}_t)_{t \geq 0}$ takes values in a certain Filipović space H_{m+1}^λ of forward curves. Using Equation (13), we obtain

$$\Sigma_t^i(T) = \frac{1}{2}(T-t)^2 \zeta_i(\boldsymbol{\theta}_t),$$

and

$$\alpha_t^i(T) = \kappa_i(\boldsymbol{\theta}_t)(T-t), \quad 0 \leq t \leq T < \infty, \quad 0 \leq i \leq m,$$

where

$$\kappa_i(h)(s) = -\zeta^i(\mathbf{h})(s)[Z^i(\mathbf{h})(s) - Z^0(\mathbf{h})(s)] + \zeta^0(\mathbf{h})(s)Z^i(\mathbf{h})(s)$$

with

$$Z^i(\mathbf{h})(s) = \int_0^s \zeta^i(\mathbf{h})(u) du.$$

Here \mathbf{h} is a function from the space H_{m+1}^λ . The system (12) takes the form

$$\theta_t^i = S_t \eta_i^0 + \int_0^t S_{t-s} \kappa_i(\boldsymbol{\theta}_s) ds + \int_0^t S_{t-s} \zeta_i(\boldsymbol{\theta}_s) dW_s,$$

where $(S_t)_{t \geq 0}$ is the shift semigroup. The operator S_t acts on the function $\mathbf{h}(s) \in H_{m+1}^\lambda$ by

$$(S_t \mathbf{h})(s) = \mathbf{h}(t+s).$$

By definition, a stochastic process $\boldsymbol{\theta}$ satisfying this equation, is a *mild solution* to the following system

$$d\theta_t^i = \left(\frac{d}{ds} \theta_t^i + \kappa^i(\boldsymbol{\theta}_t) \right) dt + \zeta_i(\boldsymbol{\theta}_t) dW_t. \quad (15)$$

Now, we construct a stochastic process Y_t using [3, Remark 3.19]. Namely, solve the system (15), return back to $\boldsymbol{\eta}_t(T)$ by $\boldsymbol{\eta}_t(T) = \boldsymbol{\theta}_t(T-t)$, define the stochastic process q_s by the consistency condition $u_i q_t = \eta_t^i(t)$, and put

$$Y_t = \int_0^t q_s ds.$$

In this way, we constructed the process Y_t without using the building block \hat{Y} .

Finally, our market model consists of:

- A stochastic basis $(\Omega, \mathfrak{F}, (\mathfrak{F}_t)_{t \geq 0}, \mathbb{P}^*)$;

- the real numbers $0 = u_0 < u_1 < u_2 < \dots < u_m$, where m is the number of tenors;
- the initial forward rates $\eta_0^0(t), \eta_0^1(t), \dots, \eta_0^m(t)$;
- the forward rate volatilities $\sigma_t^0(T), \sigma_t^1(T), \dots, \sigma_t^m(T)$;
- the stochastic processes W_t, Y_t , and

$$\begin{aligned}
 B(t) &= \exp\left(\int_0^t \eta_s^0(s) ds\right), \\
 P(t, T) &= \exp\left(-\int_t^T \eta_t^0(s) ds\right), \\
 S^{\delta_i}(t, T) &= \exp\left(u_i Y_t + \int_t^T \eta_t^i(u) du\right),
 \end{aligned}$$

where in this model, $B(t)$ is the bank account, $P(t, T)$ are the bond prices and $S^{\delta_i}(t, T)$ are multiplicative spreads between normalized FRA rates and simply compounded OIS forward rates. The results of [3, Section 4] guarantee that in the constructed model, for a fixed maturity T , the discounted bond prices $P(t, T)/B(t)$ and the spreads $S^{\delta_i}(t, T)$ are martingales.

5 Evolution of forward curves by cubature method

In this section, we will compare and contrast evolution of forward rate curves in the classical HJM model using the classical Monte Carlo simulation and our cubature formulae on Wiener space introduced in [13] and [14]. We focus on a one-dimensional Wiener process and only a set of initial forward rate data.

To begin with, we could find two set of analytical solutions to represent cubature formula of degree 5 in [13]. Furthermore, we found several numerical solutions to represent cubature formula of degree 7 in [14]. Finally, we tested both cubature formulae of degree 5 and 7 in [14] versus Black-Scholes pricing formulae. Now, we will proceed in the following steps (subsections).

5.1 Trajectories of cubature formulae of degree 5 on Wiener space

In this part, we will present the explicit solutions to the cubature formula of degree 5 and proceed further. That is, the trajectories ω_k , $k = 1, 2, 3$ of cubature formula of degree 5 on Wiener space can be generalized by

$$\omega_k(t_\ell) = 3\theta_{k,\ell}(t_\ell - t_{\ell-1}) + \omega_k(t_{\ell-1}), \quad \ell = 1, 2, 3, \quad \omega_k(0) = 0, \quad (16)$$

where for $\ell = 1, 2, 3$ we have, $0 = t_0 < t_1 < t_2 < t_3 = 1$ and $t_\ell - t_{\ell-1} = 1/3$.

Table 1 summarizes the information needed to implement Equation (16) with λ_k denoting the weight of k -th trajectory. Figure 1 depicts the trajectories.

k	λ_k	$\theta_{k,1} = \theta_{k,3}$	$\theta_{k,2}$	$\theta_{k,3} = \theta_{k,1}$
1	1/6	$(-2\sqrt{3} \mp \sqrt{6})/6$	$(-\sqrt{3} \pm \sqrt{6})/3$	$(-2\sqrt{3} \mp \sqrt{6})/6$
2	2/3	$\pm\sqrt{6}/6$	$\mp\sqrt{6}/3$	$\pm\sqrt{6}/6$
3	1/6	$(2\sqrt{3} \pm \sqrt{6})/6$	$(\sqrt{3} \mp \sqrt{6})/3$	$(2\sqrt{3} \pm \sqrt{6})/6$

Table 1: Information for cubature formulae of degree 5.

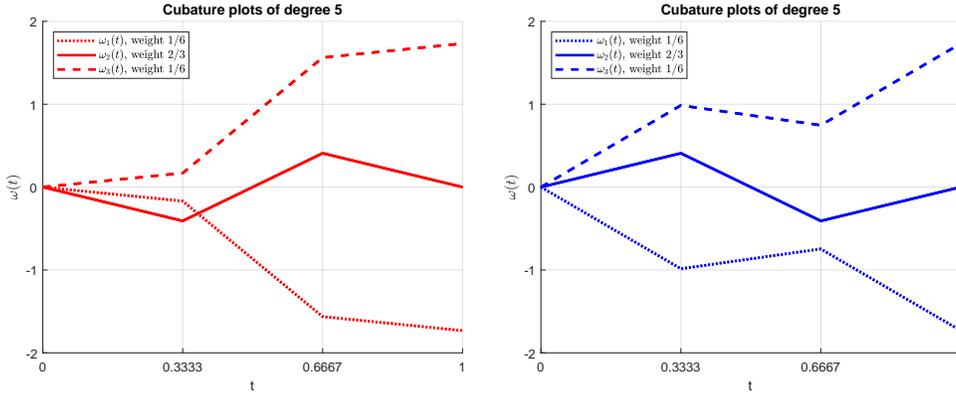


Fig. 1: Two cubature formulae of degree 5.

5.2 Itô stochastic integral versus Stratonovich stochastic integral

Let $\{\mathbf{X}(t), t \geq 0\}$ be solution to the following n -dimensional Itô SDE

$$d\mathbf{X}(t) = \boldsymbol{\alpha}(t, \mathbf{X}(t))dt + \boldsymbol{\sigma}(t, \mathbf{X}(t))d\mathbf{W}(t), \quad (17)$$

where $\boldsymbol{\alpha}$ is a n -dimensional drift function, $\boldsymbol{\sigma}$ is an $n \times m$ -dimensional diffusion function and $\{\mathbf{W}(t), t \geq 0\}$ is the standard m -dimensional Wiener process. We re-write Equation (17) in its equivalent integral form. That is,

$$X_i(t) = X_i(0) + \int_0^t \mu_i(u, \mathbf{X}(u))du + \sum_{j=1}^m \int_0^t \sigma_{ij}(u, \mathbf{X}(u))dW_u^j. \quad (18)$$

The Stratonovich representation of SDEs in Equation (17) is given by [18]

$$d\mathbf{X}_t = \tilde{\boldsymbol{\alpha}}(t, \mathbf{X}_t)dt + \boldsymbol{\sigma}(t, \mathbf{X}_t) \circ d\mathbf{W}_t, \quad (19)$$

where

$$\tilde{\alpha}_i(t, \mathbf{x}) = \alpha_i(t, \mathbf{x}) - \frac{1}{2} \sum_{j=1}^m \sum_{k=1}^n \frac{\partial \sigma_{ij}(t, \mathbf{x})}{\partial x_k} \sigma_{kj}(t, \mathbf{x})$$

is the so called *Stratonovich correction*. Now, we re-write Equation (19) in its equivalent integral form. That is,

$$X_t^i = X_0^i + \int_0^t \tilde{\alpha}_i(u, \mathbf{X}_u)du + \sum_{j=1}^m \int_0^t \sigma_{ij}(u, \mathbf{X}_u) \circ dW_u^j. \quad (20)$$

Equation (6) is an Itô SDE. We restrict our study to the case of one-dimensional Wiener process and we let volatility σ be either constant or a function of time and not a function of forward rate f . This yields to the simple case where α in Equation (7) and $\tilde{\alpha}$ in Stratonovich correction described above are equal. Thus, the Stratonovich form of Equation (7) becomes

$$df(t, T) = \left(\sigma(t, T) \int_t^T \sigma(t, u) du \right) dt + \sigma(t, T) \circ dW(t).$$

Considering the technicality presented in [14], the above equation can be written as

$$df_k(t, T) = \left(\sigma(t, T) \int_t^T \sigma(t, u) du \right) dt + \sigma(t, T) \omega_k(t, T), \quad (21)$$

where ω_k , ($1 \leq k \leq N$) is the k -th possible trajectory of $N \in \mathbb{Z}^+$. We saw that, for cubature formula of degree 5, $N = 3$.

5.3 Evolution of forward curve (implementation and comparison)

To begin with, without losing generality and for simplicity, we used the instantaneous forward rates available in *European Central Bank (ECB)* on June 24th 2020. The forward rates were already calculated based on triple A rated bonds and are given in Table 2, where we set the forward rate for $i = 0$ equal to the forward rate for $i = 1$. Moreover, we put already the results of estimations $\hat{f}(0, t_j)$ for $j \geq i$. This estimation will be explained later. Now, following [8] we consider two simple cases where σ can be constant or exponential.

Constant σ : If $\sigma(t, T) \equiv \sigma$, then Equation (8) gives

$$\alpha(t, T) = \sigma \int_t^T \sigma du = \sigma^2(T - t).$$

Clearly, for $i = 1, \dots, m$ and $j = i, \dots, m$ the discretized version becomes

$$\alpha(t_{i-1}, t_j) = \sigma^2(t_j - t_{i-1}).$$

Exponential σ : If $\sigma(t, T) = \sigma \exp(-\rho(T-t))$ for some constant σ and $\rho \in \mathbb{R}^+$, then Equation (8) gives

$$\alpha(t, T) = \sigma e^{-\rho(T-t)} \int_t^T \sigma e^{-\rho(u-t)} du = \frac{\sigma^2}{\rho} \left(e^{-2\rho(T-t)} - e^{-\rho(T-t)} \right).$$

Clearly, for $i = 1, \dots, m$ and $j = i, \dots, m$ the discretized version becomes

$$\alpha(t_{i-1}, t_j) = \frac{\sigma^2}{\rho} \left[\exp(-2\rho(t_j - t_{i-1})) - \exp(-\rho(t_j - t_{i-1})) \right].$$

i	0	1	2	3	4	5	6	7	8
t_i	0	0.25	0.50	0.75	1.00	2.00	3.00	4.00	5.00
$f(t_i, t_j)$	-0.586082	-0.586082	-0.611109	-0.637762	-0.663059	-0.719486	-0.686235	-0.583716	-0.445696
$\hat{f}(0, t_j)$	-0.5682	-0.6123	-0.6476	-0.6747	-0.7074	-0.6931	-0.6163	-0.5030	-0.3731
i	9	10	11	12	13	14	15	16	17
t_i	6.00	7.00	8.00	9.00	10.00	11.00	12.00	13.00	14.00
$f(t_i, t_j)$	-0.2998	-0.163472	-0.045439	0.051552	0.128264	0.187191	0.231425	0.264013	0.287654
$\hat{f}(0, t_j)$	-0.2411	-0.1170	-0.0069	0.0858	0.1601	0.2167	0.2576	0.2853	0.3030
i	18	19	20	21	22	23	24	25	26
t_i	15.00	16.00	17.00	18.00	19.00	20.00	21.00	22.00	23.00
$f(t_i, t_j)$	0.304581	0.316567	0.324974	0.330822	0.334859	0.337629	0.339518	0.340799	0.341665
$\hat{f}(0, t_j)$	0.3134	0.3194	0.3231	0.3263	0.3298	0.3341	0.3389	0.3436	0.3470
i	27	28	29	30	31	32	33	-	-
t_i	24.00	25.00	26.00	27.00	28.00	29.00	30.00	-	-
$f(t_i, t_j)$	0.342246	0.342636	0.342896	0.343069	0.343184	0.34326	0.34331	-	-
$\hat{f}(0, t_j)$	0.3484	0.3472	0.3436	0.3390	0.3366	0.3419	0.3419	-	-

Table 2: ECB instantaneous forward rates %.

The next step is to discretize the drift term. Let $0 = t_0 < t_1 < \dots < t_m$. Denote by $\hat{f}(t_i, t_j)$ the time- t_i of discretized forward rate for maturity t_j where $j \geq i$, then the discretized form of Equation (4) is given by [8]

$$\hat{P}(t_i, t_j) = \exp \left(- \sum_{n=i}^{j-1} \hat{f}(t_i, t_n) [t_{n+1} - t_n] \right).$$

In order to have $\hat{P}(0, t_j) = P(0, t_j)$, we compare the above equation with Equation (5) which yields to

$$\sum_{n=0}^{j-1} \hat{f}(0, t_n) [t_{n+1} - t_n] = \int_0^{t_j} f(0, u) du.$$

In other words, $\hat{P}(0, t_j) = P(0, t_j)$ if we have

$$\hat{f}(0, t_n) = \frac{1}{t_{n+1} - t_n} \int_{t_n}^{t_{n+1}} f(0, u) du, \quad n = 0, 1, \dots, m-1.$$

We estimated $\hat{f}(0, t_n)$ for $n = 0, 1, \dots, m-1$ in MATLAB[®]. That is, we used the data given in Table 2 and used MATLAB[®] syntax `polyfit` to find the coefficients of a polynomial, which fits the given data for instantaneous forward rates in a least-square sense. By trial and error we were satisfied when the degree of the described polynomial was 6. Then, we substituted the obtained polynomial as the integrand in the above equation. The results of

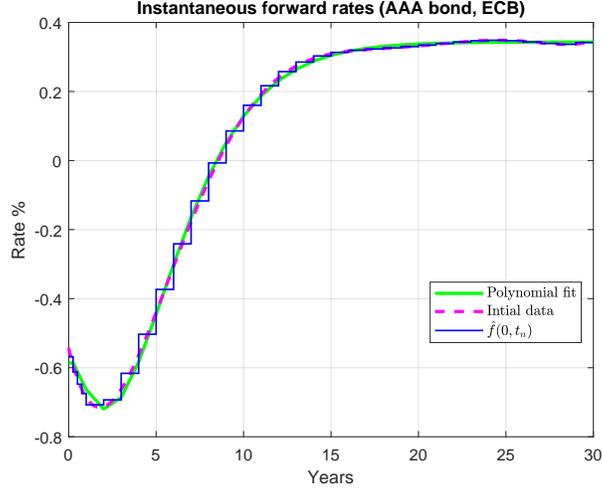


Fig. 2: Initial forward curves.

$\hat{f}(0, t_j)$ were presented in Table 2. Figure 2 illustrates these data in different graphical forms.

In the last step we will have following discretized forward evolution formulas for classical Monte Carlo and cubature methods.

Classical Monte Carlo method: For $i = 1, \dots, m$, $j = i, \dots, m$

$$\hat{f}(t_i, t_j) = \hat{f}(t_{i-1}, t_j) + \hat{\alpha}(t_{i-1}, t_j)[t_i - t_{i-1}] + \hat{\sigma}(t_{i-1}, t_j)\sqrt{t_i - t_{i-1}} Z_i,$$

where $Z_i \sim N(0, 1)$ and $\hat{\alpha}$ and $\hat{\sigma}$ depends on our choice of volatility (either constant or exponential). We will simulate the above equation for 100,000 times in MATLAB[®] and then take the average of the simulated forward rates.

Cubature method: For $i = 1, \dots, m$ and $j = i, \dots, m$ and $k = 1, 2, 3$

$$\hat{f}_k(t_i, t_j) = \hat{f}(t_{i-1}, t_j) + \hat{\alpha}(t_{i-1}, t_j)[t_i - t_{i-1}] + \hat{\sigma}(t_{i-1}, t_j) \omega_k(t_i),$$

where we divide every partition $t_i - t_{i-1}$ by 3 and run Equation (16). Then, for given value of λ_k in Table 1, we have

$$\hat{f}(t_i, t_j) = \sum_{k=1}^3 \lambda_k \hat{f}_k(t_i, t_j).$$

Similar to the Monte Carlo method $\hat{\alpha}$ and $\hat{\sigma}$ depends on our choice of volatility (either constant or exponential). Unlike classical Monte Carlo, in the cubature method one needs to perform the above calculations only once.

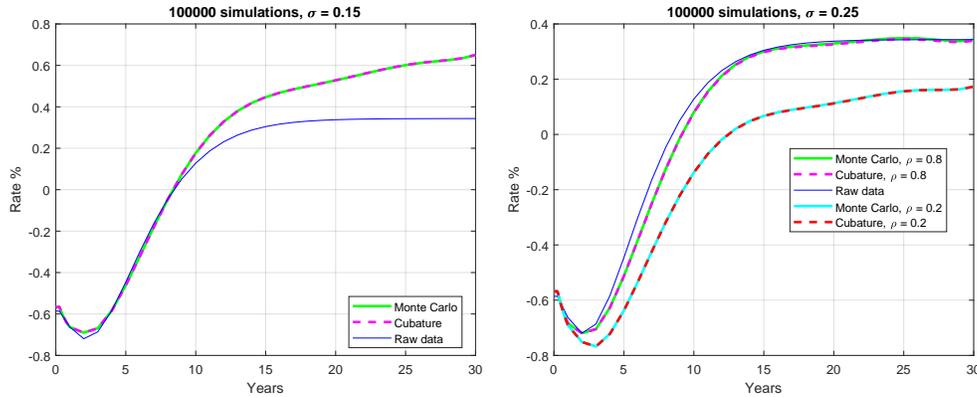


Fig. 3: Evolution of forward curve via Monte Carlo and cubature method.

Comparison: Figure 3 depicts the average result of a 100,000 Monte Carlo simulation trail versus the single performance of cubature formula of degree 5 for constant and exponential σ . In the left subfigure, $\sigma = 0.15$ and the norm error of Monte Carlo (MC) versus cubature (cub) method is $\|\hat{\mathbf{f}}_{\text{MC}} - \hat{\mathbf{f}}_{\text{cub}}\| = 0.0022$. In the right subfigure, $\sigma = 0.25$ and the norm error of Monte Carlo (MC) versus cubature (cub) method for $\rho = 0.8$ is $\|\hat{\mathbf{f}}_{\text{MC}} - \hat{\mathbf{f}}_{\text{cub}}\| = 0.0034$ and for $\rho = 0.2$ is $\|\hat{\mathbf{f}}_{\text{MC}} - \hat{\mathbf{f}}_{\text{cub}}\| = 0.0069$. The errors would vary with small amounts for different classical Monte Carlo simulation trails.

6 Discussion and future works

In this paper, we reviewed and discussed prior and post-crisis term structure models. We extended the discussion and developed a post-crisis model which can include infinitely many bonds and consequently source of uncertainties. Finally, in the HJM framework, we tested a cubature method on one-dimensional Wiener space versus classical Monte Carlo simulation.

In the constructed model in Section 4, the financial products with optionality features are priced by risk-neutral valuation (see also formulae in [3, Subsection 5.2.2]). In our future works, we would like to calculate the expected values of financial instruments' payoff functions in our model using the cubature methods introduced in our previous paper [14] and tested in Section 5. In this procedure, the main obstacles will be:

1. Re-writing Equation (11) in its Stratonovich form, i.e., for infinite dimensional case.
2. Choosing a proper volatility structure and calibrating it to the market data.

A reference to deal with the first obstacle is [1, Section 3.] For dealing with the second obstacle we refer the readers to [5], where the authors used affine models to price interest rate derivatives. In our views, the volatility structures considered and developed in [17] can also be used and be implemented in our model to construct the volatility structure we aim for.

Bibliography

- [1] C. Bayer and J. Teichmann. Cubature on Wiener space in infinite dimension. *Proceedings of The Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, **464**, 2097, 2493–2516, 2008.
- [2] F. Black. The pricing of commodity contracts. *J. Financ. Econom.*, **3**, **1**, 167–169, 1976.
- [3] C. Cuchiero, C. Fontana and A. Gnoatto. A general HJM framework for multiple yield curve modelling. *Finance Stoch.*, **20**, **2**, 267–320, 2016.
- [4] C. Cuchiero, I. Klein and J. Teichmann. A new perspective on the fundamental theorem of asset pricing for large financial markets. *Theory Probab. Appl.*, **60**, **4**, 561–579, 2016.
- [5] C. Cuchiero, C. Fontana and A. Gnoatto. Affine multiple yield curve models. *Math. Finance*, **29**, **2**, 568–611, 2019.
- [6] E. Eberlein and J. Kallsen. *Mathematical finance*, Springer, Cham, 2019.
- [7] D. Filipović. *Consistency problems for Heath–Jarrow–Morton interest rate models*, vol. 1760 of *Lect. Notes Math.*, Springer, Berlin, 2001.
- [8] P. Glasserman. *Monte Carlo methods in financial engineering*, vol. 53 of *Applications of Mathematics (New York)*, Springer, New York, 2004.
- [9] D. Heath, R. Jarrow and A. Morton. Bond pricing and the term structure of interest rates: a discrete time approximation. *J. Financ. Quantitat. Anal.*, **25**, **4**, 419–440, 1990.
- [10] D. Heath, R. Jarrow and A. Morton. Bond pricing and the term structure of interest rates: a new methodology for contingent claims valuation. *Econometrica*, **60**, **1**, 77–105, 1992.
- [11] J. Hull. *Options, futures, and other derivatives*, 10th ed., Pearson, London, 2017.
- [12] M. Kijima. *Stochastic processes with applications to finance*, 2nd ed., CRC Press, Boca Raton, FL, 2013.
- [13] A. Malyarenko, H. Nohrouzian and S. Silvestrov. An algebraic method for pricing financial instruments on post-crisis market. In *Algebraic Structures and Applications. SPAS 2017*, vol. 317 of *Springer Proc. Math. Stat.*, chapter 37, 839–856.
- [14] H. Nohrouzian and A. Malyarenko. Testing cubature formulae on Wiener space vs explicit pricing formulae. In *Stochastic processes, statistical methods and engineering mathematics. SPAS 2019*, vol. yyy of *Springer Proc. Math. Stat.*, chapter zz, xxx–xxx.
- [15] H. Nohrouzian, Y. Ni and A. Malyarenko. An arbitrage-free large market model for forward spread curves. *Applied Modeling Techniques and Data Analysis*, Wiley, 2021, xxx–xxx.
- [16] A. Pascucci. *PDE and martingale methods in option pricing*, Springer, Milan, 2011.
- [17] R. M. Verschuren. Stochastic interest rate modelling using a single or multiple curves: an empirical performance analysis of the Lévy forward price model. In *Quantitative Finance*, , 1–26, 2020.
- [18] B. Øksendal. *Stochastic Differential Equations: An Introduction with Applications*, Springer, Berlin, 2013.

Quantum approach for similarity evaluation in LSA vector space models

Alejandro Martinez-Mingo¹, Guillermo Jorge-Botana², Ricardo Olmos Albacete¹, and Jose Angel Martinez-Huertas³

¹ Department of Social Psychology and Methodology, Calle Ivan Pavlov, 6, Universidad Autonoma de Madrid (UAM), 28049, Madrid, Spain

(E-mail: alejandromartinez@uam.es)

² Developmental and Educational Psychology Department. Universidad Nacional de Educacion a Distancia (UNED), Juan del Rosal, n^o 10, 28023, Madrid, Spain

³ Department of Cognitive Psychology, Calle Ivan Pavlov, 6, Universidad Autonoma de Madrid (UAM), 28049, Madrid, Spain

Abstract. Studies on similarity between concepts have been one of the most prolific research fields. Traditionally, a geometric approach has been used in which we understand the elements as vectors that allow us to evaluate distances through different methods in an n-dimensional space (Duran et al.[1]). However, there are several critics of this approach due to certain assumptions that are not fulfilled empirically. Tversky's [2] studies already demonstrated the inconsistency of these properties, reporting violations of the assumptions of asymmetry, triangular inequality, and diagnosticity. In this study, we propose a method to put quantum similarity model or QSM (Pothos and Busemeyer [3], Pothos et al. [4], Duran [1]) into a data-driven model rails: Latent Semantic Analysis (LSA). This method was informally suggested in (Jorge-Botana et al. [34]) but in this study the idea are formally expanded. This allows to calculate QSM similarities between semantically identified concepts, that is, between subspaces with an assigned dimensionality and a basis with meaningful vectors. A preliminary use of this method confirms the hypothesis proposed by Tversky [2] in 1997, being possible to model these violations through it.

Keywords: QSM, VSM, Similarities, Human Judgement.

1 Introduction

Human judgments and the cognitive processes that lead to them have been proven to be of enormous importance within the scientific community. In this way, tremendous efforts have been made to model human behavior in decision making tasks due to the great complexity of finding suitable models to account for cognitive biases in limited rational situations.

From the area of social psychology, numerous attempts have been made to explain these biases using heuristics (e.g. representativeness or availability). Authors such as Tversky and Kahneman (Tversky [2], Tversky and Kahneman [5], Tversky and Kahneman [6], Tversky and Kahneman [7], Kahneman [8], and

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



more) have created a whole research field based on the explanation of biases in limited rationality situations. These heuristics suppose a great advance in the social psychology and consumer economy world, allowing the scientific community to explain in a more precise way the cognitive biases produced in decision making tasks. However, it is still a challenge for psychological science to develop mathematical models that are capable of explaining these biases.

This is why the use of Vector Space Models (VSM) such as Latent Semantic Analysis (LSA) (Landauer and Dummais [9]) has been one of the most successful approaches to represent some heuristics such as representativeness (Bhatia [10]). The biggest problem, however, is that many of these biases imply the violation of certain laws of classical probability theory and Boolean logic (Pothos and Busemeyer [3]), and this makes it very complicated to create models that explain them in a natural way.

Over the past decade, a number of proposals have emerged to model decision making from the quantum probability theory perspective. The greatest exponent of these proposals is the Busemeyer, Pothos, Aerts, Wang, Trueblood, Bruza, Gabora, et al. school, which culminated at the end of the last decade with the formalization of the Quantum Cognition theory (Bruza et al. [11], Busemeyer et al. [12], Busemeyer and Bruza [13]), a mathematical framework based on quantum probability theory, which provides us with tools for the modeling of human behavior from various standpoints. Specifically, similarity studies are one of the areas in which Quantum Cognition has had the greatest weight, and this is the focus of this work.

We can consider similarity as one of the most studied constructs in psychology as a science, and this is due to its wide implications to other psychological processes (attention, memory, categorization, decision making). However, Tversky's study [2] supposed a turning point in the study of similarity, being cited in almost 10,000 times. Tversky reported certain violations of the classical probability theory that we know today as asymmetry, triangular inequality, and diagnosticity. In this study, we will talk about the first two.

Asymmetry: to demonstrate asymmetry, Tversky conducted an experiment in which he asked participants "Which of the following phrases do you prefer to use?": Country A is similar to country B OR Country B is similar to country A

In his best-known example (China vs. North Korea), Tversky observed that 66 of the 69 participants judged the similarity between Korea and China $\text{sim}(\text{Korea}, \text{China})$ to be greater than $\text{sim}(\text{China}, \text{Korea})$. Many other examples are used in this article to demonstrate this effect.

The main problem lies in the impossibility of measuring the asymmetry between concepts in a space of coordinates because the distance between two concepts will always be the same in this type of space, regardless of the order. Although there are other approaches that try to parameterize this asymmetry (Tversky [2], Nosofsky [14]), it will be the Quantum Similarity Model (QSM) (Pothos and Busemeyer [3]) that provides a natural explanation for this phenomenon, not requiring parameterization.

These authors develop a similarity estimation model based on the representation of concepts using vector projections in a multidimensional space, defined

as subspaces of a Hilbert space of high dimensionality, all based on the logic of quantum probability theory. Thus, although the representation continues to be geometric, the entities are no longer represented as points but as complete subspaces. Consequently, thanks to the nature of quantum probability we are able to capture the order effect in which concepts are presented, estimating a different degree of similarity according to that order (Pothos and Busemeyer [3]).

Triangular Inequality: according to Triangular Equality, the distance between two points A and B will always be smaller than the sum of the distances between points A and C and points C and B. In terms of similarity, triangular equality is explained as $\text{Dissimilarity}(A, B)$ will always be smaller than the sum of $\text{Dissimilarity}(A, C)$ plus $\text{Dissimilarity}(C, B)$. This would be the same as saying that $\text{Sim}(A, B)$ will always be greater than the sum of $\text{Sim}(A, C)$ plus $\text{Sim}(C, B)$.

In his study, Tversky considered $A=\text{Russia}$, $B=\text{Jamaica}$, and $C=\text{Cuba}$, so that $\text{Sim}(\text{Russia}, \text{Jamaica})$ was always less than the sum of $\text{Sim}(\text{Russia}, \text{Cuba})$ and $\text{Sim}(\text{Cuba}, \text{Jamaica})$. This example suggests the violation of triangular equality, going against similarity measures if we consider that similarity is a linear transformation of distances. However, if we use a non-linear function of distances to measure similarity, we may find this violation. Kintsch [35] made a first approximation to this challenge with LSA.

Explanation of the Quantum Similarity Model

Before continuing, it is necessary to explain in more detail the proposal of the QSM. To do so, we will use the "bra-ket" notation developed by Paul Dirac [15] and will rely heavily on the original paper by Pothos and Busemeyer [3].

These authors start by describing the state vector ψ as a unit length vector within a multidimensional space. This vector represents "what a person is thinking at a particular moment". If we employ Dirac notation, then $|\psi\rangle$ is a column vector and $\langle\psi|$ is the adjoint (conjugate transpose) of this vector. Then, $|\psi\rangle\langle\psi|$ indicates an outer product and is the projector onto the one-dimensional subspace defined by $|\psi\rangle$. A projection operator is a linear operator, expressed as a matrix, which identifies the part of a vector that is restricted/ contained in a particular subspace. Also, $\langle\psi|\psi\rangle$ indicates a standard dot product (Pothos y Busemeyer [3]).

The main difference with the classical geometric perspective is that, in this model, different concepts are represented as subspaces (in other words, a concept definition is what could be spanned by a basis of a subspace), while in the classical models we represent the concepts as coordinates (particles or vectors) within the multidimensional space, that is, in a partial reality. For this reason, the most important elements of these models are the subspaces and their associated projectors, which consist of linear operators that project an state vector in a given subspace. For example, let's suppose that P_{China} is the projector of China's subspace. Then $P_{China}|\psi\rangle$ corresponds to the part of the state vector ψ that is contained (coherent) in China's subspace and reflects the probability that the vector ψ is *about* China (Pothos and Busemeyer [3]). This probability reflects the degree of consistency between the state vector and

the subspace and is therefore a measure of similarity (Tenenbaum and Griffiths [16]).

In order to evaluate the similarity between two subspaces (since we must consider the possibility that these are incompatible), we should use the method proposed by Busemeyer and Bruza [13]. Therefore, we must run successive projections from the state vector ψ to the last subspace we wish to evaluate $\text{sim}(\text{Korea, China})$ in the following way:

$$\begin{aligned}
& \|P_{China}P_{Korea}|\psi\rangle\|^2 \\
&= \|P_{Korea}|\psi\rangle\|^2 \cdot \|P_{China}|\psi_{Korea}\rangle\|^2 \\
&= \|P_{Korea}|\psi\rangle\|^2 \cdot \|P_{China} \frac{P_{Korea}|\psi\rangle}{\|P_{Korea}|\psi\rangle}\|^2
\end{aligned} \tag{1}$$

With $\|P|\psi\rangle\|^2$ being equivalent to the inner product $\|P|\psi\rangle\|^2 = \langle\psi|P|\psi\rangle$

These are characteristics of the quantum probability model, and in the case of QSM, it is assumed that the conjunction of probabilities corresponds to the similarity between two entities. Thus, $\|P_{China}P_{Korea}|\psi\rangle\|^2$ represents the similarity between North Korea and China (Pothos and Busemeyer [3]).

Based on these specifications, we already have enough information to start working with the QSM in an space model as LSA. However, it is important to note that we have to make some assumptions about this model before we continue.

Neutral initial state: in the first place, when calculating the similarity between two $\|P_{China}P_{Korea}|\psi\rangle\|^2$ concepts, we must assume that the initial vector is determined by $\|P_{China}|\psi\rangle\|^2 = \|P_{Korea}|\psi\rangle\|^2$ so that this initial vector is not biased towards either of the two concepts we are going to evaluate.

Container Space: a second assumption that we must consider from the QSM is that there must be a higher order space (Hilbert's space) that contains all the knowledge, on which the concepts can be represented. Thus, any subspace of lower dimensionality must have a representation in that container space. Our container space is the orthogonal latent space provide by LSA

Order effect: the most important implication of QSM is that the result of the similarity between two concepts is dependent on the order in which the concepts are presented so that $\text{sim}(\text{Korea, China})$ will be different from $\text{sim}(\text{China, Korea})$. This effect will occur whenever these concepts cannot be expressed with the same basic vectors, or one of the concepts is not expressed with a subset of vectors of the other concept. Thus, the quantum formalization of similarity allows us to measure the order effect.

Dimensionality effect: in order to establish a valid explanation for the asymmetry proposed by Tversky, it is necessary to assume that the participants have a greater knowledge background on one of the concepts presented to them. When formalizing this in the QSM this means that one of the concepts has a higher dimensionality than the other. If we talk about our example, the dimensionality of China will be greater than the dimensionality of North Korea, so we expect $\text{sim}(\text{Korea, China})$ to be always greater than $\text{sim}(\text{China, Korea})$.

The demonstration of the QSM is detailed in the works of Pothos and Busemeyer [3], Duran et al. [1], Pothos et al. [4, 17] and Yearsley et al. [18, 19, 20]. In these papers, the authors provide evidence of the capacity of the QSM to successfully model the cognitive processes explained above. In our case, it was necessary to explain the basis of this model for the following sections of the paper.

Running the QSM in a Vector Space Model rails

Today, several authors have proposed the use of the mathematical framework provided by quantum probability theory in order to capture the meaning of words and model associative processes within Vector Space Models or VSM (Aerts et al. [21], Balcoe et al. [22], Jaiswal et al. [23], Gonzalez and Caicedo [24]).

The studies of Aerts and Czachor [25] and Bruza and Cole [26] were a turning point in this sense since they were able to demonstrate that different VSM's such as LSA or the Hyperspace Analog to Language model (Lund and Burgess [27]) can be considered formal Hilbert's spaces (Balcoe et al. [22]).

The use of VSM, in particular LSA, will be the basis of our study to demonstrate QSM in a data-driven environment. A data driven environment has the advantage that the vectors of the basis of each subspaces are not generated by rational method. In other words, the distances between them are not set rationally by the researchers. Conversely, the locations of the vectors and their distances are extracted by the inferences of a model sensible to massive coocurrences.

2 Objectives and Hypothesis

The main objective of our study is the demonstration of the QSM assumptions through the formalization of meaningful subspaces contained in a higher order vector space that we will call "Container Space", using LSA techniques.

Specifically, our hypothesis is that we will preliminary be able to reproduce the violations of asymmetry and triangular inequality reported by Tversky [2] by using the QSM in a LSA environment.

3 Method

The proposed method will be divided into five phases:

a) Generation of the Container Space

We will begin by generating a vector space, which we will call container space, of 270 dimensions (Rehder et al. [28]) using LSA, which is a counting-words based vector space model that represent the semantics of a text corpus. In order to do that we are using a generalistic corpus based on samples from Wikipedia. It is important at this point to use a vector space model that guarantees the orthonormality of the dimensions that define it. LSA allows us to do this because, as Principal Component Analysis, the dimensions of the resulted matrices are orthonormal (Landauer et al. [29]).

b) Extraction of meaningful subspaces

From this container space with 270 latent (meaningless) and orthogonal dimensions, we will generate new subspaces for the concepts we want to represent. The method consists of the creation of a contour through the extraction of the closest neighbors to the words that we are going to analyze, understanding these close neighbors as those words that have a greater similarity (cosine) with our concepts. Potentially, the concept of a word will be extracted from contour.

The similarity phenomena in LSA and other n-dimensional models usually fit a negative logarithmic function, in which only a few neighbors have great similarities with the word (Jorge-Botana and Olmos [30], Jones et al. [31], Steyvers and Tenenbaum [32]). Therefore, we can assume that we will find the main meaning of a concept in that contour. The question is, what is the dimensionality that this contour deserves? To find the correct dimensionality, the algorithm we propose is the following (you can see the algorithm pseudocode in the Figure 1):

1. We do several descendent iterative cluster analyses using the contour of a term with 300 neighbors. We do from 300 up to 2.
2. We extract a list of the first 50 neighbors of each centroid's representative and select only the neighbors that belong to the cluster. These neighbors will be the descriptors of each centroid.
3. Finally, the base of the subspace extracted from each iteration is formed by each of the n centroids. However, it is important to highlight that the base we obtain in each iteration is not orthogonal, and as long as this base is re-orthogonalizable, it will deserve that dimensionality. If this basis cannot be properly reorthogonalized, will be not the basis to this concept.

```
Centroid Version:

Start
countour ← extractNeighbors(term, 300)
For n ← 2 to 300
  Centroids ← Clusters(countour,n)
  For each centroid in Centroids
    representative ← extractNeighbors (Centroid, 1)
    descriptors ← extractNeighbors (representative, 50)
    For each descriptor en descriptors
      If descriptor ∈ Clusters then
        Select(descriptor, clusterNum)
      End If
    End For
    basisMatrix.Add (centroid)
    clusterNum = clusterNum + 1
  End For
  report ← Orthogonalize(basisMatrix)
End For
End
```

Fig. 1. Meaningful subspaces extraction algorithm

We identify the best base to represent the subspace of a term following three criteria:

1. A basis is re-orthogonalizable if vectors remains at least 0.80 of its meaning. That is, the new re-orthogonalized vector correlates 0.80 with the original.
2. From the bases that meet the first criterion, we choose only those that have a minimum of three descriptors for each centroid
3. Of the bases that meet the first and second criteria, we choose the one with the most centroids (a maximal model).

Re-orthogonalization is carried out based on previous studies (Olmos, et al. [33], Jorge-Botana et al. [34]), that is, using the Gram-Schmidt method and the Pearson correlation to measure reliability. The basis of these new concepts subspaces have orthogonal and meaningful vectors now.

In the Figure 2, we can better understand what we have done so far.

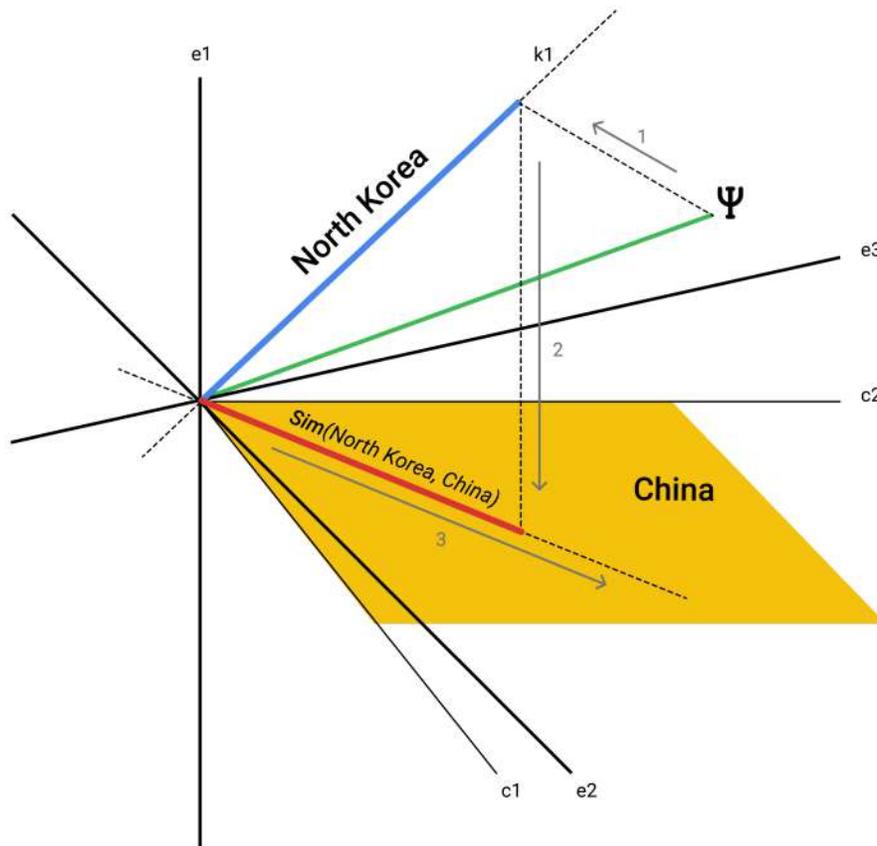


Fig. 2. Meaningful subspaces representation

Assuming that the container space U is defined according to a canonical basis $B = \{e_1, e_2, \dots, e_{270}\}$, C is a two-dimensional subspace with the base $C = \{c_1, c_2\}$ representing China and K is a one-dimensional subspace with base $K = \{k_1\}$ representing North Korea, the relationship (distances) between the vectors of China and North Korea is identified in U since the set $\{c_1, c_2, k_1\}$ shares a common reference, which is the base of the original latent space U . Note that e_1, e_2 and e_3 represent $B = \{e_1, e_2, \dots, e_{270}\}$

c) Calculation of the sub-spaces projections in the U space From the previous image, let's take the base B of the container space U with its 270 vectors $\{e_1, e_2, \dots, e_{270}\}$. China and Korea are defined by the subspaces C and K , whose vectors of each base are identified within the frame of that container space U . If we assume that the concept of China is represented by the basis (extracted by the method above) $\{v_{asia}, v_{commercial}\}$ then the transformation matrix to project something into the Chinese subspace is the sum of the outer products of each vector with itself (Busemeyer and Bruza [13]):

$$P_{China} = |v_{asia}\rangle\langle v_{asia}| + |v_{commercial}\rangle\langle v_{commercial}| \quad (2)$$

d) State Vector Estimation With the calculated projections, we can estimate the optimal parameters of the state vector ψ . To test the similarities, the state vector has to be a neutral state for both subspaces (tentatively a state of mind at an equal distance between both conceptual subspaces). Hence, its components are optimized (Pothos and Busemeyer [3]) to comply:

$$\|P_{China}|\psi\rangle\|^2 = \|P_{Korea}|\psi\rangle\|^2 \quad (3)$$

e) Similarities estimation Having the projections of both concepts and our state vector estimated we can compute the similarity according to the QSM applying the following formulas (Pothos and Busemeyer [3]):

$$Sim(Korea, China) = \|P_{Korea}|\psi\rangle\|^2 \cdot \|P_{China} \frac{P_{Korea}|\psi\rangle}{\|P_{Korea}|\psi\rangle}\|^2 \quad (4)$$

$$Sim(China, Korea) = \|P_{China}|\psi\rangle\|^2 \cdot \|P_{Korea} \frac{P_{China}|\psi\rangle}{\|P_{China}|\psi\rangle}\|^2 \quad (5)$$

4 Results

The following lines will explain the results obtained after applying the proposed method. This method has been applied to contrast the violations of asymmetry and triangular inequality reported by Tversky [2].

Asymmetry First, we have estimated the container space and extracted the China and Korea sub-spaces using the Gallito Studio tool [36]. We can interpret the meaning of these subspaces and the reorthogonalization rate through the Table 1.

As we see in the table, we can extract 7 dimensions from the concept of China and 6 from the concept of Korea. Also, we could re-orthogonalize all the dimensions with our method (see the R^2).

Dim	China	R^2	Korea	R^2
1	writing, united_states, character, neighborhood, commercial, actual	.95	December, republic, agreement	.97
2	imperial, emperor, imperial, monk, area, meaning, court	.96	Ukraine, Russia, Soviet, Soviet_Union, Russian	.93
3	Chinese, Japan, India, dynasty, Buddhism, Korea, Korean, Buddhist, Vietnam, Taiwan, Tibetan, Philippine, ming, Tibet, Peking, south_korea, Pakistan, Indonesia, Mandarin	.80	invasion, south, north	.96
4	Indian, Asian, foreign, immigrant, temple, ethnic, Hindu	.89	Chinese, Japan, India, dynasty, Korea, Buddhism, Korean, Asian, Vietnam, Buddhist, Taiwan, north_korea	.94
5	central_asia, Asia, Mongolian, west, western	.93	Cuban, country, united_states, Government, State, historic, relationship, official, right, American, popular, human, international, foreign, emperor, traditional, Model, development, system, Latest, capital, program	.93
6	karate, traditional, influence, martial, cuisine, rice	.96	socialism, communist, party, political, socialist, class, revolution	.96
7	Soviet, Russia, communist	.96	-	-

Table 1. Meaningful Subspaces

In the correlation plot below (Figure 3) we can see that there are some relationships between the two subspaces dimensions. As we notice, the correlation depends on the order. That's because of the nature of the correlation in the quantum probability framework that is given by the inner product of the vectors.

The next step is to check the adequation of the state vector. Remember that we need to optimize the function $\|P_{China}|\psi\rangle\|^2 = \|P_{Korea}|\psi\rangle\|^2$ so we can ensure the initial state is not biased, so we must comply $\|P_{China}|\psi\rangle\|^2 - \|P_{Korea}|\psi\rangle\|^2 = 0$. Also, the state vector must be unitary. We can see the results in the Table 2.

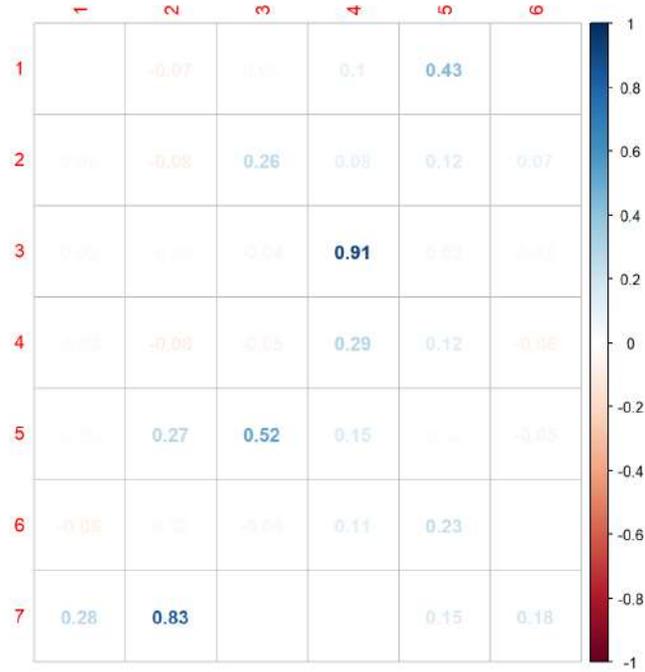


Fig. 3. China and Korea Dimensions Correlation Plot

Bias	Sum
$5.45 * e^{-10}$.893

Table 2. State Vector Evaluation

We see the state vector is totally unbiased, but the sum of its components is not 1, so we can not confirm it is a unitary vector. We should divide by the module of the vector to solve this problem.

Finally, if we use the similarity formula according to the quantum model to calculate the similarity between both concepts we obtain that $\text{Sim}(\text{Korea}, \text{China}) = .745$ and $\text{Sim}(\text{China}, \text{Korea}) = .697$. In this case, we can see the asymmetry effect between these two concepts.

Triangle inequality Once obtained the asymmetry results, the last step is to compute the similarities to check the triangle inequality. We will use the same example as Tversky [2], so we have extracted the meaningful subspaces of Jamaica, Cuba, and Russia. We will not show the dimensions table, neither the correlations plot in this case as we have not enough space. In the table below we can see the vector state bias for the three comparisons we will use in this example.

According to the QSM, the similarities are the following: $\text{Sim}(\text{Jamaica}, \text{Cuba}) = .109$, $\text{Sim}(\text{Cuba}, \text{Russia}) = .104$ and $\text{Sim}(\text{Russia}, \text{Jamaica}) = .099$

Concepts	Bias	Sum
Jamaica-Cuba	$1.35 * e^{-10}$.972
Cuba-Russia	$3.08 * e^{-7}$.945
Russia-Jamaica	$5.49 * e^{-9}$.950

Table 3. State Vectors Evaluation

Now we can check the triangle inequality property. In this case, the similarity between Russia and Jamaica (.099) is less than the sum of the other two similarities (.213), so we can confirm the triangle inequality with the same example as the Tversky [2] one.

5 Conclusions

As conclusions, we would like to focus on the QSM assumptions to justify our method. As Pothos and Busemeyer [3] specified in their paper, to implement the QSM we must have a neutral initial state and a container space in which all concepts should be represented, we must identify the order effect in the symmetry calculation and we should find a dimensionality effect, so the asymmetry could be explained.

As we could see in the results, we can meet all the QSM assumptions. Using our method to extract meaningful subspaces we are able to estimate a neutral vector state, we can represent meaningful subspaces inside the container space, we see the order effect in the China vs. Korea example and the most important, we can explain that effect with the dimensionality extracted for each concept with our method.

From our point of view, the most important achievement of this work is the method of subspaces extraction, since the dimensionality of each concept is not assigned by the researchers, but is determined by a data-driven space model and the potential re-ortogonalization of these meaningful subspaces, which is a formal criterion. Of course, this dimensionality will also depend on the amount of information the data-driven we have about each concept (it depends on our corpus of knowledge).

We could say that it is possible to apply the Quantum Similarity Model on multidimensional subspaces generated from an LSA vector space model as a container space U . This opens an exciting path to continue testing new hypotheses related to the diagnosticity effect.

Moreover, Tversky's [2] work is not the only one that reports violations of the assumptions of the classical probability theory when measuring decisional judgments. Without going any further, Tversky and Kahneman [7] also reported the "Conjunction Fallacy" effect, from which they coin the "Representativeness Heuristic". This effect also involves the violation of the distributive axiom in classical probability theory, which refers to the fact that the joint probability of two events is always lower than the probability of each of them separately.

The limitations of this study are the following:

1. These are the first preliminary one example tests performed with a very limited corpus of knowledge. We must expand our corpus to obtain more reliable results.
2. A validation study with human judgments has not been conducted. As a future line of research, we want to study if these effects are reproduced in the same way in humans and under vector space models.
3. Only a small sample of Tversky's experiments has been replicated, in the future, we must contrast these experiments with a greater number of examples. We have focused this study in the methodological aspects.

References

1. Barque Duran, A., Pothos, E. M., Yearsley, J. M., Hampton, J. A., Busemeyer, J. R., and Trueblood, J. S. (2016). Similarity judgments: from classical to complex vector psychological spaces. In *Contextuality from Quantum Physics to Psychology* (pp. 415-448).
2. Tversky, A. (1977). Features of similarity. *Psychological review*, 84(4), 327.
3. Pothos, E., and Busemeyer, J. (2011). A quantum probability explanation for violations of symmetry in similarity judgments. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33, No. 33).
4. Pothos, E. M., Busemeyer, J. R., and Trueblood, J. S. (2013). A quantum geometric model of similarity. *Psychological Review*, 120(3), 679.
5. Tversky, A., and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *science*, 211(4481), 453-458.
6. Kahneman, D., and Tversky, A. (1982). Judgments of and by representativeness (pp. pp-84).
7. Tversky, A., and Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, 90(4), 293.
8. Kahneman, D., and Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment*, 49, 81.
9. Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2), 211.
10. Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, 124(1), 1.
11. Bruza, P., Busemeyer, J., and Gabora, L. (2013). Introduction to the special issue on quantum cognition. *arXiv preprint arXiv:1309.5673*.
12. Busemeyer, J. R., Pothos, E. M., Franco, R., and Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment errors. *Psychological review*, 118(2), 193.
13. Busemeyer, J. R., and Bruza, P. D. (2012). *Quantum models of cognition and decision*. Cambridge University Press.
14. Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology*, 23(1), 94-140.
15. Dirac, P. A. M. (1939). "A new notation for quantum mechanics". *Mathematical Proceedings of the Cambridge Philosophical Society*. 35 (3): 416-418.
16. Tenenbaum, J. B., and Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and brain sciences*, 24(4), 629.

17. Pothos, E. M., Barque-Duran, A., Yearsley, J. M., Trueblood, J. S., Busemeyer, J. R., and Hampton, J. A. (2015). Progress and current challenges with the quantum similarity model. *Frontiers in psychology*, 6, 205.
18. Yearsley, J. M., Pothos, E. M., Hampton, J. A., and Duran, A. B. (2014, June). Towards a quantum probability theory of similarity judgments. In *International Symposium on Quantum Interaction* (pp. 132-145). Springer, Cham.
19. Yearsley, J. M., Pothos, E. M., Duran, A. B., and Hampton, J. A. (2015). Diagnosticity: Some theoretical and empirical progress. In *CogSci*.
20. Yearsley, J. M., Barque-Duran, A., Scerrati, E., Hampton, J. A., and Pothos, E. M. (2017). The triangle inequality constraint in similarity judgments. *Progress in Biophysics and Molecular Biology*, 130, 26-32.
21. Aerts, D., Broekaert, J., Sozzo, S., and Veloz, T. (2013, July). Meaning-focused and quantum-inspired information retrieval. In *International Symposium on Quantum Interaction* (pp. 71-83). Springer, Berlin, Heidelberg.
22. Blacoe, W., Kashefi, E., and Lapata, M. (2013, June). A quantum-theoretic approach to distributional semantics. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 847-857).
23. Jaiswal, A. K., Holdack, G., Frommholz, I., and Liu, H. (2018, September). Quantum-like generalization of complex word embedding: a lightweight approach for textual classification. *CEUR Workshop Proceedings*.
24. Gonzalez, F. A., and Caicedo, J. C. (2011, September). Quantum latent semantic analysis. In *Conference on the Theory of Information Retrieval* (pp. 52-63). Springer, Berlin, Heidelberg.
25. Aerts, D., and Czachor, M. (2004). Quantum aspects of semantic analysis and symbolic artificial intelligence. *Journal of Physics A: Mathematical and General*, 37(12), L123.
26. Bruza, P. D., and Cole, R. J. (2006). Quantum logic of semantic space: An exploratory investigation of context effects in practical reasoning. arXiv preprint quant-ph/0612178.
27. Lund, K., and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, and computers*, 28(2), 203-208.
28. Rehder, B., Schreiner, M. E., Wolfe, M. B., Laham, D., Landauer, T. K., and Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes*, 25(2-3), 337-354.
29. Landauer, T. K., McNamara, D. S., Dennis, S., and Kintsch, W. (Eds.). (2013). *Handbook of latent semantic analysis*. Psychology Press.
30. Jorge-Botana, G., and Olmos, R. (2014). How lexical ambiguity distributes activation to semantic neighbors: Some possible consequences within a computational framework. *The Mental Lexicon*, 9(1), 67-106.
31. Jones, M., Gruenfelder, T., and Recchia, G. (2011). In defense of spatial models of lexical semantics. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33, No. 33).
32. Steyvers, M., and Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, 29(1), 41-78.
33. Olmos, R., Jorge-Botana, G., Luzon, J. M., Martin-Cordero, J. I., and Leon, J. A. (2016). Transforming LSA space dimensions into a rubric for an automatic assessment and feedback system. *Information Processing and Management*, 52(3), 359-373.

34. Jorge-Botana, G., Olmos, R., and Luzon, J. M. (2020). Bridging the theoretical gap between semantic representation models without the pressure of a ranking: some lessons learnt from LSA. *Cognitive processing*, 21(1), 1–21.
35. Kintsch, W. (2014). Similarity as a function of semantic distance and amount of knowledge. *Psychological review*, 121(3), 559.
36. Jorge-Botana, G., Olmos, R., and Barroso, A. (2013, July). Gallito 2.0: A natural language processing tool to support research on discourse. In *Proceedings of the 13th Annual Meeting of the Society for Text and Discourse*.

A Decomposition Analysis of Differences in Length of Life in the Czech Republic

David Morávek¹ and Jitka Langhamrová²

¹ Department of Demography, Faculty of Informatics and Statistics, University of Economics, Prague, W. Churchill Sq. 1938/4, 130 67 Prague 3 – Žižkov, Czech Republic

(E-mail: xmord09@vse.cz)

² Department of Demography, Faculty of Informatics and Statistics, University of Economics, Prague, W. Churchill Sq. 1938/4, 130 67 Prague 3 – Žižkov, Czech Republic

(E-mail: langhamj@vse.cz)

Abstract. Nowadays, mortality is improving in almost every population. Life expectancy at birth for men and women has increased significantly over the last few decades, mainly due to a decrease of infant mortality. For the period from 1920 to 2018 in Czechia, the life expectancy at birth has increased by 29.1 years for males and by 32.1 years for females. Except for life expectancy characterizing the average length of life, the further use of the median and the modal ages at death is common. The characteristics of human length of life are based on a computation of life tables. In a population, it approximately holds true that the average length of life is lower than the median length of life, which is lower than the modal length of life. Differences in life expectancy by age and sex is computed using the decomposition method.

Keywords: lengths of life, life expectancy, mortality, life tables, decomposition method.

1 Introduction

The length of a human life tends to increase over time. In studying mortality trends, life expectancy – which represents the average length of life – is commonly used (Canudas-Romo[1]). The increase is mostly caused by a decline in infant mortality, which together with a decline in birth rate leads to the demographic ageing of a population (Arltová et al.[2]). This main indicator comes from a computation of life tables as a number, which is comparative over time and between territorial units. It is considered as a summary characteristic of mortality whose main benefit is that it is not affected by the age structure of the current population. The methodology for construction of life tables differs across countries. In Czechia, since the year 2018, a methodology used by the Czech Statistical Office (CZSO[3]) has substantially changed in terms of methods used for smoothing and for modelling mortality rates. To exclude random fluctuations in observed mortality rates, they are smoothed

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



by the generalised additive model in combination with the so-called P-splines, known as the P-GAM method (CZSO[3]). To obtain mortality rates at the oldest ages, with regard to the lower number of deaths and lower reliability of data of the mid-year population in these ages groups, it is common to use a model in the methodology of computing life tables. The model selected by CZSO[3] is based on a logistic curve, which takes into account the deceleration in mortality increase with age. Originally, it was known as the Kannisto model.

As Arriaga[4] noticed, mortality at old ages may not reflect the current mortality of those ages but rather a simplistic assumption based on a model of life tables or mathematical function. There is also the problem of the effect of the limit of the human life span on the possible change in life expectancy (Arriaga[4]). The benefit of temporary, or partial life expectancy, which was originally described by Arriaga (1984), is its dependence only on mortality rates between exact ages x and $x + i$ instead of the whole range of ages, as in the case of defining life expectancy at birth (Saikia et al.[5]). The temporary life expectancy measures the mortality of a population typically from birth to age around 75 years (Burcin[6]).

In order to describe the distribution of life table deaths, except life expectancy, which has limitation as a mean value, the median and modal ages at death are considered appropriate characteristics. It is known that life expectancy is supposed to be lower than the median age at death that is at the same time lower than the modal age at death. According to Lexis (1878), there are three parts in the distribution of life table deaths: a decrease in the high number of deaths with age after birth due to the existence of infant mortality; deaths centred around the late modal age at death; and premature deaths that occur infrequently at young ages (Canudas-Romo[7]).

Decomposing a difference in life expectancies is useful in estimating what mortality differences in a specific age group contribute to the total difference in life expectancy at birth (Preston et al.[8]). There are two main approaches in decomposing a difference in life expectancy, a continuous one according to Pollard (1982) and a discrete approach using the formula by Arriaga (1984). These two procedures formally lead to the same results, nevertheless the Arriaga formula is easier to apply to traditional life tables (Preston et al.[8]).

For the purpose of this contribution, we use life tables produced by CZSO[9] from 1920 to 2018 for Czechia. Data are based on the methodology of life tables used by CZSO since 2018. The life tables were recalculated according to the current methodology starting from 1920.

2 Distribution of Life Table Deaths

Life tables are used as a basis for computing the other characteristics representing the length of a human life, the median and the modal ages at death. To obtain the median age at death with decimal-point precision, Canudas-Romo[1] proposed a formula using values of the survival function at two contiguous ages of the value of $l(Md, t) = 0.5$, which assumes linearity in the interval between the ages x and $x + 1$, where $l(Md, t) = 0.5$ is located.

$$Md(t) = x + \frac{[0.5 - l(x, t)]}{[l(x + 1, t) - l(x, t)]}. \quad (1)$$

The modal age at death with decimal precision is calculated according to Kannisto's (2001) proposal, in whose formula a method of calibration value is used (Canudas-Romo[1]). The highest number of deaths in the life table at time t is to be found at the age x . The number of deaths at ages x , $x - 1$ and $x + 1$ are used to fit a quadratic polynomial to the function describing the death distribution (Canudas-Romo[7]).

$$M(t) = x + \frac{[d(x, t) - d(x - 1, t)]}{[d(x, t) - d(x - 1, t)] + [d(x, t) - d(x + 1, t)]}, \text{ for } x > 5. \quad (2)$$

Age distribution of life table deaths is typically bimodal with the first local mode at age 0 and the second local mode at an older age (Hirouchi[10]). To obtain the local mode at an older age, or the late modal age at death, Canudas-Romo[1] considers age $x > 5$.

The distribution of the life table deaths is not symmetrical as it is shown in Figure 1 and Figure 2. Due to high infant mortality in earlier times (in 1920 in Czechia), the life expectancy at birth for males is considerably lower (47.03) than the median age at death (57.17). Nonetheless, the majority of deaths occur at a relatively higher age (73.50). Over the almost one hundred years (in 2018), the characteristics are much closer to each other. The deaths move towards higher ages with the modal age at death equal to 84.21 years. The life expectancy at birth for males is calculated at 76.08 years for Czechia in 2018.

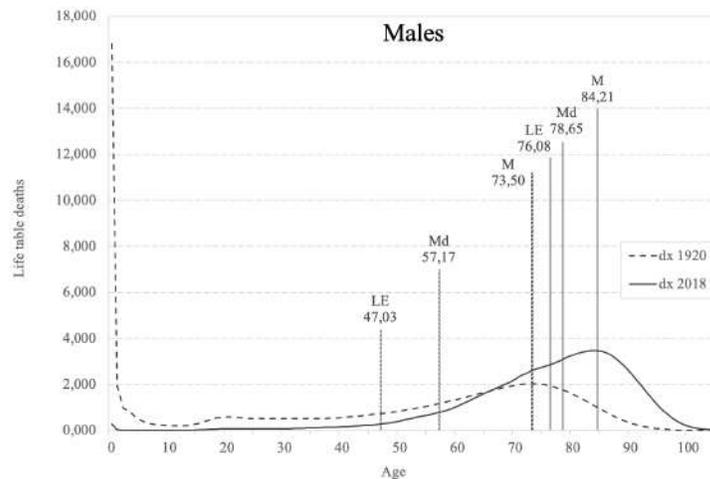


Fig. 1. Modal (M) and Median (Md) Ages at Death, and Life Expectancy at Birth (LE) for the Life Table Male Deaths in Czechia in 1920 and 2018.

Source: CZSO; Canudas-Romo[1]; author's calculations and processing.

The distribution of the life table female deaths for Czechia showed lower infant mortality. The female life expectancy at birth (49.78) is closer to the median age at death (60.52) than for males. A higher mortality at younger ages among males and females, also called the "Adult Mortality Hump" (Remund et al.[11]), is present. There are fewer female deaths than male at these ages. This demographic phenomenon has been recently described as a consequence of the reduction of infections and the increase in the share of adult mortality attributed to cancer and cardiovascular disease (Beltrán-Sánchez et al.[12]). As female life expectancy exceeds that of males, there are more deaths among females, especially at older ages with a higher modal age at death for women (84.21 vs. 87.95 in 2018).

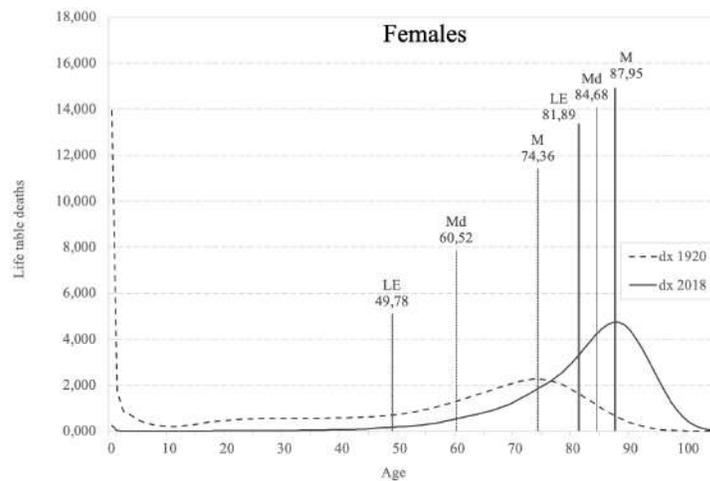


Fig. 2. Modal (M) and Median (Md) Ages at Death, and Life Expectancy at Birth (LE) for the Life Table Female Deaths in Czechia in 1920 and 2018.

Source: CZSO; Canudas-Romo[1]; author's calculations and processing.

3 Lengths of Life

The length of a human life, as measured by life expectancy and median and modal ages at death, tends to increase over time. As seen in Figure 3 for the period from 1920 to 2018 in Czechia, the life expectancy at birth has strongly increased by 29.05 years for males and by 32.11 years for females. It is shown that the characteristics of length of life tend to be closer to each other in 2018 in comparison with 1920 for both males and females. The median age at death had a similar trend over time as life expectancy at birth. However, the modal age at death for men was fluctuated during the period from 1920 to 1970, then it started to increase up to 2018. The female length of life has increased more rapidly over time. The gender difference between life expectancy at birth rose from 2.75 years in 1920 to 7.86 in 1990, then it started declining gradually (5.81

in 2018). This difference is thought to have both biological and non-biological origins (Sundberg et al.[13]).

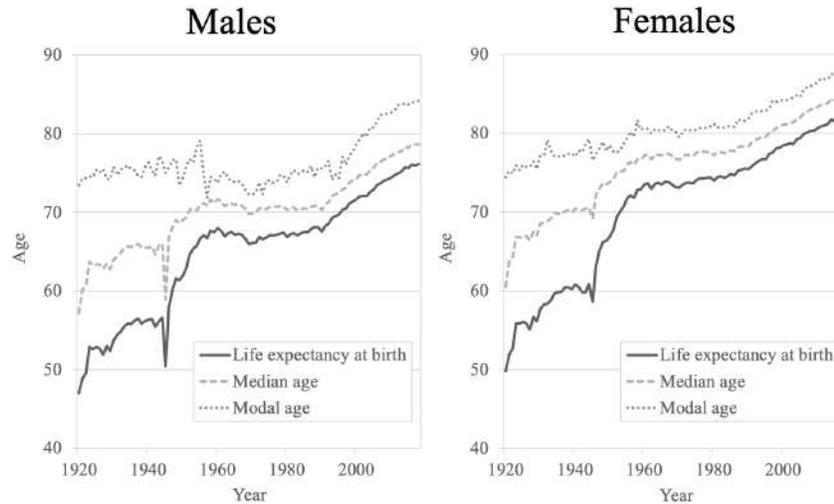


Fig. 3. Modal and Median Ages at Death, and Life Expectancy at Birth, Males and Females, Czechia, 1920–2018

Source: CZSO; author’s calculations and processing.

We computed the temporary life expectancy as the average number of years that a group of persons alive at exact age x will live from x to $x + i$, using the formula according to Arriaga[4]

$${}_i e_x = \frac{T_x - T_{x+i}}{l_x}. \quad (3)$$

We opt for calculating the temporary life expectancy from birth to 80 in order to measure mortality at ages under 80 years, separately for males and for females. We considered this to be the predominate age in which model values of mortality rates are not yet used. Trends over time in temporary life expectancy between the ages of 0 and 80 are shown in Figure 4. As seen, the difference between the age equal to 80 and the temporary life expectancy (72.7 years for males and 76.1 years for females) is negligible in 2018 compared with 1920. The increase in temporary life expectancy slightly slowed down after 1960. Since 1990, life expectancy at birth has increased faster than temporary life expectancy between the ages of 0 and 80. Therefore, the increase of life expectancy at the oldest ages over 80 years is supposed to be greater in the period from 1990.

To compute the tempo of mortality change during a period Arriaga[14] suggested a relative measure ${}_i RC_x^n$, where the observed change between temporary life expectancy is in relation to the maximum possible change. If periods of time are different, the formula should be modified (${}_i ARC_x^n$)

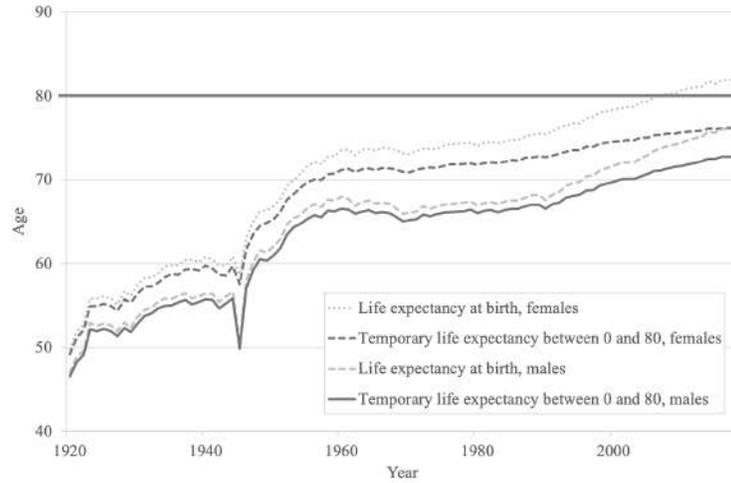


Fig. 4. Life Expectancy at Birth and Temporary Life Expectancy between Ages 0 and 80, Males and Females, Czechia, 1920–2018.

Source: CZSO; author’s calculations and processing.

$${}_iRC_x^n = \frac{{}_i e_x^{t+n} - {}_i e_x^t}{{}_i e_x^t}; \quad (4)$$

$${}_iARC_x^n = [1 - (1 - {}_iRC_x^n)^{1/n}].100. \quad (5)$$

As seen in Table 1, the calculated index of annual relative change in temporary life expectancy between ages 0 and 80 reached the highest value in the period of 1940–1960 (2.90 for males and 4.05 for females). The second highest relative mortality change was recorded among males and females in 1920–1940. The index is similar for both sexes in the period of 2000–2018 (1.90 for males and 1.92 for females). It is noticeable that the relative change in mortality is greater among females in all periods recorded compared with males.

Year	Males		Females	
	TLE	Index	TLE	Index
1920	46,51	×	49,17	×
1940	55,69	1,59	59,74	2,08
1960	66,51	2,90	71,14	4,05
1980	65,98	-0,19	71,75	0,36
2000	69,71	1,53	74,52	2,02
2018	72,72	1,90	76,13	1,92

Table 1. Temporary Life Expectancy between Ages 0 and 80 (TLE) and Index of Annual Relative Change in Temporary Life Expectancies between 0 and 80, Males and Females, Czechia, 1920–2018.

Source: CZSO; author’s calculations and processing.

4 Age and Sex Decomposition

In order to decompose the total change in life expectancy at birth by age groups, Ponnappalli[17] listed formulas in life table terms of l_x and e_x with regards to Arriaga's original proposal as follows:

$$l_x^1(e_x^2 - e_x^1) - l_{x+n}^1(e_{x+n}^2 - e_{x+n}^1) \quad (6)$$

For the open-ended age group, Arriaga suggested (Ponnappalli[17]):

$$l_x^1(e_x^2 - e_x^1) \quad (7)$$

The contributions to change in life expectancy at birth by age and sex for Czechia in the period 1920–2018 are shown in Table 2. As discussed hereinabove, the overall life expectancy at birth increased mainly due to the positive contribution at age 0 as a result of a decline in infant mortality. The contribution at the age of 0 is calculated at 11.15 years for males and 9.84 years for females in total. The highest contribution is to be found in the period 1940–1960 (5.97 for males, 5.21 for females) following World War II, when life expectancy was generally increasing, and the infant mortality rate as well as mortality caused by infectious diseases were declining (Dzúrová[16]). In the period 1940–1960, life expectancy increased by 11.60 for males and by 12.71 for females in all age groups. The second highest contribution is due to the age group 1–24 years among males (6.81 years in total). Child and adolescent mortality was slightly reduced also among females (6.90 years in total), nonetheless mortality was more substantially improved in the age group 50–79 years (8.37 years in total).

Age Group	1920–1940	1940–1960	1960–1980	1980–2000	2000–2018	Total
Males						
0	3,73	5,97	0,30	0,98	0,15	11,15
1–24	3,55	2,28	0,35	0,36	0,27	6,81
25–49	1,72	1,76	-0,23	0,63	0,80	4,69
50–79	0,36	1,38	-1,23	2,52	2,82	5,85
80+	-0,04	0,20	-0,26	0,25	0,41	0,56
Total	9,32	11,60	-1,07	4,74	4,46	29,05
Females						
0	3,49	5,21	0,27	0,76	0,09	9,84
1–24	3,55	2,80	0,19	0,19	0,16	6,90
25–49	2,61	1,98	0,24	0,34	0,34	5,51
50–79	1,35	2,47	-0,10	2,55	2,11	8,37
80+	-0,01	0,25	-0,09	0,56	0,79	1,50
Total	10,99	12,71	0,50	4,41	3,50	32,11

Table 2. Contributions to Change in Life Expectancy at Birth by Age Group and Sex, Czechia, 1920–2018.

Source: CZSO; author's calculations and processing.

The health of children and adolescents has attracted considerable political and professional attention in recent years (Kudlová[17]). The contribution to the overall change in male life expectancy at ages 0–49 years is greater than in female life expectancy in 1980–2000 and 2000–2018. The contribution of the age group 25–49 years is similar for both sexes in total (4.69 for males and 5.51 for females). As a consequence of huge advances in medicine and technologies, the mortality of the older and oldest ages has been greatly reduced (Arltová and Vrabcová[18]). The contribution of the age group 80 and older was 0.56 for males and 1.50 for females.

To decompose the changes in temporary life expectancy between exact ages 0 and 80, in the formula, the life expectancy at birth is replaced by the temporary life expectancy (Arriaga[4]). The overall change in temporary life expectancy between ages 0 and 80 is calculated at 26.20 for males and 26.96 for females (Table 3). The majority of increase was recorded in 1920–1940 and 1940–1960 as result of improving mortality among younger ages (mostly in age group 0 and 1–24 years). Child and adolescent mortality was reduced more among males. Excluding the highest ages (80+) led to a decrease in gender difference due to the greater number of women alive in these age groups. Nevertheless, the contribution of age group 50–79 is higher for males than females in 1980–2000 and 2000–2018.

Age Group	1920–1940	1940–1960	1960–1980	1980–2000	2000–2018	Total
Males						
0	3,50	5,61	0,30	0,94	0,14	10,48
1–24	3,53	2,27	0,35	0,34	0,25	6,73
25–49	1,80	1,75	-0,22	0,58	0,70	4,60
50–79	0,36	1,19	-0,95	1,87	1,92	4,39
Total	9,18	10,82	-0,52	3,72	3,01	26,20
Females						
0	3,22	4,81	0,26	0,72	0,09	9,09
1–24	3,43	2,70	0,18	0,18	0,15	6,64
25–49	2,62	1,91	0,22	0,30	0,28	5,34
50–79	1,29	1,99	-0,05	1,56	1,10	5,89
Total	10,56	11,40	0,61	2,76	1,61	26,96

Table 3. Contributions to Change in Temporary Life Expectancy between Ages 0 and 80 by Age Groups and Sex, Czechia, 1920–2018.

Source: CZSO; author’s calculations and processing.

5 Conclusions

The aim of this contribution was to show trends in length of life in the Czech Republic, not only by using life expectancy as a widely used indicator representing a mean value, but also by other characteristics, which are the median and modal ages at death. To describe the differences in length of life by age and

sex, we used a decomposition method. The contribution focused on decomposition of differences in life expectancy at birth and temporary life expectancy in Czechia 1920–2018. We calculated the temporary life expectancy to measure mortality at ages from birth under 80 years.

In comparing the distribution of life table deaths in 1920 with 2018, it is shown that the characteristics tend to be closer to each other over time. The deaths move towards higher ages. Due to high infant mortality in earlier times, the life expectancy at birth was considerably lower (in 1920 – 47.03 years for males and 49.78 years for females). Nonetheless, the majority of deaths occur at a relatively higher age. The distribution of the life table deaths is not symmetrical. The life expectancy at birth is lower than the median age at death, which is at the same time lower than the modal age at death.

Life expectancy increased rapidly due to the mortality improvements in infants, children and adolescents. For the period from 1920 to 2018 in Czechia, the life expectancy at birth has increased by 29.1 years for males and by 32.1 years for females. The overall life expectancy at birth increased mainly due to the positive contribution at age 0 as a result of a decline in infant mortality. The contribution at the age of 0 was 11.15 years for males and 9.84 years for females in total. The highest positive contribution to the overall life expectancy was registered in the period 1920–1940 and 1940–1960. The contribution of the age group 50–79 years was significant in 1960–1980 and 2000–2018. Mortality improved also among the oldest age groups above 80 years. The difference between life expectancy at birth and temporary life expectancy was higher starting from 1990 as a result of improving mortality at age above 80 years. Contribution to the change in temporary life expectancy between exact ages 0 and 80 by age group for Czechia in 1920–2018 showed that mortality was improved much more among males than females in the period 1980–2000 and 2000–2018.

In studying the trends in lengths of life over time, the question arises of how mortality improves according to other demographic characteristics. There is a difference of mortality in a population, for example in relation to marital status or education. We discussed the gender difference in length of life, which has to be further studied. A mortality analysis should be performed on a specific age group, since the phenomenon of "Adult Mortality Hump" is known. A deeper understanding could help bring about a mortality analysis by causes of death, which has a specific influence on trends in length of life.

6 Acknowledgements

The article was supported by the Internal Grant Agency of the University of Economics Prague No. 35/2020 *Decomposition analysis of mortality* and by the Czech Science Foundation No. GA ČR 19-03984S under the title *Economy of Successful Ageing*.

References

1. V. Canudas-Romo. *Three measures of longevity: Time trends and record values*, Demography, 47, 2, 299–312, 2010.
2. M. Arltová, J. Langhamrová and J. Langhamrová. *Development of Life Expectancy in the Czech Republic in Years 1920–2010 with an Outlook to 2050*, Prague Economic Papers, 22, 1, 125–143, 2013.
3. CZSO. Czech Statistical Office. *Life tables – Methodology*, Prague, Czech Statistical Office, 2018. Retrieved from : <https://www.czso.cz/csu/czso/life-tables-methodology>.
4. E. Arriaga. *Measuring and explaining the change in life expectancies*, Demography, Springer, Population Association of America (PAA), 21, 1, 83–96, 1984.
5. N. Saikia, D. Jasilionis, F. Ram and V. Shkolnikov. *Trends and geographic differentials in mortality under age 60 in India*, Population Studies, 65, 73–89, 2011.
6. B. Burcin. *Avoidable mortality in the Czech Republic in 1990–2006*, Czech Demography, 3, 2009.
7. V. Canudas-Romo. *The modal age at death and the shifting mortality hypothesis*, Demographic Research, 19, 1179–1204, 2008.
8. S. H. Preston, P. Heuveline and M. Guillot. *Demography – Measuring and Modelling Population Processes*, 2001.
9. CZSO. Czech Statistical Office. *Life Tables in Time Series – 1920–2016*, Prague, Czech Statistical Office, 2018. Retrieved from : <https://www.czso.cz/csu/czso/life-tables-in-time-series-1920-2016>.
10. S. Horiuchi, N. Ouellette, K. Cheung and J-M. Robine. *Modal age at death: Lifespan indicator in the era of longevity extension*, Vienna Yearbook of Population Research, 11, 37–69, 2013.
11. A. Remund, T. Riffe and C. G. Camarda. *A cause-of-death decomposition of the young adult mortality hump*, Demography, 55, 2017.
12. H. Beltrán-Sánchez, C. E. Finch and E. M. Crimmins. *Twentieth century surge of excess adult male mortality*, Proceedings of the National Academy of Sciences of the United States of America, 112, 29, 8993–8998, 2015.
13. L. Sundberg, N. Agahi, J. Fritzell and S. Fors. *Why is the gender gap in life expectancy decreasing? The impact of age- and cause-specific mortality in Sweden 1997–2014*, International journal of public health, 63, 6, 673–681, 2018.
14. E. Arriaga. *The Deceleration of the Decline of Mortality in LDCs: The Case of Latin America*, International Population Conference, Manila, 1981. International Union for the Scientific Study of Population, Imprimerie Derouaux, 2, 21–50, Liege, Belgium.
15. K. M. Ponnappalli. *A comparison of different methods for decomposition of changes in expectation of life at birth and differentials in life expectancy at birth*, Demographic Research, 12, 141–172, 2005.
16. D. Dzurova. *Mortality differentials in the Czech Republic during the post-1989 socio-political transformation*, Health & place, 6, 351–62, 2001.
17. E. Kudlová. *Life cycle approach to child and adolescent health*, Central European Journal of Public Health, 12, 3, 166–70, 2004.
18. J. Vrabcová and M. Arltová. *Time series analysis of the relationship between mortality and selected economic indicators in the Czech Republic*, The 9th International Days of Statistics and Economics, Prague, September 10–12. 2015.

Smooth Chain Graph Model of type II: a learning procedure

Federica Nicolussi¹

Department Economics and Quantitative Methods, via Conservatorio, 7, Universit degli Studi di Milano, Italy.

(E-mail: federica.nicolussi@unimi.it)

Abstract. Chain Graph Models (CGs) are a widely used tool to describe the conditional independence relationships among a set of variables. One of the advantages lies in the possible use undirected and directed arcs to link vertices representing variables in the graph. There are four ways to read off the conditional independencies from a chain graph. Each way differs from the other in the way of interpret the missing (un)directed arcs, (see Drton 2009). Different problems can be address with different CGs, however often it is not clear which type of CGs is the best in order to describe the multivariate system of relationships underlying the selected variables. In this work, we propose a learning algorithm, based on a Monte Carlo procedure, that consider the system of independencies underlying all four CGs and select the type and the graph which optimize a score function. When we handle with categorical variables, we take advantage of the marginal models (Bergsma and Rudas, 2002) to parametrize the joint and marginal probability distribution of the variables. Unlikely, Bergsma and Rudas, 2002 showed that particular combinations of conditional independences have no a smooth parametrization. Nicolussi and Colombi, 2013 and 2017, provide the condition according to (any type of) CG admits a smooth parametrization. In the learning procedure we consider only the smooth CGs, that is they admit a smooth parametrization. This approach is implemented to study the poverty status and particularly how this one can be affected from a group of selected variables. We took advantage of the cross-section data sets of Hungarian Household. This analysis highlighted a strong effect of the considered social variables on the poverty status.

Keywords: Chain Graph models; categorical data; learning procedure; poverty status.

1 Introduction

In social studies, there is increasing attention to multivariate models that can capture and describe multiple aspects in a simple way. In particular, it is worthwhile to observe how one or more study variables are affected by other factors. It is also plausible to think that different relationships link the variables studied (i.e. symmetrical, asymmetrical, or causal). Chain Graph models well represent complex conditional independence assumptions through a particular graph, so-called Chain Graph (CG). Moreover, they well shape both direct and indirect associations. One of the most advantages of these models

¹*6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain*



lies in their immediacy of the representativeness. Indeed, in the graph, each vertex depicts a variable, and the arcs depict association between the two variables represented by the nodes. In particular, each undirected arc pretends for the symmetric association, and each directed arc denotes an effect according to the direction of the arrow. In this work, we consider the four types of Chain Graph models presented in the literature, each of which able to depict peculiar connections, [11]). Section 1.1 is addressed to introduce this topic. In literature, it is widespread to shape Undirected Graph models for discrete variables with log-linear parameters to expound the associations among these variables, [14]. Unfortunately, log-linear parameters can not describe different kinds of relationships in the same models, and these parameters are left in favor of the more general Marginal models, [4]. Bergsma and Rudas (2002) introduced these models to model several 4dependences among a set of discrete variables. For this reason, the Chain Graph Models for discrete variables use marginal models, adding the visual tool to describe the system of relationship between variables, [17], [18], [26], [21]. The potential of these models on social and economic studies is shown by Nemeth and Rudas 2013 [19], [20], [22]. This work aims to highlight whether and how other selected variables can cause income inequality and, more specifically, the poverty status. We take advantage of the Household Monitor survey of TARKI for the Hungarian study. In literature, many works study these two data-sets under the poverty issue. Most of these work on cross-section data use among other classical log-linear models to describe the multivariate system of relationship, see, for instance, [13], [16], [6], and [23]. In this work, we also replied to some results known in the literature. The paper follows this structure. In Sections 1.1 and 1.2, we give a brief introduction to Chain Graph models and Marginal models. In Section 2, we propose a learning procedure for the final graph. Finally, in Section 3, we expound the study data sets and the used methodology step by step, and we show the results of the analysis of the data set are explained. Furthermore, in Section 4, we added a brief conclusion to summarize the output of the analysis.

1.1 Chain Graph Models

Different multivariate analysis to model the relationships among a set of variables exist in literature. Graphical models take advantage of the visual impact that does easily interpretable complex associations. A CG is a graph that includes both directed and undirected arcs while excluding any direct or semi-directed cycle. A CG can be decomposed into so-called chain components, ordered according to the direction of the arrow. Each component is an undirected sub-graphs that contains only undirected arcs, while the vertices in different components are linked to each other by directed arcs. CG Models use chain graphs to represent a system of conditional independencies in a collection of variables. Each variable would be represented as a vertex. In contrast, arcs would represent symmetrical or asymmetrical relationships between them concerning whether the arc is directed or not so that the lack of an arc represents conditional independence. There are four types of CG Models available in the literature for data analysis, which differ in the way to explain the in-

dependence statements (see [11]). However, only three of these have suitable features to describe some problems. In this work, we consider only these three. The CG models proposed by [15] and [12], hereafter LWF CGMs, unifies the directed and undirected graphs approach. The CG models proposed by [1] (AMP CGMs) describe the dependence structures among regression residuals. These two models interpret the lack of an undirect arc conditionally to the other variables in the same component. On the other hand, the CG models proposed by [8] and [25] (MR CGMs) marginalize over these last variables and are suitable to describe multivariate regression systems. Further, LWF CGMs interpret the lack of a direct arc conditionally to the other variables in the same component, while AMP CGMs and MR CGMs marginalize over these last variables. For more profound dissertations about these models and their application, see [19].

1.2 Marginal log-linear parameterization

Log-linear parameters are a useful tool to handle with categorical variables but they are not able to depict conditional independence restrictions involving subsets of variables. Since often the inherent independencies of a CG model concern subsets of variables, we need a most flexible tool, such as marginal log-linear parameters. Marginal log-linear parameters are standard log-linear parameters defined within subsets of contingency tables obtained by marginalizing over one or more variables, [4]. Bergsma and Rudas (2002) show that by building the parameters according to two specific properties (of hierarchy and completeness) the asymptotic properties of parameters hold.

Let consider for instance a set of two variables A and B collected in a contingency table of dimension $n_A \times n_B$ with probability π_{ij} where $i = 1, \dots, n_A$, $j = 1, \dots, n_B$. Let furthermore consider $\{A; AB\}$ as marginal sets. Then the marginal log-linear parameters are given by:

$$\begin{aligned} \eta_A^A &= \left\{ \log \left(\frac{\pi_{i+}}{\pi_{1+}} \right) \right\}_{i=2, \dots, n_A} \\ \eta_B^{AB} &= \left\{ \log \left(\frac{\pi_{1j}}{\pi_{11}} \right) \right\}_{j=2, \dots, n_B} \\ \eta_{AB}^{AB} &= \left\{ \log \left(\frac{\pi_{11} \pi_{ij}}{\pi_{i1} \pi_{1j}} \right) \right\}_{i=2, \dots, n_A; j=2, \dots, n_B} \end{aligned} \quad (1)$$

where η_{\bullet}^{\star} denotes the vector of log-linear parameters concerning the variables \bullet in the marginal distribution \star . The symbol $+$ in the probability π denotes the marginalization over the variables in that position.

There are many ways to aggregate the probabilities in the log-linear parameters, the widely diffuse is the baseline criterion, such as in the formula 1, that compares each probability with the probability of the so-called “reference” category, in our case the first one.

However, a more meaningful criterion to describe ordinal variables is the so-called global criterion that compares the cumulative probabilities with the retro-cumulative probabilities. For instance, the logits of an ordinal variables

A evaluated in the marginal A is

$$\eta_A^A = \left\{ \log \left(\frac{\pi(A > a_j)}{\pi(A \leq a_j)} \right) \right\}_{j=\dots, n_A-1} \quad (2)$$

where n_A is the level number of the variable A. For more details see [2]. System of independencies can be easily represented by setting to zero specific parameters defined in particular marginal distributions. In this way, each missed arc (directed or undirected) in the chain graph corresponds to a set of marginal log-linear parameters constrained to zero. In particular, given three variables A, B and C, to describe the sentence A is independent by B given C (denoted with $A \perp B|C$) the parameters η_{AB}^{ABC} and η_{ABC}^{ABC} must be constrained to zero. For more detail, see [4]. The definition of the marginal sets is crucial for representing different independencies at the same time. Rudas et al. 2010 showed how to define the marginal sets corresponding to the LWF CGMs and MR CGMs, [26]. Nicolussi and Colombi, 2017 showed how to define the set of marginals corresponding to a subset of AMP CGMs, [21].

1.3 Learning procedure

In order to select the CG models (the system of conditional relationships) best performing the data, we take advantage of a Bayesian learning algorithm, that is a variant of the posterior distribution over graphical models. The algorithm requires the evaluation of the marginal likelihood, which can be approximated through a maximum likelihood estimation of the Bayesian Information Criterion score (BIC), and the assignment of a prior probability to the graph. We carry out three parallel learning procedures one for any assumption of underlying CGM. At the end we chose the best fitting model among the resulting models from the three procedures, according to the BIC.

The used procedure is based on the algorithm proposed by [5] and it is described in Algorithm 1. Once chosen one graphical model among the ones described above, we set G_0 equal to the graph without missing arcs.

2 Poverty study

2.1 Data and methods

The results presented in this work are gained from the cross-sections Household Monitor survey carried out by TARKI Social Research Center (Monitor-TARKI) during 2012. It counts 4838 statistical units, each of which has a weight that takes into account the gender, the age, the highest education level of the subject and the reference person of the household, the settlement type, and the number of the household members. The survey considers each family member as a statistical unit. Furthermore, we computed the household equivalence income as the sum of household income weighed by the number of household members. The final contingency table with the collected data has 59 on 192 empty cells.

Algorithm 1 Learning procedure

```
 $G_t = G_0$ 
while the number of times we choose, consecutively, the graph  $G_0$  is less than two
times the number of possible edges of the model or graph has been tested against
all the other possible graphs (less than an edge). do
  Randomly select one edge  $(\gamma, \delta) \in (V \setminus E)$ 
  if if the edge is present in  $G_t$  then
    remove it
  else
    add it
  end if
  calculate the score of  $G_t$ :  $score(G_t)$ 
  calculate the probability  $P = \min[(score(G_t) - score(G_0)); 1]$ 
  set  $G_0 = G_t$  with probability  $P$ .
return  $G_3$ 
end while
```

This work aims to describe the system of relationships among factors that we use as an indicator of wealth and social inequality. In particular, the main factor in analyzing is poverty status (P). This factor refers to the household, which is defined poor whether the equivalent income of a household is less than 60% of median national income. The Employment (E) - evaluated as work intensity- the Status of the Flat (F) and the Type of household (T) were considered as social factors. Finally, the Gender (G) of the subject was considered. Below, we list the variables with their categories

P : Poverty [No, Yes];
E : Employment [0; 0.01-0.49; 0.50-0.99;1];
F : Status in the Flat [owner; rent; other];
T : Type of household [One person; Couple or other without children; Lonely parent with children; couple or other with children];
G : Gender [Male, Female].

Within the cells of the contingency table, we collect the personal weights (W) instead of the classical frequencies. In order to model the variables with the CG models we consider three groups of variables (three chain components): Anagraphical (G), Social (E, F and T) and Wealth (P), and we investigate which model is suitable for well describing the relationships among these factors. The choice of the grouping the variables supposes symmetric relationship between the variables within the same component and asymmetric, causal, relationship between variables in different components.

We test all these models by constraints to zero specific log-linear parameters in selected marginal distributions. According to most of CG models taken into account, it is sufficient to use (G); (E,F,T,G),(P,E,F,T,G) as a hierarchical partial ordered list of marginal sets. However, some independencies require addition marginal sets such as (P,E,F,T), (E,G), (F,G) and (T,G), see for more detail [26], [21]. To describe the dependence relationships between the factors we chosen the baseline logit for the categorical variables and global logit

for the ordinal variable (E). In order to select the best fitting model we adopt the procedure displayed in Algorithm 1. All analysis are carry out with the statistical software R ([24]) with the help of packages `hmm`, ([7]), `igraph` ([9]) and `gRbase` ([10]).

2.2 Results

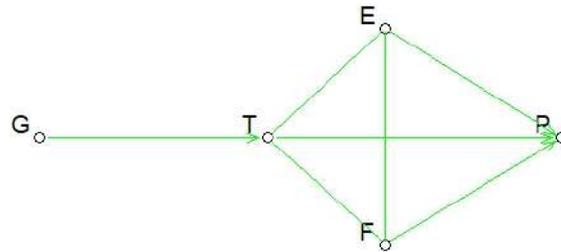


Fig. 1. Chain graph representing the best fitting model

Figure 1 shows the chain graph model which best represents the structure of independence among the factors. Indeed, the three learning procedures lead to the same graph implying the same independence statements. In particular, the graphical model represented in Figure 1 presumes that the gender does not affect the status of poverty given the three social factors considered - $P \perp G|TEF$ - and it does not affect even the work intensity and the status of the flat of the household, given by the type of household - $EF \perp G|T$ -. In Table 2.2, we reported the parameterization associated to the CG in Figure 1. Here, in the first row, we listed the constrained parameters, in correspondence of the marginal distribution where they are defined. Instead, in the second row, we reported the free parameters. The chosen model presents a likelihood ratio statistic of 145, 7348 which leads to an acceptable p-value of 0, 070 if we consider as the degree of freedom the 122 constrained parameters.

The following tables report the estimate free parameters concerning the two-order effects, conferred to the arcs of the graph in Figure 1. Higher absolute values of these parameters denote strong association. With the Wald test, we evaluated whether the parameters are singularly significant, different from zero. The symbols *, **, or *** denote the significance at the 0.05, 0.01, and 0.001, respectively. Table 2 reports the parameters η_{GT}^{GTFE} concerning the only arc (directed) starting from the gender (G). Each parameter is compared with the reference category. The influence of gender (G) on the type of household (T) is not statistically significant, but in the whole model we can not omit this link even if it is weak.

Table 3, 4 and 5 report the parameters describing the component of social variables TFE. The association between T and F is described in Table 3. Except for the parameter associated to the modalities ?rent? of F and ?couple or other without children? of T, the other parameters are not statistically significant.

Marginal	G	GTFE	TFEP	GTFEP
Effect set to zero		GE; GF; GTE; FEP; TFP; GP;GTP;GFP; GEP; GFE; GTF; GTFE	TFEP	GTFP; GTEP; GFEP; GTFEP
Free effect	G	T; E; TF; TE; GT; FE;	P; TP; FP; EP; TEp	

Table 1. Likelihood Ratio Statistic: 145,7348; df 122 (63) p-value 0,070397 (1.669027e-08)

GT	t2	t3	t4
F	-0,35	0,25	-0,37

Table 2. Monitor-TARKI survey: η_{GT}^{GTFE} based on baseline logits for both variables. The reference category is $t1=$ single with no children for the type of household (T) and Male for the gender (G). The other categories are: F=Female, $t2=$ Couple or other without children, $t3=$ lonely parents with children, $t4=$ couple or other with children.

This parameter denotes that the couples without children with a propensity to a rental house are $e^{1.88} = 0.153$ times the lonely subjects without children with a propensity to a rental house. However, the connection T–F is stronger than the G→T (the absolute values of parameters in Table 3 are greater than the ones in Table 2).

F-T	t2	t3	t4
f2	-1,88*	-23,62	1,06
f3	-19,71	0,18	-0,27

Table 3. Monitor-TARKI survey: η_{TF}^{GTFE} based on baseline logits for both variables. The reference category is ?single with no children? for the type of household (T) and ?owner? for the flat (F). f2=?renter?, f3=?other?, t2=?couple or other without children?, t3=?lonely parents with children?, t4=?couple or other with children?.

In Table4 are listed the parameters concerning the association between the work intensity of the household (E) and the type of household (T). This association is statistically significant and the parameters grow to the increasing of work intensity. The first parameter denotes that the propensity to work (work intensity greater than zero) in the couples without children is about $e^{0.92} = 2.51$ times the same in the single without children. This ratio grows when the hours of work grow except in the case of couples with children where the trend is opposite. The connection is always positive but the modality ?couple or other with children? which presents a negative trend.

The parameters in Table 5 describe the arc between the variables F and E. The connection between these two variables is weak and mainly negative in the last modalities. In the component of social variables the most substantial connection lies between the work intensity and the type of household (E ? T).

The last three tables in (6) refer to the directed arcs from the social variables T, F, and E to the poverty indicator P. The variable that strongly affects the

E-TT	t2	t3	t4
> 0	0,92***	0,51*	-0,35
> 0,49	1,34***	0,65*	-0,85*
> 0,99	2,31***	0,71***	-0,94****

Table 4. TARKI survey: η_{TE}^{GTFE} based on baseline logits for T and global logit for E. The reference category is t1: ?single with no children? for the type of household (T). The other categories are t2=?couple or other without children?, t3=?lonely parents with children?, t4=?couple or other with children?.

f2	-1,03	0,43	-1,44
f3	-0,28	-1,19	-0,07

Table 5. Monitor-TARKI survey: η_{FE}^{GTFE} based on baseline logits for F and global logit for E. The reference category is ?owner? for the type of Flat (F). f2=?renter?, t3=?other?.

poverty index is reasonably the work intensity. In particular, the propensity to be poor of employed people ($E > 0$) is $e^{-1.7} = 0.18$ times the propensity to be poor of unemployed people ($E \leq 0$). This gap increases by growing the work intensity. Indeed, the last parameter means that the the propensity to be poor in subjects with full-time job (is about $e^{-2.81} = 0.06$ times the subject having work intensity at most equal to 0.99. Even the type of household has a strong and significant influence on poverty. These parameters suggest that the propensity to be poor for a couple or other without children (T=t2) is $e^{1.12} = 0.33$ times with respect to the single without children. This trend changes when we consider a family with children. For instance, the couples or other with children (T=t4) have about 3.49 times more possibility than a single without children to be poor. Finally, there is a lack of statistical evidence that the type of contract flat (F) affects the poverty index.

TP	Yes	FP	Yes	EP	Yes
t2	-1,12 ***	f2	0,81	> 0	-1,7 ***
t3	0,37	f3	1,06	> 0,49	-2,07
t4	1,25***			> 0,99	-2,81***

Table 6. Monitor-TARKI survey: η_{TP}^{TFEP} , η_{FP}^{TFEP} and η_{EP}^{TFEP} based on baseline logits for both T, F and P and global logit for E. The reference category is *owner* for the type of Flat (F), *single without children* for (T) and *not poor* for the poverty index (P). The other categories are f2=*rent*, t3=*other*, t2=*Couple or other without children*, t3=*lonely parents with children*, t4=*couple or other with children*.

3 Conclusion

The analysis of the Hungarian study from TARKI survey (2012), Hungarian shows that the gender of the subject (G) does not affect the poverty (P) fixed the type of household (T), the work intensity (E), and the status of flat (F).

Further, all the social variables (respectively, work intensity, the status of flat and type of household) affect the poverty status. In detail, there is an effect of gender only on the type of household and the effects of all the social variables E, F, and T on the poverty status (P). The work intensity shows the strongest link with the poverty and highlights a trade-off between poverty and work intensity. The other significant connection is between the type of household and poverty. In this case, the estimated model shows that singles without children are more likely to be poor than a couple without children but are less likely to be poor than lonely parents or couples with children.

4 Acknowledgments

The research leading to these results has received support under the European Commission's 7th Framework Programme (FP7/2013- 2017) under grant agreement n. 312691, InGRID - Inclusive Growth Research Infrastructure Diffusion. The responsibility for all conclusions drawn from the data lies entirely with the authors.

References

1. Andersson, S. A., Madigan, D., and Perlman, M. D. (2001). Alternative Markov properties for chain graphs. *Scandinavian journal of statistics*, 28(1), 33-85.
2. F. Bartolucci, R. Colombi, A. Forcina, An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints, *Statistica Sinica* 17(2): 691.(2007)
3. J. Janssen and R. Manca. *Semi-Markov risk models for finance, insurance and reliability*, Springer, New York, 2007.
4. W. P. Bergsma, T. Rudas. Marginal models for categorical data. *The Annals of Statistics* 30.1: 140-159.(2002)
5. R. Breen, (2008). Statistical models of association for comparing cross-classifications. *Sociological Methods & Research*, 36(4), 442-461
6. Y. Chzhen and J. Bradshaw(2012). Lone parents, poverty and policy in the European Union. *Journal of European Social Policy*, 22(5), 487-506.
7. R. Colombi, S. Giordano, M. Cazzaro, and the R Development Core Team. hmmm: Hierarchical Multinomial Marginal Models. R package version 1.0-1 (2013)
8. Cox, D. R and Wermuth, N. (1993) Linear dependencies represented by chain graphs, *Statistical Science*, 8(3) 204—218.
9. G. Csardi, T. Nepusz. The igraph software package for complex network research, *InterJournal, Complex Systems* 1695. 2006. <http://igraph.org>
10. C. Dethlefsen, S. Hjsgaard (2005). A Common Platform for Graphical Models in R: The gRbase Package. *Journal of Statistical Software*, 14(17), 1-12. URL <http://www.jstatsoft.org/v14/i17/>.
11. M. Drton, Discrete chain graph models, *Bernoulli* 15(3): 736- 753.(2009)
12. Frydenberg, M. (1990). The chain graph Markov property. *Scandinavian Journal of Statistics*, 333-353.
13. M. de Graaf-Zijl and B. Nolan, B. (2011). Household joblessness and its impact on poverty and deprivation in Europe. *Journal of European Social Policy*, 21(5), 413-431.

14. S. L. Lauritzen. Graphical Models. Oxford University Press, 1996.
15. Lauritzen, S. L., and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, 31-57.
16. H. Lohmann(2011). Comparability of EU-SILC survey and register data: The relationship among employment, earnings and poverty. *Journal of European social policy*, 21(1), 37-54.
17. A. Forcina, M. Lupparelli and G.M. Marchetti (2010). Marginal parameterizations of discrete models defined by a set of conditional independencies. *Journal of Multivariate Analysis*, 101(10), 2519-2527.
18. G.M. Marchetti, M. Lupparelli(2011). Chain graph models of multivariate regression type for categorical data. *Bernoulli*. 17: 827-844.
19. R. Nemeth, T. Rudas, (2013a). On the application of discrete marginal graphical models. *Sociological Methodology*, 43(1), 70- 100.
20. R. Nemeth, T. Rudas(2013). Discrete Graphical Models in Social Mobility Research-A Comparative Analysis of American, Czechoslovakian and Hungarian Mobility before the Collapse of State Socialism. *Bulletin of Sociological Methodology*, 118(1), 5-21.
21. F. Nicolussi, R. Colombi (2017). "Graphical Model of type II: a smooth subclass". *Bernoulli*, 23(2), 863-883.
22. F. Nicolussi, and F. Mecatti (2014b). A smooth subclass of graphical models for chain graph: towards measuring gender gaps. *Quality & Quantity*, 1-15.
23. V. Polin and M. Raitano(2014). Poverty Transitions and Trigger Events across EU Groups of Countries: Evidence from EUSILC. *Journal of Social Policy*, 43(04), 745-772.
24. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
25. Richardson, T., and Spirtes, P. (2002). Ancestral graph Markov models. *The Annals of Statistics*, 30(4), 962-1030.
26. T. Rudas, W. P. Bergsma, and R. Nemeth(2010). Marginal log-linear parameterization of conditional independence models. *Biometrika* 97.4 : 1006-1012.

Identifiability of Finite Mixture Models with underlying Normal Distribution

Cédric Noel¹ and Jang Schiltz²

¹ University of Luxembourg and IUT of Thionville-Yutz, University of Lorraine, Espace Cormontaigne Impasse Alfred Kastler F-57970 Yutz, France (E-mail: cedric.noel@univ-lorraine.fr)

² Department of Finance, University of Luxembourg, 6, rue Richard Coudenhove-Kalergi L-1359 Luxembourg, Luxembourg (E-mail: jang.schiltz@uni.lu)

Abstract. In this paper, we show under which conditions generalized finite mixture with underlying normal distribution are identifiable in the sense that a given dataset leads to a uniquely determined set of model parameter estimations up to a permutation of the clusters.

Keywords: Identifiability, Finite Mixture Models.

1 Introduction

Identifiability of the parameters is a necessary condition for the existence of consistent estimators for any statistical model. Without identifiability, there might be several solution for the parameter estimation problem and numerical algorithms risk to find only part of these solutions. Worse, the researcher fitting the model might not even be aware that the solution his computer found is only one of many possibilities.

Identifiability of distributions has been an important research topic in the 1960s. Teicher ([6]) proved that the class of all mixtures of one-dimensional normal distributions is identifiable. Yakowitz and Spragins ([9]) extended this result five years later to the class of all Gaussian mixtures.

For a long time, it was believed that identifiability for linear regression mixtures with Gaussian errors follows directly from these results. DeSarbo and Cron ([1]) even make that claim explicitly. Hennig ([2]) only showed in 2000 that that statement is not correct in general by constructing counter-examples. Hennig investigated the identifiability of the parameters of models for data generated by different linear regression distributions with Gaussian errors.

In this paper, we extend his results to finite mixture models in which the typical trajectories in the different clusters do not just follow a line, but a polynomial of any degree.

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



The remainder of this article is structured as follows. In section two, we present the class of finite mixture models we are interested in. In section three, we present some basic results about the identifiability of mixtures of distributions. In section four, finally, we prove under which conditions finite mixture models are identifiable.

2 Finite Mixture Models

Starting from a collection of individual trajectories, the aim of finite mixture models is to divide the population into a number of homogenous sub-populations and to estimate, at the same time, a typical trajectory for each sub-population (Nagin [3]).

More, precisely, consider a population of size N and a variable of interest Y . Let $Y_i = y_{i_1}, y_{i_2}, \dots, y_{i_T}$ be T measures of the variable Y , taken at times t_1, \dots, t_T for subject number i . To estimate the parameters defining the shape of the trajectories, we need to fix the number K of desired subgroups. Denote the probability of a given subject to belong to group number k by π_k .

The objective is to estimate a set of parameters $\Omega = \{\pi_k, \beta_0^k, \beta_1^k, \dots; k = 1, \dots, K\}$ which allow to maximize the probability of the measured data. The particular form of Ω is distribution specific, but the β parameters always perform the basic function of defining the shapes of the trajectories. In Nagin's finite mixture model (Nagin [3]), the shapes of the trajectories are described by a polynomial function of age or time. Assume that for a subject in group k

$$y_{it} = \sum_{j=1}^s \beta_j^k a_{it}^j + \varepsilon_{it}, \quad (1)$$

where a_{it} denotes the age of subject i at time t , s the degree of the polynomial describing the trajectories in the different groups and ε_{it} is a disturbance assumed to be normally distributed with a zero mean and a constant standard deviation σ . The likelihood of the data is then given by

$$L = \prod_{i=1}^N \sum_{k=1}^K \pi_k \prod_{t=1}^T g_k(y_{it}), \quad (2)$$

where $g_k(y_{it})$ is the probability distribution function of y_{it} given membership in group k . In this paper we restrict ourselves to normal distributions.

The disadvantage of the basic model is that the trajectories are static and do not evolve in time. Thus, Nagin introduced several generalizations of his model in his book (Nagin [3]). Among others, he introduced a model allowing to add covariates to the trajectories. Let z_1, \dots, z_M be M covariates potentially influencing Y . We are then looking for trajectories

$$y_{it} = \sum_{j=0}^s \beta_j^k a_{it}^j + \alpha_1^j z_1 + \dots + \alpha_M^j z_M + \varepsilon_{it}, \quad (3)$$

where ε_{it} is normally distributed with zero mean and a constant standard deviation σ . The covariates z_m may depend or not upon time t .

But even this generalized model still has two major drawbacks. First, the influence of the covariates in this model is unfortunately limited to the intercept of the trajectory. This implies that for different values of the covariates, the corresponding trajectories will always remain parallel by design, which does not necessarily correspond to reality.

Secondly, in Nagin's model, the standard deviation of the disturbance is the same for all the groups. That too is quite restrictive. One can easily imagine situations in which in some of the groups all individual are quite close to the mean trajectory of their group, whereas in other groups there is a much larger dispersion. To address and overcome these two drawbacks, Schiltz ([5]) proposed the following generalization of Nagin's model.

Let x_1, \dots, x_M and z_{i_1}, \dots, z_{i_T} be covariates potentially influencing Y . Here the x variables are covariates not depending on time like gender or cohort membership in a multicohort longitudinal study and the z variable is a covariate depending on time like being employed or unemployed. They can of course also designate time-dependent covariates not depending on the subjects of the data set which still influence the group trajectories, like GDP of a country in case of an analysis of salary trajectories.

The trajectories in group k will then be written as

$$y_{i_t} = \sum_{j=0}^s \left(\beta_j^k + \sum_{m=1}^M \alpha_m^k x_m + \gamma_j^k z_{i_t} \right) a_{i_t}^j + \varepsilon_{i_t}^k, \quad (4)$$

where the disturbance $\varepsilon_{i_t}^k$ is normally distributed with mean zero and a standard deviation σ_k , constant inside group k , but different from one group to another. Since, for each group, this model is just a classical fixed effects model for panel data regression (see Wooldridge ([8])), it is well defined and we can get consistent estimates for the model parameters.

That model allows obviously to overcome the drawbacks of Nagin's model. The standard deviation of the uncertainty can vary across groups and the trajectories depend in a nonlinear way on the covariates.

Whereas the basic model is usually identified under very mild conditions, it is obvious that this is no longer true in all generality for the two generalized models. We will investigate this in the remainder of this paper.

3 Identifiability

In 1963, Teicher ([6]) showed the following result for mixtures of normal distributions.

Proposition 1. *The class of all mixtures of one-dimensional normal distributions is identifiable.*

We will use that proposition to prove under which conditions finite mixture models are identifiable.

Consider the distribution f of a finite mixture model.

$$f(y_i; \Omega) = \sum_{k=1}^K \pi_k g_k(y_i; \beta^k), \quad (5)$$

which is equivalent to

$$F(y_i; \Omega) = \sum_{k=1}^K \pi_k G_k(y_i; \theta^k), \quad (6)$$

where F and $G : k$ denote the cumulative distribution functions (cdf's) of f and g_k respectively.

Let $\mathcal{F} = \{F(y; \omega), y \in \mathbb{R}^T, \omega \in \mathbb{R}_K^{s+2}\}$ be a family of T -dimensional cdf's indexed by a parameter set ω , such that $F(y; \omega)$ is measurable in $\mathbb{R}^T \times \mathbb{R}_K^{s+2}$. The $s+2$ -dimensional cdf $H(x) = \int_{\mathbb{R}_K^{s+2}} F(y; \omega) dG(\omega)$ is the image of the above mapping, of the $s+2$ -dimensional cdf G . The distribution H is called the mixture of \mathcal{F} and G its mixing distribution. Let \mathcal{G} denote the class of all $s+2$ -dimensional cdf's G and \mathcal{H} the induced class of mixtures H .

Then \mathcal{H} is said to be identifiable if Q is a one-to-one map from \mathcal{G} onto \mathcal{H} .

The set \mathcal{H} of all finite mixtures of class \mathcal{F} of distributions is the convex hull of \mathcal{F} .

$$\mathcal{H} = \left\{ H(y) : H(y) = \sum_i c_i F(y, \omega_i), c_i > 0, \sum_i c_i = 1, F(y, \omega_i) \in \mathcal{F} \right\}. \quad (7)$$

In this context, the definition of identifiability implies that \mathcal{F} generates an identifiable finite mixture model if and only if

$$\sum_{i=1}^N c_i F_i = \sum_{i=1}^M c'_i F'_i \quad (8)$$

implies that $N = M$ and for each i , $1 \leq i \leq N$ there is some j , $1 \leq j \leq N$, such that $c_i = c'_j$ and $F_i = F'_j$.

We can then easily prove the following characterization of identifiability.

Theorem 1. *A necessary and sufficient condition for the class \mathcal{H} of all finite mixtures of the family \mathcal{F} to be identifiable is that \mathcal{F} is a linearly independent family over the field of real numbers.*

We denote by $\langle A \rangle$ the span of A over the real numbers.

Proof. Necessity.

Suppose that the family \mathcal{F} is not linearly independent. Then, there exist an integer N and N real numbers a_i , at least one of them not being zero, such that, $\sum_{i=1}^N a_i F_i = 0$. Without loss of generality, we can suppose that $a_i < 0 \Leftrightarrow i \leq M$. Thus, $\sum_{i=1}^M |a_i| F_i = \sum_{i=M+1}^N |a_i| F_i$.

Since the F_i are cdf's, this implies that

$$\lim_{y \rightarrow (+\infty, \dots, +\infty)} \sum_{i=1}^M |a_i| F_i(y) = \lim_{y \rightarrow (+\infty, \dots, +\infty)} \sum_{i=M+1}^N |a_i| F_i(y), \quad (9)$$

hence

$$\sum_{i=1}^M |a_i| = \sum_{i=M+1}^N |a_i|. \quad (10)$$

Now, define c_i for each i by

$$c_i = \frac{|a_i|}{\sum_{i=M+1}^N |a_i|}.$$

Then, $\sum_{i=1}^M c_i = \sum_{i=M+1}^N c_i = 1$ and

$$\sum_{i=1}^M c_i F_i = \sum_{i=M+1}^N c_i F_i.$$

Thus we have two different distinct representations of the same mixture and therefore \mathcal{H} is not identifiable.

Sufficiency.

If \mathcal{F} is a linearly independent family, there exists a basis of $\langle \mathcal{F} \rangle$. If we suppose that \mathcal{H} is non identifiable there exist two distinct representations of the same mixture. Therefore $\mathcal{H} \subset \langle \mathcal{F} \rangle$ which contradicts the uniqueness of the representation property of bases. \square

We will now analyze the identifiability of some classes of generalized finite mixture models.

4 Identifiability of a class of finite mixture models

We will prove the identifiability of a big subclass of the generalized finite mixture model presented in section 2. Consider indeed the model defined by

$$Y_{it} = f(a_{it}; \beta^k, \delta^k) + \varepsilon_{it}^k = \beta^k A_{it} + \delta^k W_{it} + \varepsilon_{it}^k, \quad (11)$$

that we can write as

$$Y_i = \beta^k A_i + \delta^k W_i + \varepsilon_i^k, \quad (12)$$

with $Y_i = (Y_{i1}, \dots, Y_{iT})$, $A_i = (A_{i1}, \dots, A_{iT})$, $W_i = (W_{i1}, \dots, W_{iT})$ and $\varepsilon_i^k \sim \mathcal{N}(0; \sigma_k I_T)$.

Thus, $Y_i \sim \mathcal{N}(\beta^k A_i + \delta^k W_i, \sigma_k I_T)$.

Hennig ([2]) showed the identifiability of clusterwise linear regression models in the case of a one-dimensional normal distribution. We extend this results to the case of multi-dimensional normal distributions and polynomial trajectories.

We can write

$$\mathcal{L}((Y_i)_{i \in I}) = \bigotimes_{i \in I} F_{A_i, W_i, J}, \quad (13)$$

where $F_{A_i, W_i, J}(Y_i) = \int_{T_1} \Phi_{0, \Sigma}(Y_i - \beta_k A_i - \delta_k W_i) dJ(\beta, \sigma^2)$ with $T_1 = \mathbb{R}^{s+1} \times \mathbb{R}_0^+$, $J \in \Omega_1 = \mathcal{J}(T_1)$ and $\Sigma = \sigma I_T$.

$\mathcal{J}(T_1)$ denotes the set of mixing distributions with finite support on the parameter set T . $S(J)$ is the support set of $J \in \mathcal{J}(T_1)$. Thus, $K = |S(J)|$ is the number of mixture components and the elements of $\mathcal{J}(T_1)$ are distributions generating parameter values $(\beta^1, \sigma_1^2), \dots, (\beta^K, \sigma_K^2)$ for K clusters with probability $J(\beta^1, \sigma_1^2), \dots, J(\beta^K, \sigma_K^2)$. I is some index set, here $I = \{1, \dots, N\}$ since we suppose that we analyze data from a population of size N . \bigotimes denotes the independent product of distributions.

Identifiability of a model means that knowing the data distribution $\mathcal{L}(Y_i), i \in I$, one can identify uniquely the mixing distribution J . That is, no two distinct sets of parameters lead to the same data distribution.

4.1 Nagin's base model

Nagin's base model can be written as

$$\mathcal{C}_1 = \left(F_{A, J} : F_{A, J} = \bigotimes_{i \in I} F_{A_i, J} \right)_{J \in \Omega_1}$$

In that case, identifiability means that, knowing the data distributions $\mathcal{L}(Y_i)_{i \in I}$, we can uniquely identify the mixing distribution J and two distinct sets of parameters $(\beta^1, \sigma_1^2, J(\beta^1, \sigma_1^2)), \dots, (\beta^K, \sigma_K^2, J(\beta^K, \sigma_K^2))$ and $(\beta'^1, \sigma_1'^2, J(\beta'^1, \sigma_1'^2)), \dots, (\beta'^K, \sigma_K'^2, J(\beta'^K, \sigma_K'^2))$ lead to different data distributions.

Theorem 2. *Let $h_j = \min \{q : \{A_{ij}, i \in I\} \subseteq \cup_{i=1}^q H_i \quad H_i \in \mathcal{H}_{n-1}\}$.*

If there exist j such that $|S(J)| < h_j, \forall J$ then \mathcal{C}_1 is identifiable.

Proof. We need to show only that $F_{A_i, J} = F_{A_i, \tilde{J}} \Rightarrow J = \tilde{J}$ because J contains all information to define the common distribution $F_{A_i, J}$ of $(Y_i)_{i \in I}$.

Suppose that $F_{A_i, J} = F_{A_i, \tilde{J}}$ and $J \neq \tilde{J}$. Without loss of generality we can assume that $|S(J)| \geq |S(\tilde{J})|$. Thus there exists $(\beta^1, \sigma_1) \in S(\tilde{J})$ such that

$$J\{(\beta^1, \sigma_1^2)\} \neq \tilde{J}\{(\beta^1, \sigma_1^2)\}. \quad (14)$$

$F_{A_i, J} = F_{A_i, \tilde{J}}$ implies the equality of the marginal Gaussian mixtures for all $A_i, i \in I$ and

$$F_{A_i, J}(Y_i) = \int_{T_1} \Phi_{\beta_k A_i, \Sigma}(Y_i) dJ(\beta, \sigma^2) \quad (15)$$

$$= F_{A_i, \tilde{J}}(Y_i) = \int_{T_1} \Phi_{\beta_k A_i, \Sigma}(Y_i) d\tilde{J}(\beta, \sigma^2). \quad (16)$$

The identifiability of finite Gaussian mixtures then implies, for $i \in I$

$$J\{(\beta, \sigma^2) : (\beta A_i, \sigma^2) = (\beta^1 A_i, \sigma_1^2)\} = \tilde{J}\{(\tilde{\beta}, \tilde{\sigma}^2) : (\tilde{\beta} A_i, \tilde{\sigma}^2) = (\beta^1 A_i, \sigma_1^2)\} \quad (17)$$

The idea of the proof is that the restriction to $|S(\tilde{J})|$ ensures the existence of a matrix A_i whose marginal mixture $\mathcal{N}(\beta_1 A_i, \sigma_1^2)$, parameterized by J , cannot be explained by $(\tilde{\beta}, \tilde{\sigma}^2) \in S(\tilde{J})$ if $\tilde{\beta} \neq \beta_1$. Therefore $S(\tilde{J})$ must contain (β_1, σ_1^2) .

Suppose that for all $(\beta, \sigma^2) \in S(J)$, and in particular for (β^1, σ_1^2) , there exists $i(\beta) \in I$ such that

$$\forall (\tilde{\beta}, \tilde{\sigma}^2) \in S(\tilde{J}) : \beta A_{i(\beta)} = \tilde{\beta} A_{i(\beta)} \Rightarrow \beta = \tilde{\beta}. \quad (18)$$

The definition of $A_i = A_{i(\beta^1)}$ implies that

$$\forall S(\tilde{J}) \ni (\tilde{\beta}, \tilde{\sigma}^2) \neq (\beta^1, \sigma_1^2) : (\tilde{\beta} A_i, \tilde{\sigma}^2) \neq (\beta^1 A_i, \sigma_1^2). \quad (19)$$

Thus, using (27) and (17),

$$\tilde{J}\{(\beta^1, \sigma_1^2)\} = J\{(\beta, \sigma) : (\beta A_i, \sigma^2) = (\beta^1 A_i, \sigma_1^2)\}. \quad (20)$$

But $J\{(\beta, \sigma) : (\beta A_i, \sigma^2) = (\beta^1 A_i, \sigma_1^2)\} \neq 0$ because it contains (β^1, σ_1^2) . For the same reason, $\tilde{J}\{(\beta^1, \sigma_1^2)\} \neq 0$.

Hence (19) implies that $(\beta^1 A_i, \sigma_1^2) \in S(\tilde{J})$.

By (14), $\tilde{J}\{(\beta^1, \sigma_1^2)\} \neq J\{(\beta^1, \sigma_1^2)\}$. Consequently, equation (20) implies that

$$\exists S(J) \ni (\beta^2, \sigma_2^2) \neq (\beta^1, \sigma_1^2) : (\beta^2 A_i, \sigma_2^2) = (\beta^1 A_i, \sigma_1^2). \quad (21)$$

Consider $A_i = A_{i(\beta^2)}$ and apply the same arguments than above to get $(\beta^2 A_i, \sigma_2^2) \in S(\tilde{J})$. This result leads to a contradiction between (19) and (21). Indeed, $(\beta^2, \sigma_2^2) \in S(\tilde{J})$ and $(\beta^2, \sigma_2^2) \neq (\beta^1, \sigma_1^2)$. By (19), $(\beta^2 A_i, \sigma_2^2) \neq (\beta^1 A_i, \sigma_1^2)$ and by (21), $(\beta^2 A_i, \sigma_2^2) = (\beta^1 A_i, \sigma_1^2)$.

Thus there exists some $(\beta, \sigma) \in S(J)$ such that $\forall i \in I \forall (\tilde{\beta}, \tilde{\sigma}^2) \in S(\tilde{J}) : \beta A_i = \tilde{\beta} A_i \Rightarrow \beta \neq \tilde{\beta}$.

Hence

$$\{A_{ij}, i \in I, j = 1 \cdots T\} \subset \cup_{(\tilde{\beta}, \tilde{\sigma}^2) : \tilde{\beta} \neq \beta} \{x : \beta x = \tilde{\beta} x\}. \quad (22)$$

Therefore $\cup_{(\tilde{\beta}, \tilde{\sigma}^2) : \tilde{\beta} \neq \beta} \{x : \beta x = \tilde{\beta} x\}$ is composed by $|S(\tilde{J})|$ different hyperplanes.

So for $j = 1 \cdots T$, $h_j \leq |S(\tilde{J})| \leq |S(J)|$. \square

4.2 Addition of covariates independent of the clusters

Let us now add covariates to the model that are independent of the K groups. Define

$$\mathcal{C}_2 = \left(F_{A,J} : F_{A,J} = \bigotimes_{i \in I} F_{A_i, W_{i,J}} \right)_{J \in \Omega_1}, \quad (23)$$

$$\mathcal{C}_{2A} = \left(F_{A,J} : F_{A,J} = \bigotimes_{i \in I} F_{A_i, J} \right)_{J \in \Omega_1}, \quad (24)$$

$$\mathcal{C}_{2W} = \left(F_{A,J} : F_{A,J} = \bigotimes_{i \in I} F_{W_{i,J}} \right)_{J \in \Omega_1}. \quad (25)$$

We have then the following identifiability result.

Theorem 3. *If \mathcal{C}_{2A} and \mathcal{C}_{2W} are identifiable and W_{ij} is not a multiple of A_{ij} , for all i, j , then \mathcal{C}_2 is identifiable.*

Since the covariates are just a linear addition to the model, the proof follows directly from the 2 following propositions.

Proposition 2. *\mathcal{C}_{2A} is identifiable if and only if $d_k < T$ for all $1 \leq k \leq K$ and the a_{it} are distinct, for all values of i and t .*

Proof. We need to show that $F_{A_i, J} = F_{A_i, \tilde{J}} \Leftrightarrow J = \tilde{J}$. Suppose that

$$F_{A_i, J}(Y_i) = \int_{T_1} \Phi_{\beta^k A_i, \Sigma}(Y_i) dJ(\beta, \sigma^2) \quad (26)$$

$$= F_{A_i, \tilde{J}}(Y_i) = \int_{T_1} \Phi_{\beta^k A_i, \Sigma}(Y_i) d\tilde{J}(\beta, \sigma^2). \quad (27)$$

By identifiability of finite Gaussian mixtures, the equality above is equivalent, for $i \in I$, to:

$$J \{(\beta, \sigma^2) : (\beta A_i, \sigma^2) = (\mu_1, \sigma_1^2)\} = \tilde{J} \{(\tilde{\beta}, \tilde{\sigma}^2) : (\tilde{\beta} A_i, \tilde{\sigma}^2) = (\mu_1, \sigma_1^2)\} \quad (28)$$

for some (μ_1, σ_1) .

Assume that there exists $\tilde{\beta}$ in $S(\tilde{J})$ such that $\beta A_i = \tilde{\beta} A_i$ for some $\beta \in S(J)$. This means that $1 \leq t \leq T$, $\beta A_{it} = \tilde{\beta} A_{it}$ but $\tilde{\beta} \neq \beta$ for all $\beta \in S(J)$. If $d_k < T$ and a_{it} are different $\forall t \leq T$, we have 2 different polynomials of degree strictly smaller than T that intersect in T points. Thus $\beta = \tilde{\beta}$.

If we know cluster membership for each value Y_i , we can write

$$Y_k = A_k \beta_k^t,$$

$$\text{where } Y_k = \begin{pmatrix} Y_{k11} \\ \vdots \\ Y_{kn_k T} \end{pmatrix} \text{ and } A_k = \begin{pmatrix} 1 & a_{11} & \cdots & a_{11}^{d_k-1} \\ \vdots & & & \vdots \\ 1 & a_{n_k T} & \cdots & a_{n_k T}^{d_k-1} \end{pmatrix}.$$

Since $\begin{pmatrix} 1 & a_{11} & \cdots & a_{11}^{d_k-1} \\ \vdots & & & \vdots \\ 1 & a_{n_k T} & \cdots & a_{n_k T}^{d_k-1} \end{pmatrix}$ is a Vandermonde matrix and $d_k < T$, which is required for the matrix to be invertible, the invertibility condition is guaranteed to hold if all the a_{it} values are distinct.

So

$$\beta_k = Y_k (A_k^t A_k)^{-1} A_k. \quad (29)$$

□

Proposition 3. *If for all $1 \leq t, t' \leq T$ and for $i, j \in I$, $a_{it} = a_{jt}$ and $a_{it} \neq a'_{it}$, \mathcal{C}_{2A} is identifiable if and only if $d_k < T$ for all $1 \leq k \leq K$.*

Proof. In this case the matrix $\begin{pmatrix} 1 & a_{11} & \cdots & a_{11}^{d_k-1} \\ \vdots & & & \vdots \\ 1 & a_{n_k T} & \cdots & a_{n_k T}^{d_k-1} \end{pmatrix}$ becomes $\begin{pmatrix} 1 & a_{11} & \cdots & a_{11}^{d_k-1} \\ \vdots & & & \vdots \\ 1 & a_{1T} & \cdots & a_{1T}^{d_k-1} \end{pmatrix}$ and is invertible if $d_k < T$ and $a_{it} \neq a_{it'}$, $1 \leq t, t' \leq T$. □

Numerical example

Let us illustrate the two previous propositions by an example. To keep everything the easiest possible, we consider an example with just two clusters with sizes $\pi_1 = \pi_2 = \frac{1}{2}$ and two time-points 1 and 2. For the sake of simplicity, we also suppose that the variability of the error term is the same for both groups and we take $\sigma = 0.1$.

To emphasize the difference between the identifiability of a mixture of probability distributions and the identifiability of finite mixture models, we point out that proposition 1 implies that the mixture distribution $\frac{1}{2}\mathcal{N}(\mu_{1t}; 0.1) + \frac{1}{2}\mathcal{N}(\mu_{2t}; 0.1)$ is always identifiable.

Proposition 3 tells us that finite mixture models with polynomial trajectories will be identifiable as long as the degree of the polynomials is at most 1, since $T = 2$. To illustrate this, we simulate 50 samples of 100 observations, once with linear trajectories and once with polynomials of degree 2. More precisely, we use the following parameter values

- $\beta^1 = (3, -2)$ and $\beta^2 = (0, 2)$ for the linear model ;
- $\beta^1 = (10, -12.5, 3.5)$ and $\beta^2 = (-2, 5, -1)$ for the polynomial model.

We then use our R package `trajeR` (Noel and Schiltz [4]) to fit these 100 samples and illustrate the result by means of parallel coordinate plots (Wegman [7]).

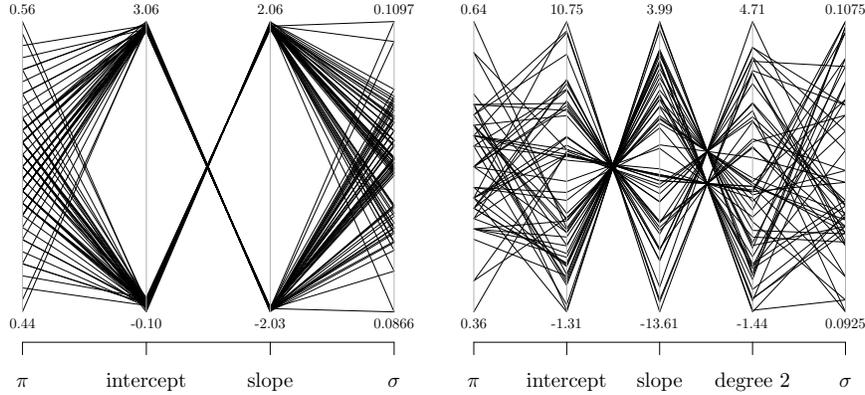


Fig. 1: Parallel coordinate plots of the estimated parameters for 50 simulated samples for linear and parabolic trajectories.

Figure 1 shows the result. On the left side, we see the different parameter estimations for the linear model. We see that in this case all parameter estimations give roughly the same result. There are 2 solutions for the different trajectory shape parameters, corresponding to the 2 clusters, one cluster with a trajectory defined by an intercept of 3 and a slope of -2 and one cluster with a trajectory defined by an intercept of 0 and a slope of 2. The estimation for the group sizes vary between 0.44 and 0.56 and the standard deviation of the error term are estimated as being between 0.087 and 0.110. The right part of the graph shows the parameter estimation for the parabolic model. In this case the trajectory shape parameters cannot be precisely estimated and there is no indication of a two-group solution. This is a clear indication of the non identifiability of the model.

4.3 The generalized model

Now consider the generalized finite mixture model

$$Y_i = \beta^k A_i + \delta^k W_i + \epsilon_i.$$

We then have the following result.

Proposition 4. *The model is identifiable if*

- $d_k < T$ for all $1 \leq k \leq K$ and all a_{it} are distinct, for all i, t ;
- W_k has full rank for all $1 \leq k \leq K$;
- $rk(A_k, W_k) = rk(A_k) + rk(W_k)$ for all $1 \leq k \leq K$ where $rk(\cdot)$ denotes the rank of a matrix, A_k is defined like in the proof of proposition 2, and W_k are the elements W_i corresponding to A_k .

Proof. If W_i does not depend on time, the trajectories of all clusters are just translations of each other. Thus, the first condition of the proposition implies that all trajectory parameters are identifiable and since $rk(A_k, W_k) =$

$rk(A_k) + rk(W_k)$ we can determine δ^k too.

In the general case, suppose that $d_k < T$ for all $1 \leq k \leq K$. Then for any integer c , a mixture of c components of the form $\sum_{k=1}^c \pi_k \mathcal{N}(\beta^k A_{it}, \sigma_k)$ is identifiable.

If we know the cluster membership of each value Y_i , we can determine β^k as in equation (29) by $\beta^k = Y_k (A_k^t A_k)^{-1} A_k$.

Denote $P_k = A_k^t (A_k A_k^t)^{-1} A_k$ and $R_k = I - P_k$. Then,

$$\beta^k A_k + \delta^k W_k = \beta^k A_k + \delta^k W_k P_k + \delta^k W_k (I - P_k) \quad (30)$$

$$= \beta^k A_k + \delta^k W_k A_k^t (A_k A_k^t)^{-1} A_k + \delta^k W_k (I - P_k) \quad (31)$$

$$= \left(\beta^k + \delta^k W_k A_k^t (A_k A_k^t)^{-1} \right) A_k + \delta^k W_k R_k \quad (32)$$

$$= \left(\beta^k + \delta^k W_k A_k^t (A_k A_k^t)^{-1} \delta^k \right) \begin{pmatrix} A_k \\ W_k R_k \end{pmatrix} \quad (33)$$

$$= \lambda_k V. \quad (34)$$

Suppose $\lambda_k V = 0$ for some λ_k . Then give $\beta^k A_k + \delta^k W_k = 0$, hence $\beta^k = \delta^k = 0$ by linear independence of the columns of A_k and W_k . So V is a $T + rk(W_k)$ matrix of full rank.

Since $Y_k = \lambda_k V + \varepsilon$, we have

$$\hat{\lambda}_k = Y_k (V V^t)^{-1} V^t \quad (35)$$

$$= (A_k W_k R_k) \begin{pmatrix} A_k A_k^t & A_k R_k^t W_k^t \\ W_k R_k A_k^t & W_k R_k R_k^t W_k^t \end{pmatrix}^{-1} \quad (36)$$

$$= (A_k W_k R_k) \begin{pmatrix} A_k A_k^t & A_k R_k W_k^t \\ W_k R_k A_k^t & W_k R_k R_k W_k^t \end{pmatrix}^{-1}. \quad (37)$$

Since $R_k = I - A_k^t (A_k A_k^t)^{-1} A_k$, we have $A_k R_k = R_k A_k^t = 0$ and $P_k^2 = P_k$. Moreover,

$$\hat{\lambda}_k = Y_k (V V^t)^{-1} V^t \quad (38)$$

$$= Y_k (A_k W_k R_k) \begin{pmatrix} A_k A_k^t & 0 \\ 0 & W_k R_k W_k^t \end{pmatrix}^{-1} \quad (39)$$

$$= Y_k \left(A_k (A_k A_k^t)^{-1} W_k R_k (W_k R_k W_k^t)^{-1} \right) \quad (40)$$

$$= \left(Y_k A_k (A_k A_k^t)^{-1} Y_k W_k R_k (W_k R_k W_k^t)^{-1} \right). \quad (41)$$

Thus

$$\hat{\delta}^k = Y_k W_k R_k (W_k R_k W_k^t)^{-1}$$

and

$$\hat{\beta}^k = Y_k A_k (A_k A_k^t)^{-1} - \hat{\delta}^k W_k A_k^t (A_k A_k^t)^{-1}.$$

Hence all parameters are identified. \square

Numerical example

Let us illustrate proposition 4 by an example. As in the previous example, we consider a model with just two clusters of sizes $\pi_1 = \pi_2 = \frac{1}{2}$, two time-points 1 and 2 and a constant variability of the error term of $\sigma = 0.1$. Furthermore, we use the shape description parameters $\beta_1 = (3, -2)$ and $\beta_2 = (0, 2)$ and fix $\delta_1 = 2$ and $\delta_2 = -3$. We will study 3 types of models, defined by the following supplementary conditions.

- The covariate W is independent of time and only takes values 0 or 1;
- the covariate is time dependent but in a nonlinear way;
- the covariate is time dependent in a linear way;

To illustrate this, we simulate 50 samples of 100 observations, We then use our R package `trajeR` (Noel and Schiltz [4]) to fit these 100 samples and illustrate the result by means of parallel coordinate plots (Wegman [7]).

We can see on figure 2 that the two first model specifications shown on the left graphs are identifiable. But the model represented on the right side is not. The linear dependence on time of the covariate has the effect that neither β nor δ can be uniquely determined.

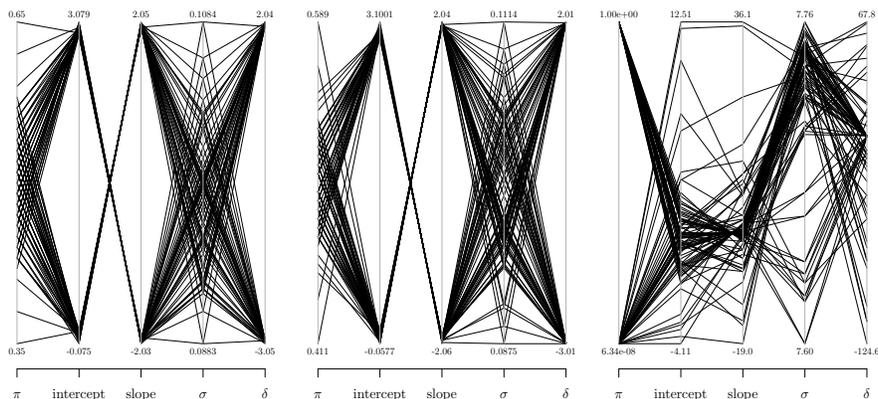


Fig. 2: Parallel coordinate plots of the estimated parameters for 50 simulated samples with different forms of the covariant.

References

1. W.S. Desarbo and W.L. Cron. A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5, 249–282, 1988.
2. C. Hennig. Identifiability of Models for Clusterwise Linear Regression. *Journal of Classification*, 17, 273–296, 2000.
3. D.S. Nagin. *Group-Based Modeling of Development*, Harvard University Press, Cambridge, 2005.
4. C. Noel and J. Schiltz. `trajeR` - an R package for finite mixture models, to appear, 2020.

5. J. Schiltz. A Generalization of Nagin's Finite Mixture Model. In: M. Stemmler, A. von Eye and W. Wiedermann. *Dependent Data in Social Sciences Research*. Springer, Heidelberg, 2015.
6. H. Teicher. Identifiability of Finite Mixtures. *Annals of Mathematical Statistics*, 34,4,1265–1269, 1963.
7. E.J. Wegman. Hyperdimensional Data Analysis Using Parallel Coordinates. *Journal of the American Statistical Association*, 85, 411, 664–675, 1990.
8. J.M. Wooldridge. *Econometric Analysis of Cross-Section and Panel Data*. 2nd edition, MIT Press, Cambridge, 2010.
9. S.J. Yakowitz and J.D. Spragins. (1968), On the identifiability of finite mixtures. *Annals of Mathematical Statistics*, 39, 209–214, 1968.

Network modeling in knowledge management systems: Superlatives and clusters in Mexican pig production. An approach

Juan Felipe Núñez-Espinoza¹, Francisco Ernesto Martínez-Castañeda², Frida Moysén Albarrán²

¹ Posgrado en Desarrollo Rural, Colegio de Postgraduados, Campus Montecillo, Montecillos, México.

(E-mail: nunezej@colpos.mx)

² Instituto de Ciencias Agropecuarias y Rurales, Universidad Autónoma del Estado de México, Instituto Literario 100, Centro, Toluca México.

(E-mail: femartinezc@uaemex.mx)

Abstract.

Pig production is one of the most important agricultural commodities due to its dynamism, volume, commerce and the complex interaction with other markets. It is possible to delineate the interactions they have among actors. But who are responsible for generate science and knowledge in the Mexican pig sector? How do they communicate this knowledge? How do they transfer it?

The objective of this work was to stablish relations and interrelations between scientific Mexican pig actors, they work, identify groups and their influence in the generation, position and communication of knowledge using network modelling.

Keywords: Community, subdisciplines, livestock, modeling.

1 Introduction

Mexican pig production shown, in general, three different main moments. An increase in volume from 1980 to 1984, a crisis from 1985 to 1991 and a sustain grow since 1992 to our days (Bobadilla *et al.* 2010). In accordance with these trends, pork meat imports have reached 42.9% of the national consumption. Even more, USDA have predicted that by 2021, the Mexican pork consumption could reach 27% and 60% of this increase consumption will be covered by imports (Desouzart, 2018). Although the Mexican pork meat export reached 124,137 tons, in 2017, the pork meat imports (mainly from EE.UU. and Canada) were about 990,000 tons, generating an important deficit of about 865,863 tons. (SIACON, 2018). Besides, the Mexican pork sector has face different challenges like dumping and unfair commercial sanctions from EE.UU. (an example is the sanction by pork fever which it was eradicated from México so long).

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



A collateral effect of these dynamics, is the prevalence of politics that support the biggest pork producers, but that inhibit the local production and the household production. So, it is worth mentioning that at these social levels the backyard pork production is not an alternative for economic progress, but just an emergency option (to pay a familiar party, to buy medicine, to pay debts, etc.), even to play an important social role (Santos-Barrios et al. 2019).

All the latest indicates an exogenous dynamic sector that cannot resolve his endogenous contradictions and this have influence the processes of social agglomeration at the pork sector. So is very common observe the conformation of corporation international lobbies at this sector in order to build competitiveness and social support strengths. An example of this is the incorporation of Mexican companies at international rankings which means get obtaining commercial certification to participate in international markets. For example, to participate in the prominent Chinese pork consumption market (which concentrates 71.51% of the world's pig population- FAO, 2017) specially today that China must get rid of an important part of its herd due to African swine fever (approximately 130 million of pigs which means 19% of it's inner pork production). So, the pork production demand will grow so fast at the southeast Asian and undoubtedly, this event will modify a lot of local social structures of pork production market in the world. It will be convenient review the status of the local social system responsible of areas like this at the agri food production, so the opportunities to modelized it's social structure and to intervene in it.

2 Social context at México

Most of these movements were because an economic and political management. So, the pig production and pork are a business and a commodity very dynamic. Many factors are involved in this production, and one of the most important has been pig health. In Mexico, pig production generated 67 billion of Mexican Pesos in 2018 (SIAP, 2019) (3,514,809,413.39 USD). It's an strategic sector with supply multipliers of 2.24 for supply and 1.16 for offer. That's mean that each million pesos invest in pig sector, a 1.24 million are generated in others sector related with supply and 160 thousand in the offer. (Sosa-Urrutia et al., 2017). It is an agri-food area linked to innumerable sectors of the Mexican economy.

In contrast, this is a complex sector, but so fragile in front of the ups and downs of the actives biological dispersion that is coming with the opening of the worldwide market and the Worldwide pig diseases impact negative in the industry. Diseases like PRRS (Porcine Reproductive and Respiratory Syndrome) cost to the industry around 75 thousand Euros in a "light" infection for a 1,000 sow heard to 698 thousands Euros (Nauthes 2017) and only in one year, 664 millions USD in USA (Holtkamp *et al.* 2013); or PED (Porcine Epidemic Diarrhoea), that can causes a loss of 1,688 piglets for each 1,000 sows breeding

heard (Goede and Morrison 2016). A lot of money. Many other diseases and other factors affect this sector also.

All of these elements permit thinking about the national and international social actors, of the pork production sectors, that are participating from multiple point of views (commerce, research, teaching, etc.), as social conglomerates that generates structural patterns that are possible analyzed and measure.

At México, one of the principal pork production mainstream is agglomerated around the Mexican Pig Veterinary Society (AMVEC: in Spanish) which is a group of Veterinarians which work and research in different areas of pig veterinary and work together with Private Producers and public sector and Universities. AMVEC is more than 50 years of being working helping the mexican pork industry. AMVEC was born in early 1968, although legally constituted in the early 1970s. The association arises at a historical moment in which the production of sorghum develops and food factories increase. Pig farming is encouraged and the government establishes zootechnical posts and the Network of Diagnostic Laboratories. Since its foundation one of the main objectives of the association is related to the academic update. Changes in production systems and the incorporation of Mexico into the globalized world pose new challenges for the veterinary profession (AMVEC 2020) and new demands.

It is important to identify how researchers are working and if this work is able to create solid research groups with high and effective responses to what the sector demand.

The objective of this work was to analyzed the kind of links between researcher and groups of research that active participate in research and studies related with the pork sector.

3 Methodology

The study of the underlying social structures in scientific collaboration it being extended to critical and strategic sectors that are substantially for societies, such as the agri-food sector, especially in front of the current conditions of climatic change that are modifying the local conditions of agri-food systems. In this context, one of the principal pillars to hold the support and resilience skills of each society, is the scientific sector, mainly for its formal abilities to stablish local and global solutions. In that direction the solidity of this scientific system will express the inner conditions of the social sector in which they work, so is going to express the inner strength of the social and technologic innovation system of the society.

To get an approach to this social skills system, it was necessary analyze the structure of influences that underlie in a particular construction of co-

authorships of a scientific community which works in the Mexican pork sector. To do that, it was involved a deductive, deterministic and by expediency scheme, using two particular tools.

1.- The archives of research published at the Proceedings of the 2018 congress AMVEC, which was integrated by 113 specialized papers from 311 authors from 77 research institutions (public and private), from 6 American Countries (Chile, Colombia, Spain, Mexico, USA and Venezuela). With this information it was possible to establish binary and unimodal matrices, nuanced with the personal attributes of each author (nationality, area of knowledge, institution) to identify the different levels of adjacency. To analyze this structural complexity with a structural point of view Social Networks Analysis (SNA) was so useful.

2.- To identify and measure the social prominence among actors (produced by their links themselves), SNA propose some categories from the matrix math that allow us to measure certain quantitative angles of the social dimensions as empathy among actors. Some of these categories are the Degree centrality, and Betweenness centrality.

The first one identified social prominence through skills of communication, empathy among actors, privileged access at the structural information streams and is calculated by the number of direct links that a particular node has (Wasserman and Faust, 1994; Paniagua, 2012). The math expression to calculated (Machín, 2012) is:

$$C_g(n_i) = \sum^A L(n_i, n_j) / (A-1)$$

where $C_g(n_i)$ is the number of nodes with which n_i is connected, and $(A - 1)$ is the width of the network.

The second one is about of the ability and power to endorse a relationship between unlinked pairs, and is determined by the number of times in which an actor appears as possible connection between these actors that are not directly linked (Wasserman and Faust, 1994). This ability is synthesized by the following equation (Machín, 2012):

$$C_I(n_i) = \sum g_{jk}(n_i) / g_{jk} \quad \forall j < k$$

where $C_I(n_i)$ is the degree of intermediation; $g_{jk}(n_i)$ is the number of geodesics between nodes j and k that pass-through node I ; and g_{jk} is the number of geodesics that join the nodes j and k .

Disciplines and subdisciplines used correspond to the Mexican Science and Technology Council (CONACyT: In Spanish) Taxonomy.

3 Results

General Structure

In swine research groups (Figure 1), as in others research areas, there is a differentiated and local access to information flows, so it's generated higher social densities to specific groups. As result of this is a fact that there are groups with much more cohesion than others, so these groups are going to have more influence in the local social clusters (Figure 1).

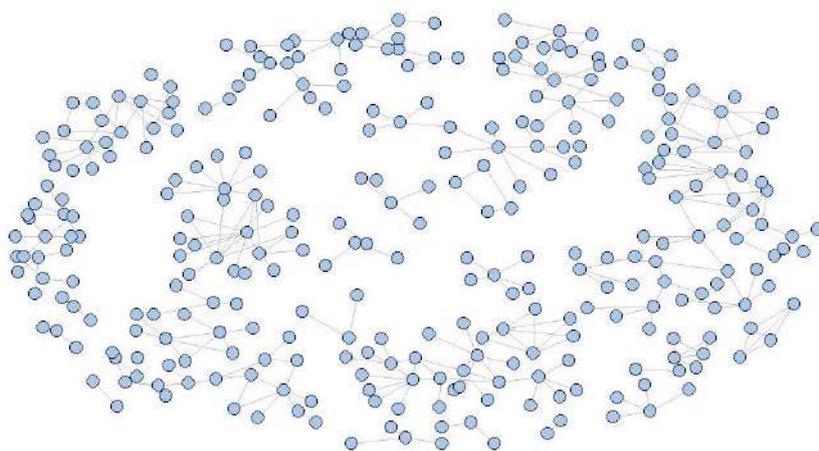


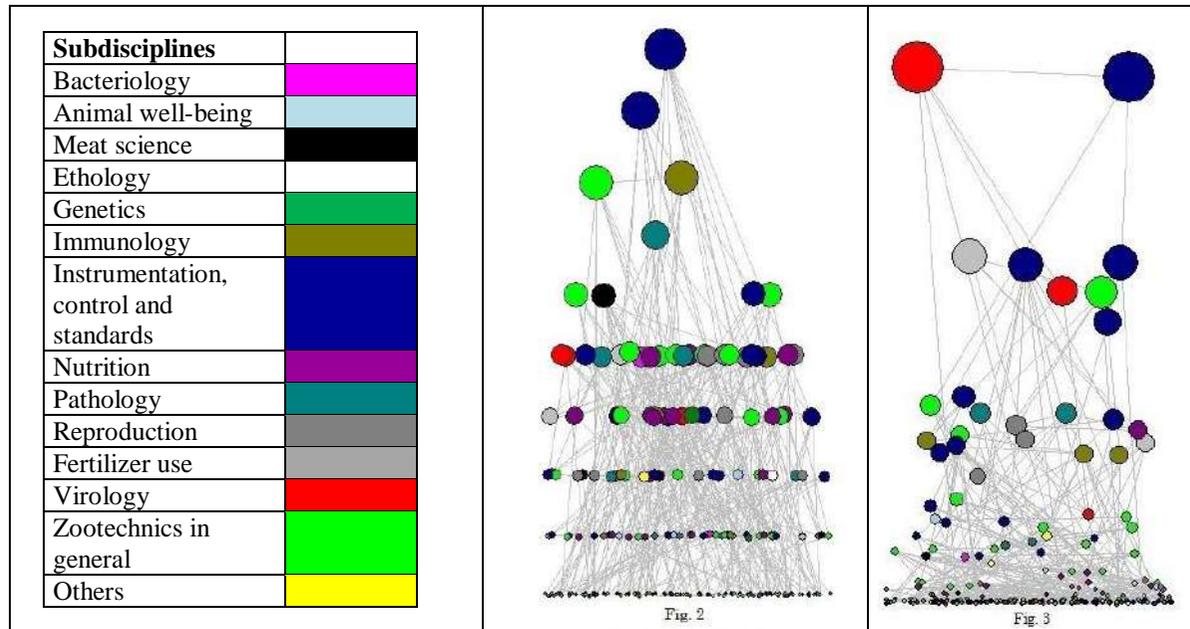
Fig. 1. Social structure in Pig research. AMVEC 2018. A proposal.

This allows us to assume a certain hierarchy (Degree), depending on the economic and intellectual resources that each group manages, therefore, a political differential is generated: some have more power than others to access the most complete information circulating in the network (Fig. 2). In this context, at the Top disciplines are: a) Instrumentation, control and standard. Probably due because main of the diagnostics technics or procedures used to do research in animal health and veterinary, have to be standardized and also because any results in this scientific area it is not valid without a standardized procedure; b) Zootechnics in general and Immunology are in second position, and is understandable if the data base analyzed correspond of Pig Veterinary issues and a portion of the researchers are linked with the pharmaceutical industry. The third structural agglomerate most important research sub-discipline correspond to Genetics.

This influence structure is further exacerbated by referring to groups that not only have privileged access to the information circulating on the network, but are also capable of regulating links (Betweenness), and even determining the predominant discourses within the research on pigs, that is, they can determine

the important issues for the pig industry and / or their particular interests, which means that they probably influence the epistemology of swine research. The top areas are Virology, Instrumentation, control and standards and Zootechnics in general. Virology now represents the most important sub-discipline probably because during 2013 and 2014, one of the most devastating diseases not only in Mexico but worldwide was caused by a Virus (it only affect pigs, not humans) “Porcine Epidemic Diarrhea” (Fig. 3). Likely to, these kind of events have caused a hierarchy adjustment among these sub-disciplines. Zootechnics in general appears as an important area in Pig research, but his explanation is more due because the Taxonomy of Disciplines and subdisciplines. Also, Zootechnics is more general subdiscipline. This may be significant because at the time when research programs and call for finance projects, is commonly directed to specifics subjects (subdisciplines like Virology, bacteriology) more than Zootechnics in general. Although we believe that more interactions and efforts should be driving to strength interdisciplinary research. It is understandable that research areas like molecular areas or including subdisciplines as virology or immunology are quite specific, studies in conjunction with mathematical modeling, or economics or environmental aspects as well as social global impact, should be done.

The process of generate any kind of knowledge (it does not matter if it is scientific or “traditional”) is a purely and solely community process. So, elements such as cooperation, solidarity, filiations, trust, concordance, mutuality, empathy and even friendship, they all confirm the subjective meaning of each scientific identity and allows the social confluence in the construction of ideas and objectives, therefore a probably promissory achievement. In this direction each academic network is a communal psychosocial system that changes continuously its profile of agglomeration as a community and structure but its inner management of these elements, permitted measure and analyze the internal patterns of association and filial complicity, so the stronghold of these communities. In this way, in the research swine network It was possible identify many (and different kind of) trust between groups of researchers and disciplines, but, dramatically, only was possible to find just one reciprocity behavior (link in red) (Figure 4).



Figures 2 and 3. Pig Research social structure by subdisciplines. Degree (Fig. 2) and Betweenness (Fig. 3).

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



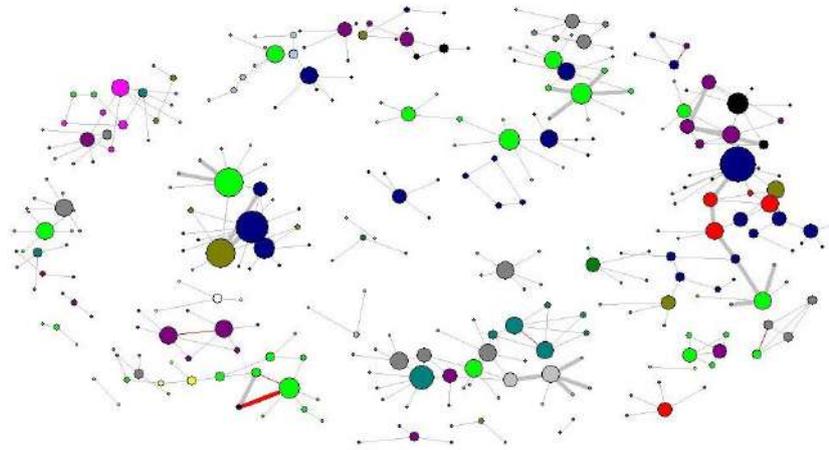


Fig. 4. Behavior of a cluster of research on pig issues based on Subdisciplines, trust, reciprocity and structural degree

4. Conclusions

At these times in which the humanity is facing enormous challenges, as climate change, economic breakdown, poverty, overpopulation, agri-food systems weaker and the zoonotic disease risks, among others, one of the most preventive strategies is the constant assessment of the local institutional systems, all of which have plenty of inner social challenges, particularly those that are in the center of vulnerable areas as the agri-food system and security sanitary. Such is the profile of the Mexican pork sector, especially for its economic, food and social importance.

In that direction is visible the prominence of some areas at the Mexican swine research, which undoubtedly is linking with some of its particular characteristics, one of which refers to an unequal distribution of the economic worries. For example, the social prominence of areas as Instrumentation, Control and Standard as well as Zootechnics like nodes that not only have the capacity to access at the structural communication but they could determine the content of such flow of information too (even taking up relevance over a strategic area as virology) could be related with the next indirect research lines that would be convenient follow in the future:

a) From a particular point of view, there are enough virologists at México or the specialists in Instrumentation, Control and Standard fulfill the epistemological



needs of the swine Mexican research? and knowing about the importance of counting on these resources (own's virologists) why an area as Instrumentation, Control and Standard has grew up in such way in a strategic sector?

b) Probably, this could be related with a general and commercial trend at the international pharmaceutical market of diagnostics techniques or procedures. According with EOB HSP (2016:13-14), just the market of the *in vitro* diagnostics represent profits of about US\$ 40–45 billion and the 80% of global sales is concentrated mainly in developed countries: United States (US\$ 19 billion), Europe (US\$ 14 billion) and Japan (approx. US\$ 4 billion). And this market is monopolized by six “key” players: Roche Diagnostics, Abbott Diagnostics, Siemens, Johnson & Johnson Medical Devices and Diagnostics, Beckman Coulter and BioMerieux. In addition, México is the second largest market at Latin América to the pharmaceutical industry (for consumption and manufacturing) and the Food and Drug Administration (FDA) is an important monitor about the standards of the medicines at the country (SE, 2013:13).

c) Undoubtedly, this type of phenomenon is linked to the increasing vulnerability of local capacities to face systemic emergencies (diseases, pandemic, famines, etc.) because the principal resources are destined to research areas that are not necessarily the responsible in the *research* of “new science” or scientific epistemology, but are responsible “just” of the implementation of standardized knowledge and this, indirectly, affect and subtract innovation and development capacity from others scientific-cognitive areas.

On a different topic, with the mathematical topology is possible get access at the social structure of trust among the people, so in this research it was possible determine the fortress inner of a brief part of this social universe, so it could be feasible establish a management politics of the intellectual capital at this sectors, so, it may be possible to stablish research groups in Mexican Pig Veterinary society leading by the subdiscipline of Instrumentation, Control and Standards and Virology, followed by Zootechnics in general (pig production), Immunology and Pathology. But mainly rethink social links between research teams in order to reinforce the innovation and development capacities in the research teams. This is because results showed a particular behavior: many groups presented trust links but in a separated group and just one showed reciprocity (probably a consolidated group).

The swine research is a complex social universe that expands or constrains according to its needs and resources, but socially is composed with a lot of social actors, so there are innumerable social incomes and conjunctions. Just in a simple search at the web, using the same sub-disciplines, the result for pig and environment was 131,209 research papers, 46,995 for the words pig and economic and 26,854 for the words pig and environment and economics. Another example is that using the subdisciplines as a key word in the Scencedirect web finder, it was possible to get around 8,513 results for pig instrumentation, control and standard, more than 227 thousand for “pig production” or 694 for Pig zootechnic (that is not a common way to refer to this discipline), 43,067 for pig immunology and 83,547 for pig pathology. But, if we type pig and virus the result was 70,447 research papers.

In relationship with the latest, this paper did not pretend dictate what should be or not where the pig veterinary research should look at. Veterinary is a large and multiple discipline profession. Instead, it pretended to picture, from a partial point of view, about what is happened in the scientific community that are working in the swine research (taking the AMVEC social circle as a sample) and what could be happened worldwide, in research matters.

We believe that is necessarily make much more research about this kind of themes, and is necessary to do much more interactions among the scientific community and efforts should be driving to strength interdisciplinary research.

In that direction it may see that some research areas may not be consider in this AMVEC scientific forum, and is environment and economy. It is quite normal to understand this behavior and is because, AMVEC, since its foundation, has been a scientific forum for pig health.

We believe too that it is necessary to consolidate what organizations like AMVEC are being doing so far. So, it is necessary to do a retrospective analysis in order to have more data. We also believe, at least in Mexico with the new Scientific and Technologic law, that more interdisciplinary research has to be done with special emphasis in human and social needs.

Finally, these kinds of papers permit different questions:

Why the veterinary community are not working in environmental issues and economic issues or why the specialist in this kind of areas are not presenting their results researches in AMVEC scientific forum?

What if we design a politic where the support and money for investigation goes no just to Instrumentation but to other strategic areas?

Undoubtedly in front of the unprecedented phenomena that are happening in the world, it is necessary incorporate big data and structural analysis at the swine research as strategic area to the agri food system and health sectors.

5 Bibliography

- AMVEC. 2018. Memory of the LII National Congress. Published: <https://www.amvec.com/blog/amvec-1/post/memorias-amvec-14#>
- AMVEC. 2020. Historia AMVEC. <https://www.amvec.com/historia>
- Bobadilla-Soto Ernesto, Espinoza Ortega Angélica, Martínez-Castañeda Francisco. 2010. Dinámica de la producción porcina en México: 1908-2008. *Rev Mex Cienc Pecu.* 1(3):251-268.
- Desouzar, Osler. 2018. Oportunidades para el mercado de la carne de cerdo. <https://slideplayer.es/slide/1092071/>, Consulted March 2020.
- Ensuring innovation in diagnostics for bacterial infection. Implications for policy. World Health Organization (WHO). Published:

- http://www.euro.who.int/__data/assets/pdf_file/0008/302489/Ensuring-innovation-diagnostics-bacterial-infection-en.pdf?ua=1 (10/04/2020)
- EOB HSP (The European Observatory on Health Systems and Policies) (2016).
- FAO (2017). FAOESTAT. Ganadería. Cerdos. Web. Disponible en: <http://www.fao.org/faostat/es/#data/QA>
- Holtkamp DJ, Kliebenstein JB, Neumann EJ, Zimmerman JJ, Rotto HF, Yoder TK, Wang C, Yeske PE, Mowrer CL, Haley CA. Assessment of the economic impact of porcine reproductive and respiratory syndrome virus on United States Pork producers. *J S Ha Prod* 2013;21(2):72-84.
- Machín, J. Redes Sociales e Incidencia en Política Pública. Estudio Comparativo México—Colombia; Sedesol-Indesol: Ciudad de México, México, 2011.
- Nauthes H, Alarcon P, Rushton J, Jolie R, Fiebig K, Jimenez M, Geurts V, Nathues. Cost of porcine reproductive and respiratory syndrome virus at individual farm level – An economic disease model. *Prev Vet Med* 2017;142:16-29.
- Paniagua, J.A. Curso de Análisis de Redes Sociales. Metodología y Estudios de Caso; Universidad de Granada: Granada, España, 2012.
- Santos-Barrios L, Ruiz-Torres M, Gómez-Demetrio W, Sánchez-Vera E, Lorga da Silva A, Martínez-Castañeda FE. 2019. An Approximation of Social Well-Being Evaluation Using Structural Equation Modeling. In: Skiadas C, Bozeman J, Editor(s): Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining, Volume 2. Pp 117-124.
- SE (Secretaría de Economía) (2013). Industria Farmacéutica. Unidad de Inteligencia de Negocios. Disponible en https://www.gob.mx/cms/uploads/attachment/file/62881/130820_DS_Farmacutica_ESP.pdf
- SIACON. 2015 SIAP-SAGARPA Servicio de Información Agroalimentaria y Pesquera <http://www.siap.gob.mx/optestadisticasiacon2016parcialsiaconzip/>, Consultado 16 Feb, 2018.
- SIAP. 2019. <http://www.siap.mx> Consulted November 2019.
- Sosa-Urrutia Manuel, Martínez-Castañeda Francisco, Espinosa-García José, Buendía-Rodríguez Germán. 2017. Contribución del sector pecuario a la economía mexicana. Un análisis desde la Matriz Insumo Producto. 8(1):31-41.
- Wasserman, S.; Faust, K. *Social Network Analysis, Methods and Applications*; Cambridge University Press: Cambridge, UK, 1994.

Completeness assessment of neonatal deaths in a region of Brazil: linkage and imputing missing data

Neir Antunes Paes¹, Carlos Sérgio Araújo dos Santos², Tiê Dias de Farias Coutinho³

¹ Health and Decision Modelling Postgraduate Course. Federal University of Paraíba, Campus I - Lot. Cidade Universitária, João Pessoa - PB, 58033-455, Brazil (E-mail: antunes@de.ufpb.br)

² Health and Decision Modelling Postgraduate Course. Federal University of Paraíba, (E-mail: carlossergioaraujo@hotmail.com)

³ Health and Decision Modelling Postgraduate Course. Federal University of Paraíba, (E-mail: tiefarias@gmail.com)

Abstract. Although coverage of neonatal deaths in Brazil is considered high, the completeness of death declaration items for several regions is a problem of concern and uncertainty, which can compromise the maternal and child health planning. Data set linkage offers considerable potential to address an extensive range of research questions, such as identifying risk factors, and quantifying mortality, morbidity and healthcare for infant health as in the neonatal period. This technique added to imputing missing data techniques is a feasible and cost-efficient way to recover data. The main aim of this paper is to evaluate the completeness of information on neonatal death declarations in the regions of Paraíba State from 2009 to 2017. The quality of data on neonatal deaths declaration was studied in two stages: in the first, the Mortality Information System and Birth Information System from the Brazilian Ministry of Health databases were matched using the deterministic linkage; in the second, the multiple imputation for missing data was carry out. In total, 5,149 neonatal deaths were computed. The following variables were investigated: gender, race/color, mother's age, weeks of gestation, birth weight, mother's educational level, number of live children, number of dead children, type of pregnancy and type of delivery. There was an important decrease in neonatal death records over time, approximately 19%. Except for the variable mother's educational level (20.0%) and gender (0,8%), all variables presented percentages of incompleteness ranging from 6.7% to 15.3%. The percentage of matched records ranged from 50.0% to 58.8% in the period. After five multiple imputations, the missing data were recovered. The Relative Efficiency of the variables with missing observations recovered was verified, whose efficiency for all variables ranged from 96.7% to 99.9%. The conclusion was an excellent and reliable imputation of missing data. The tools used here proved to be very efficient and useful for use in regions with deficient data, such as those deaths registered for Paraíba in Brazil.

Keywords: Vital Statistics, Neonatal, Mortality, Brazil.

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



1 Introduction

Death certificates contain extremely relevant information, since they serve as subsidies for the calculation of parameters for the monitoring and surveillance of deaths, in addition to favoring the planning of health actions aimed at reducing mortality and morbidity rates.

Health Information Systems (SIS) consist of a set of steps that process, store and distribute information to support the decision-making process and assist in the control of health organizations. In Brazil, there are several SIS linked to the Ministry of Health. Among them, there is the Mortality Information System (SIM), created in 1975, with the purpose of providing information on mortality in different areas, in which any and all death in Brazil must be registered through the Death Certificate (DC).

Access to reliable SIM data allows verifying, with adequate quality, the conditions of births, deaths and their determinants favoring the analysis of the health situation of a given location. However, the use of consolidated information in these systems usually presents limitations. The most compromising ones refer to the coverage of deaths, failures in filling in the variables of the DC and disagreement with common information with the Live Births Information System (SINASC) [1],[2].

While coverage refers to the number of deaths that occurred and effectively registered in a region, the incompleteness reveals the lack of care and importance given to the filling in of the variables of the collection instruments by health professionals, absence of data in medical records and even the ignorance of certain information by the respondents when the form is filled in[3].

Several authors have been using the relationship of databases (*linkage*) as a strategy to improve the quality of information, as long as they have common means of identification. This procedure allows the recovery of incomplete or inconsistent records, improving the completeness and reliability of the information provided by SIM and SINASC[4],[5],[6],[7]. However, the technique of relating the database does not meet the cases that were not paired in the two

databases, and even those that were paired, there is a risk of not being successful in capturing the missing information. A complementary technique that allows improving the completeness of information through statistical inference methods is data imputation.

Imputation consists of filling in the missing data with plausible values for further analysis of the complete data. Initially, the techniques developed were relatively simple, called single imputation: imputation by average, imputation by substitution by the nearest neighbor and by linear regression. However, from multiple imputation, where for each missing data more than one data is imputed, thus associating the variability of the data with the results, the methods became more precise and sophisticated[6].

Infant mortality is seen as one of the main targets of public health policies worldwide, whose main responsibility is neonatal mortality. In Brazil, the percentage of neonatal deaths is responsible for about 70% of infant deaths[8].

Since post-neonatal mortality is more easily controlled, and after reaching more controlled levels, the concern of public managers is directed towards neonatal mortality especially in less developed regions. The state of Paraíba is in this situation in the Northeast region of Brazil, with 223 municipalities. This State had a population of 3,9 million inhabitants in 2018. Nearly three quarters of the population live in urban areas clustered along the Atlantic coast and had an HDI of 0.722 in 2017, classified at the threshold of the high human development (0.700 – 0.799). This region is considered one of the least developed in Brazil, occupying the 20th place among the 27 states in the Country. The public health sector – the SUS – provides universal access to health care, which is the sole provider of population's health care coverage.

The neonatal mortality rate in Paraíba, which was 10,84 deaths per thousand live births (p/1000 l.b.) in 2009 decreased to 9,56 deaths p/1000 l.b. in 2017. The level of neonatal mortality in Paraíba is far from 1,1 p/thousand l.b. in Japan (the lowest in the world), but way below Sierra Leone, with 49.5 p/thousand l.b. (the highest in the world). For Brazil, this level reached 8,7 p/thousand l.b. in 2017[9].

Although it is recognized that the coverage of neonatal deaths is not complete, the focus of this work is dedicated to filling in the variables of the registered Death Certificates, whose percentages of non-completion may lead to inaccurate and incomplete assessments in the management of maternal and child health services. Thus, it is necessary to retrieve this data.

Since it has a low operating cost, the use of a database relationship (*linkage*) and multiple imputation will make it possible to use the complete information of the variables integrated in the death certificates. Thus, the main objective was to assess the integrity of information on neonatal death certificates in the regions of Paraíba from 2009 to 2017.

2 Materials and Methods

This is a cross-sectional ecological study of neonatal infant deaths in the state of Paraíba, which uses information from variables present in the Death Certificate (DC) available in SIM and in the Declaration of Live Birth (DLB) extracted from SINASC in the years of 2009 to 2017.

Data quality analysis was carried out observing the completeness of the information from the SIM and SINASC systems. Initially, the deterministic linkage between these two systems was performed to recover the missing data in the common variables of the two databases. After this process, the imputation of values that were not possible to recover was performed, using the multiple imputation technique. The scheme of the steps that were followed is shown in Figure 1.

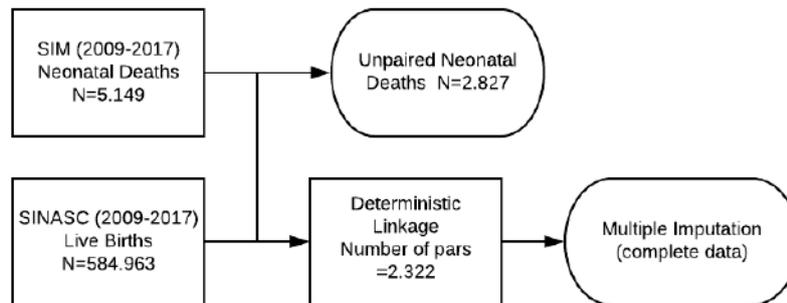


Figure 1: Structuring Steps, Linkage between SIM and SINASC and Imputation.

Deterministic Linkage

The deterministic linkage method assumes that there is information in different standardized databases that allow the connection between these databases. To make the connection between the information systems that contain information on infant deaths and births, the unifying variable of common identification, the DLB number (Declaration of Live Birth), was identified and the bases of SIM and SINASC were compared. For this purpose, the Microsoft Office Excel 2007 search and reference function (VLOOKUP) was used.

After the pairing between the DC and the DLB it was possible to manually retrieve some information from the SINASC records, both of the variables that are not included in the SIM, as well as common information to the two databases that had blank or ignored values.

Multiple Imputation

According to the analysis of the proportion of missing data, it was possible to determine the imputation method to be used [10]: i) Proportion $\leq 0,05$ \rightarrow It can assume that the data are complete or apply the single imputation; ii) Ratio between 0,05 and 0,15 \rightarrow Indication for using multiple imputation, although single imputation can also be used without losses; iii) Proportion $\geq 0,15$ \rightarrow

Indication for the use of multiple imputation. Multiple imputation was chosen, because this technique produces unbiased results and with appropriate errors.

The multiple Imputation technique creates multiple copies of the database where the missing values are replaced by different imputed values. That is, m number of different and complete databases was generated, and each one was analyzed according to the intended objective[11]. With the possession of the complete m databases, it was possible to estimate the variance of the imputed data and between the different bases.

The first step to define the application of multiple imputation was the evaluation of the non-response mechanisms and the non-response pattern. They can only be defined using data. The three non-response mechanisms are: missing completely at random (MCAR), missing at random (MAR) or not missing at random (NMAR). The non-response pattern can be classified into: monotonic pattern or non-monotonic pattern.

According to the graphical analysis of the patterns of the missing data, the pattern chosen was the general pattern, as there was a dispersion of missing units throughout the data matrix (Figure 2). The mechanism was missing completely at random (MCAR) because the value of the missing variables was not related to the variable itself or any other variable in the data set.

After the decision to choose the mechanism and the pattern, the m imputations were generated and then the Rubim Rules were applied to the database, since they can be used for any type of multiple imputation. The amount of m imputations was based on an indicator called Rubin for Relative Efficiency (RE). This indicator is a function of the amount of m imputations and the percentage of λ missing data and is given by:

$$RE = \left(1 + \frac{\lambda}{m}\right)^{-1}.$$

For the combination of individual m estimates, a general average can be obtained by: $\bar{Q} = \frac{\sum_{j=1}^m \widehat{Q}_j}{m}$, where Q_j is each of the m estimates of the parameter of interest, being $j= 1, 2, \dots, m$. The average variance within the imputations is

calculated by: $\bar{U} = \frac{\sum_{j=1}^m \hat{U}_j}{m}$, where U_j is the variance of the estimators Q_j . The variance between imputations is calculated by: $B = \frac{\sum_{j=1}^m (\hat{Q}_j - \bar{Q})^2}{m-1}$. The variance between imputations can be obtained by: $T = \bar{U} + \left(1 - \frac{1}{m}\right)B$.

The Multiple Imputation procedure was performed using the free access R version 3.6.2 statistical software, available at: <https://www.r-project.org>.

3 Results and Discussion

Between 2009 and 2017, according to SINASC, there were 522.852 births and according to SIM, 5.149 neonatal deaths in the State of Paraíba – Brazil. Based on the completion of the deterministic *linkage* between the databases using as unifying variable common to the SIM and SINASC databases, the pairing percentages are presented in Table 1. In general, from 2009 to 2015 there was growth in percentages each year, starting from 16% in 2009 and reaching 58.8% in 2015. This increase indicates an improvement in filling in the variable “number of the declaration of live birth”. In the years 2016 and 2017 there was a small decline in the percentages with values of 56.8% and 53.4%, respectively. It is noted that as of 2012, the percentage of paired records was above 50%. When observing only the neonatal death records that had the number of the declaration of live birth, the growth pattern from 2009 to 2015 is similar to the pattern that involves all records, except for 2011 and 2014 when there was a decrease.

The variable “number of the declaration of live birth” that appears in the infant death declarations is fundamental for the success of the pairing. Failure to complete this information in infant death certificates compromises the capture of information about other maternal and child variables present in death certificates. There was a deficiency in the collection of this information, referring to the need for substantial improvements in filling in this data.

Table 1: Percentage of neonatal death records paired between all records and the percentage of death records that had the DLB number in Paraíba, 2009 to 2017

Year	% (general)	% (records with n° DLB)
2009	16.0	65.4
2010	29.2	91.7
2011	44.4	90.4
2012	50.7	97.5
2013	52.3	98.7
2014	54.0	95.9
2015	58.8	99.3
2016	56.8	97.5
2017	53.4	97.0
Total	45.1	94.2

Source: Mortality Information System - Ministry of Health

The growth trend in pairing observed since 2010 is a reflection of the obligation to fill in the number of the declaration of live birth in the declaration of deaths under one year from that date[3], which explains the low percentage of pairing in 2009. When comparing the pairing percentages of the total from 2009 to 2017, it is observed that among all the neonatal death records 45.1% of them were paired, whereas considering only the records that had the completed DLB number, 94, 2% of them were paired. In summary, those neonatal death records that had the number of the Declaration of Live Birth were approximately 100% paired.

A fundamental aspect was to verify the percentage of incompleteness (fields not filled in or ignored) of the variables that were used in the neonatal death certificate. Table 2 shows the percentage of information ignored or not filled in the variables of the neonatal Death Certificate, for the period from 2009 to 2017 before and after the deterministic *linkage*.

Before applying the deterministic *linkage*, the variable “mother’s educational level” presented the highest percentage of incompleteness throughout the period and the variable “sex” presented the lowest percentage.

Table 2: Number and percentage of information ignored or not filled in according to some variables of the neonatal death Declaration before and after the linkage, and after the imputation in the State of Paraíba from 2009 to 2017

Variable	Before		After Linkage		After Imputat.
	N	%	N	%	%
Gender	40	0.78	34	0.66	0.00
0.00Race/Color	454	8.82	425	8.25	0.00
Mother’s Age	777	15.09	690	13.40	0.00
Mother’s Educational Level	1031	20.02	871	16.92	0.00
Number of live children	660	12.82	560	10.88	0.00
Number of dead children	787	15.28	633	12.29	0.00
Type of Pregnancy	344	6.68	339	6.58	0.00
Weeks of Gestation	713	13.85	572	11.11	0.00
Type of Delivery	375	7.28	360	6.99	0.00
Birth weight	480	9.32	465	9.03	0.00
Total of records	5149	-	4949	-	-

Source: Mortality Information System - Ministry of Health

The failure to fill in variables such as “mother’s educational level” and “mother’s age”, makes it difficult to analyze social inequalities in various health outcomes for women and children, with emphasis on neonatal mortality[3],[12]. It is observed, as well, that before the *linkage*, with the exception of the variable “mother’s educational level” and the “gender” of the child, all variables presented percentages of incompleteness ranging from 6.68% to 15.28%.

After applying the *linkage*, the gain of information retrieved in the Declarations of Live Birth allowed to reduce the percentage of incompleteness in all variables (Table 2). The standards remained virtually unchanged, however, the percentage levels were reduced. In the period from 2009 to 2017, the following reductions in the variables can be emphasized: “mother’s educational level” (from

20.02% to 16.92%), “mother’s age” (from 15.09% to 13.40%), “Number of live children” (from 12.82% to 10, 88%), “number of dead children” (from 15.28% to 12.29%) and “weeks of gestation” (from 13.85% % to 11.11%).

After the deterministic *linkage*, the next step was multiple imputation. Table 2 shows that after imputation all variables of neonatal death were completely filled. To apply this technique it is necessary to decide on the number of imputations required by calculating the Relative Efficiency (RE). Table 3 shows the Relative Efficiency values of the variables with missing observations that were used in the statistical modeling. By opting for five imputations, the Relative Efficiency presented a minimum percentage of 96.7% for the variable “mother’s educational level” and a maximum value of 99.9% for the variable “gender”. The high percentages of Relative Efficiency justify the option for the five imputations. Since the missing information rates for the variables used were not high, the use of a number greater than five imputations would not have a practical benefit in the estimates of the values to be imputed. It became usual to adopt five imputations, due to the experiences of researchers, who have found that a small number of imputations is sufficient for the conclusions to be statistically efficient[11],[13],[14].

To apply multiple imputation it was necessary to analyze missing data patterns present in the data set (Figure 2). This graph is useful to help find clues about the pattern of absences[15]. The y axis of the graph represents each pattern identified and on the x axis it represents each variable with missing data. Altogether, 134 distinct patterns of missing data were identified[16].

According to the classification proposed by Little and Rubin[17], it is evident that the pattern of missing data is of the “general” type, since missing data were observed in several variables in the SIM database. According to Enders[16], this pattern is considered the most frequent in practice. As for the mechanism that generates the missing data in the studied database, it is suggested that the missing data were generated by the Missing Completely At Random (MCAR), since the value of the missing variables is not related to the variable itself or any other variable in the data set. The great advantage of the mechanism being MCAR is

that the cause that led to the missing data does not need to be part of the analysis to control their influence on the research results[14]. Understanding these missing data mechanisms is critical, as the properties of the missing data methods depend very much on the nature of the dependencies of these mechanisms[18]. In addition, it is necessary to assume that the missing data were generated by one of the mechanisms for defining the imputation method used[14],[15].

Table 3: Relative Efficiency for the imputed variables according to the choice of five imputations

Variable	Relative Efficiency (RE)
Gender	99.9
Race/Color	98.4
Mother's Age	97.4
Mother's Educational Level	96.7
Number of live children	97.9
Number of dead children	97.6
Type of the Pregnancy	98.7
Weeks of Gestation	97.8
Type of Pregnancy	98.6
Birth weight	98.2

Source: Mortality Information System - Ministry of Health

In Figure 3, it can be seen that, of the total neonatal deaths, most (approximately 80%) of the records fall into pattern 1 (no variable with missing data), followed by pattern 77 (only the variable “mother’s educational level”) with missing data) whose variable had the highest percentage of missing data, right afterwards there is pattern 6 (only the variable “birth weight” with missing data) and pattern 132 with all studied variables having missing data, except for variables “gender” and “race/color” of the child.

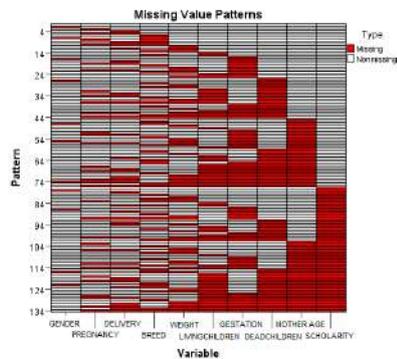


Figure 2: Non-response patterns in SIM neonatal death records in Paraíba, 2009 to 2017.

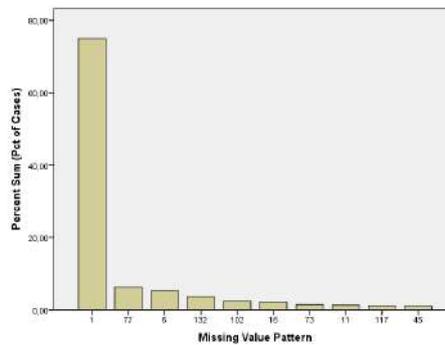


Figure 3: Percentage of cases in each pattern in SIM neonatal death records in Paraíba, 2009 to 2017.

With the combination of deterministic *linkage* and multiple imputation techniques, it was possible to determine the completeness of the information in the database on neonatal mortality. In order to have a good assessment, the deterministic *linkage* needs that the identifier variable is at least well filled in both databases used in the pairing. Because, even if the pair is formed, there is no guarantee that the incomplete information of the declarations will be captured due to its absence in both databases. In summary, given the existing limitations, the use of deterministic *linkage* led to the reduction of missing data. Its use contributed to the use of multiple imputation whose technique is supported by statistical methods collaborating to obtain more credible results estimated by the multiple imputation technique.

Conclusions

Experience in the use of deterministic *linkage* and multiple imputation in the State of Paraíba provides for the possibility of replicating it in different regions of the Country with a low operational cost and the use of relatively simple techniques by health information systems. The need for imputing missing information is greater in countries with a deficiency in the quality of death data,

generally in developing regions. The results of this application signaled the great potential for improving data quality through the combination of the two techniques. In addition to allowing for improvement in epidemiological and demographic indicators with the completeness of the data. In this way, health managers can perform more reliable analysis directing actions in the spheres of municipal, state and federal planning. The combined advantages of these two techniques that have been applied here are an excellent alternative that can be routinely implemented in health information systems in regions with data quality problems. And, even in those countries with consolidated statistics, these techniques are powerful checking tools for the validation of their results.

References

1. N. Paes. Demografia estatística dos eventos vitais: com exemplos baseados na experiência brasileira. João Pessoa: Editora do CCTA, 215p, 2018.
2. J. Dombrowski, R. Ataíde, P. Marchesini, R. Souza and C. Marinho. Effectiveness of the Live Births Information System in the Far-Western Brazilian Amazon. *Ciência & Saúde Coletiva*, 20(4), 1245-1254, 2015.
3. L. Maia, W. Souza, A. Mendes and A. Silva. Uso do linkage para a melhoria da completude do SIM e do SINASC nas capitais brasileiras. *Revista de Saúde Pública*, v. 51, p. 112, 2017.
4. L. Maia, W. Souza and A. Mendes. A contribuição do linkage entre o SIM e SINASC para a melhoria das informações da mortalidade infantil em cinco cidades brasileiras. *Rev Bras Saude Mater Infant*.15(1):57-66; 2015.
5. L. Marques, C. Oliveira and C. Bonfim. Avaliação da completude e da concordância das variáveis dos Sistemas de Informações sobre Nascidos Vivos e sobre Mortalidade no Recife-PE, 2010-2012. *Epidemiologia e Serviços de Saúde* [online]. v. 25, n. 4, pp. 849-854; 2016.
6. M. Oliveira, M. Latorre, L. Tanaka and M. Curado . Simulação e comparação de técnicas de correção de dados incompletos de idade para o cálculo de taxas de incidência. *Cadernos de Saúde Pública* [online]. v. 34, n. 6; 2018
7. A. Mendes, M. Lima, D. Sá, L. Oliveira and L. Maia. Uso da metodologia de relacionamento de bases de dados para qualificação da informação sobre mortalidade infantil nos municípios de Pernambuco. *Rev Bras Saúde Matern Infant*.12(3):243-9; 2012.
8. M. Gaíva, E. Fujimori and A. Sato. Fatores de risco maternos e infantis associados à mortalidade neonatal. *Texto Contexto Enferm*. 25(4): 2-9; 2016.
9. United Nations Inter-Agency Group for Child Mortality Estimation (UN IGME), 'Levels & Trends in Child Mortality: Report 2017, Estimates Developed by the UN Inter-Agency-Group for Child Mortality Estimation', United Nations Children's Fund, New York, 2017.
10. F. Harrell Jr. Regression modeling strategies: with applications to linear models, logistic regression and survival analysis. New York: Springer-Verlag, 2001.

11. D. Rubin. *Multiple Imputation for Nonresponse in Surveys*, Wiley Series in Probability and Statistics. New York: Wiley, 1987.
12. R. Lino, S. Fonseca, P. Kale, P. Flores, R. Pinheiro and C. Coeli. Tendência da incompletude das estatísticas vitais no período neonatal, estado do Rio de Janeiro, 1999-2014. *Epidemiologia e Serviços de Saúde* [online]. v. 28, n. 2, 2019
13. S. Buuren. *Flexible Imputation of Missing Data Interdisciplinary Statistics Series*. Boca Raton, FL: Chapman & Hall/CRC; 2012.
14. J. Graham, A. Olchowski and T. Gilreath. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, Berlin, v. 8, p. 206-213, 2007.
15. J. Schafer and J. Graham. Missing data: our view of the state of the art. *Psychological Methods*, Washington, v. 7, n. 2, p. 147-177, 2002.
16. C. Enders. *Applied missing data analysis*. New York, NY: Guilford press, 2010.
17. R. Little and D. Rubin. *Statistical analysis with missing data*. 2nd ed. Hoboken, NJ: John Wiley & Sons, 2002.
18. W. Silva Júnior. *Diferenciais regionais na mortalidade adulta por escolaridade no Brasil em 2010*. 2018. 110f. Dissertação (Mestrado em Demografia) - Centro de Ciências Exatas e da Terra, Universidade Federal do Rio Grande do Norte, Natal, 2018.

Gender, health and socio-demographic influences on updating subjective survival probabilities

Apostolos Papachristos¹ and Georgia Verropoulou²

¹ Department of Statistics and Insurance Science, University of Piraeus, Greece

(E-mail: apapachristos@webmail.unipi.gr)

² Department of Statistics and Insurance Science, University of Piraeus, Greece

(E-mail: gverrop@unipi.gr)

Abstract

The goals of the study are to investigate whether individuals update their survival expectations consistently with changes in their mortality risk profiles as well as to assess whether the degree of consistency varies by gender. We use a longitudinal dataset from the 6th and 7th Waves of the Survey of Health, Ageing and Retirement in Europe. For the statistical analysis we employ generalised linear models. Our results indicate that males and females revise their survival expectations if they experience widowhood, a deterioration or improvement in self-rated health or functional limitations, an increase in the number of chronic conditions or income or life satisfaction, and BMI transitions from normal to underweight consistently. Such revisions are in line with actual mortality patterns. Males revise their survival expectations consistently with actual mortality patterns, following a drop in income or a BMI transition from underweight to normal. On the other hand, females do not revise their SSPs consistently with actual mortality patterns for BMI transition from overweight to obese and from underweight to normal. Both genders do not revise their SSPs consistently with actual mortality patterns for BMI transition from underweight to overweight or obese. Finally, divorced females revise their survival expectations consistently with actual mortality patterns whereas divorced males do not.

Keywords: Subjective survival probabilities, SHARE, self-rated health, BMI, marital status, ADLs

This work has been partly supported by the University of Piraeus Research Center.



1. Introduction

Individuals consider their own experiences, health status, social influences as well as information through media and health campaigns when forming survival expectations (Griffin et al. 2013). More specifically individuals who are less educated, have lower income and face financial strain tend to report lower subjective survival expectations (Arpino et al. 2018; Rappange et al. 2016; Mirowsky 1999). Moreover poor physical and functional health is associated with lower survival expectations (Van Solinge and Henkens 2018; Rappange et al. 2016; Hurd and McGarry 1995). The impact of marital status on subjective survival expectations is not fully understood. Balia (2011) using data from SHARE found that widows report higher subjective survival probabilities while Liu et al. (2007) and Rarrange et al. (2016) argued that living alone is negatively associated with survival expectations. Behavioral risk factors such as obesity are also associated with lower survival expectations (Rarrange et al. 2016; Liu et al. 2007). While the associations among socio-demographic factors with subjective survival expectations are already known, it would be interesting to understand whether individuals incorporate the marginal increase or decrease in mortality risk in their survival expectations, after the occurrence of specific events which change their mortality risk profiles.

Actual mortality patterns

According to the literature there is a range of events which could change individuals' mortality risk profile. Higher socio-economic status, better self-rated and functional health and higher life satisfaction are associated with lower actual mortality (Backlund et al. 1999, Kaplan et al. 1996, St. John et al. 2015, Idler and Benyamini 1997). Moreover, an increase in the number of comorbidities is linked to higher actual mortality (Scott et al. 1997, Lee et al. 2008). Factors depending on individuals' lifestyles such as obesity, overweight or underweight are associated with higher mortality (Takala et al. 1994, Berrington et al. 2010). The transitions out of marriage through widowhood or divorce are also linked to higher actual mortality compared to being married for both males and females (Kaprio et al. 1987, Dupre et al. 2009).

Objectives of the study

The objective of this study is twofold; first to evaluate whether individuals update their survival expectations consistently with the actual mortality patterns, and second, to assess whether the degree of consistency varies by gender. More specifically, we would expect upward revisions of survival expectations after the occurrence of events such as improvement in self-rated and functional health, increase in income and increase in life satisfaction. On the other hand, we would expect downward revisions of survival expectations after the occurrence of events such as increase in the number of chronic conditions, transitions out of marriage through widowhood or divorce, transitions of BMI from normal to underweight, overweight or obese categories.

2. Methods

Data

We used data from the 6th and 7th Wave of the Survey of Health, Ageing and Retirement in Europe (SHARE). SHARE (Börsch-Supan et al. 2013) is a cross-national and multidisciplinary panel database with information on health, socio-economic status, and social and family networks. The data collection of the 6th Wave was completed in November 2015 (Börsch-Supan 2017) and the sampling was carried out in 18 countries. The SHARE Wave 7 dataset became available on April 2019 (Börsch-Supan 2019) and the data collection took place in 28 countries in 2017. More documentation and information on SHARE can be found at <http://www.share-project.org>.

The wave 6 sample covers 68231 individuals; the wave 7 sample covers 76520 individuals and the combined longitudinal sample covers 51849 individuals. It is notable that 16832 individuals who participated in wave 6 do not participate in wave 7. A supplementary analysis of the characteristics of these respondents as well as the impact on our results is carried out in this study. The longitudinal sample consists of 51245 individuals aged 50 or older. Due to SHARE rules, information about subjective survival probabilities was not collected for 1764 individuals aged 50 or older (3.4%), for whom proxy interviews were conducted. Hence, the longitudinal sample used in this study reduces down to 49505 individuals.

Variables

Subjective survival probabilities (SSPs)

In the ‘Expectations’ module of the SHARE questionnaire, respondents were asked to state their survival expectations on a scale from 0 to 100 as follows:

What are the chances that you will live to be age [T] or more?

The target age T depends on the chronological age (‘x’) of the respondent at the interview; it is set at age 75 for respondents aged 50–65, at age 80 for respondents aged 66–70, at age 85 for respondents aged 71–75, at age 90 for respondents aged 76–80, at age 95 for respondents aged 81–85, at age 100 for respondents aged 86–95, at age 105 for respondents aged 96–100 and at age 110 for respondents aged 101 or higher. The difference between the respondents’ actual age and his/her target age is the prediction interval ‘N’, in years. The reported survival expectations were divided by 100 in order to calculate the SSPs.

Dependent variable

The dependent variable is calculated as the difference in the subjective survival probabilities, reported by the same individual, between wave 6 and 7.

$$\Delta SSP = SSP_{x,N}^{Wave\ 7} - SSP_{x,N}^{Wave\ 6}$$

Explanatory variables

Demographic characteristics

This group of variables includes chronological age (in years), the increase in chronological age between waves 6 and 7, gender and the change in marital status. In particular, the change in marital status is re-coded in 4 categories (no change in marital status between waves 6 and 7, widowed, divorced and other changes in marital status). The aim of this variable is to isolate the impact of widowhood and divorce on updating SSPs.

Country of residence is used as a control variable because it does not vary between waves.

Socio-economic factors

This group of variables includes the change, between waves 6 and 7, in the “equivalised” individual income quartiles. The change in “equivalised” individual income quartiles is re-coded in three groups; no change in income quartiles between waves; a fall of income to lower quartiles, a rise of income to higher quartiles. The “equivalised” income per individual was calculated using the reported household income and the OECD-modified equivalence scale. This scale, first proposed by Haagenars et al. (1994), assigns a value of 1 to the household head, of 0.5 to each additional adult member and of 0.3 to each child. The aim of this variable is to isolate the impact income changes on updating SSPs.

Furthermore, the educational level in 4 categories, based on the ISCED-97 classification, including Primary (code 1), Lower Secondary (code 2), Upper Secondary (codes 3 & 4) Tertiary (codes 5 & 6), is used as a control variable. The educational level for respondents aged 50 or older does not vary materially between waves, hence it can be safely used as a control variable.

Physical Health

This group of variables includes the change in the number of chronic conditions, the change in the number of limitations in Activities of Daily Living and the change in self-rated health. In particular, the change in self-rated health (SRH) is re-coded in 5 levels (no change in SRH between waves 6 and 7, improvement by 1 scale, improvement by 2 or more scales, deterioration by 1 scale and deterioration by 2 or more scales). The aim of this variable is to isolate the impact of SRH improvement or deterioration on updating SSPs.

Lifestyle & Behavioral risk factors

This group includes the change BMI category of a respondent between the waves. This variable is re-coded in nine levels (no change in BMI group between waves 6 and 7, change from underweight to normal, change from underweight to overweight or obese, change from normal to underweight, change from normal to overweight or obese, change from overweight to obese, change from overweight to normal or underweight, change from obese to overweight and change from obese to normal or underweight). The aim is to investigate how SSPs are updated following a change in BMI category.

Quality of life

This group includes the change in the life satisfaction score between the waves. The aim of this variable is to investigate how SSPs are updated as life satisfaction varies during respondents’ lifespans.

Statistical modeling

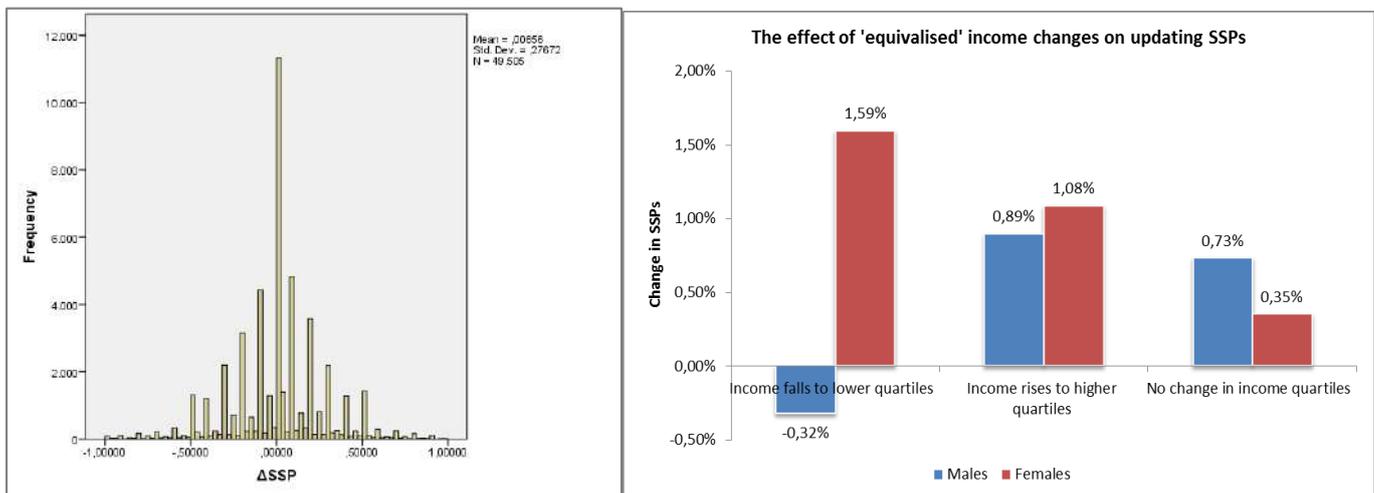
In the analysis we use two Generalised Linear Models (GLMs) with a linear link function, including ' ΔSSP ' as dependent variable, which is assumed to be normally distributed. The models are estimated for males and females separately.

3. Results

Sample

The sample characteristics are presented in Table 1. On average subjective survival probabilities between waves are marginally increased by 0.66% (Figure 1). Females revise upward their subjective survival probabilities more than males. The average respondent is 67.5 years old and males represent 43% of the sample. Approximately three out of a hundred women experience widowhood in the period during the waves. The frequency of divorces is higher for males than females. Males report a bigger increase in the number of chronic conditions and ADLs compared to females. Moreover, males report more frequently improvement in self-rated health by 2 or more scales as well as deterioration in self-rated health by 2 or more scales. On the other hand, females report no change in self-rated health more frequently than males. Overweight males decrease their BMI to normal or underweight category more frequently than females. Females on average report a larger reduction in life satisfaction score than males.

Figure 1 Distribution of the difference in SSPs; ' ΔSSP '. **Figure 2** Effect of 'equivalised' income changes on updating SSPs.



Overall there is a reduction in average 'equivalised' income between waves. This reduction is larger for males whose income remains within the same quartile or falls to lower quartiles between waves. On average, males revised downwards their SSPs following an income drop. However, this trend is not observed for females (Figure 2). Both genders revise their SSPs upwards following a rise in average 'equivalised' income.

Table 1 Sample characteristics (n = 49 505)

Descriptive measures	Males	Females	Total
Representation in the sample	43.08%	56.92%	100%
Subjective survival probabilities - Wave 6 (mean)	65.05%	64.72%	64.86%
Subjective survival probabilities - Wave 7 (mean)	65.64%	65.43%	65.52%
Dependent Variable			
Difference of subjective survival probabilities 'ΔSSP' (mean)	0.58%	0.71%	0.66%
Independent Variables			
Demographic Characteristics			
Chronological age at wave 6 (mean)	67.6 years	67.3 years	67.5 years
Increase in chronological age between waves (mean)	2.0 years	2.0 years	2.0 years
Change in Marital Status			
No change in marital status	97.51%	96.14%	96.74%
Widowed	1.23%	2.69%	2.06%
Divorced	0.30%	0.23%	0.26%
Other change in marital status	0.96%	0.92%	0.94%
Socio-economic Status			
Change in 'Equivalised' Income quartiles*			
No change in income quartiles (average change in €)	-5,114	-2,902	-3,857
Income falls to lower quartiles (average decrease in €)	-20,352	-18,835	-19,518
Income rises to higher quartiles (average increase in €)	12,865	13,864	13,452
Physical Health			
Change in the number of chronic conditions (mean)	0.17	0.13	0.14
Change in the number of ADLs (mean)	0.06	0.05	0.05
Change in self-rated health			
No change in self-rated health between waves	49.90%	52.10%	51.20%
Improvement by 2 or more scales	3.70%	3.50%	3.60%
Improvement by 1 scale	17.10%	17.50%	17.30%
Deterioration by 1 scale	23.20%	22.20%	22.60%
Deterioration by 2 or more scales	6.10%	4.70%	5.30%
Lifestyle and behavioral risk factors			
Change in BMI category			
No change in BMI category between waves	79.69%	79.40%	79.53%
Change from underweight to normal	0.14%	0.49%	0.34%
Change from underweight to overweight or obese	0.01%	0.07%	0.05%
Change from normal to underweight	0.25%	0.64%	0.47%
Change from normal to overweight or obese	6.11%	6.24%	6.18%
Change from overweight to obese	3.40%	3.63%	3.53%
Change from overweight to normal or underweight	6.39%	5.40%	5.83%
Change from obese to overweight	3.62%	3.50%	3.56%
Change from obese to normal or underweight	0.38%	0.63%	0.52%
Quality of Life			
Change in life satisfaction (mean)	-0.05	-0.07	-0.06

*This is the average income change between waves 6 and 7 in €.

A comparison of the characteristics between respondents who participated in both waves and those who participated only in Wave 6 is presented in Table 2. On average, respondents who participated only in Wave 6 report lower SSPs; are older; are less educated; have more chronic conditions and ADLs; report poorer self-rated health, are more frequently underweight and are less satisfied with their lives compare to respondents who participated in both waves. On the other hand, respondents who participated in both waves are more frequently married and have a lower frequency of widowhood. The 'equivalised' income is not materially different between respondents who participated in both waves and those who participated only in Wave 6. Overall, the combined longitudinal sample for both waves has a lower representation of older, less educated, less satisfied and less healthy individuals. Males have also lower representation in the sample.

Table 2 Characteristics of respondents who participated in both waves compared to those who participated only in Wave 6.

Descriptive measures	Respondents who participated in Waves 6 & 7	Respondents who participated in Wave 6 but not in Wave 7	Respondents who participated in Wave 6 but died before W7
Number of respondents	51,849	14,397	1,985
Subjective survival probabilities - Wave 6 (mean)	64.88%	63.48%	44.25%
Demographic Characteristics			
Gender			
Males	42.98%	44.40%	53.15%
Females	57.02%	55.60%	46.85%
Chronological age (mean)	67.4 years	67 years	79.4 years
Marital Status			
Married	68.39%	68.95%	51.94%
Partnership	1.75%	2.09%	1.21%
Separated	1.35%	1.44%	1.66%
Never married	5.61%	5.30%	5.84%
Divorced	8.46%	7.63%	5.24%
Widowed	14.45%	14.59%	34.11%
Socio-economic Status			
'Equivalentised' Income (mean value by quartile)			
1 st Quartile (mean in €)	3,880	3,681	3,997
2 nd Quartile (mean in €)	10,144	10,115	9,895
3 rd Quartile (mean in €)	19,679	19,477	18,971
4 th Quartile (mean in €)	57,518	57,011	51,922
Education level			
ISCED-97 code 0 & 1 (Primary)	21.36%	27.73%	40.20%
ISCED-97 code 2 (Lower secondary)	17.54%	16.94%	18.99%
ISCED-97 codes 3 & 4 (Upper secondary)	38.19%	34.26%	29.22%
ISCED-97 codes 5 & 6 (Tertiary)	22.90%	21.06%	11.59%
Physical Health			
Number of chronic conditions (mean)	1.77	1.79	2.70
Number of ADLs (mean)	0.22	0.35	1.34
Self-rated health			
Excellent	7.16%	6.81%	1.36%
Very good	18.29%	17.87%	4.48%
Good	36.74%	33.92%	18.04%
Fair	28.09%	28.05%	38.64%
Poor	9.72%	13.34%	37.48%
Lifestyle and behavioral risk factors			
BMI			
Underweight	1.00%	1.35%	3.43%
Normal	39.20%	41.29%	43.32%
Overweight	40.60%	40.02%	36.32%
Obese	19.19%	17.35%	16.93%
Quality of Life			
Life satisfaction (mean)	7.71	7.47	7.03

Males, those with lower socio-economic status and lower life satisfaction as well as older respondents demonstrate higher mortality compared to individuals who did not respond for reasons other than death. It is worth noting that the average SSP for respondents who died is 44%; considerably lower than the other two groups. Furthermore, poor self-rated health, more chronic conditions, more functional limitations and underweight BMI category demonstrate a strong association with in-sample mortality.

Multivariable analyses

The signs of the models' coefficients indicate the influence of predictors on updating subjective survival probabilities. Positive coefficients indicate an upward revision whereas negative coefficients indicate a downward revision of subjective survival probabilities. Chronological age is negatively associated with the difference of subjective survival probabilities ($b = -0.001$). Individuals tend to revise down their SSPs as they become older. However, the increase in chronological age is weakly associated with a positive revision of SSPs for males ($b = 0.001$) and a negative revision of SSPs for females ($b = -0.009$). Widowhood is associated with negative SSPs revisions for both genders. This association is stronger for females ($b = -0.025$). Getting a divorce is associated with positive SSPs revisions for males ($b = 0.048$) but not for females ($b = -0.002$).

Table 3 Coefficients based on generalised linear models.

Independent Variables	Males	Females
Intercept	0.105**	0.104**
Demographic Characteristics		
Chronological age at Wave 6	-0.001**	-0.001**
Increase in chronological age between waves	0.001	-0.009
Change in Marital Status (reference : No change in marital status)		
Widowed	-0.015	-0.025*
Divorced	0.048	-0.002
Other change in marital status	-0.007	-0.002
Socio-economic Status		
Change in 'Equivalised' Income quartiles (reference : No change in income quartiles)		
Income falls to lower quartiles	-0.007	0.013*
Income rises to higher quartiles	0.002	0.006
Physical Health		
Change in the number of chronic conditions	-0.005**	-0.007**
Change in the number of ADLs	-0.012**	-0.006**
Change in self-rated health (reference : No change in self-rated health)		
Improvement by 2 or more scales	0.051**	0.060**
Improvement by 1 scale	0.019**	0.023**
Deterioration by 1 scale	-0.026**	-0.016**
Deterioration by 2 or more scales	-0.055**	-0.034**
Lifestyle and behavioral risk factors		
Change in BMI category (reference : No change in BMI category)		
Change from underweight to normal	0.103*	-0.007
Change from underweight to overweight or obese	0.153	0.060
Change from normal to underweight	-0.111**	-0.006
Change from normal to overweight or obese	0.009	0.009
Change from overweight to obese	0.006	-0.002
Change from overweight to normal or underweight	0.002	0.002
Change from obese to overweight	-0.018	0.002
Change from obese to normal or underweight	0.012	-0.007
Quality of Life		
Change in life satisfaction	0.019**	0.019**

** $p < 1\%$, * $p < 5\%$. The dependent variable the difference in the subjective survival probabilities 'ASSP'. Controlling for country of residence and educational level.

Males and females who face an increase in 'equivalised' income tend to revise upwards their SSPs ($b = 0.002$ and $b = 0.006$). In contrast, following an income drop males revise downwards their SSPs ($b = -0.007$) whereas females revise upwards their SSPs ($b = 0.013$). An increase in chronic conditions and ADLs is strongly associated with negative SSPs revisions for both genders. Furthermore, an improvement of self-rated health is associated with positive SSPs revisions whereas a deterioration of self-rated health is associated with negative SSPs revisions for both genders. A deterioration

of self-rated health by 2 or more scales has a stronger negative influence on SSPs revisions for males ($b = -0.055$) whereas an improvement of self-rated health by 2 or more scales has a stronger positive influence on SSPs revisions for females ($b = 0.060$).

Males who increase their BMI from underweight to normal or from underweight to overweight or obese, tend to revise upward their SSPs ($b = 0.103$ and $b = 0.153$). Furthermore, males and females who reduce their BMI from normal to underweight tend to revise downward their SSPs ($b = -0.111$ and $b = -0.006$). A reduction of BMI from obese to overweight is associated with a decrease in SSPs for males ($b = -0.018$) and an increase in SSPs for females ($b = 0.002$). An increase in life satisfaction is associated with positive SSPs revisions for both genders.

4. Discussion

The main objective of this study is to evaluate gender, health and socio-demographic influences on updating subjective survival probabilities. A longitudinal dataset from Wave 6 and Wave 7 of the SHARE study was used, including 49505 respondents from 18 countries. The influences are examined for males and females separately, using the difference in subjective survival probabilities between waves as dependent variable. Our findings are discussed in comparison to actual mortality patterns.

Factors associated with positive revisions of subjective survival probabilities

Increases in 'equivalised' income, in life satisfaction as well as improvements in self-rated health lead to positive revisions of subjective survival probabilities. Higher socio-economic status, better life satisfaction and better self-rated health are also associated with lower actual mortality (Kaplan et al. 1996, St. John et al. 2015, Idler and Benyamini 1997) as well as with higher subjective life expectancy (Rarrange et al. 2016; Hurd and McGarry 1995; van Solinge and Henkens 2018). Therefore we conclude that individuals identify and incorporate the positive changes in these factors by revising upwards their SSPs. These revisions are consistent with actual mortality patterns.

Transitions from the BMI categories normal or underweight to overweight or obese are associated with positive SSPs revisions for both males and females. However, actual mortality is lower for individuals whose BMI is normal and higher for the other categories (Berrington et al. 2010). Therefore we conclude that the particular transitions in BMI categories lead to SSPs revisions which are inconsistent with actual mortality patterns.

Factors associated with negative revisions of subjective survival probabilities

Older chronological age, increase in the number of chronic conditions and ADLs and deterioration in self-rated health lead to negative revisions of subjective survival probabilities. The deterioration of functional health and the increase in the number of comorbidities are also linked to higher actual mortality (Scott et al. 1997, Lee et al. 2008). Widowhood is associated with negative revisions of subjective survival probabilities for both males and females. It is worth noting that mortality increases during the first year following bereavement (Kaprio et al. 1987) but it is gradually reduced over the following years. Males and females who decrease their BMI from normal to underweight category revise downwards their SSPs and it is well documented that underweight is a strong predictor of mortality (Takala et al. 1994). Overall, we

conclude that individuals identify the events above which could deteriorate their mortality risk profile and revise downwards their SSPs.

Males vs Females

The transition out of marriage through a divorce affects differently the revisions of survival expectations for males and females. Men revise upward whereas women revise downwards their SSPs after getting divorced. However, divorce is associated with higher mortality for both genders (Dupre et al. 2009). Hence we conclude that SSPs revisions following the event of a divorce are consistent with actual mortality patterns for females but not for males. A drop in income results in negative SSPs revisions for males but positive SSPs revisions for females. Reductions in income are particularly important for the mortality of individuals in the lower end of the socio-economic spectrum (Backlund et al. 1999). We conclude that males incorporate in SSPs revisions the event of income reduction, consistently with actual mortality patterns.

Males who increase their BMI from underweight to normal category or from obese to normal category tend to revise upward their subjective survival probabilities. The opposite trend is observed for females. Since all BMI categories, apart from the normal, are linked to higher mortality (Berrington et al. 2010) we conclude that males incorporate in their SSPs revisions these specific BMI transitions consistently with actual mortality. It is worth noting that according to our results, certain BMI transitions lead to SSPs revisions which are inconsistent with actual mortality patterns. This could be partially due to the small number of respondents in certain BMI transitions.

5. Limitations

Some limitations of the study should be taken into account when considering the findings. First, the longitudinal sample of Wave 6 and Wave 7 is reduced to 49505 individuals (-27%). The shrinkage of longitudinal sample size could be partially due to country-specific retention rate targets, ranging from 75% to 85%, set by SHARE (Malter et al. 2018). As a result the longitudinal sample has a lower representation of older, less educated, less satisfied and less healthy individuals. However, according to our findings the effect of chronological age, life satisfaction and health status is consistent with actual mortality patterns for both males and females. Furthermore, education is included as a control variable.

Second, the number of respondents is low for transitions between certain BMI categories. This reduces the validity of our results related to the effect of BMI on SSPs revisions. However, there is evidence that SSPs revisions related to certain BMI transitions are consistent with actual mortality patterns. Further analyses using larger datasets are required to confirm our findings regarding to the influence of BMI transitions on SSPs revisions and how these revisions vary by gender.

6. Conclusion

This study shows that individuals update their SSPs after the occurrence of specific events, consistently with actual mortality patterns. In addition, the degree of consistency varies by gender. According to our results, males and females tend to revise their SSPs after events such as self-rated and functional health deterioration or improvement, increase in chronic conditions, widowhood, increase in income, increase in life satisfaction and BMI transitions from normal to underweight consistently with actual mortality patterns. Males tend to revise their SSPs following a drop in income or a

BMI transition from underweight to normal, consistently with actual mortality patterns. Moreover only females tend to revise their SSPs after a getting divorce consistently with actual mortality patterns. Future research should involve a more detailed examination of the influences of a wide range of events on SSPs revisions, using large longitudinal datasets covering many aspects of respondents' lives.

Acknowledgements

This paper uses data from SHARE Waves 6 and 7 (DOIs 10.6103/SHARE.w6.700, 10.6103/SHARE.w7.700), see Börsch-Supan et al. (2013) for methodological details. The SHARE data collection has been funded by the European Commission through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812), FP7 (SHARE-PREP: GA N°211909, SHARE-LEAP: GA N°227822, SHARE M4: GA N°261982) and Horizon 2020 (SHARE-DEV3: GA N°676536, SERISS: GA N°654221) and by DG Employment, Social Affairs & Inclusion. Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, the U.S. National Institute on Aging (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGH4_04-064, HHSN271201300071C) and from various national funding sources is gratefully acknowledged(see www.share-project.org).

References

- Arpino, B., Bordone, V., & Scherbov, S. (2018). Smoking, education and the ability to predict own survival probabilities. *Advances in Life Course Research*, 37, 23-30.
- Backlund, E., Sorlie, P. D., & Johnson, N. J. (1999). A comparison of the relationships of education and income with mortality: the National Longitudinal Mortality Study. *Social science & medicine*, 49(10), 1373-1384.
- Balia, S. (2011). *Survival expectations, subjective health and smoking: Evidence from European countries*. University of York, Centre for Health Economics.
- Berrington de Gonzalez, A., Hartge, P., Cerhan, J. R., Flint, A. J., Hannan, L., MacInnis, R. J., ... & Beeson, W. L. (2010). Body-mass index and mortality among 1.46 million white adults. *New England Journal of Medicine*, 363(23), 2211-2219
- Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S. and Zuber, S. (2013). Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology* DOI: 10.1093/ije/dyt088
- Börsch-Supan A. (2017) Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 6. Release version: 6.0.0. SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w6.600
- Börsch-Supan, A. (2019). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 7. Release version: 7.0.0. SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w7.700
- Dupre, M. E., Beck, A. N., & Meadows, S. O. (2009). Marital trajectories and mortality among US adults. *American journal of epidemiology*, 170(5), 546-555.

study. *The Journal of Human Resources* pp. S268-S292.

Griffin, B., Loh, V. and Hesketh, B. (2013). A mental model of factors associated with subjective life expectancy. *Social science & medicine* 82, pp. 79-86.

Hagenaars, A. J., De Vos, K., & Asghar Zaidi, M. (1994). Poverty statistics in the late 1980s: Research based on micro-data.

Hurd M.D. and McGarry K. (1995). Evaluation of the subjective probabilities of survival in the health and retirement

Idler E. L. and Benyamini Y. 1997. Self-rated health and mortality: a review of twenty-seven community studies. *Journal of health and social behavior*, pp. 21-37.

Kaplan, G. A., Pamuk, E. R., Lynch, J. W., Cohen, R. D., & Balfour, J. L. (1996). Inequality in income and mortality in the United States: analysis of mortality and potential pathways. *Bmj*, 312(7037), 999-1003.

Kaprio, J., Koskenvuo, M., & Rita, H. (1987). Mortality after bereavement: a prospective study of 95,647 widowed persons. *American Journal of Public Health*, 77(3), 283-287.

Lee, S. J., Go, A. S., Lindquist, K., Bertenthal, D., & Covinsky, K. E. (2008). Chronic conditions and mortality among the oldest old. *American journal of public health*, 98(7), 1209-1214

Liu J. T., Tsou M. W., and Hammitt J. K. (2007). Health information and subjective survival probability: Evidence from Taiwan. *Journal of Risk Research* 10(2): 149-175.

Malter, F.; Schuller, K , Börsch-Supan, A. (2018). SHARE Compliance Profiles – Wave 7. Munich: MEA, Max Planck Institute for Social Law and Social Policy.

Mirowsky J. (1999). Subjective life expectancy in the US: correspondence to actuarial estimates by age, sex and race. *Social science & medicine* 49(7): 967-979.

Rappange D. R., Brouwer W. B. and Exel J. (2016). Rational expectations? An explorative study of subjective survival probabilities and lifestyle across Europe. *Health Expectations* 19(1): 121-137

Scott, W. K., Macera, C. A., Cornman, C. B., & Sharpe, P. A. (1997). Functional health status as a predictor of mortality in men and women over 65. *Journal of clinical epidemiology*, 50(3), 291-296.

St. John, P. D., Mackenzie, C., & Menec, V. (2015). Does life satisfaction predict five-year mortality in community-living older adults?. *Aging & mental health*, 19(4), 363-370.

Takala, J. K., Mattila, K. J., & Rynänen, O. P. (1994). Overweight, underweight and mortality among the aged. *Scandinavian journal of primary health care*, 12(4), 244-248.

van Solinge, Hanna, and Kène Henkens. "Subjective life expectancy and actual mortality: results of a 10-year panel study among older workers." *European journal of ageing* 15, no. 2 (2018): 155-164.

Multivariate Outlier Detection with ICS and Application to Statistical Quality Control & Autocorrelated Data

Papageorgiou Ioulia¹ and Voutsinas Stefanos²

¹ Department of Statistics, Athens University of Economics & Business, Athens, Greece

(E-mail: ioulia@aub.gr)

² Department of Statistics, Athens University of Economics & Business, Athens, Greece

(E-mail: stefanosvoutsinas@gmail.com)

Abstract. A general method for exploring multivariate data by comparing different estimates of multivariate scatter and location functionals is presented. The method is based on the eigenvalue-eigenvector decomposition of one scatter matrix relative to another. A standardization of the data is firstly conducted by using a scatter statistic and then a principal component method with a second scatter statistic is performed. Statistical Quality Control is a field that the detection of special structures is in high priority and specifically when the bulk of the data is correlated. In this case the methods for detecting multivariate outliers fail to reveal them and in this paper we illustrate the performance of the ICS, Mahalanobis distance and T^2 Hotelling chart plot when we add outliers in highly correlated data, moderate correlated data and then in moderate correlated data without outliers. The scope of this study is to compare different pairs of scatter statistics, evaluate the performance of ICS when we increase the autocorrelation and the dimension of the problem. In statistical quality control we care for the detection of the outliers (True Positive) and of course for the percentage of the False Alarms or False Positives (FP). The goal is to have a proportion close to 100 of TP and a low percentage for the FP, close to 0. As a measure of comparison we have the same data for every case we use but uncorrelated; by doing this we can see the rate of TP and FP for every case with correlated and uncorrelated data. Finally, the Mahalanobis Distance (MD) and T^2 Hotelling chart plot are used in order to make comparisons.

Keywords: Affine Equivariance, Invariance, Robust Estimates, Scatter Statistics, Breakdown Point, Mahalanobis Distance, T^2 Hotelling chart plot, Metropolis Hastings, Autocorrelation

1 Introduction

We live in a world where the data is getting bigger by the second. The value of the data can diminish over time if not used properly. Finding anomalies either online in a stream or offline in a dataset is crucial to identifying problems in the business or building a proactive solution to potentially discover the problem before it happens or even in the exploratory data analysis (EDA) phase to prepare a dataset for Machine Learning (ML). Detecting outliers or anomalies is one of the core problems in data mining. The emerging expansion and continued growth of data and the spread of Internet of Things (IoT) devices, make us rethink the way we approach anomalies and the use cases that can be built by looking at those anomalies. Another reason why we need to detect anomalies is that when preparing datasets for machine learning models, it is really important to detect all the outliers and either get rid of them or analyse them to know why you had them there in the first place.

Specifically, detecting multivariate outliers in industrial, medical and financial applications (Aggarwal 2017, Ngai et al. 2011) is an emerging topic of great importance and data mining techniques are applied so as to detect and explore the data sets. In the present paper, we proposed an alternative to Mahalanobis distance, Principal Component Analysis and Exploratory Projection Pursuit and we use the Invariant Coordinate Selection that proposed by Tyler et al.[6]. The method uses two affine equivariant scatter statistics to transform the data. The transformation gives the invariant components where some of them can be selected to interpret the outliers. The invariant components derived from ICS are affine invariant only for affine equivariant robust/nonrobust statistics, while the principal components are orthogonally invariant but scale dependent.

The detection of outliers in real time or offline analysis in a high level of quality control procedure is of great interest the last years since a small proportion of the data is categorised as outliers. To monitor this kind of data, we use ICS methodology, Mahalanobis distance and T^2 Hotelling chart plot in a casewise contamination context and when the number of observations is larger than the number of variables. The most important is that we use correlated datasets in order to show how a real problem would be. We compare the performance of ICS in correlated data of d dimension with uncorrelated data of the same dimension coming from the same mechanism. By saying same mechanism we mean that the data have the same mean and variance. We also compare the performance of ICS by using four different pairs of scatter statistics of different class and breakdown point (Maronna et al.[2]) such as Minimum Volume Ellipsoid (Rousseeuw,[12]), Minimum Covariance Determinant (Rousseeuw et al.[4]), Constrained Mestimates (Kent et al.[8]), τ estimates (Lopuhaä, [23]) and Sestimates (Lopuhaä et al.[22]).

In the statistical programme R one can find the commands for ICS methodology and a wide range of estimates in the ICSNP package (Nordhausen et al.[26]), ICS package (Nordhausen et al.[1]), ICSOutlier package (Nordhausen et al.[16]), robustbase package (Rousseeuw et al.[19]) and rrcov (Todorov et al.[30]). For a better representation of the

method with illustrated examples on ICS, the reader can check out some studies on statistical quality control conducted by Archimbaud et al.[16] and Archimbaud et al.[17]. Generally, the studies argue that ICS is a useful method for outlier detection, however, all experiments have been conducted on uncorrelated data, and the aim of this paper is to show that the method works even in this kind of datasets showing the outliers or clustering. The article is organized as follows. In Section 2 we illustrate the different pairs of scatter statistics and give the theoretical background for the ICS method. In Section 3 we describe the structure of the data, how the outliers were generated and the way we separate the cases we will work on. Also, Section 2 provides practical guide lines for the choice of the most informative invariant components where Section 4 provides results that help us monitor the performance of ICS on autocorrelated data. In section 5, we make comparisons among Mahalanobis distance, ICS and T^2 Hotelling chart plot so as to see an overall performance of the methods when we have to deal with high level quality control data that are correlated. The illustration of the results is possible with the use of boxplots where one can measure the variability of the simulated values by using the interquartile range (IQR), sample mean and if there are outliers or extreme outliers. For all three cases of data we extrapolate a conclusion for each and suggest specific pairs of scatter and location statistics.

2 Invariant Coordinate Selection Method

Invariant Coordinate Selection (ICS) is proposed by Tyler et al.[6] as a method for exploring multivariate data. The method suggests the use of two different estimates of the scatter for the data aiming to capture different aspects of the underlying distribution. Conclusions are derived by comparing the spectral decomposition of one scatter matrix against the other.

Additionally, the resulted invariant components are affine invariant on the strict condition that the scatter and location statistics are affine equivariant. Despite its attractive properties, ICS has not been extensively studied in the literature on outlier detection. An early version of ICS was proposed in Caussinus and Ruiz (1990) for multivariate outlier detection and studied further in e.g Penny and Jolliffe (1999) and Caussinus et al. (2003) for two specific matrices. The principle of ICS is quite similar to Principal Component Analysis (PCA) with coordinates or components derived from an eigenvalue decomposition followed by a projection of the data on selected eigenvectors. Recent articles by Nordhausen et al.[3] and Tyler et al.[6] argue that ICS is useful for outlier detection. Moreover, Nordhausen et al.[14] has important applications on statistical quality control using different scatter statistics. However, testing ICS on autocorrelated data was never performed and so this paper will fill this gap.

2.1 The choice of the scatter pairs

As we mentioned above, the ICS works with 2 affine equivariant scatter statistics where the scientist has to choose in order to find potential atypical values or anomalies in the data. Caussinus and Ruiz (1990), Caussinus et al. (2003) and Tyler et al.[6] recommend using Class I scatter estimators such as the classical one or some weighted scatter matrices. The main reason for this choice is that these estimators are simple and can be computed rapidly. Moreover, the nice properties of ICS given by Tyler Theorem 3 and 4 in Tyler et al.[6] are true even for nonrobust estimators such as the sample covariance matrix (Cov) and Covariance matrix of fourth moment (Cov₄). In Nordhausen et al.[17] we can see the performance of scatter statistics belong to different Class. The first pair was based on two Class I estimators $V_1 = \text{Cov}$ and $V_2 = \text{Cov}_4$, while the others are based on Class II and I with $V_1 = \text{MLC}$ (Cauchy Mestimates) and $V_2 = \text{Cov}$, Class III and I with $V_1 = \text{MCD}$ (Minimum Covariance Determinant) and $V_2 = \text{Cov}$ and Class III and II scatter estimators with $V_1 = \text{MCD}$ and $V_2 = \text{MLC}$.

In this paper we illustrate the performance of these pairs of scatter statistics:

- Class III and Class I with $V_1 = \text{MCD}$ and $V_2 = \text{Cov}_4$
- Two Class I scatter statistics with $V_1 = \text{Cov}$ and $V_2 = \text{Cov}_4$
- Class I and Class II with $V_1 = \text{Cauchy}$ and $V_2 = \text{Cov}_4$
- Class III and Class II with $V_1 = \text{MCD}$ and $V_2 = \text{Cauchy}$

In pair 1 for the MCD, the breakdown point is 30% while in pair 4 is 50%.

2.2 The Invariant Components Selection

When we perform ICS, we finally get q invariant components which are ordered from the most informative to the less informative. This happens by ordering the eigenvalues taken from the approach $V_1^{-1}V_2$ or the invariant components themselves. Both approaches test whether each invariant component is significantly relevant via sequential approaches. The nonGaussian components are selected since in the normal distributed is very rare to find outliers. When the procedure identifies a normal distributed component it stops.

Concerning the first approach with the eigenvalues, we conduct a Parallel Analysis (PA) based on Monet Carlo simulations of a large number of eigenvalues for a spherical population with the same dimensions as our data sets. The Parallel analysis has already been used by Caussinus et al. (2003) for ICS but only for a particular pair of scatters and Nordhausen et al. (2018) for four pair of scatter statistics and of different Class. The procedure can be easily adapted to any pair of scatter statistic and is made of four simple steps:

1. Simulate Independent Gaussian distributions of the same dimensions as the data matrix being considered.
2. Perform ICS with the combination of scatter matrices of interest computed on the simulated data.
3. Repeat steps 1 and 2 for a large number of times.
4. Calculate for each component j , the $1-\alpha_j$ percentile of the eigenvalues.

The percentiles are then used as cutoff values for assessing the relevance of the eigenvalues. More analytically, sequentially from $j = 1$, if the observed j th eigenvalue exceeds the cutoff associated to the j th percentile, then the j th invariant component is considered nonGaussian and so informative for the analysis. When one observed eigenvalue is smaller than its associated cutoff value, the invariant component is considered as noninformative, the test procedure stops and we get the $(j-1)^{\text{th}}$ components. In Big data problems where the algorithm is sensitive and may choose many invariant components, we make some adjustments on the initial significance level α . Dray (2008) suggested applying Bonferroni correction on the significance level and consider a level $\alpha_j = \alpha/j$ for each component $j = 1 \dots p$.

The second approach makes use of the fact that informative components for detecting outliers do not follow a Gaussian distribution and it is thus based on univariate normality tests for each component. We test for each component the null hypothesis that it is Gaussian at the level α_j , beginning with the first one as previously described for Parallel analysis method. When we find one invariant component which is not normally distributed, the test procedure stops and the non-Gaussian components are retained. The package ICSOutlier in R (Archimbaud et al.[16, 17]) contains five normality tests which are the D'Agostino test of skewness (DA), the Anscombe-Glynn (AG) test of kurtosis, the Bonett-Seier (BS) test of Geary's kurtosis, the Jarque-Bera (JB) test based on both skewness and kurtosis and the Shapiro-Wilk (SW) normality test (see Yazici and Yolacan[28] and Bonett and Seier[29] for a complete description of these five tests). From my experience, the results of the SW test are usually the best and in this paper we use this method.

As we mentioned above, these methods are automated and in the real world where thousands of data are collected we need to explore briefly their structure and see anomalies in the data. For sure, when we analyse one dataset in order to get information, those methods can be avoided since we have plenty of time to explore which components are essential and which not. Scree plots can show the most informative by checking the Generalized Kurtosis Measure (GKM) where invariant components with GKM over one are important and can be kept for further analysis. In this paper we compare all these three methods where we keep the true invariant components, the results from the Parallel Analysis and Marginal Normality tests.

2.3 Measure of Outlierness

Let q denote the number of invariant components selected by one of the two test procedures described previously or by looking at the scree-plot of the eigenvalues. The next step is to compute the squared distances, defined as the Euclidean distances of the selected centered components. For outlier detection, the squared ICS distances can be used as a measure of outlierness. The invariant coordinates are centered by the location statistic T_1 and we denote them as Z_n . Specifically:

$$Z_n = (z_1, \dots, z_n)' = (X_n - 1_n T_1'(X_n))B'(X_n)$$

Let Z_k be the k components of Z selected by the parallel analysis or the marginal normality tests. The ICS distance of the observation i is defined as:

$$ICSD^2(x_i, q) = \|Z_q\| \|Z_q\|.$$

Note that if all components are selected, the ICS distances are equivalent to the Mahalanobis distances computed with respect of the first scatter and the associated location statistic specified in S_1 .

2.4 Cut-off values

After calculating the distances, we have to create a cut-off value in order to detect the extreme ICS distances. The procedure uses simulation under a multivariate standard normal model for a specific data set-up and scatters combination. Specifically, the function extracts the dimension of the data and simulates m (for example 1000) times from a multivariate standard normal distribution, the squared ICS distances of q components. The resulting value is then the mean of the m corresponding quantiles of these distances at level $(1-\alpha)$. Finally, if the distance of an observation exceeds the expectation under the normal model, the observation is labelled as outlier.

The implementation of ICS for outlier detection in the next sections is performed in R Studio Version 1.2.1355 using the packages ICS (Nordhausen et.al.[3]), ICSOutlier (Archimbaud et.al.[14, 16]), mvtnorm (Genz et. al.[31]), moments (Komsta and Novomestky[32]), robustbase (Rousseeuw et. al.[19]) and rrcov (Todorov[30]). The Metropolis Hastings algorithm is attached in the Appendix for generating the same datasets.

3 Simulation Studies

The research of this paper is an extension of the work of Aurore Archimbaud, Klaus Nordhausen and A. Ruiz Gazen[14] with different assumptions being made in order to further analyse the effects of correlated subsamples and correlated variables on statistical process control. Some well-known charts which are used in Statistical Quality Control are the \bar{X} , R and S charts where in many cases fail to set up correct control limits because of the correlation. The data analyst has to recognize the correlation in the subsamples and set the proper limits for each chart in order to monitor the out-of-control data points. To show the importance of correlation we work with the \bar{X} chart where an illustration of the upper and lower control limits is presented in Section 3.1. Kai Yang and Walton M. Hancock[27] made a detailed study on statistical quality control for correlated samples working with \bar{X} , S, R and S^2 charts where the theoretical background for each chart is presented as well as application examples using software.

In the establishment of \bar{X} chart in SQC, the assumption is made that there is no correlation between the samples. However, in practice there are many cases where the correlation does exist within the samples. Neuhardt (1987) pointed out that when there exists a positive correlation between samples, the control limits for the \bar{X} chart will be substantially greater than, for example, the assumed α levels of 0.0027 for 3 sigma control limits. Increased, but unknown level α means that out-of-control points will occur more frequently than assumed where no reason can be found and corrected. Moreover, $\sigma_{\bar{X}}$, the standard deviation of the sample mean will be underestimated if correlation exists that will cause an over-estimation of the capability of the process. In subsection 3.1 we illustrate an example of the Type 1 error α which becomes larger when we have correlated samples. For simplicity, we use as example the \bar{X} chart and its control limits.

3.1 Control Limits for \bar{X} charts

The \bar{X} chart is a type of Shewhart control chart that is used to monitor the arithmetic means of successive samples of constant size, n . If the subgroups are positively correlated, then the control limits have to be revised. In a traditional approach, it is assumed that there is no correlation within samples and the sample mean is normally distributed with mean μ and standard deviation (Duncan 1974, Montgomery 1985), where σ is the process standard deviation and κ the length of each subgroup. The theoretical control limits (3σ) for the chart are:

$$[LCL_{\bar{x}}, UCL_{\bar{x}}] = \left[\mu - 3 \frac{\sigma}{\sqrt{\kappa}}, \mu + 3 \frac{\sigma}{\sqrt{\kappa}} \right] \quad (1)$$

Generally, the μ and σ are not known amounts and so statistical estimates, $\hat{\mu}$ and $\hat{\sigma}$, are used respectively. For the amount $\hat{\mu}$ we use as an estimator the grand average $\bar{\bar{x}}$ and $\hat{\sigma}$ will be obtained by the average range \bar{R} (See Appendix A for the notations). Now the control limits of the chart will be:

$$[LCL_{\bar{x}}, UCL_{\bar{x}}] = \left[\bar{\bar{x}} - 3 \frac{\bar{R}}{\sqrt{\kappa c_4}}, \bar{\bar{x}} + 3 \frac{\bar{R}}{\sqrt{\kappa c_4}} \right] \quad (2)$$

If there is correlation within the samples, the estimated $\hat{\sigma}$ will not give the unbiased estimate of process standard deviation σ . At Appendix B we can see that the expectation of $\hat{\sigma}$ will be equal to $\sqrt{(1-\rho)}\sigma$ and so, the actual expectation of control limits will be:

$$E[LCL_{\bar{x}}, UCL_{\bar{x}}] = \left[\mu - 3 \sqrt{\left(\frac{1-\rho}{\kappa}\right)} \sigma, \mu + \sqrt{\left(\frac{1-\rho}{\kappa}\right)} \sigma \right] \quad (3)$$

Moreover, when the subgroups are no longer independent, then the standard deviation of \bar{x} and $\sigma_{\bar{x}}$ will be equal to $\sqrt{[1 + (\kappa - 1)\rho/\kappa]}\sigma$ (See Appendix A). Finally, the theoretical control limits for the \bar{x} chart will be:

$$E[LCL_{\bar{x}}, UCL_{\bar{x}}] = \left[\mu - 3 \sqrt{\left(\frac{1+(\kappa-1)\rho}{\kappa}\right)} \sigma, \mu + \sqrt{\left(\frac{1+(\kappa-1)\rho}{\kappa}\right)} \sigma \right] \quad (4)$$

Which are the correct limits when the subsamples are correlated and there is correlation within the samples. Subsequently, the limits in equation (4) are wider than those in equation (3) by a factor of $\sqrt{1 + (\kappa - 1)\rho/(1 - \rho)}$. The reason of showing this example is to show the probability of Type 1 error, α . If $\rho > 0$, the probability of Type 1 error will be larger

than the computed when we assume that there is no correlation ($\rho=0$). For instance, when we assume that $\rho=0$, the 3σ control limits would have $\alpha = 2*0-2*\Phi(3) = 0.0027$, where $\Phi(\cdot)$ is the standard normal distribution function. On the other hand, if correlation exists, the probability of Type 1 error would be:

$$\alpha_{true} = 2 - 2\Phi(3\sqrt{(1-\rho)/[1+(\kappa-1)\rho]})$$

In summary, if we fail to detect a positive correlation and still use the limits like in eq. (2) then we will get narrower control limits than the correct ones. This means that there will be too many “out-of-control” points and the Type 1 error will be greater than assumed. For this study we chose the normal distribution and with the help of the Metropolis Hastings Algorithm we will get correlated data. When we increase the variance (σ) the autocorrelation is more intense and this helps us to create proper examples so as to show the performance of ICS. It is important here to stress that the standard functions for simulating observations from a statistical distribution when we use the statistical programme R will generate uncorrelated data points. Metropolis Hastings is adopted for the experiments in order to incorporate the autocorrelation feature among the measurements.

3.2 Metropolis Hastings

In statistics and statistical physics, the Metropolis-Hastings algorithm is a Markov chain Monte Carlo (MCMC) method for obtaining a sequence of random samples from a probability distribution from which direct sampling is difficult. This sequence can be used to approximate the distribution (e.g. to generate a histogram) or to compute an integral (e.g. an expected value). Metropolis-Hastings and other MCMC algorithms are generally used for sampling from multidimensional distributions, especially when the number of dimensions is high. For single-dimensional distributions, there are usually other methods (e.g. adaptive rejection sampling) that can directly return independent samples from the distribution, and these are free from the problem of autocorrelated samples that is inherent in MCMC methods. For this study, we did not face problem of generating problem from the normal distribution, but we wanted autocorrelated datasets to work on and test the ICS method.

3.3 Cases to Work on

To evaluate the performance of ICS, Mahalanobis distance and T^2 Hotelling chart plot we created three cases to work on.

Case A: Highly Autocorrelated

In this first case, we generated 2000 observations, in three scenarios: 10, 20 and 35 dimensions and we contaminated the 3% (60 outliers) of the data in three variables. The outliers added partially, in the 1st, 5th and 10th variable to make the

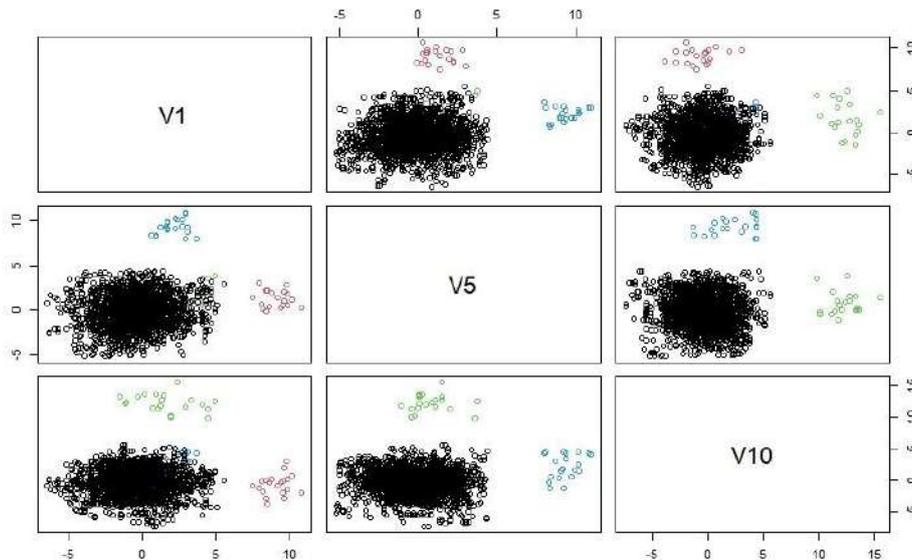


Figure 1: The three bulks of the outliers in the correlated data

problem more difficult and confusing. For a better understanding, in each of these variables (1-5-10) we contaminated 20 observations, while the rest variables do not contain any extreme value. By saying highly autocorrelated we mean that each observation in every variable is correlated on average 10% with its previous one and also it exists positive and negative correlation in the data. Moreover, all variables have approximately the same average correlation with small differences and the correlation is statistically significant. In Figure 1 the scatter plots of the variables with the augmented data are given. The three groups of outliers are plotted with different colours. The data points labelled with red are the data which have contaminated the measurements in variable 1, blue the data with contamination in variable 5 and green for those with contamination in the 10th variable. Specifically, the outliers which were added in variables 1, 5 and 10 are

normally distributed and each group follows $V_1[21:40] \sim N(9, 1)$, $V_5[761:780] \sim N(9, 1)$ and $V_{10}[501:520] \sim N(12, 1)$ respectively.

If we plot the Figure 1 without colouring the three bulks of the outliers, the scientist will be confused because it is difficult to distinguish the number of the subgroups since the blue and green subgroups follow the same distribution but they are monitored in different index and variable of the process. Also, in real conditions of an analysis, plotting all possible pairs of the variables as an approach to outlier detection cannot be possible due to large number of possible pairs and given the fact that the dimension of the problem is high, the scientist will never be able to plot and monitor anomalies. So, ICS methodology is a tool that would monitor the clusters or outliers in a high dimensional space. The main data come from a Multivariate Normal Distribution with mean equal to 0 and variance 2. The range of the data can be from -7 to +7 and we can see in Figure 1 that the values which are close to 3 standard deviations, look like outliers, but they are not. The reason for this problem is the autocorrelation, which means that every observation is connected with its previous one and the chain stays longer in the tails.

ICS and other methods for detecting outliers cannot deal with this kind of data and detect fake subgroups. On the other hand, when we treat the same data but without autocorrelation, the methods have a good performance despite the fact that the dataset has almost the same shape in the pairwise associations, range and subgroups. Figure 2 shows the data without correlation and one can see that there are small discrepancies. The most important is that we can easily distinguish the subgroups in both cases and so use the ICS methodology to detect the outliers and evaluate the method. In Figure 3, it is clear that there is strong autocorrelation and it will be very interesting to see the performance of ICS. We next illustrate the second (B) case where the autocorrelation is not so strong and the variance is lower.

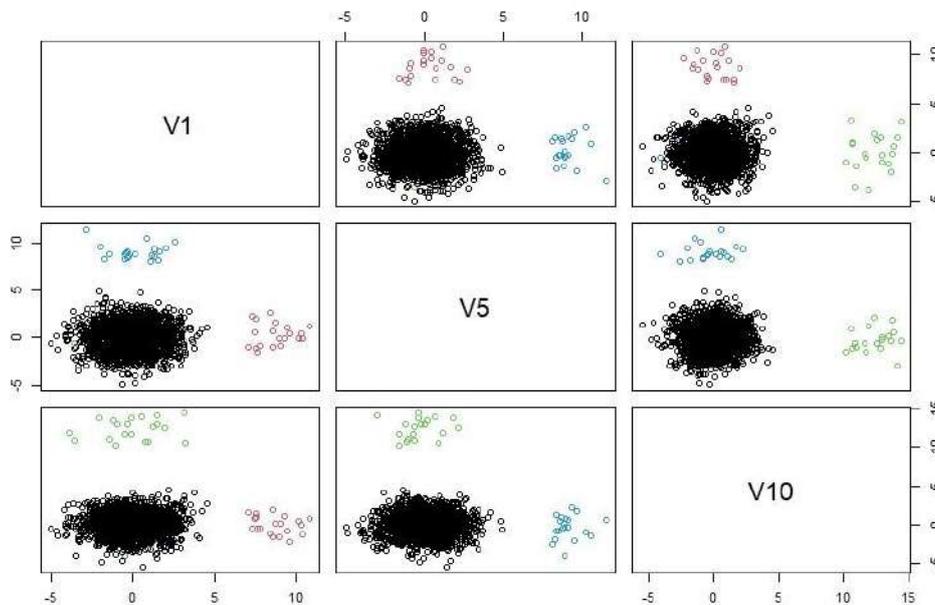


Figure 2: The data points with the bulk of the outliers in the uncorrelated data

Case B: Moderate Autocorrelated

In this case the Normal distributions generated by M-H algorithm have a mean equal to 0 and variance equal to 1. The elimination in the variance created a sample with lower autocorrelation than the case A. the majority of the data points do not have statistically significant correlation but it still exists as we can see in Figure 4. It is statistically significant but not at the same level as in case A. We expect that the ICS in these dataset will have a better performance and not fail revealing the atypical values. We also expect that the chart plots will be trapped even with this kind of autocorrelation and lead the scientist to wrong decision.

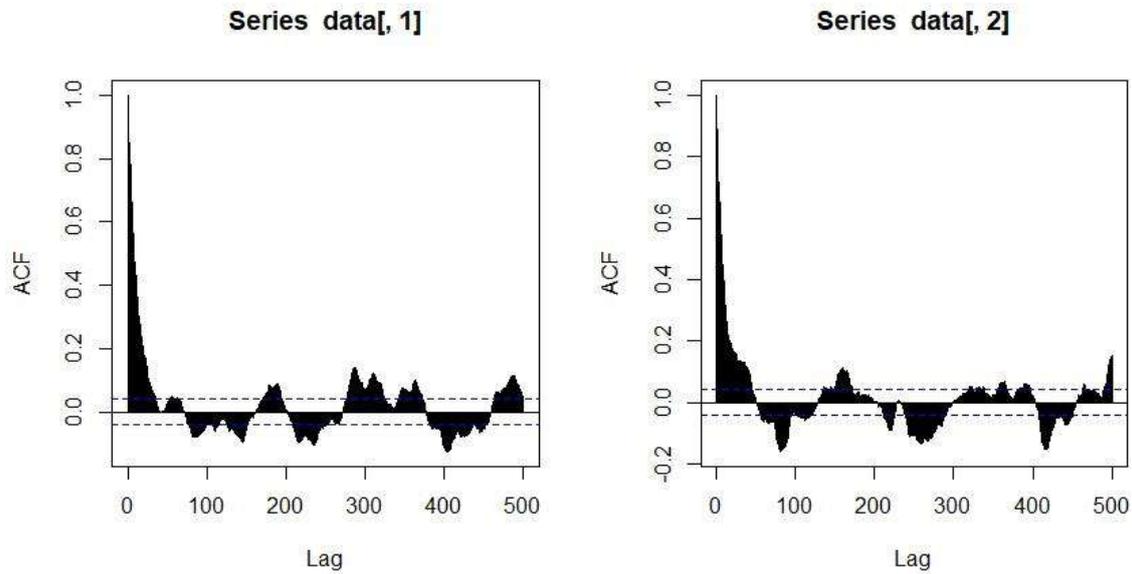


Figure 3: Autocorrelation plot for Variables 1-2 in Case A

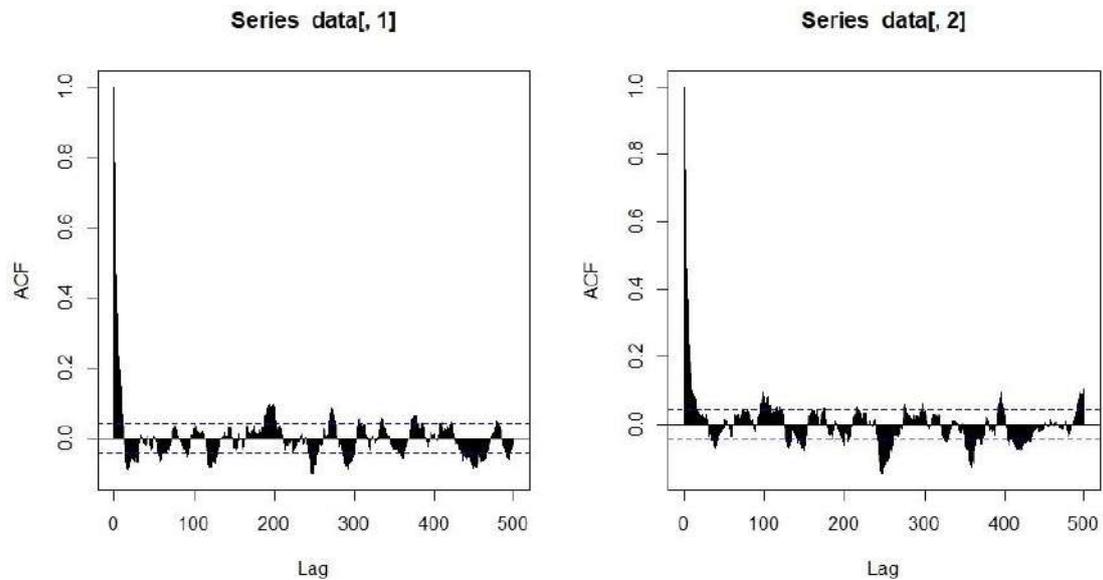


Figure 4: Autocorrelation plot for Variables 1-2 in Case B

Case C: Moderate Autocorrelated data without Outliers

The last Case is of high importance because we do not add atypical values in the data and we do not expect from the ICS to find outliers and have a small proportion of False Alarms. The data have moderate autocorrelation and their distribution is a normal with mean equal to 0 and variance 1. When we collect data from the statistical quality control or other fields where statistics is essential, we do not expect from the data to have outliers or special structures. ICS, PCA, Mahalanobis distance and other multivariate methods have been created for detecting outliers, but it is important to see their performance (FP, TP) when a dataset is not contaminated. Finally, it is important to illustrate the autocorrelation plot for the variables so as to understand the volume of the autocorrelation and make comparisons.

4 Simulations

In section 3 we have already illustrated the cases we worked on and how we created the outliers. We added a small proportion of outliers (3%) so as to depict problems that are met in the industrial applications. The outliers are multivariate and it is impossible to be identified by plotting the variables in a 1-D plot such as Box-plots. The pairwise associations could show some discrepancies in the data but not always and the structure of the data do not allow the scientist to understand how many and in which area the subgroups of outliers exist.

To compare the performance of ICS in all these cases we described above, we provide the percentage of outliers correctly identified (denoted by TP for “True Positive”) and the percentage of non-outlying observations mistakenly identified as outliers (FP for “False Positive”).

4.1 Average Selected Invariant Components

In our example, we added in specific variables (1, 5, and 10) outliers and we expect from the ICS to show only 3 invariant components that are useful for detecting the multivariate outliers. In a real time analysis it is very difficult to identify the real invariant components and for this reason we use the Parallel Analysis and the Marginal Normality tests to identify them. Before we continue with the performance of ICS in terms of TP and FP, we observe the selected dimensions using the Shapiro-Wilk normality test (SW) and the Parallel Analysis (PA) method for a level $\alpha=5\%$. In Table 1, it can be seen the average of these dimensions over 200 simulations for the different cases (A, B, C). It is important here to be stressed that we have created the same data with the same outliers but without autocorrelation for every case and they are mentioned as AA, BB and CC respectively.

Scatters	d	Case A		Case AA		Case B		Case BB		Case C		Case CC	
		SW	PA	SW	PA	SW	PA	SW	PA	SW	PA	SW	PA
MCD-COV ₄	10	4.91	3.82	3.13	3.90	4.10	3.86	3.12	3.77	2.25	1.78	1.16	1.00
MCD-COV ₄	20	7.07	6.44	3.50	4.77	5.52	5.82	3.51	4.73	4.12	3.55	1.95	1.02
MCD-COV ₄	35	10.40	11.39	4.68	6.40	7.70	9.30	4.65	5.90	6.71	7.21	3.40	1.02
COV-COV ₄	10	5.35	4.01	3.10	5.55	4.05	4.24	3.10	5.40	2.21	1.54	1.13	1.01
COV-COV ₄	20	6.98	5.99	3.32	7.98	5.24	6.25	3.53	7.03	4.00	2.89	1.82	1.00
COV-COV ₄	35	10.25	9.63	4.30	11.02	7.37	9.61	4.32	8.71	6.47	5.94	3.07	1.01
MLC-COV ₄	10	5.10	4.19	3.13	5.47	4.02	4.52	3.10	5.22	2.19	1.64	1.14	1.01
MLC-COV ₄	20	6.71	6.93	3.43	7.86	5.22	7.06	3.50	6.51	3.80	3.66	1.65	1.00
MLC-COV ₄	35	10.20	11.34	4.33	9.91	7.58	11.02	4.29	8.02	6.56	7.58	2.69	1.01
MCD-MLC	10	4.71	3.16	2.96	3.02	3.74	2.92	3.04	2.85	1.20	1.76	1.00	1.00
MCD-MLC	20	6.02	6.80	2.90	3.12	4.28	4.73	3.08	3.10	1.59	4.28	1.02	1.02
MCD-MLC	35	10.33	13.54	3.01	3.54	6.81	9.96	3.55	3.74	2.47	9.62	1.53	1.01

Table 1: Averaged numbers of selected invariant components for the PA and SW methods

When we have highly autocorrelated data the two methods choose more invariant components compared with the moderately autocorrelated data. The pair MCD-MLC chooses on average the smallest amount of components in Cases A, B and C and it over-chooses components in Case CC while the other pairs of scatter statistics choose on average only one in every dimension. The Shapiro-Wilk marginal normality test retains on average the same number of invariants with the parallel analysis when we deal with non-autocorrelated data, but retains fewer components compared with the Parallel analysis when there is not correlation among the data. Generally, the selected components are much more than those we need to analyse the multivariate outliers and as a result a kind of noise will be created but it is worse in our case to under-choose invariants. When we do not retain the proper amount of components we have lower level of TP (True Positives) because we leave out the informative ones and the D’Agostino test has this drawback when we analyse correlated data in Statistical Quality Control.

All the above comments are about the selection of the invariants but we will be able to see if the ICS works when we take a look at Tables 2-4 which illustrate the averaged TP and FP for all Cases in every dimension. The results for the cases AA, BB and CC are shown in the Appendix as extra material. Specifically, we calculated the standard error and the median for those amounts so as to show a right representation of those amounts. The mean is easily influenced by outliers and does not show the proper results.

Using the true invariant components help us detect the multivariate outliers and reach high percentage of TP and low in FP amounts. The ICS works in highly correlated data and specifically one can achieve 90% of TP when he deals with up to 20 variables. While we increase the dimension of the problem, we get lower performance but not in unnoticeable levels. When we choose only the true invariant components we get the best results for the FP amount and more specifically with the second pair of scatter statistics, Cov-Cov₄. The fact that we deal with correlated data does not mean that the true invariants will be always the first three and one can choose more than three to detect anomalies in the data. The SW method and PA do exactly this and eliminate the SD of the amounts TP and FP. It is important here to stress that when we use different kind of marginal normality tests, we get different results. The D’Agostino test failed many times to reveal the outliers and created a large variance in the estimation and so, we chose the Shapiro-Wilk normality test. All pairs except for the Cov-Cov₄ have high percentage of TP but they high values of FP would create problems in a Statistical Process. The pair MCD-Cov₄ has very satisfactory values for the amount TP but high values for the FP. Generally, there is not the perfect pair and there is a balance among the drawbacks and advantages of the pairs. In Section 5 where comparisons are conducted we illustrate the range of the values and choose some pairs in order to make comparisons with

other methods. The Parallel analysis results are almost the same with those which gained from the SW test. Small discrepancies in the SD is the only noticeable fact and the first pair MCD-Cov₄ would be used to detect the multivariate outliers.

To sum up, the PA and SW methods give better results even when we select the first three components that contain the biggest amount of information. The fact that the data are correlated does not allow the scientist to analyse only the first three invariants like we would do in data without autocorrelation (Nordhausen et. al.[3]). We continue with Case B and do exactly the same experiments so as to compare the performance of the ICS. For sure, we expect the amounts TP and FP to improve and their SD to be decreased.

<i>Measures in %</i> <i>d</i>	<i>TP (Median)</i>			<i>FP (Median)</i>			<i>TP (SD)</i>			<i>FP (SD)</i>		
	10	20	35	10	20	35	10	20	35	10	20	35
<i>True q MCD-COV₄</i>	98.33	93.33	76.67	0.77	1.70	3.64	7.80	15.11	17.29	0.62	0.86	1.15
<i>True q COV-COV₄</i>	96.67	77.50	68.33	0.15	0.67	1.03	9.60	16.57	17.90	0.31	0.49	0.61
<i>True q MLC-COV₄</i>	96.67	89.17	66.67	0.62	1.06	1.52	10.98	15.78	17.89	0.53	0.67	0.70
<i>True q MCD-MLC</i>	96.67	91.67	70.00	1.67	3.81	7.68	10.07	15.77	18.34	1.03	1.18	1.44
<i>SW MCD-COV₄</i>	97.50	93.33	86.66	1.11	2.73	5.41	6.15	12.17	13.00	0.62	0.86	1.25
<i>SW COV-COV₄</i>	90.00	81.66	68.33	0.62	1.39	2.42	9.51	13.34	14.27	0.35	0.46	0.60
<i>SW MLC-COV₄</i>	95.00	88.33	74.16	0.93	1.85	3.04	7.84	13.56	16.03	0.52	0.66	0.69
<i>SW MCD-MLC</i>	96.66	95.00	91.66	1.65	4.23	9.43	8.05	10.67	12.23	0.86	1.24	1.54
<i>PA MCD-COV₄</i>	98.33	93.33	86.66	1.16	2.83	5.57	7.92	11.50	12.71	0.70	0.86	1.07
<i>PA COV-COV₄</i>	95.00	84.17	70.00	0.59	1.34	2.45	7.53	12.11	13.27	0.45	0.54	0.58
<i>PA MLC-COV₄</i>	96.65	88.33	76.66	0.98	2.01	3.25	7.39	12.77	15.78	0.59	0.63	0.70
<i>PA MCD-MLC</i>	96.65	95.00	91.66	1.75	4.48	9.69	10.32	10.05	13.67	0.95	1.23	1.61

Table 2: Results of True Positive (TP) and False Positive (FP) amounts in Case

4.2 Averaged Results of TP and FP for Case B

As we mentioned above, in this point of the analysis, we illustrate the results of the amounts TP and FP for Case B and make comparisons between Case A and B. additionally, we try to find the most robust pair of scatter statistics for Normal distributed data with autocorrelation. Nordhausen et al.[3] suggests the pair Cov-Cov₄ for revealing outliers that exist in elliptical distributions like the Normal distribution. In Table 3 we illustrate the averaged results of True Positives and False Positives for Case B. it is obvious that when we generate moderate correlated data, the ICS gets a much better performance than in Case A where the SD of the estimations were almost double in size compared with Case B. the best performance comes when we choose to analyse the first three invariant components instead of the methods SW and PA. The SD of the estimations is lower compared with the other methods and in every pair we get results for TP over 90% and for FP values very close to 0. If we do not know how many invariants we have and the dataset is needed to be analysed in real time, then, the Parallel Analysis method and SW give robust results. The SD of the estimations is very low in both methods and one can detect small differences that will also be seen in Section 5 where comparisons are conducted by using Box-plots. We would say that the best pair is the MCD-Cov₄ but it cannot be doubted that the rest pairs are very capable. In both Cases, the results which are taken by PA method and SW show that Monte Carlo simulations give more robust results but they are time-consuming procedures. It is essential here to stress, that the pairs 1 & 4 needed a lot of time to give results compared with the other two because of the algorithmic procedures that are needed in order to get the proper sample and find the smallest determinant. We continue with Case C and some general comments about the results and we then continue with the performance of Mahalanobis distance, Robust MD, T² chart plot and the robust T² chart plot for detecting atypical values.

<i>Measures in %</i> <i>d</i>	<i>TP (Median)</i>			<i>FP (Median)</i>			<i>TP (SD)</i>			<i>FP (SD)</i>		
	10	20	35	10	20	35	10	20	35	10	20	35
<i>True q MCD-COV₄</i>	98.33	98.33	96.67	0.36	0.26	0.36	2.27	3.38	8.92	0.26	0.27	0.37
<i>True q COV-COV₄</i>	96.66	95.00	91.67	0.05	0.05	0.10	3.38	4.63	10.17	0.11	0.12	0.24
<i>True q MLC-COV₄</i>	98.33	96.67	93.33	0.26	0.21	0.21	2.06	4.70	9.53	0.22	0.19	0.24
<i>True q MCD-MLC</i>	98.35	98.35	95.00	0.82	0.95	1.03	2.77	4.57	10.57	0.37	0.49	0.49
<i>SW MCD-COV₄</i>	96.66	96.66	90.00	0.67	0.98	1.44	2.70	3.77	8.00	0.36	0.45	0.41
<i>SW COV-COV₄</i>	95.00	90.00	81.66	0.31	0.57	1.03	3.95	6.00	8.97	0.26	0.34	0.40
<i>SW MLC-COV₄</i>	96.65	93.33	86.65	0.57	0.82	1.24	2.86	4.47	7.46	0.37	0.41	0.44
<i>SW MCD-MLC</i>	98.35	98.33	93.33	0.88	1.31	2.11	2.76	3.50	7.69	0.46	0.55	0.54
<i>PA MCD-COV₄</i>	98.35	95.00	88.33	0.62	1.06	1.49	2.35	4.32	6.31	0.40	0.42	0.42
<i>PA COV-COV₄</i>	93.33	88.33	78.33	0.31	0.67	1.13	3.85	5.51	8.10	0.27	0.32	0.31
<i>PA MLC-COV₄</i>	96.66	91.66	81.66	0.62	1.03	1.55	2.85	4.98	7.51	0.38	0.39	0.43
<i>PA MCD-MLC</i>	98.33	98.33	90.00	0.82	1.29	2.32	5.85	3.50	6.48	0.40	0.55	0.53

Table 3: Results of True Positive (TP) and False Positive (FP) amounts for Case B

4.3 Averaged Results of FP for Case C

In case C, we created data with moderate volume of autocorrelation and we did not add outliers in any of the existed variables. This kind of data will show the FP proportion which is important for the company when we analyse data in real time. A small amount of FP would be ideal since we want to categorize new observations as defective or not. For the invariant selection we use the parallel analysis (PA) and the marginal normality test Shapiro-Wilk (SW) so as to evaluate the method. We expect a high performance since the volume of correlation is not so high but it is still confusing to detect outliers. In Table 4 we illustrate the results only for the FP since we do not have outliers to detect and the amount TP cannot be calculated anymore. Both methods have the same performance and keep the proportion FP in low levels. The pair of scatter statistics that has the best performance is the second in both methods. In a problem with low dimension we would choose the 4th pair which has the best performance but loses its properties in higher dimension. Finally, the PA method gives a larger standard error compared with the Marginal normality tests and only in higher dimensions the SD becomes smaller than the SW estimations. Before we continue with comparisons among the methods, it is important to point out that the performance of ICS on data without autocorrelation and without outliers is almost perfect and the FP amount is very close to zero.

<i>Measures in %</i>	<i>FP (Median)</i>			<i>FP (SD)</i>			
	<i>d</i>	<i>10</i>	<i>20</i>	<i>35</i>	<i>10</i>	<i>20</i>	<i>35</i>
<i>SW MCD-COV₄</i>		1.57	2.05	2.45	0.41	0.38	0.40
<i>SW COV-COV₄</i>		1.40	1.80	2.20	0.36	0.30	0.37
<i>SW MLC-COV₄</i>		1.45	1.95	2.35	0.42	0.47	0.40
<i>SW MCD-MLC</i>		0.90	1.70	2.50	0.87	1.00	1.14
<i>PA MCD-COV₄</i>		1.55	2.05	2.50	0.80	0.57	0.39
<i>PA COV-COV₄</i>		1.25	1.65	2.20	0.75	0.75	0.57
<i>PA MLC-COV₄</i>		1.40	1.90	2.45	0.80	0.71	0.46
<i>PA MCD-MLC</i>		1.00	2.15	3.15	0.78	0.74	0.56

Table 4: Results for False Positive (FP) amount in Case C

5 Comparing ICS with Mahalanobis distance and Hotelling control chart

In this section, we conduct some comparisons so as to evaluate other methods that exist for detecting multivariate outliers. Mahalanobis distance introduced by P. C. Mahalanobis in 1936 and measures the distance among the multivariate observations by using a covariance matrix. The sample covariance matrix is a common one for detecting multivariate outliers but it is not always effective like the robust ones. For our scope, we use robust MD and non-robust MD to monitor outliers and in Table 5 we illustrate the results for the Case A. also, a common chart for detecting outliers, the T² Hotelling test, is used so as to show the performance of methods that created only for data coming from quality control. The T² chart plot is trained in phase I with a dataset without outliers so as to create the cut-off value. We trained it for every case with the proper multivariate data and in Phase I we removed every point that was denoted as outlier. The performance was pretty good but not better than the ICS or MD.

5.1 Comparisons for Case A

To better organize the article, we separate the results for every case in order to give a full representation of the ICS performance compared with other methods. For the comparisons we keep only the results from the MCD-Cov₄ and MLC-Cov₄ pairs as we explained in Section 4.1. The chosen scatter statistic for the robust-MD is a highly robust estimator, Constrained M-estimates and for the robust Hotelling chart plot the S-estimator with breakdown point 50%. We trained the control chart with 4000 obs. And we removed the values that over passed the cut-off value. We then created 2000 obs. with 3% outliers which are labelled. Moreover, the limits kept untouched without creating new ones.

The first comments has to do with the performance of Mahalanobis Distance (MD) which did not manage to reveal a high percent of the outliers. When we increased the dimension of the problem the TP amount went dramatically down while the FP stayed close to zero. When a robust covariance matrix was used, the percent stayed high in all dimensions and had almost the same performance with the ICS. For sure, ICS has the best performance from all these methods and when we deal with only 10 variables the percentage of TP can be almost 99% while the rest can reach up to 93.5%. The chart plot does not reveal the atypical values like the ICS and RMD, but it is essential to stress that the SD of the estimations is much lower for the two amounts compared with the other methodologies. The standard error plays a crucial role and many scientists would use the chart plot instead of the other two methods, but sometimes we do not know which of the data are in control to train the algorithm and for this reason we choose methods like MD or ICS to monitor atypical values and the structure of the dataset. In Figures 5-6 are illustrated some results taken from Table 5 in order to show the range of the values.

Specifically, we use the results taken from the first (MCD-Cov₄) and third (MLC-Cov₄) pairs of scatter statistics in ICS, illustrating them with box-plots so as to show the differences there are among these methods. The red bullet shows the mean for every case and the different colour of the boxes the dimension of the problem. All methods have a good performance of revealing the outliers, but the ICS is by far the best method when we deal with up to 20 variables and focus on TP amount. In higher dimension the amount TP gets lower but it is still higher than the values of Hotelling chart. The more the range of the values of FP is closer to one (1) the better the method is, and in a problem with 10 variables, the ICS would

give almost excellent results and detect all of the outliers. The RMD does not give very often false alarms (FP) and it would be used as well for low dimension problems.

Measures in % <i>d</i>	TP (Median)			FP (Median)			TP (SD)			FP (SD)		
	10	20	35	10	20	35	10	20	35	10	20	35
ICS SW MCD-Cov ₄	97.50	93.33	86.66	1.11	2.73	5.41	6.15	12.17	13.00	0.62	0.86	1.25
ICS SW MLC-Cov ₄	95.00	88.33	74.16	0.93	1.85	3.04	7.84	13.56	16.03	0.52	0.66	0.69
ICS PA MCD-Cov ₄	98.33	93.33	86.66	1.16	2.83	5.57	7.92	11.50	12.71	0.70	0.86	1.07
ICS PA MLC-Cov ₄	96.65	88.33	76.66	0.98	2.01	3.25	7.39	12.77	15.78	0.59	0.63	0.70
Mahalanobis Dist.	73.30	43.35	21.00	1.16	0.36	0.26	9.40	11.05	8.80	0.28	0.22	0.18
Robust MD	91.65	81.65	71.65	0.98	1.50	2.83	7.98	12.50	17.07	0.53	0.71	1.10
T ² Hotelling chart	91.65	78.35	61.65	1.29	1.85	2.22	4.73	8.05	9.34	0.47	0.57	0.64
Robust T ² chart	93.35	81.65	68.35	1.44	1.96	2.50	4.09	6.63	10.00	0.52	0.58	0.69

Table 5: Results for True Positive (TP) and False Positive (FP) amounts in Case A

The Figure 6 shows the results for the FP amount which is not hazardous if it is high, but it is better to be kept in lower levels. When the method detects an outlier that it is not actually an outlier, we call it false alarm (FP) and in the statistical quality control is an important amount for a procedure that it is needed to be conducted fast without delays. Also, using the first pair in ICS is inefficient and cannot keep the FP amount in low levels, while the other three methods kept it close to zero and the RMD having the best performance even when we increase the number of variables. In this point, someone would say that the RMD could be used for statistical quality control data, but in case C where we have no outliers to detect, the methods shows that exist and has high values of FP. Also, when we remove the autocorrelation, the RMD has still the same performance while the ICS keeps the FP almost zero. For sure, we discuss further the problems in subsection 5.3. By using ICS and not RMD, the scientist can focus in the amounts TP or FP when he/she deals with correlated data and he/she will never lose the performance when the data lose the correlation, but instead, the results will be improved by this change of autocorrelation.

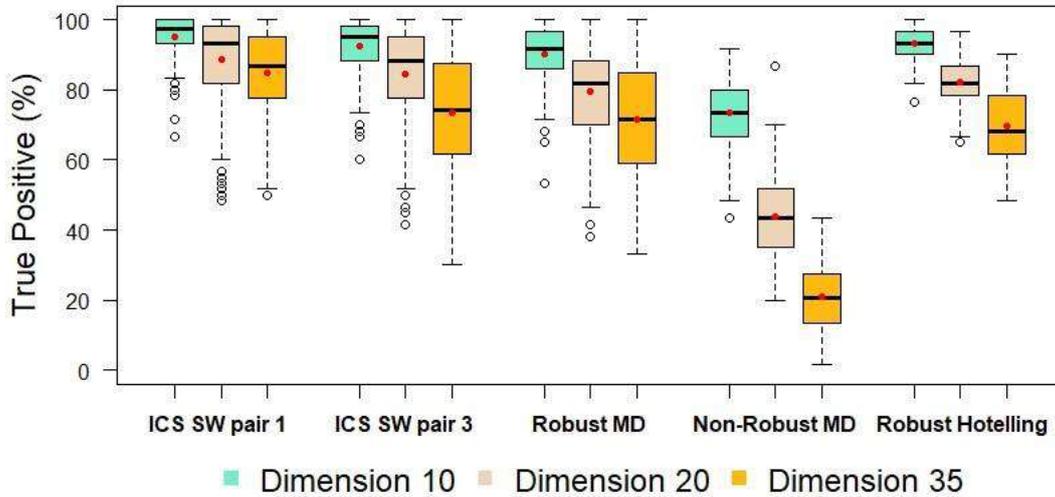


Figure 5: Box plots of the True Positive (TP) amount of Case A for comparing the methods ICS, Robust Mahalanobis distance and Hotelling chart in three scenarios. The red bullet shows the mean of the values.

5.2 Comparisons for Case B

We continue the comparisons for Case B where the amounts TP and FP have improved since the volume of autocorrelation is lower. We kept the results for the pair MCD-Cov₄ to illustrate and compare with the other methods. In Table 6 we can monitor the differences among the three methods and make some inferences: The Table shows that the Parallel Analysis and Shapiro tests have lower SD for the estimated amount TP compared with the Mahalanobis distance and Robust MD. The T² Hotelling chart plot gives satisfactory results for the TP amount when we have to analyse up to 20 variables and keeps the FP relatively low and close to zero. We can notice that when the autocorrelation among the data is moderate and

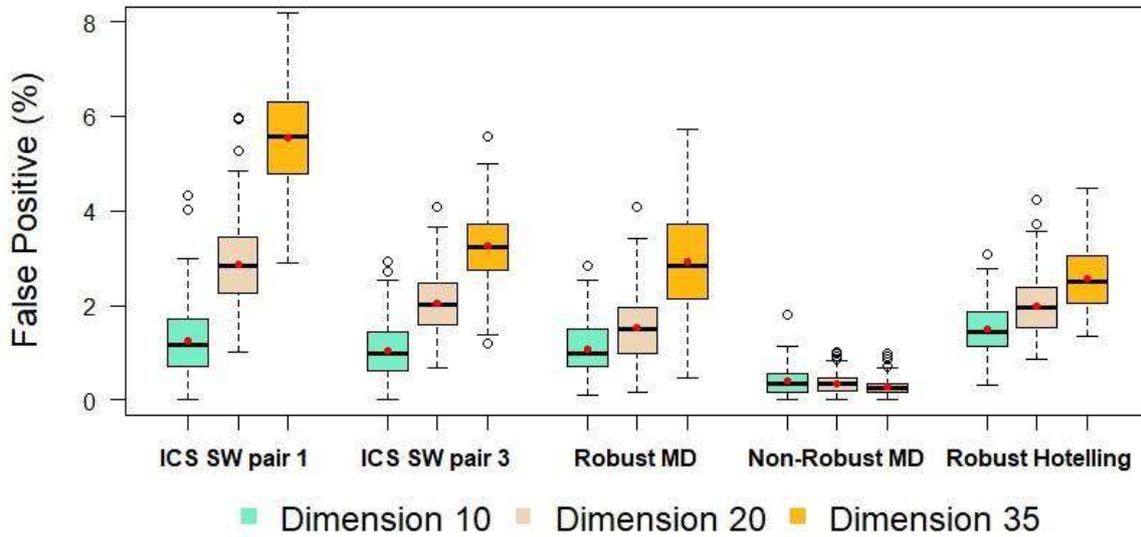


Figure 6: Box plots of the True Positive (TP) amount of Case A for comparing the methods ICS, Robust Mahalanobis distance and Hotelling chart in three scenarios. The red bullet shows the mean of the values.

Measures in % <i>d</i>	TP (Median)			FP (Median)			TP (SD)			FP (SD)		
	10	20	35	10	20	35	10	20	35	10	20	35
ICS SW MCD-Cov ₄	96.65	96.65	90.00	0.67	0.98	1.44	2.70	3.77	8.00	0.36	0.45	0.41
ICS SW MLC-Cov ₄	98.35	98.35	93.35	0.88	1.31	2.11	2.76	3.50	7.69	0.46	0.55	0.54
ICS PA MCD-Cov ₄	98.35	95.00	88.35	0.62	1.06	1.49	2.35	4.32	6.31	0.40	0.42	0.42
ICS PA MLC-Cov ₄	98.35	98.35	90.00	0.82	1.29	2.32	5.85	3.50	6.48	0.40	0.55	0.53
Mahalanobis Dist.	83.35	63.35	40.85	0.46	0.46	0.46	5.41	7.48	7.80	0.22	0.22	0.21
Robust MD	95.00	88.35	78.30	0.98	1.13	1.39	3.68	5.72	9.03	0.36	0.38	0.47
T ² Hotelling chart	95.00	88.35	78.35	1.24	1.85	2.16	3.15	4.88	7.04	0.46	0.55	0.54
Robust T ² chart	96.65	90.00	83.35	1.55	1.65	2.50	3.15	4.88	6.75	0.46	0.55	0.66

Table 6: Results of True Positive (TP) and False Positive (FP) amounts in Case B

not intense like in case A, the ICS shows an improvement in the FP amount and has almost the same capability with the RMD. The use of highly robust estimators gives the scientist the chance to reveal the outliers will all the methods and they work better with ICS and RMD. In Figures 7 & 8 one can see that the methods have excellent performance when the number of the variables do not overpass the 10 and gets lower while we increase the dimension of the problem. Finally, the first pair with MCD-Cov₄ in ICS method showed that is able in both cases to reveal the real outliers. One can notice that the illustration of the results is based on the median of the 200 values taken for the estimations FP and TP since the mean gives a spurious result. In this point of the analysis, we illustrate the results of Case C and make inferences based on them.

5.3 Comparisons for Case C

In this Case the reader has the chance to understand the importance of low values of FP amount in a procedure. There are no outliers but the methods show that an amount of observations is spuriously categorised as outliers and hopefully, this amount does not overpass the 3% of the dataset. The estimation of FP amount increases when the dimension of the problem rises too. One could consider that the Non-Robust Mahalanobis Distance has a good efficiency and the FP amount is very close to zero as we wished, but when we have a glance back to Figure 7, we are able to see the unsatisfactory performance of Non-Robust MD and conclude that there is no stability in this method. It categorises the majority of the data under the cut-off value and for this reason we get low values in both amounts. Additionally, the pair MCD-Cov₄ gives in all cases respectable results and it would be suggested for a real time analysis, where high and moderate correlated data may are mixed. The advantage of using ICS is the high rate of TP values, where it is the most important amount that has to be taken into consideration if we consider that high values of FP do not create problem to the company. We indeed stop the quality procedure to check the false alarms, but skipping a faulty product/service would lessen the quality and the standards of the company.

Measures in %	FP (Median)	FP (SD)
---------------	-------------	---------

d	10	20	35	10	20	35
ICS SW MCD-COV ₄	1.57	2.05	2.45	0.41	0.38	0.40
ICS SW COV-COV ₄	1.40	1.80	2.20	0.36	0.30	0.37
ICS PA MCD-COV ₄	1.55	2.05	2.50	0.80	0.57	0.39
ICS PA COV-COV ₄	1.25	1.65	2.20	0.75	0.75	0.57
Mahalanobis Distance	0.80	0.70	0.60	0.29	0.26	0.22
Robust MD	1.00	1.20	1.65	0.42	0.43	0.46
T ² Hotelling chart	1.20	1.85	2.27	0.45	0.58	0.57
Robust T ² chart	1.52	1.57	2.55	0.51	0.52	0.59

Table 7: Results for False Positive (FP) in Case C

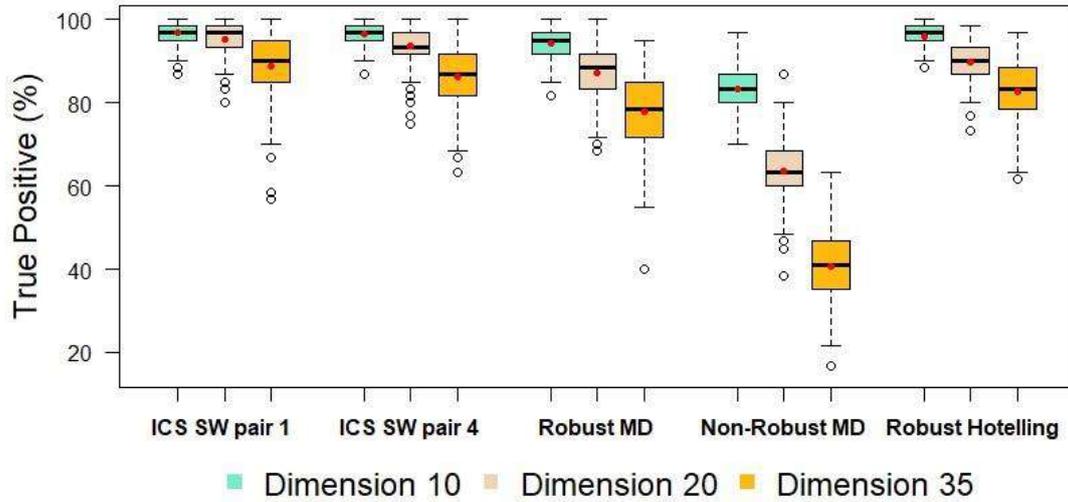


Figure 7: Box plots of the True Positive (TP) amount of Case B for comparing the methods ICS, Robust Mahalanobis distance and Hotelling chart in three scenarios. The red bullet shows the mean of the values.

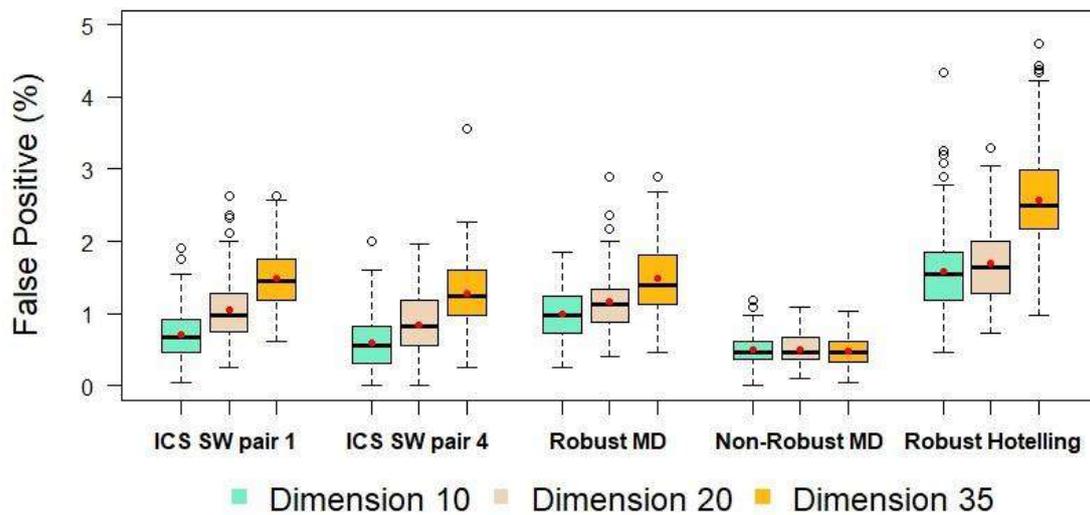


Figure 8: Box plots of the False Positive (FP) amount of Case B for comparing the methods ICS, Robust Mahalanobis distance and Hotelling chart in three scenarios. The red bullet shows the mean of the values.

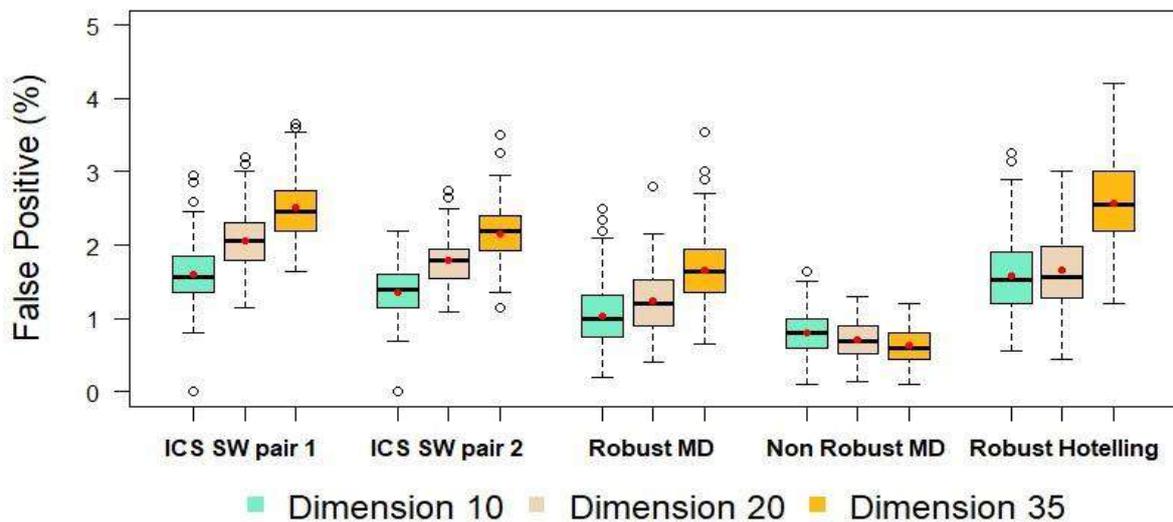


Figure 9: Box plots of the False Positive (FP) amount of Case C for comparing the methods ICS, Robust Mahalanobis distance and Hotelling chart in three scenarios. The red bullet shows the mean of the values.

6 Conclusions

All this analysis on ICS was conducted in order to illustrate a new method for dimension reduction and multivariate outlier detection. Illustration of high dimension data is impossible and for this reason, statisticians shift their interest on multivariate methods which are able to identify atypical observations and specify their indexes. In quality control, the quality of a product depends on many factors and the ICS is able to give as an output the defective products by inserting two proper scatter statistics.

In Section 2 we give more information for these scatter statistics and the ICS methodology. These ones have to be robust in order to give us valuable information for the data and we see their differences in Section 4 where simulations have been conducted. When we have to deal with contaminated datasets, robust statistics are needed. We list them in three classes, Class I, Class II and Class III and their robustness is defined as nonrobust, moderate robust and highly robust respectively. Its strength depends on the number of the atypical values that are able to tolerate. The maximum breakdown point is 50% and they target only in the inner data. In my experience, I suggest not choosing two high robust statistics for the ICS since in many cases where elliptical data are analysed, fail to reveal the structure. The scientists have to use several scatter statistics and their location before they continue to conclusions and further analysis.

In the field of statistical quality control, the data have some anomalies but it is difficult to be detected because of the autocorrelation that exists among the observations. ICS does not assume any statistical model or hypothesis and it would be very useful seeing how it works with special structures. Industry data are correlated in most cases and over the years, different kind of chart plots have been created in order to deal with multidimensional outliers. For sure, smart software for detecting an error helps the companies to control the quality of their products and monitor in which machinery the error happened, but it is still very difficult for those to recognize small changes in the production. Autocorrelated data make these procedures much more difficult and in many cases we get false alarms and the procedure is delayed since the production manager investigate the cause of the alarm. ICS does not get any information from the previous observations such as chart plots and this makes it more robust. In Section 3 there are simulation studies where one can see three different scenarios of autocorrelated data and compare which combination of scatter statistics fits better on the data. Also, it is described the methodology that we used to get the invariant components and in some cases the Shapiro Wilk and Parallel analysis methods showed encouraging results and one could monitor the atypical values with an automated way.

The large number of combinations among the scatter statistics would help the data scientist monitor the best combination for his/her data set and experiment with future observations as we can see in Section 4. One can see the results of the TP and FP amounts for the ICS, MD, RMD and Hotelling chart plot, and finally have a clue about the best pair of scatter statistics. High values of TP amount is essential and ICS had good performance in almost every case while other alternatives did not have a stable efficiency when the autocorrelation is shifted. In Sections 4-5, it is explained that in a quality process the product that is faulty has to be detected and for this reason we should focus on the TP amount and not on FP. For sure, it is essential to eliminate the FP amount but if a company lose the quality of its services/products will lead to a hazardous economic situation.

To sum up, the ICS method can be implemented on statistical quality control data with correlation and show the majority of the faulty products/services. The use of different pairs of scatter statistics has to be done and take into account the time which is needed to implement highly robust statistics. There are scatter statistics that are fast calculated and show the majority of the outliers, but highly robust estimators can reach over 90% of efficiency. The last comments have to do with some advice for the reader that may become valuable when it comes to analyse data with the ICS or other related methods. It is important for the analyst to know the nature of his data before he starts analysing and inferring the results of the analysis. During that journey of detecting outliers, I faced difficult datasets where the ICS was not able to help me understand them. The reason was simple; there is no only one statistical method that cures everything on the statistical

field. Some ordinal variables, small range measurements and discrete variables were the factors that ICS could not deal with. Continuous datasets are the ideal for this method and for sure datasets that contain outliers. Every set of data has to be analysed for a reason and not just to be analysed. For example, the Iris data are always used for scientists for classification, as we also did, and in the real world we have to deal with several kind of data and then we will try to choose from our statistical basket the proper method that may treat the problem. There is no fixed method but I can make it sure that ICS would help many analysts to identify problems in their companies.

References

1. Nordhausen K., Sirkia S., Oja H., Tyler D. E. (2007). *R Package*.
2. Maronna RA, Martin RD, Yohai VJ (2006). *Robust Statistics Theory and Methods*. John Wiley & Sons, Chichester, UK.
3. Nordhausen K, Oja H, Tyler DE (2008). "Tools for Exploring Multivariate Data: The Package ICS". *Journal of Statistical Software*, 28(6).
4. Rousseeuw PJ, Van Driessen K (1999). "A Fast Algorithm for the Minimum Covariance Determinant Estimator". *Technometrics*, 41,212-223.
5. McLachlan GJ (1999). "Mahalanobis Distance". *Resonance* 4(6)
6. Tyler David E, Critchley Frank, Dümbgen Lutz and Hannu Oja (2009). "Invariant Coordinate Selection". *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, Vol. 71, No. 3, pages 549-592
7. Edgar Acuña and Caroline Rodriguez (2004). "On detection of outliers and their effect in supervised Classification".
8. John T. Kent and David E. Tyler (1996). "Constrained Mestimation for Multivariate Location and Scatter". *The Annals of Statistics*, Vol. 24, No. 3, pages 1346-1370.
9. Hannu Oja, Seija Sirkia and Jan Eriksson (2006). "Scatter Matrices and Independent Component Analysis". *Austrian Journal of Statistics*, Vol. 35, No. 2&3, pages 175-189.
10. Peter J. Rousseeuw, Lopusuää and N. A. Campbell (1998). "On the Calculation of a Robust Sestimator of a Covariance Matrix". *Statistics in Medicine*, Vol. 17, pages 2685-2695.
11. Pauliina I., Oja H., Serfling R. (2012). "On Invariant Coordinate System (ICS) Functionals". *International Statistical Review*, Vol. 80, pages 93-110.
12. Stefan Van Aelst, Rousseeuw P. (2009). "Minimum Volume Ellipsoid". *WIREs Computational Statistics*.
13. Mia Hubert, Michiel Debruyne, Peter J. Rousseeuw (2017). "Minimum Covariance Determinant and Extensions". *Wiley Interdisciplinary Reviews: Computational Statistics*.
14. Aurore Archimbaud, Klaus Nordhausen and Anne RuizGazen (2018). "ICSOutlier: Unsupervised Outlier Detection for Low Dimensional Contamination Structure". *The R journal*, Vol. 10/1.
15. D. Fischer, A. Berro, K. Nordhausen and A. RuizGazen (2016a). "REPLab: An R package for detecting clusters and outliers using explanatory projection pursuit".
16. Aurore Archimbaud, Klaus Nordhausen and A. RuizGazen (2016). "ICSOutlier: Outlier detection using Invariant Coordinate Selection". *R package*.
17. Aurore Archimbaud, Klaus Nordhausen and A. RuizGazen (2018). "ICS for multivariate outlier detection with application to quality control". *Computational Statistics & Data Analysis*, 128:184199, 2018, pages 235-244.
18. P. Rousseeuw and M. Hubert (2013). "High Breakdown estimators of multivariate location and scatter". *Robustness and Complex data structures*, pages 4966, Springer Verlag 2013.
19. P. Rousseeuw, C.Croux, V. Todorov, A. Ruckstuhl, M. SalibianBarrera, T. Verdeke, M.Koller and Martin Machler (2017). "robustbase: Basic Robust Statistics". *R package*.
20. Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, Werner A. Stahel (1986). "Robust Statistics. The Approach Based on Influence Functions". *Wiley Series*, New York, 1986.
21. Dümbgen L. (1998). "On Tyler's Mfunctional of scatter in high dimension". *Annals of Institute of Statistical Mathematics*, 50, pages 471-491.
22. Lopuhaä H. P. (1989). "On the relation between Sestimators and Mestimators of multivariate location and scatter". *Annals of Statistics*, 17, pages 1662-1683.
23. Lopuhaä H. P. (1991). "Multivariate τ estimators of location and scatter". *Canadian Journal of Statistics*, 19, pages 310-31.
24. Mardia K. V. (1970). "Measures of Multivariate skewness and kurtosis with applications". *Biometrika*, 57, pages 519-530.
25. Nordhausen K., Oja H., Ollila E. (2008a). "Robust Independent Component Analysis based on two scatter matrices". *Austrian Journal of Statistics*, 37, pages 91-100.
26. Nordhausen K., Sirkia S., Oja H., Tyler D. E. (2007). "ICSNP: Tools for Multivariate Nonparametrics". *R package*.
27. Kai Yang and Walton M. Hancock (1990). "Statistical Quality Control for Correlated Samples".
28. Berna Yazici & Senay Yolacan (2007). "A Comparison of various tests of normality". *Journal of Statistical computation and simulation*, 77:2, pages 175-183.
29. Douglas G. Bonett & Edith Seier (2002). "A test of normality with high uniform power". *Computational Statistic & Data Analysis*, Volume 40, issue 3, pages 435-445.
30. Valentin Todorov (2018), "Scalable Robust Estimators with High Breakdown Point". *rrcov, R-package*.
31. Alan Genz, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, Bjoern Bornkamp, Martin Maechler & Torsten Hothorn (2020). *mvtnorm. R-package*.
32. Lukasz Komsta & Frederick Novomestky (2015). *moments, R-package*.

THE LEE-CARTER MODEL: A FUZZY-RANDOM VERSION

Laura González-Vila Puchades¹ and Jorge de Andrés-Sánchez²

¹ Department of Mathematics for Economics, Finance and Actuarial Science. University of Barcelona
(E-mail: lgonzalezv@ub.edu)

² Social and Business Research Laboratory. Business Management Department. Rovira i Virgili University
(E-mail: jorge.deandres@urv.cat)

Abstract and Keywords

Abstract: The Lee-Carter model is a useful dynamic stochastic model to represent the evolution of central mortality rates throughout time. This model only considers the uncertainty of the coefficient related to the mortality trend over time but not to the age-dependent coefficients. This work proposes a fuzzy-random extension of the Lee-Carter model that allows quantifying the uncertainty of both kinds of parameters. As it is common in actuarial literature, the variability of the time-dependent index is modelled as an ARIMA time series but the uncertainty of the age-dependent coefficients is quantified by using triangular fuzzy numbers. Once the fuzzy-random extension has been developed, life expectancies are obtained by using fuzzy numbers arithmetic. Finally, we make a comparative assessment of our method with the basic Lee-Carter model.

Keywords: Lee-Carter model, Fuzzy numbers, Fuzzy regression, Fuzzy-random modelling.

Fuzzy-random extension of the Lee-Carter model

Lee and Carter (1992) proposed modelling the logarithm of the central death rate for each specific age and each year with a linear function. In such a way, if $m_{x,t}$ is the central death rate of a person aged x in the calendar year t , the Lee-Carter (LC) model considers:

$$\ln(m_{x,t}) = a_x + b_x k_t + \varepsilon_{x,t} \text{ or } m_{x,t} = \exp(a_x + b_x k_t + \varepsilon_{x,t})$$

where $\exp(a_x)$ is the specific value of the central mortality rate at age x regardless of the time calendar t , b_x quantifies the sensitivity of the central death logarithm rate for age x in year t respect to changes in k_t and k_t is a specific mortality index for each year t that represents the trend of the mortality across time. Further, $\varepsilon_{x,t}$ is a random error term, with mean 0 and standard deviation σ_ε , which reflects particular age-specific historical influences not captured by the model. By adding the conditions $\sum_x b_x = 1$, $\sum_t k_t = 0$ and by applying singular value decomposition the authors obtained:

$$a_x^* = \frac{\sum_{t=0}^T \ln(m_{x,t})}{T+1} \quad k_t^* = \sum_x \ln(m_{x,t}) - \sum_x a_x^* \quad b_x^* = \frac{\sum_t (\ln(m_{x,t}) - a_x^*) k_t^*}{\sum_t (k_t^*)^2}$$

Once the parameters of the model have been fitted, for $t = 0, 1, \dots, T$ and $x = 0, 1, \dots, \omega$, the trend of the mortality across time, k_t , for $t = T + 1, \dots$ is forecasted with an ARIMA model.

Our fuzzy-random approach of the LC model considers two different sources of uncertainty. Firstly, it is supposed that historical influences of each specific age are due to fuzziness in the model structure. As a consequence a_x and b_x turn into the triangular fuzzy numbers (TFNs) $\tilde{a}_x = (a_x, l_{a_x}, r_{a_x})$ and $\tilde{b}_x = (b_x, l_{b_x}, r_{b_x})$, where a_x and b_x are their centres or cores and $l_{a_x}, l_{b_x}, r_{a_x}, r_{b_x}$ their respective left and right spreads. Secondly, the mortality index k_t follows an ARIMA stochastic process, i.e., k_t is an outcome of the random variable (RV) k_t , $t = T + 1, \dots$

In order to find the estimates $\tilde{a}_x^* = (a_x^*, l_{a_x}^*, r_{a_x}^*)$, $\tilde{b}_x^* = (b_x^*, l_{b_x}^*, r_{b_x}^*)$ and k_t^* , $t \leq T$, the fuzzy regression model of Ishibuchi and Nii, M. (2001) is used. So, following steps are implemented:

Step 1. By taking the centres of $\ln(\tilde{m}_{x,t})$, $t = 0, 1, \dots, T$ and $x = 0, 1, \dots, \omega$, we fit the centres of \tilde{a}_x^* and \tilde{b}_x^* and the outcomes of the RVs k_t , k_t^* , as they are fitted in the LC model. It is necessary to point out that the observed values of the central rate of mortality (and its logarithms) in which we will base our work (the Human Mortality Database) are crisp, i.e. $\ln(\tilde{m}_{x,t}) = \ln(m_{x,t})$.

Step 2. By applying the minimum fuzziness criterion, the spreads of \tilde{a}_x^* and \tilde{b}_x^* are adjusted. Thus, spread estimates must minimize the uncertainty of the estimated outputs and simultaneously these estimated outputs have to contain the real observations, with a membership level of at least α . Considering the relation of inclusion of FNs in Tanaka (1987), the problem to solve for a prefixed level α is:

$$\min_{l_{a_x}, l_{b_x}, r_{a_x}, r_{b_x}} (T+1) \sum_x (l_{a_x} + r_{a_x}) + \sum_t |k_t| \sum_x (l_{b_x} + r_{b_x})$$

subject to:

$$a_x^* + b_x^* k_t^* + r_{a_x} + k_t r_{b_x} \geq \ln(m_{x,t}) \text{ for } k_t > 0, \quad a_x^* + b_x^* k_t^* + r_{a_x} - k_t l_{b_x} \geq \ln(m_{x,t}) \text{ for } k_t \leq 0, \quad a_x^* + b_x^* k_t^* - l_{a_x} - k_t l_{b_x} \leq \ln(m_{x,t}) \text{ for } k_t > 0, \quad a_x^* + b_x^* k_t^* - l_{a_x} + k_t r_{b_x} \leq \ln(m_{x,t}) \text{ for } k_t \leq 0$$

$$b_x^* - l_{b_x} \geq 0 \text{ if } b_x^* \geq 0, \quad b_x^* + r_{b_x} < 0 \text{ if } b_x^* < 0, \quad l_{a_x}, l_{b_x}, r_{a_x}, r_{b_x} \geq 0, \quad x = 0, 1, 2, \dots, \omega.$$

By using \tilde{a}_x^* , \tilde{b}_x^* and k_t^* , it is possible to have a fuzzy estimate for the observed central mortality rates, $\tilde{m}_{x,t}^* = \exp(\tilde{a}_x^* + \tilde{b}_x^* k_t^*)$. Further, by applying the results in Dubois and Prade (1993), $\tilde{m}_{x,t}^*$ can be approximated by a TFN:

$$\tilde{m}_{x,t}^* \approx (m_{x,t}^*, l_{m_{x,t}^*}, r_{m_{x,t}^*}) = (\exp(a_x^* + b_x^* k_t^*), \exp(a_x^* + b_x^* k_t^*) l_{\ln(m_{x,t}^*)}, \exp(a_x^* + b_x^* k_t^*) r_{\ln(m_{x,t}^*)})$$

To forecast future central mortality rates, projecting the values of the index k_t is necessary. In our approach, these values are the outcomes of the RVs k_t for each year $t > T$ that actuarial literature commonly fits by an ARIMA($p, 1, q$) on the data set $\{k_t^*, t \leq T\}$. Finally, from forecasted values of $\tilde{m}_{x,t}^*$, and by using FNs arithmetic, other related variables such as mortality or survival probabilities and life expectancies can also be obtained as FNs.

Empirical assessment of the Fuzzy Random Lee-Carter model

We make a comparative assessment on the prediction capability of our fuzzy-random extension of the LC model (FRLC) with the basic one (BLC). To carry out the analysis, we use central mortality rates collected separately for men and women in eight Western Europe countries from the Human Mortality Database. Model parameters are fitted by using central mortality rates in the period 1970-2000 and models out-of-sample performance is tested during 2001-2012. We evaluate the capability of BLC and FRLC to predict future values of $m_{x,t}$ and $e_{x,t}$ by means of confidence intervals. As the predictor for k_t , $t=2001, 2002, \dots, 2012$, we have considered its confidence interval for the significance level of 10%. Following García and Herrera (2008) and García *et al.* (2010), an adequate non-parametrical test to carry out this kind of analysis is the Friedman rank test (Friedman χ^2 and Iman-Davenport F statistics) that may be completed by the pairwise comparisons that allow using Friedman ranks (Z-score). The results are shown in Tables 1 and 2.

The main conclusion of the comparative assessment, whose complete description can be found in Andrés-Sánchez and González-Vila (2019), is that our proposed methodology improves the BLC model.

	Pairwise Z scores from Friedman ranks		Friedman test	
	FRLC vs BLC	Friedman χ^2	Iman-Davenport F	
Austria (Men)	3.164***	8.542**	6.079***	
Austria (Women)	3.062***	14.083***	15.621***	
Belgium (Men)	2.654**	18.375***	35.933***	
Belgium (Women)	2.654**	22.167***	133.026***	
France (Men)	2.858**	20.660***	68.042***	
France (Women)	2.449**	20.000***	55.000***	
Italy (Men)	2.654**	16.420***	23.828***	
Italy (Women)	2.654**	22.167***	133.026***	
Netherlands (Men)	2.858**	20.667***	68.208***	
Netherlands (Women)	2.654**	10.830***	9.046***	
Portugal (Men)	2.654**	22.167***	133.026***	
Portugal (Women)	3.062***	15.500***	20.059***	
Spain (Men)	2.858**	20.667***	68.208***	
Spain (Women)	3.062***	14.083***	15.621***	
UK(Males)	2.654**	5.417*	3.207*	
UK (Women)	2.654**	16.417***	23.815***	

Table 1. Results of Friedman rank tests and pairwise Friedman rank tests for the accuracy of the confidence interval predictions on central mortality rates by BLC and FRLC in sample populations in the period 2001-2012.

	Pairwise Z scores from Friedman ranks		Friedman test	
	FRLC vs BLC	Friedman χ^2	Iman-Davenport F	
Austria (Men)	3.164***	19.042***	42.247***	
Austria (Women)	2.449**	7.750**	5.246**	
Belgium (Men)	2.143*	18.370***	35.892***	
Belgium (Women)	2.245*	20.100***	56.692***	
France (Men)	2.347*	12.540***	12.037***	
France (Women)	1.123	9.375***	7.051***	
Italy (Men)	2.449**	24.000***	∞ ***	
Italy (Women)	2.041	6.500**	4.086**	
Netherlands (Men)	4.695***	22.167***	133.026***	
Netherlands (Women)	2.347*	12.542***	12.041***	
Portugal (Men)	2.245*	13.040***	13.088***	
Portugal (Women)	1.429	11.420***	9.986***	
Spain (Men)	2.756**	16.250***	23.065***	
Spain (Women)	2.960***	14.083***	15.622***	
UK(Males)	2.552**	20.000***	55.000***	
UK (Women)	2.654**	22.167***	133.026***	

Table 2. Results of Friedman rank tests and pairwise Friedman rank tests for the accuracy of the confidence interval predictions on life expectancies by BLC and FRLC in sample populations in the period 2001-2012.

Notes: (1) ***, ** and * stand for the rejection of the null hypothesis with a significance level of 10%, 5% and 1% respectively. (2) Friedman χ^2 follows a Squared-Chi with 2 grades of freedom and Iman-Davenport F follows a Snedecor F with 2(24) grades of freedom.

References

- Andrés-Sánchez, J. de & González-Vila Puchades, L. (2019). A Fuzzy-Random Extension of the Lee-Carter Mortality Prediction Model. *International Journal of Computational Intelligence Systems* 12(2), 775-794.
- Dubois, D. & Prade, H. (1993) Fuzzy numbers: an overview. In: Dubois, D., Prade, H. & Yager, R.R. (eds.) *Fuzzy Sets for intelligent systems*, 113-148. Morgan Kaufmann Publishers.
- García, S. & Herrera, F. (2008) An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons. *Journal of Machine Learning Research* 9, 2677-2694.
- García, S., Fernández, A., Luengo, J. & Herrera, F. (2010) Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental Analysis of Power. *Information Sciences* 180, 2044-2064.
- Human Mortality Database. <http://www.mortality.org>.
- Ishibuchi, H. & Nii, M. (2001) Fuzzy regression using asymmetric fuzzy coefficients and fuzzified neural networks. *Fuzzy Sets and Systems* 119, 273-290
- Lee, R.D. & Carter, L.R. (1992). Modeling and forecasting US mortality. *Journal of the American Statistical Association*, 419(87), 659-675.
- Tanaka, H. (1987) Fuzzy data analysis by possibility linear models. *Fuzzy Sets and Systems* 24, 363-375.

Estimating the Healthy Life Expectancy (HLE) in the far past: The case of Switzerland (1876-2016) with forecasts to 2060 and comparisons with HALE from WHO

Christos H Skiadas¹ and Charilaos Skiadas²

¹ ManLab, Technical University of Crete, Chania, Greece
(E-mail: skiadas@cmsim.net)

² Department of Mathematics and Computer Science, Hanover College, Indiana,
USA
(E-mail: skiadas@hanover.edu)

Abstract

Healthy Life Expectancy (HLE) estimates are achieved after systematic work of a large group of researchers all over the world during last decades. The most successful estimate was termed as HALE and is provided by the World Health Organization (WHO) in the related website. Having established a methodology of data collection and handling the HLE can be estimated and provided to researchers and policy makers.

However, it remains an unexplored period of the last few centuries where, LE data exists along with the appropriate life tables, but not enough information for HLE estimates is collected and stored. The problem is now solved following a methodology of estimating the HLE from the life tables after the Healthy Life Years Lost (HLYL) estimation. Our methodology on a Direct HLYL estimation from Life Tables, is tested and verified via a series of additional methods including a Weibull parameter test, a Gompertz parameter alternative and of course a comparison with HALE estimates from WHO. The complete

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



C. H. Skiadas and C. Skiadas

methodology and estimation methods are published in the book on “*Demography of Population Health, Aging and Health Expenditures*” of Volume 50 of the Springer Series on Demographic Methods and Population Analysis.

<https://www.springer.com/gp/book/9783030446949>,

<https://doi.org/10.1007/978-3-030-44695-6>

Keywords: Life Expectancy, Healthy Life Expectancy, HALE, Logistic model, forecasts, Life Tables

Life Expectancy and Healthy Life Expectancy estimates

The full Life Tables are used not only to estimate the Life Expectancy (LE) but also the Healthy Life Expectancy (HLE) based on the methodology presented in the new book “*Demography of Population Health, Aging and Health Expenditures*” of Volume 50 of the Springer Series on Demographic Methods and Population Analysis.

<https://www.springer.com/gp/book/9783030446949>,

<https://doi.org/10.1007/978-3-030-44695-6>

Based on the data series from 1900 to 2016 for males and females in Switzerland, estimates until 2016 and forecasts to 2060 are done. The Logistic model is fitted to data series to calculate the three parameters of the model. Then forecasts to 2060 are done. For fitting and long range forecasts the Logistic Model is selected.

1900 was a milestone in health improvement in many countries. The development of the first healthcare system of modern history, started with policies introduced by the Otto von Bismarck's social legislation (1883-1911). The introduction of such systems in many countries came after important discoveries from scientists as Pasteur and Chamberland in France, Von Behring in Germany, Kitasato from Japan,

Healthy Life Expectancy in Switzerland

Descombey from France and many others. The 1901 Nobel Prize in Physiology or Medicine, the first one in that field, awarded to Von Behring for his discovery of a diphtheria antitoxin.

It looks like the health care systems and methodologies already set in 1900 follow a rather systematic trend until today. See figure 1 where Life Expectancy (LE) data series is provided by the Human Mortality Database (HMD) and Healthy Life Expectancy (HLE) is estimated with our Direct methodology (Skiadas & Skiadas 2018a,b and 2020a,b,c). The LE series from 1876-1900 is strongly fluctuating mainly due to health causes. The fluctuations become smaller after this period with a clear stabilization from 1950 until now except the strong declining during the 1918 influenza pandemic followed with a fast recovery later on. The period starting from 1950 is followed with a rather smooth trend as a result of the improvement of the health systems structure, financing, technology and pharmaceutical discoveries and production.

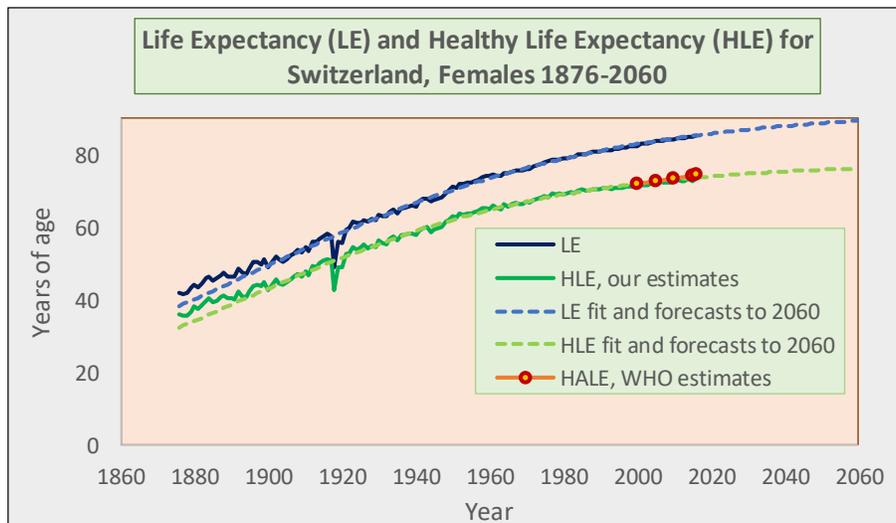


Fig. 1. Life Expectancy (LE) and Healthy Life Expectancy (HLE) in Switzerland, females (1876-2016).

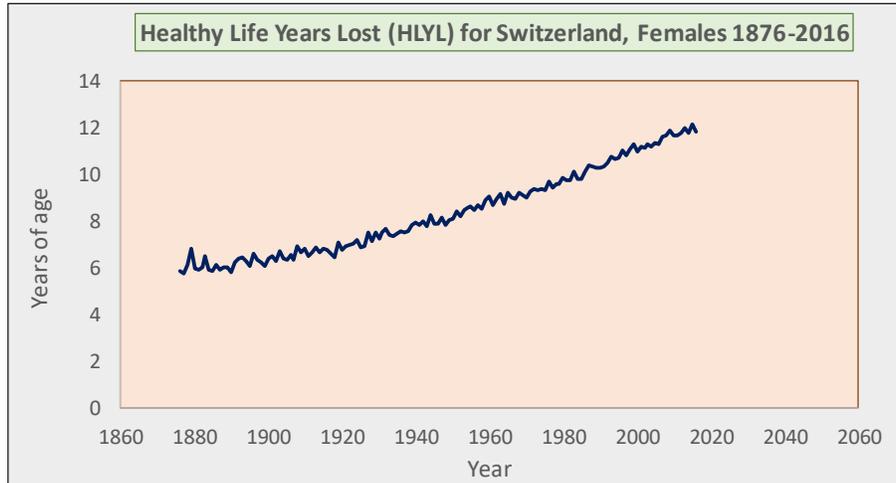


Fig. 2. Healthy Life Years Lost (HLYL) in Switzerland, females (1876-2016).

The Healthy Life Years Lost (HLYL) calculated data series is illustrated in figure 2. The HLYL trend is slightly growing until 1876 followed with a faster grow until 11.84 years of age in 2016 after 5.82 years of age in 1876.

The Logistic Model

This classical model proposed by P. F. Verhulst in 1838 to estimate the population of France is proven to be a successful tool for long range forecasting. In his first application, Verhulst predicted the population of France for almost 100 years. Pearl and Reed used this model to predict the growth of the United States Population. Applications in other countries have also done.

The three parameter Logistic Model equation form is the following

Healthy Life Expectancy in Switzerland

$$g(t) = \frac{F}{1 + \left(\frac{F}{g(0)} - 1\right) \exp(-b(t - T(0)))}$$

Where b is the trend or diffusion parameter and F is the upper level of the sigmoid logistic process and $g(0)$ is the value at time $T(0)=1900$.

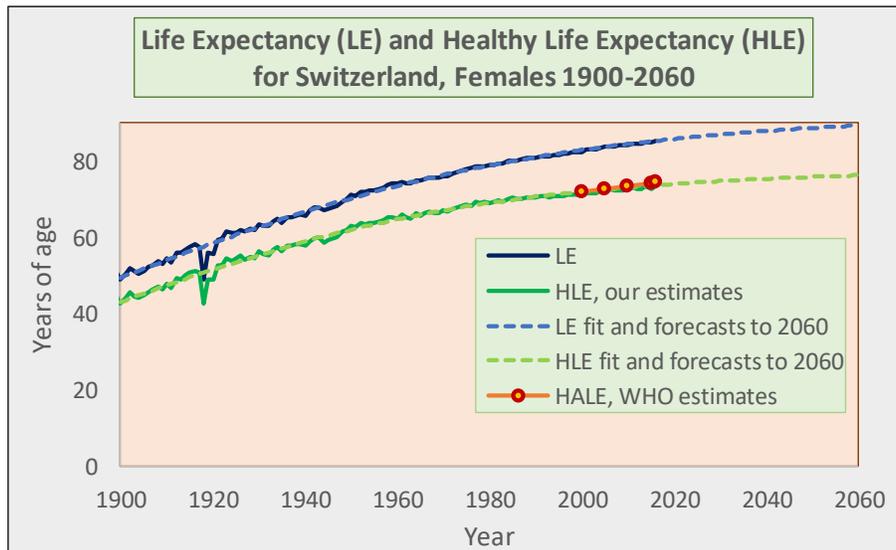


Fig. 3. Logistic model fit and forecasts to 2060 for females in Switzerland.

The HALE estimates and our Direct calculations

The latest WHO estimates for Healthy Life Expectancy called HALE are provided for the years 2000, 2005, 2010, 2015 and 2016.

These estimates perfectly fit into our calculations for the HLE and the fit results by using the Logistic model.

Our HLE calculations are based on the Direct estimates from the Life Tables of the HLYL with a formula provided in recent publications (Skiadas & Skiadas 2020a,b,c) that is:

C. H. Skiadas and C. Skiadas

$$HLYL = \max \frac{xm_x}{\sum_0^x m_x}$$

Where m_x is the mortality at age x as provided in HMD life tables.

Then $HLE=LE-HLYL$.

The Logistic model is applied to data sets for LE and HLE from our estimates from 1900 to 2016. The parameters selected appear in the following Table I.

The Healthy Life Years Lost (HLYL) are 6.40 years of age in 1900, 11.63 for 2016, 13.11 in 2060 with a maximum of $F=14.61$ years of age difference.

TABLE I. Logistic model parameters and estimates

	Logistic Model Parameters		LE and HLE in 1900, 2016 and 2060		
	b	F	1900	2016	2060
LE	0.0203	92.43	49.46	85.37	89.40
HLE	0.0231	77.82	43.09	73.74	76.29
HLYL= LE-HLE		14.61	6.37	11.63	13.11

Table II summarizes the three healthy life expectancy estimates from WHO (HALE), our Direct estimates and from Logistic fit. All three methodologies provide close results.

TABLE II. HALE and Healthy Life Expectancy Direct estimates and Logistic fit

Year	2000	2005	2010	2015	2016
WHO HALE	71.83	72.76	73.65	74.26	74.47
Direct HLE	71.59	72.35	72.69	72.75	73.40
Logistic Fit	72.05	72.64	73.17	73.65	73.74

Conclusions

We have solved the problem of finding the HLE in the far past. The case of Switzerland (1876-2016, females) with forecasts to 2060 and comparisons with HALE has explored. The selected Logistic model has a good fit while the HALE estimates from WHO compare very good to our estimates both with Direct method and the logistic fit.

References

Skiadas, C.H. and Skiadas, C. (2018). *Exploring the Health State of a Population by Dynamic Modeling Methods*. The Springer Series on Demographic Methods and Population Analysis 45, Springer, Chum, Switzerland. <https://doi.org/10.1007/978-3-319-65142-2>.

Skiadas, C.H. and Skiadas, C. (2018). *Demography and Health Issues: Population Aging, Mortality and Data Analysis*. The Springer Series on

C. H. Skiadas and C. Skiadas

Demographic Methods and Population Analysis 46. Springer, Chum, Switzerland. <https://doi.org/10.1007/978-3-319-76002-5>.

Skiadas, C.H. and Skiadas, C. (2020). *Demography of Population Health, Aging and Health Expenditures*. The Springer Series on Demographic Methods and Population Analysis 50. Springer, Chum, Switzerland. <https://www.springer.com/gp/book/9783030446949> , <https://doi.org/10.1007/978-3-030-44695-6>

Skiadas, C.H. and Skiadas, C. (2020). *Relation of the Weibull Shape Parameter with the Healthy Life Years Lost Estimates: Analytical Derivation and Estimation from an Extended Life Table*. In The Springer Series on Demographic Methods and Population Analysis 50. Springer, Chum, Switzerland. https://doi.org/10.1007/978-3-030-44695-6_2

Skiadas, C.H. and Skiadas, C. (2020). *Direct Healthy Life Expectancy Estimates from Life Tables with a Sullivan Extension. Bridging the Gap Between HALE and Eurostat Estimates*. In The Springer Series on Demographic Methods and Population Analysis 50. Springer, Chum, Switzerland. https://doi.org/10.1007/978-3-030-44695-6_3

Big Data-driven Approach to Validate Properties of Markov-Modulated Linear Regression Model's Estimator Empirically

Nadezda Spiridovska¹, Ilya Jackson²

^{1,2} Research and Development Department, Department of Mathematical Methods and Modelling, Transport and Telecommunication Institute, Lomonosova 1, Riga, LV-1019, Latvia

(¹ E-mail: Spiridovska.N@tsi.lv, ² E-mail: Jackson.I@tsi.lv)

Abstract. Markov-modulated linear regression (MMLR) model is a special case of Markov-additive processes firstly proposed by Alexander Andronov in 2012. The model assumes that unknown regression coefficients depend on an external state of the environment, but regressors remain constant. MMLR model differs from other switching models by a new analytical approach to parameter estimation and known transition intensities between the states of Markov component. This paper analyses statistical properties of MMLR model's estimator based on simulated data. The research considers the influence of the sample parameters (e.g., sample size and distribution of the initial data), as well the influence of estimation method details (e.g., different weight matrices in OLS) on the consistency of model estimates. The detailed artificial domains, experiment design and numerical results are presented. The artificial probability distributions are used, which allows one to control their complexity and expand the field of experiments.

This research is carried out within the project No. 1.1.1.2/VIAA/1/16/075 and is funded by the post-doctoral research aid programme of the Republic of Latvia.

Keywords: Markov-modulated linear regression, random environment, simulation.

1 Introduction

Currently Markov processes and models with underlying Markov chains are widely used for modeling queueing systems, transportation, inventory systems, DNA sequences and many other practical systems involving endogenous environment [1].

Markov-modulated linear regression (MMLR) was firstly proposed by Andronov and Spiridovska [2]. This model belongs to Markov-additive processes (MAPs) that are part of Markov-modulated processes. MAPs form a general class of two-component stochastic processes, which includes many influential models such as Markov-modulated Brownian motion (MMBM), Markov random walk (MRW), Markov renewal processes (MRP) and others. Such models are deeply studied by matrix-exponential methods [3]. According to Pacheco et al., a Markov-additive process $(X, J) = \{(X(t), J(t)), t \geq 0\}$ is a two-component or two-dimensional Markov process defined on the state space $R \times E$ such that, for

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



$s, t \geq 0$, the conditional distribution of increments $(X(s+t) - X(s), J(s+t))$ given $(X(s), J(s))$ depends only on Markov component $J(s)$, where (Ω, \mathcal{F}, P) is a probability space, $R = (-\infty, +\infty)$, E is a countable set and $N_+ = \{0, 1, 2, \dots\}$ is given. A component $J(s)$ is Markov, and component $X(s)$ is additive [4].

Markov-switching regression models or regime switching regression models share the same idea of varying the regression parameters randomly in accordance with external environment. However, these models have significant differences in terms of the analytical approach. Namely, MMLR describes the external environment as a continuous-time homogeneous irreducible Markov chain with known parameters. On the other hand, switching models consider Markov chain as unobservable and estimation procedure involves estimation of transition matrix.

MMLR model is inherently a strategic level model that can be successfully applied for long-term forecasting. Therefore, time required to estimate the parameters is not critical for this model. The first simulation-based studies of the model [2] demonstrated that the small size of the initial sample leads to completely unsatisfactory results. For instance, the estimates were far from the true values of the parameters. However, an increased sample size led to improved estimates. After that MMLR model was applied in two practical case studies. Based on the results the authors concluded with the statement that it is crucial for the real-world application of MMLR to have more simulation-based empirical validations of model's estimator properties, which was the main motive for this study.

The paper is structured as follows. Section 2 sheds the light on the related work and outlines the novelty. Section 3 provides the theoretical framework of the model. Section 4 describes experimental design. Section 5 covers the simulation results and validation. Finally, the paper discusses the findings and gives recommendations for future research and possible applications. The model and experiments are implemented in Python 3.7 and the source-code is available in the Github repository [5].

2 Related work and novelty

MMLR model proved its worth in practice. Namely, two cases studies were conducted and described in detail. Firstly, the model was applied to analyze coaches' delay time at the Riga Coach Terminal for the period from 2012 to 2017 [6]. Despite a large amount of initial data, for example, 5414 records only in 2012, the number of observations has significantly decreased, due to data preparation phase and the assumptions related to time-additivity of the response variable. Despite the fact that the applied model demonstrated low predictive accuracy, it was not related to the incorrectness of the proposed model, on contrary, it was caused by the low quality of the classical regression model that did not take into consideration any factor besides the day of the week, which was not the only factor determining the delay time. Additionally, sparse and repeating data could also cause unreliable results. The next approbation of the model was carried out on the data of trip validations provided by Rigas Satiksme, a provider of transport

services in Riga [7]. Initial data set covered time period of 4.5 months in 2017 and contained 1,048,001 observations. However, after data transformation and aggregation the number of observations decreased significantly and prediction did not show convincing results for the same reasons.

Since MMLR model is efficiently applicable at the strategic level for long-term planning, the quality and quantity of the initial data is critically important. This issue can be considered in the context of Big Data. On the first hand, Big Data solutions have huge potential related to more computationally efficient data processing. On the other hand, there are such drawbacks as non-representativeness and unreliability. Besides, results heavily depend on the quality of initial data [8, 9]. Therefore, in order to tailor MMLR model to real-world applications, the framework for data exploring and preparing was developed. The framework took into account the limitations and necessary assumptions unveiled after applying the model in practice [10].

Based on the case studies, the authors concluded with the statement that it is crucial for the real-world application of MMLR to have more Big Data-driven and simulation-based empirical validations of model's estimator properties. In this regard, the novelty of this research is associated with taking into consideration the influence of the sample parameters (e.g., sample size and distribution of the initial data), as well the influence of estimation method details (e.g., different weight matrices in OLS) on the consistency of model estimates.

3 Theoretical Framework

Markov-modulated linear regression is a special case of a Markov-additive process (Y, J) , where the component J is Markovian. The component Y , on the other hand, is additive and described by linear regression. The MMLR model is additionally examined by Andronov [11]. The core ideas behind the model are covered in this section.

Considering the case, in which the component J is finite ($m < \infty$) ergodic continuous-time Markov chain with the known transition rates $\lambda_{i,j}$ between states i and j . Let $\Lambda_i = \sum_{j \neq i} \lambda_{i,j}$. MMLR supposes that, if on the interval $(s, s + t)$ a Markovian component state is j , then an increment of additive component Y is described by linear regression:

$$Y(s + t) - Y(s) = xt\beta^{(j)} + Z\sqrt{t}, i = 1, \dots, n, \quad (1)$$

where $x = (x_1, x_2, \dots, x_k)$ is $1 \times k$ vector of regressors, $\beta^{(j)}$ is $k \times 1$ vector of unknown regression coefficients. These regression coefficients are different and depend on the state $\beta^{(j)} = (\beta_{1,j}, \beta_{2,j}, \dots, \beta_{k,j})^T, j = 1 \dots m$, where m stands for the number of states. Z is a disturbance scale under usual assumptions. Namely, Z does not depend on $Y(s)$ and j , distributed normally with zero mean and constant variance σ^2 .

On the considered time interval $(0, \tau)$ initial values of the components are $Y(0) = 0, J(0) = i (i = 1 \dots m)$. T_j stands for total sojourn time of Markovian component in the state j on the interval $(0, \tau)$. $\vec{T} = (T_1, \dots, T_m)$ and $T_1 + T_2 + \dots + T_m = \tau$. Z_1, \dots, Z_m are disturbance terms that take place at the moments of

time T_1, \dots, T_m respectively. Thus, additive component Y can be expressed as follows:

$$Y(\vec{T}) = x(\beta^{(1)}T_1 + \dots + \beta^{(m)}T_m) + (Z_1\sqrt{T_1} + \dots + Z_m\sqrt{T_m}). \quad (2)$$

Since Z_1, \dots, Z_m and \vec{T} are independent and taking into account normal distribution properties the equation (2) can be formulated as follows:

$$Y(\vec{T}) = x(\beta^{(1)}T_1 + \dots + \beta^{(m)}T_m) + \sqrt{\tau}Z. \quad (3)$$

Using Kronecker product \otimes and vectorization operator vec it is possible to represent the model in the standard form of linear regression:

$$Y(\vec{T}) = (\vec{T} \otimes x)vec\beta + \sqrt{\tau}Z, \quad (4)$$

where $\beta = (\beta^{(1)} \dots \beta^{(m)})$ is a matrix composed of columns $\{\beta^{(j)}\}$.

Regression coefficients $\beta = (\beta^{(1)} \dots \beta^{(m)})$ and variance of disturbance term σ^2 must be estimated based on n independent observations of the process (4). On the other hand, parameters describing Markovian component as a number of states m and transition intensities $\{\lambda_{i,j}\}$ are known. For the r^{th} observation ($r = 1, \dots, n$) the following is recorded:

- the initial state of the Markovian component $J_r(0)$;
- the duration of a single observation τ_r ;
- values of the additive response variable Y_r ;
- the vector of regressors $x_r = (x_{r,1}, x_{r,2}, \dots, x_{r,k})$.

The whole trajectory of the environment J is unknown, therefore, sojourn times in each state are also unknown for r -th observation $\vec{T}_r = (T_{r,1}, \dots, T_{r,m})$. However, the estimated conditional average sojourn times $\vec{t}_r = E(\vec{T}_r) = (E(T_{r,1}), \dots, E(T_{r,m})) = (t_{r,1}, \dots, t_{r,m})$ are used in the state j instead of actual sojourn times.

In order to improve the convergence of the regression coefficients' estimates, it was proposed to introduce an additional random component V , which would be taken into account in the weight coefficients. Thus, the model can be represented in matrix notation as follows:

$$Y = (Y_1, \dots, Y_n)^T = \begin{pmatrix} \vec{t}_1 \otimes x_1 \\ \vec{t}_2 \otimes x_2 \\ \dots \\ \vec{t}_n \otimes x_n \end{pmatrix} vec\beta + diag(\sqrt{\tau_1}, \dots, \sqrt{\tau_n})Z + V, \quad (5)$$

where $Y = (Y_1, \dots, Y_n)^T$, $Z = (Z_1, \dots, Z_n)^T$, $V = (V_1, \dots, V_n)^T$ and $\sqrt{\tau} = (\sqrt{\tau_1}, \dots, \sqrt{\tau_n})$ are n -vectors of scale responses, disturbance terms, random components (zero mean and known variance) and observation times respectively. $\vec{t}_r = (t_{r,1}, \dots, t_{r,m})$ stands for m -vector of estimated conditional average sojourn times. $X = (x_{r,k}) = (x_r: r = 1, \dots, n)$ is a $n \times k$ matrix of regressors, and $diag(\alpha)$ stands for the n -dimensional diagonal matrix with the vector α on the main diagonal.

Additionally, it is supposed that matrix $X = ((\vec{t}_1 \otimes x_1)^T (\vec{t}_2 \otimes x_2)^T \dots (\vec{t}_n \otimes x_n)^T)^T$ of size $n \times km$ has rank $r(X) = km$, thus, $(X^T X)^{-1}$ exists.

In this case generalized least squares method (GLSM) estimates regression coefficients as follows:

$$vec\tilde{\beta} = (X^T W^{-1} X)^{-1} X^T W^{-1} Y, \quad (6)$$

where $W = diag(w_1, w_2, \dots, w_n)$ is non-degenerate diagonal weight matrix, and estimation (6) is unbiased. In case of OLS weight matrix is an identity matrix ($W = I$) [12-14].

Besides, the conditional average and covariances for sojourn times in different states must be estimated. Calculations are performed using transition probabilities and eigenvectors-eigenvalues. The complete derivation of formulas and equations is described by Andronov and Spiridovska [2] and Andronov [11].

For the conditional average sojourn time $t_{i,v,j}(\tau) = E(T_v | t = \tau, J_0 = i, J_\tau = j)$ in the state $v \in S$ on the interval $(0, \tau)$:

$$t_{i,v,j}(\tau) = \frac{1}{p_{i,j}(\tau)} \int_0^\tau p_{i,v}(u) p_{v,j}(\tau - u) du, \quad (7)$$

$$p_{i,j}(\tau) = P\{J(\tau) = j | J(0) = i\} = \sum_{\eta=1}^{mp} Z_{i,\eta} \exp(\gamma_\eta \tau) \bar{Z}_{\eta,j}, \quad (8)$$

where $A = \lambda - \Lambda$ is a matrix, Λ stands for m -dimensional diagonal matrix with a vector $(\Lambda_1, \dots, \Lambda_m)$ on the main diagonal, such that $\Lambda_i = \sum_{j \neq i} \lambda_{i,j}$. v_η and Z_η for $\eta = 1, \dots, m$ are the eigenvalue-eigenvector pair of A . $Z = (Z_1, \dots, Z_m)$ is a matrix of the eigenvectors and $\bar{Z} = Z^{-1} = (\bar{Z}_1^T, \dots, \bar{Z}_m^T)^T$ its inverse. Such that \bar{Z}_η is the η -th row of \bar{Z} . This formula can be applied along with numerical integration.

Covariation of sojourn times can be expressed as follows:

$$Cov(T_v, T_\mu | \tau, J(0) = i, J(\tau) = j) = t_{i,v,\mu,j}(\tau) - t_{i,v,j}(\tau) t_{i,\mu,j}(\tau), \quad (9)$$

where $t_{i,v,\mu,j}(\tau)$ is conditional second (mixed) moment for states $v, \mu \in N$ on the interval $(0, \tau)$: $t_{i,v,\mu,j}(\tau) = E(T_v, T_\mu | \tau, J(0) = i, J(\tau) = j)$:

$$t_{i,v,\mu,j}(\tau) = \frac{1}{p_{i,j}(\tau)} \int_0^\tau p_{i,v}(u) E(T_\mu(\tau - u) I(J(\tau - u) = j) | J_0 = v) p_{i,\mu}(u) E(T_v(\tau - u) I(J(\tau - u) = j) | J_0 = \mu) du, \quad (10)$$

where $I(A)$ is an indicator function of an event A . The conditional average sojourn time spent in a particular state is calculated by the formula (7).

Expectation of response variable vector and the expression for the variance of response variable are accordingly:

$$E(Y(\vec{T})) = (E(\vec{T}) \otimes x) vec\beta = (\vec{t} \otimes x) vec\beta. \quad (11)$$

$$D(Y(\vec{T})) = vec\beta^T (Cov(\vec{T}) \otimes x^T x) vec\beta + \sigma^2 \tau \quad (12)$$

Coefficients of the regression model are estimated using formula (6). To minimize variance of the estimates the inverse value of the response variable variance (12) will be used as the weight coefficient for the current observation. Nevertheless, the expression contains unknown parameters β and σ^2 that are subject to estimation. This case usually requires an iterative evaluation procedure. Firstly, regression coefficients are estimated by the formula (6). Secondly, the variance σ^2 is estimated. Lastly, the estimations alternate until a satisfactory result is obtained. An expression for estimating variance σ^2 is the following:

$$\hat{\sigma}^2 = (\sum_{r=1}^n \tau_r)^{-1} (\bar{D} - vec\hat{\beta}^T [\sum_{r=1}^n (Cov(\vec{T}_r) \otimes x_r^T x_r)] vec\hat{\beta}), \quad (13)$$

where

$$\tilde{D} = \sum_{r=1}^n (Y_r - (\vec{t}_r \otimes x_r) \tilde{\beta})^2.$$

4 Experimental Design

As it is concluded by Kleijnen [15], since results are directly dependent on initial data, the process of designing an experiment is extremely important. Therefore, it is worth to stay away from poor practices, for example, to keep input parameters of the simulation model constant or to vary only one input at a time, while preserving others as the so-called base values. Taking into account recommendations and good practices of experimental design, three simulation experiments are conducted.

All experiments share the following setup. External environment J is presented by 3 states ($m = 3$) and the matrix with known transition rates $\lambda_{i,j}$ between states i and j :

$$\lambda = \begin{pmatrix} 0 & 0.2 & 0.3 \\ 0.1 & 0 & 0.2 \\ 0.4 & 0 & 0 \end{pmatrix}.$$

Stationary state distribution is the following: $\pi = (0.364, 0.242, 0.394)^T$, see Kijima [11]. Total number of observations $n = 30$. The number of independent variables (regressors) $k = 3$.

The following initial data is randomly generated for sampling-based estimation procedures and considered to be known:

- 1) Matrix of independent variables XF ;
- 2) Vector IF that contains initial states $J_{i,0}$ of the Markov chain $J(\cdot)$ ($i = 1, 2, \dots, n$);
- 3) Vector τF that contains total observation times $t_i = T_{i,1} + \dots + T_{i,m}$ for each observation ($i = 1, 2, \dots, n$).

The initial data is available along with all the source code in the GitHub repository [5]. It is worth to note that the total duration of observations is 25.1. The data is simulated for the following values of the parameters of the regression model:

$$\beta = \begin{pmatrix} 2.1 & 3.7 & 4.2 \\ 1.7 & 3.2 & 6.3 \\ 5.2 & 8.4 & 3.0 \end{pmatrix} \text{ and } \sigma = 2,$$

so $vec\beta = (\beta_{1,1}, \beta_{2,1}, \beta_{3,1}, \beta_{1,2}, \beta_{2,2}, \beta_{3,2}, \beta_{1,3}, \beta_{2,3}, \beta_{3,3})^T =$
 $(2.1, 1.7, 5.2, 3.7, 3.2, 8.4, 4.2, 6.3, 3.0)^T$.

Disturbance scale Z has variance $\sigma^2 = 4$ which suggests more difficult estimation conditions (comparing to previously chosen $\sigma^2 = 1$ in earlier experiments).

The simulation setup assumes that various observations are independent random experiments grouped into batches. Each batch has the structure described above. Namely, 30 observations and the number of regressors $k = 3$. The matrix

of regressors, initial state of random environment and observation times are generated randomly according to the following distributions:

- the expectation of the matrix of regressors coincides with the above obtained matrix X , all elements of the 2nd and 3rd columns are independent and uniformly distributed on intervals $(-1, 1)$ and $(-0.5, 0.5)$ correspondingly;
- the expectation of observations' times coincides with previous values t , all times are independent and time of i -th observation t_i , has uniform distribution on $(0, 2t_i)$;
- the initial states I for various observations are independent and are chosen with respect to stationary distribution of the states for the random environment J .

Therefore, the total number of the observations for one experiment equals to $30q$, where q is a number of simulated batches.

In order to evaluate the quality of the obtained models, the following indicators have been used: coefficient of determination (R^2), root mean square error (RMSE), F-statistics and mean absolute percentage error (MAPE).

In the first experiment the number of batches q varies from 100 to 1000 with the step 100, therefore the total number of observations n varies from 3000 to 30000. Regression coefficients are estimated according to formula (6). The convergence of the regression coefficients' estimates is compared for two types of weight matrices, namely identity matrix (OLS case) and matrix with the inverse values of the response variable variances $W = \text{diag}(D(Y_1)^{-1}, D(Y_2)^{-1}, \dots, D(Y_n)^{-1})$, according to the formula (12). The following hypotheses are stated: 1) the sample size significantly affects the accuracy of estimates; 2) different weight matrices significantly affect the accuracy of estimates and the convergence speed.

The core idea behind the second experiment is to insert extra noise to the matrix of regressors. When matrix X is generated, uniformly distributed noise with parameters $(-10 * NF; 10 * NF)$ is added to the elements of the 2nd column and $(-1 * NF, 1 * NF)$ to the elements of the 3rd column. Such that NF is a noise factor, which takes values from 0.05 to 0.5 with a step of 0.05 (0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50). For each NF the value vector $\text{vec}\tilde{\beta}$ is estimated. Each estimate is based on 600 iterations. In this regard the following hypothesis is put forward: instability of the initial data qualitatively affects the convergence of estimates.

5 Numerical Study

In the first experiment the convergence speed of the regression coefficients for two weight matrices W are compared. In the first case of $W = \text{diag}(1, 1, \dots, 1)$, the obtained estimates of the $\text{vec}\tilde{\beta} = (\tilde{\beta}_{1,1}, \tilde{\beta}_{2,1}, \tilde{\beta}_{3,1}, \tilde{\beta}_{1,2}, \tilde{\beta}_{2,2}, \tilde{\beta}_{3,2}, \tilde{\beta}_{1,3}, \tilde{\beta}_{2,3}, \tilde{\beta}_{3,3})^T$ with regard to the sample size are presented in the Table 1. The last column contains the true values of the parameters. In the iterative procedure, the initial

values of the vector of unknown model parameters $vec\tilde{\beta}$ equal to zero. Coefficients of the regression model are estimated using formula (6).

Table 1. Convergence speed of the regression coefficients using identity weight matrix $W = I$

	Estimated in q iterations										True
	100	200	300	400	500	600	700	800	900	1k	
$\beta_{1,1}$	1.47	1.83	1.97	1.60	2.86	1.61	2.69	1.91	2.39	1.80	2.1
$\beta_{2,1}$	1.79	1.79	1.75	1.77	1.53	1.70	1.53	1.99	1.64	1.77	1.7
$\beta_{3,1}$	5.38	5.21	5.27	5.32	5.06	5.33	5.09	5.07	5.05	5.20	5.2
$\beta_{1,2}$	4.95	5.12	4.00	3.95	2.84	2.89	1.92	2.93	2.55	2.72	3.7
$\beta_{2,2}$	2.49	2.70	2.72	3.23	3.09	3.58	4.35	3.22	3.63	3.34	3.2
$\beta_{3,2}$	8.32	8.26	8.61	8.31	8.95	8.51	8.29	8.60	8.87	8.84	8.4
$\beta_{1,3}$	4.16	3.99	4.46	4.30	4.05	3.88	3.66	3.86	4.27	4.35	4.2
$\beta_{2,3}$	6.28	6.38	6.33	6.17	6.43	6.33	6.27	6.22	6.17	6.20	6.3
$\beta_{3,3}$	3.04	3.02	2.85	3.05	2.92	3.05	3.19	3.20	3.12	3.01	3.0
MAPE %	45.1	44.6	42.7	35.0	34.8	30.7	30.3	27.7	27.9	27.5	0.0

In case of the identity weight matrix, non-uniformity of observations is not taken into account, variance $\sigma^2 = 4$ and its estimation is not required. Table 1 demonstrates that even after thousand iterations, estimates of some coefficients are distinguished by relatively large deviations from the true parameters.

Table 2 shows the case of the diagonal weight matrix, in which the diagonal elements are the inverse values of the variances of the dependent variable. Such a weight matrix is applied to minimize the variance of the estimates and to improve the convergence. In practical settings the unknown parameters β and σ^2 are subject to estimation as well. However, this experiment assumes that their values are known, for example, obtained in a previous study.

Table 2. Convergence speed of the regression coefficients using weight matrix with the inverse values of the response variable variances

	Estimated in q iterations										True
	100	200	300	400	500	600	700	800	900	1k	
$\beta_{1,1}$	2.26	1.76	2.22	2.16	1.87	2.14	1.79	1.99	2.01	2.09	2.1
$\beta_{2,1}$	1.65	1.66	1.62	1.76	1.81	1.76	1.75	1.77	1.65	1.73	1.7
$\beta_{3,1}$	4.99	5.34	5.24	5.13	5.21	5.13	5.28	5.25	5.15	5.26	5.2
$\beta_{1,2}$	4.43	2.73	4.08	4.14	3.92	3.99	3.89	3.62	3.46	2.55	3.7
$\beta_{2,2}$	3.04	3.30	3.07	2.82	3.11	3.06	3.28	3.62	3.43	3.21	3.2
$\beta_{3,2}$	8.53	8.74	8.31	8.52	8.39	8.43	8.22	8.46	8.51	8.67	8.4
$\beta_{1,3}$	3.96	4.59	4.01	4.42	3.99	3.99	4.45	4.15	4.24	4.17	4.2
$\beta_{2,3}$	6.32	6.26	6.30	6.32	6.43	6.32	6.27	6.29	6.32	6.33	6.3
$\beta_{3,3}$	3.08	2.87	3.03	2.92	3.02	3.05	2.94	3.02	2.99	2.99	3.0
MAPE %	37.1	32.4	32.1	34.6	34.5	33.3	33.3	32.9	28.7	26.9	0.0

Both the convergence speed and accuracy of estimation increased after the diagonal weight matrix $W = \text{diag}(D(Y_1)^{-1}, D(Y_2)^{-1}, \dots, D(Y_n)^{-1})$ is applied. Fig.1 reflects changes of the determination coefficients for both cases. Since R^2 naturally increases along with growing sample size, it cannot be fully reliable. In this regard, in order to verify the quality of the model, RMSE (Fig.2) and F-statistic (Fig.3) were also calculated. All three figures demonstrate the advantage of weight matrix with the inverse values of the response variable.

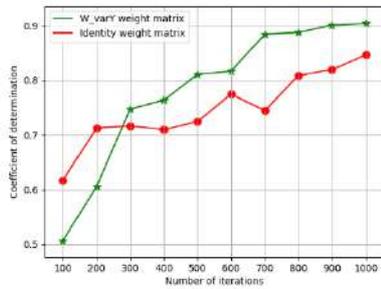


Fig. 1. Coefficient of determination

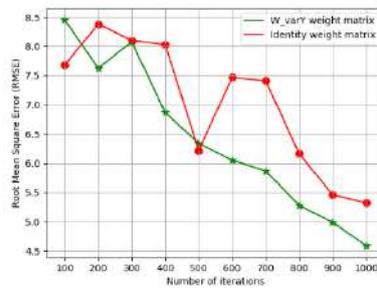


Fig. 2. RMSE

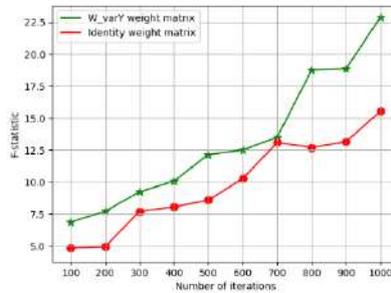


Fig. 3. F-statistic

Additionally, for both cases, the sample size is a significant component of the parameter estimation procedure. Despite the choice of weight matrices, estimates converged to the true values of the parameters, and the difference is only in speed.

If not the true value of the variance σ^2 (which is unknown) is used in (12), but an approximate, the convergence would, most likely, slow down. The same applies to the parameters β . In this experiment, in order to obtain the variance of the dependent variable $D(Y_i)$, the true values of the model parameters β are used, not the estimates $\hat{\beta}$. In a real situation true values are usually not available. Therefore, estimates of $\hat{\beta}$ must be used instead, which also complicates the estimation procedure and potentially slows down the convergence of the estimates. However, these statements require experimental validation and can be considered as the basis for further research.

In the second experiment extra noise is inserted to the matrix of regressors. The results of the estimates are presented in Table 3 depending on the value of NF.

Table 3. Convergence speed with regard to the noise factor

	Noise factor (NF)										$\beta_{i,j}$
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	
$\beta_{1,1}$	2.21	1.90	2.27	2.06	2.23	2.11	2.10	2.07	2.06	1.96	2.1
$\beta_{2,1}$	1.67	1.76	1.68	1.71	1.68	1.71	1.71	1.68	1.68	1.69	1.7
$\beta_{3,1}$	5.16	5.24	5.15	5.22	5.18	5.18	5.22	5.23	5.19	5.24	5.2
$\beta_{1,2}$	3.73	3.69	3.69	4.19	3.56	3.70	3.76	4.10	4.11	4.18	3.7
$\beta_{2,2}$	3.26	3.20	3.19	3.11	3.39	3.13	3.13	3.19	3.20	3.18	3.2
$\beta_{3,2}$	8.35	8.38	8.41	8.31	8.28	8.36	8.38	8.22	8.22	8.22	8.4
$\beta_{1,3}$	4.17	4.26	4.16	4.17	4.15	4.13	4.26	4.15	4.31	4.27	4.2
$\beta_{2,3}$	6.32	6.27	6.36	6.32	6.31	6.31	6.32	6.30	6.31	6.31	6.3
$\beta_{3,3}$	3.01	2.99	2.98	3.01	3.02	3.01	2.97	3.03	2.98	2.96	3.0
MAPE%	32.9	32.2	32.4	36.1	32.4	32.7	32.6	35.3	34.7	35.8	0.0

It is worth to point out that the estimates are already influenced by errors in the regressors, because when calculating the estimates $\tilde{\beta}$ mathematical expectations of the sojourn times of Markov chain in different states are used. Therefore, extra noise in initial date slows down the convergence speed.

Conclusions

Based on the experimental results the paper concludes with the statement that both the sample size and choice of weight matrices have significant influence on the accuracy of estimates and the convergence speed.

Application of the weight matrix with the inverse values of the response variable variances leads to faster convergence and more accurate estimation in comparison with the identity matrix. Additionally, for both cases, the sample size is a significant component of the parameter estimation procedure. Despite the choice of weight matrices, estimates converged to the true values of the parameters. The difference was only in speed.

If not the true value of the variance σ^2 is used in the estimation procedure, but an approximate, the convergence is expected to slow down. The same applies to the parameters β . In a real-world case studies true values are usually not known. In this regard, estimates of $\tilde{\beta}$ must be used instead, which also complicates the estimation procedure and potentially slows down the convergence. However, this hypothesis requires experimental validation and can be considered as the basis for further research. Besides that, further research is needed in order to estimate the variance of the random component taking into account non-uniformity of observations. It is also planned to consider the impact of deviations from the initial assumptions (e.g., incorrectly estimated transition intensities).

Acknowledgements

This work was financially supported by the specific support objective activity 1.1.1.2. "Post-doctoral Research Aid" (Project id. N. 1.1.1.2/16/I/001) of the Republic of Latvia, funded by the European Regional Development Fund.

Nadezda Spiridovska research project No.1.1.1.2/VIAA/1/16/075 "Non-traditional regression models in transport modelling".

References

1. Ching, W.K., Huang, X., Ng, M.K. and Siu, T.K. Markov chains: models, algorithms and applications (Vol. 189). Springer Science & Business Media, 2013.
2. Alexander M. Andronov, Nadezda Spiridovska. Markov-Modulated Linear Regression. In proceedings' book: International conference on Statistical Models and Methods for Reliability and Survival Analysis and Their Validation (S2MRSA), Bordeaux, France, pp.24–28., arXiv:1901.09600v1., 2012.
3. Bladt M., Nielsen B.F. Markov Additive Processes. In: Matrix-Exponential Distributions in Applied Probability. Probability Theory and Stochastic Modelling, vol 81. Springer, Boston, MA, 2017.
4. Pacheco, A., Tang, L.C., Prabhu, N.U.: Markov-Modulated Processes & Semiregenerative Phenomena. World Scientific, New Jersey – London, 2009.
5. Github repositorium, address: <https://github.com/NadezdaSpiridovska/MMLR>.
6. Spiridovska, N.: A Quasi-Alternating Markov-Modulated Linear Regression: Model Implementation Using Data about Coaches' Delay Time. International journal of circuits, systems and signal processing 12, 617–628, 2018.
7. Spiridovska, N., Yatskiv (Jackiva), I.: Public transport passenger flow analysis and prediction using alternating Markov-modulated linear regression. In 29th European Conference on Operational Research (Euro2018) handbook, p. 208, 2018.
8. Xu, G., Zong, Y., Yang, Z.: Applied Data Mining 1st Edition, 2013.
9. McGurrin, Michael. Big Data and ITS. Noblis, Inc., 2013.
10. Irina Jackiva (Yatskiv), Nadezda Spiridovska. Data Preparation Framework Development for Markov-Modulated Linear Regression Analysis. In: Kabashkin I., Yatskiv I., Prentkovskis O. (eds) Reliability and Statistics in Transportation and Communication. RelStat 2018. Lecture Notes in Networks and Systems, Springer, Cham (Scopus) DOI: 10.1007/978-3-030-12450-2_17, 2019.
11. Andronov, A. "Statistical estimation of parameters of Markov-modulated linear regression." In Статистические методы оценивания и проверки гипотез, pp. 163-180. 2012 (In Russian).
12. Rao C.R. Linear statistical inference and its applications. – New-York – London-Sydney: John Wiley&Sons Inc. – 1995.
13. Srivastava M.S. Methods of Multivariate Statistics. - New-York: Wiley-Interscience.- 2002.
14. Turkington D.A. Matrix Calculus and Zero-One Matrices. Statistical and Econometric Applications. – Cambridge: Cambridge University Press. – 2002.
15. J. P. C. Kleijnen, "Regression models and experimental designs: A tutorial for simulation analysts," 2007 Winter Simulation Conference, Washington, DC, pp. 183-194, 2007.

16. Kijima, Masaaki. Markov processes for stochastic modeling. Vol. 6. CRC Press, 1997.

The Potential of Social Impacts of Biorefinery on Blue Economy

Natalia Stepanova¹

¹ V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences ,
Moscow, Russia
(E-mail: natalia0410@rambler.ru)

Abstract. The Blue Economy aims to balance sustainable economic benefits with long-term ocean health, in a manner which is consistent with sustainable development and its commitment to intra- and inter-generational equity. The term has also been used to give greater recognition to the many, though often not priced, ocean values ranging from cultural worth and village-based subsistence economies, to commercial and industrial commodities. Under this definition, not all ocean-based activities are consistent with the Blue Economy concept, because many ocean activities are not sustainable. Blue Economy, significantly advances practice beyond existing sustainable development frameworks. A proliferation in terms adds more complexity to an already challenging management space. Nevertheless, the conceptual framework is useful for structuring evaluations of practice, and helping to reveal missing ingredients necessary for the sustainable development of oceans. This paper examines the Blue Economy concept as an analytical frame for assessing initiatives aimed at achieving sustainable oceans development and management, with a particular focus on marine biorefinery as an example of an important sector within a Blue Economy. The policy implications of a rapidly evolving Blue Economy, across multiple sectors, are highlighted.

Keywords: integrated supply chain, harvesting technologies, high value products.

1 Introduction

The General concept of the modern society development is based on the increased consumption, production and accumulation. This leads to qualitative changes, but the negative consequences of uncontrolled growth are becoming more serious. In fact, it forms the basis of a potential global crisis. In addition to the existing production system, nature does not fit into the established order, environmental damage is increasing.

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



As for oceans, they play a significant role for economic development (fish provide a critical source of protein for international markets and coastal communities), and that is why they have a high state and private interest (Abbott, et al., 2014), (Campbell, Gray, Fairbanks, Silver, & Gruby, 2013). In the oceans, humans are now carrying out all of the activities, which traditionally associate with terrestrial industrial uses and processes (Golden, et al., 2017). At the same time, ocean based economic activities create the negative externalities: overfishing (Jackson, 2001), degradation of water quality (Peters, 2000), pollution (Abel, 2014), etc. Moreover, current industrial fishing practices result in the waste of millions of tons of valuable resources every year (Kelleher, 2005). At the same time, global oceans governance is too fragmented and the high seas spaces and resources are at risk for ecological decline and inequitable allocation (Ban, et al., 2013), (Veitch, et al., 2012). An important quantity of low-value marine biomass has to be managed in an efficient manner to avoid its waste (Balina, Romagnoli, & Blumberga, 2017), (Antelo, de Hijas-Liste, Franco-Uría, Alonso, & Pérez-Martín, 2015).

2 The roles of marine biorefinery

Blue Economy is a new economy concept for delivering sustainability (Mulazzani & Malorgio, 2017), (Pauli, 2010). According to the World Bank, the blue economy is the "sustainable use of ocean resources for economic growth, improved livelihoods and jobs, while preserving the health of ocean ecosystem." It is one of potential "sustainable business models", which is important in driving and implementing corporate innovation for sustainability (Bocken, Short, Rana, & Evans, 2014). The Blue Economy concept seeks to marry ocean-based development opportunities with environmental stewardship and protection. Small scale marine fisheries: artisanal, subsistence and recreational fisheries can be a good example for Blue Economy practice (Pauly, 2018). The integration of different biomass conversion processes to produce energy and value added products into a single facility is called a biorefinery (Bikker, van Krimpen, van Wikselaar, & al., 2016). In this paper we focus on

the role of algae biorefinery, which has a great potential for the Blue Economy. Ocean algae species from seaweeds to planktons play a vital role in carbon capture, producing up to 90 percent of the planet's oxygen. For example, macroalgae cultivation along coastlines could sequester about 1 billion tons of carbon annually (Chung, Beardall, Mehta, Sahoo, & Stojkovic, 2011). On the other hand, macroalgae can provide a sustainable alternative source of biomass for food, feeds, fuel and chemicals generation.

3 For an equitable access to marine biorefinery resources

Today fossil resources and their derivatives are used in all productive sectors of the economy. Unfortunately, fossil fuel extraction, processing, and end product uses are involved in numerous negative environmental impacts (Smith, 2005). So, we need to develop new sources of energy and materials, which will support long-term development of a human civilization while preserving ecosystems and their biodiversity (Ghisellini, Cialani, & Ulgiati, 2016). Biofuel is seen to be a good alternative for fossil fuel (Demirbas, 2008). The high carbohydrate content of macroalgae makes them suitable for bioconversion into biofuel, and even considered them as a “third or even fourth generation” biofuel feedstock (Lehahn, Kapilkumar, & Alexander, 2016). Over the years, many researchers have examined biofuel production from various types of macroalgae. They conclude that macroalgae do not contain as much phenolic material in the biofuels owing to the absence of lignin type materials but the fate of nitrogen and low quality are potentially problematic in using these as fuels (Chen, Zhou, Luo, Zhang, & Chen, 2015). Apart from biofuel, macroalgae have a potential for additional end uses (Ghadiryfar, Rosentrater, Keyhani, & Omid, 2016). Today they are widely used to co-produce food and high value chemicals (Chew, et al., 2017). That is why, macroalgae have recently attracted attention as a possible feedstock for biorefinery. Nevertheless, several major challenges should be taken into consideration for successful macroalgae economy (Palatnik & Zilberman, 2017). Water temperature, carbon dioxide, minerals and light are all important factors in cultivation of macroalgae, besides, different algae have

different specific requirements. Also the biomass productivity is the main constraint against being competitive with other energy and protein producing technologies (Seghetta, Hou, Bastianoni, Bjerre, & Thomsen, 2016). Another possible feedstock for biorefinery is microalgae. Because of their remarkably high growth rates compared to most other photosynthetic organisms, the use of marginal land for cultivation, valuable commercial cellular compounds and potential high lipid accumulation, microalgae have been the focus of intense research in the last 50 years. It is estimated that more than 50,000 species of microalgae exist, but only a limited number, have been studied and analyzed (Richmond, 2004).

This large variety of different microalgae makes it possible to be selected for use in a broad diversity of applications, such as value added products for pharmaceutical purposes, food crops for human consumption and as energy source. The microalgae species are most used for biofuels production (Ahmad, Yasin, Derek, & Lim, 2011). Microalgae can be converted directly into energy, such as biodiesel, and therefore appear to be a promising source of renewable energy. Moreover, the utilization of microalgae for biofuels production can also serve other purposes: reducing of CO₂, wastewater treatment (Wang, Li, Wu, & Lan, 2008), etc. Depending on the microalgae species other compounds may also be extracted, with valuable applications in different industrial sectors, including a large range of fine chemicals and bulk products, such as fats, polyunsaturated fatty acids, oil, natural dyes, sugars, pigments, antioxidants, high-value bioactive compounds, and other fine chemicals and biomass (Raja, Hemaiswarya, Kumar, Sridhar, & Rengasamy, 2008), (Rizwan, Mujtaba, Memon, Lee, & Rashid, 2018). The main advance of the biorefinery approach is the combination of processes and technologies for conversion of biomass to value added products with a minimal amount of waste. Another important advantage is renewable resource utilization. Besides, the biomass is grown in the sea does not compete for land, nutrients and fresh water. So, the biorefinery approach is considered sustainable (D'Amato, Droste, Allen, & al., 2017), and indeed could play an important role for the Blue Economy and sustainability (Philp, 2018).

However, to make the implementation of these new technologies possible, we need that both socioeconomic and environmental objectives will be considered simultaneously (Scarlat, Dallemand, Monforti-Ferrario, & Nita, 2015), (Schütte, 2018). Seaweed biomass is a renewable, but limited resource largely depended on weather conditions. That is why seaweed farming has evolved into a successful commercial endeavour in a number of tropical countries (Valderrama, Cai, Hishamunda, & Ridler, 2013). By the late 1960s, dwindling availability of wild seaweed stocks led to successful culturing in tanks. But these techniques soon proved to be economically unfeasible. The first seaweed farm was established in the south of the Philippines in 1969. And today, seasonality is one of the main causes of production fluctuations (van Hal, 2014). Diseases – is another major problem, which not only discourages farmers but also contributes to supply uncertainty for processors.

Conclusions

Today, the liquid biofuels production is the main implementation of biomass in different countries. Even the large-scale production of second-generation ethanol is already a reality. However, there are still some problems related to incomplete breakdown of biomass, incomplete sugars fermentation, production of undesirable by-products, and requirement of starter culture every new batch (Bell, 2017). For example, approximately 60 to 70% resultant solid fraction is considered today as waste while carrageenan extraction (Uju Wijayanta, Goto, & Kamiya, 2015). So, the cost-effective cultivation and dehydration difficulties currently limit the scale of microalgae technologies implementation.

Whatever the final application of microalgae, its production is based on the same principles as light availability, enough heat transfer and adequate control of culture parameters (Acién Fernández, Fernández Sevilla, & Molina Grima, 2013). Even though modeling for biorefineries is an active field of research (Santibanez-Aguilar, Morales-Rodriguez, Gonzalez-Campos, & Ponce-Ortega, 2015), (Seghetta, Hou, Bastianoni, Bjerre, & Thomsen, 2016), (Ling, et al., 2014), very few models have been

applied to marine biomass (Konda, Singh, Simmons, & Klein-Marcuschamer, 2015). Moreover, marine farming takes place in rural areas, where technologies to convert this biomass to chemicals and biofuels are not available (Ingle, et al., 2018). It is therefore important to analyze the demand for biomass in relation to the existing potential. The detailed assessment of the integrated supply chain that includes cultivation and harvesting technologies, conversion processes to fuels and high value products is performed in (Palatnik, Freer, Zilberman, Golberg, & Levin, 2018).

References

- 1 Abbott, J., Anderson, J. L., Campling, L., Hannesson, R., Havice, E., Lozier, S., & Wilberg, M. J. (2014). Steering the global partnership for oceans. *Marine Resource Economics*, 29(1), 1-16.
- 2 Abel, P. (2014). *Water Pollution Biology*. Taylor & Francis Ltd.
- 3 Acién Fernández, F. G., Fernández Sevilla, J. M., & Molina Grima, E. (2013). Photobioreactors for the production of microalgae. *Reviews in Environmental Science and Bio/Technology*, 12(2), 131–151.
- 4 Ahmad, A. L., Yasin, N. H., Derek, C. J., & Lim, J. K. (2011). Microalgae as a sustainable energy source for biodiesel production: A review. *Renewable and Sustainable Energy Reviews*, 15(1), 584–593.
- 5 Antelo, L. T., de Hijas-Liste, G. M., Franco-Uría, A., Alonso, A. A., & Pérez-Martín, R. I. (2015). Optimisation of processing routes for a marine biorefinery. *Journal of Cleaner Production*, 104, 489–501.
- 6 Balina, K., Romagnoli, F., & Blumberga, D. (2017). Seaweed biorefinery concept for sustainable use of marine resources. *Energy Procedia*, 128, 504–511.
- 7 Ban, N. C., J, B. N., Gjerde, K. M., Devillers, R., Dunn, D. C., Dunstan, P. K., & Halpin, P. N. (2013). Systematic Conservation Planning: A Better Recipe for Managing the High Seas for Biodiversity Conservation and Sustainable Use. *Conservation Letters* 7(1), 41-54.
- 8 Bell, G. (2017). Second-Generation Biorefineries: Optimization, Opportunities, and Implications for Australia. *Industrial Biotechnology*, 13(2), 76–84.
- 9 Bikker, P., van Krimpen, M. M., van Wikselaar, P., & al., e. (2016). Biorefinery of the green seaweed *Ulva lactuca* to produce animal feed, chemicals and biofuels. *Journal of Applied Phycology*, 3511–3525.
- 10 Bocken, N. M., Short, S. W., Rana, P., & Evans, S. (2014). A literature and practice review to develop sustainable business model archetypes. *Journal of Cleaner Production*, 65, 42–56.

- 11 Campbell, L. M., Gray, N. J., Fairbanks, L. W., Silver, J. J., & Gruby, R. L. (2013). Oceans at Rio þ 20. 3) *Campbell, L. M., Gray, N. J., Fairbanks, L. W., Silver, J. Conservation Letters*, 6(6), 439–447.
- 12 Chen, H., Zhou, D., Luo, G., Zhang, S., & Chen, J. (2015). Macroalgae for biofuels production: Progress and perspectives. *Renewable and Sustainable Energy Reviews*, 47, 427–437.
- 13 Chew, K. W., Yap, J. Y., Show, P. L., Suan, N. H., Juan, J. C., & Ling, T. C. (2017). Microalgae biorefinery: High value products perspectives. . *Bioresource Technology*, 229, 53–62.
- 14 Chung, I., Beardall, J., Mehta, S., Sahoo, D., & Stojkovic, S. (2011). Using marine macroalgae for carbon sequestration: a critical appraisal. *J. Appl. Phycol.* 23, 877–886.
- 15 Cook, J., Oreskes, N., Doran, P., Anderegg, W., Verheggen, B., Maibach, E., & al., e. (2016). Consensus on consensus: a synthesis of consensus estimates on human-caused global warming. *Environmental Research Letters*, 11, 048002.
- 16 D’Amato, D., Droste, N., Allen, B., & al., e. (2017). Green, circular, bio economy: A comparative analysis of sustainability avenues. *Journal of Cleaner Production*, 168, 716–734.
- 17 Demirbas, A. (2008). Biofuels sources, biofuel policy, biofuel economy and global biofuel projections. *Energy Conversion and Management*, 49(8), 2106–2116.
- 18 Ghadiryanfar, M., Rosentrater, K. A., Keyhani, A., & Omid, M. (2016). A review of macroalgae production, with potential applications in biofuels and bioenergy. *Renewable and Sustainable Energy Reviews*, 54, 473–481.
- 19 Ghisellini, P., Cialani, C., & Ulgiati, S. (2016). A review on circular economy: the expected transition to a balanced interplay of environmental and economic systems. *Journal of Cleaner Production*, 114, 11–32.
- 20 Golden, J. S., Virdin, J., Nowacek, D., Halpin, P., Bennear, L., & Patil, P. G. (2017). Making sure the blue economy is green. *Nature Ecology & Evolution* 1(2).
- 21 Ingle, K., Vitkin, E., Robin, A., Yakhini, Z., Mishori, D., & Golberg, A. (2018). Macroalgae Biorefinery from *Kappaphycus alvarezii*: Conversion Modeling and Performance Prediction for India and Philippines as Examples. *BioEnergy Research* 11, 22-32.
- 22 Jackson, J. B. (2001). Historical Overfishing and the Recent Collapse of Coastal Ecosystems. *Science*, 293(5530), 629–637.
- 23 Kamm, B., & Kamm, M. (2004). Principles of biorefineries. *Applied Microbiology and Biotechnology* 64, 137–145.
- 24 Kelleher, K. (2005). *Discards in the World's Marine Fisheries: an Update*. Rome, Italy: FAO Fisheries Technical Paper No. 470. Food and Agricultural Organization of the United Nations.

- 25 Konda, N., Singh, S., Simmons, B., & Klein-Marcuschamer, D. (2015). An investigation on the economic feasibility of macroalgae as a potential feedstock for biorefineries. *Bioenergy Res* 8, 1046– 1056.
- 26 Lehahn, Y., Kapilkumar, N. I., & Alexander, G. (2016). Global potential of offshore and shallow waters macroalgal biorefineries to provide for food, chemicals and energy: feasibility and sustainability. *Algal Research* 17, 150–160.
- 27 Ling, T., Eric, C., Robert, M., Min, Z., Andy, A., & Xin, H. (2014). Techno-economic analysis an life-cycle assessment of cellulosic isobutanol and comparisons with cellulosic ethanol and n-butanol. . *Biofuels Bioprod Biorefin* 8, 30–48.
- 28 Mulazzani, L., & Malorgio, G. (2017). Blue growth and ecosystem services. *Marine Policy*, 85, 17–24.
- 29 Palatnik, R. R., & Zilberman, D. (2017). Economics of Natural Resource Utilization - the Case of Macroalgae. *Springer Proceedings in Mathematics & Statistics*, (pp. 1–21).
- 30 Pauli, G. A. (2010). *The blue economy: 10 years, 100 innovations, 100 million jobs*. Paradigm Publications.
- 31 Pauly, D. (2018). A vision for marine fisheries in a global blue economy. *Marine Policy*, 87, 371–374.
- 32 Peters, N. E. (2000). Water Quality Degradation Effects on Freshwater Availability:Impacts of Human Activities. . *Water International*, 25(2), 185–193.
- 33 Philp, J. (2018). The bioeconomy, the challenge of the century for policy makers . *New Biotechnology*, 40, 11–19.
- 34 Raja, R., Hemaiswarya, S., Kumar, N. A., Sridhar, S., & Rengasamy, R. (2008). A Perspective on the Biotechnological Potential of Microalgae. *Critical Reviews in Microbiology*, 34(2) , 77–88.
- 35 Richmond, A. (2004). *Handbook of microalgal culture: biotechnology and applied phycology*. Blackwell Science Ltd.
- 36 Rizwan, M., Mujtaba, G., Memon, S. A., Lee, K., & Rashid, N. (2018). Exploring the potential of microalgae for new biotechnology applications and beyond: A review. *Renewable and Sustainable Energy Reviews*, 92, 394–404.
- 37 Santibanez-Aguilar, J., Morales-Rodriguez, R., Gonzalez-Campos, J., & Ponce-Ortega, J. (2015). Stochastic design of biorefinery supply chains considering economic and environmental objectives. . *J Clean Prod.* .
- 38 Scarlat, N., Dallemand, J.-F., Monforti-Ferrario, F., & Nita, V. (2015). The role of biomass and bioenergy in a future bioeconomy: Policies and facts. *Environmental Development*, 15, 3–34.
- 39 Schütte, G. (2018). What kind of innovation policy does the bioeconomy need? *New Biotechnology*, 40, 82–86.
- 40 Seghetta, M., Hou, X., Bastianoni, S., Bjerre, A.-B., & Thomsen, M. (2016). Life cycle assessment of macroalgal biorefinery for the production of ethanol, proteins and fertilizers – A step towards a

- regenerative bioeconomy. *Journal of Cleaner Production*, 137 , 1158–1169.
- 41 Silver, J. J., Gray, N. J., Campbell, L. M., Fairbanks, L. W., & Gruby, R. L. (2015). Blue Economy and Competing Discourses in International Oceans Governance. *The Journal of Environment & Development*, 24(2), 135–160.
- 42 Smith, K. (2005). *Powering our future: an energy sourcebook for sustainable living*. . New York: iUniverse.
- 43 Uju Wijayanta, A., Goto, M., & Kamiya, N. (2015). Great potency of seaweed waste biomass from the carrageenan industry for bioethanol production by peracetic acid-ionic liquid pretreatment. *Biomass Bioenergy* 81, 63–69.
- 44 Valderrama, D. C. (2013). *Social and economic dimensions of carrageenan seaweed farming*.
- 45 Valderrama, D., Cai, J., Hishamunda, N., & Ridler, N. (2013). Social and economic dimensions of carrageenan seaweed farming. *Fisheries and Aquaculture Technical Paper No. 580. Rome, FAO. 204 pp.*, 61-89.
- 46 van Hal, J. W.-C. (2014). Opportunities and challenges for seaweed in the biobased economy. *Trends in biotechnology*, 32(5)., 231-233.
- 47 Veitch, L., Dulvy, N. K., Koldewey, H., Lieberman, S., Pauly, D., & Roberts, C. M. (2012). Avoiding Empty Ocean Commitments at Rio+20. *Science*, 336(6087), 1383–1385.
- 48 Wang, B., Li, Y., Wu, N., & Lan, C. Q. (2008). CO₂ bio-mitigation using microalgae. *Applied Microbiology and Biotechnology*, 79(5), 707–718.

Improving Stepwise Logistic Regression Using a SAS Macro

Jian Sun^{1,2}

¹ School of Public Health, University of Alberta, Edmonton, Canada
(E-mail: Jsun9@ualberta.ca; sunjian@hotmail.com)

² Department of Medicine, University of Calgary, Calgary, Canada

Abstract. Stepwise covariate selection is a popular method for multivariable regression model building. Based on the different significance levels pre-specified by statisticians, different covariates are included in the model. Further analyses with these models might introduce biases. This paper proposes a novel method to select covariates for stepwise logistic regression without pre-setting a significance level. Multiple models containing different numbers of covariates were outputted for final model selection. A user-oriented SAS macro was developed. Users of the macro may determine the final models, based on estimated characteristic changes of the overall models, the variances of the covariate effects on the response variable and their special needs. With this method, model selections are much easier than with purposeful or the best subsets method. This method improved stepwise covariate selection processes. Broad applications are expected.

Keywords: logistic regression, model building, multivariate statistics, SAS Programming, Statistical computation.

1 Introduction

Logistic regression is the most frequently used statistical model for the analysis of data with a discrete outcome in various research areas. If the data consist of a few covariates, there may be only one best model, which is easy to build. However, if covariates increase, multiple “best” models may exist. Building a best model with data consisted of numerous covariates is difficult.

What is the best model? A best regression model should be a model with the correct covariates and the most precise estimates for them. As Hosmer et al. described, “The traditional approach to statistical model building involves seeking the most parsimonious model that still accurately reflects the true outcome experience of the data” [1]. A model with fewer covariates is numerically stable and easier to use. In contrast, the more covariates included in a model the greater the standard errors of the coefficients and the wider the confidence intervals of the corresponding odds ratios.

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



Stepwise selection is a popular and effective statistically driven method to reduce covariates of multivariable model including logistic regression. In each stage of the selection procedure, a covariate is added to (or subtracted from) the set of covariates based on a pre-specified significance level by the statistician. With different significance levels, different models are determined. Further analyses with these models might introduce biases.

This paper proposes a novel method to select covariates for backward logistic regression without pre-setting a significance level. A user-oriented SAS macro was developed. To validate the proposed covariate selection method and the macro, a randomly generated hypothetical dataset was used.

2 Material and methods

The proposed method is an improvement of the traditional backward elimination method. All covariates were included in the model first. SAS LOGISTIC procedure was used to remove covariates one at a time. The process was repeated until only one covariate was left in the model. Multiple models contained different numbers of covariates were outputted for final model selection. The detail of the method is described by means of the macro.

Table 1 is a SAS macro for data generation. The first DATA step (lines 2-8) generated nine uniform distributed probabilities (p s) ranged from zero to one. The second DATA step (lines 11-23) generated nine random samples as covariates x_1 , x_2 , ..., x_8 and response variable y with 100 observations from Bernoulli distribution with the variant p s derived from the first DATA step. Seeds for both RAND functions were set to "2".

Table 1. SAS macro for data generation

1	%macro case(seed=2,obs=100,cov=8);	14	array x{%eval(&cov+1)} x1-x{%eval(&cov+1)};
2	data uniform(keep=p);	15	call streaminit(&seed);
3	call streaminit(&seed);	16	do i=1 to &obs;
4	do i=1 to %eval(&cov+1);	17	do j=1 to %eval(&cov+1);
5	p=rand("Uniform");	18	x(j)=rand("Bernoulli",p(j));
6	output;	19	end;
7	end;	20	output;
8	run;	21	end;
9	proc transpose data=uniform out=p prefix=p;	22	rename x{%eval(&cov+1)}=y;
10	run;	23	run;
11	data Bernoulli (keep=x1-x&cov y);	24	%mend;
12	set p;	25	%case();
13	array p{%eval(&cov+1)} p1-p{%eval(&cov+1)};		

To determine whether there were any complete or quasi-complete separations [2, 3], missing values, or all '0' or all '1' variables in the dataset, a LOGISTIC procedure for the model containing all variables in the dataset was performed. After

evaluating the data, the Macro Variate (Table 2) with file name Bernoulli was invoked.

Table 2. SAS macro for model building

1	%macro variate(mydata=Bernoulli);	53	if max=1 then do;
2	data _null_;	54	call symput(variable,'');
3	set &mydata;	55	delete;
4	column=n(of _all_)-1;	56	end;
5	call symput('num',put(column,2.));	57	if variable='_2LogL' then variable='-2 Log L';
6	run;	58	run;
7	%macro x;	59	%end;
8	%do i=1 %to #	60	%do k=2 %to #
9	%global x&i;	61	data compare&k(drop=estimate1 ProbChiSq1);
10	%let x&i=x&i;	62	merge output1_2 output&k._1(in=a drop=max);
11	%end;	63	by variable;
12	%mend x;	64	if a;
13	%macro names;	65	delta&k= 100 *(estimate2- estimate1)/estimate1;
14	%do i=1 %to #	66	rename estimate2=estimate&k
15	&&x&i		ProbChiSq2=ProbChiSq&k;
16	%end;	67	if variable='-2 Log L' then delta&k=estimate2;
17	%mend names;	68	run;
18	%macro logistic;	69	%end;
19	%x;	70	data output1_3;
20	%do j=1 %to #	71	set output1_1(drop=max);
21	ods select none;	72	rename estimate2=estimate1
22	ods output ParameterEstimates=_pe		ProbChiSq2=ProbChiSq1;
	FitStatistics=_fit;	73	run;
23	proc logistic data=&mydata;	74	%macro names2;
24	model y(event='1')=%names;	75	%do m=2 %to #
25	run;	76	compare&m
26	ods output close;	77	%end;
27	proc sql;	78	%mend names2;
28	create table output1 as	79	data delta;
29	select variable, estimate, ProbChiSq	80	merge output1_3 %names2;
30	from pe	81	by variable;
31	where variable^='Intercept'	82	if variable='-2 Log L' then variable='D';
32	order by ProbChiSq desc;	83	temp=input(substr(variable,2,3),3.);
33	quit;	84	run ;
34	data output2;	85	%mend logistic;
35	set output1;	86	%logistic ;
36	max=0;	87	proc sort data=delta out=delta1(drop=temp);
37	if n=1 then max=1;	88	by temp;
38	run ;	89	run ;
39	proc sql;	90	ods select all;
40	create table output&j._1 as	91	proc print data=delta1 noobs;
41	select variable, estimate as estimate2,	92	title 'Table 3. Delta-beta-hat-percent between full
	ProbChiSq as ProbChiSq2, max		and reduced model';
42	from output2	93	var variable delta;
43	union all	94	where variable^='D';
44	select Criterion as variable,	95	run ;
	InterceptAndCovariates as estimate2,	96	proc print data=delta1 noobs;
	as ProbChiSq2, 0 as max	97	title 'Table 4. P-values of the Wald test;
45	from fit	98	var variable ProbChiSq;
46	where Criterion='-2 Log L'	99	where variable^='D';
47	order by variable;	100	run ;
48	quit;	101	proc print data=delta1 noobs;
49	data output&j._2;	102	title 'Table 5. Deviance and coefficient estimate';
50	set output&j._1;	103	var variable estimate;
51	rename ProbChiSq2=ProbChiSq1	104	run ;
	estimate2=estimate1;	105	%mend variate;
52	if variable='-2 Log L' then variable='_2LogL';	106	%variate ();

LOGISTIC procedure (lines 23-25) estimated the coefficient and the p -value of the Wald statistic [1, 4] for each covariate. The least significant effect (i.e. a covariate with the largest p -value) was removed. For the remaining covariates, the LOGISTIC estimated their effects, and the covariate with the largest p -value was removed again. The process was repeated until only one covariate was left in the model.

DATA step Compare&k (lines 61-68) merged the full model with each reduced model, and calculated the *delta-beta-hat-percent* [1]

$$\Delta\hat{\beta}_{ik}\% = 100 * (\hat{\beta}_{i1} - \hat{\beta}_{ik}) / \hat{\beta}_{i1}, i = 1, 2, \dots, 8, k = 2, 3, \dots, 8, \quad (1)$$

where $\hat{\beta}_{i1}$ and $\hat{\beta}_{ik}$ are the estimated coefficient of the covariate x_i in the full model and the k th reduced model, respectively. In this paper, a full model means a model containing all covariates in the dataset rather than that defined by McCullagh and Nelder [5]. It is not a saturated model, in which there are as many estimated parameters as observations [6]. The reduced model is the model obtained by setting certain parameters in the full model equal to zero [7].

DATA step Delta (lines 79-84) merged the deviances $D = -2\text{Log}L$ [1], the *delta-beta-hat percent*, p -values and coefficient estimates of each covariate in the full and the reduced models. L is the likelihood of a fixed model. PRINT procedures (lines 91-104) printed them out for final model selection. Each column except the “Variable” column in any of the three output tables represented a model. They were named model 1, model 2, ..., and model 8. They are in descending order of the number of covariates from left to right.

Based on the outputs of the macro, the process of final model selection, which had three steps, was performed:

(1) Observe p -values of each covariate in each model. Choose the smallest model that contains at least one covariate with $p < 0.05$. If the statistician wants to include some important covariates, choose the smallest model containing those covariates.

(2) Check *delta-beta-hat-percent*, $\Delta\hat{\beta}_{ik}\%$ for the covariates in the model with $p < 0.05$. If any $|\Delta\hat{\beta}_{ik}\%| > 20\%$ or another criterion pre-set based on special requirement, shift one column left in the output tables and repeat this step, until arriving at a column that does not include any $|\Delta\hat{\beta}_{ik}\%|$ which is larger than the pre-set criterion for the covariates with $p < 0.05$.

(3) Compare deviance $D = -2\log L$ for each model with χ^2 in the Chi-Squared distribution table for $p = 0.05$. If any $D > \chi^2$ with the same degree of freedom (DF), shift one column left and then compare D with χ^2 again with the reduced DF . The comparisons will repeat until we find a model with $D < \chi^2$ and then return to Step 2. The iteration between steps 2 and 3 continues until a model that meets both $\Delta\hat{\beta}_{ik}\%$ and D requirements is found.

An alternative method for step 3 is to conduct the likelihood ratio test [7]:

$$LR_k = -2\text{Log}L_k - (-2\text{Log}L_{k-1}), k = 2, 3, \dots, 8, \quad (2)$$

where L_k and L_{k-1} are the likelihoods of the models k and $k-1$, respectively. Because deviance $D = -2\log L$, which provided by LOGISTIC directly, and $-2\log L$ cannot be used as a SAS variable name, D was used for output instead of $-2\log L$.

Although the process described above appears complicated, the steps can be completed within minutes.

To validate the methodology and the macro, the best subsets regression [1] was performed with the same dataset using a SAS code created by King [8]. The method and macro were also used for a project to detect factors associated with recommendations for rare disease drug in Canada [9].

All analyses were performed using SAS 9.4 (SAS Institute Inc., Cary, NC).

3 Results

Table 3 shows the *delta-beta-hat-percent* $\Delta\hat{\beta}_{ik}\%$ between the full model and each reduced model. *delta2*, *delta3*, ..., *delta8* are the variable names of $\Delta\hat{\beta}_{ik}\%$. Tables 4 and 5 show *p*-values (*ProbChiSq1*, *ProbChiSq2*, ..., and *ProbChiSq8*) of the Wald tests and the coefficient estimates (*estimate1*, *estimate2*, ..., and *estimate8*) of all covariates in each model, respectively. Deviance (D), a criterion rather than a variable is included in Table 5 for convenience.

Table 3. Delta-beta-hat-percent between full and reduced model

Variable	delta2	delta3	delta4	delta5	delta6	delta7	delta8
x1	4.67241
x2	0.60020	0.4229	7.21031	13.4218	18.0088	10.5127	.
x3	3.56317	13.9809
x4	-0.71575	-0.5464	0.86465	-2.1822	-3.2992	.	.
x5	0.41012	1.9325	-7.05382
x6	0.42983	-2.0565	-5.36167	-2.3109	.	.	.
x7	-0.07275	-0.6350	0.81951	1.2553	-3.6413	-10.9315	-18.3892
x8

Table 4 shows that in any of the eight models only $x7$ is significant ($p < 0.05$). Considering the definition of the best model, i.e. smaller is better if no other significant effect changes, the model with only one covariate $x7$ was chosen first and

its *delta-beta-hat-percent* $\Delta\hat{\beta}_{78}\%$ ($= -18.3892$) in the column *delta8* of Table 3 was checked. If $|\Delta\hat{\beta}_{ik}\%| \leq 20\%$ was used as a criterion, this univariate model would be the final model. If the 18% coefficient change were not satisfied, this model would not be chosen. Instead, if 15% coefficient changes were accepted, *x2* would be added to the model. More rigorously, if even a 10% variation for the coefficient of *x7* were not permitted, $|\Delta\hat{\beta}_{ik}\%| \leq 10\%$ would be set and the three covariate model with *x2*, *x4* and *x7* would be chosen. It would not matter that the value of *delta6* for *x2* increased to 18.0088% $>10\%$, because *x2* had no significant effect ($p = 0.20086$) on the outcome. Although *x2* had no direct significant effect on outcome, combining with *x4*, it adjusted the coefficient of *x7* from 1.39711 to 1.64959. The odds ratio of *x7* increased from 4.044 to 5.205 (not shown in the table).

Table 4. *P*-value of the Wald test

Variable	ProbChiSq1	ProbChiSq2	ProbChiSq3	ProbChiSq4	ProbChiSq5	ProbChiSq6	ProbChiSq7	ProbChiSq8
x1	0.90516	0.90072
x2	0.30450	0.30100	0.30185	0.25468	0.22172	0.20086	0.22428	.
x3	0.81040	0.80319	0.77825
x4	0.35729	0.35958	0.35939	0.35271	0.36482	0.36932	.	.
x5	0.65400	0.65237	0.64712	0.67263
x6	0.56257	0.56116	0.56824	0.58055	0.56805	.	.	.
x7	0.00943	0.00957	0.00944	0.00787	0.00731	0.00808	0.01114	0.016159
x8	0.91743

Table 5. Deviance and coefficient estimate

Variable	estimate1	estimate2	estimate3	estimate4	estimate5	estimate6	estimate7	estimate8
D	75.6505	75.6615	75.6767	75.7538	75.9349	76.2615	77.1366	78.6116
x1	0.1535	0.1606
x2	0.6631	0.6671	0.6659	0.7109	0.7521	0.7825	0.7328	.
x3	0.2364	0.2448	0.2694
x4	0.6900	0.6850	0.6862	0.6959	0.6749	0.6672	.	.
x5	0.2857	0.2868	0.2912	0.2655
x6	-0.3598	-0.3614	-0.3524	-0.3405	-0.3515	.	.	.
x7	1.7119	1.7107	1.7011	1.7260	1.7334	1.6496	1.5248	1.3971
x8	0.1203

If n (= 100 here) is the observation number and q (= 8 here) is the covariate number in the full model, the degree of freedom (DF) for the deviance test is $n - q = 92$. In Table 5, the D for each model is smaller than 80, while χ^2 with $DF = 92$ is 115.39 for $p = 0.05$. For reduced models, the DF s for the deviance tests are larger than 92 and the corresponding $\chi^2 > 115.39$. Therefore, all models here have no significant differences from the saturated model at significance level 0.05.

The result of the likelihood test was the same as that of the deviation test. The difference of D s between any two adjacent columns in Table 5 was much smaller than 3.84, the critical value of the Chi-Square statistic with $DF = 1$. Therefore, removing any covariates except x_7 did not change the overall model significantly at $p = 0.05$ level.

If it were important to keep covariate x_4 , for example, in the model, the iteration should be started from the model containing x_2 , x_4 and x_7 for the model selection process.

Finally, the result of the best subsets regression (not shown) verified that among all possible models with the same number of covariates, all reduced models generated by the macro with the hypothetical data had the smallest Mallows' C_p [1,10].

The method and macro were used for the project to detect factors associated with recommendations for rare disease drug coverage in Canada [9]. The real data consist of 92 observation and 15 covariates. All variables were binary. Based on the proposed method, five covariates were remained in the model. Although they were the same as using stepwise method with that the significance levels for entry and stay were set at 0.2 in chance, with the proposed method, the coefficient changes (< 20%) of significant covariates in the model were derived. The overall model did not change significantly at $p = 0.05$ level after 10 covariates were removed [9].

4 Discussion

4.1 Why a new method?

The SAS LOGISTIC procedure is a convenient tool for us to perform stepwise regressions. Statisticians only need to specify significance levels for variable entry and/or stay, then LOGISTIC will automatically select covariates to build the model for us. However, the significance levels are set arbitrarily. The arbitrariness may result in bias for the following analyses. Because of different significance levels, different sets of covariates will remain in the fixed models. If the significance level were too low, important effects would be missed. In contrast, if the significance level were set too high, some covariates, which have little effect on the outcome would be included in the model, diluting the important relationships between other covariates and the outcome.

A remedial strategy is using a higher significance level to run the LOGISTIC first, and then manually removing one covariate with the largest p -value and running the LOGISTIC again. This process will repeat until there are no covariates with p -values larger than the pre-set significance level in the model. Every time after removing a covariate, statisticians have to evaluate the characteristic changes of the remaining covariates and the overall model. The higher the significance level pre-set, the more covariates will be included. Therefore, the manual workload to eliminate redundant covariates will increase. While the significance level increases to 1.0, regardless of forward, backward, combined stepwise or purposeful selection methods, the selection process becomes a backward selection without a significant level. The idea of the proposed method came from this strategy.

4.2 What is new?

Traditional backward elimination regression needs a pre-set significance level to control the covariate selection process. The process stops when there are no covariates that meet the criteria for removal. The proposed method does not need a pre-set significance level. Therefore, the process will not stop until only one covariate remains in the model.

The *delta-beta-hat-percent* $\Delta\hat{\beta}\%$ proposed by Hosmer et al. [1] was calculated for each covariate in each reduced model. The numerators are the differences of the coefficient estimates between the full model and each reduced model [11]. As we have seen from Table 5, the estimates of the coefficients vary depending on the presence or absence of other covariates included in the model. However, the variations are not monotonic. Removing a covariate may increase or decrease other covariates' effects on outcome. During the process of the backward selection, even if at a stage a coefficient changed a lot, the direction of the change might toward to its initial value in the full model. In contrast, even if the change were minor at a stage of the process, the estimated coefficient might quite differ to its original value. Therefore comparing reduced models with the full model appears more reasonable.

A user-oriented macro was developed. To validate the proposed method and the macro, a hypothetical dataset was generated. For simplicity, all variables are binary. Referring to the three output tables, statisticians can decide their final models quickly and confidently. The decision is more flexible than with the traditional stepwise method. Statisticians can change their criteria and build different final models to meet their special requirements. Same as the best subsets method, the output of the macro provides multiple models for further selections. However, with the proposed method, the number of covariates included in the final model is easy to determine.

The results of deviance and likelihood tests are the same. If observation number is larger, the latter is more convenient. Instead of checking the Chi-Squared distribution table with larger DF , statisticians only need to compare the difference

between each pair of adjacent deviances D_s with 3.84, the critical value of Chi-Square statistic with $DF = 1$.

4.3 How to use the macro?

As with traditional logistic regressions, the data should not contain any variables with all “0” or all “1” values. Complete or quasi-complete separation should not occur. In addition, the data should not contain any variables with missing values. Because the LOGISTIC in the macro runs iteratively, the dataset should always be the same during the iteration. If the data contained a variable with missing values, the effective observation number will increase while this variable removes.

After data preparation, users need to change their variable names to y , x_1 , x_2 , x_3 , and so on, exclude variables not for the regression, and then input their file names to invoke the macro. The number of variables and observations are arbitrary.

Readers can also use the macro in Table 1 to generate a dataset with a different seed to validate the macro in Table 2. I chose seed = “2” for convenience. If you choose another seed number, you may need to deal with complete or quasi-complete separation before running the macro. For illustrative purposes, I set the covariate number to “8”. Because the number was relatively small, the characteristic changes of the overall model and the coefficient estimates during the process were not obvious.

After running the macro, statisticians can screen the output tables to select their final models. The macro focuses on the main effect model building. If statisticians want to identify interactions, they can add corresponding interactions to the fixed model and run the LOGISTIC procedure to build their final models with interactions easily.

In practice, one or more important covariates may have to stay in the model. Suppose that the purpose of a medical study is to compare two treatment effects. If the covariate represented the treatment were removed, the analysis would be meaningless. In this case, the statistician should select a final model containing the treatment from the macro outputs.

Conclusions

This paper proposed a novel method for stepwise logistic regression. A user-oriented SAS macro was included. With this method, model selection is much easier than with purposeful or the best subsets method. This method improved the stepwise covariate selection process. Broad applications are expected.

Acknowledgement

The author thanks Dr. Tania Stafinski, School of Public Health, University of Alberta, for the initial manuscript revision.

References

1. D. W. Hosmer, S. Lemeshow and R. X. Sturdivant. Applied logistic regression. 3rd ed., John Wiley & Sons, Inc., Hoboken, New Jersey, 2013.
2. A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71, 1, 1-10, 1984.
3. C. Rainey. Dealing with separation in logistic regression model. *Political Analysis*, 24, 339-355, 2016.
4. W. H. Greene. *Econometric analysis*, 6th ed., Prentice Hall, Boston, 2008.
5. P. McCullaph and J. A. Nelder. *Generalized linear model*, 2nd ed., Chapman and Hall, London, 1989.
6. A. Agresti. *Categorical data Analysis*, 3rd ed., John Wiley & Sons, Inc., Hoboken, New Jersey, 2013.
7. D. G. Kleinbaum. *Logistic regression: a self-learning text*, Springer-Verlag, 1994.
8. J. E. King. Running a best-subsets logistic regression: an alternative to stepwise methods. *Educational and Psychological Measurement*, 63, 392-403, 2003.
9. F. N. I. Nagase, T. Stafinski, Sun J, G. Jhangri and D. Menon. Factors associated with positive and negative recommendations for cancer and non-cancer drugs for rare diseases in Canada. *Orphanet Journal of Rare Diseases*, 2019; <https://doi.org/10.1186/s13023-019-1104-7>.
10. C. L. Mallows. Some comments on Cp. *Technometrics*, 15, 661-675, 1973.
11. Z. Bursac, C. H. Gauss, D. K. Williams and D. W. Hosmer. Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine*, 3, 17, 2008, (no page numbers).

Alcohol Consumption and Marital Status in the Czech Republic

Kornélia Svačinová¹, Markéta Pechholdová², and Jana Vrabcová³

¹ Department of Demography, Prague University of Economics and Business, Czech Republic (e-mail: k.csefalvaiova@seznam.cz)

² Department of Demography, Prague University of Economics and Business, Czech Republic (e-mail: marketa.pechholdova@seznam.cz)

³ Department of Statistics and Probability, Prague University of Economics and Business, Czech Republic (e-mail: vrabcova.jana@post.cz)

Abstract

Background: Associations between alcohol consumption and marital status (living alone/ living with a partner) are broadly discussed. The protective mechanism highlights the social integrative function of marriage and the role of social control over risk-taking behaviour. Selected studies have documented an association between marriage and lower alcohol consumption, but the evidence is not straightforward. **Aims:** This paper reviews documented interactions between marriage/ partnership in relation to alcohol consumption and adds empirical evidence on this relationship from Czechia, where alcohol consumption is common and well tolerated. **Data and methods:** Literature review and meta-analyses was conducted. Divorces due to alcohol were enumerated. Alcohol consumption was compared between those living with and without partner based on data from the Czech Household Panel Survey 2017 and the SHARE – Survey of Health, Ageing and Retirement in Europe. **Results:** Divorces due to alcohol decrease. Living with a partner is however associated with higher and riskier alcohol consumption in both sexes. Living alone is more linked with being an abstainer. **Conclusion:** Higher alcohol consumption among those with partner stems from complex nature of relationships. The present findings however need to be validated by further analyses on more robust data, such as death rates.

Keywords: Alcohol, Czech Republic, Marital Status, Loneliness, Partnership.

1 Introduction

Alcohol is one of the most commonly used psychoactive and dependence-producing substances in European countries. Problems it causes have not only individual, but also societal and family-wide repercussions. In many cultures, including Czech, the consumption of alcohol is not an isolated event. Society has developed a positive attitude to alcohol consumption and this position isn't optimal. It is woven into the fabric of marriage and family life. Through our shared experiences, we have developed expectations regarding whether and how drinking can have positive effects. There is a general belief that alcohol

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



enhances social situations and increases interpersonal warmth and intimacy. At the same time it increases conflicts, violence, and deprivation. Although alcohol is often associated with certain positive effects, excessive drinking and alcohol disorders can exert a negative effect on the marital development and on the development of children in the context of the family. It is very difficult to find relevant data for the case of Czechia regarding to evidence that alcohol influences and is influenced by marital/ partnership status. Crucial problem is that studies in certain key areas are missing.

2 Literature Review

Two competing hypotheses propose opposite effects for the relation between alcohol consumption and marital functioning. First hypothesis conceptualizes alcohol use as maladaptive and proposes that it serves as a chronic stressor that causes marital dysfunction and subsequent dissolution. An opposing hypothesis proposes that alcohol use is adaptive and temporarily relieves stressors that cause marital dysfunction, stabilizing the marital relationship, and perhaps preventing dissolution (Marshal, M.P., 2003).

Alcohol addiction associates with many indicators, such as age, gender, education, culture, religion, social, health, economic or family status. Impact of alcohol addiction on marriage and quality of marital life of the spouse living with individual with an alcohol addiction has remained an area of interest in research.

2.1 Alcohol and Marital Status

The term alcoholic marriage (McCrary and Epstein, 1995; Paolino and McCrary, 1977) is often used to describe a partnership in which one or both of the partners have a history of alcoholism that interferes with successful, day-to-day marital functioning.

Some existing studies have documented an association between marriage and lower alcohol consumption. As such, marriage has a protective effect in reducing the risk of an addictive substance, especially alcohol and drugs (e.g. Bachman, Wadsworth, O'Malley, Schulenberg and Johnston, 1997; Leonard and Rothbard, 1999). On the other hand, results from initial opponents' studies point to the increased alcohol consumption after marital breakup.

Marshal, M.P. (2003) review altogether 60 studies that tested the relation between alcohol use and one of three marital functioning domains (satisfaction, interaction, violence). Results provide overwhelming support for the notion that alcohol use is maladaptive, and that it is associated with dissatisfaction, negative marital interaction patterns, and higher levels of marital violence. A small subset of studies found that light drinking patterns are associated with adaptive marital functioning.

Saxena, et al. (2003). Authors have carried out a research study in which the study has compared two groups of families: at least one family member was

consuming alcohol and families without consuming alcohol) in an urban community in Delhi (India). The study clearly indicated that higher number of illnesses or injury in the previous year was reported by the non-drinking families. This is connected with the position, that families with drinkers were less likely to seek health services.

Brennan et al. (1994) conducted a prospective study of drinkers' wives with a problem of late life. In the initial assessment of 87 late-life drinkers, health and social functioning was worse and more reliant on cognitive coping strategies and more shared management of cognitive avoidance than 87 non-alcoholic drinkers.

Crisp and Barber (1995) studied about the deprivation experienced by the wives of alcoholics. They used the Drinker's Partner Distress Scale (DPDS). Two dimensions of alcohol induced problems i.e. depression and marital conflict are measured. Both sub-scales demonstrated internal consistency and predictions of convergent and discriminant validity were supported in relation to both sub-scales.

2.2 Alcohol and Loneliness

Social isolation and living alone are increasingly common in industrialised countries. Form of family cohabitation (in a marriage, living with a partner, living alone) is closely connected to the mental state, e.g. people living alone may according to many publications experience the effect of loneliness or feeling alone rather than married or cohabiting couples. There is consistent evidence linking social isolation and loneliness to worse cardiovascular and mental health outcomes. Less is known on the role of other conditions and wider socio-economic consequences (N. Leigh Hunt, 2017).

A systematic meta-analysis was undertaken to determine the associations between alcohol and loneliness. Research findings provided by Herttua, K., et al. (2011) in Finland indicate that living alone is associated with an increased risk of alcohol-related mortality. Because of the study design, it is impossible to say whether living alone is a cause or a consequence of alcohol abuse, but the greater increase in alcohol-related deaths among people living alone compared to married and cohabiting people after the alcohol price reduction suggests that people living alone are more vulnerable to the adverse effects of increased alcohol availability. Living alone should be regarded as a potential risk factor for death from alcohol-related causes.

According to Varga, S. and Piko, B.F. (2015) who studied Hungarian adolescents' health risk behaviours, loneliness is negatively related with drinking.

Results from a Swiss national survey came to conclusion that loneliness is not associated with chronic alcohol consumption, binge drinking, or diet awareness (Richard, A., et al., 2017).

Canham, S. L., et al. (2016) based on data from Health and Retirement Study in the United States for respondents aged 50+ have found that being lonely is associated with reduced odds of weekly alcohol consumption 4–7 days per

week, but not 1–3 days per week, compared to average alcohol consumption 0 days per week in the last 3 months. No association was found between at-risk drinking or binge drinking and loneliness. Loneliness was associated with reduced alcohol use frequency, but not with at-risk or binge drinking.

Schonfeld, L. and Dupree, L.W. (1991) however, suggest that older adults with problematic alcohol consumption drink at home or while alone.

3 Alcohol as Cause of Divorce

In modern world, increasing proportions of individuals do not marry, but live together without marriage. The age of marriage and the likelihood of entering marriage with a child have increased. After a long period of increase, the likelihood of divorce has stabilized at about 45% in Czechia. Fertility has decreased and the age of childbearing has generally increased. Almost one half of children are born outside marriage. The average duration of marriage is 13 years (2018), from which almost half of the couples end with divorce. The most frequent divorce category is 'different characters, views and interests'. Infidelity and alcoholism belong among the five most common causes of divorce. The consequence of these trends is the tremendous diversity of family structures and caregiving arrangements among families. In longitudinal studies, there are multiple transitions without a clear developmental sequence. The problem is further complicated by the fact that alcohol disorders are associated with family structures and the likelihood of transitioning to alternate family structures (Leonard, K. E. and Eiden, R. D., 2007).

3.1 Results for Czechia

First aim of our study is to evaluate the interplay of alcohol and marital status/partnership status in Czechia: *is alcohol a frequent cause of divorce?* Excessive use of alcohol is routinely monitored as one of the causes of divorces. In 2018, alcohol was reported as a primary cause in 377 cases of divorces (1.5%) (CZSO¹, 2019). The proportion of alcohol-caused divorces has declined since 2010 (3.2%), especially on the part of males. However, influence of alcohol might be recently underreported since there is increasing proportion of divorces when cause is not given (5.6% in 2010 and 28.3% in 2018) (Tab. 1).

Alcohol was found a significant predictor of domestic violence. According to Nešpor and Csémy (2005), alcohol was present in 2/3 of cases of domestic violence. Martinková et al. (2014) found 72% of domestic violence offenders were under the influence of alcohol at the moment of the offence/ attack. Recent research showed that risk of aggressive behaviour against partner increases with drinking frequency of 5 or more glasses of alcohol (OR=15–20 for daily excessive alcohol users compared to abstainers/rare alcohol users) (Pikálková, S., et al., 2015). Extent of alcohol-related domestic violence in Czechia is

¹ Czech Statistical Office.

substantial taking into account that about 17–40% of women and 10–38% of men have experienced domestic violence in their lifetime and 2–9% in the last 12 months (Dohnal, D., et al., 2017; In: Mravčík, V., et al., 2019). Results of the survey provided by Martinková, M. et al. (2014) confirm that quarter of female respondents (30 females) had been always afraid of their partners before she left him; more than two thirds of the respondents were afraid of their partner often or very often (68.3 %; 82 females). Also according to *The National Survey on Substance Use*, alcohol-related social harm is quite prevalent since negative impact of alcohol consumption in the last 12 months was by the Adverse Social Consequences scale (Moskalewicz J, Sieroslawski J., 2010; In: Mravčík, V., et al., 2019) reported by 19.4% of the population aged 15–64 years (28.8% males and 11.2% females), including primarily financial problems (12.9%), family problems (7.4%), and impact on relations with friends and social life (5.7%) (Mravčík, V., et al., 2017).

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018
Number of divorces	30 783	28 113	26 402	27 895	26 764	26 083	24 996	25 755	24 313
Alcohol as the cause of divorce	989	768	691	633	554	466	447	496	377
Alcohol as the cause of divorce on the part of male	719	555	491	451	366	322	315	317	245
Alcohol as the cause of divorce on the part of female	270	213	200	182	188	144	132	179	132
Relative ratio (%)	3,21	2,73	2,62	2,27	2,07	1,79	1,79	1,93	1,55

Tab. 1. Alcohol as cause of marriage breakdown between 2010–2018
Source: Czech Statistical Office

Alcohol as cause of marriage breakdown is in long term view statistics higher for males than for females (245 divorces on the part of male, compared to 132 divorces on the part of female in 2018) (see Tab. 1). Distribution of divorces by age categories reaches its maximum for the age group 40–44 years for both sexes when the alcohol-related divorce rate is the highest. The most registered cases of alcohol-related divorces were registered after 9–10 years of marriage. Similarly, partnerships after 20 years of marriage show the same trend. It is evident that one decade seems to be a limit when the number of divorces has increased. Results by education are influenced by the possibility of not stating education, resp. provide this information voluntarily. In 2018, the highest educational attainment was not reported for almost 37% of men and 40% of women (from the total 377 divorces where alcohol was the cause of marriage breakdown). For both sexes (25% for males; 11% for females), the proportion of alcohol-related divorces according to education is generally highest among people with a secondary education, without A-level examination. The proportion of alcohol-related divorces decreases with the increasing level of education.

4 Alcohol Consumption and Partnership

Another aim of this paper is to evaluate alcohol consumption in relation to partnership status: *do people with a partner drink more or less compared to those living without a partner?*

4.1 Methods and Results

For the analysis we used data from the Czech Household Panel Survey 2017 and the SHARE – Survey of Health, Ageing and Retirement in Europe. The goal of our research is to compare and interpret results from two different databases.

4.1.1 Outcomes from the SHARE database

This paper uses data from the generated easySHARE panel dataset for Czechia covering 13 722 observations. The easySHARE release 7.0.0 is based on SHARE Waves 1, 2, 3 (SHARELIFE), 4, 5, 6 and 7 (Börsch-Supan, A. and Gruber, S., 2019). Please find below the outputs from the SHARE database (Fig. 1; Tab. 2; Tab. 3). Fig. 1 represents the results for alcohol consumption per day during last six months by partnership (living with a partner in a household/ living without a partner in a household). On the X axis there are observations obtained for males and females (living without partner/ living with partner) and on the Y axis there is 'days a week consumed alcohol last 6 months'. Previous studies demonstrated no clear evidence for higher alcohol consumption when living alone, which is identical to our findings. This can be explained by the fact that in developed modern society, including Czechia, living alone doesn't unconditionally mean loneliness (associated with depression and alcohol as stress relief). But it's the result of individuals' choice and in many cases a lifestyle.

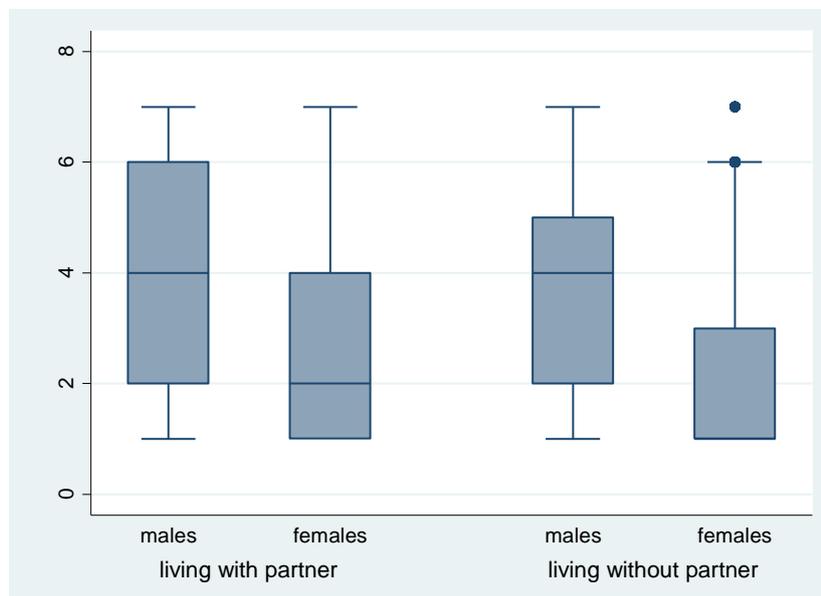


Fig. 1. Alcohol consumption by partnership
Source: SHARE, data: easySHARE release 7.0.0

We ran Two-way ANOVA using the statistical software STATA to determine whether the difference between alcohol consumption of those with and without partner is significant (Tab. 2). Selected variables from the SHARE database were following: '*female*' (gender: female = 1; male = 0), '*partnerinhh*' (living with partner in household), '*br010_mod*' (days a week consumed alcohol last 6 months). Based on the p-values and a significance level of 0.05, we can conclude the following: the p-value for '*female*' is 0.0000, which indicates that gender is associated with different alcohol consumption during last 6 months; the p-value for '*partnerinhh*' is 0.2095, which indicates that living with/ without a partner in a household is not associated with different alcohol consumption during last 6 months; the p-value for the interaction between '*female*partnerinhh*' is 0.0697, which is not statistically significant.

	Number of obs = 13722 Root MSE = 1.88553		R-squared = 0.1322 Adj R-squared = 0.1320		
Source	Seq. SS	df	MS	F	Prob > F
female	7238.75512	1	7238.75512	2036.08	0.0000
partnerinhh	166.957857	1	166.957857	46.96	0.2095
partnerinhh*female	23.8164786	1	23.8164786	6.70	0.0697
Residual	48770.7805	13718	3.55523987		
Total	56200.3099	13721	4.09593397		

Tab. 2. Alcohol consumption by partnership (Two-way ANOVA)
Source: SHARE, data: easySHARE rel. 7.0.0

We used multinomial logistic regression to predict the probabilities of the different possible outcomes of the dependent variable '*br010_mod*', given a set of independent variables ('*female*'; '*partnerinh*'; '*isced1997_r*'). The '*female*' predictor is in all of the cases negative and significant, indicating that females (coded 1) are at lower risk of alcohol consumption during last 6 months, and at higher risk of alcohol consumption 'not at all' (set as baseline) as compared to males. The results of logistic regression for the variable '*partnerinh*' show that respondents living together with partner are at higher risk of alcohol consumption in all categories. Statistical significance was confirmed ($p = 0.000$). We included the variable '*isced1997_r*' (International Standard Classification of Education 1997) to the regression analysis to see how the level of education affects the rate of alcohol consumption. We conclude that people with lower education are at higher risk of common alcohol consumption, and at lower risk of alcohol consumption 'not at all'. The results are significant for the cases 'once or twice a month' ($p = 0.049$); 'once or twice a week' ($p = 0.001$); 'three or four days a week' ($p = 0.003$) and 'five or six days a week' ($p = 0.009$). Please find results in Tab. 3 below.

Multinomial logistic regression		Number of obs =		13722		
Log likelihood = -22507.955		LR chi2(18) =		1960.12		
		Prob > chi2 =		0.0000		
		Pseudo R2 =		0.0417		
br010_mod	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
1__not_at_all	(base outcome)					
2__less_than_once_a_month						
female	-.1503952	.0623565	-2.41	0.016	-.2726117	-.0281787
isced1997_r	.0068651	.0035728	1.92	0.055	-.0001374	.0138676
partnerinhh	-.1388086	.0302659	-4.59	0.000	-.1981286	-.0794886
_cons	-.7121029	.0665911	-10.69	0.000	-.8426191	-.5815868
3__once_or_twice_a_month						
female	-.3574059	.0591398	-6.04	0.000	-.4733178	-.2414941
isced1997_r	.0069109	.0035154	1.97	0.049	.0000209	.0138009
partnerinhh	-.1356317	.0295828	-4.58	0.000	-.1936129	-.0776505
_cons	-.4955721	.0631926	-7.84	0.000	-.6194273	-.3717169
4__once_or_twice_a_week						
female	-1.136668	.0541878	-20.98	0.000	-1.242874	-1.030462
isced1997_r	.0103762	.0032344	3.21	0.001	.0040369	.0167154
partnerinhh	-.1315868	.0290245	-4.53	0.000	-.1884737	-.0746999
_cons	.1038427	.0569847	1.82	0.068	-.0078452	.2155306
5__three_or_four_days_a_week						
female	-1.566008	.0845793	-18.52	0.000	-1.731781	-1.400236
isced1997_r	.0129643	.0044165	2.94	0.003	.0043081	.0216204
partnerinhh	-.2386814	.0491027	-4.86	0.000	-.334921	-.1424418
_cons	-.684061	.0840884	-8.14	0.000	-.8488712	-.5192508
6__five_or_six_days_a_week						
female	-2.031798	.1324895	-15.34	0.000	-2.291473	-1.772123
isced1997_r	.0152496	.0058497	2.61	0.009	.0037845	.0267148
partnerinhh	-.1839923	.0722034	-2.55	0.011	-.3255085	-.0424762
_cons	-1.455793	.1192657	-12.21	0.000	-1.689549	-1.222036
7__almost_every_day						
female	-2.066788	.0671201	-30.79	0.000	-2.198341	-1.935235
isced1997_r	-.0010826	.0054991	-0.20	0.844	-.0118605	.0096954
partnerinhh	-.2104242	.036739	-5.73	0.000	-.2824314	-.1384171
_cons	.2374378	.0650326	3.65	0.000	.1099763	.3648994

Tab. 3. Alcohol consumption during last 6 months observations (Multinomial logistic regression)

Source: SHARE, data: easySHARE rel. 7.0.0

4.1.2 Outcomes from the Czech Household Panel Survey

Another source of data information was obtained from the Czech Household Panel Survey 2017. Also in this case alcohol consumption by partnership was analysed. On the X axis there are observations obtained for males and females (without partner/ with partner) and on the Y axis there is 'alcohol consumption in litres per year (log)'. According to results we can conclude that those living with a partner drink on average more than those without a partner (Fig. 2). In the case of alcohol consumption by partnership and age men with a partner drink the same as those without a partner from the age 50 years. Younger men drink more when they have a partner. Females with a partner drink more in every age

group except of the youngest (Fig. 3). On the X axis there are observations obtained for males and females (without partner/ with partner) by age and on the Y axis there is 'alcohol consumption in litres per year (log)'.

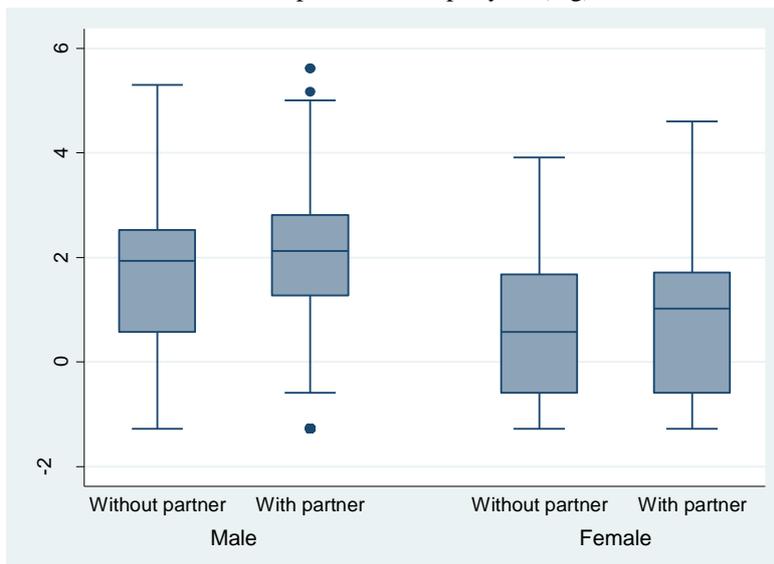


Fig. 2. Alcohol consumption by partnership
Source: the Czech Household Panel Survey, 2017

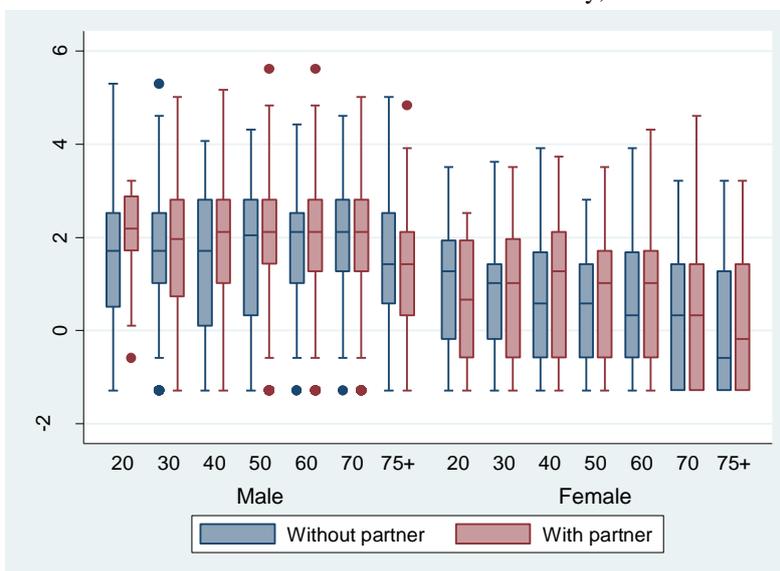


Fig. 3. Alcohol consumption by partnership and age
Source: the Czech Household Panel Survey, 2017

In the next step Two-way ANOVA was used to determine whether the difference between alcohol consumption to those with and without a partner is significant. Results are separated for males and females (Tab. 4 and Tab. 5). Findings: the difference is significant only for females.

	Delta-method		t	P> t	[95% confidence interval]	
	Margin	Std. Error				
partner						
Without partner	10.07468	.7338821	13.73	0.000	8.635521	11.51384
With partner	10.98402	.5105169	21.52	0.000	9.982885	11.98515

Tab. 4. Alcohol consumption by partnership (males)
Source: the Czech Household Panel Survey, 2017

	Delta-method		t	P> t	[95% confidence interval]	
	Margin	Std. Error				
partner						
Without partner	3.227768	.177056	18.23	0.000	2.880593	3.574943
With partner	4.08223	.1590233	25.67	0.000	3.770414	4.394046

Tab. 5. Alcohol consumption by partnership (females)
Source: the Czech Household Panel Survey, 2017

Next method that was used was the logistic regression to see the associations between binge drinking and partnership. Binge drinking means in this case consuming more than 60 grams of ethanol at a single occasion. Results are presented in Tab. 6 and Tab. 7. As reference group we set '*without partner*'. Unlike expected, binge drinking is more prevalent among those living with a partner (although the difference is not significant among females).

Partnership	Odds ratio	95% confidence interval
Without partner	1 (reference)	
With partner	1.36	(1.06-1.75)

Tab. 6. Binge drinking and partnership (males)
Source: the Czech Household Panel Survey, 2017

Partnership	Odds ratio	95% confidence interval
Without partner	1 (reference)	
With partner	1.23	(0.93-1.65)

Tab. 7. Binge drinking and partnership (females)
Source: the Czech Household Panel Survey, 2017

Logistic regression was used also for the analysis of associations between being an abstainer and partnership. Being an abstainer means in this case consuming less than 1 gram of ethanol daily. Please find results in Tab. 8 and Tab. 9.

Partnership	Odds ratio	95% confidence interval
Without partner	1 (reference)	
With partner	0.44	(0.31-0.65)

Tab. 8. Being an abstainer and partnership (males)
Source: the Czech Household Panel Survey, 2017

Partnership	Odds ratio	95% confidence interval
Without partner	1 (reference)	
With partner	0.67	(0.52-0.85)

Tab. 9. Being an abstainer and partnership (females)
Source: the Czech Household Panel Survey, 2017

Conclusion

Several demographic trends have created a multiplicity of family structures that complicate the study of family processes (Bumpass, 2004, In: Leonard, K. E. and Eiden, R. D., 2007). Marital status groups are known to differ in terms of health and mortality in different societies, with non-married persons being in a disadvantaged position compared with married persons (Hu and Goldman, 1990; Joung et al., 1996; Martikainen et al., 2005, In: Joutsenniemi, K., et al. 2007). Alcohol is a major global contributing factor to death, disease and injury. Alcohol consumption affects not only the individuals, but also their families, partners and society at large (WHO, 2011). Alcohol was found a significant predictor of domestic violence and frequent cause of marriage breakdown in Czechia (377 divorces in 2018). We analyzed two different sources of information: SHARE database (version easySHARE rel. 7.0.0) and the Czech Household Panel Survey 2017. Our findings are identical to results in literature review: the association between alcohol consumption and living without a partner was not confirmed. Unlike expected, living with partner increases alcohol consumption. Results (p-values) from SHARE database using multinomial logistic regression show a statistically significant association. In other outcomes (e.g. data obtained from the Czech Household Panel Survey) significance wasn't confirmed. The results contradict a common belief that living alone is associated with heavier or riskier drinking. The findings need to be confirmed using mortality data.

Acknowledgment

This article was supported by the Czech Science Foundation, Grant No. GA ČR 19-23183Y, on a project titled '*Alcohol burden in the Czech Republic: mortality, morbidity and social context*'.

References

1. Leonard, K. E., Eiden, R. D. (2007). *Marital and family processes in the context of alcohol use and alcohol disorders*. Annual review of clinical psychology, 3, 285–310. <https://doi.org/10.1146/annurev.clinpsy.3.022806.091424>
2. Kaisla Joutsenniemi, Tuija Martelin, Laura Kestilä, Pekka Martikainen, Sami Pirkola, Seppo Koskinen, Living arrangements, heavy drinking and alcohol dependence, Alcohol and Alcoholism, Volume 42, Issue 5, September 2007, Pages 480–491, <https://doi.org/10.1093/alcalc/agm011>
3. N. Leigh Hunt (2017). *An overview of systematic reviews on the public health consequences of social isolation and loneliness*. <https://www.sciencedirect.com/science/article/abs/pii/S0033350617302731?via%3DiHub>
4. Shodhganga, *Impact of alcohol addiction on quality of marital life. Literature review*. https://shodhganga.inflibnet.ac.in/bitstream/10603/185094/14/10_chapter%202.pdf
5. Canham, S. L., Mauro, P. M., Kaufmann, C. N., & Sixsmith, A. (2016). Association of Alcohol Use and Loneliness Frequency Among Middle-Aged and Older Adult Drinkers. *Journal of aging and health*, 28(2), 267–284. <https://doi.org/10.1177/0898264315589579>. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4681688/>
6. Herttua, K., Martikainen, P., Vahtera, J., & Kivimäki, M. (2011). Living alone and alcohol-related mortality: a population-based cohort study from Finland. *PLoS medicine*, 8(9), e1001094. <https://doi.org/10.1371/journal.pmed.1001094>. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3176753/>
7. Varga, S., & Piko, B. F. (2015). Being lonely or using substances with friends? A cross-sectional study of Hungarian adolescents' health risk behaviours. *BMC public health*, 15, 1107. <https://doi.org/10.1186/s12889-015-2474-y>. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4637146/>
8. Mravčík V, Chomynová P, Nechanská B, Černíková T, Csémy L. Alcohol use and its consequences in the Czech Republic. *Cent Eur J Public Health*. 2019;27(Supplement):S15-28. doi: 10.21101/cejph.a5728. PubMed PMID: 31901189. Available at: https://cejph.szu.cz/artkey/cjp-201988-0003_alcohol-use-and-its-consequences-in-the-czech-republic.php
9. Sharmin, Sonia & Kypri, Kypros & Khanam, Masuma & Wadolowski, Monika & Bruno, Raimondo & Mattick, Richard. (2017). Parental Supply of Alcohol in Childhood and Risky Drinking in Adolescence: Systematic Review and Meta-Analysis. *International Journal of Environmental Research and Public Health*. 14. 10.3390/ijerph14030287. Available at: https://www.researchgate.net/figure/Meta-analysis-forest-plot_fig2_314441757
10. Bumpass L. L., 2004. Social change and the American family. *Annals of the New York Academy of Sciences*. Acad Sci. 2004;1038:213–19.
11. Saxena, S., Sharma, R., & Maulik, P. (2003). Impact of alcohol use on poor families: A study from North India. *Journal of Substance Use*, 8(2), 78–84.

12. Brennan , Penny, L.; Moos, Rudolf, H. and Kelly , Kristma, M. (1994). Spouse of late life problem dmkers. Functioning, coping, responses and family contexts, Journal of Family Psychology, Dec. , Vol. 8(4), 447-457.
13. Crisp, B. R. and Barber , J. G. (1995). The Drinker's Partner Distress Scale: an instrument for measuring the distress causal by drinkers to their partners, International Journal of Addic fian, June, 3 0(8), 1009- 1017.
14. McCrady BS, Epstein EE (1995). *Marital therapy in the treatment of alcohol problems*. In: Jacobson NS, Gurman AS, editors. Clinical handbook of couple therapy. New York: Guilford Press; 1995. pp. 369–393.
15. Paolino TJ, McCrady BS (1977). *The alcoholic marriage: Alternative perspectives*. New York: Grune & Stratton.
16. WHO, 2019. *Status report on alcohol consumption, harm and policy responses in 30 European countries 2019*. Available at: http://www.euro.who.int/__data/assets/pdf_file/0019/411418/Alcohol-consumption-harm-policy-responses-30-European-countries-2019.pdf?ua=1
17. Marshal M. P. (2003). For better or for worse? The effects of alcohol use on marital functioning. *Clinical psychology review*, 23(7), 959–997. <https://doi.org/10.1016/j.cpr.2003.09.002>
18. Czech Statistical Office. *Divorces: by cause of marriage breakdown on the part of male and female and by educational attainment of divorces*. Demographic Yearbook 2010-2018. Prague: CZSO. 2019 (In Czech). Available at: <https://www.czso.cz/csu/czso/demographic-yearbook-of-the-czech-republic-2010-v1lnbg8yv8>
19. *National Strategy for Prevention and Reduction of Damage Related to Addiction Behavior 2019–2027*. 2019. Prague: Secretariat of the Government Council for the Coordination of Drug Policy (In Czech). ISBN 978-80-7440-231-9.
20. Csémy, L., et al., 2019. *Tobacco and alcohol consumption in the Czech Republic, 2018*. Prague: National Institute of Public Health.
21. Džurová, Dagmar & Spilková, Jana & Pikhart, Hynek. (2010). *Social inequalities in alcohol consumption in the Czech Republic: A multilevel analysis*. Health & place. 16. 590-7. 10.1016/j.healthplace.2010.01.004. Available at: https://www.researchgate.net/publication/41424124_Social_inequalities_in_alcohol_consumption_in_the_Czech_Republic_A_multilevel_analysis
22. Richard A, Rohrmann S, Vandeleur CL, Schmid M, Barth J, et al. (2017) Loneliness is adversely associated with physical and mental health and lifestyle factors: Results from a Swiss national survey. *PLOS ONE* 12(7): e0181442. <https://doi.org/10.1371/journal.pone.0181442>
23. Canham, S. L., Mauro, P. M., Kaufmann, C. N., & Sixsmith, A. (2016). Association of Alcohol Use and Loneliness Frequency Among Middle-Aged and Older Adult Drinkers. *Journal of aging and health*, 28(2), 267–284. <https://doi.org/10.1177/0898264315589579>
24. Schonfeld L, Dupree LW. Antecedents of drinking for early- and late-onset elderly alcohol abusers. *Journal of Studies on Alcohol*. 1991;52(6):587–592.
25. Börsch-Supan, A. & S. Gruber (2019): easySHARE. Release version: 7.0.0. SHARE-ERIC. Dataset. doi: 10.6103/SHARE.easy.700
26. WHO, 2011. *Global status report on alcohol and health*. Geneva, Switzerland. ISBN 978-92-4-156415-1. Available at: https://www.drugsandalcohol.ie/14675/1/Global_status_report_on_alcohol_and_health.pdf

A relative entropy measure of divergences in labour market outcomes by educational attainment

Maria Symeonaki*

Department of Social Policy, School of Social and Political Sciences,
Panteion University of Social and Political Sciences, Athens, Greece
*(E-mail: msymeon@panteion.gr)

Abstract. The present study¹ proposes a new way of examining cross-country differences in labour market outcomes for young individuals aged between 15-29 in relation to their educational attainment, using raw data drawn from the European Union's Labour Force Survey (EU-LFS) and three different Kullback–Leibler divergence measures. We assume that if educational attainment did not play a decisive role to whether young individuals were employed (or unemployed), the probability of him/her being “low”, “medium” or “highly” educated would be equal and therefore their distribution to the educational attainment categories would be the discrete Uniform distribution. Having accepted this hypothesis, we estimate the direct divergence between the way employed (and unemployed) young individuals are distributed to the educational attainment categories and the discrete Uniform distribution (with three possible outcomes relating to those categories). The divergence between the distributions of employed and unemployed individuals by educational attainment is also explored. Countries are ranked according to the relative entropy values of these measures for the latest at the time available raw data drawn from the EU-LFS for the year 2016.

Keywords: relative entropy, Kullback–Leibler divergence measure, EU-LFS, labour market outcomes

1 Introduction

In recent years, there has been an increasing interest in early job insecurity and the labour market outcomes of young individuals in order to examine the labour market position of youth in Europe and recognise factors explaining divergences between member states aiming at informing social policy making and provide

¹ The study is implemented under the research project HARMONIA funded by a grant (No. 2018/30/M/HS4/00744) from the National Science Centre in Poland



applicable knowledge. It is well accepted that young people are under greater risk of unemployment, involuntary part-time employment, and take on precarious and flexible jobs more easily in the process of moving from school to the labour market. Nevertheless, there are cross-national discrepancies in the patterns of transitions from school to employment. A considerable amount of literature has been published on school-to-work transitions and the labour market outcomes of young individuals: Quintini et al. (2007) discovered that there is much turnover between labour market categories in all OECD countries, but the average length of young graduates' transition differs considerably amongst these countries. A more recent study (Quintini and Martin (2014)) studied the school-to-work transitions for sixteen emerging and advanced countries and observed that the employability of young individuals is lower in emerging economies, where school leavers of a young age have a longer transition to the labour market, characterised by a higher percentage of the NEET rate (individuals that are Not in Employment, Education or Training) and informal employment. School to work transition was investigated in a following ILO report (Mathys, 2019) where it was suggested that there is a large variation across sixty countries that were studied, based on their level of development and their income, while education was found to be a strong, positive factor that influenced young individuals' transition, especially in developed countries. Early job insecurity and the dynamic process of transiting into the labour market is addressed by multiple indicators and models such as time elapsed between graduation and the first job, rate of transitions from employment to unemployment or inactivity, transition probabilities from school to the labour market (Bosch and Maloney (2007), Brzinskey-Fay (2014), Christodoulakis and Mamatzakis (2009), Eurofound (2014), Flek and Mysikova (2015), Karamessini et al. (2016, 2019a, 2019b), Symeonaki and Stamatopoulou (2020a, 2020b, 2020c), Gallie et al (2017), among others). Other studies have considered the construction of a multidimensional index of early job insecurity: Symeonaki et al. (2019a, 2019b, 2019c), in Karamessini et al. (2016; 2019) and Symeonaki et al. (2018). It has conclusively been shown that higher educational attainments is a major contributing factor for a smother transition into the labour market.

The present study examines cross-country differences in labour market outcomes for young individuals aged between 15-29 in relation to educational attainment, using raw data drawn from the European Union's Labour Force Survey (EU-LFS) and three different Kullback–Leibler divergence indicators (Kullback–Leibler (1951), Kullback (1987)). The Kullback-Leibler divergence measure has been extensively used since its definition in various fields of studies including economics, engineering, statistics, physics, psychology, etc. More precisely, we measure the direct divergence between the distributions of employed (and unemployed) young individuals to the educational categories (i.e. Low, Medium and High) and the discrete Uniform distribution where the elements of the finite set of educational categories are equally likely. The divergence between the distributions of employed and unemployed individuals is also explored. Countries are ranked according to their relative entropy values of these measures for the latest at the time available raw

data for the year 2016. The Kullback–Leibler relative entropy (or divergence) measure is suggested as a very practical way to measure equality of opportunities of young individuals in employment in respect to their educational achievements and the differences in the distributions of employed and unemployed individuals to the educational levels.

The paper is structured as follows: Section 2 provides information on the data and the indicators used in the subsequent analysis, while Section 3 presents the results for the countries under study, draws conclusions and makes further research suggestions.

2 Preliminaries, Data and Measurement

In the present study the Kullback–Leibler divergence measure is suggested as a means to measure equality of opportunities of young individuals in employment and the differences in employment and unemployment in respect to the different educational levels. Raw data from the EU-LFS survey is used for implementing the suggested methodology for 25 European countries. The EU-LFS is a unique data source, providing detailed information on labour market participation and the working conditions in European countries. It allows multivariate analysis by sex, age, educational attainment and other socio-demographic characteristics, while common principles and guidelines are used to ensure cross-country comparability. For the purpose of the present research, the focus is on individuals that are aged between 15 and 29. Normally, in EUROSTAT's definitions as well, a young person is an individual aged between 15 and 24. It was decided that the upper limit was extended to the age of 29 to incorporate more information on the post-graduation employment and simultaneously increase the country samples. Other studies have used the suggested definition of a young individual acknowledging the fact that school-to-work transition has been gradually belated in many countries and frequently is completed after the age of 25 (Karamessini et al. (2016, 2019a, 2019b), Symeonaki and Stamatopoulou (2020a, 2020b, 2020c)). The fact that some young individuals do remain in education beyond the age of 24 years is in general well-accepted (OECD 1998, Chapter 3, p. 91).

The Kullback–Leibler divergence measure, denoted by D_{KL} and also known as relative entropy measure, is a measure of how one probability distribution is different from a second probability distribution, which is called the reference distribution. The Kullback–Leibler divergence was firstly introduced by Solomon Kullback and Richard Leibler in 1951 as the directed divergence between two distributions. However, Kullback preferred the term discrimination information measure (Kullback (1987)). More specifically, D_{KL} is given by the following definition:

Definition 1: For discrete probability distributions P and Q defined on the same probability space X , the Kullback–Leibler divergence from Q to P is defined as:

$$D_{KL}(P \parallel Q) = \sum_{x \in X} P(x) \ln \left(\frac{P(x)}{Q(x)} \right).$$

Apparently this is equivalent to:

$$D_{KL}(Q \parallel P) = - \sum_{x \in X} P(x) \ln \left(\frac{Q(x)}{P(x)} \right).$$

Moreover, D_{KL} can be interpreted as the expectation of the logarithmic difference between the probabilities P and Q , where the expectation is taken using the probabilities P . It is proven that, the Kullback–Leibler divergence is defined only if:

$$\forall x, Q(x) = 0 \Rightarrow P(x) = 0 \text{ (absolute continuity)}.$$

The minimum value is equal to 0 ($D_{KL} = 0$) when the two distributions are identical.

In this study we use D_{KL} in order to quantify the equality of opportunities of youth into the labour market. More specifically, we estimate the direct divergence between the distributions of employed individuals to the educational categories and the Uniform distribution, $D_{KL}(E \parallel Unif)$ and unemployed individuals to the educational categories and the Uniform distribution, $D_{KL}(U \parallel Unif)$, assuming the discrete Uniform distribution with three possible outcomes relating to the educational levels (i.e. Low, Medium and High), each having an equal probability of $p = 1/3$. E denotes the distribution of employed young individuals into the three educational categories, denoting low, medium and high educational attainment. U is the respective distribution for unemployed young individuals. The divergence between the distributions of employed and unemployed individuals $D_{KL}(E \parallel U)$ is also explored.

The Kullback–Leibler divergence is a special case of a broader class of statistical divergences called f -divergences as well as the class of Bregman divergences. It is in fact the only such divergence over probabilities that is a member of both classes. Hobson (1971) proved that the Kullback–Leibler divergence is the only measure of difference between probability distributions that satisfies some desired properties, which are the canonical extension to those appearing in a commonly used characterization of entropy.

The values of the educational attainment were recoded to reflect low, medium and high educational level and harmonised according to the latest version of International Standard Classification of Education (ISCED) and the divergence measures were estimated for the latest at the time available data for the year 2016.

3 Results, Interpretation and Future Work

In this section the results of our analysis are presented. More specifically, Table 1 reveals the respective D_{KL} values that were estimated using raw data drawn from EU-LFS for the year 2016. $D_{KL}(E \parallel Unif)$ ($D_{KL}(U \parallel Unif)$) is the relative entropy of E (or U) with respect to a distribution that reflects equal distribution of employed (unemployed) young individuals to the three educational categories. $D_{KL}(E \parallel U)$ measures the discrepancies between employed and unemployed individuals in relation to their distribution to the educational levels. Apparently, lower values would indicate smaller divergences.

Table 1. The values of $D_{KL}(E \parallel Unif)$, $D_{KL}(U \parallel Unif)$ and $D_{KL}(E \parallel U)$

	$D_{KL}(E \parallel Unif)$	$D_{KL}(U \parallel Unif)$	$D_{KL}(E \parallel U)$
Switzerland	0.063	0.059	0.036
Estonia	0.126	0.106	0.136
Czech Republic	0.325	0.140	0.153
Austria	0.090	0.080	0.118
Latvia	0.141	0.126	0.091
Lithuania	0.233	0.222	0.127
Denmark	0.042	0.061	0.018
Sweden	0.120	0.111	0.258
the Netherlands	0.032	0.168	0.177
Hungary	0.208	0.185	0.159
Norway	0.008	0.101	0.134
Poland	0.269	0.233	0.076
Slovakia	0.352	0.158	0.126
Belgium	0.108	0.047	0.140
Finland	0.189	0.126	0.107
Slovenia	0.241	0.155	0.012
Bulgaria	0.221	0.077	0.143
France	0.132	0.084	0.138
Romania	0.101	0.169	0.021
Portugal	0.029	0.029	0.016
Cyprus	0.163	0.158	0.014
Croatia	0.407	0.396	0.042
Spain	0.019	0.052	0.110
Italy	0.160	0.149	0.025
Greece	0.150	0.130	0.006

Source: EU-LFS, 2016.

Figures 1, 2 and 3 present the respective relative entropy measures for the year 2016 for the European countries under study.

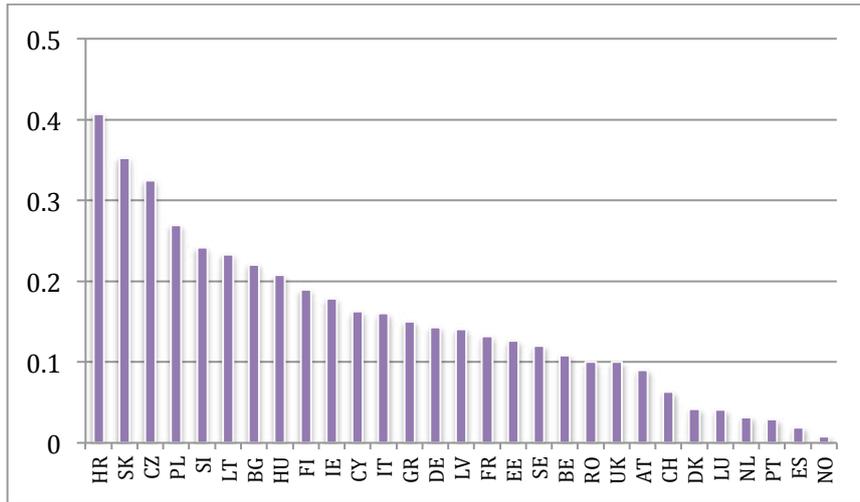


Fig. 1. $D_{kl}(E \parallel Unif)$, EU-LFS, 2016

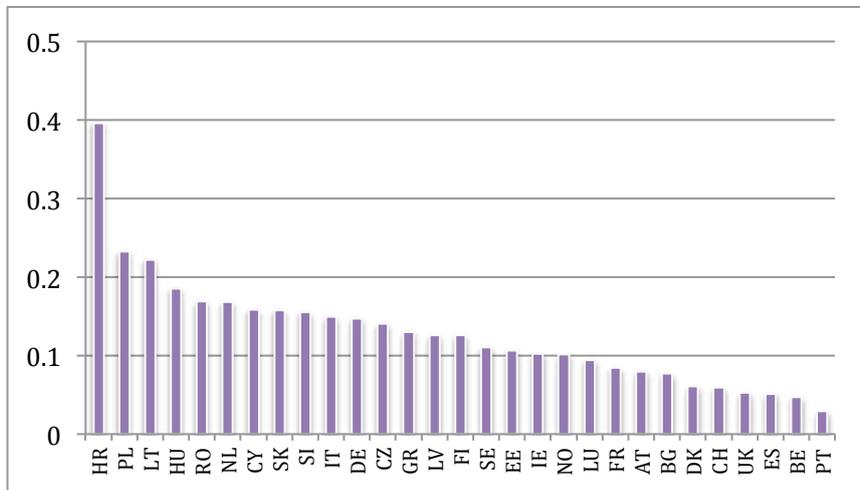


Fig. 2. $D_{kl}(U \parallel Unif)$, EU-LFS, 2016

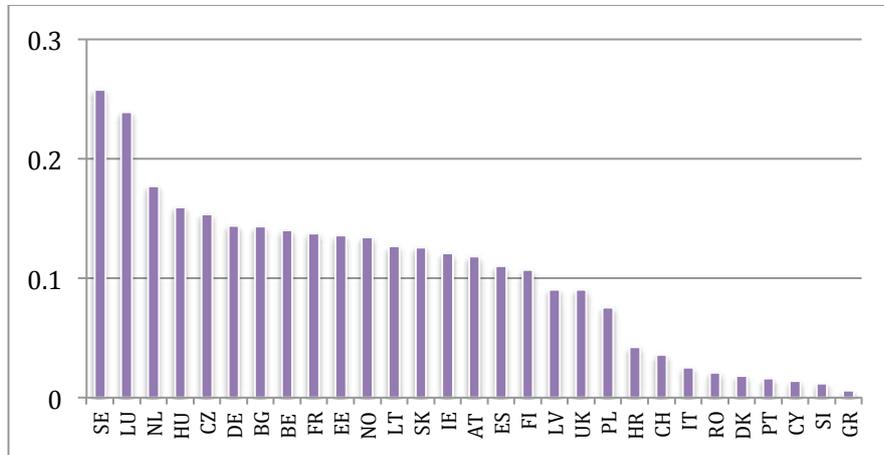


Fig. 3. $D_{KL}(E||U)$, EU-LFS, 2016

Expressed in the language of Bayesian inference, $D_{KL}(E||Unif)$ measures the information gained by revising one's beliefs from the prior probability distribution $Unif$ to the posterior probability distribution E . In other words, it is the amount of information lost when an equal opportunity distribution $Unif$ is used to approximate E . In the general case where the divergent measure $D_{KL}(A||B)$ is estimated, A represents the "actual" distribution of data, observations, or a precisely calculated theoretical distribution, while B typically represents a theory, a model, a description or an approximation of A . Consequently, it gives a measure of divergence of "reality from a model" and it provides estimation of how much the model has yet to learn.

It is evident from the results exhibited in Table 1 and Figures 1 that in countries belonging to the social democratic regime (e.g. Norway, the Netherlands, Switzerland, Denmark) the values of $D_{KL}(E||Unif)$ are very low, with Norway exhibiting the smallest value (equal to 0.008) denoting an almost identical distribution of E and $Unif$. Some southern European countries (Spain and Portugal) perform in a similar way, also exhibiting small values. However, it is important to note here that in first eight countries (Figure 1) with the highest values of divergence measures are post-socialist countries (namely Croatia, Slovakia, Czech Republic, Poland, Slovenia, Lithuania, Bulgaria and Hungary). This means that in these countries the way employed young individuals are distributed to educational categories is influenced by educational attainment the most. We detect a similar behaviour when looking at the divergence between the unemployed distribution to educational categories and the Uniform distribution. Again, the first five countries with the highest scores are post-socialist countries

(namely Croatia, Poland, Lithuania, Hungary and Romania), whereas no country in this welfare regime is seen in the right hand side of the graph (Figure 2).

When the divergences between the distributions of employed and unemployed young individuals to the three educational levels is examined one notices that there are three distinct clusters of countries (Figure 3). In the right-hand side cluster Portugal, Cyprus, Slovenia and Greece are included showing very small differences between the distribution of employed and unemployed individuals to the educational categories. This means that educational attainment in these countries could not predict whether a young individual would be employed or unemployed and therefore would not increase or decrease the chances of being employed or unemployed.

The relationship to other probability distances (distance measures) could be explored as further research, while the values of the Kullback–Leibler divergence could be estimated for the following or preceding years using the EU-LFS raw data to detect changes or evolution through time.

References

1. M. Bosch, and W. Maloney. *Comparative Analysis of Labor Market Dynamics Using Markov Processes: An Application to Informality*. Discussion Paper no. 3038. Bonn: IZA, 2007.
2. A. Hobson. *Concepts in statistical mechanics*. New York: Gordon and Breach. ISBN 978-0677032405, 1971.
3. C. Brzinsky-Fay, C. Lost in Transition? Labour Market Entry Sequences of School Leavers in Europe. *European Sociological Review*, 23(4), 409-422, 2007.
4. C. Brzinsky-Fay, C. The measurement of school-to-work transitions as processes. About Events and Sequences. *European Societies*, 16(2), 213-232, 2014.
5. K.P. Burnham, D.R. Anderson, D. R. (2002). *Model Selection and Multi-Model Inference* (2nd ed.). Springer. p. 51. ISBN 9780387953649.
6. G. Christodoulakis, G. and C. Mamatzakis, C. *Labour Market Dynamics in EU: a Bayesian Markov Chain Approach*. Discussion paper series no. 2009-07, Department of Economics, University of Macedonia, 2009.
7. Eurofound. *Mapping youth transitions in Europe*. Luxembourg: Publications Office of the European Union, 2014.
8. V. Flek, and M. Mysikova. Unemployment Dynamics in Central Europe: A labour flow approach. *Prague Economic Papers*, 24(1), 73-87, 2015.
9. D. Gallie, A. Felstead, F. Green and Inanc, H. The hidden face of job insecurity, *Work, employment and society*, 31(1), 36-53, 2017.
10. M. Karamessini, M. Symeonaki, G. Stamatopoulou. The role of the economic crisis in determining the degree of early job insecurity in Europe, Negotiate working paper 3.3, 2016.
11. M. Karamessini, M. Symeonaki, G. Stamatopoulou and A. Papazachariou. *The careers of young people in Europe during the economic crisis: Identifying risk factors*, Negotiate working paper no. D3.2. Retrieved from <https://blogg.hioa.no/negotiate/files/2015/04/NEGOTIATE-working-paper-no-D3.2-The-careers-of-young-people-in-Eurpa-during-the-economic-crisis.pdf>, 2016.

12. M. Karamessini, M. Symeonaki and G. Stamatopoulou, G. *The role of the economic crisis in determining the degree of early job insecurity in Europe*, Negotiate working paper no D3.3. Retrieved from <https://blogg.hioa.no/negotiate/files/2015/04/NEGOTIATE-working-paper-D3.3.pdf>, 2016.
13. M. Karamessini, M. Symeonaki, D. Parsanoglou and G. Stamatopoulou. Mapping Early Job Insecurity impacts of the Crisis in Europe. In B. Hvinden, T. Sirovatka and J. O'Reilly (Eds.) *Youth Unemployment and early job insecurity in Europe: Concepts, consequences and policy approaches?* (pp.24-44). Edward Elgar, 2019a.
14. M. Karamessini, M. Symeonaki, G. Stamatopoulou and D. Parsanoglou. Factors explaining youth unemployment and early job insecurity in Europe. In B. Hvinden, T. Sirovatka and J. O'Reilly (Eds.), *Youth Unemployment and early job insecurity in Europe: Concepts, consequences and policy approaches?* (pp. 45-69). Edward Elgar, 2019b.
15. S. Kullback and R.A. (1951). *On information and sufficiency*. Annals of Mathematical Statistics. **22** (1): 79–86. doi:10.1214/aoms/1177729694. JSTOR 2236703. MR 0039968.
16. S. Kullback (1959), *Information Theory and Statistics*, John Wiley & Sons. Republished by Dover Publications in 1968; reprinted in 1978: ISBN 0-8446-5625-9.
17. S. Kullback, S. (1987). *Letter to the Editor: The Kullback–Leibler distance*. The American Statistician. **41** (4): 340–341. doi 10.1080/00031305.1987.10475510. JSTOR 2684769.
18. Mathys, Q. (2019). From school to work: an analysis of youth labour market transitions, ILOSTAT spotlight on work statistics, no. 9. Geneva: ILO.
19. D. MacKay (2003). *Information Theory, Inference, and Learning Algorithms* (First ed.). Cambridge University Press. p. 34. ISBN 9780521642989.
20. G. Quintini, G. and S. Martin (2014), Same Same but Different: School-to-work Transitions in Emerging and Advanced Economies, *OECD Social, Employment and Migration Working Papers*, No. 154, OECD Publishing, Paris, <https://doi.org/10.1787/5jzbb2t1rcwc-en>.
21. M. Symeonaki, M. Karamessini, and G. Stamatopoulou. Measuring School-to-Work Transition Probabilities in Europe with Evidence from the EU-SILC. In J. Bozeman, T. Oliveira, C. Skiadas, and S. Silvestrov (Eds.), *Data Analysis and Applications: New and Classical Approaches*, ISTE Science Publishing, to appear, 121-136, 2019a.
22. M. Symeonaki, M. Karamessini and G. Stamatopoulou, G. Gender-based differences on the impact of the economic crisis on labour market flows in Southern Europe. In J. Bozeman and C. Skiadas, (Eds.), *Data Analysis and Applications: New and Classical Approaches*, ISTE Science Publishing, 107-120, 2019b.
23. M. Symeonaki and G. Stamatopoulou, Describing labour market dynamics through Non Homogeneous Markov System theory, in *The Springer Series on Demographic Methods and Population Analysis*, 2020 (in press).
24. M. Symeonaki and G. Stamatopoulou, Assessing labour market mobility in Europe, in *Demography of Population Health, Aging and Health Expenditures: The Springer Series on Demographic Methods and Population Analysis*, 2020 (in press).

25. M. Symeonaki, G. Stamatopoulou and M. Karamessini. On the measurement of early job insecurity. In C. H. Skiadas and C. C. Skiadas (Eds.), *Demography and Health Issues - Population Aging, Mortality and Data Analysis*, 275-288, Springer. DOI 10.1007/978-3-319-76002-5, 2018.
26. M. Symeonaki and G. Stamatopoulou, G. A Markov system analysis application on labour market dynamics: The case of Greece. Paper presented at IWPLMS, Athens, Greece, 22-24 June, 2015.
27. M. Symeonaki, D. Parsanoglou, D. and Stamatopoulou, G. The evolution of early job insecurity in Europe, *SAGE Open*, 1-23, <https://journals.sagepub.com/doi/pdf/10.1177/2158244019845187>, 2019.

Assessing the intergenerational educational mobility in European countries based on ESS data: 2002 – 2016

Maria Symeonaki* and Paraskevi Tsinaslanidou**

Department of Social Policy, School of Social and Political Sciences,
Panteion University of Social and Political Sciences, Athens, Greece

*(E-mail: msymeon@panteion.gr)

** (E-mail: voulatsinaslanidou@outlook.com)

Abstract. The present paper examines the impact of the immediate educational family environment on individuals' educational achievements, focusing on the comparison of the effectiveness of the distinct European social protection systems. To achieve this, primary data are drawn from Rounds 1-8 of the European Social Survey for 24 countries and the upward mobility index, as well as the Bartholomew and Prais-Shorrocks mobility indices are calculated. Key findings of the paper are the high intergenerational educational mobility of the social-democratic welfare states and the extremely low mobility of the Southern-European welfare regimes.

Keywords: intergenerational educational mobility, welfare regimes, European Social Survey

1 Introduction

Although the principle of social justice is theoretically promoted in the prevailing system of economic and social organization, research on social mobility demonstrates the difficulty of the lower social classes for upward and especially long-distance mobility and their intergenerational stagnation. The phenomenon of social stratification, and therefore social inequality, is not a characteristic of our time, but has extensively preoccupied societies in the past. It is a fact that there is a very wide range of different and conflicting theories about the causes that give rise to social inequality and the impact it has on both the individual and society as a whole. The approaches/policies that have prevailed in advanced countries to interpret and address social inequalities are essentially wavering between equality of opportunity and equality of conditions. The approach of equality of opportunities stems from the liberal notion and asserts that policies such as public education equate existing inequalities and give individuals the same chances of achieving social growth. In this way, the development of each one is a result of his/her abilities and does not depend on monetary, racial, geographical or other barriers related to social origin. Equality of opportunities is usually studied in relation to the educational and professional success of the individual, compared to her/his social and demographic characteristics (Breen and Jonsson [6]). On the contrary, equality of conditions concerns income inequality, wealth and, more broadly, material goods between individuals and concerns intervention policies such as high taxation (Breen and Jonsson, [6]). The approach of equality of opportunities seems to have prevailed in the field of social policy in recent decades, while the approach of equality of conditions has been significantly sidelined. But to what extent are the policies pursued by European states sufficient to create equal opportunities?

The research presented in the present paper is based on the above reflection, on the fact that the welfare state is one of the key institutional forces influencing stratification and mobility by intervening interactively in the relationship of family origin - individual successes (Papatheodorou and Papanastasiou, [20]; Papanastasiou, [19]) and on studies where social origin provably continues to affect educational success (Atkinson, [2]; Blanden et al., [4]; Chevalier et al., [7]; OECD, [18]; Peter et al., [22]; Woessmann [33]). More specifically, using primary data drawn from the European Social Survey, the absolute upward mobility index and the relevant Bartholomew (MB) and Prais-Shorrocks (MP-S) mobility indices are calculated for the years 2002 to 2016, in order to compare and evaluate the effectiveness of the European welfare states in enhancing intergenerational educational mobility.

The paper is structured as follows. Section 2 provides information on the connection between social mobility, education and social policy. Section 3 presents the data used, their limitations and the methodology of the study. Section 4 presents the results and Section 5 draws conclusions and makes further research suggestions.

2 Social mobility, the importance of education and social policy

The key role of education in the study of social mobility is first proven by its direct connection with status and professional perspective. According to Heath [13], education is considered as very



important, especially for people belonging to the lower classes, as it equips themselves with the necessary provisions to achieve social growth. Similar results were obtained from Glass's [12] study, in which the likelihood of mobility of working or middle-class individuals increased when (they) completed secondary education, while Blau and Duncan [5] proved that education has the strongest impact on professional success.

The fact that the individual's social background continues to be a key factor in his/her educational career pathway proves the necessity of intergenerational research. According to Atkinson [2], education is class-based, as middle-class children are more likely to achieve educational achievement than working-class children. Similarly, Blanden et al. [4] found that children from low income families are less likely to succeed in education, while Chevalier et al. [7] concluded that extensive access to higher education took place while the impact of family background on individuals' educational achievement had increased. The OECD [18] study, which compared the results of 15-year-old students to their performance in math, literature and science, concluded that the educational level of parents has the greatest impact on educational performance. The same results were provided by Woessmann [33], who compared the effects of the family environment on children's educational achievement in 17 European countries and in the United States. Finally, it is commonplace among social researchers that children from disadvantaged families tend to have less success prospects than children from families with more advantages. The fact that regardless of society and historical period, family and social origins affect the future success of individuals has been characterized by Erikson and Goldthorpe [8] as constant flux.

The comparison of educational systems and their results has been studied worldwide in the light of their sociological, economic and historical dimension (West and Nikolai [32]). In her research, Allmendinger [1], classified educational systems according to their degree of standardization and stratification and found out that institutions with a high degree of stratification are those that have a decisive influence on the professional status of their graduates. Hoffmeyer-Zlotnik and Warner [17], studying the layout of different education systems, came up with the typology of four educational models. In the first type, with a representative in Germany, primary education is short-lived and there is a distinction between lower and upper secondary education. Higher education includes parallel schools (which provide additional vocational education), academic high schools and technical and non-technical universities. In the second type, represented by Luxembourg, primary education is longer, while in the lower secondary education there are a limited number of school types. In higher education there are different types of general and vocational schools and in higher education there are academic-vocational institutions and universities. In the third type, represented by Denmark, there is no distinction between primary and secondary education. More specifically, primary and lower secondary education are included in a single school, while in secondary education there is one type of vocational school and several types of general education. In higher education, the distinction between vocational and university education is slight. Finally, in the fourth type, represented by France, a characteristic feature is the large participation in kindergarten and pre-kindergarten. The duration of primary education is longer and there are no subdivisions of lower secondary education. In higher secondary education there is little vertical differentiation, while higher education has several differences.

Although education is one of the first state policies adopted in the 19th century and is at the heart of social services, it has been studied in a relatively limited way in the light of different social protection systems (Hega and Hokenmaier [15]). One of the first approaches to the study of educational and social policy is that of Heidenheimer [16], who, by examining the public policies developed in Europe and America, argues that the development of Western welfare states is characterized by an "exchange" between public investment in secondary education and investment in social security programs. Hecló [14] came to similar conclusions about the "compensation" between public investment in education and the expansion of social programs. In the light of the well-accepted typology of Esping-Andersen [9], Hega and Hokenmaier [15] studied the relationship of "compensation" between education spending and social security spending (in 18 OECD countries for the period 1960-1990). They concluded that welfare states with similar social security policies are grouped in the same way in the field of educational policy. They found that the countries with the highest expenditures in both education and social security were those with a social democratic system of social protection. Countries with conservative-corporate welfare states have higher spending than countries with a liberal social welfare state in their social security programs, while states with a liberal social protection system tend to spend more on their education policy. Another important finding of the research is the increased participation in general education programs at the secondary level of liberal social states, compared to the other welfare states.

In relation to educational inequality, Peter et al. [22], using the typology of Esping-Andersen, studied the "within" and "between" school differences of students compared to their socioeconomic

background, and found that socioeconomic origin influenced more educational outcomes of students in conservative-corporatist welfare states, less so in liberal ones while the lowest influence was found in social democracies. To assess the effect that the educational level of father has on children's educational success, in their study Papatheodorou and Papanastasiou [21] applied the method of generalized regression in the 14 oldest Member States of the European Union, combining the Esping-Andersen typology and the south-European model. According to the results of the research, the education of the father has a significant effect on the education of the individuals of all countries under examination. The countries in which intergenerational educational mobility is higher are those with a social democratic system, while the lowest intergenerational educational mobility is found in the countries of the south-European model.

Symeonaki and Stamatopoulou [28] studied the patterns of intergenerational educational mobility in Greece and their changes for different birth cohorts, while also investigating the transmission of educational attainments from both father and mother through generations over time, based on data drawn from the European Social Survey. Distance and similarity measures were proposed to complement traditional methodologies. Symeonaki and Stamatopoulou [27] explored the transition to higher education as an issue of intergenerational educational mobility, while they regarded intergenerational mobility as a distance measure between probability distribution functions (Symeonaki and Stamatopoulou [29]). Stamatopoulou et al. [25] studied the intergenerational transmission of education with evidence from the ESS and Symeonaki et al. [30] used the European Survey on Income and Living Conditions (EU-SILC) to study intergenerational occupational mobility. Fuzzy Markov Systems and symbolic, heuristic knowledge were used in Symeonaki et al. [26] to study intergenerational educational mobility in Greece with data drawn from the ESS.

To conclude educational intergenerational mobility defined as the trajectories observed from one generation to another and between different social classes has been used in the literature to measure whether and to what extent the socio-economic status of origins (measured in terms of parental education) transmit from parents to children and consequently can be seen as an indicator of equality of opportunities.

3 Data and methodology

The data used in the present study are drawn from the European Social Survey (ESS)¹, which is a long-term comparative research survey designed to record and document the attitudes, beliefs and behavioural patterns of the European populations and to produce comparable social indicators, able to be used for European social policy. Started in 2002, the ESS is conducted every two years in more than 20 European countries. It involves strict random probability sampling, a minimum target response rate of 70% and rigorous methodological criteria and collects data from nationally representative samples of persons aged 15 years and above, regardless of their nationality, citizenship or legal status. The "homeless" and people living in collective dwellings are excluded from the sample. The main advantage for choosing the ESS data is the fact that it provides the necessary information on parental social status and the surveyed individuals, even if they do not live in the same residence.

More specifically, our focus is on the relation of highest educational attainment of both parents and individuals, which were harmonised according to the latest version of International Standard Classification of Education (UNESCO [31]) and then recoded to reflect low, medium and high educational attainment. Data used were drawn from the years 2002 up to the latest at the time available data, i.e. the year 2016, for all countries that participated for at least 4 rounds. The analysis includes therefore 24 countries, which were categorised according to their welfare state into (Esping-Andersen [9]; Fenger [10]; Ferrera[11]): social democratic (Sweden, Norway, Finland, the Netherlands and Denmark), conservative-corporatist (Belgium, France, Germany and Austria), liberal (Ireland and UK), southern-European (Spain, Portugal, Greece and Italy) and post-communist (European sub-type: Poland, Czech republic, Hungary, Slovakia and Bulgaria and former USSR sub-type: Estonia, Ukraine, Russia and Lithuania). The design weight (dweight) was applied, in order to correct the different probabilities of selection and to make the sample more representative.

4 Results

Representative results of the transition matrices estimated for different welfare systems are depicted in Table 1. For 2004 in the social democratic, conservative-corporatist and post-communist (former USSR type) welfare states, individuals with low educational parents are more likely to move to the next educational category, while individuals with parents of middle and high educational level are most

¹ <http://www.europeansocialsurvey.org>

likely to remain in the same educational categories as those of their parents. Similar transitions are observed in the European post-communist regime with one difference: the downward transition probability relating to individuals whose mother has a high level of education. In the south-European and liberal welfare states, individuals with highly educated parents are more likely to remain in the same state, while those with low-level parents are more likely to remain in the lower education category. Individuals with medium educated parents are more likely to have a downward shift in liberal regimes, while in the south-European states, in the case of the father, they are more likely to move upwards, while in the mother's, remaining at the same level is more likely.

Table 1. Educational intergenerational transition probabilities, father and mother, ESS, 2004

Regime	Ed. level of father/ ed. level of respondent			Ed. level of mother/ ed. level of respondent		
	Low	Medium	High	Low	Medium	High
Social democratic						
Norway						
Low	0.309*	0.428	0.263	0.291	0.438	0.271
Medium	0.133	0.510	0.357	0.104	0.462	0.434
High	0.102	0.292	0.605	0.136	0.341	0.523
Conservative						
Germany						
Low	0.323	0.559	0.118	0.213	0.613	0.174
Medium	0.134	0.665	0.201	0.123	0.589	0.288
High	0.136	0.408	0.456	0.242	0.344	0.414
South-European						
Portugal						
Low	0.783	0.154	0.064	0.783	0.147	0.071
Medium	0.233	0.302	0.465	0.189	0.453	0.358
High	0.232	0.275	0.493	0.180	0.320	0.500
Liberal						
UK						
Low	0.702	0.153	0.145	0.697	0.145	0.158
Medium	0.516	0.226	0.258	0.447	0.224	0.329
High	0.300	0.251	0.450	0.286	0.261	0.453
Post-socialist – European type						
Poland						
Low	0.346	0.606	0.048	0.334	0.612	0.055
Medium	0.203	0.665	0.132	0.216	0.625	0.159
High	0.140	0.376	0.484	0.211	0.408	0.382
Post-socialist – USSR type						
Estonia						
Low	0.291	0.496	0.213	0.292	0.497	0.212
Medium	0.225	0.438	0.337	0.205	0.471	0.324
High	0.155	0.356	0.490	0.222	0.330	0.449

* Probability of the respondent having a low educational level given that his/her father has a low educational level in Norway, 2004. The remainder numbers depicted in the Table reflect respective probabilities.

In Table 2 the transition probabilities for 2016 show small differences compared to those of 2004. In the social democratic states, intergenerational educational mobility is extremely high, as respondents with parents of all educational levels are more likely to achieve a higher level of education. The transitions observed in the post-communist countries of the former USSR are also on the rise, with people with parents with low and middle level education moving to the next level and individuals with high educational attainments remaining in the same state. In conservative-corporate welfare regime countries, individuals with low-level parents are more likely to move to the middle level, while respondents with high-level parents are more likely to achieve a level of education similar to that of their parents. With regard to the second educational level, the influence of the father is most likely to lead to immobility, while the influence of the mother seems to lead to an upward transition. In liberal and post-communist European social states, people with low and high levels of education are more likely to stay at the same level of education as their parents, while people with middle-level parents are more likely to succeed an upward transition. Intergenerational educational mobility in the south-European model appears to be extremely limited, as the only possible movement observed concerns individuals with a middle-level education of the father.

Table 2. Educational intergenerational transition probabilities, father and mother, ESS, 2016

Regime	Ed. level of father/ ed. level of respondent			Ed. level of mother/ ed. level of respondent		
	Low	Medium	High	Low	Medium	High
Social democratic						
Norway						
Low	0.252*	0.329	0.419	0.225	0.344	0.431
Medium	0.151	0.402	0.446	0.132	0.341	0.527
High	0.096	0.237	0.667	0.125	0.254	0.621
Conservative						
Germany						
Low	0.271	0.381	0.348	0.163	0.451	0.386
Medium	0.092	0.482	0.426	0.086	0.416	0.498
High	0.109	0.256	0.636	0.140	0.254	0.606
South-European						
Portugal						
Low	0.589	0.204	0.208	0.592	0.206	0.202
Medium	0.193	0.398	0.410	0.236	0.389	0.375
High	0.087	0.261	0.652	0.151	0.205	0.644
Liberal						
UK						
Low	0.406	0.214	0.379	0.393	0.218	0.389
Medium	0.152	0.255	0.593	0.086	0.278	0.636
High	0.106	0.218	0.675	0.129	0.205	0.666
Post-socialist – European type						
Poland						
Low	0.511	0.291	0.198	0.537	0.283	0.180
Medium	0.178	0.357	0.465	0.170	0.360	0.470
High	0.157	0.200	0.643	0.192	0.260	0.548
Post-socialist – USSR type						
Estonia						
Low	0.254	0.396	0.349	0.278	0.408	0.315
Medium	0.157	0.401	0.441	0.143	0.421	0.436
High	0.074	0.230	0.696	0.102	0.248	0.650

* Probability of the respondent having a low educational level given that his/her father has a low educational level in Norway, 2016. The remainder numbers depicted in the Table reflect respective probabilities.

We now focus on estimating the intergenerational mobility of individuals in Europe and its evolution for the years 2002 to 2016. More specifically, three different mobility indices are calculated in order to reveal the extent of the transitions within generations. The ones used in the present analysis are the well-established mobility indices:

The Prais – Shorrocks mobility index (Prais [23]; Shorrocks [24]): $M_{PS} = \frac{1}{n-1} (n - tr(\mathbf{P}))$, where n is the number of states and $tr(\mathbf{P})$ denotes the trace of the transition matrix \mathbf{P} , i.e. the sum of its diagonal elements.

The Bartholomew [3] mobility index defined by $M_B = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n p_{ij} |i - j|$ and the upward mobility index defined by $u = \frac{1}{N} \sum_{j>i} n_{ij}$, where n_{ij} is the absolute number of individuals with the j -th educational level whose father/mother has the i -th educational level and N denotes the total number of respondents.

Figures 1 and 2 show the results of the Prais-Shorrocks index for intergenerational educational mobility from 2002 to 2016. It is clear that, in all social protection systems, intergenerational educational mobility in relation to the mother's educational level is higher than that of the father. More specifically,

regarding the influence of the father, it is observed that until 2006 there is a tendency of convergence between the results of liberal and social democratic social states, while after 2006 the educational mobility in liberal regimes is much lower. Similar results are presented between post-communist and conservative social states for most years under study. Overall, the social democratic social states have consistently one of the highest levels of mobility, while the south-European states have by far the lowest educational mobility.

The above trend of higher and lower educational mobility of social democratic and south-European countries is also observed in relation to the educational level of the mother. Regarding that transition, very high mobility is observed in the conservative regimes, while the post-communist and liberal social states are placed in an intermediate category.

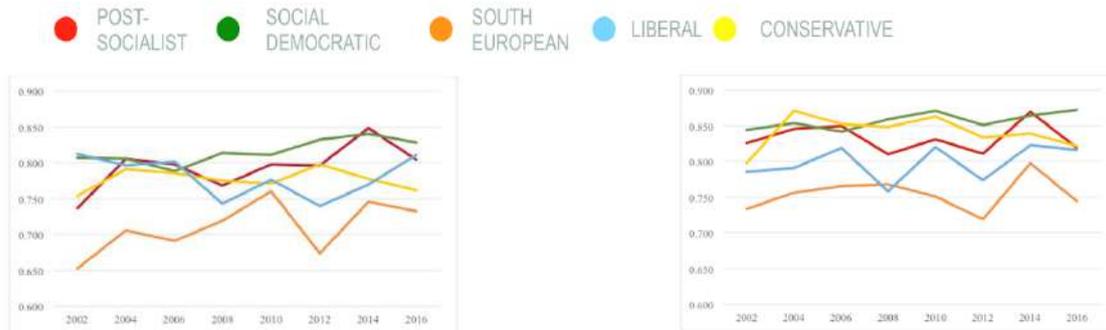


Fig. 1. M_{PS} in relation to the educational level of father, Rounds 1-8

Fig. 2. M_{PS} in relation to the educational level of mother, Rounds 1-8

Figures 3 and 4 depict the Bartholomew mobility index for intergenerational educational mobility from 2002 to 2016. In relation to the effect exerted by the father's educational level, the highest mobility is observed in the liberal states until 2008 and in the social democratic states for the remaining years. A relative convergence of results is seen between post-communist and conservative welfare states, while for the south-European regimes are much lower. Regarding the effect of the mother's level of education, the highest mobility is observed in the social democratic welfare systems, while the conservative, post-communist and liberal regimes show similar results. The south-European countries consistently show the lowest intergenerational educational mobility.

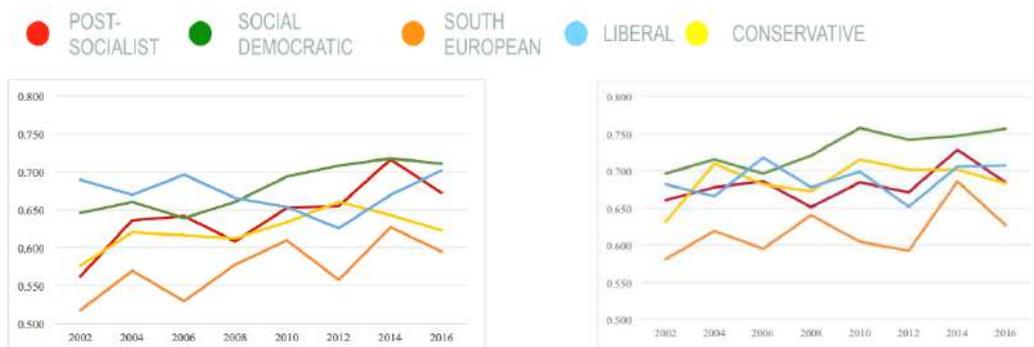


Fig. 3. M_B in relation to the educational level of father, Rounds 1-8

Fig. 4. M_B in relation to the educational level of mother, Rounds 1-8

Finally, Figures 5 and 6 show the upward mobility index in relation to the educational level of the father and mother for the years 2002 to 2016. Regarding the influence of the father's level of education, the highest educational upgrade is in the social democratic, liberal and post-communist countries. Conservative regimes are next, while the south-European countries exhibit the lowest mobility. In relation to the educational level of the mother, the results differ significantly. For all the years under study, the highest educational transitions take place in the social democratic and conservative systems. A relatively high upward mobility is recorded in the post-communist regimes until 2008, while after 2008 quite high mobility is recorded in the liberal states. In the south-European social protection systems, the lowest upward mobility is again identified.

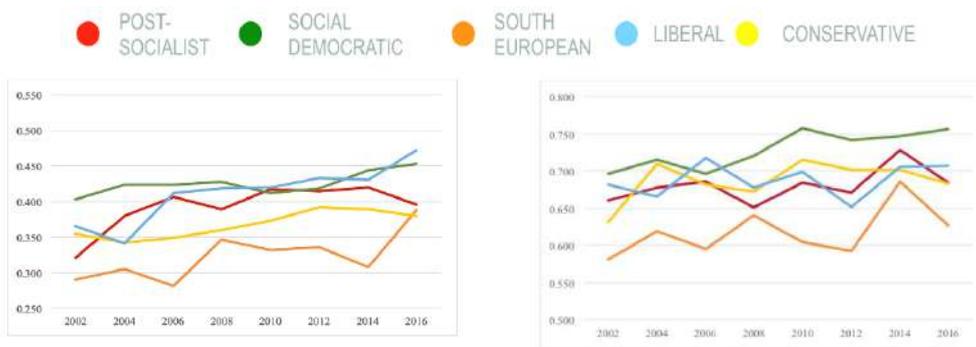


Fig. 5. Upward mobility index in relation to the educational level of father, Rounds 1-8

Fig. 6. Upward mobility index in relation to the educational level of mother, Rounds 1-8

Conclusions and further research

The aim of this study was to examine the effectiveness of European social protection systems in enhancing intergenerational educational mobility. Using primary data from Rounds 1-8 of the European Social Survey, the analysis focused on comparing the impact of the immediate family educational environment on the educational outcomes of individuals for social democratic countries (Denmark, Finland, the Netherlands, Sweden and Norway), conservative-corporatist (Austria, Belgium, France and Germany), liberal (United Kingdom and Ireland), south-European (Greece, Italy, Spain and Portugal) and post-socialist countries (Poland, Czech Republic, Hungary, Slovakia, Bulgaria, Ukraine, Russia, Estonia and Lithuania).

In order to obtain a first picture of the educational intergenerational fluidity of the systems under study, the Prais-Shorrocks mobility index was calculated. Based on the results of the index, it was observed that all social states had high levels of fluidity compared to the levels of education of both parents. Considering the findings of the educational levels of both mother and father, the most “fair” systems are those with a social democratic regime. In the next positions are the post-communist and conservative social states, presenting quite high levels of mobility, while with a small difference, the liberal systems follow in the ranking. The systems of the south-European model have the lowest performance.

Bartholomew’s mobility index was also calculated to consider the distances travelled. The findings prove relatively high efficiency for all systems under study. More specifically, in relation to the educational levels of both parents, the social democratic regimes show the highest efficiency, while in the ranking the regimes of the liberal, conservative and post-communist models follow with similar results. In the south-European countries, the level of education of the family of origin exerts the greatest influence on the distances of travel.

Through the calculation of the upward mobility index, the clearest possible impression of the educational upgrades was attempted. For all systems studied, the values of the index show moderate mobility. Higher education upgrades are taking place in the social democratic welfare states, while the results of the conservative, liberal and post-communist regimes are close to those of the countries belonging to the social democratic regime. In the south-European countries, the results of educational upgrades are quite limited, which proves the strong influence of parental educational background on the educational achievements of individuals.

It is clear that social democratic welfare states, which are characterized by extensive universal interventions and high levels of redistribution, achieve the highest attenuation of intergenerational educational immobility. Liberal regimes significantly reduce the impact of family backgrounds on educational success, as (although generally characterized by low and selective social benefits) in the field of their educational policy they significantly promote equal opportunities by presenting very high government spending. Slightly lower than those of the liberal states, are the results of the post-communist and conservative social states, which are characterized by extended social policy benefits, while simultaneously have high standardization and stratification levels in their educational models. Southern-European welfare states, where social protection is extremely residual and the burden is significantly borne by the family, show the highest educational stagnation, hindering individuals’ efforts for educational upgrading and creating “dependency paths” between intergenerational successes.

It is recognized that the present study does not take into account all the complexity of the factors that may affect the intergenerational educational movements of individuals, nor the complexity of all the effects that the specific social protection system has on them. Although an attempt was made to examine the correlation between intergenerational educational mobility and economic inequality (of all years and countries under study), both with indicators of economic inequality and poverty from EUROSTAT and OECD databases, and with the calculation of poverty indicators from income variables in the ESS, the results did not show the required levels of statistical significance. Further investigation of the phenomenon is indicated. In conclusion, what we consider to be clear from the present analysis and is of utmost importance is the fact that social interventions in the intergenerational cycle can serve as essential defence factors in the promotion of educational equality and mobility, when the policies implemented give the required weight primarily on creating equal conditions and secondarily on creating equal opportunities.

References

1. J. Allmendinger. Educational Systems and Labour Market Outcomes. *European Sociological Review*, 5, 3, 231{250, 1989.
2. W. Atkinson. Beck, individualization and the death of class: A critique. *The British Journal of Sociology*, 58, 3, 349{366, 2007.
3. D.J. Bartholomew. *Stochastic Models for social processes*. Wiley, London, 1982.
4. J. Blanden, P. Gregg and S. Machin. *Intergenerational Mobility in Europe and North America*. Centre for Economic Performance, London, 2005.
5. P. Blau and O. Duncan. *The American Occupational Structure*. Wiley, New York, 1976.
6. R. Breen and J. Jonsson. Inequality of opportunity in comparative perspective: Recent Research on Educational Attainment and Social Mobility. *Annual Review of Sociology*, 31, 223{243, 2005.
7. A. Chevalier, K. Denny and D. McMahon. *A Multi-country Study of Inter-generational Educational Mobility*. Institute for the Study of Social Change, Tasmania, 2003.
8. R. Erikson and J. Goldthorpe. *The constant flux. A study of class mobility in industrial societies*. Clarendon, Oxford, 1993.
9. G. Esping-Andersen. *The three worlds of welfare capitalism*. Policy, Oxford, 1990.
10. H. J. M. Fenger. Welfare regimes in Central and Eastern Europe: Incorporating post-communist countries in a welfare regime typology. *Contemporary Issues and Ideas in Social Sciences*, 3, 2, 1{30, 2007.
11. M. Ferrera. The Southern Model of Welfare State in Social Europe. *Journal of European Social Policy*, 6, 1, 17{37, 1996.
12. D. Glass. *Social Mobility in Britain*. Routledge, London, 1954.
13. A. Heath. *Social Mobility*. Fontana, London, 1981.
14. H. Hecllo. *The welfare state in hard times*. APSA, Washington, 1985.
15. G. Hega and K. Hokenmaier. The Welfare State and Education: A Comparison of Social and Educational Policy in Advanced Industrial Societies. *German Policy Studies*, 2, 1, 143{173, 2002.
16. A. J. Heidenheimer. Education and social security entitlements in Europe and America, in: *The development of welfare states in Europe and America* (P. Flora & A.J. Heidenheimer (Ed)). Transaction, London, 1981.
17. J. H, P. Hoffmayer-Zlotnik and U. Warner. How to Survey Education for Cross-National Comparisons: The Hoffmayer-Zlotnik/Warner-Matrix of Education. *Metodoloski Zvezki*, 4, 2, 117{148, 2007.
18. OECD. *A Family Affair: Intergenerational Social Mobility across OECD Countries*, in: *Economic Policy Reforms Going for Growth*. OECD, Paris, 2010.
19. S. Papanastasiou. *Intergenerational social mobility and types of welfare state in Europe*. Gutenberg, Athens, 2018.
20. Ch. Papatheodorou and S. Papanastasiou. Family origin and poverty in EU countries: The role of social protection systems, in: *Social Research Paths* (M. Petmezidou and T. Kallinikaki (Ed)). Pattern, Athens, 2016.
21. Ch. Papatheodorou and S. Papanastasiou. *Intergenerational mobility in the EU*. INE-GSEE, Athens, 2011.
22. T. Peter, J. D. Edgerton and L. W. Roberts. Welfare regimes and educational inequality: a cross-national exploration. *International Studies in Sociology of Education*, 20, 3, 241{264, 2010.
23. S. Prajs. Measuring social mobility. *Journal of the Royal Statistical Society, Series A*, 118, 56{66, 1955.
24. Shorrocks. The measurement of social mobility. *Econometrica*, 46, 1013{1024, 1978.
25. G. Stamatopoulou, M. Symeonaki and C. Michalopoulou. Intergenerational transmission of education in Greece: evidence from the European Social Survey 2002-2010, 15th Conference of the Applied Stochastic Models and Data Analysis International Society (ASMDA), Barcelona, Spain, 25-28 June, 2013.
26. M. Symeonaki, O. Filopoulou and G. Stamatopoulou. Measuring Intergenerational Educational Mobility in Greece, 14th Conference of the Applied Stochastic Models and Data Analysis International Society (ASMDA), Rome, Italy, 7-10 June, 2011.

27. M. Symeonaki and G. Stamatopoulou. Exploring intergenerational educational mobility in Greece with data drawn from EU-SILC. Proceedings of the University of Cyprus Conference on Social Justice and Participation: The Role of Higher Education, Cyprus, 2011.
28. M. Symeonaki and G. Stamatopoulou. Exploring the transition to Higher Education in Greece: issues of intergenerational educational mobility, *Policy Futures in Education*, 12, 5, 681{694, 2014.
29. M. Symeonaki and G. Stamatopoulou. Intergenerational mobility as a distance measure between probability distribution functions, in: *Theoretical and Applied Issues in Statistics and Demography* (C. H. Skiadas (Ed)), 2014.
30. M. Symeonaki, G. Stamatopoulou and C. Michalopoulou. Intergenerational Occupational Mobility in Greece: Evidence from EU-SILC. Demographic Analysis and Research International Conference, Chania, Greece, 5-8 June 2012.
31. United Nations Educational, Scientific and Cultural Organization Institute for Statistics. International Standard Classification of Education ISCED 2011. UNESCO Institute for Statistics, Canada, 2012.
32. A. West and R. Nikolai. Welfare Regimes and Education Regimes: Equality of Opportunity and Expenditure in the EU (and US). Cambridge University Press, 42, 3, 469{493, 2013.
33. L. Woessmann. How Equal Are Educational Opportunities? Family Background and Student Achievement in Europe and the United States. CESifo (No. 1162), Munich, 2004.

Estimating alcohol-attributable mortality in Czechia

Vrabcová Jana¹, Pechholdová Markéta², and Svačinová Kornélia³

¹ Department of Statistics and Probability, Prague University of Economics and Business, Czech Republic (E-mail: vrabcova.jana@post.cz)

² Department of Demography, Prague University of Economics and Business, Czech Republic (E-mail: marketa.pechholdova@seznam.cz)

³ Department of Demography, Prague University of Economics and Business, Czech Republic (E-mail: k.csefalvaiova@seznam.cz)

Abstract.

Background: Czechia ranks among countries with the highest alcohol consumption worldwide in both men and women. High although decreasing alcohol consumption among young people are also specific for the Czech population. The drinking culture is however mainly built on regular drinking and consumption of drinks with low alcohol content, such as beer or wine. **Aims:** This paper assesses the burden of alcohol consumption on population health. Mortality associated with alcohol consumption, both directly and indirectly, is assessed and compared with previous studies. **Data and methods:** Alcohol-attributable fractions (AAFs) were computed based on the 2017 routine mortality data and data on alcohol consumption derived from the 2017 Czech Household Panel Survey. **Results:** Overall, 4.3% of deaths were attributable to alcohol consumption in Czechia in 2017, with 6.8% in men and 1.7% in women. The impact of alcohol varies by age with younger ages being more affected in both sexes by direct effects whereas among elderly, indirect effects are more pronounced. **Conclusions:** Compared to previous studies, alcohol-related harm in Czechia has decreased within the past decade, but remains high compared to the Western countries.

Keywords: Alcohol, Mortality, Morbidity, Alcohol-Attributable Fractions, Czechia.

1 Introduction

Alcohol consumption is associated with morbidity and mortality. Alcohol use is a leading risk factor for global disease burden and causes substantial health loss (GBD,2016). In 2016, the harmful use of alcohol resulted in some 3 million deaths (5.3% of all deaths) worldwide. Over 200 health conditions are linked to harmful alcohol use, ranging from liver diseases, road injuries and violence, to cancers, cardiovascular diseases, suicides, tuberculosis and HIV/AIDS (WHO, 2018). In 2016, of all deaths attributable to alcohol consumption worldwide, 28.7% were due to injuries, 21.3% due to digestive diseases, 19% due to cardiovascular diseases, 12.9% due to infectious diseases, and 12.6% due to cancers (Mravčík et al., 2019).

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



In the international comparison, Czechia has long been characterized by high alcohol consumption. Beer and drinking alcohol are considered an important part of Czech culture, society and history, and the beer industry is seen as part of the national heritage (Dzúrová et al., 2010). According to the European School Survey Project on Alcohol and Other Drugs, Czech young people consume alcohol significantly more often than in the other 28 European countries that participated in this study (Csémy et al., 2006). Within the HBSC study in 2014, the alcohol consumption in the last 30 days was reported by 8% girls and 15% boys of age 11. According to the ESPAD study carried out in 2015 and 2016, 69% and 65% of the 16-year-olds, respectively, reported any alcohol use in the last 30 days (Csémy, et al., Chomynová et al., 2016, Mravčík et al., 2019)

Figure 1 shows the consumption of pure alcohol in Czechia in the period 1948-2018. Until the early 1980s, overall alcohol consumption has been increasing. Consumption of alcohol in activities that could endanger health, life or property has been prohibited under the communist rule. In the 1960s, the fight against excessive alcohol consumption and domestic alcohol production began. In 1973, the Ministry of Industry and Trade ordered a ban on the sale of alcohol before 10 a.m. Despite these strict measures, Czech society still remained very tolerant of alcohol consumption (Hnilicová et al., 2017).

In 1985, the so-called “Dry law” was introduced. In 1986-88 the ruling Communist party tried to reduce alcohol consumption in the country by a Gorbachev-inspired anti-alcohol campaign. Alcohol consumption in workplaces was no longer tolerated and the display of alcohol in shop windows was reduced. All these restrictions resulted in a decrease in alcohol consumption during the 1980s (Hnilicová et al., 2017).

After 1989 the re-establishment of democracy and a market economy abolished all political control over drinking and deregulated prices of beverage alcohol. After the revolution in 1989, the price of spirits rose more slowly than other alcoholic beverages, which is probably the reason for a greater increase in the consumption of spirits in the early 1990s compared to beer and wine. (Kubička et al., 1998). Political and economic changes have not focused much on alcohol policies. The social and political environment is becoming very liberal towards alcohol, which has an impact on its consumption. Undoubtedly, the increase is also due to the fact that in many cases the post-revolutionary euphoria was replaced by fear of economic and political change. After a few years’ alcohol consumption returned to its original level and continued to rise slowly. After 2000, the consumption of spirits decreased, wine consumption stagnated and beer consumption increased. In the last decade there has been some stagnation of alcohol consumption overall, there is a slight increase in wine consumption, on the contrary, a slight decrease in beer consumption.

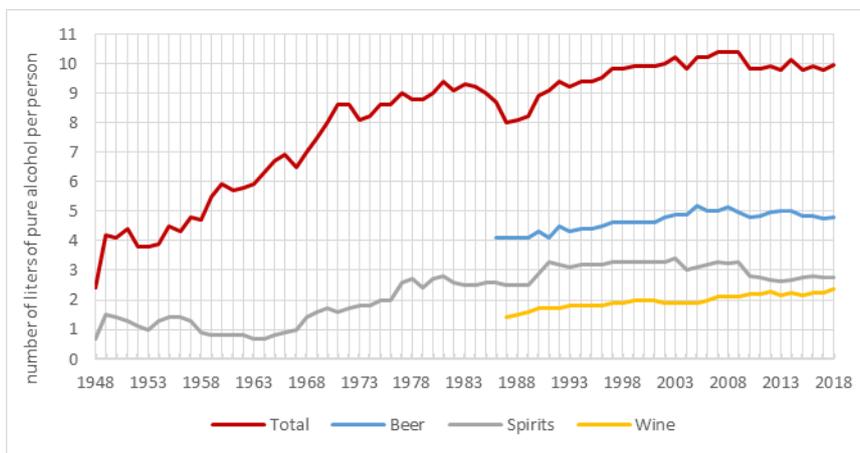


Fig. 1. Alcohol consumption in Czechia in 1948-2018
Data source: Czech Statistical Office

Social tolerance to alcohol consumption is quite high in Czechia (and also in other European countries) and so far, alcohol consumption has not been legally or practically restricted either by the Government or by any social organization. Long-term tolerance to alcohol both in cultural and political environment resulted in unfavourable trend of increasing alcohol-related mortality for both sexes. According to information provided in the Czech National Strategy for the Prevention and Reduction of Damage Related to Addictive Behavior 2019–2027, tobacco and alcohol use is one of the main causes of morbidity and mortality in the Czech Republic, with almost 20% of total mortality attributable to smoking; about 6% due to alcohol use. The greatest health burden related to tobacco and alcohol is seen in middle-aged and elderly people (SZÚ, 2018). Mravčík et al., 2019 compares the results of several surveys and found, that the excessive alcohol use has been consistently 3 times higher among males than females in all surveys. Significant differences were observed in the alcohol use by age groups – while the use of excessive amounts of alcohol on one occasion has been much more prevalent among younger age groups, especially among young adults (aged 15–34 years), the daily alcohol use has been increasing with age, being the highest among 55–64-year-olds and 65 and older respondents.

Routine mortality statistics capture the alcohol burden only partially, through the so-called direct effects of alcohol. Epidemiological studies point at indirect effects of alcohol in many other medical conditions (accidents, violence, hypertension, ischemic heart disease, neoplasms). Statistically, these indirect effects can be measured with the so-called population-attributable fractions (PAF). The PAF method necessarily requires information on alcohol consumption patterns, namely the distribution of population according to the amount of alcohol consumed (the extent of indirect alcohol-related harm depends on the frequency and the amount of alcohol consumption). The present

paper aims at quantifying the alcohol-attributable mortality using the 2017-year data. In order to compare with the past, we are applying similar design of the alcohol-attributable mortality measurement as a previous study covering the period 2008 (Kohoutová, 2012).

2 Data and methods

Mortality data were obtained from routine vital statistics by the Czech statistical office, same as population exposures for 2017. Data on alcohol consumption survey were retrieved from the Czech Household Panel Survey (CHPS) conducted in four waves on a panel of Czech households between 2015 and 2018. The present study is based on the third wave of the survey from the year 2017. Individual personal data for persons aged 18 and over were collected from self-reported paper questionnaire (PAPI). Individual alcohol consumption was measured based on a combination of frequency-quantity questions. The frequency question asked “How often do you consume alcoholic beverages?” (more than once a day, daily, several times a week, once a week, several times a month, less frequently never). The quantity questions asked “How many alcoholic drinks do you consume on one occasion on average?”, with answers scaled between 1 and 11+. One alcoholic drink was defined as either one beer, one glass of wine or one shot of spirit. According to the content of alcohol, one drink corresponds to approximately 18 grams of pure ethanol. A total of 5,036 individuals provided answer about the alcohol consumption in the CHPS. Tab. 1 provides an overview of the shares of the CHPS respondents according to the age, sex and alcohol consumption category.

Age	Men					Women				
	0	1-19	20-39	40-74	75+	0	1-19	20-39	40-74	75+
15-24	10%	51%	21%	15%	3%	20%	63%	11%	6%	0%
25-34	9%	50%	21%	15%	6%	21%	61%	13%	4%	1%
35-44	5%	45%	24%	20%	5%	20%	57%	18%	4%	1%
45-54	7%	44%	24%	19%	7%	24%	59%	13%	3%	0%
55-64	6%	39%	29%	16%	10%	30%	52%	15%	2%	1%
65-74	7%	40%	24%	21%	8%	39%	47%	12%	2%	0%
75+	13%	50%	18%	14%	5%	60%	30%	9%	0%	0%

Tab. 1. Representation of the share of CHPS respondents by age, sex and alcohol consumption

Data source: CHPS, author’s calculation

To calculate the effect of alcohol, the alcohol-attributable fractions (AAF) were computed, i.e. the proportion of events that would not have occurred if the population had not been exposed to alcohol. There are two ways

to calculate AAF - direct and indirect. The direct method can be used if the numbers of alcohol-related events are known.

The direct method of calculating AAF was used to determine the proportion of transport accidents and attacks under the influence of alcohol. Since 2000, police statistics on crime in Czechia have been available, which monitor the number of acts committed under the influence of alcohol and since 2003 statistics on transport accidents under the influence of alcohol have been available (Police of the Czech Republic, 2000-2011, Kohoutová, 2012). The proportion of assaults (X85-Y09) under the influence of alcohol was used as a direct AAF. As the AAF for transport accident (V01-99), the proportion of deaths caused by alcohol-related transport accidents in the total number of deaths in transport accidents in the given year was used (see Tab. 2).

<i>Cause (code)</i>	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Transport accidents (V01-V99)	0.06	0.06	0.06	0.08	0.05	0.05	0.04	0.03	0.08	0.15	0.14
Assaults (X85-Y09)	0.14	0.16	0.23	0.16	0.26	0.22	0.26	0.22	0.32	0.23	0.27

Tab. 2. Attributable fractions calculated by the direct method for deaths due to transport accidents and assaults in Czechia, 2000-2010

Data source: Police of the Czech Republic, 2000-2011, Kohoutová, 2012

<i>ICD10</i>	<i>Condition</i>
E24.4	Alcohol-induced pseudo-Cushing syndrome
F10	Mental and behavioural disorders due to use of alcohol
G31.2	Degeneration of nervous system due to alcohol
G62.1	Alcoholic polyneuropathy
G72.1	Alcoholic myopathy
I42.6	Alcoholic cardiomyopathy
K29.2	Alcoholic gastritis
K70	Alcoholic liver disease
K86.0	Alcohol-induced chronic pancreatitis
T51, X45, X65, Y15	Toxic effects of substances chiefly nonmedicinal as to source, Accidental poisoning and intentional self-poisoning by and exposure to alcohol, Poisoning by and exposure to alcohol, undetermined intent

Tab. 3. List of conditions wholly attributable to alcohol

Data source: ÚZIS, 2011, Kohoutová, 2012

However, for most events, direct data is not available and an indirect method must be used. The indirect method calculates AAF by combining the prevalence of alcohol consumption in the population and relative risks obtained from the results of epidemiological studies. The relative risk is the probability of an event occurring in alcohol consumers compared to the probability of its occurrence in the reference group, such as abstainers in this case. The AAF is a positive function of the prevalence of drinking and the relative risk function of each alcohol-related condition, and its calculation enables

the estimation of the proportion of cases of a disease or type of injury that may be attributed to the consumption of alcohol (Jones, et al., 2008). The following causes of death are identified by the Institute of Health Information of the Czech Republic as causes wholly attributable to alcohol (see Tab. 3).

Tab. 4 summarizes partially attributable causes of death considered in the analysis along with the respective relative risks stratified by alcohol consumption category.

ICD10	Condition	Men				Women				Source
		1-19	20-39	40-74	75+	1-19	20-39	40-74	75+	
C00-C14	Malignant neoplasms of lip, oral cavity and pharynx	1.45	1.45	1.85	5.39	1.45	1.85	5.39	5.39	Rehm et al., 2004
C15	Malignant neoplasm of oesophagus	1.80	1.80	2.38	4.36	1.80	2.38	4.36	4.36	Rehm et al., 2004
C18	Malignant neoplasm of colon	1.03	1.05	1.10	1.21	1.03	1.05	1.10	1.21	Jones et al., 2008
C20	Malignant neoplasm of rectum	1.05	1.09	1.19	1.42	1.05	1.09	1.19	1.42	Jones et al., 2008
C22	Malignant neoplasm of liver and intrahepatic bile ducts	1.45	1.45	3.03	3.60	1.45	3.03	3.60	3.60	Rehm et al., 2004
C32	Malignant neoplasm of larynx	1.22	1.43	2.02	3.86	1.22	1.43	2.02	3.86	Jones et al., 2008
C50	Malignant neoplasm of breast up to 45 years	-	-	-	-	1.15	1.41	1.46	1.46	Rehm et al., 2004
	Malignant neoplasm of breast over 45 years	-	-	-	-	1.14	1.38	1.62	1.62	
D00-D48	Other neoplasms	1.10	1.10	1.30	1.70	1.10	1.30	1.70	1.70	Rehm et al., 2004
E10-E14	Diabetes mellitus	1.00	1.00	0.57	0.73	0.92	0.87	1.13	1.13	Rehm et al., 2004
G40-G41	Epilepsy, Status epilepticus	1.23	1.23	7.52	6.83	1.34	7.22	7.52	7.52	Rehm et al., 2004
I10-I15	Hypertensive diseases	1.15	1.43	2.04	4.15	1.15	1.43	2.04	4.15	Jones et al., 2008
I20-I25	Ischaemic heart diseases	0.82	0.85	0.98	1.53	0.85	0.90	1.10	1.87	Rehm et al., 2004
I47-I48	Cardiac arrhythmias	1.51	1.51	2.23	2.23	1.51	2.23	2.23	2.23	Jones et al., 2008
I60-I62, I69.0-I69.2	Haemorrhagic stroke	1.10	1.19	1.82	4.70	1.10	1.19	1.82	4.70	Jones et al., 2008
I63-I66, I69.3-I69.4	Ischaemic stroke	0.85	0.90	1.17	4.37	0.85	0.90	1.17	4.37	Jones et al., 2008
I85	Oesophageal varices	1.95	2.90	7.13	26.53	1.95	2.90	7.13	26.53	Jones et al., 2008
K73	chronic hepatitis, not elsewhere classified	1.95	2.90	7.13	26.53	1.95	2.90	7.13	26.53	Jones et al., 2008
K74	liver fibrosis and cirrhosis	1.30	1.30	9.50	13.00	1.30	9.50	13.00	13.00	Rehm et al., 2004
K80	Cholelithiasis	0.82	0.82	0.68	0.50	0.82	0.68	0.50	0.50	Jones et al., 2008
K85, K86.1	Acute and chronic pancreatitis	1.12	1.34	1.78	3.19	1.12	1.34	1.78	3.19	Jones et al., 2008
O03	Spontaneous abortion	-	-	-	-	1.20	1.76	1.79	1.49	Jones et al., 2008

Tab. 4. Estimates of relative risks for diseases partially caused by alcohol, by sex and individual categories of alcohol consumption (grams / day)

Note: a) for category 40-50 g/day, b) for category 60 and more g/day

Data source: Rehm et al., 2004, Jones et al., 2008, Kohoutová, 2012

AAFs were calculated across five categories of alcohol consumption (the same as in the table 2): abstainers, 1 to 19 g/day, 20 to 39 g/day, 40 to 74 g/day and 75 or more g/day according to the following formula:

$$AAF = \frac{\sum_{i=1}^k P_i (RR_i - 1)}{\sum_{i=0}^k P_i (RR_i - 1) + 1}$$

where i means a category according to the level of alcohol consumption (for abstainers $i = 0$), RR_i expresses the relative risk of occurrence for the category of consumption and P_i represents the proportion of the population exposed in each group of alcohol consumption in Czechia (Jones et al., 2008).

AAFs are calculated for each age group and separately for men and women.

3 Results

Alcohol-attributable mortality

A total of 4,750 deaths in Czechia in 2017 were caused by alcohol consumption (4.3% of all deaths); 1,650 deaths occurred from conditions fully caused by alcohol consumption, 2,989 were caused in part by alcohol consumption, and 110 deaths were due to acute consequences. In men, 6.8% of deaths were caused by alcohol, while in women it was only 1.7% of all deaths. Although most deaths occur in older age groups, young people are also greatly affected by alcohol consumption. For example, in men aged 25-34, it is estimated that 9.4% of all deaths were due to alcohol consumption, compared with 3.6% of those aged 75 and over.

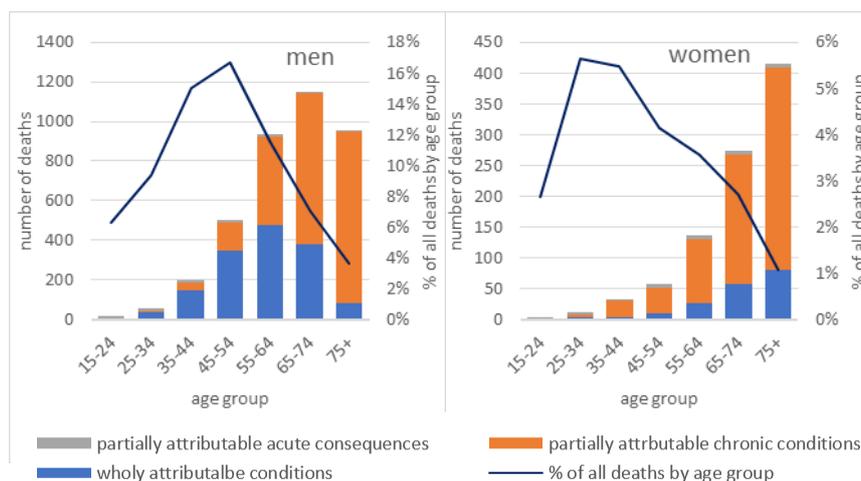


Fig. 2. Number (% of all deaths in each age group) of deaths attributable to alcohol consumption by sex, age and type of condition in Czechia in 2017
Data source: CZSO, CHPS, Rehm et al., 2004, Jones et al., 2008, Kohoutová, 2012

It is similar for women, with the highest proportion of deaths due to alcohol in the age group 25-34 (5.6%) compared to women aged 75 and over (1.1%) (see **Error! Reference source not found.**). In the youngest age groups, men are more at risk of deaths directly affected by alcohol (AAF = 1). The distribution of deaths caused entirely by alcohol differs between men and women. In men, the mode (474 cases) is in the age group 55-64 years, where they account for 51% of alcohol-attributable deaths. On the contrary, the highest proportion of deaths caused wholly by alcohol in women is in the age group 75+ (19% of alcohol-attributable deaths). In both men and women, the number of deaths due to partially attributable chronic conditions increases with age. Partially attributable acute consequences (transport accidents and injuries) make up a relatively small proportion of all alcohol-attributable deaths, but if we look at men aged 15-24, these deaths are the most common (70% of all alcohol-attributable deaths).

Alcohol-attributable deaths by cause

The three most common causes of alcohol-attributable deaths by age group are listed in the Tab. 5.

age group	men		women	
	condition	N	condition	N
15-24	Transport accidents	81	Epilepsy, Status epilepticus	1
	Assault	4	Haemorrhagic stroke	0
	Epilepsy, Status epilepticus	2	Other neoplasms	0
25-34	Transport accidents	87	Transport accidents	17
	Alcoholic liver disease	19	Assault	5
	Accidental poisoning by and exposure to alcohol	12	Malignant neoplasm of breast	2
35-44	Alcoholic liver disease	95	Malignant neoplasm of breast	11
	Transport accidents	74	Transport accidents	7
	Accidental poisoning by and exposure to alcohol	34	Epilepsy, Status epilepticus	4
45-54	Alcoholic liver disease	250	Transport accidents	24
	Transport accidents	73	Malignant neoplasm of breast	16
	Accidental poisoning by and exposure to alcohol	65	Assault	8
55-64	Alcoholic liver disease	355	Transport accidents	36
	Malignant neoplasms of lip, oral cavity and pharynx	92	Malignant neoplasm of breast	32
	Transport accidents	82	Malignant neoplasm of liver and intrahepatic bile ducts	16
65-74	Alcoholic liver disease	310	Malignant neoplasm of breast	53
	Hypertensive diseases	128	Malignant neoplasm of liver and intrahepatic bile ducts	39
	Malignant neoplasm of liver and intrahepatic bile ducts	108	Hypertensive diseases	33
75+	Hypertensive diseases	255	Hypertensive diseases	139
	Ischaemic stroke	124	Malignant neoplasm of breast	52
	Cardiac arrhythmias	120	Degeneration of nervous system due to alcohol	47

Tab. 5. Top three causes of alcohol-attributable deaths by age groups and sex in Czechia in 2017

Data source: CZSO, CHPS, Rehm et al., 2004, Jones et al., 2008, Kohoutová, 2012

Transport accidents are the most common alcohol-attributable cause of death for younger men (15-24 years), and in women aged 25-34 and 45-64 years. Malignant neoplasm of breast is the most common among women aged 35-44. Malignant neoplasm of breast, together with traffic accidents, remain in top 3 alcohol-attributable causes for women till age 75, at which age it is their turn of hypertensive diseases. From the age of 25 and over, the most common causes of men are alcoholic liver disease, transport accidents and accidental poisoning by and exposure to alcohol. These three causes alternate in the first rungs for men until the age of 54, to be succeeded later by malignant neoplasms of lip, oral cavity and pharynx, malignant neoplasms of liver and intrahepatic bile ducts, hypertensive diseases or ischaemic stroke.

Alcohol-attributable deaths according to the consumptions levels

The impact of the levels of alcohol consumption differs among causes of death. In males, alcohol-related deaths for malignant neoplasm of rectum or colon were spread evenly across the four categories of consumption (see Fig. 3).

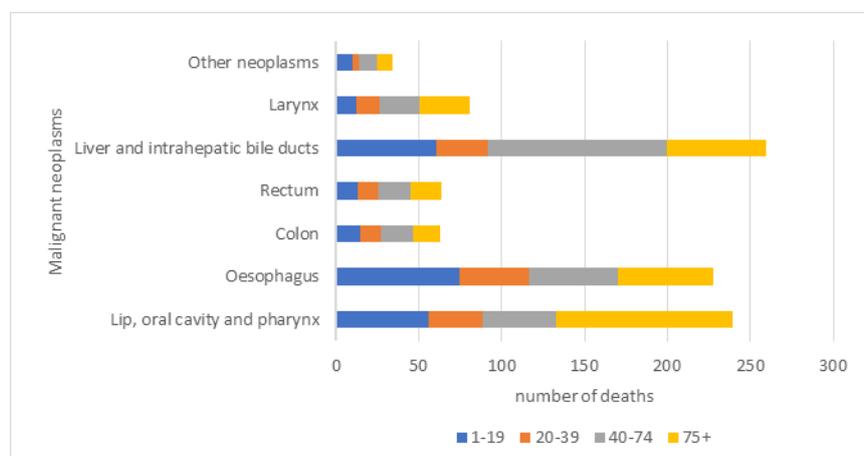


Fig. 3. Number of men's deaths from malignant neoplasm attributable to different levels of alcohol consumption in Czechia in 2017

Data source: CZSO, CHPS, Rehm et al., 2004, Jones et al., 2008, Kohoutová, 2012

For malignant neoplasm of lip, oral cavity and pharynx, and unspecified liver cirrhosis, ischaemic heart diseases, ischaemic or haemorrhagic stroke, the majority of men's deaths were attributable to alcohol consumption greater than 75 g/day (see Fig. 3 and Fig. 4). For ischaemic heart disease and ischaemic stroke, the results show that when consumed up to 74 g/day (up to 39 g/day), protective effects of alcohol are observed in men.

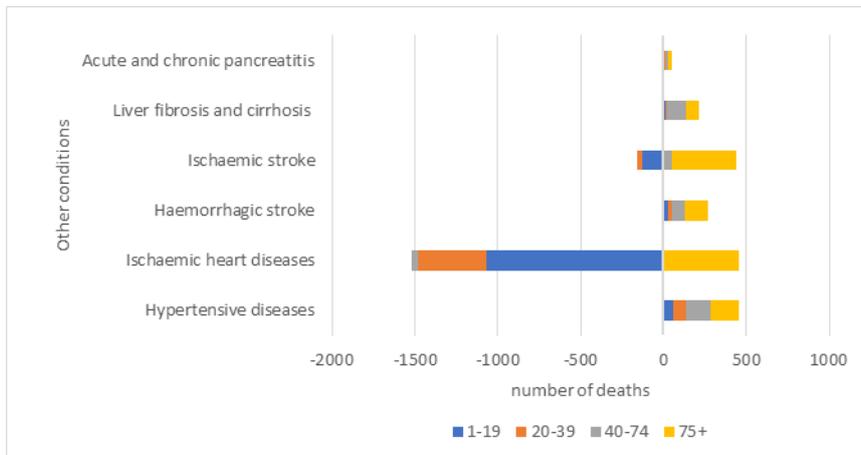


Fig. 4. Number of men's deaths from selected conditions attributable to different levels of alcohol consumption in Czechia in 2017

Data source: CZSO, CHPS, Rehm et al., 2004, Jones et al., 2008, Kohoutová, 2012

In women, a large portions of deaths are caused by consuming 1-19 g of alcohol per day (see Fig. 5 nad Fig. 6). However, this distribution may be partly due to the input values of the representation of respondents according to consumption, where in the first group of alcohol is represented by more than 50% across the ages (often more than 60%), except for the oldest age group, where abstainers were abundantly represented and women consuming 1-19 g/day make up only 30%. Most alcohol-attributed deaths in women occur at conditions liver and intrahepatic bile ducts (100 cases).

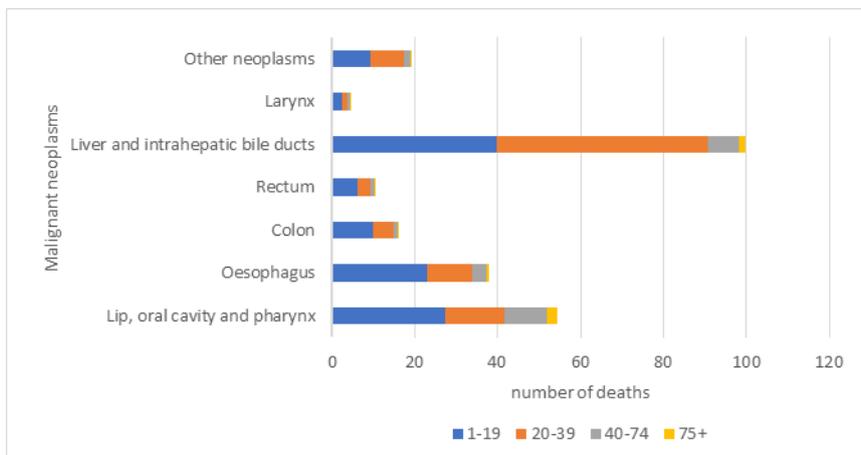


Fig. 5. Number of women's deaths from malignant neoplasm attributable to different levels of alcohol consumption in Czechia in 2017

Data source: CZSO, CHPS, Rehm et al., 2004, Jones et al., 2008, Kohoutová, 2012

For ischaemic heart diseases and ischaemic stroke, the results show that there should be some protective effects of alcohol in women too.

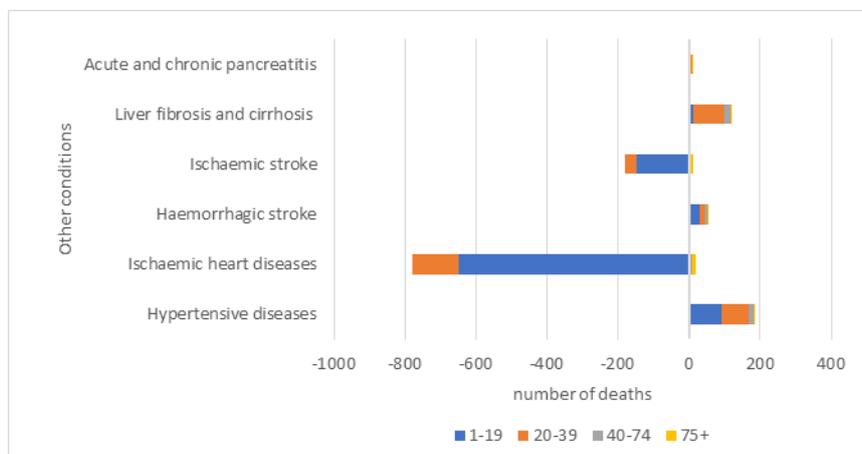


Fig. 6. Number of women's deaths from selected conditions attributable to different levels of alcohol consumption in Czechia in 2017

Data source: CZSO, CHPS, Rehm et al., 2004, Jones et al., 2008, Kohoutová, 2012

3.1 Protective effects of alcohol

Some epidemiological evidence supports the protective effect of low alcohol consumption on the risk of ischaemic heart diseases (IHD), but much is debated about the true extent of the protective effects of alcohol. According Jones et al. 2012 from recent meta-analyses, alcohol consumption was found to have protective effects on the risk of four conditions: IHD; ischaemic stroke, type II diabetes; and cholelithiasis. However, the vast majority are from the prevention of IHD deaths among individuals aged over 75 years and studies that have examined how the risks of heart disease change with increasing age have found that, at least in men, there is no evidence for a protective effect of alcohol in those aged 75 and older. However, many studies seek to address possible errors in the classification of abstainers and low-alcohol consumers. These studies suggest that many of the protective effects of alcohol are caused by unmeasured or residual confounding (Jackson et al., 2005, Fillmore et al., 2006; 2007).

Other conditions also speak of a protective effect of alcohol, such as ischaemic stroke, diabetes or cholelithiasis. However, many studies suggest that there is no clear evidence of the protective nature of moderate alcohol consumption.

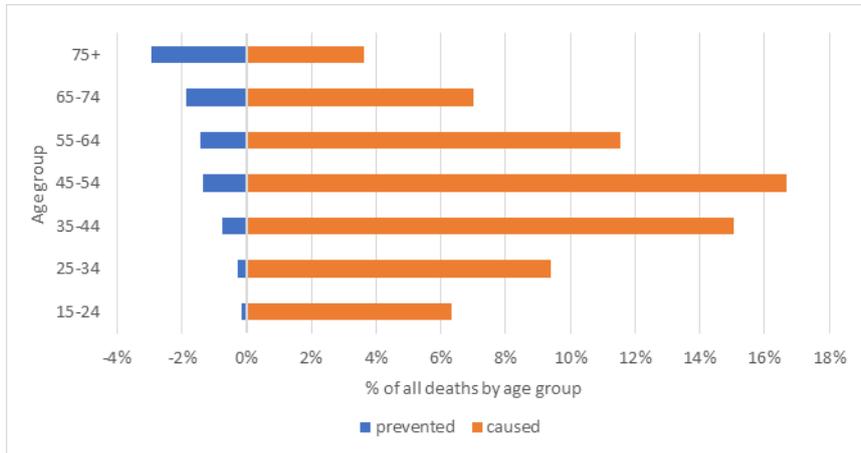


Fig. 7 Percentage of men's deaths attributable to alcohol consumption by age in Czechia in 2017

Data source: CZSO, CHPS, Rehm et al., 2004, Jones et al., 2008, Kohoutová, 2012

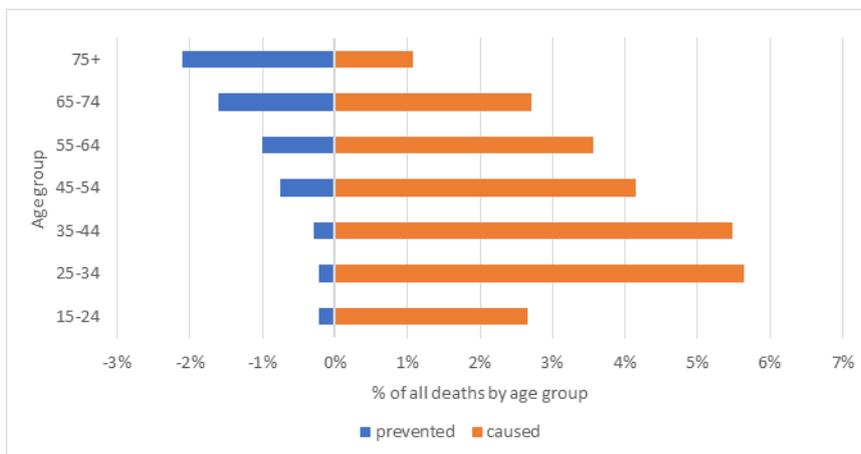


Fig. 8. Percentage of women's deaths attributable to alcohol consumption by age in Czechia in 2017

Data source: CZSO, CHPS, Rehm et al., 2004, Jones et al., 2008, Kohoutová, 2012

4 Conclusions

In this paper, we evaluated the burden of alcohol consumption on the health of the population in Czechia in 2017. Alcohol affects our health and also affects mortality. To calculate the effect of alcohol on mortality were computed the alcohol-attributable fractions (AAF) - the proportion of events that would not have occurred if the population had not been exposed to alcohol. Mortality associated with alcohol consumption, both directly and indirectly, was assessed

and compared with previous studies by Rehm et al., 2004, Jones et al., 2008, Kohoutová, 2012.

Overall, 4.3% of deaths were attributable to alcohol consumption in Czechia in 2017, with 6.8% in men and 1.7% in women. 2.7% (2,990 cases) of all deaths were caused partially by alcohol consumption, 1.5% (1,650 cases) of all deaths occurred from conditions fully caused by alcohol consumption, and 1% (110 cases) of deaths were due to acute consequences. The impact of alcohol varies by age with younger ages being more affected in both sexes by direct effects whereas among elderly, indirect effects are more pronounced.

If we compare the AAF values calculated for 2017 with data from (Kohoutová, 2012) in periods 1994-1999 and 2000-2010, we find that in most cases of these conditions there was a decrease in both men and women. This improvement is usually more pronounced in men. For example, in Malignant neoplasms of lip, oral cavity and pharynx (C00-C14), the AAF values for men were 0.59 (period 1994-1999), 0.62 (period 2000-2010) and according to our calculations there was an improvement up to 0.42 in 2017. That is, from almost 60% of deaths from this condition (C00-C14), which were partially caused by alcohol consumption at the beginning of the reference period, there was a decrease to 42% in 2017. For women, in C00-C14 condition the AAF values were 0.42 (period 1994-1999), 0.41 (period 2000-2010) and 0.33 in 2017. There was also a big improvement in the Ischaemic stroke condition, where these values were recorded in men 0.40 (period 1994-1999), 0.45 (period 2000-2010) and 0.13 by our calculations in 2017. For women, there is also a noticeable improvement, even such that results here show the protective effects of alcohol: 0.01 (period 1994-1999), 0.05 (period 2000-2010) and -0.07 in 2017.

As mentioned earlier, the protective effects of alcohol are somewhat controversial. Our results showed, for example, that in the case of Diabetes mellitus, the protective effect of alcohol for men decreased in 2017 (0.15, -0.18, -0.10). In the case of these conditions - malignant neoplasm of breast, cardiac arrhythmia and spontaneous abortion, there was an increase in AAF for women between periods. However, the changes are minimal.

The most common causes of death due to alcohol consumption in 2017 are different for men and women and change with age. Transport accidents are the most common alcohol-attributable cause of death for younger men (15-24 years), and in women aged 25-34 and 45-64 years. Malignant neoplasm of breast, together with traffic accidents, remain in top 3 alcohol-attributable causes for women till age 75. In old age, it begins to predominate in women hypertensive diseases. In men aged 25-54, the most common causes are alcoholic liver disease, traffic accidents and accidental alcohol poisoning. In older age, they are followed by malignant neoplasms of lip, oral cavity and pharynx, malignant neoplasms of liver and intrahepatic bile ducts, hypertensive diseases or ischaemic stroke.

The impact of the levels of alcohol consumption differs among causes of death. In women, a large proportion of deaths are due to the consumption of small amounts of alcohol (1-19 g of alcohol per day), however, this distribution may

be partly due to share of respondents according to consumption (more than 50% of respondents with low alcohol consumption).

For ischaemic heart diseases and ischaemic stroke, the results show that there should be some protective effects of alcohol for both sexes. However, many studies suggest that there is no clear evidence of the protective nature of moderate alcohol consumption and many of the protective effects of alcohol are caused by unmeasured or residual confounding. Some studies seek to address possible errors in the classification of abstainers and low-alcohol consumers.

Compared to previous studies, alcohol-related harm in Czechia has decreased within the past decade, but remains high compared to the Western countries.

Acknowledgment

This article was supported by the Czech Science Foundation, Grant No. GA ČR 19-23183Y, on a project titled ‘Alcohol burden in the Czech Republic: mortality, morbidity and social context‘.

References

1. Csémy, Ladislav et al. 2006. Evropská školní studie o alkoholu a jiných drogách (ESPAD): výsledky průzkumu v České republice v roce 2003. Praha: Úřad vlády České republiky, 2006. 120 s. Výzkumná zpráva; no. 4. ISBN 80-86734-94-3.
2. Csémy L, Kázmér L, Dvořáková Z. Substance use among school-aged children: results of HBSC study 1994-2014. Presentation at regular meeting of Society for Addictive Diseases of Czech Medical Association JEP in Prague. (2016 April 06)
3. ČSÚ. CZSO 2008. Spotřeba alkoholických nápojů a cigaret v letech 1920 až 2006. IN Retrospektivní údaje o spotřebě potravin v letech 1920 – 2006. [cit. 2020-04-21]. Available from WWW: <<https://www.czso.cz/csu/czso/retrospektivni-udaje-o-spotrebe-potravin-v-letech-1920-2006-n-7sg9bp0osn>>
4. ČSÚ. CZSO 2011a. Spotřeba alkoholických nápojů a cigaret (na obyvatele za rok). IN Spotřeba potravin 2010. [cit. 2020-04-21]. Available from WWW: <<https://www.czso.cz/csu/czso/graf-spotreba-alkoholickych-napojuna-1-obyvatele-v-ceske-republice>>
5. Dzurova, Dagmar & Spilková, Jana & Pikhart, Hynek. (2010). Social inequalities in alcohol consumption in the Czech Republic: A multilevel analysis. *Health & place*. 16. 590-7. 10.1016/j.healthplace.2010.01.004.
6. GBD 2016 Alcohol Collaborators. Alcohol use and burden for 195 countries and territories, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet*. 2018;392(10152):1015-35
7. Hnilicová, Helena, Siri Nome, Karolína Dobiášová, Miroslav Zvolský, Roger Henriksen, Elena Tulupova a Zuzana Kmecová. Comparison of Alcohol Consumption and Alcohol Policies in the Czech Republic and Norway. *Central European Journal of Public Health* [online]. 2017, 25(2),

- 145-151 [cit. 2020-04-25]. DOI: 10.21101/cejph.a4918. ISSN 12107778.
Available from WWW: <http://cejph.szu.cz/doi/10.21101/cejph.a4918.html>
8. Chomynová P, Csémy L, Mravčík V. European School Survey Project on Alcohol and Other Drugs (ESPAD) 2015. *Zaostřeno*. 2016;14(5):1-16.
 9. Kubička, L., Csémy, L., Duplinský, J., Kožený, J., 1998. Czech men's drinking in changing politic climate 1983–93: a three-wave longitudinal study. *Addiction* 93, 1219–1230.
 10. Policie ČR. 2000-2011. Přehledy kriminality. [cit. 2020-04-22]. Available from WWW: <https://www.policie.cz/clanek/statisticke-prehledy-kriminality-za-rok-2010.aspx>
 11. Policie ČR. 2003-2011. Dopravní nehody.
 12. World Health Organization. Global status report on alcohol and health 2018. Geneva: WHO; 2018.

Fuzzy theories and statistics

— fuzzy data analysis —

Norio Watanabe

Department of Industrial and Systems Engineering, Tokyo, Japan
(E-mail: watanabe@indsys.chuo-u.ac.jp)

Abstract. Fuzzy theories are not well accepted in the field of statistics. However, fuzzy theories are important from the statistical viewpoint. In this note we first discuss the statistical applications of the fuzzy theories briefly. The main is the fuzzy set theory and we do not refer the fuzzy measure theory. Secondly, we introduce some statistical tools for analyzing fuzzy data. The fuzzy data analysis is important in the fields related to human sensitivity. Furthermore we define the fuzzy directional data as a special case of fuzzy data. The statistical analysis of fuzzy directional data is also discussed.

Keywords: fuzzy system, fuzzy data, fuzzy random variable, directional data.

1 Introduction

Fuzzy theories are not accepted well by statisticians necessarily, though most statisticians accept the probability theory. For example, Zadeh [15] discussed the relationship between fuzzy set theory and probability theory. However, fuzzy theories and statistics can coexist and the fusion of these theories is important from the statistical viewpoint. In this note we first discuss the statistical applications of the fuzzy theories briefly. The main is the fuzzy set theory and we do not refer the fuzzy measure theory. Secondly, we introduce some statistical tools for analyzing fuzzy data.

Applications of fuzzy theories in the field of statistics can be divided into three types. The first is the case where data are not fuzzy but fuzziness appears in inference results. A typical example is the fuzzy clustering. The second is the case where fuzziness is absent in data and inference results, but a fuzzy system is used as a statistical model. Nonlinear statistical models can be obtained based on fuzzy systems. The last is the statistical analysis of fuzzy data. We call this fuzzy data analysis. The statistical treatment of fuzzy data can be studied based on the probability theory and the fuzzy set theory. The fuzzy data analysis is important in the fields related to human sensitivity.

Each fuzzy datum is represented by a fuzzy set, which is determined by a membership function. However, the usual functional data analysis cannot be applied to fuzzy data. Therefore the fuzzy set theory should be incorporated into the statistical methods. Theoretically, fuzzy data can be regarded

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



as realizations of the fuzzy random variables (see [4], for example). The fuzzy random variable is studied theoretically by many authors. However, statistical methods are not established well. We introduce some statistical tools for fuzzy data from the practical viewpoint. One essential problem in the fuzzy data analysis is how to observe membership functions. Precise observation of fuzziness is very difficult. Therefore consideration is required for overcoming this paradoxical situation.

Furthermore we define the fuzzy directional data as a special case of fuzzy data. An example of fuzzy directional data is data related to the color circle. A new definition of the average fuzzy set for fuzzy directional data is introduced, since the definition of the average fuzzy set for usual fuzzy data is not appropriate for fuzzy directional data. An application is demonstrated by using real data. Our aim is to introduce some visualization tools for extracting information on characteristics of fuzzy data.

2 Fuzzy set theory and statistics

Applications of fuzzy theories in the field of statistics can be divided into three types as was discussed in [14]. The first is the case where data are not fuzzy but fuzziness appears in inference results. A typical example is the fuzzy clustering (see [1] and [12], for example). Usually, data are crisp but results are fuzzy. Another example is a fuzzy statistical test. Watanabe and Imaizumi [9] discussed the statistical test of a fuzzy hypothesis for crisp data. Compared to fuzzy clustering, the concept like fuzzy hypothesis is not well accepted by statisticians, since the fuzzy set theory has to be used explicitly.

The second is the case where fuzziness is absent in data and inference results, but a fuzzy system is used as a statistical model. A typical fuzzy system is the Takagi-Sugeno's model ([6]). Fuzzy systems are useful in the nonlinear regression and nonlinear time series analysis. For example, the Takagi-Sugeno's model can be used as the nonlinear time series model instead of the TAR (threshold autoregressive) model. Another example is the fuzzy trend model for time series ([7]). In this situation the fuzzy theory is applied in the background and models can be used similarly to usual statistical models.

The last is the statistical analysis of fuzzy data. We call this fuzzy data analysis. The statistical treatment of fuzzy data can be studied based on the probability theory and the fuzzy set theory.

In the first and second situations data are crisp. On the other hand data are fuzzy in the third situation. This means that an essential application of the fuzzy theory is required for handling data, since data themselves are fuzzy and then classical statistical methods can not be applied directly. In the following we discuss the fuzzy data analysis from the practical viewpoint.

3 Fuzzy data analysis

3.1 Fuzzy data

Let $\{(X_n, Y_n) | n = 1, 2, \dots, N\}$ be bivariate fuzzy data. We assume that X_n and Y_n are fuzzy sets defined on intervals of real numbers respectively. Moreover

we assume that the α -level sets of X_n and Y_n are given by intervals as follows:

$$\begin{aligned} X_n(\alpha) &= [X_n^L(\alpha), X_n^U(\alpha)] \\ Y_n(\alpha) &= [Y_n^L(\alpha), Y_n^U(\alpha)] \end{aligned} \quad (1)$$

for all n and $\alpha \in (0, 1]$, where boundaries can take values $\pm\infty$.

3.2 Observation

The important issue is how to observe fuzzy sets. Precise observation of fuzziness is very difficult, since much more information is required for membership functions compared crisp data. A practical approach is to assume that the membership functions are piecewise linear.

There are two simple ways for observation of piecewise linear membership functions. The first is to observe the boundaries of the α -level set for each grade α in a predefined finite set $\{\alpha_0, \alpha_1, \dots, \alpha_K\}$, where $\alpha_0 = 0 < \alpha_1 < \dots < \alpha_K = 1$. Note that the α -level set A_0 is defined as the closure of $\{x | \mu_A(x) > 0\}$. The simplest example is the set $\{0, 1\}$. In this example the membership function is assumed to be a trapezoid. The second is to observe the grade α for each element in a pre-defined finite subset of X . In this case the number of elements should be somewhat large.

Both ways require direct evaluation on fuzziness. When a questionnaire survey is planned, careful attentions should be paid for creating questionnaires so that people who are unfamiliar with the concept of fuzziness can evaluate the degree of fuzziness. A simple example appears in Section 5.

Sometimes the empirical distribution function obtained from crisp data is regarded as the monotone membership function. However, this function is meaningless as the membership function ([8]).

3.3 Mean value

First we introduce the average set of fuzzy data X_1, \dots, X_N . The sample mean \bar{X} of X_1, \dots, X_N can be obtained by applying the extension principle in the fuzzy set theory (see [5], for example) to the N -variable function:

$$f(x_1, x_2, \dots, x_N) = (x_1 + x_2 + \dots + x_N)/N. \quad (2)$$

Under the assumption (1) we have

$$\bar{X}(\alpha) = \left[\sum_{n=1}^N X_n^L(\alpha)/N, \sum_{n=1}^N X_n^U(\alpha)/N \right], \quad (3)$$

where \bar{X}_α is the α -level set of \bar{X} .

3.4 Variance

Various types of variances for fuzzy random variables are discussed by Couso and Dubois [2]. However, statistical applications to fuzzy data are not considered well. In this note we consider statistics for data only apart from the theoretical aspect.

Fuzzy variance defined by the extension principle is not appropriate usually. Consider an example where X_n depends on a location parameter x_n and shapes of membership functions are the same. That is, $\mu_{X_n}(x) = \mu_X(x - x_n)$ for some membership function μ_X , where μ_{X_n} is the membership function of X_n . Then the information on stochastic fluctuation is included in $\{x_n\}$ only, since the fuzziness does not depend on samples. However, the fuzzy variance defined by the extension principle is usually quite fuzzy and this fuzziness is not related to stochastic fluctuation. Thus information on fluctuation cannot be derived appropriately from this fuzzy variance.

On the other hand the usual crisp variance can be calculated by introducing some defuzzification method. However, crisp statistics will lose information on fuzziness usually.

Therefore we adopt a different approach and propose two tools. The first is another fuzzy variance and the second is a simple combination of usual variances for some values characterizing fuzzy data.

Fuzzy variance

The proposed fuzzy variance consists of variances of the points in each α -level set.

Let $s^2(\alpha, \lambda) = \text{Var}(\lambda X_n^L(\alpha) + (1 - \lambda)X_n^U(\alpha)|n = 1, \dots, N)$ for all $\lambda \in [0, 1]$, where ‘Var’ means the sample variance in the usual sense.

We consider the interval of the variances:

$$I_s(\alpha) = \left[\inf_{\lambda \in [0,1]} s^2(\alpha, \lambda), \sup_{\lambda \in [0,1]} s^2(\alpha, \lambda) \right], \quad (4)$$

and then define the interval:

$$S_F^2(\alpha) = \bigcap_{\epsilon > 0} \bigcup_{1 \geq \beta \geq \alpha - \epsilon} I_s(\beta), \quad (5)$$

for $\alpha \in (0, 1]$. It can be shown that $\{S_F^2(\alpha)|\alpha \in (0, 1]\}$ determines the unique fuzzy set S_F^2 whose α -level set is $S_F^2(\alpha)$ ([10]). We can obtain the information on the fluctuation of X_n from S_F^2 . The entire set is $[0, \infty)$. In this note we call S_F^2 the fuzzy variance of $\{X_n\}$. The fuzzy standard deviation S_F is defined in the same way. The fuzzy set S_F^2 can be approximated by using finite points of α , β and λ . The operation using ϵ in (5) is not required for finite sets of α and β .

In this note we also consider the simplified fuzzy variance s_f^2 . We replace Eq. (4) by the following:

$$\left[\text{Var}(X_n^L(\alpha)|n = 1, \dots, N), \text{Var}(X_n^U(\alpha)|n = 1, \dots, N) \right]. \quad (6)$$

Then the fuzzy set is defined in the same way as S_F^2 . We denote this fuzzy set by s_f^2 and call it the simple fuzzy variance of $\{X_n\}$. This fuzzy variance is

simple but useful in practical analysis, since the boundaries are more important than other points in each α -level set.

Variance chart

Our fuzzy variances are summaries on fluctuation of data and can be regarded as visualization tools. However, it becomes difficult to interpret fuzzy variances sometimes. Therefore we consider a chart of variances on fuzzy data as another visualization tool.

We represent the fuzzy data by location parameters and the width parameters as following:

$$\begin{aligned} c_n^x &= (X_n^L(1) + X_n^U(1))/2 \\ x_n^L(\alpha) &= c_n^x - X_n^L(\alpha) \\ x_n^U(\alpha) &= X_n^U(\alpha) - c_n^x. \end{aligned} \quad (7)$$

Note that c_n^x does not depend on α and $x_n^L(\alpha)$ and $x_n^U(\alpha)$ are the left and right side width parameters respectively. Then we consider three kinds of usual variances:

$$\begin{aligned} s_c^2 &= \text{Var}(c_n^x | n = 1, \dots, N) \\ s_L^2(\alpha) &= \text{Var}(x_n^L | n = 1, \dots, N) \\ s_U^2(\alpha) &= \text{Var}(x_n^U | n = 1, \dots, N). \end{aligned} \quad (8)$$

The proposed chart consists of these variances as illustrated by Fig. 1. We call

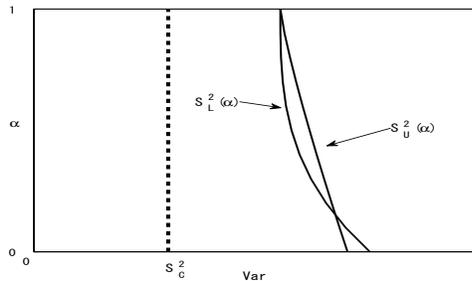


Fig. 1. F-Var chart.

this chart ‘F-Var chart.’ The ‘F-Std chart’ can be defined by using standard deviations.

3.5 Correlation

A fuzzy correlation coefficient can be defined by the extension principle ([3]). However, this measure is not appropriate for statistical analysis similarly to the fuzzy variance. We consider two correlation analyses in a similar way to the variance.

Fuzzy correlation coefficient

The fuzzy correlation coefficient between $\{X_n\}$ and $\{Y_n\}$ is determined by considering correlation between two points in α -level sets $X_n(\alpha)$ and $Y_n(\alpha)$. Put

$$r_\alpha(\lambda_1, \lambda_2) = \text{Cor} \left(\begin{array}{l} \lambda_1 X_n^L(\alpha) + (1 - \lambda_1) X_n^U(\alpha), \\ \lambda_2 Y_n^L(\alpha) + (1 - \lambda_2) Y_n^U(\alpha) \mid n = 1, \dots, N \end{array} \right) \quad (9)$$

for all $(\lambda_1, \lambda_2) \in A$, where ‘Cor’ means the sample correlation coefficient in the usual sense and A is a subset of the direct product $[0, 1] \times [0, 1]$. Typical examples of A are $[0, 1] \times [0, 1]$ and $\{(t, t) \mid t \in [0, 1]\}$. The selection of A should depend on the situation. For example, $A = [0, 1] \times [0, 1]$ is appropriate when data can include plus and minus values. When minus values are meaningless and never appear, $A = \{(t, t) \mid t \in [0, 1]\}$ is appropriate. By considering the interval of the correlation coefficients:

$$I_r(\alpha) = \left[\inf_{(\lambda_1, \lambda_2) \in A} r_\alpha(\lambda_1, \lambda_2), \sup_{(\lambda_1, \lambda_2) \in A} r_\alpha(\lambda_1, \lambda_2) \right], \quad (10)$$

we define the interval:

$$r_F(\alpha) = \bigcap_{\varepsilon > 0} \bigcup_{1 \geq \beta \geq \alpha - \varepsilon} I_r(\beta), \quad (11)$$

for $\alpha \in (0, 1]$. The family $\{r_F(\alpha) \mid \alpha \in (0, 1]\}$ determines the unique fuzzy set r_F . In this note we call r_F the fuzzy correlation coefficient between $\{X_n\}$ and $\{Y_n\}$. The entire set of r_F is the interval $[-1.0, 1.0]$. We can obtain the information on the relationship between X_n and Y_n from r_F . Similar fuzzy correlation coefficient for bivariate fuzzy random variable is discussed by [11].

In a similar way to the simple fuzzy variance we also define the simple fuzzy correlation coefficient r_f by setting $A = \{(0, 0), (1, 1)\}$.

Correlation chart

We consider a chart of correlation coefficients on fuzzy data in a similar way to ‘F-Var chart’ and call it ‘F-Cor chart.’ Let $c_n^y, y_n^L(\alpha)$ and $y_n^U(\alpha)$ denote the location and width parameters similarly to Eq. (7). The ‘F-Cor chart’ is illustrated by Fig. 2, where $r_c = \text{Cor}(c_n^x, c_n^y \mid n = 1, \dots, N)$, $r_L(\alpha) = \text{Cor}(x_n^L, y_n^L \mid n = 1, \dots, N)$ and $r_U(\alpha) = \text{Cor}(x_n^U, y_n^U \mid n = 1, \dots, N)$.

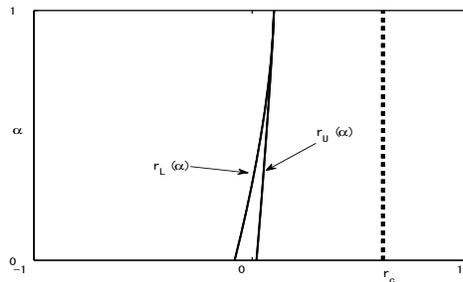


Fig. 2. F-Cor chart

4 Fuzzy directional data

In this note we consider fuzzy directional data satisfying $\mu_{X_n}(x) = \mu_{X_n}(x-2\pi)$. We assume that μ_{X_n} is unimodal on an interval whose length is 2π as follows:

$$\mu_{X_n}(x) = \begin{cases} \nearrow & \text{for } d_n \leq x \leq c_n - a_n \\ 1 & \text{for } c_n - a_n \leq x \leq c_n + a_n \\ \searrow & \text{for } c_n + a_n \leq x \leq d_n + 2\pi, \end{cases} \quad (12)$$

where

$$-2\pi < d_n < c_n - a_n \leq c_n + a_n < d_n + 2\pi \leq 2\pi \quad (13)$$

for each n . The middle point of the peak is c_n and $-\pi < c_n \leq \pi$. Fig. 3 shows an example. The fuzzy set X_n in Fig. 3 can be interpreted as “about

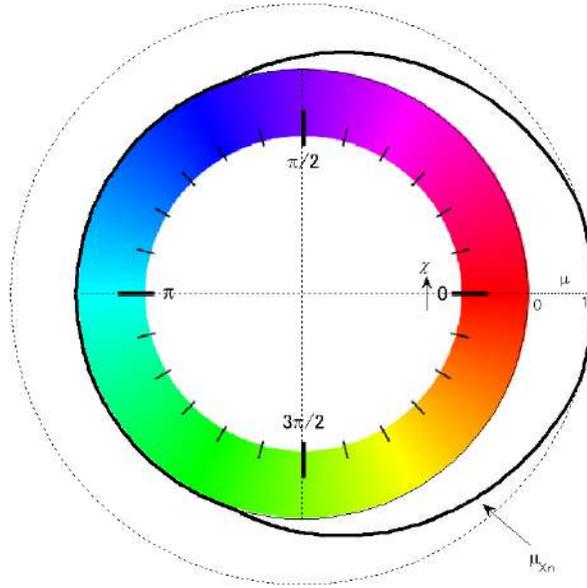


Fig. 3. Fuzzy directional data

zero degree.”

Let consider the α -level on the interval $[d_n, d_n + 2\pi]$ and put

$$(X_n)_\alpha = [c_n - x_n^L(\alpha), c_n + x_n^U(\alpha)], \quad (14)$$

where $0 \leq x_n^L(\alpha) \leq 2\pi$, $0 \leq x_n^U(\alpha) \leq 2\pi$ and $x_n^L(\alpha) + x_n^U(\alpha) \leq 2\pi$. When $\alpha_1 = \mu_{X_n}(d_n) > 0$, we set $x_n^L(\alpha) = x_n^L(\alpha_1)$ and $x_n^U(\alpha) = x_n^U(\alpha_1)$ for $\alpha \leq \alpha_1$. This means that $(X_n)_\alpha = [d_n, d_n + 2\pi]$. Hereafter we consider that the entire set is $[-2\pi, 2\pi]$ and define $\mu_{X_n}(x) = 0$ for $x < d_n$ and $x > d_n + 2\pi$. Fig. 4 shows an example.

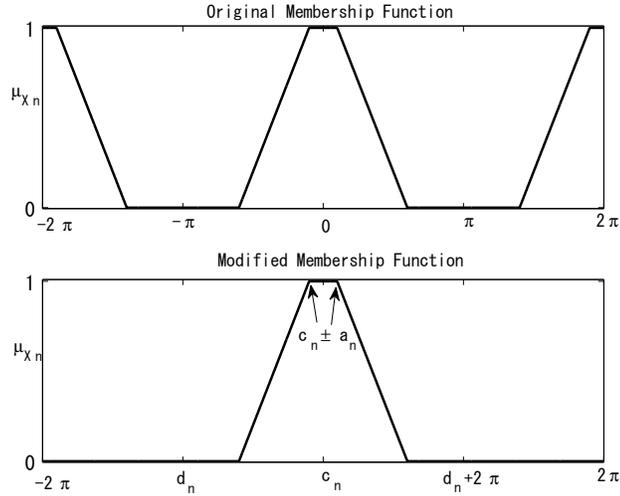


Fig. 4. Modification of fuzzy directional data

The location parameter c_n should be treated as directional data. However, it is shown that boundaries of $(X_n)_\alpha$ should not be treated as usual directional data. Therefore new approach is required for fuzzy directional data.

Let \bar{x} denote the mean value of $\{c_n\}$ obtained by the method in statistical analysis of directional data. That is,

$$R(\cos \bar{x}, \sin \bar{x}) = \left(\frac{1}{N} \sum_1^N \cos c_n, \frac{1}{N} \sum_1^N \sin c_n \right) \quad (15)$$

(the case $\sum \cos c_n = \sum \sin c_n = 0$ is omitted). Then we introduce the mean value of directional fuzzy set by using the interval:

$$\bar{X}(\alpha) = \left[\bar{x} - \frac{1}{N} \sum_1^N x_n^L(\alpha), \bar{x} + \frac{1}{N} \sum_1^N x_n^U(\alpha) \right]. \quad (16)$$

Then the mean fuzzy set \bar{X} is obtained. Note that $x_n^L(\alpha)$ and $x_n^U(\alpha)$ do not have periodicity.

For variance and correlation analysis we transform data based on the distance between angles c_n and \bar{x} and apply the methods in the previous section. Put $c_n^* = (c_n - \bar{x}) \bmod 2\pi$. If $c_n^* > \pi$, then shift as $c_n^* = c_n^* - 2\pi$. The transformed fuzzy data X_n^* is defined by the α -level set $[c_n^* - x_n^L(\alpha), c_n^* + x_n^U(\alpha)]$, where $-\pi < c_n^* \leq \pi$.

The quantity $V = 1 - R$ (see Eq. (15)) is called variance in statistical analysis of directional data. The variance V is based on the length of the chord corresponding to two angles. On the other hand usual variance is based on the length of the arc. There is no clear reason to apply V to boundaries of α -level set of fuzzy directional data. Thus we adopt the variances in the previous section to $\{X_n^*\}$.

5 Application

A questionnaire survey was conducted on the meaning of some words by using the color circle in Fig. 3. Areas of “red”, “blue” and “green” are located around the angles 0 , $2\pi/3$ and $4\pi/3$ respectively on this circle. Respondents are college students and the numbers of male and female students are 50 and 40 respectively. By assuming trapezoid membership functions they were asked boundaries from which the meaning of a word fulfills almost surely, and does not fulfill almost surely. Some impression words related the color were selected including “AKAI(red)” and “ATATAKAI(warm)” in Japanese.

The membership functions of fuzzy directional data on “red” and “warm” are plotted in Fig. 5. The bold dotted lines are the membership functions of the mean fuzzy sets obtained by the proposed method in Section 4.

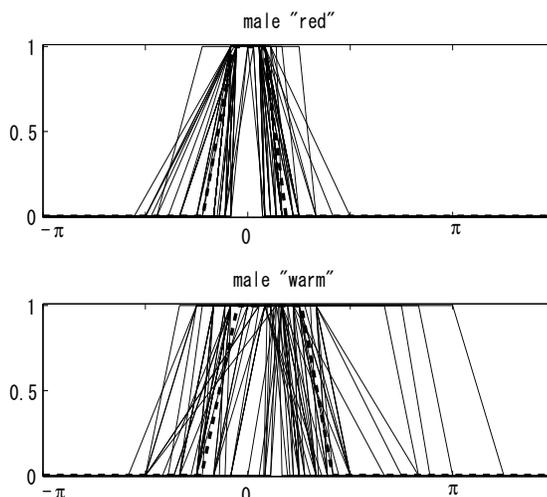


Fig. 5. Fuzzy directional data (male)

The fuzzy statistics and charts on variance and correlation proposed in Section 3 are calculated after the transformation stated in Section 4. Results are shown in Figs. 6-8.

For example, it is found that the “red” and “warm” is correlated but the correlation between location parameters is weak from the F-Cor chart. This information can not be derived from the fuzzy statistics. Note that variances of location parameters in this example are small. This is a characteristic of fuzzy directional data on the color circle.

References

1. J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum, New York, 1981.

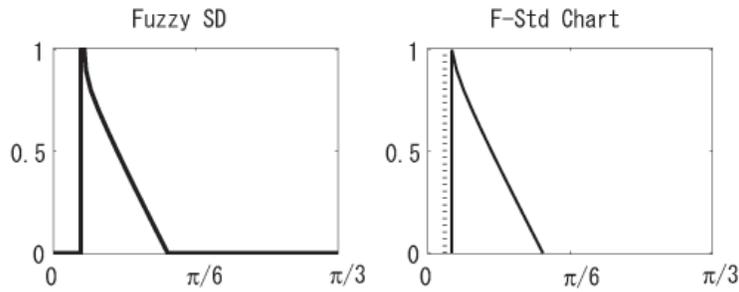


Fig. 6. Mean value and variances (male "red")

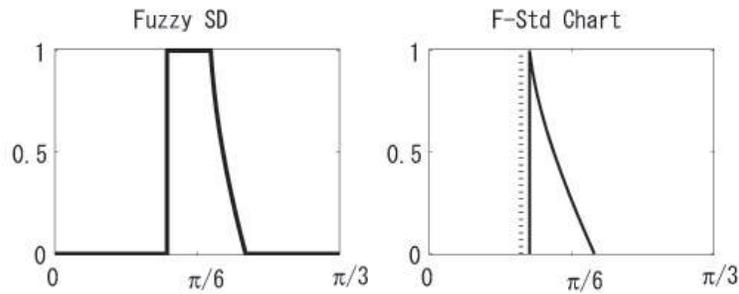


Fig. 7. Mean value and variances (male "warm")

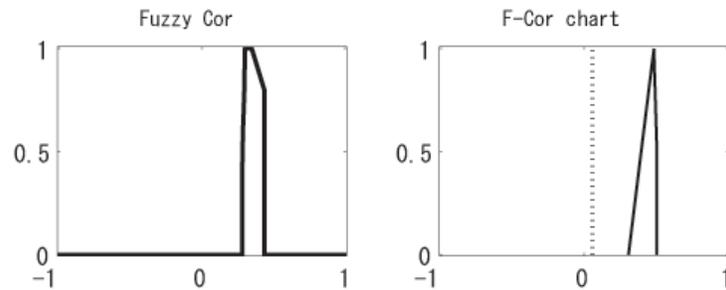


Fig. 8. Mean values and correlation (male "red"-"warm")

2. I. Couso and D. Dubois. On the variability of the concept of variance for fuzzy random variables, *IEEE Trans. of Fuzzy Systems*, 17, 5, 1070–1080.
3. S.-T. Liu and C. Kao. Fuzzy measures for correlation coefficient of fuzzy numbers, *Fuzzy set and Systems*, 128, 2, 267–275, 2002.
4. M. L. Puri and D. A. Ralescu. Fuzzy random variables, *J. Math. Anal. Appl.*, 114, 409–422, 1986.
5. M. L. Puri. Fuzzy sets: an introduction, in *International Encyclopedia of Statistical Science*, Springer, 566–571, 2011.
6. T. Takagi and M. Sugeno. (1985) Fuzzy identification of systems and its applications to modeling and control, *IEEE Trans. on Systems, Man, and Cybernetics*, 15, 1, 116–132, 1985.

7. E. Watanabe and N. Watanabe. Weighted multivariate fuzzy trend model for seasonal time series, in *Stochastic Modeling, Data Analysis and Statistical Applications* (L. Filus et al. eds.), pp. 443–450, ISAST, 2015.
8. N. Watanabe. Statistical method for estimating membership functions, *Japanese J. of Fuzzy Theory and Systems*, 5, 4, 589–601, 1993.
9. N. Watanabe and T. Imaizumi. A fuzzy statistical test of fuzzy hypotheses, *Fuzzy Sets and Systems*, 53, 2, 167–178, 1993.
10. N. Watanabe and H. Sugie. A formulation of a linear programming problem in a fuzzy environment (in Japanese), *J. of Japan Soc. for Fuzzy Theory and Systems*, 10, 375–379, 1998.
11. N. Watanabe and T. Imaizumi. A fuzzy correlation coefficient for fuzzy random variables, *The 8th IEEE International Conference on Fuzzy Systems*, Proceedings Vol. II, 1035–1038, 1999.
12. N. Watanabe, T. Imaizumi and T. Kikuchi. A hyperbolic fuzzy k -means clustering and algorithm for neural networks, in *Data Analysis, Classification, and Related Topics* (H. A. L. Kiers et al. eds.), 77–82, Springer, 2000
13. N. Watanabe and E. Watanabe. Some fuzzy correlation coefficients for bivariate fuzzy data, *SMTDA 2014*, Book of Abstracts 192, 2014.
14. N. Watanabe. Fuzzy theory and statistics, (in Japanese) *J. Japan Soc. for Fuzzy Theory and Intelligent Informatics*, 31, 3, 59–64, 2019.
15. L. A. Zadeh. Fuzzy set theory and probability theory: what is the relationship? in *International Encyclopedia of Statistical Science* (Lovric, M. ed.) 563–566, Springer, 2011.

Assessing the Performance of the European Socio-economic Classification (ESeC) in Eight European Countries for 2018

Aggeliki Yfanti¹, Anastasia Charalampi², and Catherine Michalopoulou³

¹ Ph.D., Department of Social Policy, Panteion University of Social and Political Sciences, Athens, Greece

(E-mail: aggelikiyfanti@panteion.gr)

² Postdoctoral Fellow, Department of Social Policy, Panteion University of Social and Political Sciences, Athens, Greece

(E-mail: acharalampi@panteion.gr)

³ Professor of Statistics, Department of Social Policy, Panteion University of Social and Political Sciences, Athens, Greece

(E-mail: kmichal@panteion.gr)

Abstract. In social sample survey research and the census, international classifications have been developed for the measurement of background variables such as the level of educational attainment (ISCED), economic activities (ISIC) and occupations (ISCO) to ensure the cross-national and overtime comparability of measurement. In this respect, Eurostat developed the European socio-economic classification (ESeC) as a “vehicle ... [to] monitor social structure and social change, one of the most crucial purposes of social statistics”. However, although the conceptual derivation of ESeC has been thoroughly validated, there is no evidence in the literature on its performance. In this paper, we assess the implementation of ESeC by investigating the demographic “profile” of both the employment status based on the size of the organization and the ESeC. The analysis is based on the 2018 European Social Survey datasets for eight European countries: Belgium, France, Germany, Italy, Netherlands, Poland, Switzerland and the UK.

Keywords: Employment status, social class, ESeC, European Social Survey.

1 Introduction

In social sample survey research and the census, international classifications have been developed under the auspices of international bodies to provide standardized measurement of certain background variables and thus ensure their cross-national and overtime comparability (d’Errico [5]; Kish [13]; Yfanti *et al.* [25]). The most commonly used such classifications are the International Standard Classification of Education (ISCED) developed by UNESCO based on the level of educational attainment, the International Standard Industrial Classification (ISIC) of all economic activities developed by the United Nations Statistics Division and the International Standard Classification of Occupations

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



(ISCO) developed by the International Labour Organization. These classifications have been updated over the years and their most recent revisions are ISCED 2011, ISIC Rev.4 and the ISCO-08 which replaced ISCO-88. In this context, Eurostat commissioned in 1999 an Expert Group “to make recommendations for the development of a new statistical tool for understanding differences in social structures and socio-economic inequalities across the European Union” (Rose *et al.* [20: 3]). In 2006, after a large-scale international effort involving national statistical offices, expert groups and academic researchers work was completed and the European socio-economic classification (ESeC) was proposed (Eurostat [7]; Harrison and Rose [10]) as a “vehicle ... [to] monitor social structure and social change, one of the most crucial purposes of social statistics” (Rose *et al.* [20: 6]).

The ESeC is a categorical schema classifying the adult population into a number of categories, i.e. social positions, (Harrison and Rose [10]; Rose *et al.* [20]). It “has been developed from a sociological classification that has been widely used in pure and applied research, known as the Erikson-Goldthorpe-Portocarero (EGP) Schema. The decision to adopt the EGP classification as the basis for ESeC was made because it is widely used and accepted internationally, is conceptually clear, and has been reasonably validated both in criterion terms as a measure and in construct terms as a good predictor of health and educational outcomes. ESeC improves on the EGP Schema in terms of more thorough validation and better documentation for comparative purposes” (Harrison and Rose [10: 4]; see also Rose *et al.* [20]).

The ESeC “aims to differentiate positions within labour markets and production units in terms of their typical ‘employment relations’. Therefore, ESeC recognises four basic positions: employers, the self-employed (own account workers), employees and those involuntarily excluded from the labour market” (Rose *et al.* [20: 3]; see also Harrison and Rose [10]; Lambert and Bihagen [14, 15]). The full version of ESeC defines 10 classes as follows: large employers, higher grade professionals, administrative and managerial occupations (1); lower grade professionals, administrative and managerial occupations and higher grade technician and supervisory occupations (2); intermediate occupations (3); small employers and self-employed occupations, excluding agriculture etc. (4); small employers in agriculture and self-employed occupations, agriculture etc. (5); lower supervisory and lower technician occupations (6); lower services, sales and clerical occupations (7); lower technical occupations (8); routine occupations (9) and unemployed (10).

In Figure 1, the conceptual derivation of ESeC is presented. As shown, the classification distinguishes between employers (large and small, professional and non-professional), the self-employed (professional and non-professional), employees according to their employment relations and the two categories of the unemployed and those that had never worked which are excluded.

Rose and Harrison [19], Harrison and Rose [11] and Davies and Elias [4] provided evidence for the operational validity of the ESeC because it was derived from existing variables, especially the ISCO-88. By operational validity is defined, what “in simple terms ... ‘works’ in the sense that it can be

constructed and deployed on a variety of datasets” (Rose *et al.* [20: 21]). Moreover, Rose and Harrison [19] and Rose *et al.* [20]) discussed the ESeC criterion and construct validity.

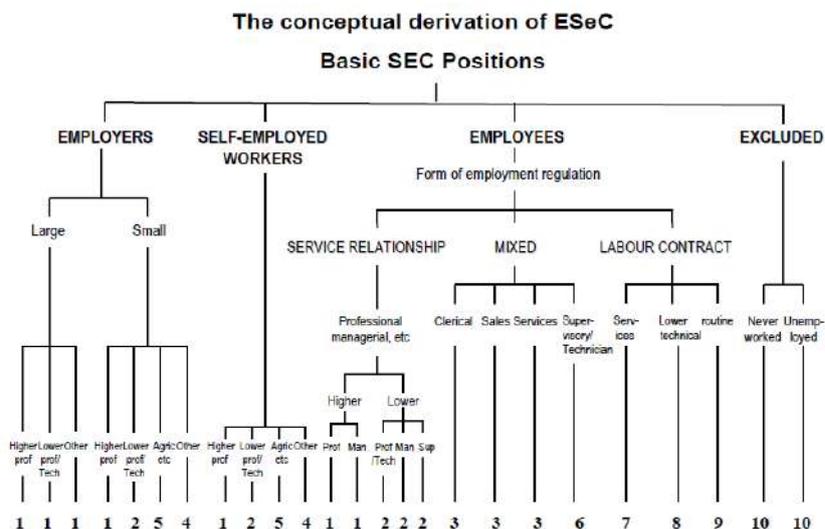


Figure 1. The conceptual derivation of ESeC. Reproduced from “The European socio-economic classification (ESeC) user guide,” by Harrison and Rose, Institute for Economic and Social Research, University of Essex, Colchester, UK, 2006, p. 22.

In the context of attitude scaling, a measurement instrument is considered to be valid if it measures “what it sets out to measure, so that differences between individuals’ scores can be taken as representing true differences in the characteristic under study” (Moser and Kalton [18: 355]). Construct validity is assessed on “the basis of theoretical considerations, [when] the researcher postulates the types and degrees of association between the scale and other variables and he then examines these associations to see whether they confirm his expectations” (Moser and Kalton [18: 356]). “Criterion validity is usually assessed by correlating the subscales (or overall scale) with another scale or index measuring the relevant underlying construct. High values of the correlation coefficients indicate adequate evidence of criterion validity” (Michalopoulou [17: 11]). Bihagen *et al.* [1] and Wirth *et al.* [24] investigated the measurement of social class thus bearing on the criterion validity of ESeC. However, although the conceptual derivation of ESeC has been thoroughly validated and discussed (Connelly *et al.* [3]; Filhon *et al.* [8]; Goedemé [9]; Maloutas [16]; Rose and Harrison [19]; Watson *et al.* [23]), there is no evidence in the literature on its performance. In this paper, we assess the implementation

of ESeC by investigating the demographic “profile” of both the employment status based on the organisation size and the ESeC. The analysis is based on the 2018 European Social Survey datasets for eight European countries: Belgium, France, Germany, Italy, Netherlands, Poland, Switzerland and the UK.

2 Method

2.1 Participants

The analysis was based on the European Social Survey Round 9 Data [6] for the following eight countries: Belgium, France, Germany, Italy, Netherlands, Poland, Switzerland and the UK. These countries were selected from the 19 participants in Round 9 because they had statistical significant results for both demographic characteristics (gender and age) under consideration (see section 2.3). The ESS implements all the strict methodological prerequisites for comparability over time and cross-nationally (Kish [13]) by applying probability sampling, minimum effective achieved sample sizes in all participating countries and a maximum target response rate of 70% (The ESS Sampling Expert Panel [21]). Face-to-face interviewing is used for data collection. The survey population is defined as all persons aged 15 and over residing within private households in each country, regardless of their nationality, citizenship or language. In Table 1, the demographic characteristics of participants are presented.

Table 1. The demographic characteristics of participants: European Social Survey, 2018

Country	<i>N</i>	Men (%)	Women (%)	Mean age (SD)
Belgium	946	49.1	50.9	47.90 (19.190)
France	5481	47.0	53.0	50.10 (18.967)
Germany	7162	51.4	48.6	49.63 (19.062)
Italy	5241	47.3	52.7	51.32 (19.410)
Netherlands	1442	49.7	50.3	48.45 (18.768)
Poland	3219	47.4	52.6	47.62 (18.880)
Switzerland	722	50.3	49.7	47.50 (18.859)
UK	5440	47.1	52.9	50.15 (18.141)

Data weighted by the design and population size weights.

As shown, there are more women than men in the datasets of Belgium, France, Italy, the Netherlands, Poland and the UK, whereas in those of Germany and Switzerland there are slightly more men than women. The mean age ranges from 47.50 (Switzerland) to 51.32 (Italy) years of age.

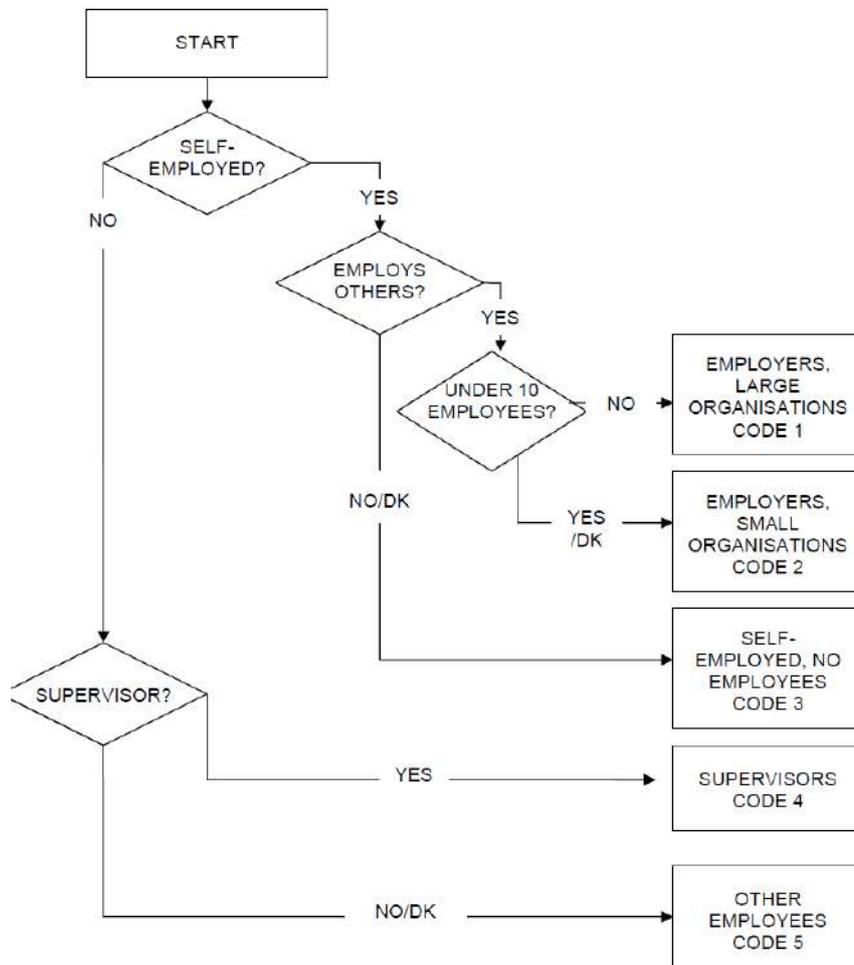


Figure 2. The derivation of the employment status variable based on the size of organization. Reproduced from “The European socio-economic classification (ESeC) user guide,” by Harrison and Rose, Institute for Economic and Social Research, University of Essex, Colchester, UK, 2006, p. 16.

2.2 Measures

Harrison and Rose [10:3] pointed out that the ESeC is “an occupationally based classification but has rules to provide coverage of the whole adult population. The information required to create ESeC is:

- occupation coded to the minor groups (i.e. 3-digit groups) of EU variant of the International Standard Classification of Occupations 1988 (ISCO88 (COM));

- details of employment status, i.e. whether an employer, self-employed or employee;
- number of employees at the workplace;
- whether a worker is a supervisor.

As mentioned before, ESeC distinguishes between large and small employers and therefore the size of the organization is required — any establishment with 10 or more employees is considered as large. In Figure 2, the derivation of the employment status variable based on the size of the organization is presented (see also Bohr [2]; Tijdens [22]). ESeC was computed using the SPSS script provided by the Institute of Social and Economic Research at the University of Essex.

2.3 Statistical analysis

The analysis pertained to obtain the demographic “profile” of ESeC and compare it cross-nationally. The demographic profiling was based on gender (men and women) and age (15-24, 25-34, 35-44, 45-54, 55-64 and 65+). Only statistically significant results ($p < .001$) were to be included in the analysis. In this respect, eight of the 19 countries participating in Round 9 [6] of the ESS had significant results for both gender and age. In four countries (Austria, Czechia, Finland and Serbia), although the results for gender were significant those of age were not. In the remaining seven countries (Bulgaria, Cyprus, Estonia, Hungary, Ireland, Norway and Slovenia) none of the results were significant. Therefore, the following eight datasets were retained in the analyses: Belgium, France, Germany, Italy, Netherlands, Poland, Switzerland and the UK.

3 Results

3.1 The employment status based on the size of the organization

In Table 2, the frequency distribution of the employment status based on the size of the organization is presented. As shown, the percentages of respondents classified as self-employed with more than 10 employees ranged from 0.7% (France) to 1.5% (Netherlands). The percentages of those classified as self-employed with less than 10 employees ranged from 3.3% (Poland) to 8.3% (Italy). The percentages of those classified as self-employed with no employees ranged from 5.1% (Germany) to 12.6% (Poland). More than 46.4% of respondents were classified as employees overall the eight countries. Italy and Poland had the largest percentages of respondents classified as employees. The Netherlands had the smallest percentage of respondents classified as employees but the largest percent of those classified as supervisors.

In Table 3, the demographic “profile” of the employment status based on the size of the organization is presented. As shown, in Belgium, France, Germany,

Table 2. The frequency distribution of the employment status based on the size of organization: European Social Survey, 2018

Country	Employment status/size of organization frequency (valid %)				
	SE10+E	SE<10E	SEnoE	Supervisors	Employees
Belgium	0.9	4.1	7.4	29.3	58.3
France	0.7	3.5	8.0	31.7	56.0
Germany	1.1	4.1	5.1	34.5	55.2
Italy	0.5	8.3	11.6	12.2	67.3
Netherlands	1.5	4.0	7.5	40.6	46.4
Poland	0.9	3.3	12.6	12.9	70.3
Switzerland	1.3	4.0	5.8	31.5	57.5
UK	-	4.0	12.5	33.8	49.8

SE10+E = self-employed with more than 10 employees; SE<10E = self-employed with less than 10 employees; SEnoE = self-employed with no employees. Missing values were as follows: 10.4% (Belgium); 8.3% (France); 5.8% (Germany); 24.9% (Italy); 3.7% (Netherlands); 12.6% (Poland); 5.0% (Switzerland) and 4.1% (UK). Data weighted by the design and population size weights.

Table 3. The demographic “profile” of the employment status based on the size of organization: European Social Survey, 2018

Country/ESeC employment status	Gender (%)		Age (%)					
	Men	Women	15-24	25-34	35-44	45-54	55-64	65+
Belgium (N=849)								
SE 10+ employees	1.4	0.2	1.4	0.0	0.7	1.3	1.3	1.5
SE <10 employees	5.4	2.9	1.4	3.6	2.2	6.9	4.7	4.6
SE no employees	9.6	5.3	4.2	6.5	9.6	6.9	8.0	7.7
Supervisors	36.6	21.7	12.5	26.8	30.1	30.2	33.3	32.5
Employees	47.1	69.9	80.6	63.0	57.4	54.7	52.7	53.6
France (N=5025)								
SE 10+ employees	1.3	0.2	0.0	0.0	0.4	0.3	1.5	1.3
SE <10 employees	5.0	2.1	0.0	0.0	3.4	2.9	2.5	7.2
SE no employees	8.9	7.2	2.5	10.7	7.3	8.7	8.4	8.0
Supervisors	37.1	26.8	12.7	31.4	35.0	32.6	30.9	34.9
Employees	47.6	63.7	84.8	57.9	53.9	55.5	56.7	48.6

Table 3. (continued)

Country/ESeC employment status	Gender (%)		Age (%)					
	Men	Women	15-24	25-34	35-44	45-54	55-64	65+
Germany (<i>N</i> =6733)								
SE 10+ employees	1.4	0.7	0.0	0.3	1.7	1.2	0.4	2.0
SE <10 employees	5.6	2.5	0.5	1.3	3.3	4.8	5.0	6.3
SE no employees	5.8	4.4	2.7	2.3	3.7	5.8	7.8	5.7
Supervisors	41.2	27.3	13.8	36.1	39.8	37.4	32.6	38.7
Employees	46.1	65.0	83.1	60.0	51.6	50.7	54.2	47.3
Italy (<i>N</i> =3914)								
SE 10+ employees	0.9	0.1	0.0	0.9	0.9	0.5	0.0	0.7
SE <10 employees	10.6	5.8	2.6	5.2	7.9	10.7	10.7	7.7
SE no employees	13.7	9.3	5.2	8.3	10.7	12.1	13.3	12.6
Supervisors	14.7	9.5	2.6	7.8	13.5	11.9	14.0	13.4
Employees	60.1	75.3	89.6	77.8	66.9	64.8	62.0	65.6
Netherlands (<i>N</i> =1385)								
SE 10+ employees	2.6	0.4	0.0	0.0	1.5	2.6	1.6	2.2
SE <10 employees	5.4	2.6	1.7	1.7	4.5	6.0	3.6	5.1
SE no employees	8.9	6.1	4.0	7.2	7.0	8.2	10.0	7.4
Supervisors	48.4	32.6	24.1	44.2	42.5	43.8	40.2	44.2
Employees	34.7	58.3	70.1	47.0	44.5	39.3	44.6	41.0
Poland (<i>N</i> =2812)								
SE 10+ employees	1.5	0.3	0.0	0.4	1.1	0.0	1.5	1.4
SE <10 employees	5.1	1.6	1.8	1.7	5.1	2.5	4.6	3.0
SE no employees	13.4	11.8	7.8	7.2	7.5	14.6	16.9	17.6
Supervisors	16.0	10.0	1.8	13.1	14.8	11.4	10.8	17.6
Employees	64.0	76.3	88.5	77.6	71.4	71.5	66.2	60.4
Switzerland (<i>N</i> =679)								
SE 10+ employees	2.3	0.3	0.0	1.0	0.0	2.3	0.9	2.1
SE <10 employees	5.2	2.7	0.0	1.9	2.6	4.7	6.4	6.2
SE no employees	5.4	6.2	1.3	2.9	4.4	7.8	10.0	6.9
Supervisors	39.0	23.7	11.8	29.5	35.1	38.0	28.2	37.9
Employees	48.1	67.1	86.8	64.8	57.9	47.3	54.5	46.9
UK (<i>N</i> =5186)								
SE <10 employees	6.4	1.8	1.3	3.2	2.9	2.6	3.6	6.9
SE no employees	15.8	9.5	4.1	11.0	12.4	14.9	16.1	11.3
Supervisors	32.9	34.6	9.3	35.2	36.7	34.9	35.2	36.6
Employees	45.0	54.2	85.3	50.6	47.9	47.6	45.1	45.2

SE = self-employed. Data was weighted by the design and population size weights. All the results were significant at $p < .001$.

the Netherlands, Switzerland and the UK, women were mainly classified as employees with percentages ranging from 54.2 % (UK) to 69.9% (Belgium).

Men were classified as employees to lesser extent than women — with percentages ranging from 34.7% (Netherlands) to 48.1% (Switzerland) — having also more positions as supervisors — with percentages ranging from 32.9% (UK) to 48.4% (Netherlands). This trend was less pronounced in the cases of Italy and Poland where over 75.3% of women were classified as employees. Also, more than 60.1% of men were classified as employees and more than 14.7% as supervisors. The great majority of younger respondents (more than 70.1%) were classified as employees in all countries. Respondents were classified less as employees and more as supervisors as they were getting older. Again, this trend was less pronounced in the cases of Italy and Poland.

3.2 The ESeC

In Table 4, the frequency distribution of ESeC is presented. As shown, the percentages of the respondents classified as large employers, higher grade professionals, administrative and managerial occupations (ESeC1) ranged from 9.2% (Italy) to 16.0% (UK). The percentages of the respondents classified as

Table 4. The frequency distribution of the European socio-economic classification (ESeC): European Social Survey, 2018

Country	ESeC frequency (valid %)								
	1	2	3	4	5	6	7	8	9
Belgium	15.8	19.0	3.5	13.2	0.9	22.8	12.2	5.5	7.0
France	10.6	17.5	7.2	10.8	3.0	26.7	9.5	6.0	8.6
Germany	11.4	19.9	4.9	10.5	1.1	32.3	5.3	7.3	7.4
Italy	9.2	9.1	5.8	24.4	1.1	13.4	11.9	10.8	14.4
Netherlands	12.0	14.7	5.1	12.9	1.1	38.8	7.4	2.5	5.5
Poland	10.7	13.1	2.2	10.5	13.7	15.2	8.2	11.1	15.3
Switzerland	12.9	18.0	7.4	10.2	1.3	27.7	6.7	8.6	7.1
UK	16.0	12.4	2.8	18.1	1.1	29.9	8.1	2.5	9.0

The ESeC nine classes are defined as follows: 1 = large employers, higher grade professionals, administrative and managerial occupations; 2 = lower grade professionals, administrative and managerial occupations and higher grade technician and supervisory occupations; 3 = intermediate occupations; 4 = small employers and self-employed occupations (excluding agriculture etc); 5 = small employers in agriculture and self-employed occupations (agriculture etc); 6 = lower supervisory and lower technician occupations; 7 = lower services, sales and clerical occupations; 8 = lower technical occupations; 9 = routine occupations. Missing values were as follows: 34.1% (Belgium); 31.3% (France); 31.5% (Germany); 47.4% (Italy); 26.9% (Netherlands); 42.6% (Poland); 29.4% (Switzerland) and 30.6% (UK). Data weighted by the design and population size weights.

lower grade professionals, administrative and managerial occupations and higher grade technician and supervisory occupations (ESeC2) ranged from 9.1% (Italy) to 19.9% (Germany). The percentages of the respondents classified as having intermediate occupations (ESeC3) ranged from 2.2% (Poland) to 7.4%

(Switzerland). The percentages of the respondents classified as small employers and self-employed occupations excluding agriculture (ESeC4) ranged from 10.2% (Switzerland) to 24.4% (Italy). The percentages of the respondents classified as small employers in agriculture and self-employed occupations in agriculture (ESeC5) ranged from 0.9% (Belgium) to 13.7% (Poland). The percentages of the respondents classified as having lower supervisory and lower technician occupations (ESeC6) ranged from 13.4% (Italy) to 38.8% (Netherlands). The percentages of the respondents classified as having lower services, sales and clerical occupations (ESeC7) ranged from 5.3% (Germany) to 12.2% (Belgium). The percentages of the respondents classified as having lower technical occupations (ESeC8) ranged from 2.5% (Netherlands and the UK) to 11.1% (Poland). The percentages of the respondents classified as having routine occupations (ESeC9) ranged from 7.0% (Belgium) to 15.3% (Poland).

In Table 5, the demographic “profile” of the ESeC is presented. As shown, in most cases men were classified in higher socio-economic positions than women except for the Polish dataset where this trend is reversed. The same trend was detected for age as the older respondents were classified in higher socio-economic positions than the younger respondents.

In the dataset of Belgium, men and women were classified as having mainly lower supervisory and lower technician occupations (ESeC6) and lower services, sales and clerical occupations (ESeC7), respectively. In the cases of France, Germany, the Netherlands, Switzerland and the UK, although men and women were in the main classified as having mainly lower supervisory and lower technician occupations (ESeC6), women were positioned systematically lower than men. This trend is most marked in the case of the Italian dataset where men and women were classified mainly as small employers and self-employed occupations excluding agriculture (ESeC4) and lower services, sales and clerical occupations (ESeC7), respectively. In the Polish dataset, men and women were classified as having routine occupations (ESeC9) and small employers in agriculture and self-employed occupations in agriculture (ESeC5), respectively.

In the cases of Germany and the Netherlands, younger respondents were classified as having mainly lower supervisory and lower technician occupations (ESeC6) whereas the older respondents were classified as lower grade professionals, administrative and managerial occupations and higher grade technician and supervisory occupations (ESeC2). In the cases of Belgium, Italy and the UK, younger respondents were classified as having mainly lower services, sales and clerical occupations whereas the older respondents were classified as having mainly lower supervisory and lower technician occupations (ESeC6). In the cases of France and Switzerland, younger respondents were classified as having mainly lower technical occupations (ESeC8) — and in the case of France the same percentage was classified also as having routine occupations (ESeC9) — whereas the older respondents were classified as having mainly lower supervisory and lower technician occupations (ESeC6). In the case of Poland, younger respondents were classified as having mainly lower technical occupations (ESeC8) as was the case in France and Switzerland.

However, although they were changing into a higher socio-economic position later in life, after they were 55+ years old they returned to their initial socio-economic position.

Table 5. The demographic “profile” of the European socio-economic classification (ESeC) classification: European Social Survey, 2018

Country/ESeC classes	Gender (%)		Age (%)					
	Men	Women	15-24	25-34	35-44	45-54	55-64	65+
Belgium (<i>N</i> =623)								
ESeC1	15.9	15.6	4.7	20.7	21.6	16.0	14.3	13.4
ESeC2	20.0	17.7	16.3	25.0	17.6	14.4	19.3	20.4
ESeC3	1.2	6.4	4.7	3.3	2.9	2.4	5.0	3.5
ESeC4	15.7	10.3	9.3	10.9	15.7	14.4	12.6	13.4
ESeC5	1.4	0.4	0.0	2.2	0.0	1.1	0.8	1.4
ESeC6	25.5	19.5	18.6	19.6	20.6	23.2	24.4	25.4
ESeC7	3.2	23.0	30.2	9.8	11.8	15.2	10.1	7.7
ESeC8	8.7	1.8	9.3	4.3	4.9	6.4	4.2	6.3
ESeC9	8.4	5.3	7.0	4.3	4.9	6.4	9.2	8.5
France (<i>N</i> =3766)								
ESeC1	11.7	9.4	3.5	13.3	11.1	11.3	11.7	9.4
ESeC2	18.8	16.1	11.4	17.3	19.3	19.6	16.8	16.8
ESeC3	3.5	11.1	9.0	5.5	9.0	7.0	6.6	7.0
ESeC4	11.9	9.6	1.5	11.8	10.1	10.7	10.6	12.7
ESeC5	3.9	2.1	3.0	1.8	0.9	2.5	4.1	4.5
ESeC6	28.2	25.1	14.4	32.4	30.3	25.5	23.9	27.0
ESeC7	2.5	17.1	14.4	4.7	9.2	10.1	13.6	7.9
ESeC8	9.2	2.6	21.4	5.5	3.8	5.8	5.2	5.4
ESeC9	10.2	6.9	21.4	7.8	6.3	7.6	7.5	9.3

Table 5. (continued)

Country/ESeC classes	Gender (%)		Age (%)					
	Men	Women	15-24	25-34	35-44	45-54	55-64	65+
Germany (<i>N</i> =4902)								
ESeC1	12.2	10.5	9.7	18.5	14.4	11.0	9.3	8.8
ESeC2	17.7	22.8	13.2	23.1	16.7	22.0	20.4	19.8
ESeC3	1.2	9.9	6.5	5.8	4.3	3.6	6.4	3.9
ESeC4	11.5	9.0	4.0	3.1	10.6	11.3	12.5	13.9
ESeC5	1.8	0.1	0.0	0.9	0.9	0.7	1.1	1.9
ESeC6	33.9	30.0	20.5	28.5	37.9	33.7	29.1	36.9
ESeC7	2.0	9.7	13.2	3.6	4.3	5.4	5.8	3.9
ESeC8	11.3	1.9	17.3	7.3	3.8	6.7	8.1	5.7
ESeC9	8.4	6.1	15.6	9.1	7.3	5.7	7.3	5.3
Italy (<i>N</i> =2747)								
ESeC1	7.5	11.5	3.5	12.8	2.6	10.0	7.8	7.6
ESeC2	8.9	9.3	7.0	4.5	10.1	10.9	7.6	9.9
ESeC3	3.7	8.7	8.8	6.6	8.9	7.1	5.3	3.4
ESeC4	26.8	21.3	8.8	14.9	23.2	27.6	29.9	24.7
ESeC5	1.4	0.6	0.0	1.4	0.5	0.7	1.0	1.8
ESeC6	15.4	10.6	3.5	10.1	14.5	13.1	15.6	14.0
ESeC7	5.3	20.6	36.0	21.9	10.4	12.0	8.0	8.3
ESeC8	15.6	4.3	12.3	12.8	9.7	9.4	9.5	11.9
ESeC9	15.4	13.1	20.2	14.9	10.1	9.3	15.2	18.4
Netherlands (<i>N</i> =1051)								
ESeC1	12.5	11.5	6.0	12.7	11.9	16.3	11.1	11.5
ESeC2	14.4	15.0	11.0	19.7	13.2	10.2	16.8	16.2
ESeC3	1.6	9.6	6.0	5.6	2.6	5.1	4.7	6.3
ESeC4	14.7	10.6	7.0	8.5	14.6	16.3	15.8	11.5
ESeC5	1.6	0.4	1.0	0.0	0.0	1.4	1.1	2.4
ESeC6	42.1	34.9	36.0	39.4	44.4	39.1	37.4	37.2
ESeC7	2.6	13.2	20.0	7.0	7.3	4.2	5.3	6.7
ESeC8	3.6	1.3	3.0	4.2	2.0	1.9	3.7	1.6
ESeC9	6.9	3.5	10.0	2.8	4.0	5.6	4.2	6.7
Poland (<i>N</i> =1845)								
ESeC1	6.5	16.5	1.7	16.9	18.0	9.7	9.0	5.6
ESeC2	15.5	9.8	11.3	16.9	12.1	14.3	12.4	11.7
ESeC3	1.0	3.8	0.0	3.8	2.5	1.5	1.2	2.7
ESeC4	13.3	6.8	9.6	10.0	11.5	8.9	16.2	7.3
ESeC5	11.1	17.2	7.8	4.5	7.9	16.6	15.9	21.9
ESeC6	14.8	15.8	6.1	12.4	18.3	14.7	12.1	19.4
ESeC7	1.8	16.9	20.9	9.0	5.3	6.6	6.6	8.8
ESeC8	17.2	2.7	26.1	11.7	9.6	12.8	17.7	21.2
ESeC9	18.8	10.5	16.5	14.8	14.9	17.8	16.2	13.6

Table 5. (continued)

Country/ESeC classes	Gender (%)		Age (%)					
	Men	Women	15-24	25-34	35-44	45-54	55-64	65+
Switzerland (N=507)								
ESeC1	14.0	11.5	4.8	19.5	15.7	11.8	17.4	8.5
ESeC2	16.4	20.3	14.3	18.2	22.9	19.6	11.6	19.7
ESeC3	3.8	12.4	9.5	5.2	4.8	5.9	8.1	11.1
ESeC4	9.9	10.6	0.0	6.5	8.4	9.8	16.3	13.7
ESeC5	2.0	0.5	0.0	0.0	1.2	2.0	1.2	1.7
ESeC6	30.0	24.4	16.7	24.7	28.9	33.3	23.3	30.8
ESeC7	3.1	11.5	19.0	9.1	4.8	6.9	4.7	2.6
ESeC8	3.0	2.8	23.8	11.7	6.0	3.9	8.1	6.8
ESeC9	7.8	6.0	11.9	5.2	7.2	6.9	9.3	5.1
UK (N=3755)								
ESeC1	13.8	18.5	6.5	16.6	16.1	17.2	20.1	14.0
ESeC2	13.0	11.8	10.7	13.6	11.1	12.5	11.7	13.6
ESeC3	2.0	3.8	2.3	2.2	3.3	2.9	2.4	3.3
ESeC4	21.8	13.9	6.5	14.4	19.2	19.9	20.3	19.2
ESeC5	1.6	0.5	0.0	1.6	0.0	1.3	1.3	1.0
ESeC6	25.6	34.9	12.1	33.8	36.5	31.1	26.6	29.1
ESeC7	4.7	12.1	37.2	6.7	5.0	5.9	6.9	7.0
ESeC8	4.3	0.5	5.1	1.6	3.2	2.7	1.2	3.1
ESeC9	13.3	4.1	19.5	9.5	5.6	6.5	9.5	9.6

ESeC1 = large employers, higher grade professionals, administrative and managerial occupations; ESeC2 = lower grade professionals, administrative and managerial occupations and higher grade technician and supervisory occupations; ESeC3 = intermediate occupations; ESeC4 = small employers and self-employed occupations (excluding agriculture etc); ESeC5 = small employers in agriculture and self-employed occupations (agriculture etc); ESeC6 = lower supervisory and lower technician occupations; ESeC7 = lower services, sales and clerical occupations; ESeC8 = lower technical occupations; ESeC9 = routine occupations. Data weighted by the design and population size weights. All the results were significant at $p < .001$.

4 Conclusions

In this paper, we assessed the implementation of ESeC by investigating the demographic “profile” of both the employment status based on the size of the organisation and the ESeC. The analysis was based on the 2018 European Social Survey datasets for eight European countries: Belgium, France, Germany, Italy, Netherlands, Poland, Switzerland and the UK.

The findings showed that of the five categories of the employment status based on the size of the organization, i.e. self-employed with more than 10 employees, self-employed with less than 10 employees, self-employed with no employees, supervisors and employees, more than 79.5% of the respondents were classified in the two categories of supervisors and employees. In most cases, women were mainly classified as employees. Men were classified as

employees to lesser extent than women having also more positions than the women as supervisors. More than 70.1% of younger respondents were classified as employees in all countries. Respondents were classified less as employees and more as supervisors as they were getting older. Both these trends were less pronounced in the cases of Italy and Poland.

In most countries, the largest category of ESeC was that of the lower supervisory and lower technician occupations (ESeC6) except for Italy where the largest category was that of the small employers and self-employed occupations excluding agriculture (ESeC4). In the case of Poland, the largest categories with the same percentages were those of the lower supervisory and lower technician occupations (ESeC6) and routine occupations (ESeC9). In all cases, men were classified in higher socio-economic positions than women. Although, the younger respondents were mostly classified in lower socio-economic positions than the older respondents, the findings did not result in any clear overall pattern.

These results suggested that the ESeC full version of the ten categories could be problematic for samples of these sizes. Therefore, researchers should have better collapse the full version to the lesser categories of seven, six, five or three class models so as to obtain more meaningful results (Connelly *et al.* [3]; Lambert and Bihagen [15]; Rose *et al.* [20]). As the cross-national comparisons did not result in any clear geographical pattern and because the ESeC classification relies on the occupational variables, further research is necessary especially of the labour market in each country, so as to establish the necessary framework for such comparisons (Goedemé [9]; Maloutas [16]).

References

1. E. Bihagen, M. Neramo and R. Erikson. Social class and employment relations: comparisons between the ESeC and EGP class schema using European data. In D. Rose and E. Harrison (Eds.), *Social class in Europe: An introduction to the European socio-economic classification* (pp. 89-113), Routledge, New York, 2010.
2. J. Bohr. EU-AES tools: implementation of the European socioeconomic groups classification (ESeG) using Adult Education Survey microdata (GESIS Papers, 2018/14). GESIS - Leibniz-Institut für Sozialwissenschaften, Köln, 2018. <https://doi.org/10.21241/ssar.57622>
3. R. Connelly, V. Gayle, & P.S. Lambert. A Review of occupation-based social classifications for social survey research. *Methodological Innovations*, 9, 1–14, 2016.
4. R. Davies and P. Elias. The application of ESeC to three sources of comparative European data. In D. Rose and E. Harrison (Eds.), *Social class in Europe: An introduction to the European socio-economic classification* (pp. 61-86), Routledge, New York, 2010.

5. A. d'Errico, F. Ricceri, S. Stringhini, C. Carmeli, M. Kivimaki, M. Bartley, et al. Socioeconomic indicators in epidemiologic research: a practical example from the LIFEPAATH study. *PLoS ONE*, 12, 5, 1-32, 2017.
6. European Social Survey Round 9 Data. Data file edition 1.2. NSD - Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC, 2018. doi:10.21338/NSD-ESS9-2018.
7. Eurostat. Task Force on core social variables: final report (Theme: Population and social conditions; Collection: Methodologies and working papers). Office for Official Publications of the European Communities, Luxembourg, 2007.
8. A. Filhon, J. Deauvieu, L. de Verdalle, A. Pelage, T. Poullaouec, C. Brousse, M. Mespoulet, & K. Sztandar-Sztanderska. European classification project: an assessment of national variations in the perception of social space. *Comparative Sociology*, 15, 3, 275-299, 2016.
9. T. Goedemé. A note on the replication of the European socio-economic classification (ESeC) in the EU Statistics on income and living conditions (EU-SILC). Institute for New Economic Thinking (INET) Oxford Working Paper No. 2019-17, Department of Social Policy and Intervention, University of Oxford, 2019.
10. E. Harrison and D. Rose. The European socio-economic classification (ESeC) user guide, Institute for Economic and Social Research, University of Essex, Colchester, UK, 2006.
11. E. Harrison and D. Rose. From derivation to validation: evidence from the UK and beyond. In D. Rose and E. Harrison (Eds.), *Social class in Europe: An introduction to the European socio-economic classification* (pp. 39-60), Routledge, New York, 2010.
12. Institute of Social and Economic Research. ESeC, University of Essex, <https://www.iser.essex.ac.uk/archives/esecc>
13. L. Kish. Multi-population survey designs: Five types with seven shared aspects. *International Statistical Review*, 62, 2, 167-186, 1994.
14. P. S. Lambert, & E. Bihagen. Concepts and measures: empirical evidence on the interpretation of ESeC and other occupation-based social classifications. Paper presented at the Research Committee 28 (RC28) on Social Stratification and Mobility Summer Meeting, Montreal, 14-17 August 2007.
15. P.S. Lambert, & E. Bihagen. Using occupation-based social classifications. *Work, Employment & Society*, 28, 3, 481-494, 2014.
16. T. Maloutas. Socio-economic classification models and contextual difference: the 'European socio-economic classes' (ESeC) from a Southern European angle, *South European Society and Politics*, 12, 4, 443-460, 2007.
17. C. Michalopoulou. Likert scales require validation before application — Another cautionary tale, *Bulletin de Méthodologie Sociologique*, 134, 5-23, 2017.
18. C. Moser and G. Kalton. *Survey Methods in Social Investigation*, Heinemann Educational Books, London, 1975.

19. D. Rose, E. Harrison. The European socio-economic classification: a new social class schema for comparative European research, *European Societies*, 9, 3, 459-490, 2007.
20. D. Rose, E. Harrison and D. Pevalin. The European socio-economic classification. In D. Rose and E. Harrison (Eds.), *Social class in Europe: An introduction to the European socio-economic classification* (pp. 3-38), Routledge, New York, 2010.
21. The ESS Sampling Expert Panel. Sampling guidelines: principles and implementation for the European Social Survey, ESS ERIC Headquarters, London, 2016. <http://www.europeansocialsurvey.org/>
22. K.G. Tijdens. ESEG-2014 coding scheme + explanatory note. Deliverable 8.13 of the SERISS project funded under the European Union's Horizon 2020 research and innovation programme GA No: 654221, 2016. Available at: www.seriss.eu/resources/deliverables
23. D. Watson, C.T. Whelan, & B. Maître. Validating the European socio-economic classification: cross-sectional and dynamic analysis of income poverty and lifestyle deprivation, ESRI Working Paper, No. 201, The Economic and Social Research Institute (ESRI), Dublin, 2007.
24. H. Wirth, C. Gresch, W. Müller, R. Pollak and F. Weiss. Measuring social class: the case of Germany. In D. Rose and E. Harrison (Eds.), *Social class in Europe: An introduction to the European socio-economic classification* (pp. 114-137), Routledge, New York, 2010.
25. A. Yfanti, C. Michalopoulou and S. Zachariou. The impact of definitions in classifying the employed, unemployed and inactive when comparing measurements from different sources, *Communications in Statistics: Case Studies, Data Analysis and Applications*, 5, 1, 46-5, 2019.

Improved insurer's capital adequacy of reserve risk using copula approach and hypothesis tests

Ilze Zariņa¹, Irina Voronova², and Gaida Pettere³

¹ Department of Entrepreneurship and Management, Faculty of Engineering Economics and Management, Riga Technical University, Riga, Latvia
(E-mail: ilzezarina@inbox.lv)

² Department of Entrepreneurship and Management, Faculty of Engineering Economics and Management, Riga Technical University, Riga, Latvia
(E-mail: irina.voronova@rtu.lv)

³ Department of Engineering Mathematics, Faculty of Computer Science and Information, Riga Technical University, Riga, Latvia
(E-mail: gaida@latnet.lv)

Abstract. Putting aside adequate amount of capital and absorbing losses even during recession times are important for financial stability management and for shareholders. There are non-linear dependence and heavily skewed loss distributions in insurance. Copula as risk-aggregation measure is not yet widely used in the insurance sector. Therefore, we are going to study how to choose the most appropriate type of copula for non-life reserve risk, calculate adequate capital by applying value-at-risk at 99.5% which is mandatory in EU market, and select the copula and hypothesis tests to choose the most appropriate copula type for reserve risk. A case study based on actual data will be discussed.

Keywords: VaR, copula approach, insurance economic modelling, internal capital modelling, hypothesis tests, selected copula tests, stability management.

1 Introduction

The insurance industry is facing new trends—pandemic, digitalisation, climate change, increased cost of capital and economic slowdown. These trends have created new emerging risks, which the industry is facing.

Major consulting companies like KPMG [13], PwC [16], and Deloitte [3] have published articles on the pandemic topic by highlighting employees' and partners' health protection, reviewing claim-handling policy for exclusions and therefore minimising reputation risk, and being prepared for further extreme volatility in the financial markets. Digital transformation in insurance industry could lead in the shorter time horizon to personalised products. These are premium risk pricing by avoiding overpricing and also individual capital modelling based on companies' individual risk profile.

In order to maintain financial stability, the first step is to prepare forward-looking capital projection forecasts taking into account uncertainty of net asset value, capital requirement, solvency coverage ratio, and foreseeable dividends. It

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



cannot be done without a proper risk and stability management culture, which includes insurance products' risk-aggregation assessment. Standard approach (standard formula) uses linear correlation matrix, but there are non-linear dependence and heavily skewed loss distributions in the insurance sector. One of the solutions is to use the copula approach for underwriting risks by partly solving this issue with the internal model. Solvency II regime [20] regulates that capital must be assessed within one-year period value-at-risk at 99.5% confidence level for any type of model—standard, internal, partial internal.

We are focusing on appropriate copula model with practical case study example and not describing insurance industry basics, proofs, actuarial reserving, claim distributions and mandatory regulators' requirements (Solvency II framework) for internal model. We are modelling necessary capital for reserve risk which answers the question 'how large are capital needs in absolute amounts for potential reserve shifts from reserve set aside in economical balance sheet and simulations at 99.5% confidence level in 12-month time horizon, e.g. from car accident. Capital for reserve risk is only one but a major part of the risk many insurers are facing such as market risk, counterparty default, natural and man-made catastrophes' risk. The aim of the paper is to investigate how to develop an internal capital model methodology by using a copula approach and choose the most appropriate type of copula for non-life reserve risk. The object is improved capital requirement of insurance company for reserve risk. The subject is copula approach and goodness-of-fit tests for capital requirement for reserve risk.

By searching for the terms "insurance & capital & copula" in the journal database Scopus, we would like to highlight just published research by Mejdoub and Arab [8] in which authors attempt to apply capital modelling to a portfolio of non-life insurance risks for a Tunisian insurance company by using bivariate copula approach. Results shows that the insurance companies must be careful in the choice of the suitable copula and take into consideration an absence of tail dependence and an upper tail dependence for specific copula families and that t-Student copula gives the highest capital requirements. Tail-dependence application and the skew-copula was defined on the basis of multivariate skew t-distribution by Kollo and Pettere [12] following Azzalini and Capitanio [1]. Kolmogorov Smirnov and Cramer-von-Mises statistic goodness-of-fit tests are often used to test copula models. We have used cross validation (AIC principle) and parametric bootstrap (method-of-moments estimation principle) tests. We have not identified research papers where these specific tests are applied for insurance sector copula model testing.

The paper is organised as follows. First, the copula theorem is presented with copula families that are used in modelling (Section 2). Second, types of goodness-of-fit tests are summarised for choice of copula family (Section 3). Next, the simulation results are presented and results show how improved capital adequacy can be reached (Section 4). The last section concludes the paper and indicates the future challenges.

2 The improved capital adequacy models and its assumptions

We have used copulas for risk aggregation. Copulas are particular multivariate distribution functions. Let us recall that the distribution function H of a d -dimensional random vector $\mathbf{X} = (X_1, \dots, X_d)$ is the function defined by

$$H(\mathbf{x}) = \mathbb{P}(\mathbf{X} \leq \mathbf{x}) = \mathbb{P}(X_1 \leq x_1, \dots, X_d \leq x_d), \mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d. \quad (1)$$

The distribution function F_j of X_j , $j \in \{1, \dots, d\}$, can be recovered from the multivariate distribution function \mathbf{H} by $F_j(x_j) = H(\infty, \dots, \infty, x_j, \infty, \dots, \infty)$, $x_j \in \mathbb{R}$. This is why F_1, \dots, F_d are also called the univariate margins of \mathbf{H} or the marginal distribution functions of \mathbf{X} . Sklar's [18] and [19] theorem is the central theorem and is used to create copula families from existing families of multivariate distribution functions. Sklar's theorem is attributed to Sklar [18].

Normal copula

The d -dimensional normal copula C_P^n is the copula defined by Sklar's theorem from the multivariate normal distribution $N_d(\mathbf{0}, P)$ with correlation matrix P . If Φ_P denotes the distribution function of the latter, $C_P^n(\mathbf{u})$ is given, for any $\mathbf{u} \in [0, 1]^d$, by

$$\begin{aligned} C_P^n(\mathbf{u}) &= \Phi_P(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)) = \\ &= \int_{-\infty}^{\Phi^{-1}(u_d)} \dots \int_{-\infty}^{\Phi^{-1}(u_1)} \frac{\exp(-\frac{1}{2}\mathbf{x}'P^{-1}\mathbf{x})}{(2\pi)^{\frac{d}{2}}\sqrt{\det P}} d\mathbf{x}_1 \dots d\mathbf{x}_d, \end{aligned} \quad (2)$$

where Φ^{-1} denotes the quantile function of $N(0,1)$ [7].

t -copula

The d -dimensional t copula $C_{P,v}^t$ is the copula defined by Sklar's theorem from the multivariate t distribution with location vector $\mathbf{0}$, scale matrix P and $v > 0$ degrees of freedom. If $t_{P,v}$ denotes the distribution function of the latter, $C_{P,v}^t(\mathbf{u})$ is given, for any $\mathbf{u} \in [0, 1]^d$, by

$$\begin{aligned} C_{P,v}^t(\mathbf{u}) &= t_{P,v}(t_v^{-1}(u_1), \dots, t_v^{-1}(u_d)) = \\ &= \int_{-\infty}^{t_v^{-1}(u_d)} \dots \int_{-\infty}^{t_v^{-1}(u_1)} \frac{\Gamma(\frac{v+d}{2})}{\Gamma(\frac{v}{2})(\pi v)^{\frac{d}{2}}\sqrt{\det P}} \left(1 + \frac{\mathbf{x}'P^{-1}\mathbf{x}}{v}\right)^{-\frac{v+d}{2}} d\mathbf{x}_1 \dots d\mathbf{x}_d, \end{aligned} \quad (3)$$

where t_v^{-1} denotes the quantile function of t_v of the univariate Student t distribution with v degrees of freedom [7].

Other methods used and assumptions

Assume now that the copula C has been selected. We are interested in the value-at-risk (VaR) of a position by using the Monte Carlo method which generates a number N of such scenarios and the sample α -quantile is then the one period value-at-risk with confidence α defined by:

$$VaR_\alpha = F_{L^+}^{\leftarrow}(\alpha), \quad (4)$$

where L^+ is aggregate loss, F_{L^+} is known as loss distribution and α describes confidence level where $\alpha \in (0, 1)$. The capital needs are the difference between reserve in economical balance sheet BE_{total} and the value-at-risk with confidence level $\alpha=0.995$:

$$Capital_{total} = |BE_{total} - VaR_{0.995}|. \quad (5)$$

We have used $N=10000$ and $\alpha=0.995$ and R functions in the package *copula*: *normalcopula()*, *tcopula()*.

3 Model Selection tests

We have used hypothesis tests in order to validate various copulas' models. In same case basic graphical diagnostics can be enough in practice for risk assessment approximations. It is not a case of internal capital model methodology, documentation package for national regulators, and financial market authorities. Formal statistical tests which compute p -values that can help to guide the choice of the hypothesized copula family C are crucial and play an important role. We assume this goodness-of-fit issue for adequate parametric copula family amounts formally to testing

$$H_0: C \in \mathcal{C} \text{ versus } H_1: C \notin \mathcal{C}, \quad (6)$$

where H_0 explains that the choice of the hypothesised copula family C cannot be rejected and H_1 explains that the choice of the hypothesised copula family C can be rejected.

3.1 Parametric Bootstrap

As suggested in Fermanian [4], Quessy [17], and Genest and Rémillard [5], a natural goodness-of-fit test consists of comparing C_n with an estimate C_{θ_n} of C

obtained under the assumption that $C \in \mathcal{C}$ holds. The estimated margins are used to form the sample

$$\mathbf{U}_{i,n} = (F_{n,1}(X_{i1}), \dots, F_{n,d}(X_{id})), i \in \{1, \dots, n\}, \quad (7)$$

where for any $j \in \{1, \dots, d\}$, F_j is estimated by using component samples of $\mathbf{X}_1, \dots, \mathbf{X}_n$

$$F_{n,j(x)} = \frac{1}{n+1} \sum_{i=1}^n 1(X_{ij} < x), x \in \mathbb{R}. \quad (8)$$

In the previous statement, $\boldsymbol{\theta}_n$ is an estimator (parameter vector) of $\boldsymbol{\theta}$ computed from the pseudo-observations $\mathbf{U}_{1,1}, \dots, \mathbf{U}_{n,n}$ such as the maximum pseudo-likelihood estimator.

We are using an approach that appears to perform particularly well according to the large scale simulations carried out in Genest et al. [6] where procedure is based on Cramer-von Mises statistic

$$S_n^{gof} = \int_{[0,1]^d} n (C_n(\mathbf{u}) - C_{\theta_n}(\mathbf{u}))^2 dC_n(\mathbf{u}) = \sum_{i=1}^n (C_n(\mathbf{U}_{i,n}) - C_{\theta_n}(\mathbf{U}_{i,n}))^2, \quad (9)$$

An approximate p -value for the test based on S_n^{gof} can be obtained by means of a parametric bootstrap whose asymptotic validity is investigated in Genest and Remillard [5]. Advantage is its conceptual simplicity.

Parametric Bootstrap algorithm is summarised by Hofert et al. [10] as follows:

1. Compute the pseudo-observations $\mathbf{U}_{1,1}, \dots, \mathbf{U}_{n,n}$.
2. Compute an estimate $\boldsymbol{\theta}_n$ of $\boldsymbol{\theta}$ from the pseudo-observations $\mathbf{U}_{1,1}, \dots, \mathbf{U}_{n,n}$.
3. Compute the test statistic S_n^{gof}
4. For some large integer N , repeat the following steps for every $k \in \{1, \dots, N\}$:
 - 4.1 Generate a pseudo-random sample $\mathbf{U}_1^{(k)}, \dots, \mathbf{U}_n^{(k)}$ from the fitted copula C_{θ_n} and compute the corresponding pseudo-observations $\mathbf{U}_{1,n}^{(k)}, \dots, \mathbf{U}_{n,n}^{(k)}$.
 - 4.2 Compute an estimate $\boldsymbol{\theta}_n^{(k)}$ of $\boldsymbol{\theta}$ from the pseudo-observations $\mathbf{U}_{1,n}^{(k)}, \dots, \mathbf{U}_{n,n}^{(k)}$ using the same (rank-based) estimator as in Step 2.
 - 4.3 Compute the corresponding version $S_n^{gof,(k)}$ of S_n^{gof} as:

$$S_n^{gof,(k)} = \sum_{i=1}^n (C_n^{(k)}(\mathbf{U}_{i,n}^{(k)}) - C_{\theta_n^{(k)}}(\mathbf{U}_{i,n}^{(k)}))^2, \quad (10)$$

where

$$C_n^{(k)}(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^n 1(\mathbf{U}_{i,n}^{(k)} \leq \mathbf{u}), u \in [0,1]^d. \quad (11)$$

Under H_0 , $S_n^{gof,(k)}$ can be thought of an approximately independent copy of S_n^{gof} .

5. An approximate p -value for the test is given by

$$\left(\frac{1}{2} + \sum_{k=1}^N \mathbf{1}(S_n^{gof,(k)} \geq S_n^{gof})\right) / (N + 1). \quad (12)$$

We have used **R** function `gofcopula()` for carrying out goodness-of-fit tests in the package `copula`

3.2 Cross-validation criterion

There could be situation that all candidate parametric copula families are rejected if sample size is large or none of candidate parametric copula families are rejected if sample size is small. Test that uses *Akaike information criterion* (AIC) and performs the selection of the best ranked family can be justified by using formula

$$AIC = 2(l_{n,max} - p), \quad (13)$$

where $l_{n,max}$ is the maximized likelihood and p is the total number of marginal and copula parameters.

Grønneberg and Hjort [7] has defined cross-validation copula information criterion, up to a multiplicative constant, the first-order equivalent of the cross validation criterion

$$\widehat{xv}_n = \frac{1}{n} \sum_{i=1}^n \log c \theta_{n,-1}(F_{n,-i}(\mathbf{X}_i)), \quad (14)$$

where $\theta_{n,-1}$ is the maximum pseudo-likelihood estimate computed from the sample $\mathbf{X}_1, \dots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \dots, \mathbf{X}_n$ and

$$\mathbf{F}_{n,-i}(\mathbf{x}) = \left(F_{n,1,-i}(x_1), \dots, F_{n,d,-i}(x_d) \right), \quad \mathbf{x} \in \mathbb{R}^d, \quad (15)$$

with

$$F_{n,j,-i}(x) = \begin{cases} \frac{1}{n \sum_{k=1, k \neq i}^n \mathbf{1}(X_{kj} \leq x)}, & \text{if } x \geq \min_{k \in \{1, \dots, n\} \setminus \{i\}} X_{kj} \\ 1/n, & \text{otherwise.} \end{cases} = \pi r^2$$

This test will leave out and penalise copula families with too many parameters that tend to overfit. Papers of authors such as Claeskens and Hjort [2], Grønneberg and Hjort [7], Jordanger and Tjøstheim [11], McNeil et al. [14] help to improve AIC formula approach and historical development in copula theory in a more detailed way. We have used **R** for carrying out AIC tests in the package function called `xvcopula()`.

4 Simulation studies

Data are gathered from EU insurance company and claims triangles are changed by using log function, which does not change risk dependence structure, which is needed for finding appropriate copula model and risk aggregation calculation. Capital by standard model under Solvency II regime [20] is calculated as follows where 3 standard deviations represent parametric VaR at 99.5% confidence level for log-normal distribution with given correlation matrix:

$$\sigma_{total} = \frac{\left(\sqrt{\text{SUMPRODUCT}(3.6 \times 0.09 \quad 0.5 \times 0.1 \quad 2.6 \times 0.11); \begin{pmatrix} 1 & 0.25 & 0.25 & 3.6 \times 0.09 \\ 0.25 & 1 & 0.25 \times 0.5 \times 0.1 \\ 0.25 & 0.25 & 1 & 2.6 \times 0.11 \end{pmatrix}} \right)}{(3.6 + 0.5 + 2.6)} = 0.082,$$

and finally

$$\text{Capital}_{total} = 3 \cdot \sigma_{total} \cdot \text{BE}_{total} = 3 \times (3.6 + 0.5 + 2.6) \times 0.082 = 1.6.$$

Reserve risk distributions, its parameters and correlations are known for three insurance products - motor third party liability for private persons, commercial property insurance against fire and natural catastrophes, and professional third party liability (see Table 1).

Table 1. Input parameters and capital in mEUR with various copula models

<i>Insurance products</i>	<i>Motor third party liability for private persons</i>	<i>Commercial property insurance against fire and natural catastrophes</i>	<i>Professional third party liability</i>
Reserve risk distribution	<i>Log-normal and correlation between every two insurance products is the same ρ^*</i>		
Mean and standard deviation of the distribution on the log scale with default values of 0 and 1 respectively	15; 0.09	13; 0.1	15; 0.11
BE_{total} - Best estimate for reserve in economical balance sheet after reinsurance	3.6	0.5	2.6
Capital_{total} - Capital by using standard model	1.6		

*where $\rho = \begin{pmatrix} 1 & 0.25 & 0.25 \\ 0.25 & 1 & 0.25 \\ 0.25 & 0.25 & 1 \end{pmatrix}$ - correlation matrix used in simulation

We have used R version 3.5.1 and package copula by Hofert et al. [9] and package *gofCopula* by Okhrin et al. [15]. Copula models are considered in internal capital modelling—normal copula, and *t-copula* with fixed 4 degree of freedom (see Table 2).

Table 2. Potential candidates for improved capital adequacy in mEUR - Copula models

Assumptions: N=10 000, 3 – dimensional observation, $\rho = 0.25$, value-at-risk 99.5%	
Option A: <i>Capital by using improved model - Normal copula,</i>	1.4
Option B: <i>Capital by using improved model - t-copula if degree of freedom is 4</i>	1.5

Finally, model selection is done by using goodness-of-fit tests in line with Section 3. It can be seen below that t-copula with 4 degrees of freedom (df) can be rejected (Section 3, H_0 defined in formula (6)) with significance at the 0.05 level (Option B). Models with normal copula (Option A) are plausible and cannot be rejected.

Table 3. Goodness of fit results for various copula models

<i>Approach</i>	<i>Maximum pseudo-likelihood</i>	<i>Goodness of Fit - Parametric Bootstrap</i>	<i>Cross-validation - Model selection</i>	<i>Conclusions</i>
<i>R package functions</i>	<i>fitCopula()</i>	<i>gofcopula()</i>	<i>xvcopula()</i>	
Option A: <i>Normal copula</i>	8.114	0.0192	5.246	cannot be rejected, plausible
Option B: <i>t-copula df=4</i>		0.2135	-623.243	reject H_0

This example shows that the most basic copula—normal copula—can be used as the solution in order to decrease capital requirement and reach improved capital adequacy ratio. There are no need for more complicated models for the main business lines and products: motor third party liability insurance for private persons, commercial property insurance against fire and natural catastrophes, and professional third party liability insurance.

Copula models have large data sets. Calculation tests have the advantage of being easy to implement with R but the disadvantage of being computationally expensive.

Conclusions

Internal capital model with copula approach can be assessed with goodness-of-fit tests—cross validation (AIC principle) and parametric bootstrap (method-of-moments estimation principle). Both tests are easily implemented in R software, but calculation is computationally expensive and time consuming for the large scale of insurance data.

Simulation studies showed that we should accept normal copula with a 0.05 significance level. Basic copula family—normal—can be used for non-life insurance market if underlying assumptions hold and with given data sample. Improper risk aggregation approach and split by-products can lead to bad business decisions by stopping underwriting for a certain product and improper capital

planning during budgeting process. Also it is important to take into consideration that normal copula ignores upper tail dependence when the Pearson correlation coefficient is smaller than one.

The suggested topics for further research include other copula families (e.g. skew t-copula) and include finding an appropriate type of copula (including tail dependence aspects) insurance sector in Baltic states for reserve risk modelling if reserve risk is distributed with more narrow size distribution e.g. Gamma, Weibull, and Pareto. Also further studies should continue if dimensions, insurance products are more than three.

References

1. Azzalini, A., Capitanio, A. (2003), *Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution*. J. R. Stat. Soc. Ser. B Stat. Methodol. 65, 367-389.
2. Claeskens, G., & Hjort, N. L. (2009). *Model selection and model averaging*. London: Cambridge University Press. Psychometrika 76, 507–509 (2011). <https://doi.org/10.1007/s11336-011-9219-3>
3. Deloitte. (2020). *Potential impact of Covid19 to insurance industry*, Retrieved from <https://www2.deloitte.com/uk/en/insights/economy/covid-19/impact-of-covid-19-on-insurers.html?id=us:2em:3pa:financial-services:eng:di:031720>.
4. Fermanian, J.-D. (2005). *Goodness-of-fit tests for copulas*. Journal of Multivariate Analysis, 95(1), 119–152. doi:10.1016/j.jmva.2004.07.004
5. Genest, C., & Rémillard, B. (2008). *Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models*. Annales de l'Institut Henri Poincaré: Probabilités et Statistiques, 44, 1096–1127, 2008, doi:10.1214/07-AIHP148
6. Genest, C., Rémillard, B., & Beaudoin, D. (2009) *Goodness-of-fit tests for copulas: A review and a power study*. Insurance: Mathematics and Economics, 44. 199-213. 10.1016/j.insmatheco.2007.10.005
7. Grønneberg, S., & Hjort, N. L. (2014). *The copula information criteria*. Scandinavian Journal of Statistics, 41, 436–459, <https://doi.org/10.1111/sjos.12042>
8. Mejdoub, H., & Arab, M.B. 2019. *Insurance risk capital and risk aggregation: bivariate copula approach*. International Journal of Computational Economics and Econometrics, Inderscience Enterprises Ltd, vol. 9(3), pages 202-218.
9. Hofert, M., Kojadinovic, I., Maechler M, Yan J. copula: Multivariate Dependence with Copulas. R package version 1.0-0, <https://CRAN.R-project.org/package=copula>, 2020
10. Hofert, M., Kojadinovic, I., Mächler, & M., & Yan, J. (2018). *Elements of Copula Modeling with R*, 267. Switzerland: Springer International Publishing, <https://doi.org/10.1007/978-3-319-89635-9>
11. Jordanger, L. A., & Tjøstheim, D. (2014). *Model selection of copulas: AIC versus a cross validation copula information criterion*. Statistics & Probability Letters, 92, 249–255, <https://doi.org/10.1016/j.spl.2014.06.006>
12. Kollo, T., Pettere, G. (2010). *Parameter estimation and application of the multivariate skew t-copula*. In: Copula Theory and Its Applications. Eds. P. Jaworski et al. Springer-Verlag, Berlin, 289-298.

13. KPMG. (2020). *Do insurers have COVID-19 covered?* , Retrieved from <https://home.kpmg/xx/en/home/insights/2020/03/do-insurers-have-covid-19-covered.html>
14. McNeil, A. J., Frey, R., & Embrechts, P. Quantitative risk management: Concepts, techniques and tools (2nd ed.). Princeton, NJ: Princeton University Press, 2015.
15. Okhrin, O., Trimborn, S., Waltz, M. (2020). *gofCopula: Goodness-of-Fit Tests for Copulae*. Discussion Paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3560825
16. PwC. (2020). *COVID-19 and the insurance industry*. Retrieved from <https://www.pwc.com/us/en/library/covid-19/covid-19-and-insurance-industry.html>.
17. Quessy, J.-F., *Méthodologie et application des copules: tests d'adéquation, tests d'indépendance, et bornes sur la valeur-à-risque*. PhD thesis, Université Laval, Québec, Canada, 2005.
18. Sklar, A. (1959). *Fonctions de répartition à n dimensions et leurs marges*. Publications de l'Institut de Statistique de l'Université de Paris, 8, 229–231.
19. Sklar, A. (1996). *Random variables, distribution functions, and copulas – A personal look backward and forward*. Distributions with Fixed Marginals and Related Topics, 28, 1–14.
20. Solvency II: European Insurance and pensions authority. (2011). *Calibration of the Premium and Reserve Risk Factors in the Standard Formula of Solvency II* , Retrieved from: https://register.eiopa.europa.eu/Publications/Reports/EIOPA-11-163-A-Report_JWG_on_NL_and_Health_non-SLT_Calibration.pdf.

Does the death postponement phenomenon really exist?

Sergei Zuyev and Holger Rootzén

Chalmers University of Technology and University of Gothenburg,
412 96 Gothenburg, Sweden
(E-mails: sergei.zuyev@chalmers.se and holger@chalmers.se)

Abstract. It is a common belief that people close to death from natural causes can postpone their imminent death if they see a strong reason to survive a bit longer. This is known as the Postponement hypothesis: that a meaningful occasion can act as a motivator to prolong life for a short amount of time. A few studies have already addressed this hypothesis but their conclusions are contradictory.

To check the postponement hypothesis, we analysed almost 249 thousand cases in the dataset for South African people who died in the year 2015. We took a person's birthday as the meaningful occasion and analyse the death rate around this date using statistical models offered by survival analysis. If the hypothesis is true, it can be expected that the mortality rate should be lower a period just before the birthday and, perhaps, higher shortly afterwards.

The results of our analysis show that no postponement of death can be seen for the examined dataset. In fact, to the contrary, the data suggest that the mortality rate is higher both before and after the birthday. Speculations as to why this is the case might be a higher risk associated with the stress of expectations for the birthday as well as an earlier start of celebrations with associated departure from the recommended regime.

Keywords: death rate, hazard function, postponement, longevity, survival.

1 Introduction

The Death Postponement theory has been with us for some time. It states that a person is able to postpone his/her death for a while if there is a strong reason for this. Such a reason could be personal events like the soon coming birthday or public events like important historical or religious dates close-by. The theory is popular because it is easy to believe that a person is, at least, in some control of one's own life even in critical circumstances. A number of studies has been carried out so far to see if it has any statistical grounds for the postponement to exist. Perhaps, the main proponent of the theory is David P. Philips who has published several articles with different co-authors studying the postponement phenomenon.

In the study [5], the authors investigated whether the postponement phenomenon can be detected using the Jewish holiday of Passover, Pesach, as a

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

© 2020 ISAST



significant event. A dataset of 1919 adult humans, which died in California of natural causes during the years 1966–1984 was used as a basis for their analysis and was compared with a control group consisting of non-Jewish people. Application of regression analysis and binomial tests allowed the authors to detect a significant decrease (dip) in the number of deaths observed just before Pesach and their increase (peak) just after, while the control group did not show the same pattern. The results were later criticised by Gary Smith in [8] who noted that in the selection process, people were assumed to be Jewish only by their names. Another study by D.P. Philips examined the mortality of 1288 Chinese who died between the years 1960 and 1984 around the Moon Festival, a traditional celebration of Chinese culture [6]. Again, the authors claim to show a similar dip-peak pattern that was significant for Chinese and did not occur in a non-Chinese control group. The authors, however, noted that the sample size in both studies, of Pesach and the Moon party, were small in size, the statement which puts in doubts the very conclusion of these studies.

In a later work, entitled “The Birthday: Lifeline or Deadline?” [7], the authors analysed much larger samples of 1.3 and 1.4 million people aged 18+ who died of natural causes between 1969 and 1990 in California using birthday as a significant event. The analyses seem to demonstrate the postponement phenomenon for female population (so the birthday for females is a ‘lifeline’), while for the male population the death rate is actually higher both before and after the birthday (a ‘deadline’).

The work of other authors mainly confirm this ‘birthday deadline phenomenon’ that the mortality actually increases before the birthday. The authors of [1] analyse Swiss mortality databases from years 1969-2008 containing over 2 million death records with the help of the ARIMA model. They show a 13.8% increase in death rate around birthdays with variations of between 11 to 18 per cent in men and women older than 60. In the group of natural causes deaths, the heart disease and cancer were the main causes of death. The conclusion is that birthdays increase the death rate mainly for the heart disease patients (infarction and stroke), certainly, due to extra stress. Another study [4] analyses more than 4 million death records in Germany during the period of 1992–2011. As the meaningful event, both Christmas and birthday were used. The conclusion is similar: there is no such phenomenon as an intentionally postponed death, at least for birthdays and on the scale of a few days.

A frequent source of error in data interpretation leading to belief in existence of the postponement phenomenon can be illustrated with the following example.

Consider a typical situation when the number of recorded deaths decreases with age, like on the left histogram in Figure 1. If we combine these data into the monthly death statistics, essentially by cutting the histogram on the left by the year start, we obtain the histogram on the right which would also demonstrate a decaying pattern. This may lead to a wrong interpretation as the postponement phenomenon that less people die before their birthday than after it. This, however, just reflects the fact that there are fewer people who survive to their next birthday. The same statement is true for any other cut-off day in the year rather than the birthday provided a decaying with age histogram pattern.

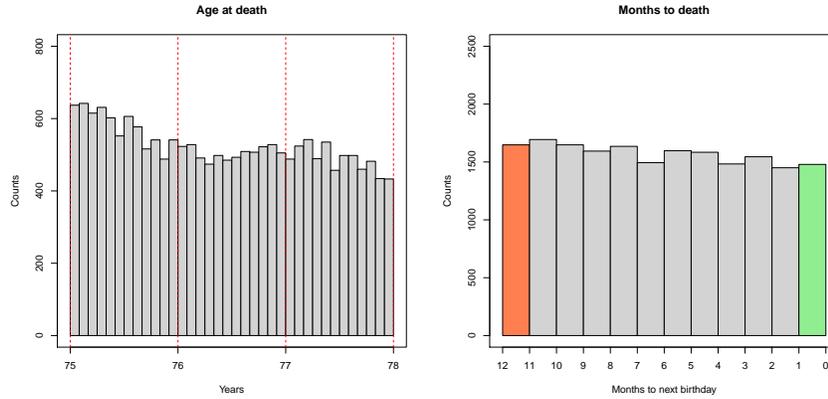


Fig. 1. Left plot: histogram of the number of deaths during a year for different ages. Right plot: histogram of the same data grouped into months. MCD-2015 dataset [2].

2 Data and pre-processing

We worked with the Mortality and Causes of Death in South Africa MCD-2015 dataset [2] which contains over 422 thousand records of death occurred in 2015 (or earlier, but reached Stats SA governmental agency in 2015) in South Africa. Just over 1700 records had missing birth dates and were excluded from the analysis. Also excluded were deaths caused by accidents since no death postponement could be expected for these. We also limited ourselves to people who died aged between 49 and 81 years. This left us with the total number of 204,367 natural death records to work with. We think that older people should be more concerned about their soon eminent death and express the postponement more clearly, if it exists. As for the people who died at age 82 or more (about 13% of all the deaths), the number of their deaths around birthday is too small to make any statistically credible statements about a subtle death rate changes assumed by the postponement phenomenon.

3 Theoretical model

We assume that each living person aged x is exposed to an instantaneous death rate $h(x)$ depending on the age and possibly on the absolute time, so that the probability that a person of age x survives to the age $x + \Delta x$ is $1 - h(x)\Delta x + o(\Delta x)$. Thus, if τ denotes the life duration of a randomly uniformly selected newborn, then the *Survival function* is given by

$$S(x) = \mathbf{P}\{\tau > x\} = \exp\left\{-\int_0^x h(y) dy\right\}. \quad (1)$$

Obviously, the p.d.f. $f_\tau(x)$ of τ and $S(x)$ are related through

$$h(x) = \frac{f_\tau(x)}{S(x)} \quad (2)$$

which is the definition of the *hazard function*. Assuming that new births happen in time as a Poisson point process with a rate ν , generally depending on time, the death records $\{(t_i, x_i)\}$, describing a person i to die at time t_i aged x_i , is a Poisson point process with the intensity function $\mu(t, x) = \nu f_\tau(x)$ in $\mathbb{R} \times \mathbb{R}_+$. Thus the number of deaths of people aged between x and $x + \Delta x$ years during a time interval of length s is Poisson distributed with parameter $s\nu(S(x) - S(x + \Delta x)) = s\nu S(x)h(x)\Delta x + o(\Delta x)$ and the death counts are independent for disjoint age ranges.

The birth rates are known to exhibit a seasonality. Assuming that people born in different time of the year have the same tendency towards the death postponement, when counted from an anniversary, the yearly pattern averages to a profile which would show the same susceptibility to the postponement. Thus, for its detection, we may assume that the parameters ν and h do not depend on the absolute time t since, when related to an anniversary, they are averaged over the year.

A popular distribution to describe the remaining lifetime after attaining a certain age a is the *generalised Pareto GP* (ξ, a, σ) -distribution, for which

$$\begin{aligned} S(x) &= \left(1 + \frac{\xi(x-a)}{\sigma}\right)^{1/\xi}; \\ \frac{1}{h(x)} &= \sigma + \xi(x-a). \end{aligned} \tag{3}$$

When $\xi = 0$, the hazard is constant and $S(x) = \exp\{-(x-a)/\sigma\}$ so that $\tau - a$ is Exponentially $\text{Exp}(1/\sigma)$ -distributed. As we will see in Section 4.2 below, the linear form of the inverse hazard $1/h$ agrees with our data suggesting a GP-distribution for the remaining lifetime after a person turns 50 years old.

4 Methods and analysis

We employ two methods to detect possible postponement phenomenon: Poisson regression to estimate the Poisson process intensity $\mu(t, x)$ above and fitting the hazard function suggested by the GP-distribution.

4.1 Modeling the counts

The counts of deaths for each day lived after the 49th anniversary is shown at the upper plot on Figure 2.

Our approach consists in fitting a linear model to

$$\log \mu(t, x) = \log \nu + \log f_\tau(x)$$

not counting the people who died within 7 days before, at and within 7 days after the birthday. We then analyse the residuals produced by the fitted model for these days around the birthday to see if their mean is significantly different from the residuals of the other days. The death postponement would mean that the residuals for the days prior to birthdays are on average smaller and within

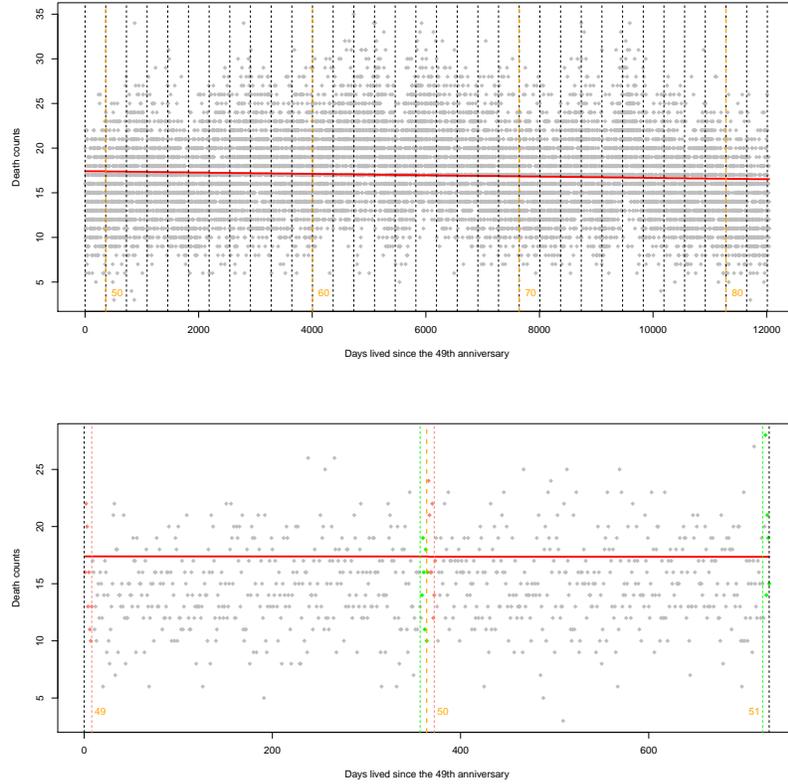


Fig. 2. Top: the death counts for each day between the ages from 49 to 81 years old. Bottom: similar, but from 49 to 51 years old only.

the week after, perhaps, larger than the residuals for the other days. That would indicate a dip in the number of deaths prior to the birthdays possibly followed by a peak just after it.

The estimated exponential curve for μ using generalised linear model fitting with Poisson distribution is shown in red on the top plot on Figure 2. The two-sample t -test was used for assessment of the presence of dips or peaks. We found that there is no significant peak after the birthday (p -value = 0.13), but contrarily to the postponement hypothesis, we found a significant *peak before* the birthday (p -value = 0.05). The residual deviance was 13529 on 11547 degrees of freedom which indicates over-dispersion in the Poisson regression. Therefore the quasi-Poisson family as well as the negative binomial regression were also tried which, however, did not vary much from the Poisson model and gave the same result. We may argue that the age only partly explains the mortality: a significant part of the variation is due to other factors, in particular, the mentioned seasonality of the true birth rate.

The shape of the data on the upper plot on Figure 2 suggests that there may be some extra curvature not explained by the Poisson model on the whole data range and hence its conclusions may not be accurate. Therefore we proceed with analysis of each year individually: for y ranging from 50 to 81, we fit a Poisson model to the number of deaths at the ages from anniversary $y - 1$ plus one week to anniversary $y + 1$ minus one week *excluding* deaths at the age of one week around anniversary y . Only one anniversary showed a significant dip before birthday (p -value less than 0.05). Also one anniversary showed a peak before birthday and 4 had p -value below 10%. Four anniversaries showed a peak after birthday (and 7 had p -value less than 10%). For 32 years considered, 4 peaks or dips is not a significant figure: the probability for the binomially distributed random variable with $n = 32$ and $p = 0.05$ to take values 4 or more is 0.074, this is the Binomial test. On the other hand, it is a 90%-significant figure, since the probability for the binomially distributed random variable with $n = 32$ and $p = 0.1$ to take values 7 or more is 0.036. Strictly speaking, the binomial test is not applicable since the consecutive years $y, y + 1$ use the same data between y and $y + 1$ to estimate the regression. However, taking just the even years still gives the same number of peaks after the birthday.

We also noted that excluding two weeks before and after the birthday when constructing the model or excluding 3 days before and after instead of one week marks as significant exactly the same years as one week before-after analysis does.

To conclude, the counts modelling using Poisson and negative binomial regression around each anniversary indicates that more people are dying in the week after the birthday and the all years range model indicates that more people than could be expected are dying within one week before birthday. There is no postponement phenomenon exhibited by the data.

4.2 Modelling the hazard function

Another method to verify the postponement phenomenon is to detect if the hazard function h , which is the death rate at each age, has a dip prior to birthdays. It is common to use the Kaplan–Meier estimator [3] of the hazard function, but in the absence of censoring (all the people in the dataset have died in 2015), it is equivalent to the following estimators:

$$\hat{h}(k) = \frac{d_k}{\hat{S}(k)}, \quad k = 0, 1, \dots \quad \text{with}$$

$$\hat{S}(k) = \frac{1}{n} \sum_{i=k+1} d_i,$$

where d_i is the number of deaths at the age of $i = 0, 1, \dots$ days and $n = \sum_{i=0} d_i$.

The estimated hazard \hat{h} and its inverse $1/\hat{h}$ for the people who died at the age of 50 to 80 years are presented in Figure 3. The observed parabola-like curves correspond to the same number of deaths recorded during one day: if, say, $d_k = d_{k+1} = d$, then $\hat{h}_k = d/(d + \hat{S}_{k+2})$ and $\hat{h}_{k+1} = d/\hat{S}_{k+2}$ are lying on a

parabola and $|\hat{h}_k - \hat{h}_{k+1}| = O(1/\hat{S}_{k+2}^2)$. It would be approximately a parabola if the days with the same number of counts are not consecutive. If $d_k = k$ and $d_{k+1} = d + 1$, then $|\hat{h}_{k+1} - \hat{h}_k|$ is of order $1/\hat{S}_{k+2}$ and \hat{h}_{k+1} is lying on the next curve above corresponding to $d + 1$ deaths during a day.

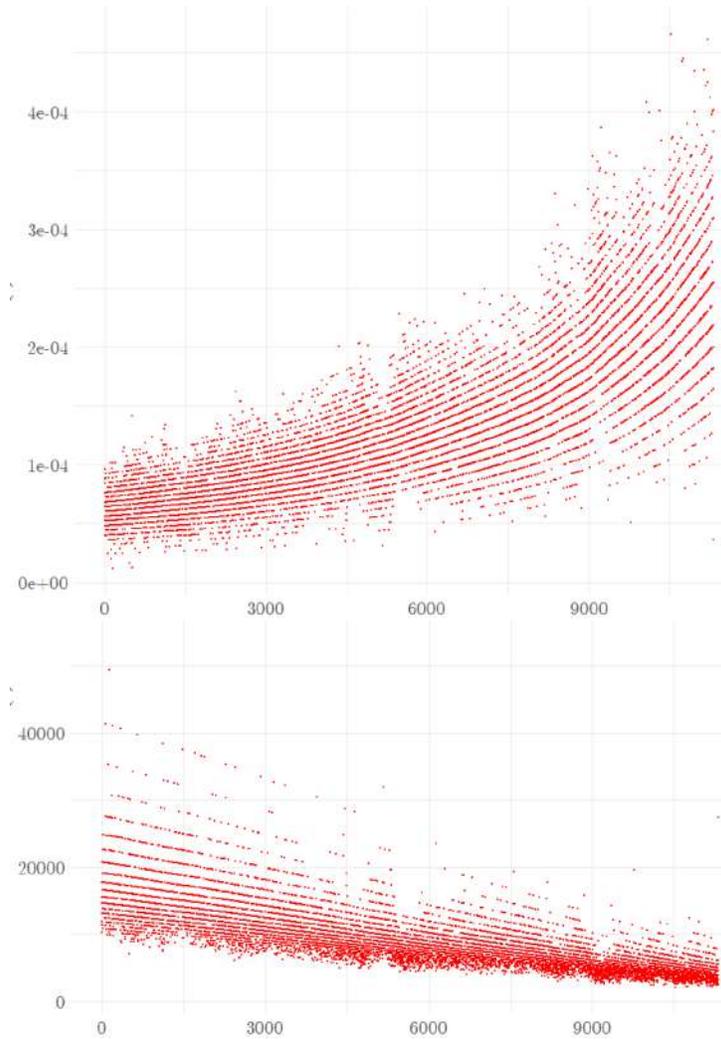


Fig. 3. Estimated hazard function h and $1/h$ for the days after the 50th anniversary.

The shape of the lower graph suggests that $1/h$ is linear, i.e. the lifetime remaining after attaining 50th anniversary conforms to a Generalised Pareto distribution (3). To make it more certain, we smoothed $1/\hat{h}$ over one week intervals and computed 95% confidence intervals using bootstrap method. The resulting graph, shown on Figure 4, prompts to use a linear model to fit $1/\hat{h}$.

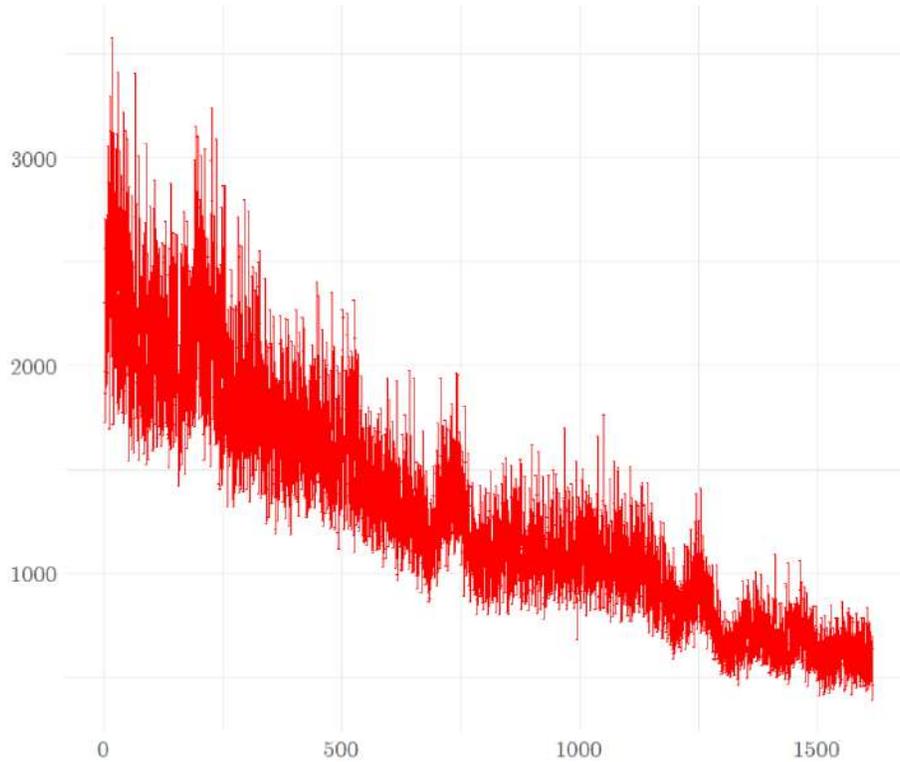


Fig. 4. Graph of $1/h$ smoothed over one week interval with vertically drawn 95% bootstrap estimated confidence intervals. The timescale – weeks.

Another approach consists in using the generalised Gamma-family linear model to fit h directly since the inverse transformation is its natural link. Once the linear model is accepted, it would mean the generalised Pareto distribution for the residual lifetime over 50 years of age.

Arguing as in the Poisson model fitting above, even if the linear model for $1/h$ might be a too strong assumption for the whole range of deaths at 50-80 years age, it should be a good approximation for shorter periods. So we analyse separately deaths at the ages $y - 1, y + 1$, where y ranges from 50 to 80. We fit both a linear model to $1/h$ and a generalised gamma-family linear model to h *excluding* the deaths 2 weeks before and 2 weeks after the anniversaries. We decided that 2 weeks is more appropriate here than 1 week in the Poisson model analysis above to account for dependence of \hat{h}_i 's: close days i have a similar denominators \hat{S}_i 's.

As for the Poisson regression in the precious section, we proceed with examining the residuals produced by the predicted values of these models for omitted data before and after birthday and compare their means to the rest of the residuals. We obtained the following results presented in Figure 5. As we see that according to linear modeling of $1/h$, there are 13 significant peaks before birthday and 12 after birthday that is, according to the binomial test

Age	Before BD		After BD		Age	Before BD		After BD	
	LM	GLM	LM	GLM		LM	GLM	LM	GLM
50	Peak	-	-	-	65	-	-	Peak	-
51	Peak	-	-	-	66	Peak	-	-	-
52	Peak	Peak	Peak	Peak	67	-	-	-	-
53	Peak	-	-	-	68	-	-	Peak	Peak
54	-	-	-	-	69	Peak	Peak	-	-
55	-	-	Peak	-	70	-	-	-	-
56	Peak	-	-	-	71	-	-	Peak	Peak
57	-	-	-	Dip	72	Peak	Peak	-	Dip
58	-	Dip	-	-	73	Dip	Dip	Peak	-
59	-	-	-	-	74	Peak	-	-	Dip
60	-	-	Peak	-	75	-	-	Peak	Peak
61	Peak	Peak	-	-	76	Peak	-	-	-
62	-	-	-	Dip	77	-	Dip	Peak	Peak
63	-	Dip	Peak	Peak	78	Peak	Peak	-	-
64	Peak	Peak	-	-	79	-	-	Peak	-

Fig. 5. 95%-significant periods before and after birthday according to a linear model (LM) fitting $1/h$ and according to a generalised linear model (GLM) fitting h for people died from 49 to 80 years old.

with p -value of 5%, a highly significant figure for 31 anniversaries considered. Notably, the same years which were identified by the Poisson regression in the previous section are also appearing here. The generalised linear model is more conservative, but still gives 6 significant peaks before and 6 after birthday. The observed 4 dips is not a significant number when working with 95% confidence level.

5 Conclusion and critique

We verified validity of the death postponement theory on a dataset consisting of over 204 thousand records of deaths from natural causes in South Africa in 2015. We took person's birthday as an important event a person tries to survive to, and studied the death rates just before and after the birthday. If a death postponement phenomenon exists, it would manifest itself in a lower than expected death rate just before birthday (a dip), perhaps, followed by a higher rate (peak) at the birthday and just after it. We employed a Poisson model to estimate the number of deaths at each day to verify this hypothesis and also analysed the hazard function by linear and generalised linear models. We found *no confirmation* of the postponement phenomenon for this dataset, for the birthday as an important event and for the timescale of a few days. On the contrary, the death rate is found to be higher both before and soon after the birthday. This might be explained by the stress of expectations for the birthday and/or an earlier start of celebrations with associated departure from the recommended regime.

As a byproduct of our analysis, we found that the models of mortality based only on the age explained at most 60% of the variation in the death counts. Thus the age is an important, but not the only determining factor of the death hazard.

Acknowledgement

SZ thanks his Master students: Natalia Andreeva, Furqan Farooqi, Ingrid Ingemarsson and Lucas Lazaroo, for data mining and the pictures used in Sec. 4.2.

References

1. V. Ajdacic-Gross, D. Knöpfli, K. Landolt, M. Gostynski, S.T. Engelter, P.A. Lyrer, F. Gutzwiller, and W. Rössler. Death has a preference for birthdays – an analysis of death time series. *Annals of Epidemiology*, 22:603–606, 2012.
2. DataFirst. South Africa – mortality and causes of death 2015 study description, 2017. <https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/610/study-description>.
3. E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *J. Amer. Stat. Assoc.*, 53:457–481, 1958.
4. H. Leerhoff and U. Rockmann. Death won’t wait. Cancer deaths around birthdays and religious holidays. In *Proceedings of the 59th World Statistics Congress of the International Statistical Institute*, Hong Kong, 25-30 August 2013. ISI. <http://2013.isiproceedings.org/Files/CPS204-P41-S.pdf>.
5. D.P. Phillips and E.W. King. Death takes a holiday: mortality surrounding major social occasions. *The Lancet*, 332(8613):728–732, 1988.
6. D.P. Phillips and D.G. Smith. Postponement of death until symbolically meaningful occasions. *J. American Medical Association*, 263:1947–1951, 1990.
7. D.P. Phillips, C.A. van Voorhees, and T.E. Ruth. The birthday: lifeline or deadline? *Psychosomatic medicine*, 54:532–542, 1992.
8. G. Smith. Asian-American deaths near the Harvest Moon festival. *Psychosomatic medicine*, 66:378–381, 2004.