A Weibull Accelerated Life Model Approach to the Analysis of Factors Affecting Transition Time from Graduation to First Job

Aglaia G. Kalamatianou, Dimitrios Kalamaras and Foteini Kougioumoutzaki

Department of Sociology, Panteion University of Social and Political Science,
Athens, Greece
Email: AglaiaKalamatianou@gmail.com
Email: dkalam@panteion.gr
Email: fkouyoum@hotmail.com

Abstract: Duration models have been proved useful to the study of the distribution of the transition time from university to work which is usually influenced by various factors. In this paper, an analysis of such determinants is performed by means of a parametric Weibull Accelerated Life Model, for the case of the transition from university to first job of Greek social science graduates. Results confirm that individual characteristics such as gender, parental educational level, field of studies, post graduate studies and foreign language skills are among those which have a significant impact to the transition time.
Keywords: Survival analysis, Accelerated Life Models, Weibull baseline, Transition to first job, Social science graduates

1 Introduction

The study of the transition process from higher education to the labour market is many researchers' major concern in recent years, mainly due to its link with the issue of unemployment of graduates which followed the increase of participation in tertiary education noticeable since the late 1990s in almost all developed countries. As stated by Salas-Velasco (2006), '*…the transition from university to work acquires quite relevance because new cohorts of recent graduates may have greater difficulties to find a first 'good' job in comparison to the smooth transitions experienced by young people in the past. Nowadays in many countries, the transition period is becoming longer and transition patterns are becoming less defined and less certain than they once were* OECD (1999, 2000). A long period of unemployment after graduation has multiple social, economical, political and personal effects (loss of human capital, waste of money for individuals and society, graduates' frustration being unemployed) making the thorough examination of the transition process of great importance.

A study of this transition process includes mainly the estimation of its duration and the assessment of the factors involved. Relevant studies reveal that this duration may be influenced by several factors, e.g. type of degree, duration of studies, postgraduate studies, previous working experience, job search intensity, gender, family background etc., see, Salas-Velasco (2007), Biggeri *et al.* (2001), among others. For handling corresponding questions descriptive statistical methods and survival analysis methods have been implemented; a short and clear review on where such methodologies have been used for analyzing data concerning transition process

from university to work is given in Salas-Velasco (2007) and in Sciulli and Signorelli (2011).

Briefly speaking, survival analysis, also known as event history analysis, duration analysis or hazard modelling, is an appropriate statistical methodology for studying duration time till the occurrence of one or more types of event, where the event may not occur (during a predefined follow up period) to a proportion of sample units, leading to censored duration times for the corresponding units. Given this background, the effects of explanatory variables on the duration time can also be examined on the basis of suitable developed models.

In the context of studying transition process from university to work survival analysis is applied, although yet not extensively. In the main, when the case concerns the time duration of the transition process, the survival function can provide estimations of the probability an individual remains unemployed up to a point of time t after graduation while the hazard function (hazard risk, hazard rate) can give the (instantaneous) risk an individual finds a job within a very small interval of time (given that the individual has not already found a job). Both functions can be estimated using non parametric methods (Kaplan-Meier procedure is a usual one) and parametric methods. For the evaluation of factors related to the duration of transition, continuous-time or discrete-time regression type event history models can be used. Among these, proportional hazards models and in particular the semiparametric Cox proportional hazards model, as well as accelerated life models are the most commonly used. In the above context,

> *the aim of this paper is to report on the factors* (*explanatory variables*)*, concerning mainly graduates' characteristics, that affect transition time from graduation to working life of Greek social science graduates by means of accelerated life models*.

The focus on this particular group of graduates is due to fact that they face severe employment difficulties compared to other Greek graduates, Kalamatianou & Kougioumoutzaki (2012a). The obtained results would make us more aware of reaching conclusions concerning these graduates' challenges, while they could also be integrated into policies aiming to facilitate transition to working life. This paper adds to a series of others, (Kalamatianou & McClean (2003); Kalamatianou & Kougioumoutzaki (2012a,b); Kougioumoutzaki, F. & A. Kalamatianou (2012a,b), concerning employment and educational aspects of Greek social sciences graduates as they are represented of the graduates of Panteion University of Social and Political Sciences. In addition, the paper is a contribution to the study of the transition from university to the labour market by using a parametric Weibull accelerated life model which is not widely used in such applications.

The remainder of this paper is set out as follows: Section 2 describes the data, the variables and some descriptives. Section 3 presents theoretical results concerning the accelerated life models as well as results concerning the choice of the appropriate model for our data representation. Section 4 is devoted to the implementation of the chosen model to the data and the results. The final section 5 draws some conclusions.

2 Data, Variables and Descriptive Evidence

For the present analysis a sample of 683 graduates from the population of graduates of 1988-2000 of the Panteion University of Social and Political Sciences of Athens were selected and interviewed. The data was collected using a questionnaire filled in by graduates' answers to corresponding questions and by elements retrieved from the official student records. The whole procedure of data collection was carried out during 2005-07. Given the difficulties in reaching graduates plus the fact that there were no previous corresponding statistical results, it is necessary to note that we examined all graduates of the most recent years of the abovementioned period ie., 1997-2000, while for the rest we considered all graduates every three years of graduation (ie., years 1988, 1991, 1994), in order to facilitate sampling. This way of collecting data also facilitates comparisons between more recent and past years' graduates. Since the object of interest is the time to obtain the first job the sample consists of all graduates (of the aforementioned years) who we managed to conduct and accepted to participate in the survey and also at the time of the interview were either unemployed searching for a job or employed and their job was the first after graduation. In this context given that the size of the corresponding total population of interest is unknown our sample can be considered as representative to give the trends on the issue under study.

To response to the aim of this paper, described in the previous section, necessary information was collected and recorded for each graduate as follows:

Dependent variable:

- *Time Duration from graduation to first job*, measured (in months) as the time between graduation date and a) date of obtaining the first job, for the employed graduates (complete observations) b) interview date, for the unemployed graduates searching for a job (right censored observations). Note, that in this analysis we did not distinguish between part-time and full-time jobs or temporary and permanent jobs or related and non related to corresponding university studies jobs. Also note, the above duration may include spells where graduates were not searching for job as well as duration of military service for male graduates.

Explanatory variables:

Here we considered variables -assumed time independent-, corresponding to graduates' characteristics (mainly demographic and educational) regarded as affecting transition to work in similar studies, -see Biggeri et al.,(2001), Salas-Velasco (2007), Sciulli and Signorelli (2011) among others- and also apply to our case.

- *Gender* (1, male; 0, female)
- *Parental educational level* (1, at least one of the parents had a secondary school or higher level degree; 0, otherwise)

VARIABLES RELATED TO FIRST DEGREE:

- *Field of studies* (1, the graduate has an economic oriented degree, ie. -graduates from the departments of Public Administration, Economic & Regional Development; 0, the graduate has a degree oriented to social or political sciences –ie., all other departments of Panteion university (Sociology, Political Science and History, International and European

Studies, Communication, Media and Culture, Psychology, Social Anthropology, Social Policy)

- *Practical training during studies* (1, the graduate was trained; 0, otherwise)
- *Final degree mark*, recorded on the basis of three dummy variables as follows:
  (a) *Degree1* (1, for "Well"; 0, otherwise)
  (b) *Degree2* (1, for "Very Well"; 0, otherwise)
  (c) *Degree3* (1, for "Excellent"; 0, otherwise)
- *Age at graduation*, recorded on the basis of three dummy variables as follows:
  (a) *G_Age22* (1, graduation took place on the 22nd year of age; 0, otherwise)
  (b) *G_Age26* (1, graduation took place between 23rd and 26th year of age; 0, otherwise)
  (c) *G_Age27* (1, graduation took place after the 27th year of age; 0, otherwise)
- *Duration of studies*, recorded on the basis of three dummy variables as follows:
  (a) *GMRT* (1, graduation took place at the minimum required time for graduation –institutional time ie., four academic years; 0, otherwise)
  (b) *GDP2* (1, graduation took place 2 years after the minimum required time; 0, otherwise)
  (c) *GDDelay* (1, graduation took place at least 3 years after the minimum required time; 0, otherwise)
- *Work during studies* (1, the graduate had a job during studies; 0, otherwise)

OTHER QUALIFICATIONS:

- *Post graduate studies* (1, the graduate performed post graduate studies; 0, otherwise).
- *Foreign language skills* (1, yes; 0, no)
- *Computer skills* (1, yes; 0, no)
- *Seminars* (1, the graduate followed extra educational seminars; 0, otherwise)
- *Channels* -Search intensity for finding a job- (1, the graduate used simultaneously more than one ways to find the first job; 0, otherwise)

Given the type of the data concerning the dependent variable (censored data) it follows that survival analysis methods should be applied to respond to our purpose. For the total sample the mean and median duration of the transition time are estimated to about 33 and 25 months accordingly while for recent graduates 30 and 26 months and for past years' graduates 38 and 23 months. Combined graphical information (omitted here due to lack of space) regarding the empirical hazard functions, -based on Kaplan and Meier (1958) method-, for the different values of the explanatory variables, confirms that PH assumption is not satisfied in several cases, pointing thus to the accelerated life models for the evaluation of factors related to the transition time, than the proportional hazards models.

3. The Accelerated life Model

General results: The accelerated life models belong to the survival models that are appropriate for incorporating explanatory variables (factors or covariates) that influence survival time of the items under consideration and they can be used to determine which covariate has the most significant impact on the survival or to compare survival patterns for different groups of items. On the lines of Leemis (1995), Kalbfleisch and Prentice (1980), Box-Steffensmeier and Jones (2007), assuming a vector $\mathbf{z} = (z_1, z_2, \mathrm{K}, z_q)$ of q explanatory variables the survivor function for duration time T in the accelerated life model is

$$S(t) = \big(S(t,\mathbf{z})\big) = S_0\big(t\psi(\mathbf{z})\big) \qquad t \geq 0,$$

where S_0 is a baseline model and $\psi(\mathbf{z})$ is a link function satisfying $\psi(\mathbf{0}) = 1$ and $\psi(\mathbf{z}) > 0$ for all $\mathbf{z}$ implying $S(t) \equiv S_0(t)$ when $\mathbf{z} = \mathbf{0}$. In this context the hazard rate is $h(t) = \psi(\mathbf{z})h_0\big(t\psi(\mathbf{z})\big) \qquad t \geq 0$

The most general case is to let $\psi(\mathbf{z})$ be any function of the explanatory variables. A popular option, also assumed in this paper's application, is the log linear form $\psi(\mathbf{z}) = e^{\boldsymbol{\beta}'\mathbf{z}}$, where $\boldsymbol{\beta}$ is a $q \times 1$ vector of regression coefficients corresponding to the q explanatory variables. Here if $\psi(\mathbf{z}) > 1$ the explanatory variables accelerate the rate at which the item move through time with respect to the baseline case or decelerate it when $\psi(\mathbf{z}) < 1$.

Concerning the baseline model two distributions, the exponential and the Weibull are among the most commonly used. In the case of Weibull, there are two parameters for the baseline distribution, a scale parameter λ and a shape parameter κ so that $h_0(t) = \lambda\kappa t^{\kappa-1}$. Then the hazard function for the accelerated life model is

$$h(t,\mathbf{z},\kappa,\boldsymbol{\beta}) = \kappa e^{\kappa\boldsymbol{\beta}'\mathbf{z}} t^{\kappa-1} \qquad t \geq 0 \qquad (1)$$

where, $\mathbf{z} = (z_0, z_1, z_2, \mathrm{K}\ , z_q)'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \mathrm{K}\ , \beta_q)'$; here an additional covariate $z_0 = 1$ is defined for all items to facilitate the parameter $\lambda = e^{\beta_0 z_0}$ to be included in the vector of regression coefficients. The log likelihood function that guides to parameter estimation of model (1) is given as follows

$$\log L(\kappa,\boldsymbol{\beta}) = r\log\kappa + \sum_{i\in U}\boldsymbol{\beta}'\mathbf{z}_i + (\kappa-1)\sum_{i\in U}\log x_i - \sum_{i=1}^{n} x_i^{\kappa} e^{\kappa\boldsymbol{\beta}'\mathbf{z}_i}$$

where n is the number of items in the test, $i = 1,2\mathrm{K}\ ,n$, $x_i = \min\{t_i, c_i\}$ is either the failure time $t_1, t_2, ..., t_n$ or the censoring time $c_1, c_2, ...c_n$ and U is the set of the indexes of the uncensored observations.

For $\kappa = 1$ the above Weibull model reduces to the exponential one.

Results concerning our application - The appropriate baseline distribution: Assuming the accelerated life time model to be an appropriate representation of our data we examined farther the suitability of the exponential and the Weibull model for the baseline distribution. Figure 1 displays the empirical hazard rates –according to Kaplan and Meier (1958) method- for the total sample as wells as for the subgroups of recent and past years' graduates. Clearly, they are time depended and in particular monotonic decreasing indicating that the likelihood for a graduate to get his/her first job decreases as duration time spent to unemployment increases. These graphical results plus those presented in Figures 2a,b point to the Weibull model rather than exponential as appropriate to describe the baseline distribution in our application. However the suitability of the Weibull model examined further by estimating the parameters κ and λ. If $\kappa = 1$ exponential is the appropriate model otherwise it is Weibull. One technique for this estimation, described by Leemis

(1995), p. 216, utilizes the Newton – Raphson procedure –see also Keats and Lawrence (1997).

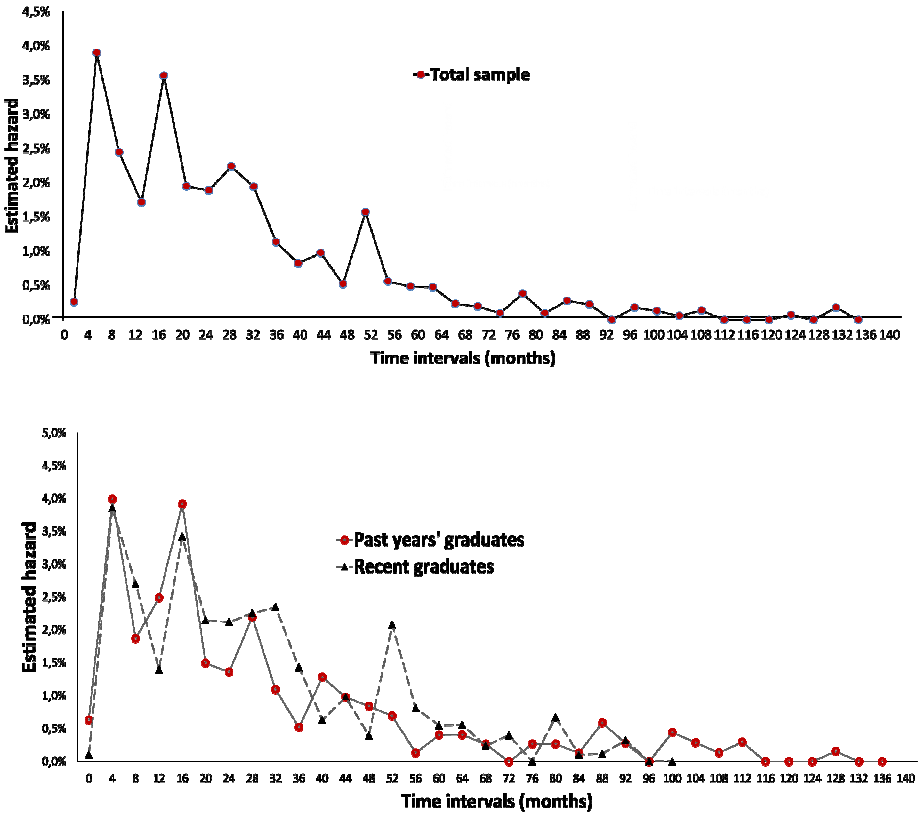


Fig. 1: Empirical hazard function

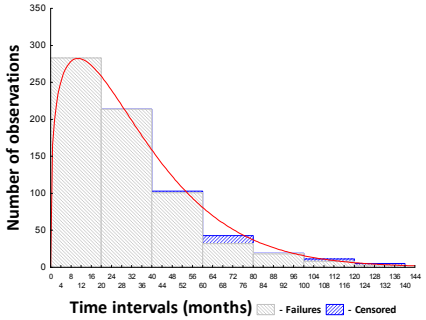


Fig.2a. Failure times and censored observations fitting on Weibull distribution (total sample: shape =1.28, scale =35.70)

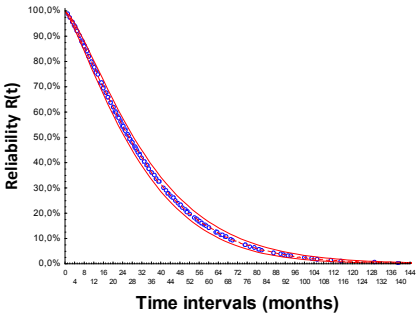


Fig. 2b. Weibull and reliability failure time for ML Parameter estimates (total sample: shape =1.28, scale =35.70)

Table 1 gives ML estimates for the parameters κ and λ their 95% C.I., applying the abovementioned technique to the total sample of our data as well as to the subsamples of recent and past years' graduates. Obviously in all cases, the value $\kappa = 1$ (shape parameter) is not included in the corresponding intervals indicating that Weibull distribution is suitable for describing the baseline distributions.

Table 1: M.L. Estimates of the Weibull parameters and 95% confidence intervals

	Total sample			***Past years' graduates***			***Recent graduates***		
Parameters	***Est.***	***L. lim.***	***U. lim.***	***Est.***	***L. lim.***	***U. lim.***	***Est.***	***L. lim.***	***U. lim.***
Shape $(\hat{\kappa})$	1.28	1.20	1.35	1.21	1.09	1.34	1.47	1.37	1.58
Scale $(\hat{\lambda})$	35.70	33.54	38.01	36.52	32.36	41.22	33.81	31.69	36.07

4. Implementation of the model and results

Under model (1), Table 2 reports the estimates of the regression coefficients corresponding to explanatory variables of the time duration from graduation to first job (described in section 2), for the total sample of social science graduates analyzed in this paper, as well as for the subsamples of recent and past years' graduates. Based on these results, one could argue the followings: Considering the total sample, gender, parental educational level, field of studies, postgraduate studies, and foreign language skills appear to have a significant impact on the duration of finding the first job. In particular, men's transition time to the first job is expected 22% higher than that of women, $100\times(1.22-1)=22$, other things being equal. This favorable outcome for women which is inconsistent with that shown in several relevant studies (see, Salas-Velasco (2007), Biggeri *et al.* (2001) among others), but closer to that shown in some others (see, Galero and Galeiro (2011), Sciulli and Signorelli (2011), Allen and Van Der Velden (2007), among others) that however don't focus on social science graduates and adopt different statistical methodologies, must be evaluated taking into consideration that here transition time for male graduates may possibly include duration of their military service. Quoting the negative sign of $\hat{\beta}_2$, it follows that graduates, whose at least one parent has a secondary level education degree or higher, have sooner transition time $(13\%, 100\times(1-0.87)=13)$ than those with parents of lower level education. Similarly, the transition time of graduates with a degree oriented to economics (see section 2) is found 11% faster $(100\times(1-0.89)=11)$, than that of graduates with a degree in the rest disciplines. Foreign language skills also cause acceleration to the transition period that is, such skilled graduates have a 21%, $100\times(1-0.79)=21$, faster transition than the unskilled. Finally, graduates with a postgraduate degree are expected to have longer transition period than those without such degree. In fact, the duration of the former is estimated 35% higher than that of the latter, $100\times(1.35-1)=35.$ This peculiar result may reflect the fact that corresponding graduates may have chosen to continue

their studies instead of starting searching for a job just after completing first degree, investing in a probable better job in the future.

Table 2. Determinants of finding the first job: *Weibull accelerated life model estimates*

Explanatory variables	***Coef.***	***Total sample***		***Past years' grad.***		***Recent grad.***	
		$\boldsymbol{\beta}$	e^{β}	$\boldsymbol{\beta}$	e^{β}	$\boldsymbol{\beta}$	e^{β}
Constant		3.49	32.88	3.19	24.32	3.78	43.76
Gender	$\hat{\beta}_1$	***0.20****	1.22	0.03	1.03	***0.27****	1.32
Parental educational level	$\hat{\beta}_2$	***-0.13****	0.87	-0.02	0.98	-0.10	0.90
Field of studies (Econ.Vs.Soc)	$\hat{\beta}_3$	***-0.12****	0.89	-0.13	0.87	-0.09	0.91
Practical training during studies	$\hat{\beta}_4$	-0.17	0.85	-0.12	0.89	-0.12	0.88
Post Graduate Studies	$\hat{\beta}_5$	***0.30****	1.35	***0.41****	1.50	***0.32****	1.37
Final degree mark							
Degree1(well)	$\hat{\beta}_6$	0.23	1.26	***0.68****	1.97	0.17	0.84
Degree2 (very Well)	$\hat{\beta}_7$	0.25	1.28	***0.66****	1.93	0.00	1.00
Degree3 (excellent)	$\hat{\beta}_8$	0.00	1.00	0.00	1.00	0.00	1.00
Age at graduation							
G_Age22	$\hat{\beta}_9$	0.07	1.08	0.26	1.29	-0.02	0.98
G_Age26	$\hat{\beta}_{10}$	-0.07	0.93	-0.11	0.90	-0.01	0.99
G_Age27	$\hat{\beta}_{11}$	0.00	1.00	0.00	1.00	0.00	1.00
Duration of studies							
GMRT	$\hat{\beta}_{12}$	0.07	1.07	0.13	1.14	-0.07	0.93
GDP2	$\hat{\beta}_{13}$	0.10	1.11	0.25	1.28	-0.05	0.95
GDDelay	$\hat{\beta}_{14}$	0.00	1.00	0.00	1.00	0.00	1.00
Work during studies	$\hat{\beta}_{15}$	-0.09	0.91	-0.10	0.91	-0.08	0.92
Foreign language skills	$\hat{\beta}_{16}$	***-0.24****	0.79	***-0.51****	0.60	-0.11	0.90
Computer skills	$\hat{\beta}_{17}$	0.01	1.01	0.46	1.59	0.00	1.00
Seminars	$\hat{\beta}_{18}$	-0.03	0.97	-0.16	0.86	0.02	1.03
Channels	$\hat{\beta}_{19}$	-0.22	0.80	-0.26	0.77	-0.32	0.73
Scale parameter (κ)		0.67		0.75		0.60	
Log(κ)		***-0.39****		***-0.29****		***-0.51****	
Loglik(model)		-2423.50		-828.60		-1571.00	
Loglik(intercept only)		-2458.20		-841.30		-1601.80	
LR χ^2		69.50		25.44		61.47	
df		20.00		20.00		20.00	
Prob > χ^2		0.00		0.02		0.00	

* Statistically significant at 5% level

Considering the subgroups of recent and past years' graduates, the results show similar trends to the above, for almost all cases of explanatory variables however statistical significance appears in fewer cases. As regards past years' graduates the significant factors are: Post graduate studies, foreign languages skills and final degree mark showing a shorter transition period for those with a well and very well mark degree. Considering recent graduates, gender and postgraduate studies appear to have a significant impact to the transition period in the way indicated before. However these differences between recent and past years' graduates as well as comparing them to the total sample must be examined further taking into consideration additional information on the conditions of the labour market and on studies' characteristics for safer evaluation.

5 Conclusion

Transition from higher education to labour market, a serious problem in many countries nowadays, is a complicated process involving various factors. In this paper attempt is made to analyse the time duration from graduation to first job and the involved factors, concerning Greek social science graduates as they are represented by graduates of Panteion University of Social and Political Science, Athens, Greece, which is the only Greek university offering degrees exclusively on those sciences. The main points emerged from the analysis are the following: The likelihood of graduates to get their first job decreases as time spent on unemployment increases. This result remains stable between recent and past years' graduates. The groups of graduates with a higher hazard of exit unemployment are those of women, whose at least one parent has a secondary level education or higher degree, those who have an economic oriented degree, those who have not a postgraduate degree –possibly reflecting the fact that they may choose to continue their studies instead of starting searching for a job-, and those who have foreign language skills. However differences are noticed considering recent and past years' graduates showing to the necessity of additional information on the conditions of the labour market and on studies' characteristics for safer evaluation. Finally, as regards the methodological approach used in the present analysis, Kaplan-Meier method for estimating the empirical Hazard rates and parametric accelerated life models with Weibull as the baseline distribution for the assessment of the related factors, proved effective to undertake our application.

References

1. Allen, J. and Van Der Velden, Transitions from higher education to work, in *Careers of University Graduates*, (U. Teichler, ed.), Springer, (2007)

2. Biggeri, L., Bini, M. and Grilli, L., The transition from university to work: a multilevel approach to the analysis of time to obtain the first job, *Journal of the Royal Statistical Society* (Series A), **164**, 293–305, (2001).
3. Box-Steffensmeier, J. M., and Jones, B. S., *Event History Analysis. A Guide for Social Scientists*, Cambridge University Press, Cambridge, (2007).
4. Galero, A. and Galeiro, A., Understanding the transition to work for first degree university graduates in Portugal, *Nota & Economicas*, Junho, 44-61, (2011)
5. Kalamatianou, A. G. and Kougioumoutzaki, F., Employment Status and Job-Studies Relevance of Social Science Graduates: The Experience from a Greek Public University, *International Journal of Economic Sciences and Applied Research* (*IJECAR*), **5** (1), (2012a, forthcoming).
6. Kalamatianou, A. G. and Kougioumoutzaki, F., Aspects of employment of Panteion University's graduates: job satisfaction, in *Sociology and Social Transformations in Modern Greece* (provisional title) (A. Moissidis, D. Papadopoulou & G. Petraki, eds.), Gutemberg, Athens, (2012b, forthcoming).
7. Kalamatianou, A. G. & McClean, S., The Perpetual Student: Modelling Duration of Undergraduate Studies Based on Lifetime-Type Educational Data, *Lifetime Data Analysis*, **9**, 311-330, (2003).
8. Kalbfleisch, J. D., and Prentice, R. L., *The Statistical Analysis of Failure Time Data*, John Wiley & Sons, Inc., New York, (1980).
9. Kaplan, E. L., and Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457-481, (1958).
10. Keats, J. B. & Lawrence, F. P., Weibull maximum likelihood parameter estimates with censored data, *J.of Quality Technology*, Vol. **29**, No. 1, 105-110, (1997).
11. Kougioumoutzaki, F. and Kalamatianou, A. G., Sociology Graduates: Studies Satisfaction and Factors Involved, *Proceedings, 3rd Annual Congress of the Hellenic Sociological Society,* Athens (2012a, forthcoming).
12. Kougioumoutzaki, F. and Kalamatianou, A. G., Studies Satisfaction of Social Sciences' Graduates: A Descriptive Analysis, *Proceedings, Scientific Association of Social Policy, 4th International Scientific Symposium, The role of Social Policy Today: Approaches and Challenges*, Athens (2012b, forthcoming).
13. Leemis, L., M., *Reliability Probabilistic Models and Statistical Methods*, Prentice-Hall, Inc, New Jersey, (1995).
14. OECD (1999), Preparing Youth for the 21st Century: The transition from Education to the Labour Market. Paris: OECD.
15. OECD (2000), From Initial Education to Working Life: making Transitions Work. Paris: OECD.
16. Salas-Velasko,.M., The transition from higher education to employment in Europe: the analysis of the time to obtain the first job. *Higher Education* **54**, 333–360, (2007)
17. Sciulli, D., and Signorelli, M., University-to-work transitions: an empirical analysis on Perugia graduates, *Eur. J. of Higher Education.*, **1** (1), 39-65, (2011).

Modeling tax evasion of the self-employed in Hungary

Ildikó Ritzlné Kazimir, Gergely Horváth, Johanna Giczi

Hungarian Central Statistical Office Departement of National Accounts
Budapest, Hungary
Hungarian Central Statistical Office Departement of Methodology
Budapest, Hungary
Email: gergely.horvath@ksh.hu

Abstract:
Nowadays various kinds of hidden economy cause difficulties for statistical offices in the estimation of GDP. Among them tax evasion plays crucial role in many countries, not only for the tax authority, but for anybody who wants to measure the real output of an economy. In this paper we describe a new method modeling tax evasion of sole proprietors. The main target of the model is to give a solid base for the computation of some components of the GDP.

Keywords: Stochastic modeling, Modeling, K nearest neighbour method, Official statistics

1 Introduction

The non-observed economy includes several kinds of economic activities, which require different estimation methods due to their diversity. The non-observed economic activities can be classified into seven categories by the recommendation of EUROSTAT. (Eurostat's Tabular Approach) This recommendation considers the reasons why a certain economic activity might fall into the non-observed category. One of the categories is the non-observed production which comes from the deliberate misreporting of administrative data to minimize taxes and contributions. The gross value added of non-observed economy has to be made available exhaustive for sectors and industries in the GDP calculation. (GNI Inventory 2011) Therefore it has become necessary to develop a model for the subsector of sole proprietors which can estimate the non-observed gross value added because of the deliberate misreporting (i.e. tax evasion). (Murai et al 2011) Our model makes an attempt at determining the non-observed gross value added of sole proprietors considering value added tax (VAT) evasion on the basis of the individual tax audits. There are several steps in the model. In the first step the population of value added tax evaders is estimated, and then the undeclared VAT is calculated for this population with the help of the additional undeclared VAT, which is the result of the tax audits. Finally the undeclared gross value added is estimated. Our paper describes the options of the first step, and examines, how the population of VAT-evaders might be estimated as precisely as possible.

2. Data sources

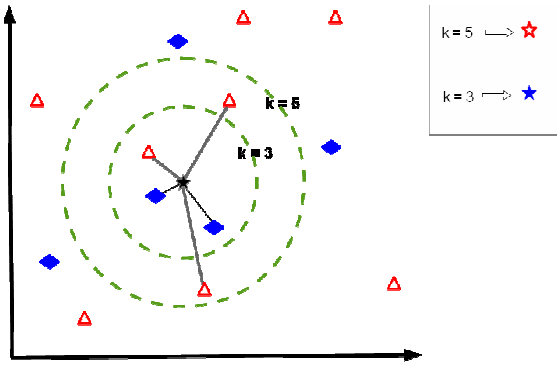
In this research we have two main sources of data. The first one is our own database which is regularly used in our researches, and VAT tax audit database of

the sole proprietors. This latter one belongs to the tax authority; we are only allowed to use it specially for this purpose. In HCSO (Hungarian Central Statistical Office) the main source of data is the business register, which contains the main attributes of any enterprises and apart from this there are regularly transferred various tax data. We have to mention some problems concerning tax audit data of the private entrepreneurs. The dataset transferred to the HCSO contains only the final tax audits. It is not known how many businesses were selected for supervision, and this problem is related with an other one: there are a lot more data from the earlier years (e.g. for 2006 there are more than for 2009) than in recent ones. In the 2006-2009 data, the year of the tax audit is denoted, but the additional sum of the tax (VAT) and the fine can not be split into different years. Even the method by which the enterprises are selected for supervision is unknown for us. Considering all these attributes of the tax data we have to admit that the tax audit data cannot be taken for a set of probability sample. In this research, we only use tax audit data to identify the tax evading enterprises. The self-entrepreneurs who were compelled to pay some added VAT were identified (by us) as tax evaders, and at the same time these enterprises were supposed to have tax evading as their permanent behaviour towards tax liability. With this simplification, we have succeeded in getting a much bigger dataset.
We also have to add a last remark on the use of these datasets. There is not much data about the hidden economy and tax evasion. The best data sources of this topic are based on special researches, carried out by research institutes, not by official statistical agencies. (Allingham et al 1972, Becker 1974, Hasseldine et al. 1991, Slemrod 1985) The task described in this paper is a part of a bigger process, the computation of GDP. We have to use data of the official statistical system, but not that of the researches of an institute which is not part of the official statistical system.

3. Modeling tax evasion

Before presenting the model, it is important to clearly state the target of this modeling effort. Our target is to produce estimates of the hidden GVA of the self proprietors, not to find every possible tax-evader entrepreneur. It could be the goal of a tax authority. If the results of the model fit into the production of the GDP, we achieved our goals, even knowing some weaknesses of the model. (Szabó ez al. 2009)
To describe the attributes of the tax evader enterprises only tax audits can be used, but the inherency in this field cannot be applied to the full multitude. In order to illustrate the tax evading behaviour of the enterprises we have tried to set up an explanatory model on the basis of the data from the supervised entrepreneurs. We do not expect very good results, because, on the one hand, the starting data have some inconsistencies, and on the other hand, there are a number of other variables that would have been very important to use, but they were not available for us. In order to test the effects of the variables a logistic regression was applied. In this way the co-efficiency of the variables can be seen. (Hajdu 2003)

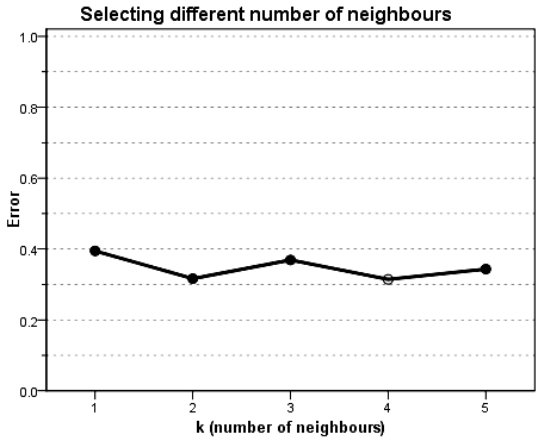
Applying some more variables the illustrative force of the model could be improved a bit (thanks to the more properly chosen threshold value).With the help of this explanatory model we have reached our most important goal, i.e. we have succeeded in finding an explanation to the phenomenon of tax evasion. Using this explanatory model we have succeeded in getting a solid base for the research. We have empirical results that these factors have significant influence on tax evasion, therefore we can use these variables.
As we have already mentioned, the traditional modeling techniques cannot be applied in our case, because the tax audit data cannot be used as a set of probability sample, and correct weighting could not be carried out. That is why we searched for other modeling methods, and at last we have found a simple machine learning algorithm, the k nearest neighbor method (kNN). (Bodon 2010)



1. figure The kNN method

The kNN method is one of the most simple 'sluggish processes' of machine learning methods. (Han et al, 2004) It estimates the properties of unknown individuals by the properties of known individuals in the entire population. Similarity is based on distance – the nearer the more akin. To determine distances, each case is placed in a multidimensional space as defined by its known properties; to classify an individual, matching neighbours are sought.
Classifiable individuals receive a category which is most frequent among their neighbours, in the space of classified cases (Figure 1).
Building our kNN model, we started to use some default parameters and we let the software choose the best combination of variables (we included all the variables chosen by the logistic regression model) and the number of neighbors (1 to 7 neighbors). The goal of the model was to achieve the highest rate of correct classification. But the results were not suitable enough in several aspects. Using the default parameters the method selected only two variables (if-full-time and grouped NACE category) and 4 neighbors, and predicted all cases to be tax evader. If this model had been accepted we would have overestimated the number of tax-

evaders among the sole proprietors, because each of them might have been classified as tax evader. The percent of the law-abiding enterprises and the tax evaders in the tax-audit data was: 29% law-abiding and 71% tax-evader. In this case the method classified every non tax evader into the tax evaders, but all the tax evaders were classified correctly, the error ratio was only 29%. But this result is not usable for us. On the one hand we do not think that every sole proprietor are tax evader, and on the other we were not able to explain such data.
After discussing these results, we had to find some other starting point.
Before proceeding we remark the following on the effects of the number of selected neighbors according to their parity. In case of even number of neighbors, in that special case when the frequency of the two categories is equal, the prediction will be the more frequent category in the training population (tax evader). In this study our target variable has only two categories, so it is obvious that in the case of even number of neighbours, the classification always will be tax evader.
Considering the results of various alternatives in this model, we have found a characteristic difference.



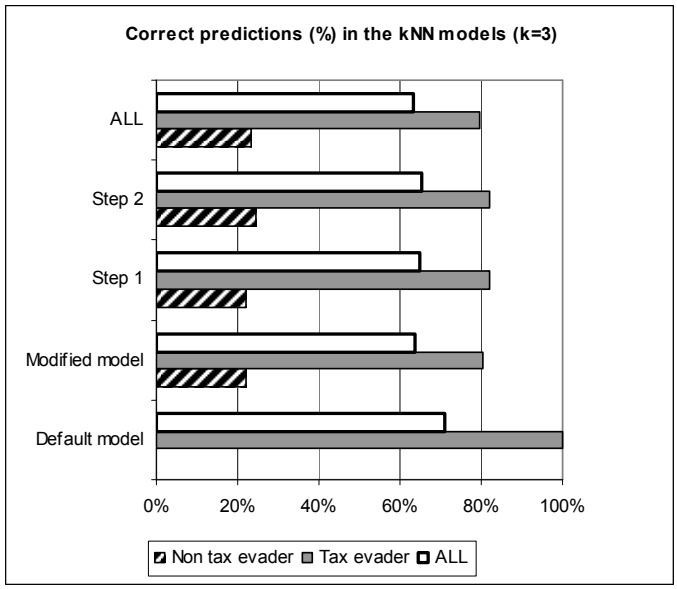
2. figure Error ratio with different number of neighbours

Choosing more neighbors, we got lower overall error ratio but increased error ratio of the law-abiding sole proprietors. If we had chosen even number of neighbors, the overall error ratio was lower, but that of the law-abiding sole proprietors increased. The graph of the error ratios draws a “W” shape, where the peaks represent the odd, and the bottoms represent the even numbers. That is why we decided, to use an odd number of neighbors.

Table 1 Variables in the kNN models

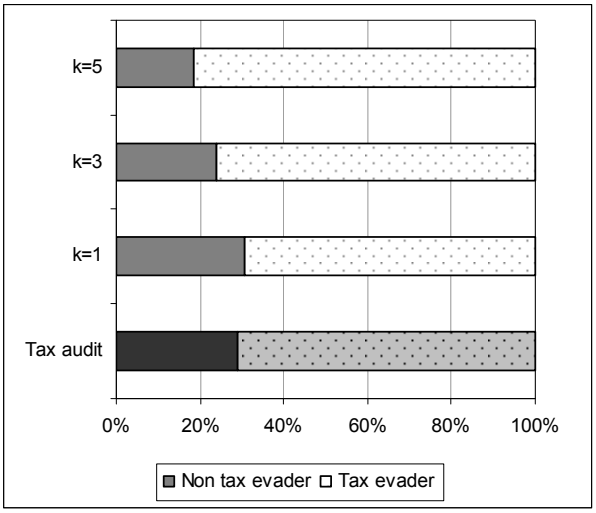
				Variables
Default			IF-FT	If full-time employed
			NACE	NACE (grouped)
	Modified		TREV	Total revenue
			REGION	Region
		Step 1	PSR	Possibility of sales without receipt
			AVGSAL	Avarage salary
		Step 2 (optimal model)	R48	ratio of cost of raw material by total cost
			SUMSAL	Sum of salaries
Not included in the kNN model				
			RTS-TR	Ratio of total supply by total revenue
			TCOST	Total costs
			OCOST	Other costs

We redesigned the model and used three fixed features — if-full-time and grouped NACE category and the full total revenue of the enterprise as a variable to summarize several attributes of an economic organization — and fixed three neighbors. The model was broadened through a step-by-step approach. In the first step we used the three variables mentioned above. This model served as the basis. In the next step, the three variables were fixed, and the software was allowed to select more (from the remaining ones). The selected variables have been added to the fixed features for the next step. We repeated this loop twice. Because of the nature of the data, we weren't able to find an absolute best solution therefore an optimal solution was targeted. The main goal was to maximize the correct predictions for each (tax evader, non tax evader) category at the same time. Comparing the models we considered two indicators: the overall error ratio and the error ratio of the smaller group (non tax-evaders). For us, the latter was a bit more important, as the correct prediction for this category was much lower. After two steps we have found an optimal solution. As a result the correct prediction was: 24.32%, 82.04%, 65.38% (non tax-evader, tax evader, overall), and for that we used 8 features in the optimal model.
As a control experiment, we used a full model where all the selected variables were fixed, but the results showed slightly higher error rates than in the optimal model detailed above.



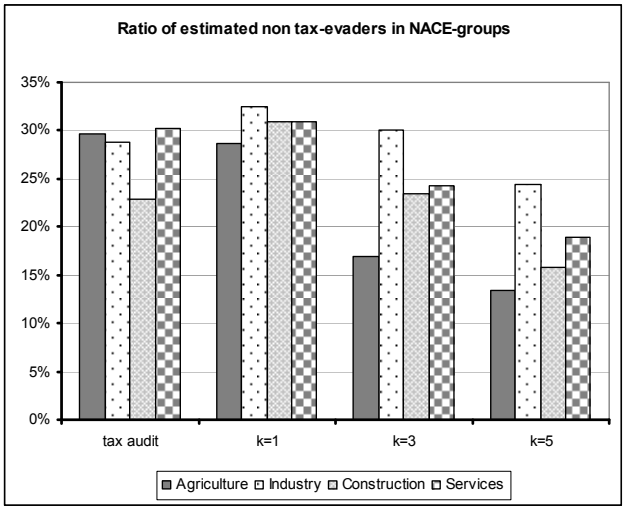
3. figure

After finding an acceptable combination of variables, we compared the results of the optimal model for one, three and five neighbors. We have already mentioned earlier that in this study the correct prediction of both categories are of great importance, because we assume, there are non tax evader entrepreneurs, and though the dataset of the tax audits is not a probability sample, it is quite incredible that the tax authority found all the non tax evaders. So we rejected any models without non tax evaders.
The results showed that models utilizing one or three neighbors give better results. With one neighbor, we can achieve nearly 30% correct predictions among the non tax evaders, and about 70% in the other category. Increasing the number of neighbours, the number of correct predictions among non tax evaders decrease, with five neighbours it is lower more than 10% (comparing the one neighbour).



4. figure Estimated ratio of tax evaders with 1, 3, 5 neighbours

The estimated proportion of tax evaders slightly increase as we use more neighbours (but this is expected). The maximum share of non tax evaders (in the case of one neighbour) is a bit higher than in the training data.



5. figure

After describing the effect of the chosen neighbours to the portion of tax-evaders, we can take a look at the distribution of the phenomenon among the grouped NACE categories.
There is a big difference in the pattern of tax evasion. In the case of one neighbour, the pattern is quite similar to that one in the training data, except the construction, in which the estimated ratio of tax-evasion is lower than in the tax audit data. All other groups seem to be quite equal. In contrary, the pattern differs if three or five neighbours are chosen. And these patterns look similar (except the decreasing ratio of non tax evaders).

4. Conclusions

We have attained doubtful results, and in this paper we do not suggest using any certain one. All the acceptable models have some advantages and some trade-offs. The most obvious weakness of the model is its quite inaccuracy of the classification. But the ultimate target (of this model) is not the prediction of the tax evaders, but estimating the sum of the hidden gross value added of the sole proprietors.
It seems that using one or three neighbours gives the most acceptable results. In these models there is a credible ratio of tax evaders, so it can serve as a solid base for the estimation of the undeclared VAT and the non-observed gross value added of sole proprietors.

References

1. Allingham, M. G. – Sandmo, A (1972): Income Tax Evasion: A Theoretical Analysis, Journal of Public Economics 1: North-Holland Publishing Company, http://old.nhh.no/sam/stabssem/2004/sandmo.pdf pp 323-338. old.
2. Becker, G. (1974): Crime and Punishment: An Economic Approach, in: Becker, G. – Landes, W. M.:Essays in the Economics of Crime and Punishment,. UMI, http://www.nber.org/books/beck74-1 pp: 1-54
3. Dr. Bodon Ferenc (2010): Adatbányászati algoritmusok (Data mining algorithm) http://www.cs.bme.hu/~bodon/magyar/adatbanyaszat/tanulmany/adatbanyaszat.pdf
4. Eurostat's Tabular Approach to Exhaustiveness Guidelines, http://www.unescap.org/stat/isie/reference-materials/National-Accounts/Eurostat-Guidelines-Tabular-Approach.pdf
5. Han, J – Kamber, M. (2004): Adatbányászat Koncepciók és technikák (Data Mining. Concepts and Techniques) Budapest: Panem Kiadó
6. Hajdu O. (2003): Többváltozós statisztikai számítások, Statisztikai módszerek a társadalmi és gazdasági elemzésekben, Budapest: Központi Statisztikai Hivatal
7. Hasseldine, D. J. – Bebbington, K. J. (1991): Blending economic deterrence and fiscal psychology models in the design of responses to tax evasion: The New Zealand experience, Journal of Economic Psychology, 12. North-Holland Publishing Company pp 299-324
8. Johanna Giczi - Gergely Horváth - Ildikó Ritzlné Kazimir (forthcoming 2012): Value Added Tax Evasion of Sole Proprietors, Acta Oeconomica
9. KSH (Központi Statisztikai Hivatal) (2011): GNI Inventory of Hungary Version 2.2. Budapest. http://portal.ksh.hu/pls/ksh/docs/eng/xftp/modsz/gni_inventory_ver_2_2.pdf
10. Murai B. – Ritzlné Kazimir I. (2011): A nem megfigyelt gazdaság mérésének lehetőségei (Measuring the Non-Observed Economy) Statisztikai szemle, 89(5): 501-522.
11. Slemrod, J. (1985): An Empirical Test for Tax Evasion, The Review of Economic and Statistics, 67(2): 232-238. http://www.jstor.org/stable/1924722 .
12. Szabó A. – Gulyás L. – Tóth I. J. (2009): Az adócsalás elterjedtségének változása – becslések a TAXSIM ágensalapú adócsalás-szimulátor segítségével. (Changes in the incidence of tax evasion – estimations with TAXSIM, an agent-based tax evasion simulator) In. Semjén A. – Tóth I. J. (eds.): Rejtett gazdaság. Be nem jelentett foglalkoztatás és jövedelemeltitkolás – kormányzati lépések és a gazdasági szereplők válaszai (Undeclared employment and non-reported income:Government policies and the reaction of economic agents) KTI Könyvek Budapest: MTA Közgazdaságtudományi Intézet
http://www.wargo.hu/tij/news/ktik11_rejtett_gazdasag.pdf

MORTALITY TABLE PROBLEMS IN THE LIFE INSURANCE SECTOR: AN ADVISORY ALTERNATIVE SOLUTION FOR TURKEY

Erdem KIRKBEŞOĞLU*
İsmet KOÇ†

The main objective of this study is to develop an updated mortality table that will be used in the insurance sector in Turkey and test the functionality of this table by comparing it with American CSO 1980, which is a life table commonly used by life insurance companies in Turkey. The main contribution of the study is to put forth to what extent the premium and compensation calculated in the insurance sector in Turkey will differentiate from premium and compensation calculated from the mortality data of Turkey. The Orphanhood method, an indirect method, was used to generate this table, and in this way, estimated death rates, life expectations and insurance premiums were compared with American CSO 1980. The results indicate that the life expectancies of CSO 1980 show a similarity with the average of male and female life expectancies of Turkey. On the other hand, 15 years death rates of CSO 1980 have higher values than male and female death rates of Turkey until age 50. Additionally, it can be observed that the endowment life insurance, the most preferred life insurance type, whose premiums are calculated from the Turkish male and female mortality table has the lowest premiums for the most suitable starting ages (25-40) for insurance. This situation shows that insurance companies have increased the premiums in crucial age groups, leading to unjustified losses for the insured.

Key words: mortality table, life insurance, orphanhood method

INTRODUCTION

The insurance sector, which is among the sectors where the trust factor is of great importance in social and economical life, has to have the purpose of ensuring stability and order both for its employees in its internal structure and for the clients outside. In addition this stability and order will provide the application of the right plan, policies and strategies and require regulations, laws and rules both on an organizational and industrial basis. Therefore, regulations that will enhance trust in society in the sector are required for the insurance companies, which are considered as a warranty of lives and properties of the individuals, to be stable and efficient. In Turkey, governments intervening legally and politically in sectors such as accounting and banking could not manage to realize the necessary regulations for the insurance sector until the end of 2000. More clearly, the government and its supervisory body in the insurance sector, Turkish Undersecretariat of Treasury, were insufficient for a long time in realizing the legal regulations necessary both for insurance companies and insurers. Although the "Insurance Law", effective in 2007, provided several legal regulations for the insurance firms, it does not cover a regulation towards developing the current mortality tables, which is the most basic problem, especially for life insurance firms (Akmut, 1980; Özsoy, 1970; Kırkbeşoğlu, 2007). Not having updated mortality tables developed by using the reliable mortality data or information of the insured in Turkey will cause a decrease in trust in and demand for the sector, and miscalculations of premiums; developing reliable mortality tables requires reliable mortality data.

Mortality data have a great importance in the insurance sector, social security system and various institutions' provident funds, where actuarial balances are calculated. Besides this, it makes it difficult to use mortality ratios and commutation (insurance premium- income) tables

* Instructor, Başkent University, Faculty of Commercial Sciences
† Prof. Dr., Hacettepe University Institute of Population Studies

by insurance firms. None of the insurance firms in Turkey uses a table developed by using Turkish mortality data. Instead, mortality tables from abroad reflecting mortality ratios of different countries are preferred. Therefore, there is a two-way problem at this point. The first is that reliable mortality data is needed in Turkey and alternative studies are necessary to realize this. The second is that mortality ratios of foreign countries that are used particularly by the insurance companies do not represent the real mortality profile of Turkey, leading to premiums that are either too high or too low. Therefore, reliable studies that will provide alternative mortality data and mortality tables developed using these data are needed. These tables facilitate the calculation of premiums and the compensation to be paid to them. Hence, it is important, and of critical importance for insurance companies, that these tables should include the most reliable mortality ratios possible for ensuring actuarial balance – more clearly, for premiums to be charge at a rate equal to the compensation paid. Therefore, mortality data to be used by life insurance companies and hence mortality tables need to involve reliable mortality data reflecting the demographical characteristics of that country.

The main aim of this study is to develop an updated mortality table to be used in the insurance sector in Turkey and test the functionality of this table by comparing it with foreign – origin – mortality tables used by the insurance sector. The Orphanhood method, an indirect method, is used to generate this table (Zlotnik and Hill, 1981). In this way, the insurance premiums will be calculated and an insurance premium comparison will be made with American CSO 1980, which is a life table commonly used by life insurance companies in Turkey. Hence, the main contribution of the study is to put forth to what extent premiums and compensation calculated in the insurance sector in Turkey will differentiate from premiums and compensation calculated based on the mortality data of Turkey.

MORTALITY DATA IN LIFE INSURANCES

Mortality tables are defined as tables that list how many people will die and how many will survive in a year at every age as foreseen from the results obtained based on life and death statistics developed by keeping a certain group of the population under observation (Life Insurances Regulation, 1996). Mortality tables are used by particularly actuaries, demographist or people dealing with public health to make various studies on issues like migration, fertility, population estimations, orphanhood or widowing, life expectation, marriage and working life. In the insurance sector where mortality tables are of great importance (especially in life insurances), it is significant how reliable these tables are in calculating premiums and compensation. In addition, mathematical allowances and shares that insurance companies have are closely associated with the reliability of these tables.

Even though mortality tables are so important in life insurance, today there is no mortality table in Turkish the insurance sector developed from mortality data belonging to Turkey. On the other hand, although limited in number, there are a few studies where mortality tables were developed from Turkish mortality data starting from the year 1950. However, none of these tables were taken into consideration for use in the insurance sector by the Undersecretariat of Treasury Directorate General of Insurance Department.

Mortality tables generated at certain intervals in developed countries over years are used both in these countries and in developing or less developed countries in various fields. That there is no table generated from the Turkish data in the insurance sector in our country so far shows that the same situation is valid for Turkey. Developed countries, such as the USA, Germany, France, Switzerland and the UK are using mortality tables that include the mortality ratios of

their own citizens. Such tables are far from representing the mortality ratios of Turkish society. These countries' current and previous mortality ratios have no similarities with the mortality ratios of Turkey today; particularly as the developments in medicine, economical improvements and developments in life standards have led to the extension of the average lifetime in Turkey as well.

In particular, retirement insurances having been sold based on the mortality tables of the 1950s will lead to companies facing a great amount of risk in the future. Since tables designed with old data involve high mortality ratios, it is assumed that retirement income is paid in a shorter time and it does not take into account that the insured may live longer than expected (Osmançavuşoğlu, 1999: 52). Therefore, insurance companies have the risk of encountering financial problems. Also, high mortality ratios cause the premiums to be higher than they should actually be in insurances covering death indemnity. Indeed, foreign tables, as they involve higher mortality ratios, assume that the insured may die at an earlier age.

Consequently, that the mortality tables used in the insurance sector in a country put forth the exact premiums and compensation depends on two factors. The first of these is the requirement that mortality data by which the mortality table is prepared should belong to that country and the second is that the mortality data is reliable. Thus, the reliability of mortality data seems to be of great importance.

The rest of the study deals with calculating mortality ratios according to ages with the Orphanhood Method, which is often preferred in countries where it is not possible to obtain direct mortality ratios. Turkish mortality and insurance premium tables will be developed by using the mortality ratios obtained.

DATA SOURCES AND METHODOLOGY

Data Sources: 1998 and 2003 Turkish Demographic and Health Survey

In Turkey, mortality data can be taken from the several different sources. These are Central Record Systems (MERNIS), Turkish Statistical Institute mortality statistics, burial records, Turkish Ministry of Health records and census. However, taking reliable mortality data related to Turkey includes several problems. One of the problems is different registration procedures in urban and rural areas. Muhtar is responsible for reporting deaths to the DDP in rural areas if medical doctors or medical personnel are not present there owing to the fact that different persons can be responsible for reporting deaths in rural areas. Additionally, muhtars may not have a good level of education, so reliable information cannot be obtained for the most of rural areas. The other problem is that the MERNIS is based on the family ledgers system. Migration movements do not affect the MERNIS records if persons do not change their family ledgers. Therefore, although any person lives permanently in one district, he/she can appear in another district in the family ledger (Wunsch and Hancıoğlu, 1995). In addition, death records provided by health personnel or muhtar are not filled out; therefore, many people still appear as alive who died several decades ago. The other mortality data source is the national population census. Although the last national population census, 2000 General Population Census, contained mortality questions for adults, the results have not yet been published. Therefore, it can be seen that obtaining information pertaining to mortality data is problematic.

Because of these reasons, in this study, 2003 Turkish Demographic and Health Survey will be used as a main data and 1998 Turkish Demographic and Health Survey will be used as a supplementary data to make mortality estimations and to construct mortality and commutation tables. TDHS-2003 is a national representative sample survey which is designed to provide information about infant and child mortality, levels and trends on fertility, family planning and maternal and child health. The results of the survey are presented as an urban and rural residence at the national level for the five regions in the country. Besides, the TDHS-2003 sample also provides information about 12 geographical regions (NUTS1) to analyze and compare with European Country within the context of Turkey's move to join the European Union.

The primary objective of the TDHS-2003 is to provide information about socioeconomic characteristics of households and women, fertility, mortality, marriage patterns, family planning, maternal and child health, nutritional status of women and children, and reproductive health. Detailed information is provided from a sample of ever-married women in reproductive ages (15-49). The TDHS-2003 was designed to produce information in the field of demography and health which cannot collect from other sources. Especially it has been useful to obtain direct and indirect factors that determine levels and trends infant and childhood mortality.

The sample design and sample size of the TDHS-2003 make possible to analyze for whole Turkey, urban and rural areas and five demographic regions of the country which are West, East, Central, South and North regions. The TDHS-2003 sample also provides information about 12 geographical regions (NUTS1) for a limited number of variables which were adopted at the second half of the year 2002 to analyze and compare with European Country within the context of Turkey's move to join the European Union (HUIPS, 2004). In the selection of the TDHS-2003 sample, the approach of a weighted, multistage, stratified cluster sampling method was used. The distribution of the target sample was based on the results of the 2000 Turkey General Population Census. The target size of the TDHS-2003 was set as 13,160 households which is 30 percent larger than of the TDHS-1998.

On the other hand, 1998 Turkey Demographic and Health Survey (TDHS-1998) will be used as a secondary data in this study. TDHS-1998 shows similarities with TDHS-2003. The results of the survey are presented as an urban and rural residence at the national level for the five regions. But 12 geographical regions (NUTS 1) were not included in this survey. In the selection of the TDHS-1998 sample, the approach of a weighted, multistage, stratified cluster sampling method was used. The target size of the TDHS-1998 was set as 9,970 households and 8,596 of them were applied in the survey. The interview was completed successfully with 94 percent of 8,596 households. 9,468 women were determined to interviewing. 8,576 of women were identified as eligible. In the half of the selected households, husbands of currently married eligible women who were present in the household on the nigh before the interview or who usually lived in that particular household were eligible for the husband survey. (HUIPS, 1999).

Data Quality

The important indicator to determine the data quality from the surveys is missing on key variable. Among births in the 15 year preceding the TDHS-2003, 4 percent are missing information on year of birth. Information on age at death is missing for just 1 percent of these births. Marriage age or date was not taken from less than 1 percent of ever-married women. Height or weight measurements are missing for nearly 8 percent of the children under age 5.

Compared with data from TDHS-1998, these figures show that the missing information is very limited in the survey (HUIPS, 2004).

One of the most powerful interviewing tools is the birth history for collecting information on births and deaths. Complete information on birth dates were collected almost all births occurring since 2001 and nearly 94 percent of births during 1998-2000. On the other hand it can be said that the complete information on birth dates were collected accurately from the TDHS-1998 data. The TDHS-2003 and the TDHS-1998 data appear to be good quality with respect to the completeness of the information collected on dates of birth and ages at death. A detail inspection of the birth history data from the TDHS-2003 and the TDHS-1998 point out age heaping at death was also minimal. One of the commonly observed failures of the sampling surveys is age heaping at death to 6, 12, 18 and 24 months. Therefore, infant deaths may record as a child dates because of the respondents heaping the age at death to 12 months or interviewers recording ages of death as "1 year". This situation causes calculated bias rates. These biases are not seen much more during 10 years before the 2003 survey (HUIPS, 2004).

The other evaluation related to reliability of birth history is calculation of sex ratios at birth for all five births. This ratio is expected to fluctuate around 105 male births per 100 female births. The overall sex ratio is calculated 104.4 for all births in the birth history for the TDHS-2003 and 105.6 for the TDHS-1998. On the other hand, sex ratio of death for age 0-1 interval is 1.08 for ten years preceding the TDHS-2003. This shows that male mortality rate is higher than female mortality rate for infant mortality, which is expected situation for early age mortality. When investigating the sampling errors of TDHS-2003 and 1998, infant mortality rate (IMR) during 4 years before the survey is calculated 28,767 for the TDHS-2003 and 42.702 for the TDHS-1998. Standard error of the IMR is calculated 2.914 for the TDHS-2003 and its confidence limits are calculated between 22.938 and 34,596 which refer 11.658 years interval in 95 percent confidence limit. Standard error of the IMR is calculated 4.659 for the TDHS-1998 and its confidence limits are calculated between 33.384 and 52.020. This numeral refers to 18,636 years interval which is larger interval than the TDHS-2003 in 95 percent confidence limit.

METHODOLOGY

Orphanhood Method

Besides the increase in population studies in recent years, the requirements of basic data for such studies are not available or too deficient in many countries. The inadequacy of registration statistics and difficulty in collecting accurate data directly causes that indirect methods of analysis, particularly those based on orphanhood, represent an important source of adult mortality estimates in developing countries (Timæus, 1990). Especially the lacks of accurate vital registration and censuses have led demographers to project indirect methods to estimate basic demographic parameters from incomplete or inaccurate data. Although, these methods can not be considered as substitutes for the mortality measures obtained from complete and accurate vital registration system, in the absence of such data, these would provide reasonable basis for demographic analyses and other purposes (Sivamurthy and Seetharam, 1980).

The plausible information about orphanhood for the measurement of adult mortality was first explored by Henry in 1960. He developed the ideas of Lotka, who had considered the reverse problem, estimating orphanhood from data on mortality (Timæus, 1990). He argued that if the information on infant and child mortality can be obtained by asking the mothers about the

survival of their children, why can not be done the same for adult mortality by asking the children about the survival of their parents? (Blacker, 1977). Henry's idea was taken up by Brass, who established an equation relating the female probability of surviving from age 25 to age 25 + n to proportions of respondent in two contiguous five year age groups whose mother was still alive at the time of the interview (United Nations, 1983). Final version of Brass's methods is published by Brass and Hill (1973). Later, Hill and Blacker, working under the Brass' guidance, developed an equation to estimate adult male mortality from proportions of persons with fathers alive.

Orphanhood approach has obvious advantages. The questions, "Is your mother alive?" and "Is your father alive?" are simple and easy to answer. These questions does not include date of the death or reference period and can be answered by outspokenly "Yes" or "No"; they takes little place on the questionnaires and the results are simple to code, punch and tabulate. Furthermore, every additional question inserted in survey has an additional cost, but the cost of the orphanhood questions are minimal (Blacker, 1977). These questions applied in the survey questionnaire of African countries in 1960's with the opinion of Brass and Hill. The quality of the answers for these questions would be better than the direct questions about the deaths 12 months preceding the survey.

For orphanhood method, it is needed the mean age at maternity for females and mean age at conception for males and the proportion of not orphaned respondent by five year age groups form males and females. Estimates of adult mortality taken from close relatives that represent averages of the mortality experienced during the relatives were exposed to the risk of dying (United Nations, 1983). Respondents' mothers must have been alive at the birth of the respondents. Thus, the period of exposure to die is the age of respondent (Coşkun, 2002). But there is a difference for the paternal orphanhood. Although, the risk of dying for mothers has started with the birth of their child, the risk of dying for fathers has started with conception of the child. For this reason, approximately nine moths should be incorporated in calculations.

Brass established an equation relating to the female probability of surviving from age 25 to 25+n. This equation has the form:

$$l_{(25+n)} / l_{(25)} = W_{(n)} \cdot S_{(n-5)} + (1-W_{(n)}) \cdot S_{(n)}$$

Where, S(n) is the proportion of respondents aged from n to n+4 with mother alive. W(n) is a weighting factor which is employed to make allowance for typical age patterns of fertility and mortality. Brass and Hill (1973) are calculated the set of W(n) values from the African standard mortality pattern and model fertility schedules of fixed shape but variable age locations. The weighting factor depends on value of n and the mean age of childbearing.

Later, Hill and Trussell (1977) have performed another estimation by using regression coefficients. This equation has the form:

$$l_{(25+n)} / l_{(25)} = a_{(n)} + b_{(n)} \cdot M + c_{(n)} \cdot S_{(n-5)}$$

a(n), b(n) and c(n) are the regression coefficient which is calculated with four different Coale-Demeny mortality patterns as standards. These regression coefficients were calculated only for females. Thus, estimating of adult mortality is not available with this equation for males (Coşkun, 2002).

In case of calculating adult male mortality, the value is replaced by the values 32.5 or 37.5. Because the men are usually older than women at the birth of their children. So the survivorship probabilities are estimated from the following two formulas:

$$l_{(35+n)} / l_{(32,5)} = W_{(n)} \,.\, S_{(n-5)} + (1-W_{(n)}) \,.\, S_{(n)}$$

$$l_{(40+n)} / l_{(37,5)} = W_{(n)} \,.\, S_{(n-5)} + (1-W_{(n)}) \,.\, S_{(n)}$$

If the mean age of paternity is less than 36, the first equation is used; on the other hand if the mean age is greater than 36, the second equation is used.

On the other hand, the reference date of the mortality level is calculated. The formula is changed separately for females and males. For females, the formula is:

$$t_{(n)} = n\,(1.0 - u_{(n)}) / 2.0$$

where,

$$u_{(n)} = 0.3333.\ln({}_{10}S_{(n-5)}) + Z_{(M+n)} + 0.0037.(27-M)$$

The value of $Z_{(M+n)}$ is provided by interpolation by using value of the standards function table.

For males, the formula is:

$$t_{(n)} = (n + 0.75) \,.\, (1.0 - u_{(n)}) / 2.0$$

Where,

$$u_{(n)} = 0.3333.\ln({}_{10}S_{(n-5)}) + Z_{(M+n)} + 0.0037.(27-M+0.75)$$

In this case ${}_{10}S_{(n-5)}$ represents the proportion of respondents in the age group from n-5 to n+4 with mother alive; on the other hand, n is the mid-point of the 10 year age group being considered.

According to Zlotnik and Hill (1981), if information of orphanhood has been collected by two censuses or surveys between the five or ten years periods, the hypothetical intersurvey cohort of respondents can be calculated. Shortly, “hypothetical cohort” method is occurred to estimate the adult mortality between the censuses or surveys. This method has additionally advantages than the Brass’ method. Firstly time reference time of the estimates stay between two surveys. The other advantage of synthetic cohorts is that deaths are reported fully during the intersurvey and omission of more distant deaths will have no effect on the result. Therefore, synthetic cohort data on the survival of the parents are less vulnerable than lifetime data to the adoption effect that is underreporting of orphanhood by respondents (Timæus, 1991).

The proportions of not orphaned among a hypothetical intersurvey cohort of respondents are calculated from this formula:

$$S_{(n,s)} = S_{(n,2)} \qquad \text{for } n < T;$$

$$S_{(n,s)} = S_{(n-T,s)} \,.\, S_{(n,2)} \,/\, S_{(n-T,1)} \qquad \text{for } n \geq T.$$

In this formula, "T" is defined by length of interval between the surveys. S(n,1) is the proportion of persons in the age group from n to n+4 whose mother alive at the time of the first survey and S(n , 2) is at the second survey with same rationale. Proportions of not orphaned among a hypothetical intersurvey cohort is constructed with these formulas and also the other calculations are the same with Brass's orphanhood method.

In Turkey, the questions about survival of the parents were not asked in any censuses but it has been applied in the surveys. In years 1973, and in 1974-75 TPS, 1978 TFS, 1983 TPHS, 1988 TPHS, 1993 TDHS, 1998 TDHS and 2003 TDH surveys.

Life Insurance Premium Calculations

The main responsibility of the insured person in life insurances is to pay the premium to the insurer. The insurance premium calculated thanks to the commutation or annuity tables are defined as net premium or risk premium. Like all organizations, the insurance companies also have some expenditure. Therefore, the manufacturing and administrative costs of the company should be added to the net premium. Furthermore, the reserved amounts put aside for unexpected losses should be added to the net premium as well. All these additions are called loading, and by adding the loading to the net premium the gross premium is measured.

Gross premium = Net premium + Loading (The expenditures of the insurer + the commission of the insurer)

The net premium, which is in the gross premium, is based on the fundamental of equality in the responsibilities of insurer and insured person (Moralı, 1997). That is, the net premiums calculated from the commutation tables according to the types of life insurance are thought to be equal to the present value of the life insurance guarantee that is taken by insurer.

The present value of the net premiums = the present value of the insurer's responsibility.

The insured can make their premium payments both in cash and in installments. For both payment styles the insurers have to make different calculations. The premium payments are calculated differently according to the life insurance types. But in this study, it is assumed that the premium payments are made at the moment of policy starting for net single premium payments.

In this study, the comparisons for insurance premiums will be done over endowment life insurance, which is the most sold type of insurance by the insurance companies in Turkey. As known, endowment life insurance is a type of life insurance where the amount in the policy is paid to the relatives of the insured in case s/he dies in a certain period of time or to the insured in case he survives during this period of time. As it can be understood from the definition, endowment life insurance covers the outcome of both death and survival. The formula for the net amount of single premium that the insured pays is given below (Moralı, 1997: 131),

$$A_{x:\overline{n}|} = \sum_{t=0}^{n-1} v^{n+1}{}_{n|}q_x + v^n {}_n p_x$$

In this formula, notations represent:

(A) present value of the life insurance premium
(p) the probability of living,

(q) the probability of dying,
(v) present value of 1 TL at the end of the year. (V= 1 / (1+i)),
(n) life insurance period.

Net premium refers to the net risk except all expenses and profit margin.

CONSTRUCTION OF MORTALITY TABLES FOR TURKEY

Background of Mortality Tables in Turkey

The death rates of any country should represent the population characteristics of that country to use in mortality tables for the calculation of premiums and reserves in the insurance sector. Although mortality tables have gained importance in the insurance sector, there are no mortality tables constructed from Turkish data, and what is more, this deficiency has been loosely compensated by foreign mortality tables. In Turkey, sixteen different mortality tables had been used until 1978 (Duransoy, 1993). These tables were

1. American CSO 1953-1958
2. English Hm
3. English Om
4. French AF
5. French PF 1952-1956
6. French PM
7. French PMF 1931
8. French RF
9. German Abel
10. German General ADST
11. Swiss SM 1901-1910
12. Swiss SM 1921-1930
13. Swiss SM 1941-1950
14. Swiss SM 1948-1953
15. Swiss SM 1958-1963
16. Swiss TG 1960

But the surplus of these tables has caused the blocking of standardization use in the life insurance sector. Therefore, a lot of research has been applied to decrease these tables. On 5th May 1978, with Article 14(950.1/7) 12665 and Clause 28 of law no. 7397, the Insurance Inspection Committee (Sigorta Murakabe Kurulu) of the Turkish Treasury Undersecretariat reduced the mortality table total from sixteen to just three (Duransoy, 1993). These are;

1. Swiss Male (SM) Mortality Table (1948-1953)
2. Allgemeine Deutsche Sterbetafel Tabelle (ADST) General German Mortality Table (1949-1951)
   and, the
3. Commissioners Standard Ordinary (CSO) Mortality Table (1953-1958).

Later, the Undersecretariat added the American Commissioners Standard Ordinary (CSO) 1980 mortality table to include gender differences with Article B.02.1HM.O.SGM.0.2.1.2 /Gen/99/62885 on 31st July 2001 (Ataman, 2002).

The Undersecretariat of the Turkish Treasury has undertaken the control and inspection of mortality tables for life insurance companies in Turkey. The mortality tables used to determine risk premiums have been chosen with great care. The Undersecretariat of the Turkish Treasury can determine the mortality and the morbidity tables according to the results of the portfolio of life insurance companies or the data of the Turkish Statistics Institute. The Under-secretaryship can request the table and the results of the portfolio, which are

constructed by the Association of Insurance and Reinsurance Companies of Turkey, from the insurance companies at the end of the year. The Association of the Insurance and Reinsurance Companies of Turkey consolidate these tables and then they are sent to The Undersecretariat of the Turkish Treasury and insurance companies (Life Insurance Regulation, 1996).

If insurance companies have ten years or more mortality experience, these companies can construct their own mortality tables or they can replace current mortality tables with their own mortality experience. But these companies have to send the new constructed mortality tables to the Undersecretariat of the Turkish Treasury with related formulas, methods of calculation, and assumptions in order to receive confirmation from the Undersecretariat of the Turkish Treasury (Life Insurance Regulation, 1996).

Although the insurance companies can construct their own mortality tables, no insurance company has yet used their own mortality table. All insurance companies have used four mortality tables accepted by The Undersecretariat of the Turkish Treasury. But these four different mortality tables are not used with same frequency. Insurance companies use these mortality tables for different payments of different life insurance policies.

The life insurance companies use 1980 CSO mortality tables more than the tables of the 1950s because they are more up to date, being not so old, and it is important not lose the most up to date measures for insurance policies.

On the other hand, when investigating profitability, out dated mortality tables are more profitable for whole life insurance policies and term life insurance policies for insurance companies; old mortality tables contain higher death rates and therefore, risk premiums for death will be charged at an equally higher rate. However, insurance companies cannot always easily use these high death rate mortality tables as a result of growing life insurance competition. A company can ensure that they have a large client base by using a mortality table with lower death rates.

Turkey Male and Female Mortality Tables (2001)

In this study, it is aimed to construct the mortality table from the join of both sexes with orphanhood method. In this study, 1998 and 2003 TDHS results were used to adult mortality. In two surveys, the question about the survival status of the parents is asked in the household questionnaire. The question about the survival status of mothers is asked at question number 10 and the question about the survival status of fathers is asked at question number 12 in household questionnaire. These questions are;

- Is ……’s natural mother alive? (Question 10)
- Is ……’s natural father alive? (Question 12)

In these questions “natural” word is used to avoid the adoption effect. If this word is not used in questionnaires, it can affect the results of orphanhood method negatively. According to these questions, the respondents give an answer as “yes”, “no” or “don’t know” and these answers assist to calculate proportions of alive mother and proportions of alive father.

Table 1 shows that the proportion of persons whose mother/father alive in 1998 and 2003 TDHS. The synthetic orphanhood method assumes that proportion of respondents’ age (n) of the first survey will decrease when they pass the next age group (n+5) in the second survey.

But for synthetic orphanhood method, intersurvey cohort of respondents should be calculated. Table 2 shows that 1998 and 2003 intersurvey proportion of persons whose father and mother alive.

Table 1: Proportion of Persons Whose Mother/Father Alive in 1998 and 2003 TDHS

	Proportion of Alive Mother		Proportion of Alive Father	
age	**1998 TDHS**	**2003 TDHS**	**1998 TDHS**	**2003 TDHS**
0 - 4	0.9972	0.9978	0.9942	0.9959
5 - 9	0.9923	0.9944	0.9824	0.9858
10 - 14	0.9823	0.9891	0.9589	0.9658
15 - 19	0.9753	0.9750	0.9203	0.9377
20 - 24	0.9547	0.9585	0.8863	0.8936
25 - 29	0.9228	0.9377	0.7937	0.8276
30 - 34	0.8650	0.8804	0.7083	0.7069
35 - 39	0.8035	0.8061	0.5959	0.5914
40 - 44	0.7190	0.6938	0.4606	0.4653
45 - 49	0.5863	0.6230	0.3247	0.3586
50 - 54	0.3761	0.4577	0.1650	0.2185
55 - 59	0.2113	0.2534	0.0704	0.0976
60 - 64	0.0894	0.1008	0.0266	0.0332
65 - 69	0.0517	0.0463	0.0179	0.0178
70 - 74	0.0233	0.0151	0.0033	0.0043
75+	0.0255	0.0132	0.0153	0.0095

Brass established an equation relating to the female probability of surviving; Hill and Blacker established an equation relating to the male probability of surviving from age 25 to 25+n. These equations have the form:

$$l_{(25+n)} / l_{(25)} = W_{(n)} \cdot S_{(n-5)} + (1-W_{(n)}) \cdot S_{(n)} \qquad \text{for females}$$

$$l_{(35+n)} / l_{(32,5)} = W_{(n)} \cdot S_{(n-5)} + (1-W_{(n)}) \cdot S_{(n)} \qquad \text{for males}$$

Where, S(n) is the proportion of respondents aged from n to n+4 of alive mother or father. W(n) is a weighting factor which is calculated according to mean age at maternity and paternity using interpolation formula. Mean age at maternity can be calculated directly by using TDSH-1998 and 2003 Individual Questionnaire. But it is not possible for the mean age of paternity which cannot be calculated directly from data. Therefore, first age of marriage of males subtract from the first age of marriage of females. And than it added to mean age at maternity and later, nine-month pregnancy periods is subtracted from this number, since males are also under the risk of dying during this pregnancy period.

Table 2: 1998 and 2003 Intersurvey Proportion of Persons Whose Mother/Father Alive

	1998-2003 intersurvey proportion of alive mother	1998-2003 intersurvey proportion of alive father
5-9	0.9950	0.9874
10-14	0.9918	0.9708
15-19	0.9845	0.9494
20-24	0.9675	0.9219
25-29	0.9503	0.8609
30-34	0.9066	0.7667
35-39	0.8449	0.6401
40-44	0.7296	0.4999
45-49	0.6322	0.3892
50-54	0.4936	0.2619
55-59	0.3327	0.1549
60-64	0.1587	0.0730
65-69	0.0822	0.0489
70-74	0.0239	0.0117
75+	0.0136	0.0334

In this study, it is used only the last one year data preceding the 1998 and 2003 surveys. So, the average mean age of childbearing can be calculated for 2001.2 as a reference time. Table 3 shows that the mean age at maternity and paternity for 1998 and 2003. Table 4 and Table 5 indicate female and male adult survivorship probabilities which are calculated by Brass formula.

Table 3: Mean Age of Childbearing for Females and Males

	Mean Age of Childbearing	
	Females	**Males**
1998	27,09	30,52
2003	26,79	30,73
Average	26,94	30,62

Female and male adult survivorship probabilities assist to determine suitable mortality level and then e_{20} values from the west model life table between 1998 and 2003. So, orphanhood method is important to estimate adult mortality. The value of e_{20} provides already crucial information about adult mortality.

Table 4: Female Adult Survivorship Probability l(25+n)/l(25)

Age (n)	Weighting factor W(n)	Proportion with mother surviving S(n-5)	Complement of weighing factor (1-W(n))	Proportion with mother surviving S(n)	Female adult survivorship probability l(25+n)/l(25)
20	0.8334	0.9845	0.1666	0.9675	0.9817
25	0.9072	0.9675	0.0928	0.9503	0.9659
30	0.9502	0.9503	0.0498	0.9066	0.9481
35	0.9781	0.9066	0.0219	0.8449	0.9053
40	0.9412	0.8449	0.0588	0.7296	0.8381
45	0.8742	0.7296	0.1258	0.6322	0.7174
50	0.6887	0.6322	0.3113	0.4936	0.5891

Table 5: Male Adult Survivorship Probability l(35+n)/l(35)

Age (n)	Weighting factor W(n)	Proportion with father surviving S(n-5)	Complement of weighing factor (1-W(n))	Proportion with father surviving S(n)	Male adult survivorship probability l(35+n)/l(35)
20	0.3610	0.9494	0.6390	0.9219	0.9318
25	0.2817	0.9219	0.7183	0.8609	0.8781
30	0.1011	0.8609	0.8989	0.7667	0.7762
35	-0.1090	0.7667	1.1090	0.6401	0.6263
40	-0.4452	0.6401	1.4452	0.4999	0.4375
45	-0.7096	0.4999	1.7096	0.3892	0.3106

Table 6: Mean Mortality Level, and e_{20} values for 1998-2003 Intersurvey Cohort

	Mean West Level	e_{20}
Female	22,62	55,99
Male	22,17	52,14

In this study, COMBIN application of MORTPAK is used to construct abridged life tables which joins the infant and the child mortality rate and estimated e_{20} values by using west model life table (United Nations, 1988). This application produces mortality tables; however these mortality tables are constructed as an abridged mortality table. Thus, UNABR application of MORTPAK is used to transform abridged mortality tables to unabridged mortality tables which consist of single ages (United Nations, 1988).

Table 7: Turkey Female Mortality Table (2001)

x	lx	dx	qx	px	mx	Sx	Lx	Tx	ex
0	100,000	2,886	0.02886	0.97114	0.02928	0.98331	98,557	7,285,398	72.85
1	97,114	403	0.00415	0.99585	0.00416	0.99706	96,912	7,186,841	74.00
2	96,711	166	0.00172	0.99828	0.00172	0.99867	96,628	7,089,928	73.31
3	96,545	90	0.00093	0.99907	0.00093	0.99924	96,500	6,993,301	72.44
4	96,455	56	0.00058	0.99942	0.00058	0.99951	96,427	6,896,801	71.50
5	96,399	38	0.00039	0.99961	0.00039	0.99966	96,380	6,800,374	70.54
6	96,361	27	0.00028	0.99972	0.00028	0.99975	96,348	6,703,994	69.57
7	96,334	21	0.00022	0.99978	0.00022	0.99980	96,324	6,607,646	68.59
8	96,313	16	0.00017	0.99983	0.00017	0.99984	96,305	6,511,322	67.61
9	96,297	14	0.00015	0.99985	0.00015	0.99985	96,290	6,415,017	66.62
10	96,282	13	0.00014	0.99986	0.00014	0.99986	96,276	6,318,728	65.63
11	96,269	13	0.00014	0.99986	0.00014	0.99986	96,262	6,222,452	64.64
12	96,255	14	0.00015	0.99985	0.00015	0.99984	96,248	6,126,190	63.65
13	96,241	17	0.00018	0.99982	0.00018	0.99980	96,232	6,029,942	62.65
14	96,224	21	0.00022	0.99978	0.00022	0.99976	96,213	5,933,710	61.67
15	96,202	25	0.00026	0.99974	0.00026	0.99972	96,190	5,837,497	60.68
16	96,177	29	0.00030	0.99970	0.00030	0.99968	96,163	5,741,307	59.69
17	96,149	33	0.00034	0.99966	0.00034	0.99964	96,132	5,645,144	58.71
18	96,116	37	0.00038	0.99962	0.00038	0.99960	96,098	5,549,012	57.73
19	96,079	40	0.00042	0.99958	0.00042	0.99957	96,059	5,452,914	56.75
20	96,039	43	0.00045	0.99955	0.00045	0.99954	96,017	5,356,855	55.78
21	95,996	45	0.00047	0.99953	0.00047	0.99953	95,973	5,260,837	54.80
22	95,951	46	0.00048	0.99952	0.00048	0.99952	95,928	5,164,864	53.83
23	95,905	47	0.00049	0.99951	0.00049	0.99951	95,881	5,068,937	52.85
24	95,858	48	0.00050	0.99950	0.00050	0.99950	95,834	4,973,055	51.88
25	95,810	48	0.00050	0.99950	0.00050	0.99950	95,786	4,877,222	50.91
26	95,762	49	0.00051	0.99949	0.00051	0.99949	95,737	4,781,436	49.93
27	95,713	49	0.00051	0.99949	0.00051	0.99949	95,689	4,685,699	48.96
28	95,664	50	0.00052	0.99948	0.00052	0.99948	95,639	4,590,010	47.98

Table 7: Turkey Female Mortality Table (2001) (Continued)

x	lx	dx	qx	px	mx	Sx	Lx	Tx	ex
29	95,614	51	0.00053	0.99947	0.00053	0.99947	95,589	4,494,371	47.01
30	95,564	52	0.00054	0.99946	0.00054	0.99945	95,538	4,398,782	46.03
31	95,512	53	0.00056	0.99944	0.00056	0.99943	95,485	4,303,244	45.05
32	95,459	56	0.00059	0.99941	0.00059	0.99940	95,430	4,207,759	44.08
33	95,402	59	0.00062	0.99938	0.00062	0.99936	95,373	4,112,328	43.11
34	95,343	63	0.00066	0.99934	0.00066	0.99932	95,312	4,016,956	42.13
35	95,280	68	0.00071	0.99929	0.00071	0.99926	95,246	3,921,644	41.16
36	95,213	73	0.00077	0.99923	0.00077	0.99920	95,176	3,826,397	40.19
37	95,139	80	0.00084	0.99916	0.00084	0.99912	95,099	3,731,222	39.22
38	95,059	87	0.00092	0.99908	0.00092	0.99904	95,016	3,636,122	38.25
39	94,972	96	0.00101	0.99899	0.00101	0.99894	94,924	3,541,107	37.29
40	94,876	106	0.00112	0.99888	0.00112	0.99882	94,823	3,446,183	36.32
41	94,770	118	0.00125	0.99875	0.00125	0.99868	94,710	3,351,360	35.36
42	94,651	132	0.00139	0.99861	0.00139	0.99853	94,585	3,256,649	34.41
43	94,520	147	0.00155	0.99845	0.00155	0.99836	94,446	3,162,064	33.45
44	94,373	163	0.00173	0.99827	0.00173	0.99817	94,292	3,067,618	32.51
45	94,210	183	0.00194	0.99806	0.00194	0.99795	94,119	2,973,326	31.56
46	94,027	204	0.00217	0.99783	0.00217	0.99770	93,925	2,879,208	30.62
47	93,823	229	0.00244	0.99756	0.00244	0.99742	93,709	2,785,282	29.69
48	93,594	256	0.00273	0.99727	0.00273	0.99710	93,466	2,691,574	28.76
49	93,339	287	0.00307	0.99693	0.00307	0.99675	93,195	2,598,107	27.84
50	93,052	320	0.00344	0.99656	0.00345	0.99635	92,892	2,504,912	26.92
51	92,732	359	0.00387	0.99613	0.00388	0.99590	92,553	2,412,020	26.01
52	92,373	401	0.00434	0.99566	0.00435	0.99540	92,173	2,319,467	25.11
53	91,972	448	0.00487	0.99513	0.00488	0.99483	91,748	2,227,295	24.22
54	91,524	501	0.00547	0.99453	0.00549	0.99419	91,274	2,135,546	23.33
55	91,024	560	0.00615	0.99385	0.00617	0.99348	90,744	2,044,272	22.46
56	90,464	624	0.00690	0.99310	0.00692	0.99268	90,152	1,953,529	21.59
57	89,840	696	0.00775	0.99225	0.00778	0.99177	89,492	1,863,377	20.74

Table 7: Turkey Female Mortality Table (2001) (Continued)

x	lx	dx	qx	px	mx	Sx	Lx	Tx	ex
58	89,143	776	0.00871	0.99129	0.00875	0.99076	88,755	1,773,885	19.90
59	88,367	864	0.00978	0.99022	0.00983	0.98962	87,935	1,685,130	19.07
60	87,503	961	0.01098	0.98902	0.01104	0.98835	87,022	1,597,195	18.25
61	86,542	1,066	0.01232	0.98768	0.01240	0.98693	86,009	1,510,173	17.45
62	85,476	1,182	0.01383	0.98617	0.01393	0.98533	84,885	1,424,164	16.66
63	84,294	1,308	0.01552	0.98448	0.01564	0.98354	83,640	1,339,279	15.89
64	82,985	1,445	0.01741	0.98259	0.01756	0.98154	82,263	1,255,640	15.13
65	81,541	1,592	0.01953	0.98047	0.01972	0.97930	80,744	1,173,377	14.39
66	79,948	1,751	0.02190	0.97810	0.02214	0.97679	79,073	1,092,632	13.67
67	78,197	1,920	0.02455	0.97545	0.02486	0.97398	77,237	1,013,559	12.96
68	76,278	2,099	0.02752	0.97248	0.02790	0.97085	75,228	936,322	12.28
69	74,178	2,287	0.03083	0.96917	0.03131	0.96735	73,035	861,094	11.61
70	71,891	2,482	0.03452	0.96548	0.03513	0.96346	70,651	788,059	10.96
71	69,410	2,682	0.03864	0.96136	0.03940	0.95911	68,069	717,409	10.34
72	66,728	2,885	0.04323	0.95677	0.04419	0.95427	65,285	649,340	9.73
73	63,843	3,086	0.04834	0.95166	0.04954	0.94889	62,300	584,054	9.15
74	60,757	3,282	0.05402	0.94598	0.05552	0.94292	59,116	521,754	8.59
75	57,475	3,467	0.06032	0.93968	0.06220	0.93630	55,741	462,638	8.05
76	54,008	3,635	0.06730	0.93270	0.06964	0.92897	52,191	406,897	7.53
77	50,373	3,780	0.07503	0.92497	0.07795	0.92087	48,483	354,706	7.04
78	46,594	3,893	0.08356	0.91644	0.08720	0.91194	44,647	306,223	6.57
79	42,700	3,970	0.09297	0.90703	0.09750	0.90211	40,715	261,576	6.13
80	38,731	4,002	0.10332	0.89668	0.10895	0.89131	36,730	220,860	5.70
81	34,729	3,983	0.11468	0.88532	0.12166	0.87949	32,738	184,131	5.30
82	30,746	3,908	0.12710	0.87290	0.13573	0.86658	28,792	151,393	4.92
83	26,838	3,775	0.14066	0.85934	0.15130	0.85252	24,951	122,601	4.57
84	23,063	3,584	0.15541	0.84459	0.16850	0.83727	21,271	97,650	4.23
85	19,479	3,339	0.17140	0.82860	0.18747	0.82078	17,810	76,379	3.92
86	16,140	3,045	0.18866	0.81134	0.20831	0.80302	14,618	58,569	3.63

Table 7: Turkey Female Mortality Table (2001) (Continued)

x	lx	dx	qx	px	mx	Sx	Lx	Tx	ex
87	13,095	2,714	0.20723	0.79277	0.23118	0.78398	11,738	43,951	3.36
88	10,382	2,358	0.22711	0.77289	0.25620	0.76365	9,203	32,213	3.10
89	8,024	1,992	0.24830	0.75170	0.28350	0.74205	7,028	23,010	2.87
90	6,031	1,633	0.27078	0.72922	0.31318	0.71922	5,215	15,983	2.65
91	4,398	1,295	0.29450	0.70550	0.34535	0.69521	3,751	10,768	2.45
92	3,103	991	0.31938	0.68062	0.38007	0.67011	2,607	7,017	2.26
93	2,112	729	0.34534	0.65466	0.41742	0.64401	1,747	4,410	2.09
94	1,383	515	0.37225	0.62775	0.45738	0.61706	1,125	2,663	1.93
95	868	347	0.39997	0.60003	0.49995	0.58938	694	1,537	1.77
96	521	223	0.42836	0.57164	0.54511	0.56114	409	843	1.62
97	298	136	0.45722	0.54278	0.59272	0.53252	230	434	1.46
98	162	79	0.48637	0.51363	0.64265	0.50370	122	204	1.26
99	83	43	0.51562	0.48438	0.69473	0.32632	62	82	0.98
100	40	40	1.00000	0.00000	2.00000	0.00000	20	20	0.50

Table 8: Turkey Male Mortality Table (2001)

x	lx	dx	qx	px	mx	Sx	Lx	Tx	ex
0	100,000	3,010	0.03010	0.96990	0.03056	0.98253	98,495	6,865,775	68.66
1	96,990	432	0.00445	0.99555	0.00446	0.99660	96,774	6,767,280	69.77
2	96,558	227	0.00235	0.99765	0.00235	0.99804	96,445	6,670,506	69.08
3	96,331	150	0.00156	0.99844	0.00156	0.99864	96,256	6,574,061	68.24
4	96,181	111	0.00115	0.99885	0.00115	0.99897	96,126	6,477,804	67.35
5	96,071	87	0.00091	0.99909	0.00091	0.99917	96,027	6,381,679	66.43
6	95,983	72	0.00075	0.99925	0.00075	0.99930	95,947	6,285,652	65.49
7	95,911	61	0.00064	0.99936	0.00064	0.99939	95,880	6,189,705	64.54
8	95,850	55	0.00057	0.99943	0.00057	0.99945	95,822	6,093,824	63.58
9	95,795	50	0.00052	0.99948	0.00052	0.99949	95,770	5,998,002	62.61
10	95,745	47	0.00049	0.99951	0.00049	0.99951	95,722	5,902,231	61.65
11	95,698	47	0.00049	0.99951	0.00049	0.99950	95,675	5,806,509	60.68
12	95,652	49	0.00051	0.99949	0.00051	0.99946	95,627	5,710,834	59.70
13	95,603	54	0.00057	0.99943	0.00057	0.99939	95,576	5,615,207	58.73
14	95,548	62	0.00065	0.99935	0.00065	0.99930	95,517	5,519,632	57.77
15	95,486	72	0.00075	0.99925	0.00075	0.99919	95,450	5,424,114	56.81
16	95,415	83	0.00087	0.99913	0.00087	0.99907	95,373	5,328,664	55.85
17	95,332	95	0.00100	0.99900	0.00100	0.99894	95,284	5,233,291	54.90
18	95,236	107	0.00112	0.99888	0.00112	0.99883	95,183	5,138,007	53.95
19	95,130	116	0.00122	0.99878	0.00122	0.99874	95,072	5,042,824	53.01
20	95,013	124	0.00131	0.99869	0.00131	0.99866	94,951	4,947,753	52.07
21	94,889	130	0.00137	0.99863	0.00137	0.99861	94,824	4,852,802	51.14
22	94,759	134	0.00141	0.99859	0.00141	0.99858	94,692	4,757,978	50.21
23	94,625	136	0.00144	0.99856	0.00144	0.99856	94,557	4,663,285	49.28
24	94,489	137	0.00145	0.99855	0.00145	0.99855	94,421	4,568,728	48.35
25	94,352	138	0.00146	0.99854	0.00146	0.99854	94,283	4,474,307	47.42
26	94,214	138	0.00146	0.99854	0.00146	0.99854	94,146	4,380,024	46.49
27	94,077	137	0.00146	0.99854	0.00146	0.99854	94,008	4,285,879	45.56
28	93,939	137	0.00146	0.99854	0.00146	0.99854	93,871	4,191,870	44.62

Table 8: Turkey Male Mortality Table (2001) (Continued)

x	lx	dx	qx	px	mx	Sx	Lx	Tx	ex
29	93,802	138	0.00147	0.99853	0.00147	0.99852	93,733	4,097,999	43.69
30	93,664	140	0.00149	0.99851	0.00149	0.99850	93,595	4,004,266	42.75
31	93,525	142	0.00152	0.99848	0.00152	0.99846	93,454	3,910,671	41.81
32	93,383	147	0.00157	0.99843	0.00157	0.99840	93,309	3,817,218	40.88
33	93,236	152	0.00163	0.99837	0.00163	0.99833	93,160	3,723,908	39.94
34	93,084	160	0.00172	0.99828	0.00172	0.99824	93,004	3,630,748	39.01
35	92,924	168	0.00181	0.99819	0.00181	0.99813	92,840	3,537,744	38.07
36	92,756	179	0.00193	0.99807	0.00193	0.99800	92,666	3,444,904	37.14
37	92,577	193	0.00208	0.99792	0.00208	0.99784	92,481	3,352,238	36.21
38	92,384	207	0.00224	0.99776	0.00224	0.99767	92,281	3,259,757	35.28
39	92,177	223	0.00242	0.99758	0.00242	0.99748	92,066	3,167,476	34.36
40	91,954	242	0.00263	0.99737	0.00263	0.99725	91,833	3,075,411	33.45
41	91,712	263	0.00287	0.99713	0.00287	0.99700	91,581	2,983,577	32.53
42	91,449	287	0.00314	0.99686	0.00314	0.99672	91,306	2,891,997	31.62
43	91,162	313	0.00343	0.99657	0.00344	0.99640	91,006	2,800,691	30.72
44	90,849	343	0.00377	0.99623	0.00378	0.99605	90,678	2,709,685	29.83
45	90,507	374	0.00413	0.99587	0.00414	0.99567	90,320	2,619,007	28.94
46	90,133	409	0.00454	0.99546	0.00455	0.99524	89,928	2,528,687	28.06
47	89,724	448	0.00499	0.99501	0.00500	0.99477	89,500	2,438,759	27.18
48	89,276	489	0.00548	0.99452	0.00550	0.99425	89,032	2,349,259	26.31
49	88,787	535	0.00603	0.99397	0.00605	0.99367	88,519	2,260,227	25.46
50	88,252	585	0.00663	0.99337	0.00665	0.99304	87,959	2,171,708	24.61
51	87,666	640	0.00730	0.99270	0.00733	0.99234	87,346	2,083,749	23.77
52	87,026	699	0.00803	0.99197	0.00806	0.99157	86,677	1,996,403	22.94
53	86,328	763	0.00884	0.99116	0.00888	0.99072	85,946	1,909,726	22.12
54	85,564	833	0.00973	0.99027	0.00978	0.98978	85,148	1,823,779	21.31
55	84,732	907	0.01071	0.98929	0.01077	0.98875	84,278	1,738,631	20.52
56	83,824	988	0.01179	0.98821	0.01186	0.98762	83,330	1,654,353	19.74
57	82,836	1,074	0.01297	0.98703	0.01305	0.98638	82,299	1,571,023	18.97

Table 8: Turkey Male Mortality Table (2001) (Continued)

x	lx	dx	qx	px	mx	Sx	Lx	Tx	ex
58	81,762	1,167	0.01427	0.98573	0.01437	0.98502	81,178	1,488,724	18.21
59	80,595	1,266	0.01571	0.98429	0.01583	0.98351	79,962	1,407,545	17.46
60	79,329	1,371	0.01728	0.98272	0.01743	0.98186	78,644	1,327,583	16.74
61	77,958	1,482	0.01901	0.98099	0.01919	0.98005	77,217	1,248,940	16.02
62	76,476	1,599	0.02091	0.97909	0.02113	0.97806	75,677	1,171,723	15.32
63	74,877	1,721	0.02299	0.97701	0.02326	0.97588	74,016	1,096,046	14.64
64	73,156	1,849	0.02528	0.97472	0.02560	0.97348	72,231	1,022,030	13.97
65	71,306	1,982	0.02779	0.97221	0.02818	0.97085	70,315	949,799	13.32
66	69,325	2,117	0.03054	0.96946	0.03101	0.96798	68,266	879,484	12.69
67	67,207	2,255	0.03355	0.96645	0.03412	0.96483	66,080	811,218	12.07
68	64,953	2,394	0.03685	0.96315	0.03754	0.96138	63,756	745,138	11.47
69	62,559	2,531	0.04046	0.95954	0.04130	0.95761	61,294	681,382	10.89
70	60,028	2,666	0.04441	0.95559	0.04542	0.95348	58,695	620,088	10.33
71	57,362	2,795	0.04873	0.95127	0.04995	0.94897	55,965	561,393	9.79
72	54,567	2,916	0.05344	0.94656	0.05491	0.94406	53,109	505,429	9.26
73	51,651	3,026	0.05858	0.94142	0.06035	0.93870	50,138	452,320	8.76
74	48,625	3,121	0.06418	0.93582	0.06631	0.93288	47,065	402,182	8.27
75	45,504	3,198	0.07027	0.92973	0.07283	0.92654	43,906	355,117	7.80
76	42,307	3,253	0.07690	0.92310	0.07998	0.91965	40,680	311,211	7.36
77	39,053	3,284	0.08409	0.91591	0.08778	0.91218	37,411	270,531	6.93
78	35,769	3,287	0.09189	0.90811	0.09632	0.90409	34,126	233,120	6.52
79	32,483	3,259	0.10034	0.89966	0.10564	0.89534	30,853	198,994	6.13
80	29,223	3,199	0.10947	0.89053	0.11581	0.88589	27,624	168,141	5.75
81	26,024	3,105	0.11932	0.88068	0.12689	0.87572	24,472	140,517	5.40
82	22,919	2,978	0.12992	0.87008	0.13895	0.86478	21,430	116,046	5.06
83	19,941	2,818	0.14132	0.85868	0.15206	0.85303	18,532	94,616	4.74
84	17,123	2,629	0.15354	0.84646	0.16631	0.84047	15,809	76,083	4.44
85	14,494	2,415	0.16661	0.83339	0.18175	0.82705	13,287	60,275	4.16
86	12,079	2,181	0.18056	0.81944	0.19848	0.81276	10,989	46,988	3.89

Table 8: Turkey Male Mortality Table (2001) (Continued)

x	lx	dx	qx	px	mx	Sx	Lx	Tx	ex
87	9,898	1,934	0.19540	0.80460	0.21656	0.79758	8,931	35,999	3.64
88	7,964	1,682	0.21115	0.78885	0.23607	0.78150	7,123	27,068	3.40
89	6,282	1,431	0.22781	0.77219	0.25709	0.76454	5,567	19,945	3.17
90	4,851	1,190	0.24537	0.75463	0.27968	0.74669	4,256	14,378	2.96
91	3,661	966	0.26383	0.73617	0.30392	0.72798	3,178	10,122	2.76
92	2,695	763	0.28315	0.71685	0.32985	0.70844	2,314	6,944	2.58
93	1,932	586	0.30330	0.69670	0.35752	0.68810	1,639	4,630	2.40
94	1,346	436	0.32424	0.67576	0.38698	0.66702	1,128	2,991	2.22
95	910	315	0.34591	0.65409	0.41825	0.64526	752	1,864	2.05
96	595	219	0.36824	0.63176	0.45134	0.62289	485	1,111	1.87
97	376	147	0.39114	0.60886	0.48623	0.60000	302	626	1.67
98	229	95	0.41454	0.58546	0.52293	0.57668	181	324	1.41
99	134	59	0.43833	0.56167	0.56136	0.35966	105	142	1.06
100	75	75	1.00000	0.00000	2.00000	0.00000	38	38	0.50

COMPARISON OF THE RESULTS WITH AMERICAN CSO 1980 TABLES

The crucial objective of this study is to compare Turkish mortality tables (Tables 7 and 8) which are constructed with Turkish mortality data for both sexes, and four other foreign mortality tables which are used in Turkey. The premium payments are important for both insurer and insured. Thanks to this comparison and analysis, the real insurance premium will be calculated for Turkey.

The other important point is that calculation of net single premium will include only endowment life insurance. The results of the endowment life insurance have great importance because endowment life insurance is the most preferential type of life insurance policy by insurance companies in Turkey. Therefore, the results of this life insurance will exhibit the usability of Turkish mortality tables more precisely as other types of life insurance have a single probability factor, such as dying or living but endowment life insurance includes both probabilities.

In Table 9, the mortality ratios in CSO 1980 which is used by the insurance companies in Turkey and Turkey mortality table developed for both genders were compared.

Table 9: Comparison of 15-Year Death Rates (per thousand)

	American CSO 1980	Turkey Female 2001	Turkey Male 2001
15q30	34.21	11,97	29,66
15q40	84.77	40,62	78,55
15q50	190.51	123,71	192,00
15q60	422.71	343,17	426,39

The comparison of mortality ratios was done for the ages of 30, 40, 50 and 60 in 15-year intervals. Accordingly the mortality ratios of CSO 1980 mortality table have values close to those of Turkey Male Mortality table and higher than the values of Turkey female mortality table.

A second comparison was made on the life expectations as shown in Table 10. Accordingly, life expectations at birth in American CSO 1980 table are higher than Turkey Male Mortality table and lower than Turkey Female Mortality table.

Table 10: Comparison of Expectation of Life at Age x (Year)

	American CSO 1980	Turkey Female 2001	Turkey Male 2001
e0	70,83	72,85	64,58
e20	52,37	55,78	50,34
e40	34,05	36,32	32,32
e60	17,51	18,25	16,20
e70	10,96	10,96	9,84

Another comparison was made over the net single premium of endowment life insurance. Annual technical interest regarding the calculation of net single premium is 9 %, which is the ratio determined by Undersecretariat of Treasure in Turkey. Table 11 shows that how much money insured have to pay, in order to he/she can take 10,000 TL compensation if he/she survives or dies at the end of the 15 years.

Table 11: Endowment Life Insurance Net Single Premium Payments for 15 Years (TL)

Age	American CSO 1980	Turkey Female 2001	Turkey Male 2001
20	2,817	2,765	2,801
25	2,816	2,767	2,805
30	2,827	2,773	2,816
35	2,857	2,788	2,844
40	2,910	2,817	2,896
45	2,991	2,871	2,984
50	3,115	2,967	3,124
55	3,307	3,132	3,339
60	3,588	3,408	3,660

According to Table 11, Turkey female mortality table has minimum insurance premium and Turkey male mortality table follows that. But after age 50, it is seemed that CSO 80 mortality table has lower premiums than Turkey male mortality table. For example, any insured who is 50 years old has to pay 2,967 TL for Turkey female mortality table, 3,115 TL for CSO 80 mortality table, and 3,124 TL for Turkey male mortality table in order to he/she can take 10,000 TL compensation if he/she survives or dies at the end of the 15 years. Consequently, it can be observed that the endowment life insurance whose premiums are calculated from Turkey female or male mortality table has the lowest premiums for the most suitable starting ages (25-40) for insurance. This situation shows that insurance companies have taken higher premiums from insured in crucial age groups. And it also causes loss for insured unjustly.

CONCLUSION

The most significant application field for mortality tables is life insurance. The insurance premiums can only be calculated with this essential element; however, in our country, mortality tables which do not represent the reality in Turkey have been used both in the insurance sector and in security institutions. Hence, it is evident that the insurance premiums calculated by foreign tables will be different. In order to eliminate this deficiency, Turkish mortality tables constructed from Turkish mortality data are needed.

Therefore there are two main objectives of this study. The first objective is to construct mortality tables which represent the reality of Turkey and to construct commutation tables which assist in the calculation of life insurance premiums. The second objective is also the calculation of life insurance premiums according to types of life insurance by using these tables. Thanks to this, four mortality tables which are used in the life insurance sector in Turkey are compared with a Turkish mortality table prepared from Turkish mortality data.

In this study, the synthetic maternal and paternal orphanhood method devised by Zlotnik and Hill (1981) was used to estimate adult mortality for males and females. This technique is also applied to 1998 and 2003 TDHS data sets for both genders and it gave an estimation for the year 2001/2. Both methods' mortality tables have been transformed from abridged mortality tables to complete mortality tables using UNABR application of MORTPAK. However, the mortality tables are not accurate enough to calculate insurance premiums. Therefore, insurance premiums have been formed by adding a technical interest on the mortality tables. At the end of the study, a premium comparison was made between Turkish commutation tables and the CSO 1980 mortality table, which is used in Turkey. These comparisons were applied to endowment life insurance.

The other important point is that life insurance companies can construct their own mortality table using their own mortality data. Ten years mortality data is sufficient to prepare this table legally in Turkey. Most life insurance companies in Turkey have been in business for more than ten years. So, most of them have ten years mortality data to date. Thus, it is not difficult to construct a mortality table using their own mortality data for life insurance companies which have a large insured portfolio. Lack of early and old age mortality data may create problems in the construction of this table for those companies. Thus, this deficiency in early and old ages can be solved by using certain mathematical calculations, such as the CSO mortality table.

When investigating the results of this study, they do not seem profitable for insurers. The premiums must equal the present value of compensation. Turkish mortality tables are a suitable for the demographic condition of Turkey. Thus, premium-compensation balance is provided with these tables. But, for other foreign mortality tables, insurance companies have placed higher premiums than the real premiums, especially for life insurance in terms of the probability of dying. For this reason, using foreign tables may seem more profitable for insurers but actually, if Turkish mortality tables have been used in the life insurance sector, more coherent estimations will be made to reflect current demographic condition of Turkey, premium-compensation balance will be provided, and this situation will increase the demand of life insurance, decrease passing from life insurance to private pension systems or leaving life insurance, increase the dividend amounts of the insured, increase the money directed to investments, and increase the share of life insurance in the GDP leading to higher confidence in the insurance sector. Consequently, this study clearly shows that the life insurance sector in Turkey needs reliable mortality tables prepared on the basis of the current mortality levels and patterns of Turkey.

REFERENCES

Özsoy, A., (1970), **Türkiye için Ölüm Tabloları**, Ordu Yardımlaşma Kurumu Yayınları, Ankara.

Akmut, Ö., (1980), **Hayat Sigortası Teori ve Türkiye'deki Uygulama**, Sevinç Matbaası, Ankara.

Kırkbeşoğlu, E. (2007), "Türkiye'de Hayat Sigortalarında Uygun Mortalite Tablosu Kullanımına Yönelik Karşılaştırmalar", **Sigorta Araştırmaları Dergisi**, Sayı:3: 141–152

Zlotnik, H. and Hill, K., (1981), "The Use of Hypothetical Cohorts in Estimating Demographic Parameters Under Conditions of Changing Fertility and Mortality", **Demography***, 18(1),* 103-22.

Life Insurance Regulation, (1996), 3 April 1996.

Osmançavuşoğlu, Ö. (1999), **Hayat Sigortalarında Aktüerya**, Ankara.

Aydın, H., (2003), **Doğum ve Ölüm Verilerinin Sistemli Olarak Derlenebilmesi İçin Model Geliştirme**, Unpublished Doctoral Dissertation, Hacettepe University Institute of Health Sciences, Ankara.

Demirbüken, D., (2001), **An Evaluation of Burial Records of Ankara City Cemeteries**, Unpublished Master Thesis, Hacettepe University Institute of Population Studies, Ankara.

Demirci, M., (1987), **Türkiye'nin Ölümlülük Yaş Yapısına Model Yaşam Tablolarından En Uygun Kalıbın Seçilmi**, Unpublished Master Thesis, Hacettepe University The Institute for Graduate Studies in Science and Engineering Statistics Program, Ankara.

Hacettepe University Institute of Population Studies (HUIPS), (1980), **Turkish Fertility Survey: 1978**, Vol.1, Hacettepe University, Ankara.

Hacettepe University Institute of Population Studies (HUIPS), (1987), *1983* **Turkish Population and Health Survey**. Hacettepe University, Ankara

Hacettepe University Institute of Population Studies (HUIPS), (1989), **1988 Turkish Population and Health Survey**. Hacettepe University, Ankara

Hacettepe University Institute of Population Studies (HUIPS), (1994), **1993 Turkey Demographic and Health Survey**. Hacettepe University, Ankara.

Hacettepe University Institute of Population Studies (HUIPS), (1999), **1998 Turkey Demographic and Health Survey**. Hacettepe University, Ankara.

Hacettepe University Institute of Population Studies (HUIPS), (2004), **2003 Turkey Demographic and Health Survey**. Hacettepe University, Ankara.

Wunsch, G. and Hancıoğlu, A. (Eds.), (1995), **Morbidity and Mortality Data: Problems of Comparability***,* Proceedings of the European Association for Population Studies and the Hacettepe Institute of Population Studies Workshop.

Timæus, I.M., (1990), **Advances in the Measurement of Adult Mortality from Data on Orphanhood**, Centre for Population Studies, London School of Hygiene and Tropical Medicine, University of London, Research Paper 90-1.

Sivamurthy M. And Seetharam K.S., (1980), **Handbook of Indirect Methods for Mortality Estimation**, Cairo Demographic Centre, Cairo.

Blacker, J.G.C., (1977), "Estimation of Adult Mortality in Africa from Data on Orphanhood", **Population Studies**, *31*, 107-128.

Brass, W. and Hill, K., (1973), "Estimating Adult Mortality from Orphanhood", *in International Population Conference*. **IUSSP**, Liege, Vol.3, pp:111-123.

United Nations, (1983), **Manual X:Indirect Techniques for Demographic Estimation**, New York.

Coşkun, Y., (2002), **Estimation of Adult Mortality by Using The Orphanhood Method from the 1993 and 1998 Turkish Demographic and Health Surveys**, Unpublished Master Thesis, Hacettepe University Institute of Population Studies, Ankara.

Timæus, I.M., (1991), "Estimation of Mortality from Orphanhood in Adulthood", **Demography***, Vol.28/2,* 213-227.

Moralı, N., (1997), **Hayat Sigortaları için Aktüeryal Teknikler**, Genç Sigortacılar Derneği Yayınları, İstanbul.

Duransoy, M.L., (1993), **Türk Mortalite Tablosu (1980-1990)**, Unpublished Doctoral Dissertation, Mimar Sinan University, İstanbul.

Ataman Y.E., (2002), **Çoklu Hayat Tabloları ve Türkiye'ye Yönelik Bir Uygulama**, Unpublished Master Thesis, Hacettepe University The Institute for Graduate Studies in Science and Engineering Actuarial Sciences, Ankara.

Economical Aspects of Acceptance Sampling

Jindřich KLŮFA

University of Economics, Prague, Czech republic
Email: klufa@vse.cz

Abstract: In this paper we shall deal with the AOQL single sampling plans when the remainder of rejected lots is inspected. We shall consider two types of AOQL plans - for inspection by variables and for inspection by variables and attributes (all items from the sample are inspected by variables, remainder of rejected lots is inspected by attributes). These plans we shall compare with the corresponding Dodge-Romig AOQL plans by attributes. We shall report on an algorithm allowing the calculation of these plans when the non-central t distribution is used for the operating characteristic. The calculation is considerably difficult, we shall use an original method and software Mathematica. From the results of numerical investigations it follows that under the same protection of consumer the AOQL plans for inspection by variables are in many situations more economical than the corresponding Dodge-Romig attribute sampling plans (saving of the inspection cost is 70% in any cases).
Keywords: Acceptance sampling, AOQL plans, inspection by variables, software Mathematica.

1 Introduction

Under the assumption that each inspected item is classified as either good or defective (acceptance sampling by attributes) in [1] are considered sampling plans which minimize the mean number of items inspected per lot of process average quality

$$I_s = N - (N-n)\cdot L(\overline{p};n,c) \tag{1}$$

under the condition

$$\max_{0<p<1} AOQ(p) = p_L \tag{2}$$

(AOQL single sampling plans), where N is the number of items in the lot (the given parameter), $\overline{p}$ is the process average fraction defective (the given parameter), p_L is the average outgoing quality limit (the given parameter, denoted AOQL), n is the number of items in the sample $(n<N)$, c is the acceptance number (the lot is rejected when the number of defective items in the sample is greater than c), $L(p)$ is the operating characteristic (the probability of accepting a submitted lot with fraction defective p), $AOQ(p)$ is average outgoing quality (the mean fraction defective after inspection when the fraction defective before inspection was p).
Condition (2) protects the consumer against the acceptance of a bad lot. The AOQL plans for inspection by attributes are in [1] extensively tabulated.

2 AOQL plans by variables and attributes

The problem to find AOQL plans *for inspection by variables* has been solved in [2] under the following assumptions:

Measurements of a single quality characteristic X are independent, identically distributed normal random variables with unknown parameters μ and σ^2. For the quality characteristic X is given either an upper specification limit U (the item is defective if its measurement exceeds U), or a lower specification limit L (the item is defective if its measurement is smaller than L). It is further assumed that the unknown parameter σ is estimated from the sample standard deviation s.

The inspection procedure is as follows: Draw a random sample of n items and compute $\bar{x}$ and s. Accept the lot if

$$\frac{U-\bar{x}}{s} \geq k, \text{ or } \frac{\bar{x}-L}{s} \geq k. \tag{3}$$

We have determine the sample size n and the critical value k. There are different solutions of this problem. In paper [2] we used for determination n and k a similar conditions as Dodge and Romig in [1].

Now we shall formulate this problem. Let us consider *AOQL plans for inspection by variables and attributes* – all items from the sample are inspected by variables, but the remainder of rejected lots is inspected only by attributes. Let us denote

c_s^* - the cost of inspection of one item by attributes,

c_m^* - the cost of inspection of one item by variables.

Inspection cost per lot, assuming that the remainder of rejected lots is inspected by attributes (the inspection by variables and attributes), is $n \cdot c_m^*$ with probability $L(p;n,k)$, and $n \cdot c_m^* + (N-n) \cdot c_s^*$ with probability $1 - L(p;n,k)$. The mean inspection cost per lot of process average quality is therefore

$$C_{ms} = n \cdot c_m^* + (N-n) \cdot c_s^* \cdot \left[1 - L(\bar{p};n,k)\right] \tag{4}$$

Now we shall look for the acceptance plan (n,k) minimizing the mean inspection cost per lot of process average quality C_{ms} under the condition (2). The conditions (2) is the same one as used for protection the consumer Dodge and Romig in [1]. Let us introduce a function

$$I_{ms} = n \cdot c_m + (N-n) \cdot \left[1 - L(\bar{p};n,k)\right], \tag{5}$$

where

$$c_m = c_m^* / c_s^*. \tag{6}$$

Since

$$C_{ms} = I_{ms} \cdot c_s^*, \tag{7}$$

both function C_{ms} and I_{ms} have a minimum for the same acceptance plan (n,k). Therefore, we shall look for the acceptance plan (n,k) minimizing (5) instead of (4) under the condition (2).

For these AOQL plans for inspection by variables and attributes *the new parameter* c_m was defined – see (6). This parameter must be estimated in each real situation. Usually is

$$c_m > 1. \tag{8}$$

Putting formally $c_m = 1$ into (5) (I_{ms} in this case is denoted I_m) we obtain

$$I_m = N - (N-n) \cdot L(\bar{p}; n, k), \tag{9}$$

i.e. the mean number of items inspected per lot of process average quality, assuming that both the sample and the remainder of rejected lots is inspected by variables. Consequently *the AOQL plans for inspection by variables* are a special case of *the AOQL plans by variables and attributes* for $c_m = 1$. From (9) is evident that for the determination AOQL plans by variables it is not necessary to estimate c_m ($c_m = 1$ is not real value of this parameter).

Summary: For the given parameters p_L, N, $\bar{p}$ and c_m we must determine the acceptance plan (n,k) for inspection by variables and attributes, minimizing I_{ms} under the condition (2).

Solution of this problem is in the paper [2], now we shall report on an algorithm allowing the calculation of these plans. In the first place we shall solve the equation (2), in the second place we shall determine the acceptance plan (n,k) minimizing I_{ms} under the condition (2). For given sample size n (and given N, p_L) we shall look for the critical value k for which holds (2), i.e. (see [2])

$$M_n(k) - p_L / (1 - \frac{n}{N}) = 0. \tag{10}$$

Under suitable assumptions solution of the equation (10) exists and is unique – see [2]. This solution is considerably difficult (explicit formula for *k* does not exist), we must solve (10) two times numerically (in the first step we determine x_M as a solution of equation $G'(x) = 0$, in the second step we determine *k* as a solution (10) – see [2]).

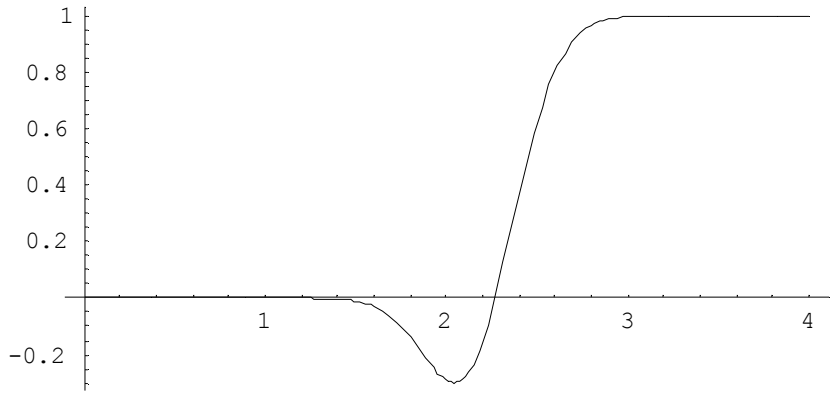


Figure 1. The function G'(x) for n=60 and k=2,2

From Figure 1 is evident that numerical solution of equation $G'(x) = 0$ depends on good first approximation x_0. In [4] is proved that solution x_M of this equation is between

$$x_r = \frac{k}{1+A}, \qquad x_a = \frac{k + A\sqrt{k^2 - 2(1-A^2)\ln A}}{1-A^2}. \tag{11}$$

Therefore we choose for x_0 following point (numerical investigations show that this point is good start value)

$$x_0 = \frac{(100+n)x_r + nx_a}{2n+100}. \tag{12}$$

2.1 AOQL plans by variables and attributes – numerical solution

For calculation of the AOQL plans by variables and attributes we shall use software Mathematica - see [5].

Example. Let N=1000, p_L =0.0025, $\overline{p}$ =0.001 and c_m =1.8 (the cost of inspection of one item by variables is higher by 80% than the cost of inspection of one item by attributes). We shall look for the AOQL plan for inspection by

variables and attributes. Furthermore we shall compare this plan and the corresponding Dodge-Romig AOQL plan for inspection by attributes.

Solution. In the first step we shall determine x_M as a solution of equation $G'(x) = 0$. According to (11) and (12) we have (*G'(x)* see [2])

```
In[1]:= << Statistics`ContinuousDistribution`
In[2]:= ndist = NormalDistribution[0, 1]
In[3]:= cm = 1.8
In[4]:= AOQL=0.0025
In[5]:= pL = 0.0025
In[6]:= pbar = 0.001
In[7]:= nbig = 1000
In[8]:= A[n_, k_] := Sqrt[1/n + k^2/(2n - 2)];
G'[x_, n_, k_] := CDF[ndist, (x - k)/A[n, k]] - CDF[ndist, -x]*
                    Exp[-((1 - A[n, k]^2) x^2 - 2k x + k^2)/
                                        (2A[n, k]^2)]/A[n, k];
xr[n_, k_] := k/(1 + A[n, k]);
xa[n_, k_] := (k + A[n, k]*Sqrt[k^2 - 2(1 - A[n, k]^2)*
                                        Log[A[n, k]]])/(1 - A[n, k]^2);
x0[n_, k_] := ((100 + n)*xr[n, k] + n*xa[n, k])/(2n + 100);
FR[n_, k_] := FindRoot[G'[x, n, k] == 0, {x, x0[n, k]}];
xM[n_, k_] := x /. FR[n, k];
```

Using Newton's method for solution (10) ($M_n(k), M_n'(k)$ see [2]) with start point o=1.6 we have

```
c[n_, k_] := -(CDF[ndist, -xM[n, k]]*CDF[ndist, (xM[n, k] -k)/
                    A[n, k]] – pL/(1 - n/nbig))/
                        (-CDF[ndist, -xM[n,k]]*(1/n + k xM[n, k]/
                        (2n - 2))*Exp[-(xM[n, k] - k)^2/(2A[n, k]^2)]/
                        (A[n, k]^3*Sqrt[2Pi]));
    o = 1.6;
fRecAux[n_,i_]:= fRecAux[n,i]=fRecAux[n,i-1]+c[n,fRecAux[n,i-1]];
                    fRecAux[n_,0]=o;
k[n_]:=fRecAux[n,7];
```

The acceptance plan (n, k) minimizing I_{ms} (see (5)) under the condition (2) is:

```
    a[n_] := CDF[ndist, (k[n] - Quantile[ndist, 1 - pbar])/Sqrt[1/n +
                        k[n]^2/(2n - 2)]];

Ims[n_] := n cm + (nbig - n)*a[n];
```

```
FMinSearch[nl_,nu_]:= nl /; nl == nu;
FMinSearch[nl_,nu_]:= FMinSearch[nl, nl+Floor[(nu-nl)/2]] /;
Ims[nl+Floor[(nu-nl)/2]]<= Ims[nl+Floor[(nu-nl)/2]+1];
FMinSearch[nl_,nu_]:= FMinSearch[nl+Floor[(nu-nl)/2]+1, nu];

 n=FMinSearch[7, nbig/2];
```

Correction for non-central t distribution (the operating characteristic is e.g. in [3]):

```
In[25]:= lambda[p_]:=Quantile[ndist,1-p]*Sqrt[n]
In[26]:= nonctdist[p_]:=NoncentralTDistribution[n-1,lambda[p]]
In[27]:= L1[p_]:=1-CDF[nonctdist[p], k[n]*Sqrt[n]]
In[28]:= AOQ[p_]:=(1-n/nbig)*p*L1[p]
In[29]:= d=0.00001
In[30]:= fMSmodq[pl_,pu_]:=pl /; pl==pu
      fMsmodq[pl_,pu_]:=fMSmodq[pl,pl+Floor[(pu-pl)/(2d)]*d] /;
      -AOQ[pl+Floor[(pu-pl)/(2d)]*d] <= -AOQ[pl+Floor[(pu-pl)/(2d)*d]+d]
      fMSmodq[pl_,pu_]:=fMSmodq[pl+Floor[(pu-pl)/(2d)]*d+d, pu]
In[33]:= pLtrue := AOQ[fMSmodq[0.00001, 0.01]]
In[34]:= pcpL = 0.00000001
In[35]:= samplan := {n, k[n]} /; (konst=pLtrue; Abs[konst – AOQL] < pcpL);
      Samplan := (pL=pL+AOQL-konst; Clear[fRecAux]; fRecAux[n_,i_] :=
      fRecAux[n,i] = fRecAux[n,i-1]+ c[n,fRecAux[n,i-1]];
      fRecAux[n_,0]=o; samplan)
In[37]:= samplan
Out[37]= {49, 2.57617}
```

The AOQL plan for inspection by variables and attributes is n = 49, k = 2.57617.

The corresponding AOQL plan for inspection by attributes we find in [1]. For given parameters N, p_L and $\overline{p}$ we have $n_2 = 130$, $c = 0$. For the comparison these two plans from an economical point of view we use parameter e (see [2]). The Mathematica gives

```
In[38] := n = 49
In[39] := k = 2.57617
In[40] := n2 = 130
In[41] := c = 0
In[42] := L1[p_]:= 1 – CDF[nonctdist[p], k*Sqrt[n]]
In[43] := L2[p_]:= Sum[Binomial[nbig*p, i]*
      Binomial[nbig - nbig*p, n2 - i]/Binomial[nbig, n2],{i, 0, c}])
In[44] := e = 100*(n*cm+(nbig-n)*(1-L1[pbar]))/(nbig-(nbig-n2)*L2[pbar])
Out[44]:= 52.2008
```

Since e = 52.2008%, using the AOQL plan for inspection by variables and attributes

(49, 2.57617) it can be expected approximately **48% saving of the inspection cost** in comparison with the corresponding Dodge-Romig plan (130, 0).

Further we compare the operating characteristics of these plans:

```
In[45]:= Table[{p, N[L1[p], 6], N[L2[p], 6]}, {p, 0.001, 0.031, 0.002}]
In[46]:= TableForm[%]
Out[46]//TableForm=
```

0.001	0.959306	0.87
0.003	0.734422	0.658207
0.005	0.522047	0.497674
0.007	0.365862	0.376067
0.009	0.256813	0.284003
0.011	0.181453	0.214346
0.013	0.129244	0.161675
0.015	0.0928235	0.121872
0.017	0.0672038	0.0918112
0.019	0.049026	0.0691225
0.021	0.0360195	0.0520083
0.023	0.0266387	0.039107
0.025	0.019822	0.0293876
0.027	0.0148338	0.0220699
0.029	0.0111597	0.0165638
0.031	0.00843709	0.0124235

For example we get $L_1(\bar{p}) = L_1(0.001) = 0.959306$, i.e. the producer's risk for the AOQL plan for inspection by variables and attributes is therefore approximately $\alpha = 1 - L_1(\bar{p}) = 0.04$.

The producer's risk for the corresponding Dodge-Romig plan is

$$\alpha = 1 - L_2(\bar{p}) = 1 - 0.87 = 0.13.$$

Finally graphic comparison of the operating characteristics of these plans (see Figure 2):

```
In[47]:= oc1 = Plot[L1[p], {p, 0, 0.025}, AspectRatio -> 0.9,
      AxesLabel -> {"p", "L(p)"}, PlotStyle -> Thickness[0.0045]]
In[48]:= oc2 = ListPlot[Table[{p, L2[p]}, {p, 0, 0.025, 0.0003}]]
In[49]:= Show[oc1, oc2]
```

3 Conclusions

From these results it follows that the AOQL plan for inspection by variables and attributes is more economical than the corresponding Dodge-Romig AOQL attribute sampling plan (48% saving of the inspection cost). Furthermore the OC curve for the AOQL plan by variables and attributes is better than corresponding OC curve for the AOQL plan by attributes - see Figure 2 (for example the producer's risk for the AOQL plan by variables and attributes $\alpha = 0.04$ is less than for the corresponding Dodge-Romig plan $\alpha = 0.13$).

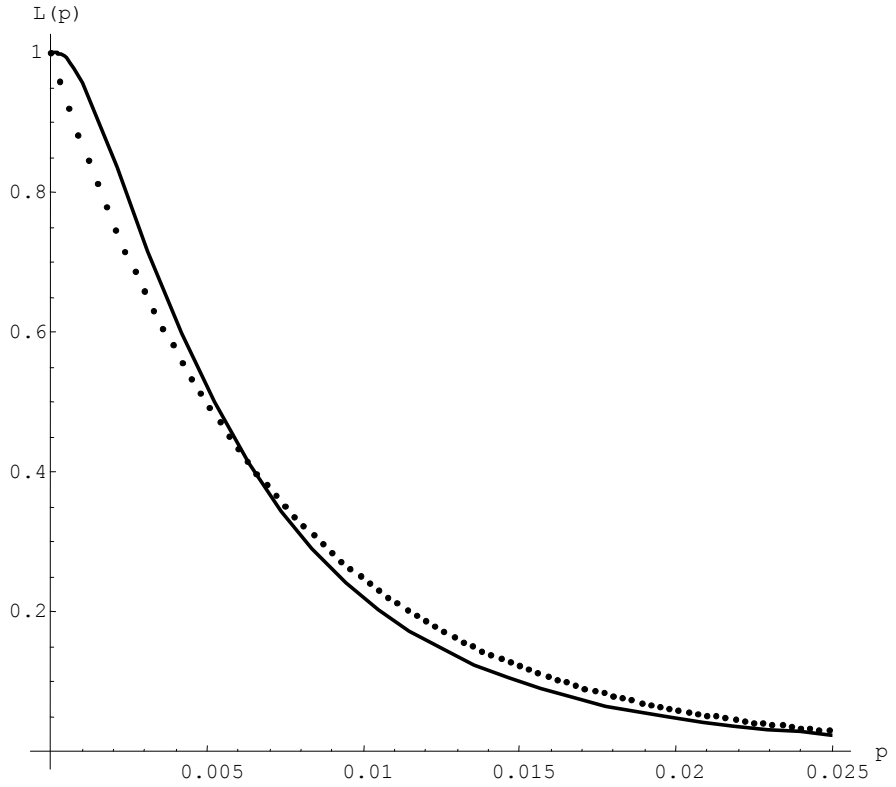


Figure 2. OC curves for the AOQL sampling-plans for inspection by variables and attributes (49, 2.57617) ——— for inspection by attributes (130,0) .....

References

[1] Dodge H.F.–Romig H.G.: Sampling Inspection Tables. John Wiley, New York, 1998.

[2] Klůfa J.: Dodge-Romig AOQL single sampling plans for inspection by variables. Statistical Papers 38, 1997, 111 – 119.

[3] Klůfa J.: Economical Aspects of Acceptance Sampling. Ekopress, Prague,1999.

[4] Klůfa J.: Dodge-Romig AOQL plans for inspection by variables from numerical point of view. Statistical Papers 49, 2008, 1-13.

[5] Wolfram S.: Mathematica. Addison-Wesley, 1991

Regime-switching models based on Aggregation functions

Jozef Komorník[1] and Magdaléna Komorníková[2]

[1] Faculty of Management, Comenius University
Bratislava, Slovakia
(e-mail: Jozef.Komornik@fm.uniba.sk)
[2] Faculty of Civil Engineering, Slovak University of Technology
Bratislava, Slovakia
(e-mail: Magdalena.Komornikova@stuba.sk)

Abstract. A synthesis of recent development of regime-switching models based on aggregation functions is presented together with its extension and interesting applications. It comprises traditional procedures for model specification and identification, parameter estimation and model adequacy testing. In our paper, we put stress on the use of the outputs of aggregation functions in the role of threshold variables of the regime–switching time series models. The practical modelling results indicate that such models provide promising methods for analysis of various economic and financial time series (in comparison with traditional SETAR and STAR models).
Keywords: Aggregation function, BDS test of independence, Regime-switching model.

1 Introduction

The aim of this paper is to sum up and extend a series of studies of its authors and some of their collaborators on regime-switching models. A special attention will be paid to threshold variables and their construction via aggregation operators. Some further problems related to comparison of these models with traditional regime–switching models will be investigated and discussed.

The paper is organized as follows. The second section is devoted to a brief theory of the regime–switching models, the aggregation functions and the BDS test. Because the limited space, we only briefly indicate the utilized estimation and testing procedures for regime–switching models that can be found in details e.g. in Franses and Dijk [2]. The third section contains an application part of the paper. Finally, some conclusions are presented.

2 Theoretical background

2.1 Regime–switching models with regimes determined by observable variables.

We utilize the regime–switching models using also selected aggregation functions for constructions of their threshold functions.

Let us recall (Franses and Dijk [2], Tong, 1978 [10], Tong, 1990 [11]) that **T**hreshold **A**uto**R**egressive models (TAR) with 2 regimes for a time series $y_t,\ t = 1, \ldots, n$ with AR(p_1) and AR(p_2) and observable *threshold variable* q_t (that determines one of 2 regimes relative to a *threshold value* c) have the form

$$y_t = \Phi_1 \cdot \mathbf{X}'_t + (\Phi_2 - \Phi_1) \cdot \mathbf{X}'_t * I[q_t > c] + \epsilon_t, \tag{1}$$

where $\Phi_i = (\varphi_{0,i}, \cdots, \varphi_{p,i})$, $i = 1, 2$ are autoregressive coeficients, $\mathbf{X}_t = (1, y_{t-1}, \cdots, y_{t-p})$, $I[A]$ is an *indicator function* and ϵ_t's is the strict white noise process with $E[\epsilon_t] = 0,\ D[\epsilon_t] = \sigma_\epsilon^2$ for $t = 1, \cdots, n$. We will use the notation $p = \max(p_1, p_2)$.
A direct generalization to 3–regimes TAR models (with 2 threshold constants $c_1 < c_2$) has been described e.g. in Franses and Dijk [2].

The class of **S**elf-**E**xciting TAR (SETAR) models is a special case of TAR models, when the threshold variables q_t are taken to be lagged values of the time series itself: $q_t = y_{t-d}$ for certain fixed integer $d > 0$ (see e.g. Hansen, 1997 [4], Hansen, 2000 [5], Franses and Dijk [2], Chen and Tsay [6]).

TAR models can be considered as a limit cases of so–called **S**mooth **T**ransition AR (STAR) model (see, e. g. Franses and Dijk [2], Teräsvirta [9]), where the transition between different regimes can be obtained by a continuous function $G(q_t, \gamma, c)$, which changes smoothly from 0 to 1. The resulting 2–regimes model with AR(p_1) and AR(p_2) have the form

$$y_t = \Phi_1 \cdot \mathbf{X}'_t + (\Phi_2 - \Phi_1) \cdot \mathbf{X}'_t * G(q_t, \gamma, c) + \epsilon_t, \tag{2}$$

We will apply two classes of the so-called *transition function* $G(q_t, \gamma, c)$:

1. *The logistic function* $G(q_t, \gamma, c) = \frac{1}{1+e^{-\gamma(q_t - c)}},\ \gamma > 0$. The resulting model is called a *Logistic STAR* (LSTAR) model.
2. *The exponential function* $G(q_t, \gamma, c) = 1 - e^{-\gamma(q_t - c)^2},\ \gamma > 0$ and the resulting model is called an *Exponential STAR* (ESTAR) model.

The parameter γ in (2) determines the *smoothness* of the change in the value of the transition function.
Generalizations to m–regimes STAR models have a form:

$$\begin{aligned} y_t = \Phi_1 \cdot \mathbf{X}'_t + (\Phi_2 - \Phi_1) \cdot \mathbf{X}'_t * G(q_t, \gamma_1, c_1) + \cdots + \\ +(\Phi_m - \Phi_{m-1}) \cdot \mathbf{X}'_t * G(q_t, \gamma_{m-1}, c_{m-1}) + \epsilon_t, \end{aligned} \tag{3}$$

where $p = \max(p_1, p_2, \cdots, p_m),\ \Phi_i = (\varphi_{0,i}, \cdots, \varphi_{p,i}),\ i = 1, 2, \cdots, m,\ c_1 < c_2 < \cdots < c_{m-1},\ \gamma_i > 0,\ i = 1, \ldots, m-1$.
We will use two alternative approaches to the construction of the threshold variable q_t. The first one is a traditional self–exciting (SE), where $q_t = y_{t-d}$

for a suitable time delay $d > 0$. The second approach is using aggregation functions (AF) and trying to model q_t in the form $q_t = \mathcal{A}(y_{t-1}, \ldots, y_{t-k})$. We determine the integer $k > 0$ of inputs in the aggregation function according to the results of the BDS test of independence (see Brock et al. [1]).

2.2 Aggregation functions

In this section we briefly recall the basic concept of the aggregation functions. For a thorough study of the aggregation functions, we recommend e.g. the book by Grabish *et. al.* [3].

Let $\mathcal{I}$ be a nonempty real interval, k an integer as the number of variables and $\mathbf{y} = (y_1, ..., y_k) \in \mathcal{I}^k$. An aggregation function in $\mathcal{I}^k$ is a function $\mathcal{A}^{(k)} : \mathcal{I}^k \to \mathcal{I}$ that

(i) is nondecreasing in each variable,
(ii) fulfills the boundary conditions $\inf_{\mathbf{y}\in\mathcal{I}^k} \mathcal{A}^{(k)}(\mathbf{y}) = \inf \mathcal{I}$ and $\sup_{\mathbf{y}\in\mathcal{I}^k} \mathcal{A}^{(k)}(\mathbf{y}) = \sup \mathcal{I}$.

Typical continuous aggregation functions on the real line are

- the arithmetic mean: $\mathcal{AM}(\mathbf{y}) = \frac{1}{k}\sum_{i=1}^{k} y_i$;
- the weighted mean: $\mathcal{WA}(\mathbf{y}) = \frac{1}{k}\sum_{i=1}^{k} w_i\, y_i$, where $w_i \in [0, 1]$, $\sum_{i=1}^{k} w_i = 1$;
- the OWA operators: $\mathcal{WA}\prime(\mathbf{y}) = \frac{1}{k}\sum_{i=1}^{k} w_i\, y'_i$ with weights as in the case of weighted means, but with y'_i as non-decreasing permutation of y_i inputs, i.e., $y'_1 \leq \ldots \leq y'_k$.

In the class of OWA operators we can find the *MIN* and *MAX* operators, corresponding to extremal cases $w_1 = 1$ and $w_i = 0$ otherwise, resp. $w_k = 1$ and $w_i = 0$ otherwise.

2.3 BDS test of independence

In order to determine the orders $k > 0$ of the suitable AFs we utilize the BDS test for independence of the pairs of (y_t, y_{t+i}), $i \geq 1$. The BDS test is a powerful tool for detecting serial dependence in time series. It tests the null hypothesis, that a time series sample comes from i.i.d. process against an unspecified alternative. This portmanteau test based on the correlation integral was presented in the paper by Brock et al. [1].

2.4 Empirical specification procedure for nonlinear models

Inspired by the recommendation from Franses and Dijk [2] we have proceeded in the following steps:

1. specify an appropriate linear AR(p) model for the time series under investigation
2. test the null hypothesis of linearity against the alternative regime-switching nonlinearity
3. estimate the parameters in the selected nonlinear model
4. evaluate the model using diagnostic tests
5. modify the model if necessary

6. use the model for descriptive or forecasting purposes.

When we test the null hypothesis of linearity against the alternative of SETAR-type nonlinearity, we need to know estimation of parameters of the nonlinear model. Therefore, in this case we start a specification procedure for the model with estimation of parameters.

3 Application to real data modelling

We focused on the application of the modelling procedures described above to the residuals obtained after removing trend and periodical components from the original data, in our case 15 time series (macroeconomic indicators, exchange rates and other economic and financial time series). All data used in this paper can be downloaded from the pages of National Bank of Slovakia and European Central Bank. For 5 of them we present some details of the results:
No. 1. Quarterly mean nominal incomes in the Slovak Republic (SR) in EUR in the period 1991 – 2011;
No. 2. Monthly exchange rates EUR/USD, January 1999 - February 2012;
No. 3. Monthly exchange rates EUR/ChF, January 1999 - February 2012;
No. 4. Monthly exchange rates EUR/GBP, January 1999 - February 2012;
No. 5. Monthly oil prices, January 2002 - March 2011.

All calculations was performed using the system *Mathematica*, version 8.

The basic descriptive statistics of the time series No. 1 – 5 are in Table 1.

time series	Number of data	Average value	Standard deviation
No. 1.	84	442.704	215.531
No. 2.	158	1.206	0.194
No. 3.	158	1.509	0.111
No. 4.	158	0.721	0.097
No. 5.	99	26.003	15.826

Table 1. Descriptive statistics of the selected data

For all considered time series we left out the most recent 5 values for the assessment the predictive qualities of the models that we measured by the criteria RMSE (Root Mean Square Error) and MAE (Mean Absolute Error) for one–step–ahead predictions.

For the models based on the AF approach we first determine the number of inputs in aggregation functions. When applying BDS test, we utilise so called thin sampling (see Rakonczai [8]). We choose only every s–th pair from this set of pairs due to possible dependences between them. A hint for the choice of the thinning can be obtained from the autocovariance function. The new

thinned set of time points is denoted as $\mathcal{T} = \{1, s+1, 2s+1, ..., rs+1 \leq n\}$, where $|\mathcal{T}| = r+1$ is the new thinned sample size. Then we use for testing the pairs $\{(y_t, y_{t+i}), t \in \mathcal{T}\}$. We have selected the value of $k > 0$ (number of arguments for AFs) equal the largest value of i for which the independence of y_t and y_{t+i} has been rejected.

For both above approaches (SE and AF models) and for each model, we consider 2, as well as 3 regimes. First we test linearity against the alternative of a regime-switching type of nonlinearity (SETAR, LSTAR, ESTAR) with the threshold variable in SE or AF form. We have chosen (for each time series and each model's type) the models with a suitable number of the smallest p-values. For these models we have applied conditional ordinary least squares (OLS) estimates of regression parameters (separately for each considered regime) and the corresponding conditional residual variance. Further (in case of SE threshold variable) we select the optimal form of the threshold variable by minimizing the Bayes's information criterion BIC (Franses and Dijk [2], Liew and Chong [7]). Finally, the residuals were tested for serial correlations and remaining nonlinearity. For each of the time series, model's type and both SE and AF approaches we calculated 20 best models according to the BIC. In each of these groups we selected 5 best models according to the RMSE criterion for prediction.

The following Table 2 contains p–values for the BDS tests of independence for the pairs of time series (y_t, y_{t+i}) for the time series No. 1 (leading to the conclusion $k = 4$).

i	1	2	3	4	5
p–value	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$	0.399

Table 2. Results of BDS for the Mean nominal incomes SR

Table 3 contains p-values of LM-type tests against SETAR (for SE approach) and ESTAR (for AF approach) type of nonlinearity as well as the tests for serial correlation and remaining nonlinearity for time series No. 1.

Model type		**p-value for test**		
Approach	**type**	**linearity**	**serial correlation**	**remaining nonlinearity**
SE ($d = 2$)	ESTAR	0.002	0.147	0.454
AF=Min ($k = 4$)	SETAR	$< 10^{-6}$	0.274	0.013

Table 3. Results of test of linearity, serial correlation and remaining nonlinearity for the Mean nominal incomes SR

Table 4 presents the characteristic of the best models for the SE– and AF–approaches for the time series No.1 (m = number of regimes).

Approach	**Type**	**m**	**p**	**c**	γ	**RMSE**	**MAE**	σ_{res}
SE ($d = 2$)	ESTAR	2	5	725	0.5	23.77	20.57	28.07
AF=Min ($k = 4$)	SETAR	3	5	391.2; 526.2	x	6.52	6.02	9.27

Table 4. The SE– and AF–approach models with minimal RMSE for the Mean nominal incomes SR

In the Figure 1 and Figure 2 we can observe values of the time series No. 1 (black) and their estimates by SE approach model (red) and AF approach model (blue) for the modelled time period and the 5 last values and their one–step–ahead predictions according both models.

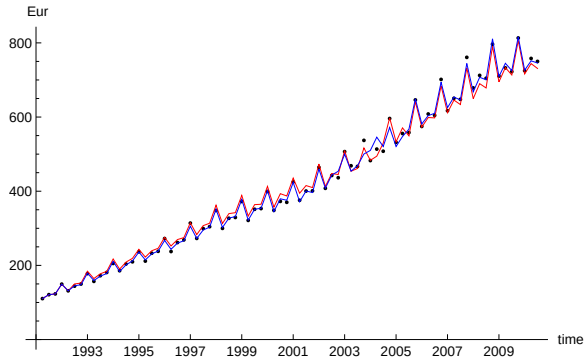


Fig. 1. The Mean nominal incomes SR (black) and their estimates by SE approach model (red) and AF approach model (blue)

In the next Table 5 we summarize the results for the best models for both SE– and AF–approaches for the time series No. 2, 3, 4, 5.

We can observe that for all time series No. 1 – 5, the values of all 3 variables RMSE, MAE and σ_{res} are reduced for the best AF–approach models in comparison to the corresponding values for the best SE–approach models. In the next Table 6 we present the sizes of these reductions (in %).

4 Conclusion

From the above results, we can conclude that using suitable aggregation functions for constructions of threshold variables can lead to regime–switching models of improved qualities.

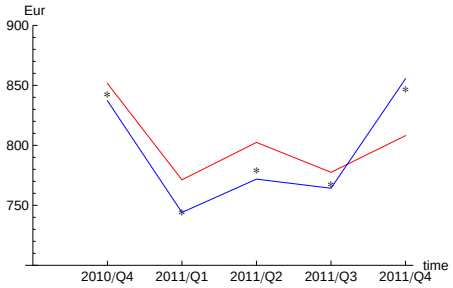


Fig. 2. The 5 last values of the Mean nominal incomes SR (black) and their one–step–ahead predictions by SE approach model (red) and AF approach model (blue)

Approach	**Type**	**m**	**p**	**c**	γ	**RMSE**	**MAE**	σ_{res}
Exchange rates EUR/USD								
SE ($d = 3$)	LSTAR	2	4	1.31	5	0.018	0.015	0.020
AF=Min ($k = 10$)	LSTAR	2	10	1.29	10.5	0.011	0.010	0.013
Exchange rates EUR/ChF								
SE ($d = 3$)	LSTAR	2	4	1.31	5	0.018	0.015	0.020
AF=Min ($k = 10$)	LSTAR	2	10	1.29	10.5	0.011	0.010	0.013
Exchange rates EUR/GBP								
SE ($d = 2$)	ESTAR	3	2	0.61; 0.67	4.5; 0.5	0.025	0.021	0.015
AF=Min ($k = 10$)	ESTAR	2	10	0.70	0.5	0.018	0.017	0.011
Oil prices								
SE ($d = 1$)	LSTAR	3	2	36.38; 52.31	2.5; 4.5	4.88	4.57	3.019
AF=Max ($k = 6$)	SETAR	2	7	52.57	x	4.67	4.07	2.750

Table 5. The SE– and AF–approach models with minimal RMSE for time series No. 2 – No. 5

Time series No.	**RMSE**	**MAE**	σ_{res}
1	72.57	70.72	64.42
2	69.30	67.65	0.22
3	36.11	28.95	29.99
4	25.06	15.17	22.39
5	4.12	10.90	8.72

Table 6. Reductions of RMSE, MAE and σ_{res} for the best AF–approach models in comparison with the best SE–approach models (in %)

Let us add that the average improvement (through all 15 investigated time series) of AF–approach optimal models in comparison with corresponding SE–approach models was 26.14% for RMSE, 25.32% for MAE and 20.33% for σ_{res}.

Concerning the individual types of the aggregation functions for the optimal AF–approach models for individual time series the most frequent was Min and the less frequent Mean.

Acknowledgement: This work was partially supported by the grants APVV No. LPP-0111-09 and VEGA 1/0143/11.

References

1. Brock, W. A., Dechert, W. D., Scheinkman, J. A., and Le Baron, B., "A test for independence based on the correlation dimension", *Econometric Reviews*, 15, 197–235 (1996).
2. Franses, P.H., Dijk, D., *Non-linear time series models in empirical finance*, Cambridge University Press (2000).
3. Grabisch, M., Marichal, J.-L., Mesiar, R., and Pap, E., *Aggregation Functions*, Cambridge University Press, Encyclopedia of Mathematics and its Applications, No 127 (2009).
4. Hansen, B. E., "Inference in TAR models". *Studies in Nonlinear Dynamic and Econometrics* 2, 1 – 14 (1997).
5. Hansen, B. E., "Sample splitting and threshold estimation". *Econometrica* 68, No.3, 575–603 (2000).
6. Chen, R., Tsay, R.S., "On the ergodicity of TAR(1) processes", *Ann. Appl. Probab.* 1, 613–634 (1991).
7. Liew, V.K., Chong, T.T., "Effect of STAR and TAR types nonlinearities on order selection criteria", http://econwpa.wustl.edu:80/eps/em/papers/0307/0307005.pdf (2003).
8. Rakonczai, P., "On modeling and prediction of multivariate extremes with applications to enviromental data", *Licentiate Theses in Mathematical Sciences* 3, ISSN 1404-028X, 83–109 (2009).
9. Teräsvirta, T., "Specification, estimation, and evaluation of smooth transition models", *Journal of American Statistical Association* 89, 208 – 218 (1994).
10. Tong, H., "On a threshold model", in: *Pattern Recognition and Signal Processing*,C. H. Chen (ed.), Amsterdam, 101 – 141 (1978).
11. Tong, H., *Non-linear time series: A dynamical systems approach*, Oxford University Press, Oxford (1990).

Improving of Railway Information Analytical Systems Using Newest Databases Technologies

Eugene Kopytov, Vasilijs Demidovs, Natalia Petukhova

Transport and Telecommunication Institute
E-mail: kopitov@tsi.lv, Riga, Latvia
State Join-Stock Company "Latvian Railway"
E-mail: dem@ldz.lv, natalia@ldz.lv

Abstract: The work presents the results of the research dedicated to the enhancement of Information Analytical Systems (IAS) on the railway transport. The paper demonstrates the solutions for some important practical tasks: the bulk data processing system designing, the development of the designing fundamentals of the data warehouses in the decision support systems on the railway. The special attention is paid to the problem of the temporal data employment in the warehouses. The correctness of the proposed models and methods has been tested in various tasks of the IAS of the Latvian Railway.
Keywords: information system, railway transport, decision support system, data warehouse, temporal data, interval form, point form.

1 Introduction

Transport in the Latvian Republic is an important economic field of the country, which together with the industry of communication provides about 12,1 % of the state national gross output (Latvian Statistical book 2011). The Latvian Railway as a part of the transport industry of Latvia is a large and complex system and has some specific features, which make a sufficient impact on to the efficiency of its performance. Among them we should note the following: high dynamics of processes; random factors; high reliability and safety; large financial, labour and material resources; large distances between related objects; a complex, far-flung railways network. Therefore, searching optimal decisions for the forthcoming periods of railway functioning is quite a complicated job, the received result depending greatly on the completeness and trustworthiness of the required source data provided by different Information Systems (IS). At the present time the Latvian Railway employs sixty IS, which are divided into five big groups shown in figure 1: cargo transportation (KRAIS), passenger transportation (PASS), business management and support (BVA), economics and finances (SAP), infrastructure and trains traffic control (INIS). The information generated by these systems allows to evaluate the efficiency of performance and to exercise prospective planning. But these systems are mainly oriented towards performing account-controlling and enquiries functions and the final user doesn't actively interact with them in the course of making optimal managing decisions. And most of railway IS programs have been developed on the basis of outdated technologies, and they are difficult to support and can work only with small volumes of data. As an example a structure of information support of a cargo transportation system is

presented in figure 1. This group KRAIS combines six large (APOVS, APIKS, etc) and many small IS developed at different times on different program and apparatus basis. These IS are bulk data processing systems and generate ten-million transactions per year. The data received from these IS have different format and degree of aggregation.

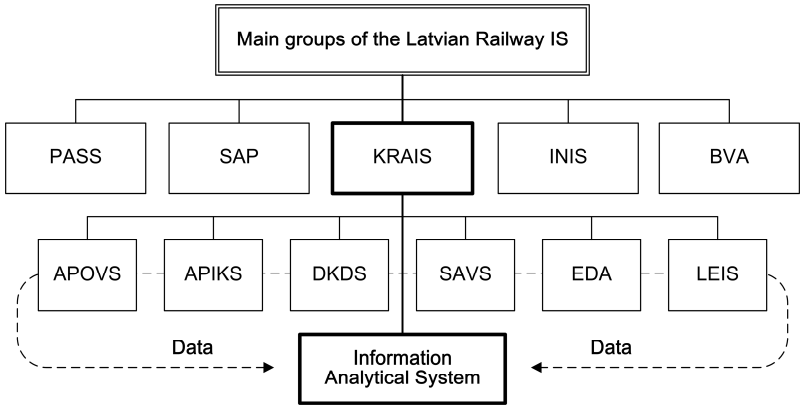


Fig. 1. The structure of the Latvian Railway IS

Effective complex planning of the railway transport enterprise activity is hampered due to the following problems: non-homogeneity of the employed information support; huge data volumes; complicacy of simultaneous application of the same data of IS for resolving the tasks of decision-making; incompleteness and errors in source data decrease of the information reliability. The methodology of building and improving the Decision Support Systems (DSS) in the railway suggested by the authors implies synergy of modern technologies and approaches in the sphere of databases and their integration in the existing technologies. In spite of the advantages of the data warehouses, Data Marts, intellectual data processing, temporal databases technologies, in developing DSS on the Latvian Railway the authors have encountered a number of problems, which solving required research and developing new methods and approaches (Kopytov et al. 2002, 2011). Some results of the research are given in the sections below.

2 Data Warehouses application in DSS on the Latvian Railway

To analyze the dynamics of the situation change and to reveal the regularities, it is necessary to keep the collected data for a long time providing an easy access to them. To provide effective management of an enterprise on the whole, we need DSS using the aggregated data of all enterprise IS. It's more reasonable to build modern DSS on the basis of data warehouse (DW) where big volumes of data from different sources are accumulated. W. Inmon (2005) in his fundamental work defined DW as the foundation of DSS. W. Inmon was the first to give the definition of the DW term as subject-oriented, integrated, constant, supporting the data set chronology, devised for the decision-making support. The questions of

data modelling for DW have been considered in detail by R. Kimball (2002). DW as compared with databases of On-Line Transaction Processing (OLTP) systems most fully describes the system functioning, it contains historically interrelated data of several OLTP systems life through the whole period of their existence as well as the data from outside sources.

A transport enterprise can be defined as a complicated system represented by the aggregation of all its IS. To analyze this system functioning we need to have its real model, which is built on the basis of the data in DW received from different sources, as it is shown in figure 2.

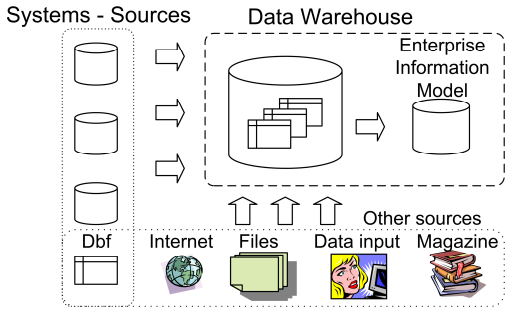


Fig. 2. Conceptual scheme of the Information Analytical System

The principles of building DW suggested in the paper are aimed at improving the data trustworthiness employed in DSS for analyzing and predicting the Latvian Railway company activities. Recently DSS has stopped being the prerogative of the certain layer of managers and has a form of pyramidal model (see figure 3), and, therefore, the increased requirements of accessibility and data safety and reliability are laid down to it. Reliability of prediction depends on the reliability of data stored in DW, which architecture is based on the schemes type “star” and “snowflake”. It is peculiar for these schemes to have tables of factors where all transactions and all aggregates of transactions are described, as well as tables of measurements for each entity. Here, the concept of transaction can differ from the similar concept in the initial data obtained from On-Line Analytical Processing (OLAP) systems (see Korovkin et al.). The used methods give sufficient enhancement of DW capabilities and allow to get from the stored data more detailed information about both the past and the present situation of the company and about the possible future situation. The advantageous aspect of the suggested approach is that the DW is not locked in the process of forecasting and remains accessible both to the users and to some forecast processes. Loss and distortion of data is excluded.

One of the obligatory measurements in AIS is a measurement of time. Thus, a time indicator is present in the data prepared for analysis. This time indicator is formed periodically once per week, decade, month, quarter or year during the data processing. Forming of data for the certain period is performed on the basis of Referenced Data System (RDS). Data situated in RDS are the subjects to continual

changes, for example, old trains are cancelled and new ones are appointed, schedule and routes of trains are changed. The preparation of data for the long period of time demands registration of all changes in RDS for the processed period. When there is a necessity of the creation of the new physical data model, which is impossible to receive on the basis of the existing ones, we shall need all initial data. Therefore, for the flexibility of the system in DW it is necessary to store data received from the initial systems and transformed into one form. Therefore, time factor reflected in temporal model RDS should be taken into account (see Date et al. 2002, Jensen 2000).The decision of the problem of using "historical" data in OLTP systems and IAS is seen by the authors in employing the principles of temporal databases, which supplement the stored data with the time characteristic and have temporal superstructure. For the storage of historical data, the authors have suggested the relational databases with temporal extensions (Kopytov et al. 2002). The tasks of creating a data model involving the time factor and developing a time logic instrument including components of temporal data management and functions of the access to multiversional objects are performed.

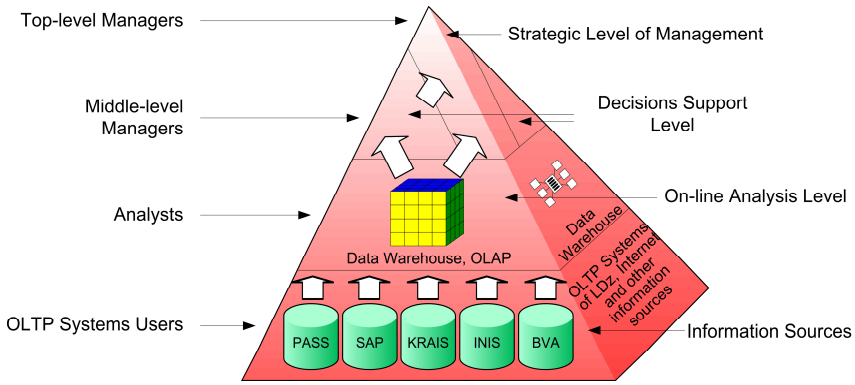


Fig. 3. Pyramidal Model of DSS at Latvian Railway

3. Temporal Data Presentation

Let us consider the variant at which temporal data measurement is taken into account not only in the DW but in the source database of an operative IS as well. This variant is pretty popular since management of some particular objects is impossible without accounting of their temporal measurement, for example, trains schedule. In this case temporal data on these objects are presented in the database in one of temporal forms. Let us consider some familiar forms of temporal data in terms of providing data integrity in the course of loading into the warehouse, storing and using as the source of data for analysis. Classification of temporal forms given in figure 4 distributes popular forms according to the principle of the temporal measurement presentation (we denote: R is the name of relation; OID is the object abstract identifier; A is the set of attributes; VT is valid time interval, TT is transaction time interval, t is the date when the fact is true; t_f is the time of the

transaction fixation). The primary importance is whether time is presented by an interval (interval form) or a discrete point (point form).

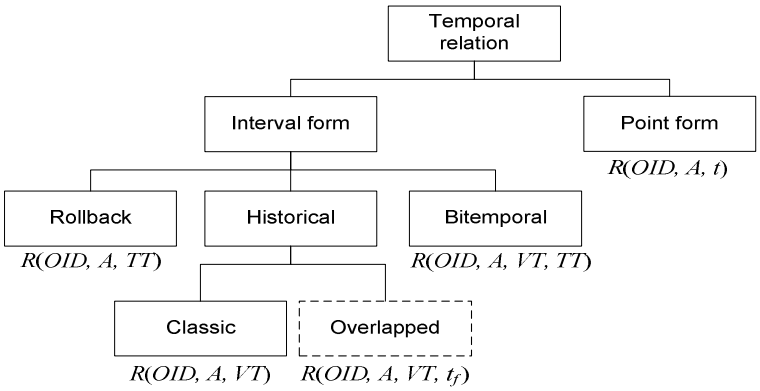


Fig. 4. General classification of the temporal forms

Let us illustrate presentation of the railway schedule data by both the interval and the point forms. An example of data presentation in point form is shown in Table 1.

Table 1. Train schedule: data presentation in point form

Train ID	Station	Time	Date
…	…	…	…
4387	Riga	17:00	07.03.2011
4387	Riga	17:00	08.03.2011
…	…	…	…
4387	Riga	16:59	14.04.2011
4387	Riga	17:00	15.04.2011
4387	Riga	17:20	16.04.2011
4387	Riga	17:20	17.04.2011
4387	Riga	17:00	18.04.2011
4387	Riga	16:59	19.04.2011
…	…	…	…

The data presented in *the point form* are very convenient to perform analytical operations on them but this form has three problems. Firstly, this is a big volume of data; secondly, without introducing additional attributes, the operation of synchronizing such elements with the storage is likely to be pretty cumbersome since it will be necessary either to compare or to replace the whole set of data presented in the source; and thirdly, this set of data does not contain any information on the introduction of new schedule versions (simple change of the time of the train departure does not contain such information).

Since the schedule is usually made for a long period of time (i.e. an entire season), it is more convenient to use *the interval form* of data presentation. This form allows substantial reduction of the volume of stored data. For the same example using interval form, it is necessary to enter only 3 records shown in Table 2, where the attributes of the valid time t_s^{VT} and t_e^{VT} define respectively the beginning and the end of the period *VT* of the schedule validity. However, the interval form is inferior

to the point form in processing schedule queries and processing analytical queries in particular.

Table 2. Train schedule: data presentation in interval form (Historical–Classic)

Train ID	*Station*	*Time*	t_s^{VT}	t_e^{VT}	*Periodicity*
4387	Riga	17:00	01.03.2011	31.05.2011	Work days
4387	Riga	17:20	01.03.2011	31.05.2011	Days-off
4387	Riga	16:59	12.04.2011	18.05.2011	Tue, Thu

It has been mentioned that this form has some weak points concerning synchronization with the DW. The set of records in the source database will be constantly changing which means, that for synchronization we'll have either to develop a complicated procedure of comparing two data sets with the further replacement of the modified fragments or to replace the whole set of data completely which is not a good decision either. We suggest the special temporal form without the above mentioned drawbacks. It is the relation form with the overlapping lifespans Historical-Overlapped (figure 4, the dotted line). It comprises the principal properties of the classic temporal relations forms, exactly the possibility of existence of the several versions of the same object, becoming the active one consecutively. At the same time the investigated form permits the object versions to become active more than one time.

The logic of interaction with the relation of the Historical-Overlapped form is explained on the example of the train traffic schedule, described by the diagram presented in figure 5. It is assumed that t_n means the time of observation. The train schedule (the version v_1) is assigned from t_1 to t_4. Later for the period of time from t_2 to t_3 the schedule alteration (the version v_2) is introduced. In case of the query on the train traffic at the moment of time t_{n1}, there will be delivered the version of the schedule v_1, and at the moment t_{n2} – the schedule version v_2, and at the moment t_{n3} – again the version of the schedule v_1. It is important that only two tuples exist in this relation: tuples with the schedule versions v_1 and v_2. The active version of the schedule is calculated on the basis of t_f. The version that has been introduced the latest has the priority in case there is a choice comprising several versions.

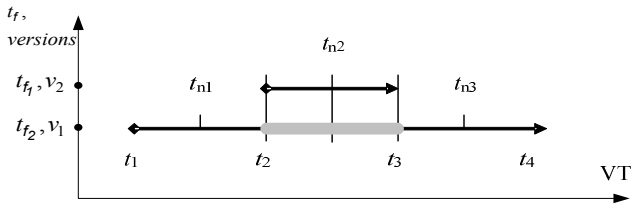


Fig.5. Overlapping the life cycles of the object

The efficiency of implementing the suggested temporal form Historical-Overlapped is examined from the position of the support of data integrity in the process of the data synchronisation in DW. Let us illustrate that on example of the task of storing the schedule of the train departure from the station. The schedule of the traffic of the majority trains possesses the prolonged nature but as a result of

the various situations there are possible pre-planned and unplanned changes of the departure time. The task is the following one: at any moment of time on any day in the past or in the future to determine the active for this day departure time of the specified train from the station under observation. Figure 6 presents the example of the scenario of altering the train schedule. This example is considered in the details. The core fundamental seasonal schedule *(Base)* is assigned for the train number 123P for the summer period from 01.05 to 30.08. Afterwards in the connection with the necessity of conducting the repair works for the period from 03.07 to 15.07 the train schedule was changed (*Repair works*). And on the City Anniversary 06.07 the special timetable was assigned (*City anniversary*).

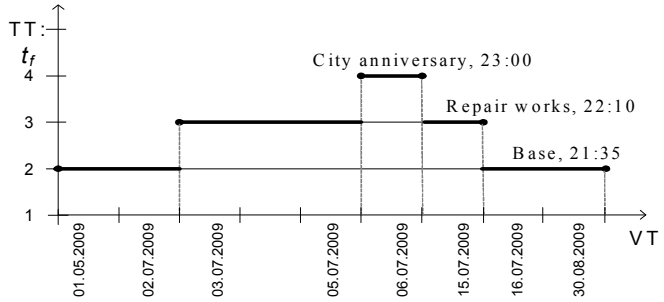


Fig. 6. Example of overlapping of object lifespans in the train schedule

Figure 7 demonstrates two ways of train departure time storing of the above example: in classical relation Historical-Classic with non-overlapping lifespans and in the relation Historical-Overlapped with the overlapping lifespans.

Historical-Classic

OID Train	*Departure time*	t_s^{VT}	t_e^{VT}
123P	21:35	01.05.2009	02.07.2009
123P	22:10	03.07.2009	05.07.2009
123P	23:00	06.07.2009	06.07.2009
123P	22:10	07.07.2009	15.07.2009
123P	21:35	16.07.2009	30.08.2009

Historical-Overlapped

OID Train	*Departure time*	t_s^{VT}	t_e^{VT}	t_f
123P	21:35	01.05.2009	30.08.2009	1
123P	22:10	03.07.2009	15.07.2009	2
123P	23:00	06.07.2009	06.07.2009	3

Fig. 7. Train schedule in the forms Historical-Classic and Historical-Overlapped

In the considered example the relation Historical-Overlapped is more *economical* from the point of view of the amount of tuples. It is not difficult to define that, in case the number of tuples (versions) in relation Historical-Overlapped equals g, the minimal number of tuples in the relation of the Historical-Classic form, reflecting the same information, will be in the diapason from g to $2g$-1 depending on the situation. The greater amount of the overlapped factors and consequently the greater number of cases of coming into operation of one and the same version, the greater is the benefit of employment the Historical-Overlapped form. It also should be mentioned that the relation Historical-Overlapped expresses the semantics of the task without introduction of the additional attributes. The benefit gained in the process of the semantics reporting is achieved due to the fact that the tuple in the

relation Historical-Overlapped represents additionally the factor of the exchange appearance, but not only the order of the schedules alteration as it happens in the relation Historical-Classic.

There considered the support of *the integrity in the process of the data synchronisation*. The principle of managing the versions of the objects in the case of objects changing in the relation Historical-Overlapped differs from the management in the relation Historical-Classic. The distinctive peculiarity of the Historical-Overlapped form lies in the fact that the procedure of processing the values of the temporal attributes of the adjacent (the previous and/or the following) object versions do not precede the fact of appearing the new object version or the fact of altering the existing object version. On the contrary, the creation of the new object version or the modification of the existing one in the relation Historical-Classic are connected with the processing the boarders of the lifespans of other versions, in case the lifespans of theses versions are overlapped fully or partially as a result of the modification (in the majority of cases it refers to the previous and to the following versions of the object). Thus, the records (in the source database and in the DW) presented in the Historical-Classic form are being changed and those presented in the Historical-Overlapped form are only being added while every record is marked by the time of addition which is very convenient for the procedure of synchronization. There is the comparison of the procedures of adding the new version of the object for these two relation forms. Three situations can appear in the process of the versions adding shown in figures 8-10.

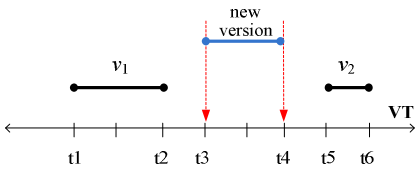


Fig. 8. New version does not overlap any of the existing ones in valid time

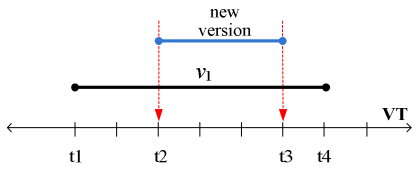


Fig. 9. Lifespan of the new version is completely within the lifespan of another version

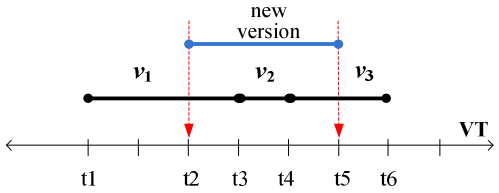


Fig. 10. Lifespan of the new version covers the lifespans of other versions

The results of the comparison of the temporal table data change after adding the new version of the object for the relations Historical-Classic and Historical-Overlapped are presented in Table 3. They are the activities that are to happen obligatory during adding the new version in the temporal relation of the above mentioned forms.

Table 3. Data change in the temporal table after adding a new version of the object

Temporal Form / *Situation*	***Historical-Classic***	***Historical-Overlapped***
Situation 1, shown in figure 8	1 new tuple is added	1 new timestamped tuple is added
Situation 2, shown in figure 9	1 tuple is changed and 2 tuples are added	1 new timestamped tuple is added
Situation 3, shown in figure 10	1-2 tuples are changed, d tuples are removed and 1 new tuple is added	1 new timestamped tuple is added

Therefore, the procedure of synchronization of the data in the Historical-Overlapped form comes down to a trivial addition of new records which appear in the source database after the last synchronization.

Conclusions

The authors have considered of data warehouses application in IAS on the railway transport. A special attention is paid to the problem of the temporal data employment in DW. The given paper suggests the Historical-Overlapped form which possesses better indicators of data loading into the data warehouse (one of the basic processes in DW) and contains more information useful for analysis than the interval Historical-Classic and point forms.

References

1. Date, C.J., H. Darwen, N.A. Lorentzos. *Temporal Data and the Relational Model*. Morgan Kaufmann (2002).
2. Inmon, W. *Building the Data Warehouse*, 4th Ed., NY, John Wiley& Sons (2005).
3. Jensen C. *"Temporal Database Management"*. Dr. techn. thesis, defended on 14.04.2000, Available at: http://www.cs.auc.dk/~csj/Thesis/
4. Kimball R., Ross M. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modelling*. 2nd Ed. NY, John Wiley & Sons (2002).
5. Kopytov E., Demidovs V., Petukhova N. Data Presentation and Transformation in Train Schedule Information Systems. *Proceedings of the 10th International Conference on Modeling and Applied Simulation*, Rome, 216-225 (2011).
6. Kopytov E., Demidovs V., Petukhova N. Method of Temporal Databases Design Using Relational Environment. *Scientific Proceedings of Riga Technical University. Computer Science: Applied Computer Systems*. 5,13, 236-246 (2002).
7. Korovkin S., Levenets I., Romanov I., and etc. Resolving of Problem of Complex On-Line Analysis of Information in Data Warehouse. Available at: http://www.citforum.ru/database/articles/art_11.shtml
8. *The Latvian Statistical Yearbook*, Latvian Central Statistical Bureau (2011).

Analytical Data Processing of the Results of Experiments on Producing the Multi-component Nanostructured Protective Coating

Eugene Kopytov, Alexander Urbach, Sergey Yunusov, Konstantin Savkov

Transport and Telecommunication Institute, Riga, Latvia
Email: kopitov@tsi.lv
Riga Technical University, Riga, Latvia
Email: Aleksandrs.Urbahs@rtu.lv
Transport and Telecommunication Institute, Riga, Latvia
Email: yunusov@inbox.lv
Riga Technical University, Riga, Latvia
Email: sakon@inbox.lv

Abstract: The paper considers the analytical processing the data which have been received as the results of experiments on applying the nanostructured protective covering on different materials. There has been developed the information system for saving the experiments results and their following processing. The analytical module of the information system is supposed for solving the tasks of determining the thickness of covering depending on the specified mode of plant operation (pressure, voltage, current); it is also employed for determining what operation mode is required for receiving the covering with certain designated properties and parameters. Consequently, the task of scientific prediction of the covering properties is solved.
Keywords: nanotechnology, information system, analytical processing

1. Introduction

The various covering applied to the materials are very important for human activities. The range of these coverings is sufficiently broad and diversified; they are used for protection of the material surface from the environment exposure, for erosion decrease, for safety and reliability gaining, for durability growth, etc. Nevertheless, the traditional technologies for creating various coverings depleted their opportunities and the nanotechnologies are put forward. Nanotechnologies are one of the essential directions of modern production development allowing the ample opportunities for obtaining the objects with certain specified properties. The process involves the objects construction not from the natural materials but directly from atoms and molecules implementing the devices-assemblers, equipped with the artificial intellect systems. The materials obtained by the way of nanotechnologies application often surpass the natural materials with their quality. According to the scientists' and experts' forecasts, the nanotechnologies will be in the basis of the economy in the XXI century (Roko 2002). The issue of obtaining the nanomaterials with specified properties attracts great attention of the experts in the area of medicine, electronics, material science, ecology, etc. Nowadays the nanotechnologies take up the dominant positions in various fields of industrial production.

The nanotechnologies designed in the latest decades comprise such technology as the process of vacuum plasma deposition of nano-covering. The nano-materials conventionally involve disperse and massive materials with some content of structural elements (grains, crystals, blocks, clusters), if their geometrical sizes do not exceed 100 nm at least in one dimension and possessing the brand new properties, and functional and operational features. An increase in the vehicles mechanical components durability is one of the primary tasks for aerospace mechanical engineering. This area is characterized with operation conditions of the highest dynamic loads, as well as possibility of erosion abrasive particles ingress in the friction knots (Samotugin and Leszczynski 2002). The deeper investigation of tribotechnical processes taking place inside the vehicles limiting components as well as their controlling is the great reserve for the resource increase and upturn of the efficiency of machinery employment. The structure and the properties of durable covering depend in a great extent on the technique and technology of its employment. The methods of employment by the means of deposition are divided into two big groups: physical (PVD, physical vapor deposition) and chemical ones (CVD, chemical vapor deposition). There are a lot of varieties within these two groups.

All these processes are divided into two big groups: the processes utilizing the evaporation procedure, and the processes utilizing the sputtering procedure. Among all facilities dealing with the sputtering procedure, the most widely used ones are the plants on the basis of magnetrons MSIP (Magnetron Sputtering Ion Plating). In the process of implementing the high voltage in the atmosphere of inert gas (it is argon as a rule) the glow-discharge occurs. The inert gas ions from plasma usually possess the high energy and impact the target, inserted as a cathode. At the expense of impact impulse the release of the material takes place, and then it is sputtered escaping the intermediate liquid stage.

The magnetron method is the sub-kind of the cathode sputtering method, which assumes that the layer of plasma is formed at the surface of the sputtering cathode (the target) employing the crossed magnetic and electrical fields. The density of this plasma is significantly bigger compared to the common (non-magnetic) systems of cathode sputtering. Accordingly, there is a sufficient growth of the density of ion current on cathode and the velocity of cathode sputtering. The magnetron method allows applying the wide spectrum of covering of various metals and their compounds with the high uniformity of properties including the highly rigid durable coatings. The sputtered particles contain the neutral atoms on 75-95%, that is why the bottom layer (a detail, an item, a tool, etc) is heated faintly, and this permits to deposit the coatings on the detail having the low fusion temperature (low-melting-point metals and alloys, plastics, organic substances). The magnetron sputtering is supposed to be the prospective method and at the same time it is a method which has been well-tried and developed in the laboratory of vacuum technologies in the Institute of Transport Machinery Technologies in Riga Technical University. The experiments results are entered into the purposefully generated information system and further the accumulated data are

analytically processed; this processing allows discovering the dependence of covering parameters and the plant mode.

The analytical processing also permits discovering the new dependencies between the plant modes parameters, and this gives opportunity to obtain the dependencies allowing forecasting the covering properties and operation mode.

2. Prerequisites and Means of Problem Solution

The research group is dealing with the issues of applying thin layers (nano coatings) on different surfaces for long period of time. The experimental plant employed for applying coating (see figure 1) presents the modified vacuum plant NNV-6,6-I1 implementing the technology CIB – condensation of the substance from the steam phase with employing the effect of ion bombing (Kopytov et al. 2011).

Two arc evaporators and one magnetron dispenser are used as the sources of evaporation material. The refined details subjected to evaporation are inserted into the plant camera. The plant is extracted into high vacuum, then final refining and heating the details up to the temperature of deposition takes place with the employment of the ion bombing effect. Then the plant is transferred into the mode of evaporation in accordance with the preset program, the required composition and properties of coating are determined by the materials of vapor source, magnetron target and gas, injected into the operating camera. The coating thickness is regulated with the help of power of vapor source and the length of evaporation process.

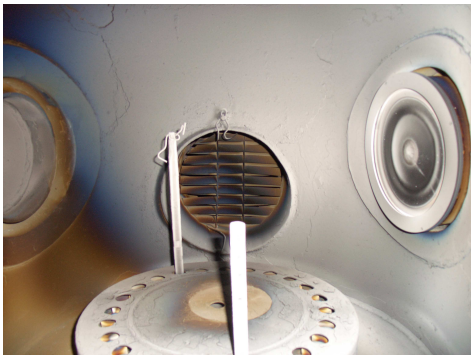


Fig. 1. Plant for applying the coating

The employment of different materials and injected gases in the process of evaporation allows obtaining the multi-component and multi-layer coatings, possessing the preset properties on section. So it is possible to obtain durable and corrosion-resistant, as well as ornamental coatings, and reestablish the geometrical sizes.

The diffractometer ADP-1 is used for evaluating the phase composition of obtained nano-coatings. Nevertheless the output analog signal was directed to the computer and data plotter. The investigated detail of this diffractometer can be irradiated and rotated on the operating table for 60^0 at a pitch of $0{,}001^0$. Consequently, under investigation of one sample only, the massive of 60 000 signals, describing the deposition uniformity and coating structure is obtained. It is necessary to compare the amplitudes of output signals with the base (reference) values for determining the quality of coating with preset properties. The diffraction patterns and special tables are employed on this purpose.

The measurement of the following parameters is necessary for evaluating the quality of precipitated coating:

- coating thickness;
- microhardness of coating or composition of coating-backing, GPa;
- power of coating and backing adhesion;
- roughness of the surface with coating;
- coefficient of friction.

The below listed parameters can be shown for comparative characteristics of the sample parameter values before and after applying the layer:

- hardness;
- surface roughness;
- friction coefficient.

The measurement of values of the above listed parameters is done by the measuring complexes and the values are registered in the logbooks.

3. Information System Development

The necessity of information support for implementing the experiments is determined by the whole range of reasons. Information on the results of the conducted experiments is kept separately; part of it exists in electronic form, and part in the hardcopies. It takes a lot of time to process the results and to determine the coating thickness according to the tables, and that is why the device, capable of transforming the analog signals from the diffractometer into the digital ones has been designed, and the data are registered in the special database; it is done on the purpose of computerizing the results processing.

The presented business requirement has been formulated for the system: the specialists of the research group need the information system capable of providing the registration and accumulating information on the conducted experiments, processing data, received from the measuring devices and the complex statistical processing of the obtained results. This system allows increasing the quality of measurements processing, to register and store centrally the information on investigated samples, to take the statistical analysis of the obtained results.

The structural analysis has been employed on the purpose of finding out the full set of functions and tasks of designed information system; there also has been

developed the functional model of the covering application process with IDEF0 methodology implementation (Grover and Kettinger 2000). The functional model shown in figure 2 permits accumulating the information on the process of covering application in so called "information bulk" within the frameworks of task formalisation.

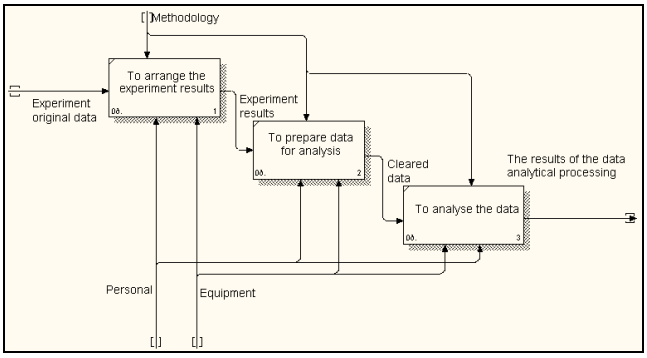


Fig. 2. Functional model of the covering application process

In the procedure of performing the experiment the special database (see figure 3), registering the following processes, has been developed for the initial data acquisition:

- *sample preparation* – in accordance with the instructions and methodological guide the sample is prepared for the experiment, the parameters measurements are done and then entered into the database;
- *nanocovering application* – the sample is placed into the plant, the certain specified parameters for covering application are designated and registered in the database;
- *parameters measurement after covering application* – the measurements of covering application evenness, its hardness, friction coefficient and strength of adhesion with padding are done with implementation of special equipment and then entered into the database.

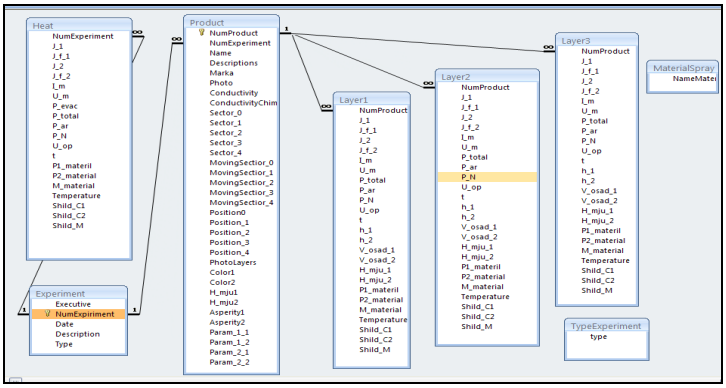


Fig.3. Model of the database

The data kept in the database then are transferred to the data warehouse (Inmon 2005). The essential processes of the data warehouse functioning are the following: data unloading, transformation loading, further data processing and the data marts production. There is the fundamental form for registering the basic data on the experiment – the performer, the date, the type and the experiment purpose, and so on (see figure 4). The additional forms register the plant mode parameters and the experiments results for every of three covering layers; this fact demonstrates that different parameters of the plant operation mode can be designated. Everything depends on the covering type; that is why there can be either one or two or three layers; the bigger number of layers is applied rather rarely in the experiment.

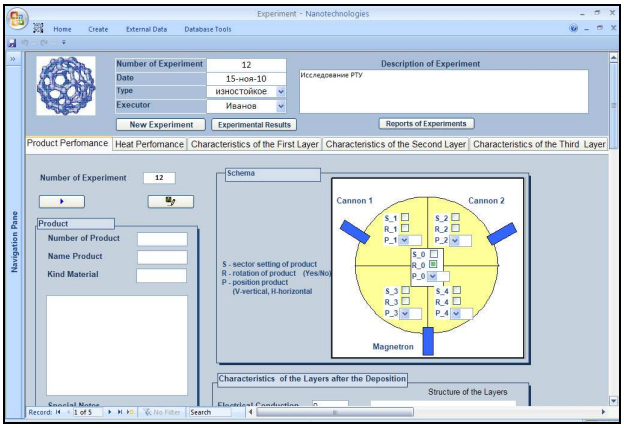


Fig.4. Application Interface

4. Data Analytical Processing

The different modes of the plant operation (see figure 5) were set in the process of performing the experiments: pressure, voltage, current; two parameters were set as constant values, while the third parameter was changeable. The performed experiments resulted in the obtained dependencies of the covering settling velocity on the plant operation mode. The above presented graphs were obtained after the data processing in this analytical module, and the equations describe these dependencies. The graphs presented in figures 6-8 are the results of the experiments investigating the procedure of applying the covering on the different sides of the gas-turbine engine blade.

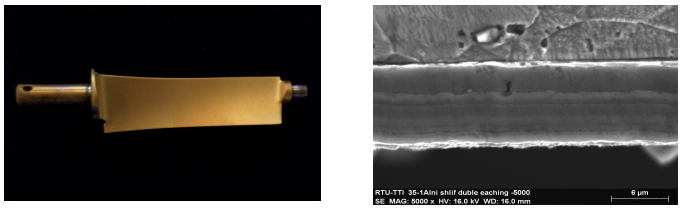


Fig. 5. Sample of blade with nanocoating

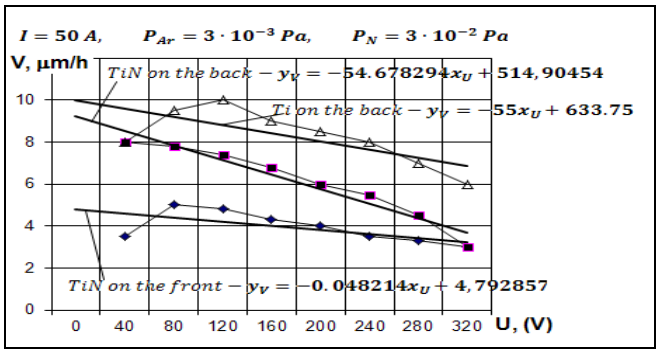


Fig. 6. The dependence of the coating rate on the voltage

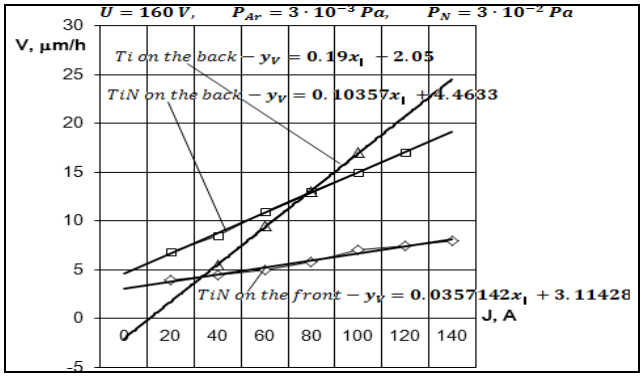


Fig. 7. The dependence of the coating rate on the current strength

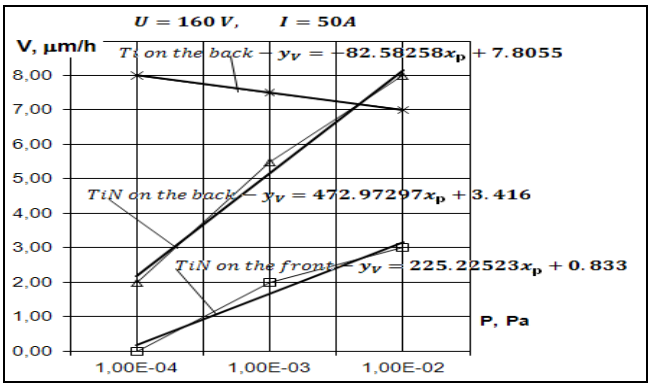


Fig. 8. The dependence of the coating rate on the pressure

The obtained dependencies allow the researcher to determine the plant operation mode, suitable for obtaining the covering with the specified properties, such as, for instance, thickness of covering. The determining parameter is the gas pressure in the plant. The covering settling velocity will determine the voltage and the current intensity in the plant employing the received dependencies. The analytical module allows identification of the plant mode parameters automatically and to a high precision. In the process of accumulating the experiments data in the different plant operation modes, the automatic correction of the received dependencies aimed at their updating and specification is done on the regular basis. Consequently, the analytical module allows efficient and prompt imposing the correcting procedures in the process of performing the experiment. Comparison of the forecast results and this forecast verification manifested a rather high rate of the results convergence.

3. Conclusions

The research under consideration covers the complex approach towards arranging the procedure of the experiments on applying the nanocoverings on various materials. This approach comprises the full provision of the experiments information support, including the database containing the entire information on the experiments results. The data kept in the database are “cleared” and then transferred to the data warehouse; consequently, the data are “downstocked” for the following analytical processing. The researcher employing the system analytical module has possibility to analyse this information, discover the hidden dependencies, and make the results forecasts and so on. This approach has provided the high quality level of performing the experiments on applying the nanocoverings.

Acknowledgment

The present article was executed within the framework of the Joint research project of Latvian Council of Science No. 10.0009 “Development of Technology for Industrial Multicomponent Nanostructure Sheeting” and with the financial assistance of European Social Fund. Project No. 009/0159/1DP/1.1.2.1.2/09/IPIA/VIAA/006 (The Support in Realization of the Doctoral Program "Telematics and Logistics" of the Transport and Telecommunication Institute).

References

1. Grover, V. and Kettinger, W.J., Ed., *Process Think: Winning Perspectives for Business Change in the Information Age*, Idea Group Publishing, Inc. (2000).
2. Inmon, W. *Building the Data Warehouse*, 4th ed, NY: John Wiley & Sons (2005).
3. Kopytov, E., Labendik, V., Yunusov, S., Urbach, F., Savkov, K. Informational support of carrying out experiments on producing nano-coverings. *Proceedings of VII International “Congress Machines, Technologies, Materials 2011”,* Varna, Bulgaria, **8**/128, 194-196 (2011).
4. Roko, M., *Nonotechnology in the next decade*, Moscow, Mir (2002).
5. Samotugin, S., Leszczynski, L., *Plasma hardening of tool materials*, Donetsk, Novij Mir (2002).

Reliability Evaluation of 3D Motion Detector by using Monte-Carlo Simulation

Samuel Kosolapov

Signal and Image Processing Laboratory, ORT Braude Academic College of Engineering, Karmiel, Israel
Email: ksamuel@braude.ac.il

Abstract: This article analyzes reliability of 3D Motion Detectors (3DMD) containing two video cameras constantly monitoring 3D zone in attempt to detect intrusion of the material object(s) into this zone. Achieving a high 3D accuracy is not a goal in this case. Goal of this research is to evaluate reliability of intrusion detection for the customer specified setup by using simulation. Mathematical simulation describing 3DMD was built by using MAPLE. Simulation takes into account geometrical parameters of the 3DMD setup, camera noise, digitization effect and parameters of the 3D zone to be protected. First unit of the simulator generates a series of the "Left" and "Right" images in the situation when target (object of specified sizes) is moving over specified path with a selected speed. Second unit of the simulator detects target motion by analyzing above series of images and builds 3D motion path of the target. Third unit of the simulator evaluates if security event (violation of the volume to be protected) happen by using results of partial (simplified) calibration procedure. Simulations are executed a number of times (classical Monte-Carlo approach) in order to evaluate importance of setup assembly tolerances on the resulted 3DMD reliability). Additionally, influence of errors in inter-camera synchronization was analyzed. Results of simulations of a practically interesting 3DMD setup were summarized as a set of 2d and 3d plots enabling to recommend 3DMD setup parameters for the specified 3D zone to be protected.

Keywords: Image Processing, 3D Imaging, Stereo Camera, Motion Detector, Monte-Carlo simulation, MAPLE

1 Introduction

Motion Detectors (MD) utilizing one video camera are well-known and widely used for detection of physical objects intrusion into protected zone [1, 2]. Disadvantage of most MD is that any significant change in the content of the frame grabbed by video camera is treated as "security violation event". This means that MD cannot be used in case when only part of the zone seen by camera must be protected against intrusion of the physical objects of the pre-specified sizes. 3D Motion detector (3DMD), containing two video cameras (left and right) can measure 3D positions and sizes of the physical objects (clearly distinguishable as by left as by right cameras) by using well-known triangulation technique (after proper system calibration).
A number of practical 3DMD implementations are known [1, 2, 3, 4].

Typically, 3DMD calculates a difference of the two consecutive frames from corresponded cameras. This step effectively eliminates "constant" background, which make 3D triangulation from "matched" points on the left and right image more reliable [3].
Despite advantages of 3DMD over one-camera MD are obvious, practical implementation of 3DMD is not trivial.
Full-fledged "standard' camera calibration procedures are not always feasible in the field conditions, so that only partial calibration can be used [4]. During partial calibration procedure some parameters of camera and setup are assumed as known with some assembly error, whereas other parameters can be measured.

2 Simulation

Achieving a high 3D accuracy is not a goal in this case. Goal of this research is to evaluate reliability of intrusion detection for the customer specified setup by using simulation.
In this research previously developed MAPLE simulator [5] was modified: some of the 3DMD setup parameters (camera's parameters: focal length, CCD size, pixel size) are considered as known (say, from camera' specification, or from laboratory measurements) with modest accuracy, whereas other 3DMD setup parameters (setup assembly parameters) are considered as known with poor accuracy.
Additionally, in the discussed setup, two cameras were fixed on the rigid frame, so that directions of the optical axis of both cameras can be considered as known with modest accuracy.
Considering mentioned above fact that full-fledge calibration (for example by presenting to both cameras some 3D test object of known sizes in the known positions and orientations) is problematic in the real-life conditions, partial simplified calibration was executed by marking a number of points near the borders of the zone to be protected. Generally, this procedure is executed by presenting to the both cameras clearly distinguishable object (for example color sphere of small size) in a number of 3D positions. In this way borders of the zone to be protected can be set in acceptable accuracy [4]. In his research simplified calibration was simulated by generating a series of images containing white circles. Positions of their centers are known with pixel accuracy (resulting in "digitization noise" – lowering 3D accuracy of the 3D triangulation).
Quasi-calibration procedure does not improve accuracy of 3D measurements, but it improves reliability of motion detection inside the marked zone.
Simulations were executed for a number of practically interested scenarios by using classical Monte-Carlo Simulation approach
First unit of the simulator generates series of "left" and "right" digital images in the situation when object of specified size is moving over specified path (seen by both cameras) with a selected speed.

In the real-life situation, acquisition time of the right and left images are slightly differ (inter-camera synchronization error). Influence of this inter-camera synchronization was analyzed by "rounding" acquisition time (which results in an additional error in 3D triangulation).
Second simulator unit is attempting to "find" the object motion by using differences of the two consecutive frames from corresponded cameras.
Third unit of the simulator evaluates if security event (violation of the volume to be protected) happen (by using results of partial calibration procedure). Simulations are executed a number of times (classical Monte-Carlo approach) in order to evaluate importance of setup assembly tolerances on the resulted 3DMD reliability. Fig. 1 presents typical results of Monte-Carlo simulations (5000 runs). Configuration used was: right and left camera positioned on the base line at the distance 10 m. Optical exes of both cameras are intersected at the distance of 5 m from the base line. No calibration procedure was used (which means that all relevant setup parameters were assumed as known with specified tolerance). During every run of Monte-Carlo procedure, every setup parameter was changed in a pseudo-random fashion in accordance to its tolerance. Results are summarized as histogram.

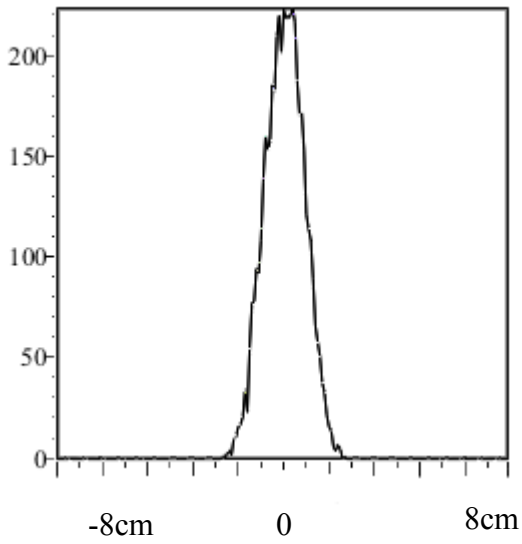


Fig. 1. 3DMD reliability evaluation by using
Histogram of 3D Distance "Measurements".
X-Axis: Distance from "security zone"; Y-Axis: Number of Events
Histogram bin size: 1cm X 1cm
No Calibration. No Markers.

By using results presented on the Fig. 1 it is possible to analyze FPR (False Positive Rate - "The zone to be protected" was NOT violated, but "the alarm" WAS raised – situation that typically forces the user of MD to switch off the system)
In case extra to "security zone" was set to 1 cm FPR was 0.42 (3DMD was tested 5000 times and "Alarm" was raised in 2088 cases). For the 10 cm FPR was 0.
This means that small expansion of the zone to be protected (having practically no meaning for the customer) can drastically improve reliability of the 3DMD.

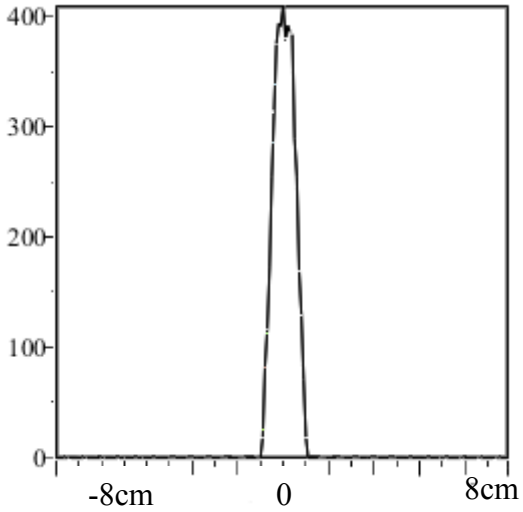


Fig. 2. 3DMD reliability evaluation by using
Histogram of 3D Distance "Measurements".
X-Axis: Distance from "security zone"; Y-Axis: Number of Events
Histogram bin size: 1cm X 1cm
No Calibration. 1 Marker.

Fig. 2 presents results of the Monte-Carlo simulation for the case when 1 marker was used (that is one white ball was presented to the system to mark "center of the protected zone). It can be seen, that even 1 marker significantly improves reliability of the 3DMD.

3 Conclusions

Results of Monte-Carlo simulations enable to evaluate for selected 3DMD setup (assuming that assembly tolerances are known) expected reliability, and, thus, decide if selected setup is fitted for the specified goals or not. It was shown that typical for real systems inter-camera synchronization error may significantly lower 3DMD reliability, hence special techniques minimizing this influence must be developed to make 3DMD approach feasible.

References

1. Jain R., Kasturi, R., Schnuck, B.G., *Machine Vision*. McGraw-Hill Inc., 289-297 (1995).
2. Azad, P., Gockel, T., Dillman R., *Computer Vision*, *Principles and Practice. Elector Interational Media BV* , 137-142 (2008)
3. Hee-Sung Kim, Kurillo, G., Bajcsy, R. Hand tracking and motion detection from the secuence of stereo color image frames, *IEEE International Conference on Industrial Technology, 1-6, (2008)*
4. Satoshi Kawabata, Intrusion Detection System using Contour-based Multi-Planar Visual Hull Method, *Doctorial work, 2003*
5. Kosolapov S., Lomes A., Dead-Zone and 3D Accuracy Evaluation of Modified Two-Camera Stereo Setup by using Monte-Carlo Simulation. *Proceedings ASMDA 2011, 744*, (2011).

Monte-Carlo Simulation of Poorly Calibrated Stereo Tracking System

Samuel Kosolapov and Alexander Lomes

Signal and Image Processing Laboratory, ORT Braude Academic College of Engineering, Karmiel, Israel
Email: ksamuel@braude.ac.il

Abstract: In this paper we present and analyze the feasibility of poorly-calibrated stereo-tracking system (STS) – using two video cameras to evaluate path of the moving object. It is known, that in order to get better 3D accuracy, wide base between two cameras must be used. This, however, leads to problems with camera calibration – so called "poor calibration": in which only part of the calibration parameters can be measured in the field conditions. Previously developed MAPLE simulator was modified so that some STS setup parameters (mostly camera's parameters: focal length, CCD size, pixel size) are considered as known from camera' specification with modest accuracy, whereas other STS setup parameters (setup assembly parameters) are considered as known with poor accuracy. In order to evaluate feasibility of poorly calibrated STS, classical Monte-Carlo approach was used to evaluate accuracy and reliability of STS in a number of practically interesting situations. Results of simulations were organized as a set of 2d and 3d plots enabling to recommend customer-tailored STS configuration for the selected tracking ranges and velocities.

Keywords: Image Processing, 3D Imaging, Stereo Camera, Video tracking, Monte-Carlo simulation, MAPLE

1 Introduction

Video tracking - process of locating a moving object and evaluating its motion path by analyzing consecutive video frames is well known and is widely used in a number of applications like security and surveillance. A number of tracking algorithms of a different computation complexity are known [1, 2, 3].
It is known, that Video Tracking Systems utilizing one video camera can provide useful tracking only in a limited number of cases. Stereo Tracking Systems (STS), containing two video cameras, after proper calibration, can provide more exact 3D tracking in a wider range of situations [1, 2, 3].
In order to increase working range and 3D accuracy of STS, large stereo base can be used. Unfortunately, in this case full-fledged stereo calibration is problematic, or even impossible, especially in the field conditions: for example, placing flat test-object of known sizes in a number of orientations (classical MATLAB approach) in the field of view of both cameras is problematic for distances between cameras more than, say, 50 m.

In this research previously developed MAPLE simulator [4] was modified: some of the STS setup parameters (mostly camera's parameters: focal length, CCD size, pixel size) are considered as known from camera' specification with modest accuracy, whereas other STS setup parameters (setup assembly parameters) are considered as known with poor accuracy. Fortunately, we know that poor-calibration stereo vision is feasible: animals and humans uses a pair of eyes to evaluate distances; exact orientation of eyes is assumed to be known by a brain, but with a very poor accuracy.
Hence, goal of this research was to evaluate accuracy and reliability of poorly-calibrated STS, and those evaluations were executed for a number of practically interested situations by using classical Monte-Carlo Simulation approach
First unit of the simulator generates series of "left" and "right" digital images in the situation when object of specified size is moving over specified path (seen by both cameras).
Second unit is attempting to track the object in the "left" and "right" images and evaluate its 3D motion path. "Tracking" is repeated selected number of times; every "tracking attempt" is executed with a different set of selected parameters, whereas every parameter in the set is varied in accordance with its tolerance.

2 Simulation results

Classical two-camera setup is symmetrical: left and right camera are positioned at an equal distances from the Z-axis. This configuration is inconvenient when distance between cameras became large (position of the middle point may be physically inaccessible).
On Figure 1 and Figure 2 presented modified (non-symmetrical) setup: center of XYZ coordinate system is associated with the left camera.
Every parameter of the setup is considered as known with its error marked by a letter "E". For example, distance between left camera and right camera (d) is shown as (d+Ed). Used model contains 22 parameters. Error vector (E-vector) contains 19 components.
During Monte-Carlo simulation all 22 setup parameters and 19 E-vector components were set (in accordance with selected situation: for example: d = 100m, Ed = 1m).
Results of simulation were presented as a set of 2D and 3D graphs.
For practically reasonable assembly errors, results of 3D accuracy evaluation demonstrates, that absolute XYZ coordinates have a low (in some situations inacceptable) accuracy (for example, for d =100m, 16 MegaPixels camera, and for the distance to target 3km, absolute Z-error is about 600m, absolute X (and Y) errors are about 10m).
In the same situation relative XYZ accuracy (evaluated as "size of the object" was overwhelmingly better: about 5 m for Z and less then 1 m for X and Y.
It is clear, that for a smaller distance to target values of those errors are smaller.

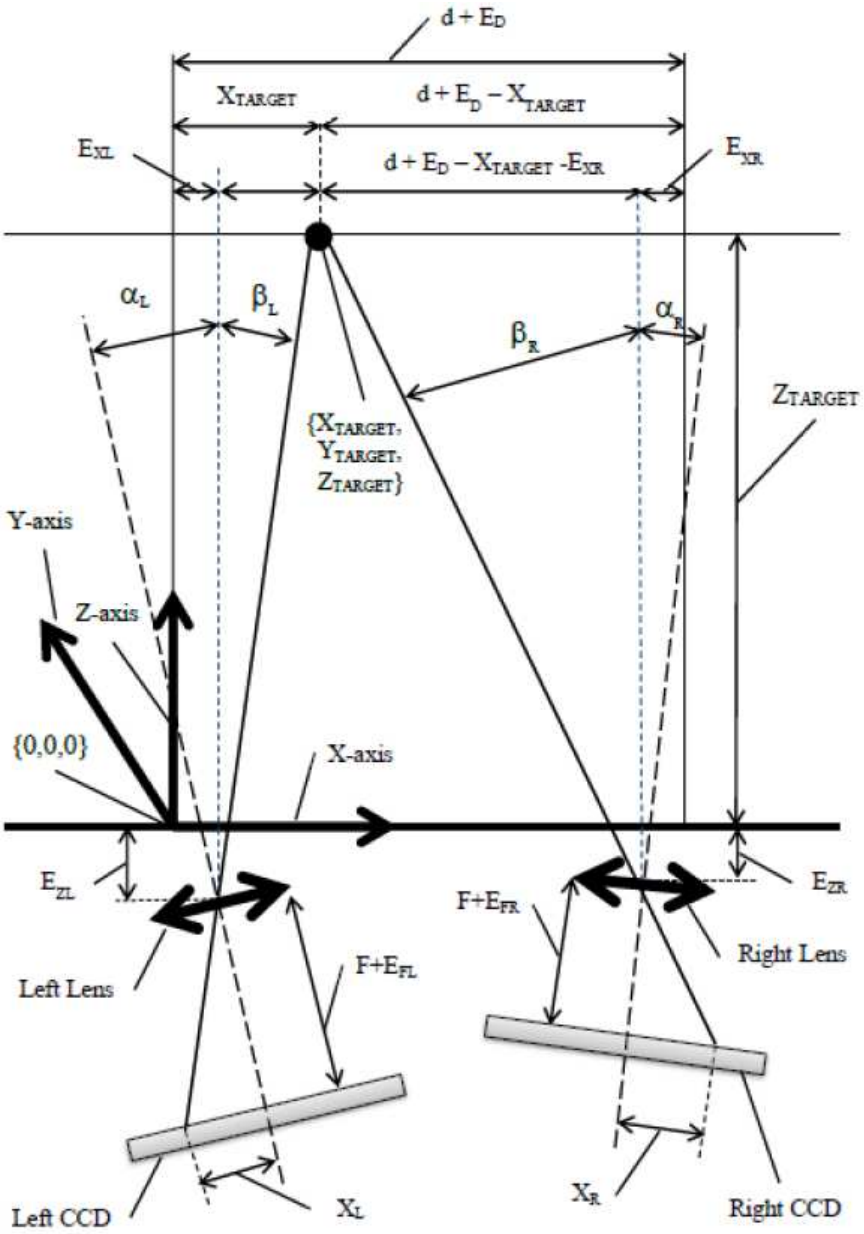


Fig. 1. Modified Parallel Configuration with Assembly Errors. X-Z Plane

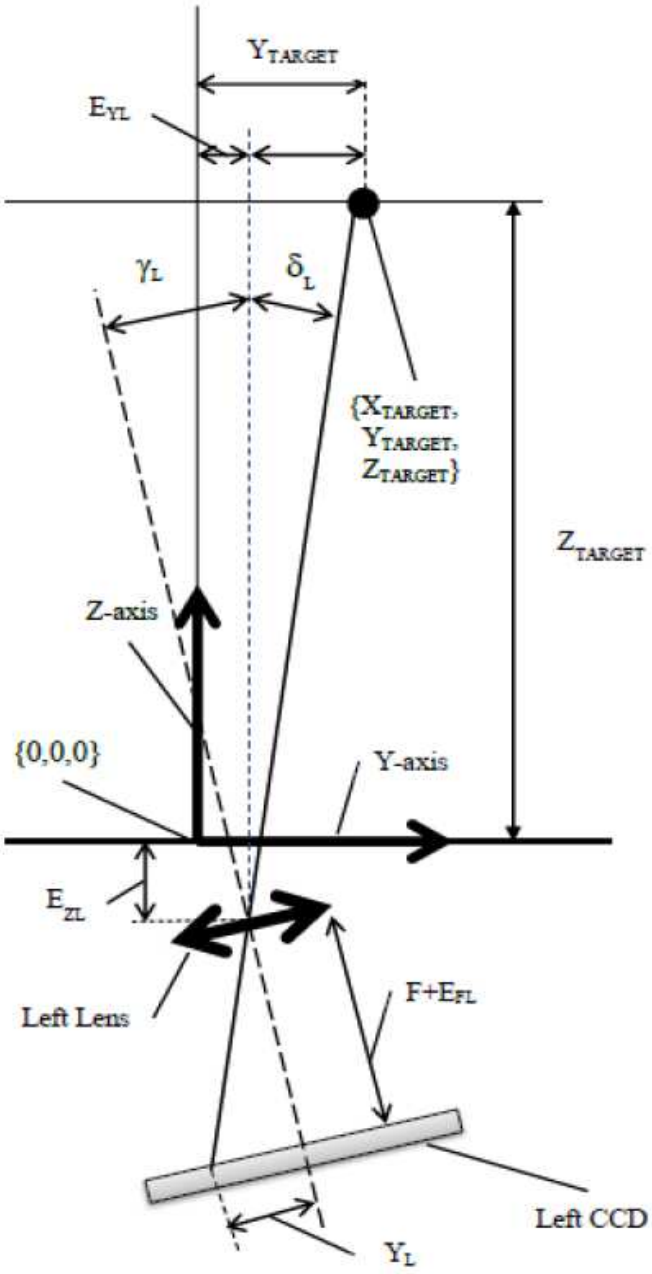


Fig. 2 Modified Parallel Configuration with Assembly Errors. Y-Z plane.

In one analyzed exemplary (and practically interesting) tracking scenario, target object is heading to STS, which means that distance to tracking object is decreasing in time. In this situation poorly calibrated STS may be useful for target path prediction: when a distance to target is large, single XYZ point of a target path is known with a significant error, but, by collecting a plurality of single XYZ points, obtained in a different periods of time, mathematical description of the target path can be known with significantly better accuracy.

3 Conclusions

Results of Monte-Carlo simulations enable to evaluate for selected setup (assuming that assembly tolerances are known) expected STS accuracy for the selected camera parameters (like focal length, CCD size and pixel size), and, thus, decide if selected setup is fitted for the specified goals or not. Results of poorly calibrated STS simulations show that even small assembly errors lead to significant errors in absolute {X, Y, Z} coordinates evaluations, whereas distances between objects can be measured with adequate accuracy..

References

1. Jain R.,, Kasturi, R., Schnuck, B.G., *Machine Vision*. McGraw-Hill Inc., 289-297 (1995).
2. Azad, P., Gockel, T., Dillman R., *Computer Vision*, *Principles and Practice. Elector Interational Media BV* , 137-142 (2008)
3. Kai Ni, Dellaert F.,Stereo Tracking and Three-Point/One-Point Algorithms – a Robust Approach in Visual Odometry, *2006 IEEE International Conference on Image Processing*, *2777-2780*, (2006).
4. Kosolapov S., Lomes A., Dead-Zone and 3D Accuracy Evaluation of Modified Two-Camera Stereo Setup by using Monte-Carlo Simulation. *Proceedings ASMDA 2011, 744*, (2011).

Statistical analysis of fractal point processes: an application to Northern Aegean (Greece) seismicity

Eleftheria C. Kotti and George M. Tsaklidis

Department of Mathematics
Aristotle University of Thessaloniki (AUTH)
541 24 Thessaloniki, Greece
(e-mail: riakotti@math.auth.gr, tsaklidi@math.auth.gr)

Abstract. The point process describing the earthquakes sequence for the Northern Aegean area in the time period 1970-2011 is studied. The abrupt cutoff distribution is employed to model the earthquakes sequence. Time-scaling analysis of the sequence is performed using the statistical measures of coefficient of variation, Fano factor, Allan factor, and Hurst exponent. The purpose of statistical analysis is the detection of clustering structures and the correlation of the occurrence times. The point process of the earthquakes sequence can be considered as a fractal process with high degree of time-clusterization of the events.
Keywords: Clustering, Power-law, Coefficient of variation, Fano factor, Allan factor, Hurst exponent, Earthquakes sequence.

1 Introduction

Several models have been proposed to represent the earthquake genesis of process. The most simple statistical model concerning the number of the earthquakes in some time interval is the Poisson model. Other statistical models, which have been applied are based on the Gaussian (Rikitake, 1974), the Weibull (Hagiwara, 1974; Utsu, 1984; Rikitake, 1999), the gamma (UdiasandRice, 1975; Utsu, 1984), double exponential (Utsu, 1972) and the lognormal (Nishenko, Bullard, 1987) distribution. The study of earthquakes occurred in North Aegean is of outmost importance because of the great number of recorded earthquakes of magnitude greater than or equal to 4.1, which are studied in this paper. The aforementioned earthquakes can be considered as a fractal point process.

2 Statistical methods

2.1 Coefficient of variation (C_ν)

A widely applied measure associated with the cluster behaviour of point processes is the coefficient of variation (C_ν), which is defined as

$$C_\nu = \frac{\sigma_\tau}{< \tau >}, \tag{1}$$

where $< . >$ denotes the expected value of the interevent times and σ_τ is the corresponding standard deviation. A clustered process is characterized by $C_\nu > 1$. Especially, the Poisson process has $C_\nu = 1$ (Kagan and Jackson, 1991 [4]). Note that the coefficient of variation does not provide any information concerning the timescale.

2.2 Fano Factor (FF)

The Fano Factor (FF) or normalized variance is defined as

$$FF(T) = \frac{< Z_k^2(T) > - < Z_k(T) >^2}{< Z_k(T) >}, \tag{2}$$

where $Z_k(T)$ represents the number of events falling into the kth window of duration T (Thurner et al., 1997 [11]).

The FF of a fractal point process with $0 < a < 1$ varies with the counting time T following the power-law

$$FF(T) = 1 + \left(\frac{T}{T_0}\right)^a, \tag{3}$$

where T_0 is the fractal onset time and a is the scaling exponent which is called *fractal exponent*. T_0 characterizes the lower limit of the significance scaling behaviour of the FF (Teich et al., 1996 [8]).

The linear fit of $FF(T)$ vs T on log-log scale can be used to estimate the fractal exponent a. If $a > 0$, then the sequence consists of clusters of points over a relatively large set of timescales (Lowen et al., 1996 [6]). However, if $a \approx 0$, then the process is Poissonian and the occurrence times are uncorrelated.

2.3 Allan Factor (AF)

The Allan Factor is an alternative measure connected to the variability of successive counts. The Allan Factor (AF) (Allan, 1966 [2]) is defined as the variance of the successive counts at the specific counting time T, divided by twice the mean number of events in this counting time, namely

$$AF(T) = \frac{< (Z_{k+1}(T) - Z_k(T))^2 >}{2 < Z_k(T) >}. \tag{4}$$

The AF of a fractal point process with $0 < a < 3$ varies with the counting time T according to the power-law

$$AF(T) = 1 + \left(\frac{T}{T_1}\right)^a, \tag{5}$$

where T_1 is the fractal onset time estimated as the crossover timescale between Poissonian and scaling behaviour (Lowen and Teich, 1995 [5]).

2.4 Hurst exponent (H)

We use the detrended fluctuation analysis (DFA) proposed by Peng (1994). Let $F_{DFA}(n)$ denote the square root of the mean of fluctuations for all parts into which the time series is separated. Then the slope of the graph $\log(F_{DFA}(n))$ vs $\log(n)$ provides an estimator of the Hurst exponent. If $0 < H < 1/2$, the differences between the successive values are called antipersistent, if $H = 1/2$ they are uncorrelated and if $1/2 < H < 1$ they are called persistent. The nearer the H-value is to unit, the more previous values affect future ones.

2.5 Simulation

It is suggested by (Lowen et al., 2005 [7]) to examine if the interevent times of an earthquake sequence follow an abrupt cutoff power-law density with cutoffs at the highest and lowest values

$$f(t) = \frac{a}{A^{-a} - B^{-a}} \begin{cases} t^{-(a+1)}, & A < t < B \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where t is the interevent time and A, B are the cutoffs.

The respective cumulative density function is

$$F(t) = \frac{a}{A^{-a} - B^{-a}} \int_A^t y^{-(a+1)} dy = c(t^{-a} - A^{-a}), A < t < B$$

where $c = \frac{-1}{A^{-a} - B^{-a}}$. F is reversible on the interval (A, B) and solving the equation $F(t) = y$ yields $F^{-1}(t) = (\frac{c}{y + A^{-a}c})^{1/a}, A < t < B$. According to the inverse method, if $U \sim U(0, 1)$, then

$$T = F^{-1}(u) = \left(\frac{c}{u + A^{-a}c} \right)^{1/a}. \tag{7}$$

Formula (7) may be applied to fractal stochastic point process simulation.

2.6 Gutenberg-Richter law

The linear relation between the magnitude and the cumulative frequency, is defined by the Gutenberg-Richter law (Gutenberg and Richter, 1994 [3])

$$logN = a - b \cdot M, \tag{8}$$

where N is the number of earthquakes of magnitude greater than or equal to M, a and b are parameters.

3 Study region and identify data

The seismic activity in North Aegean from 1970 until 2011 concerning $N = 1664$ earthquakes with magnitude greater than or equal to 4.1 is studied (figure 1). The data of the basic focal parameters (generation time, focal point, focal depth, magnitude) of the earthquakes are recorded by the Geophysics Laboratory, Department of Geology, Aristotle University of Thessaloniki, Greece. The total length of the sequence of earthquake events is nearly 42 years ($1.319.607.976s$). All the following calculations and figures are made in MATLAB. From figure 2, it can be seen that the majority of

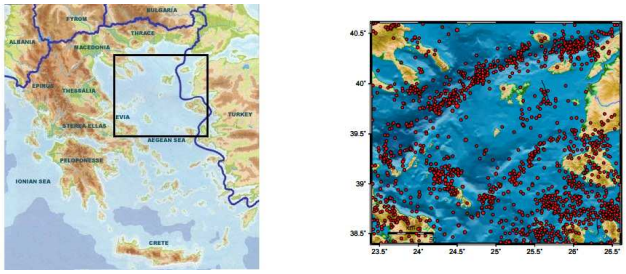
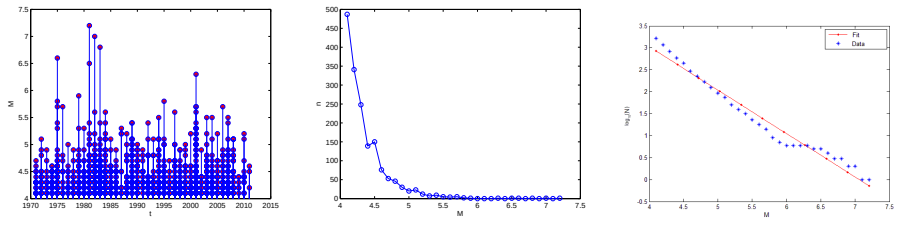


Fig. 1: Spatial distribution of earthquakes in Northern Aegean in the period 1970-2011.

earthquakes have magnitude between 4.1 and 5 and the number of earthquakes decreases abruptly with earthquake magnitude. Roughly, around half of the earthquakes have magnitude between 4.1 and 4.2.



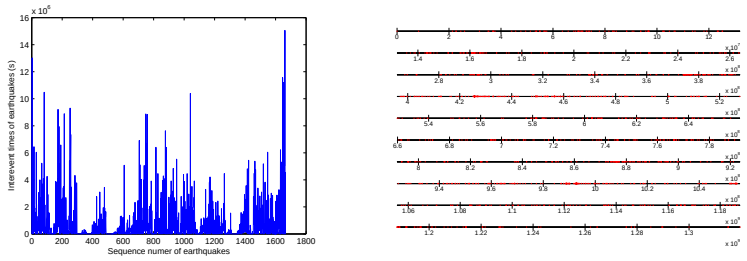
(a) (b) (c)

Fig. 2: (a) Earthquakes magnitude, (b) Frequencies occurrence earthquakes n versus magnitude M, (c) Gutenberg-Richter law plot for the earthquakes of Northern Aegean, 1970-2011.

The estimation of the parameters α and β in the formula (8) representing the Gutenberg-Richter law is carried out by the least squares method, implying $\log N = 6.985 - 0.9893 \cdot M$. The linear regression of the logarithm of cumulative frequency on the earthquakes magnitude is depicted at figure 2c. It is concluded that the power of the logarithm of cumulative frequency decreases when the magnitude increases.

4 Data analysis

To continue the data analysis, it is necessary to convert the occurrence times of earthquakes to seconds. The calculations are accurate taking into account leap years. We calculate the interevent times. In figure 3a there is a clear indication that there are clusters. This is demonstrated in figure 3b, where there are capacitors for example from $8.7 \cdot 10^8 s$ to $9 \cdot 10^8 s$ and dilutions from $9 \cdot 10^8 s$ to $9.2 \cdot 10^8 s$.



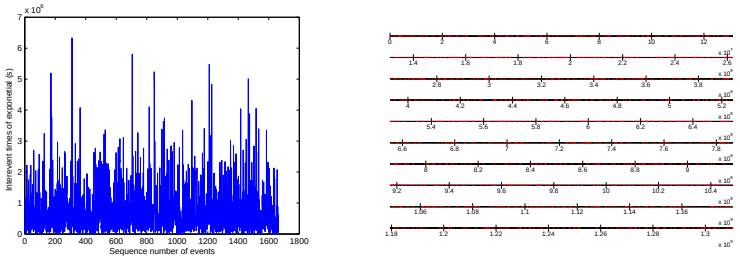
(a) (b)

Fig. 3: The (a) interevent times and (b) the point process, of the earthquakes on the North Aegean, 1970-2011.

We calculate the coefficient of variation for the real data and we get that $C_\nu = 1.824$, which indicates cluster appearance. The Hurst exponent for the real data is calculated via the slope of the line of $\log F_{DFA}(n)$ vs $\log n$. To calculate it we apply the square root method to conclude that H-value is 0.7706. This means that, if there is an increase of the interevent time at the precedent step, then it is expected that there will be an increase at the next step. The same is true for the decreases.

We study the simulated abrupt cutoff distribution for $A = 10^{3.3}$, $B = 10^7$ and $a = 0.38$, which can be considered as a fractal point process of earthquakes. It exhibits similar behaviour with the sequence of real data. In this case, the average for the coefficient of variation is $C_\nu = 3.832$ and for the Hurst exponent equals $H = 0.5570$. The aforementioned statistical measures indicate that the simulated sequence is clustered.

In contrast, we study the simulated Poisson distribution with rate equal to the rate estimated by the real data, $\lambda = 1.261 \cdot 10^{-6}$ events per second. In this case, the coefficient of variation is $C_\nu = 1.003$ and the Hurst exponent is $H = 0.4767$. These statistical measures indicate that the occurrence times are uncorrelated. The difference between the real and Poissonian sequence are shown in figure 4.



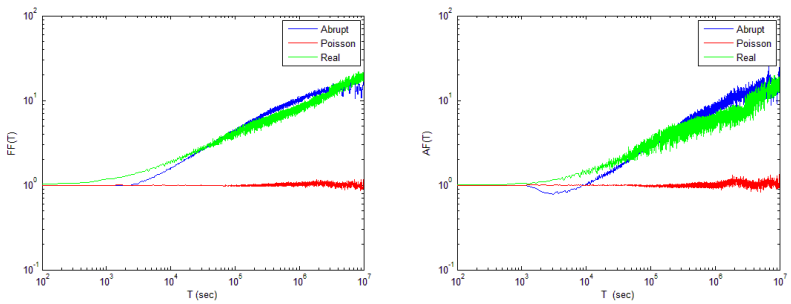
(a) (b)

Fig. 4: The (a) interevent times and (b) the point process, of the Poisson distribution for $N = 1664$ events.

Figure 5a sketches the FF curve as a function of T for the real data. The power-law behaviour is clearly shown, since the curve increases linearly from $10^4 s$ to $10^{6.6} s$. In this case, the estimated fractal exponent equals the slope of the linear fit calculated using the least square method. This slope is $a_{FF} \approx 0.362$. The exponent indicates the presence of time correlated structures for the earthquakes sequence under study. Moreover, similarity between the curve of real data and the curve of the simulated sequence is obvious. In the last sequence, one can notice the power-law behaviour, since the FF curve increases linearly starting from $10^4 s$,, with mean $a_{FF} \approx 0.338$, which is close to the a_{FF} of the real sequence. In contrast, the FF curve for the Poissonian sequence does not follow power-law behaviour (as expected), since the fractal exponent is $a_{FF} \approx 0.003$, indicating that the occurrence times are uncorrelated.

In figure 5b the AF curves for the above three sequences are plotted. The real data clearly reveal power-law behaviour, since the curve increases linearly starting from $10^4 s$. The limiting time scale $10^4 s$ indicates the low limit for the significance of the time scale behaviour. This indicates that for counting times larger than the limiting time scale the sequence of earthquakes can be considered as the realization of a fractal stochastic point process. In this case, the estimated fractal exponent equals the slope of the fitted line calculated by the least square method and we get that $a_F \approx 0.391$. This exponent indicates the presence of time correlated structures for the examined

earthquakes sequence. The average of the fractal exponent is $a_F \approx 0.400$ for simulated sequence and follows the power-law. In contrast, the fractal exponent for the Poissonian sequence is $a_F \approx 0.015$, i.e. the occurrence times are uncorrelated.



(a) (b)

Fig. 5: The (a) FF and (b) AF curves on double logarithmic scale for the real data of earthquakes, the simulated and the Poissonian sequence.

In general, for the linear regression case it holds that $y' = a_1 + b_1 x'$. By applying the transformation $y' = \log_{10} y$ and $x' = \log_{10} x$, last formula becomes

$$\log_{10} y = a_1 + b_1 \log_{10} x$$

or

$$y = 10^{a_1 + b_1 \log_{10} x}$$

and

$$y = 10^{a_1} \cdot x^{b_1}.$$

The power-law formula is $y = c \cdot x^d$. By equating the last two formulas, we get

$$c = 10^{a_1} \text{ and } d = b_1. \tag{9}$$

5 Conclusions

The analytic formula (7) is useful in order to get a random sample from the abrupt cutoff distribution and to simulate the fractal point process. The

point process of the earthquakes sequences can be considered as a fractal process with high degree of time-clusterization of the events. The sequence of the data of North Aegean evidences time-clustering and specific positive autocorrelation. So, if there are increases of the occurrence frequency of earthquakes on some time period, then there is a large probability for this trend to be continued in the next period. The FF and AF reveal power-law behaviour. In this case, we calculate the relationship between the parameters of the linear and the power-law regression, which is given by the formula (9).

6 Thanks

We want to thank the Foundation for Education and European Culture (IPEP), for the financial support provided to Kotti Eleftheria, academic years 2010-2012.

References

1. Aldroubi A., Unser eds. M., "Wavelets in Medicine and Biology," *CRC Press, Boca Raton, Florida* (1996).
2. Allan D.W.,"Statistics of atomic frequency standards," *Proc. IEEE* 54, 221-230 (1966).
3. Gutenberg B., Richter C.F. "Frequency of earthquakes in California," *Bull. Seism. Soc. Am.* 34, 185-188 (1944).
4. Kagan Y.Y., Jackson D.D., "Long-term earthquakes clustering," *Geophys. J. Int.* 104, 117-133 (1991).
5. Lowen S.B., Teich M.C., "Estimation and Simulation of Fractal Stochastic Point Processes," *Fractals* 3, 183-210 (1995).
6. Lowen S.B., Teich M.C., "The periodogram and Allan variance reveal fractal exponents greater than unity in auditory-nerve spike trains," *J. Acoustical Society of America* 99, 3585-3591 (1996).
7. Lowen S.B., Teich M.C., Jost B.M., Vibe-Rheymer K., Heneghan C., "Fractal-Based Point Processes," *Wiley Series in Probability and Statisctis* (2005).
8. Teich M.C., Heneghan C., Lowen S.B., Ozaki Y., Kaplan E., "Fractal Character of the Neural Spike Train in the Visual System of the Cat," *J. Opt. Soc. Am. A* 14, 529-546 (1997).
9. Telesca L., Cuomo V., Lasenna V., Macchiato M., "Statistical analysis of fractal properties of point processes modeling seismic sequences," *Physics of the Earth and Planetary Interiors* 125, Elsevier, 65-83 (2001).
10. Telesca L., Amatulli G., Lasaponara R., Lovallo M., Santulli A., "Time-scaling properties in forest-fire sequences observed in Gargano area (Southern Italy)," *Ecological Modelling* 185 Issue 2-4, Elsevier, 531-544 (2005).
11. Thurner S., Lowen S.B., Feurstein M.C., Heneghan C., Feichtinger H.G., Teich M.C., "Analysis, synthesis, and estimation of fractal-rate stochastic point processes," *Fractals* 5, 565-595 (1997).
12. Scharf R., Meesmann M., Boese J., Chialvo D.R., Kniffki K.D. "General relation between variance-time curve and power spectral density for point processes exhibiting $1/f^{\beta}$-fluctuations, with special reference to heart rate variability," *Biol. Cyb.* 73, 255-263 (1995).

13. Sung-IL K., Jaeseung J., Yongho K., Yang I.K., Seung H.J., Kyoung J.L.,"Fractal Stochastic Modeling of Spiking Activity in Suprachiasmatic Nucleus Neurons," *Journal of Computational Neuroscience* 19, 39-51 (2005).
14. Votsi I., Limnios N., Tsaklidis G., Papadimitriou E. "On the study of non-parametric semi-Markov Model used for earthquake prediction," *IWAP* (2010).

Childbearing and Labor Market: Time and Space Dynamics

Elena Kotyrlo[1]
[1]Department of Economics, Umeå University, Umeå, Sweden
E-mail: elena.kotyrlo@econ.umu.se

Abstract: Fertility is one of the important determinants of population growth and labor market situation in the long-run. We focus on time and space dynamics in the description of fertility in Sweden, which is presumably generated by labor mobility across municipalities. Time dynamics as postponing or accelerating of childbearing has been considered in two effects of earnings. The first effect has been considered within one generation, when families compare their current earnings with the earnings in previous and nearest future periods. Another effect or the Easterlin hypothesis is observed through the generations, when households belonging to the younger generation compare their earnings relatively to earnings of the parental generation. The hypotheses have been tested for the period 1981-2008. The study is based on estimating of space and time dynamics by estimating the SAR(2,1) model and using the general method of moments for aggregate panel data. Comparing different specifications positive spatial autocorrelation of fertility has been found. We found that current earnings negatively affect fertility rates within municipalities and in the long-run across them. The inverted Easterlin hypothesis is weakly supported within municipalities. Theoretical contributions contain the stationary condition and the long-run effect in the direct, indirect and total forms in the model SAR (2,1) with a second-order autoregressive and a first-order spatial disturbances.

Introduction

The XXth century has been characterized by dramatic decline in fertility in developed countries, which comes with spreading of treatment for contraception and growing involvement of women into labor market. Despite this main trend, variation of total fertility rates differs from one country to another. Fertility varies as a consequence of economic shocks, increasing uncertainty in life decisions such as childbearing and changing the cost of children. In this paper we considered fertility time and space dynamics in Sweden with an assumption of relation with labor market.
In the short-run this relation can be explained by two controversial effects of earnings, when child is considered as a "normal good" in the maximization of household utility (1; 2). *Income effect* appears when increasing income

encourages families having more children and contra versa. *Substitution effect* occurs because of childbearing, considered as "household production", substitutes labor supply of a woman. That is why the bigger wage raises the opportunity cost for having a child.

In the long-run, age structure of the population affects total fertility rates and economic explanation of this phenomenon, called Easterlin hypothesis (3), relays to a negative response of fertility on a gain of labor market tightness, when the number of people in their working ages grows. According to this hypothesis, decisions about childbearing, presumably concentrated in the younger generation, have been made on the basis of income potentials compare the younger generation to the older one.

Labor mobility, characterized by fast growth during the last decades in Sweden, contributes in space dynamics of fertility. Flows of in-migrated and out-migrated people increase probability of matches of couples, which entail marriage or cohabiting and having a child. Also, those labor flows generate earning flows across space and affect families' income potential for childbearing.

The purpose of the paper is to study how labor market situation, transformed in households' earnings, influences fertility with taking into account dynamics in time and space. Space diffusion has been considered in transition of fertility norms across municipalities as a first-order spatial autocorrelation; influencing of relative cohort sizes in surrounding municipalities on fertility norms in a given one; and cross-municipal influence of space diffusion of income generated by labor mobility. Time dynamics has been presented by testing of second-order serial autocorrelation of total fertility rates and testing of the direct effect of Easterlin hypothesis and effect of current earnings in a given municipality, as well. For measurement of space diffusion two types of weight matrices have been used to avoid multi-correlation in explanatory variables. The row-standardized contiguity weight matrix has been used for lagged total fertility rates and matrices with spherical distances weighted by population size have been applied for estimating of influencing of the average annual income per capita and relative cohort size on fertility rates.

The contribution of the paper is the analysis of spatial interdependence of fertility in the open municipal demo-economic systems using panel data. In contrast to the previous studies, we consider the Easterlin hypothesis as the long-run impact of earnings and, at the same time, we consider the relation between current earnings and fertility. Thus, panel data allow monitoring space diffusion of fertility norms across time in "three dimensions". We focus on the effects of earnings on fertility by taking into account factors such as educational level and migration flows. In a difference with other papers, we consider municipal level of data, which gives us more detailed result compare to the national or the regional levels. The theoretical contribution of the paper is devoted to such properties of

the model SAR(2,1) as the condition of stationarity and the long-run effect of an explanatory variable in direct, spatial and total forms.
The paper is structured as follows. Section 2 provides the main results of econometric methods for dynamic models in space and time; previous studies of fertility in spatial diffusion; studies of Easterlin hypothesis for Sweden; and results concerning to impact of earnings on fertility in Sweden. Section 3 explains the approach to the estimation and post estimation interpretation such as the short-run and the long-run effects within municipalities, across them and total effects, and stationarity condition for the estimated model. The section 4 describes the data set. Section 5 reveals empirical specification. In the section 6, the results have been presented, and then, they have been summarized in the last section.

Previous Literature

Spatial econometrics of panel data, as a new rapidly growing branch of econometrics, become more and more popular, while the assumption of importance of nearest neighbors on social and economic behavior and activity seems quite evident in many cases. The form of the estimated model, approach to estimation and post-estimation analysis, as well as interpretation of the model and results, were defined by papers of Anselin (2002) (4), Brueckner (2003) (5), Elhorst (2001) (6), LeSage and Pace (2008) (7), and Yua et al. (2008) (8). Brueckner (2003) (5) classifies spatial interactions as the direct (a spillover model), when spatial units reciprocally affect each other, and the indirect (a resource-flow model), when they share some limited resources. Labor market may be considered from both points of view. It gives spillover effect, when matching increases fertility, but the value of demand on labor market, as well as a total earning potential of labor market, to be considered as a share of a resource model. While we cannot say that labor mobility generates only one of the effects, it is helpful to find, that analytically both models go ahead to spatial lag econometric specification (4, p.250). Spatial reaction function can be presented in a lag model, reflecting a global form of spillovers or spatial autocorrelation, or in a local form of spillovers, obtained from spatial lag for an explanatory variable, in our case, in a first-order spatial lags for the average income and cohort ratio in municipalities. Elhorst (2001) (6) thoroughly considers the first-order autoregressive distributed lag model in both space and time, including taxonomy of the model, approach to estimation, stationarity in time conditions, and spatial equilibrium correction model, which provides the static long-run equilibrium relationship between endogenous and explanatory variables. The paper of Yua et al. (2008) (8) gives a better understanding of a general approach to finding stationarity in time condition for spatial modeling. It is important for the paper,

while a second-order time and first-order spatial autoregressive distributed lag model has been estimated. The paper of LeSage and Pace (2008) (7) provides with an interpretation of direct, the indirect or spatial effect and total effects, accepted by other researchers as a standard for such interpretations. The problem of interpretation appears, while for each pair of spatial units the indirect effect is specific, and size of a matrix of effects corresponds to the number of units. They suggested to measure the spatial effect compare to the direct and the total effect. The average direct effect is interpreted as a mean of diagonal elements. The indirect effect for unit i of a variable x is defined as a sum of off-diagonal elements j=1..n of the row i. The average indirect effect is a mean of i=1..n indirect effects. The average total effect is a sum of the average direct and indirect effects.

There exist a number of studies, where spatial diffusion of fertility has been tested as a spreading of knowledge about contraception (9; 10; 11). These papers encompass developing countries or historically remote period of a developed country. However, the number of papers where space dynamics of fertility was studied in a developed country is significantly less (12; 13; 14). The inspiring paper for measuring fertility in spatial dimension is a paper of Waldorf and Franklin (2002) (14), where the Easterlin hypothesis has been tested with consideration of spatial diffusion of fertility by using data for 18 Italian regions. Two kinds of spatial diffusion between spatial units have been estimated in their model, namely space interaction of fertility norms and labor mobility influencing fertility.

Easterlin hypothesis for cross-sectional and panel data is one of the popular measuring of effects of earning on fertility. That is why here we only mention two summing up studies. Macunovich (1998) (15) referred 185 published articles with 76 empirical analyses of the Easterlin hypothesis and concluded, that the results are mixed. Such ambiguous results may be explained by using relative cohort size without any other controls, considering households' incomes and male earnings as interchangeable. Waldorf and Byun (2005) (16) made meta-analysis of 334 empirical papers devoted to the Easterlin hypothesis, which showed more robust negative effect, despite positive effects were more frequent. Negative effect means, that the bigger shear of young generation is connected with a higher total fertility rate.

Easterlin hypothesis for Sweden among other countries was studied by Artzrouni and Easterlin (1966) (17), Baird (1987) (18), Pampel (1993) (19), Wright (1989) (20), Sevilla (2007) (21). The inverted Easterlin hypothesis or a positive relationship between fertility and the share of the younger generation was supported in most of papers, except papers of Artzrouni and Easterlin (1982) (17) and Baird (1987) (18).

Pampel (1993) (19) concluded, that institutional structures such as family policy, and increased female labor force participation, influence cohort relative economic status and, thus, explain the negative effect of relative cohort size. Thus, relative cohort size positively influences fertility only when it is connected with poor opportunities for employment, wages, and promotion. Pampel approved, that state policy for keeping unemployment at a minimum and guarantying jobs, as well as policies against sex discrimination in the labor market, child care system, and maternity leave entail insignificance of the cohort size effect on fertility.
This conclusion is very valuable for understanding relative effect of earnings on fertility in Sweden, where family policy plays a major role in reducing the cost of children. The substitution effect weakens not only due to maternity benefits closely related to woman's own pre-birth earnings and urging women to be employed before childbearing, but also opportunity to come back the labor market despite having small children due to child care system. Studies based on Swedish micro-data support positive effect of earnings on fertility (see 22; 23).
Educational level affects studied relations, and we assume ambiguous character of it. On the one hand, higher level of education serves a guarantee for larger salary, which increases the opportunity costs of childbearing and consequently it may reduce the number of children. Moreover, households with a higher level of education invest more per child (24). Thus, a dilemma of "quantity" and "quality" of children may entail a negative total effect of earnings on fertility for the well educated group. On the other hand, Andersson et al. (2003) (25) found that women with university education have a higher probability of having a third birth than women with lower education. It also can be explained by the family policy in Sweden where women with university education have better opportunities than others to combine work and family due to higher salaries and more flexible working hours. Another interpretation of relation between fertility and education is connected with getting education when time for taking care of children reduces, as well as earnings. That is why status of university cities is usually controlled in regression analysis (25; 26).

Hypotheses to be explored

Interactions between adjacent municipal labor markets, provided by out-migrated and in-migrated people, affect the balance on labor markets, as well as, the spatial income distribution. Particularly, local labor markets' tightness, generated by entering of a bigger new generation, can be smoothed by labor mobility, but worsens the situation on surrounding labor markets. Thus, the inverted Easterlin hypothesis in spatial dimension is expected to be supported by taking into account income potential, connected with tightness of surrounding labor markets,

which reduces opportunities to find a good job or to compete for the higher earnings.

We assume that current incomes and expected earnings changes in them give the effect of earnings, which implies that higher incomes help the family to deal with the direct costs of childbearing. We find that postponing or accelerating of childbearing in the short-run takes place because the Swedish completed cohort fertility rate is rather a constant, but the variation of the total fertility rate about 10% higher than in the other Nordic countries in the studied period. The spatial effect of earnings can be explained as an influence of the average annual income per capita in surrounding municipalities on the total fertility rate in considered municipality. This effect is also generated by migrants for whom earnings statistically observed in the place (municipality) of work, despite they may live and have their families in other municipalities.

Three different hypotheses have been tested in the paper. The first one is **the Easterlin hypothesis,** when dynamics of age structure affecting tightness of labor market and dependence upon income potential of the young generation relatively to the older generation explains decisions concerning childbearing. Cohort ratio is the main explanatory variable. The second hypothesis is the short-run **effect of earnings** when households compare their present income potential with their earnings in the nearest past and expected earnings in the future. This effect should be shifting up total fertility rates during better macroeconomic boom-periods or better family policy regimes and local labor market conditions. The main explanatory variable in view of the second hypothesis is income. Thus, the third hypothesis concerns **spatial effects** in relations between fertility and age structure and between fertility and current earnings. We presume existing of autocorrelation of fertility rates across municipalities generated by labor movements and "earnings' flow". The hypotheses have been considered in the assumption of space diffusion of fertility rates across space. The endogenous variable in the model is the annual total fertility rate in a municipality.

Our research strategy is based on the Arellano-Bond linear dynamic panel-data estimation for LAG type of spatial modeling. We can use it instead of maximum likelihood method in the assumption of only time lagged spatial terms in the model while childbearing is a lagged process in relation to factors affecting decision concerning fertility. The model has been estimated with different sets of municipal variables discussed in the second section.

Different types of matrices, summarizing the spatial morphology of fertility across municipalities, were empirically tested on the best approximation of diffusion of fertility rates and incomes and cohort ratios as the main explanatory variables. Finally, two approaches, suggested in the paper of Waldorf and Franklin (2002) (14) and based on a matrix of contiguity $\mathbf{W}_N$ and a set of matrices $\mathbf{V}_N(t)$ with spherical distances between municipalities, weighted by population,

have been used in the model. In such a case, it allows to escape multi-collinearity and to take into account influence of bigger cities (in a size of population) in the second type of matrices.

The row-standardized matrix $\mathbf{W}_N$ has been constructed in the assumption that fertility is a spatial stationary process, when covariation of fertility rates in two municipalities is only a function of the distance between them. Here $w_{ij} = 1/k_i$ if i,j share a common border, k_i is the number of municipalities bordering i; and $w_{ij} = 0$ otherwise, N is the number of municipalities.

Row-standardized weight matrices $\mathbf{V}_N(t)$ (1), based on spherical distances between municipalities, have been used for summarizing the spatial morphology, generated by influencing of cohort ratio and incomes on fertility:

$$v_{ij}(t) = \frac{Pop_j(t)/d_{ij}}{\sum_k Pop_k(t)/d_{ik}} \qquad (1)$$

where d_{ij} is a spherical distance between i and j municipality. Here we use standard approach (4) for constructing weight matrix, if a potential variable, affecting spatial interaction, exists. In our case, it is a population size of municipalities. In a panel data analysis, bigger cities cannot be controlled directly by a dummy variable or incorporating of a size of population (closed to a constant for a given unit) into a model. The set of v_{ij}, j=1..N for a given i provides higher weights for bigger cities, thus, dynamics, generated by size of the municipality, has been captured in the weight coefficients.

Correspondently, for panel data we use $\mathbf{W}_{NT} = \mathbf{I}_T \otimes \mathbf{W}_N$ [1] and $\mathbf{V}_{NT} = \mathbf{I}_T \otimes \mathbf{V}_N(t)$, t = 1..T, where T is 28 for the period 1981-2008, $\mathbf{I}_T$ is the identity matrix of dimension T.

The model where fertility has a spatial-temporal lag structure and where spatial influencing of age structure and income are extracted has been presented by formula (2). The lagged total fertility rate **W·TFR** reflects recent space-autoregressive dependence of fertility rates. The weighted lagged total fertility rate has been considered as a factor of space diffusion of fertility norms due to possible balancing of age-gender inequality and interactions between people, living in different municipalities, but potentially matching during work and leisure.

$$\mathbf{TFR} = f(\mathbf{TFR}, \mathbf{W}\cdot\mathbf{TFR}, \mathbf{R}, \mathbf{V}\cdot\mathbf{R}, \mathbf{I}, \mathbf{V}\cdot\mathbf{I}, \mathbf{X}). \qquad (2)$$

where **X** is a matrix of explanatory variables. Despite cohort ratios **R** and incomes **I** are already included in the specification implicitly by **W·TFR** and the influence of key variables has been observed by autoregressive component, we

[1] ⊗ is a Kronecker product

can extract their spatial influence by incorporating those variables with spatial lag.

Data

Municipality level data has been used in this study. 276 municipalities for the period 1981-2008 have been included in the analysis (Statistics of Sweden)[2]. Description of variables and descriptive statistics have been presented in tables A1 and A2 in the appendix.
The dependent variable is the total fertility rate (*TFR*), which is accessible from Statistics of Sweden in contrast to cohort fertility ratio, defined as the sum of the age-specific fertility rate (*SFR*) for women in the ages 16-49 in the municipality (3).

$$TFR = \sum_{i=16}^{49} SFR_i \qquad (3)$$

The discussion around introducing relative cohort size, as an indicator of relative economic status, mostly focuses on the age span defining the young generation. Researchers vary the upper boundary from 29 to 34 years in order to maximize experimentally correlation between fertility and age structure. Waldorf and Byun (2005) (16) conclude that a wider range defining the young cohort increases the likelihood of a negative correlation. We also examined mean age of men at birth of the first child, which increased from 26.66 in 1970 to 31.46 in 2006. Based on the previous results by Macunovich (1998) (15) and Waldorf and Byun (2005) (16) and changes of mean age of men at birth of the first child, we consider the cohort ratio (*R*), defined as the number of men of age 35 to 64 year, divided by the number of men of age 15 to 34 and 3-years smoothed cohort ratio (4).

$$R_{jt} = \sum_{\tau=t-2}^{t} M_{35-64,\,j\tau} \Big/ \sum_{\tau=t-2}^{t} M_{15-34,\,j\tau} \qquad (4)$$

The average income per capita in the municipality for people aged 20 and above has been deflated by CPI (1981=100) and taken in logarithm because of its log-normality (ln(*I*)). We assume that it is strongly positively correlated with the average earnings. We cannot specify incomes by gender in the data; thus, only mean households' effects of earnings on fertility can be measured. We emphasis, that *R* measures relative effect of earnings between generations, whenever the

[2] The table A3 provides some information about constructing data with respect to restructuring of municipalities in 1992.

average income responds to this effect in short-run dynamics and across municipalities.
The shares of in-migrated and out-migrated women in the ages 16-49, the shares of net migration, and flow of out-/in-migrated women in the ages 16-49, which indirectly reflect labor market situation, were used in part of the specifications. By preliminary analysis, we found that the share of the total flow of out-/in-migrated women is better in econometric modeling, because the shares of out-migrated and in-migrated women were strongly correlated and influence of the shares of net migration was insignificant.
For control educational levels, we included variables such as the share of women with primary and secondary education 9 years and less (ISCED97 1) (in the ages 16-49), the share of women with post-graduate education (ISCED97 6) and the share of women with post-secondary education 3 years or more (ISCED97 5A) (in the ages 20-49) in the model.
We do not examine the status of marriage or divorce in fertility using the conclusions about insignificance of it for fertility, made by Hoem et al. (2006) (27). With the absence of data for the entire period, we did not take into account relation between the share of employed women and fertility. However, previous studies demonstrate justification of doing it because ambivalent dependence has been found (28; 26). Following earlier papers we expected importance of variables such as size of labor force, mean age of first birth, earned income quintiles, which were not supported. We examined the relation of each variable to total fertility rates by consideration 5-year cohorts (16-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49) for males and females. We found that the character of relations was the same for all cohorts and both sexes. That is why we focus below on aggregate data only for women in the ages 16-49.
One of the tasks was eliminating of business cycle component from the analysis. Sweden gross domestic product (GDP) in constant prices (indexmundi.com) has been considered as a variable reflecting business activity changes.

Empirical Specification

We take into account space component by lagged terms and assume that residuals are not spatially autocorrelated. Several specifications were used in assumption, that total fertility rate dynamics is a stationary process having a cyclical component, which argues incorporating two time lags. Both interesting explanatory variables, incomes and cohort ratio, also contain cyclical components. The cyclical component in incomes is connected with business cycles; the cyclical component in cohort ratio can be explained by demographic cycles. The cohort ratio has a trend, generated by aging of the population. Preliminary analysis of incomes, migration and educational levels data in the

shares shows that they also have trend components. Those features of the data persuade us to use a model with explanatory variables in lags. It is possible by applying Arellano-Bond method of estimation of panel data models with spatially lagged dependent variables. It is based on generalized method of moments' techniques (GMM), which allows finding asymptotically efficient estimates in assumption of autocorrelation between explanatory variables and errors when the stationary process has been considered.

Cyclical component in incomes was controlled by three types of terms in the model. Time t in polynomial structure of the first, second and third order was incorporated in the first approach. Another approach was based on using model with time specific fixed effects and the last approach was based on including of GDP growth as a control variable for time and business cycles. A second-order time and first-order spatial autoregressive distributed lag model SAR (2,1) for estimation has been presented below explicitly in a vector form (5):

$$\begin{aligned}\mathbf{TFR}_t &= \varphi_1\mathbf{TFR}_{t-1} + \varphi_2\mathbf{TFR}_{t-2} + \gamma\mathbf{W}_{NT}\mathbf{TFR}_{t-1} + \mu\mathbf{R}_{t-1} + \lambda\mathbf{V}_{NT}\mathbf{R}_{t-1} + \theta\ln\mathbf{I}_{t-1} \\ &+ \vartheta\mathbf{V}_{NT}\ln\mathbf{I}_{t-1} + \mathbf{X}_{t-1}\beta + g(t)\mathbf{i} + \boldsymbol{\varepsilon}_t\end{aligned} \tag{5}$$

where μ, θ are parameters for cohort ratio and log-income, λ, ϑ are time-space autoregressive parameters for them, φ_k and γ are parameters at time and space lagged total fertility rate, $g(t)$ is one of the specifications describing time component in the model and $\mathbf{i}$ is identity vector of size N, $\boldsymbol{\varepsilon}_t$ is a normally distributed, reciprocally independent vector of errors of size N.

The short-run effect of the lagged *TFR* is $\frac{\partial\mathbf{TFR}_t}{\partial\mathbf{TFR}_{t-1}} = \gamma\mathbf{W}^{/} + \varphi\mathbf{I}_N$. Marginal effect of cohort size ratio and log-income on fertility is $\frac{\partial\mathbf{TFR}_t}{\partial\mathbf{R}_{t-1}} = \lambda\mathbf{V}^{/} + \mu\mathbf{I}_N$ and $\frac{\partial\mathbf{TFR}_t}{\partial\ln(\mathbf{I})_{t-1}} = \vartheta\mathbf{V}^{/} + \theta\mathbf{I}_N$. The long-run effects can be estimated only in models with time specific fixed effects or incorporation of GDP growth instead of time trend. The calculation procedure for the long-run effects, presented in formulae 6-8, is given in the appendix 5. Formulae 6-7 provide spatial effects in the presence of non-zero parameter at spatially lagged variable. If these parameters are insignificant, the formula 8 can be used for estimating the long-run effects of cohort ratio and earnings with substituting of μ and θ instead of β. Interpretation of the results in the average direct, indirect (spatial) and total effects have been calculated and given in the next section.

$$\frac{\partial\mathbf{TFR}}{\partial\mathbf{R}} = \left[((1-\sum_{k=1}^{2}\varphi_k)\mathbf{I}_N - \gamma\mathbf{W})^{-1}(\lambda\mathbf{V} + \mu\mathbf{I}_N)\right]^{/} \tag{6}$$

$$\frac{\partial \mathbf{TFR}}{\partial \ln(\mathbf{I})} = \left[((1 - \sum_{k=1}^{2} \varphi_k)\mathbf{I}_N - \gamma \mathbf{W})^{-1} (\vartheta \mathbf{V} + \theta \mathbf{I}_N) \right]' \quad (7)$$

$$\frac{\partial \mathbf{TFR}}{\partial \mathbf{X}} = \left[((1 - \sum_{k=1}^{2} \varphi_k)\mathbf{I}_N - \gamma \mathbf{W})^{-1} \beta \right]' \quad (8)$$

In the next section we use linear coefficients of elasticity for interpretation changes of fertility when an explanatory variable drifts on 1% from the average value $e = b\frac{\overline{x}}{\overline{y}}$, where b is an estimate of the parameter at x and y is an endogenous variable.

We considered stationarity condition for *TFR,* supposing that the eigenvalues of the matrices $\mathbf{A} = \begin{pmatrix} \varphi_1 \mathbf{I}_N + \gamma \mathbf{W} & \varphi_2 \mathbf{I}_N \\ \mathbf{I}_N & \mathbf{0}_N \end{pmatrix}$, corresponding to SAR(2,1) in a vector-autoregressive form, are less than 1 in absolute value. The result is given in the formula (9), where $\{\mu_i\}$, i=1..N is a set of the eigenvalues of the matrix W. Thus, it is sufficient to estimate maximum and minimum value of λ. The routines have been shown in the app. 6.

$$\lambda_j = \frac{\varphi_1 + \gamma\mu_i \pm \sqrt{(\varphi_1 + \gamma\mu_i)^2 + 4\varphi_2}}{2} \text{ and } |\lambda_j| < 1 \quad (9)$$

Estimated models were tested on absence of serial autocorrelation in residuals. Wald values of the test for accepting or rejecting the hypothesis that each parameter is equal to 0 have been shown for each model in table A4.

Results

The results of estimation with two sets of variables and different time specifications have been presented in table A4. The basic specification for the period 1981-2008 contains relative cohort size *R*, log(*I*), the share of out-/in-migrated women in the ages 16-49. The second specification for the period 1985-2008 includes the same explanatory variables and, in addition, the educational levels for women as the share of women in the ages 16-49 with post secondary education more than 3 years and the share of women in the ages 16-49 with education less than 9 years.

The main result is that the parameter at the weighted TFR_{t-1} is significant and positive in all of the specifications. This fact supports the existence of a spatial positive autocorrelation of *TFR* across municipalities. It means that declining or rising of fertility in one municipality affects neighboring municipalities in the same direction. However, we cannot specify factors generating this spatial

autocorrelation. Below comments to one of the estimates of parameters of the time specific model have been shown in the table 1 and discussed.

The parameter at spatially lagged *TFR* is greater than at time lagged variable (0.032 and 0.43 correspondently). It may be wrongly interpreted as a bigger role of the indirect effect in fertility dynamics than the direct one. However, the average short-run effect of interactions between municipalities is 0.014, which is less than the time lagged parameter or the direct effect. Thus, the indirect effect and the direct effect in the short-run explain 0.014% and 0.032% of relative changes of total fertility rates, correspondently.

The estimates weakly support the direct effect of the inverted Easterlin hypothesis for the analyzed period (1981-2008), while parameter is significant in part of the specifications. It is equal to -0.042 in the given estimation (Table) or 0.034% of TFR's change. Spatially lagged term was not significant, but with taking into account equilibrium condition in the long-run, the indirect effect provides 0.076% of the relative change of fertility.

The parameter at income is negative in each specification, but it is insignificant at spatially lagged income. Thus, we can make a conclusion concerning the direct effect of incomes in the short- and long-run only. For the considered specification (Table), they are equal to -0.319 and -0.376 correspondently. Thus, there exists a weak negative direct effect of earnings on total fertility rates or dominating of the substitution effect between wages and childbearing as a production in households, despite strong Swedish family policy encouraging fertility. The direct short-run effect appears in 0.67% of variation of fertility, but the long-run effect increases it up to 1.49%.

Among control variables, the most important is a share of a sum of in-migrated and out-migrated people in size of the population. Higher labor mobility in both directions is a signal of many processes such as changes of age-gender structure, changes in level of education in a municipality, growing number of matches entailing to cohabiting and marriages, and growing dynamics of labor market demand and supply, as well. We cannot extract the main factor, but evidently internal migration is positively related to childbearing and gives significant direct, indirect and total effects in the long-run such as 1.594, 1.343 and 2.937 correspondently. While the share of in-migrated and out-migrated people in the size of the population is rather low, the total long-run effect captures only 0.001% of relative changes of fertility.

The shares of women with educational level higher or lower than the mean level, positively connected with total fertility rates. Those results are relevant to the results of the previous papers (28; 22) and can be interpreted as a low alternative cost of childbearing for women with low educational level. For women with higher level of education probability of having more than two children is higher than in average (28). It can be explained by their higher benefits during maternity

leave and their better options in distribution of time between being employed and being taken care of a child. With taking into account time and space interactions in the long-run the positive effects of high and low educated shares of women on total fertility rates became greater almost in two times and totally explain 0.105% and 0.054% of variation of fertility in the long-run, correspondently.

Table. Estimates of time-space dynamics in the model describing TFR as a SAR (2,1) process.

Variables			The average short-run effect			The average long -run effect		
	Coeff.	St. dev.	Direct	indirect	Total	Direct	Indirect	Total
Number of observations				6322				
Wald χ^2				9615				
TFR_{t-1}	0.032*	0.015						
Weighted TFR_{t-1}	0.430****	0.032	0.032	0.014	0.132	0	0	0
TFR_{t-2}	0.086****	0.014	0.086					
$Ln(I_{t-1})$	-0.319*	0.154	-0.319	0	-0.319	-0.382	-0.323	-0.706
Weighted $Ln(I_{t-1})$	0.328	0.272						
R_{t-1}	-0.042	0.065						
Weighted R_{t-1}	-0.708	0.610	-0.042	0	-0.042	-0.050	-0.042	-0.093
The share of flow of out/inmigrated women 16-49 age $_{t-1}$	1.328**	0.526	1.328			1.594	1.343	2.937
The share of women 16-49 with postsecondary education more than 3 years $_{t-1}$	0.978****	0.281	0.978			1.174	0.989	2.163
The share of women 16-49 with education less than 9 years $_{t-1}$	1.067****	0.274	1.067			1.281	1.079	2.360
Intercept	1.798****	0.541	1.798					
Stationarity condition $\lambda_j \in [0.16;0.63], j=1..2N$								

[1] Weight matrix has been taken for the period 2007-2008.

The process, described in the model, can be interpreted as a stationary process while eigenvalues vary from 0.16 to 0.63 in the model.
In contrast with the case for Italy (14) where total fertility rates is connected with changes of average age of first birth we do not see any relationship for Sweden. While the average age of first birth increases over time[3], it can be interpreted as postponing of reproduction period but then household plans for kids have been fulfilled.

Conclusions

There is no doubt that changes in fertility may be explained by economic factors. A large number of papers are devoted to theoretical and empirical explanations of reproductive behavior considering female employment, and choices of alternative investment goods like getting an education. New econometric instruments, such as spatial econometrics for panel data, developed recently, allow studying these problems deeper, by taking into account the diffusion of fertility norms and influencing of economic factors across space.
What we did in this studying of fertility in Sweden and what is the contribution of the paper is the analysis of changes in childbearing caused by some factors, which are important for settlements located near to each other. We found that autoregressive component in spatial dimension is positively significant in all of the specifications. It reflects that the whole set of factors, determining total fertility rates, influences it in the same direction as in a given municipality so in the surrounding municipalities.
The main aim of this paper was to test empirically the relation between fertility and earnings in time and space dimension. We find that the inverted Easterlin hypothesis has been supported. It has the short-run direct effect, and the long-run direct and indirect effects. Effect of earnings has the negative direct effect, but the indirect effect across space was not found in the short-run. The specification of the model allows getting the long-run spatial effect, which is negative, as well as the direct effect. It can be explained by dominating of the substitution effect in choice between female labor supply and childbearing as a households' production, even though the family policy in Sweden provides better opportunity to combine them.
The stationary condition as a theoretical contribution of the paper contains an explicit form of a matrix of a size $2N \times 2N$, which has been got on basis of the spatial weights matrix and corresponds to the vector-autoregressive form of the

[3] For the period 1970-2008 it was increasing about 1.48 month or 45 days a year, from 24.03 in 1970 to 29.02 in 2006.

model SAR (2,1). Eigenvalues for the matrix have been obtained in an analytical form. By using this form, stationarity of fertility in time and space dynamics, measured by total fertility rates, has been empirically proved.
The long-run effect of an explanatory variable in the model SAR (2,1), which has been analytically obtained, implies direct, indirect and total effects. It has been empirically shown that the spatial effect in the long-run almost doubles the influence of the log-income and the cohort ratio on fertility. That is why consideration of spatial disturbances is important for modeling fertility process.

References

1. G.S. Becker, Barro, R.J. *A Reformulation of the Economic Theory of Fertility.* The Quarterly Journal of Economics, Vol. 103, No. 1: 1-25 (1988).
2. G.S. Becker *An economic analysis of fertility - Demographic and Economic Change in Developed Countries.* Universities-National Bureau: 225–256 (1960).
3. R.A. Easterlin *On the relation of economic factors to recent and projected fertility changes*. Demography, 3:135-153 (1966).
4. L. Anselin *Under the hood. Issues in the specification and interpretation of spatial regression models.* Agricultural Economics 27: 247-267 (2002).
5. J.K. Brueckner *Strategic Interaction Among Governments: An Overview of Empirical Studies*. International Regional Science Review 26, 2: 175–188 (2003).
6. J.P. Elhorst *Dynamic Models in Space and Time*. Geographical Analysis, Vol. 33, No. 2 (2001).
7. J.P. LeSage, Pace, R.K. *Introduction to spatial econometrics*, Taylor & Francis CRC Press, Boca Raton (2008).
8. J. Yua, de Jong, R., Lee, L. *Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and T are large*. Journal of Econometrics, 146: 118–134 (2008).
9. J.R. Weeks, Getis, A., Hill, A.G., Gadalla, M.S. and Rashed, T. The *Fertility Transition in Egypt: Intraurban Patterns in Cairo*. Annals of the Association of American Geographers, 94:1: 74-93 (2004).
10. J. Bongaarts and Watkins, C.S. *Social Interactions and Contemporary Fertility Transitions*. Population and Development Review, Vol. 22, No. 4: 639-682 (1996).
11. R. Woods *Social Class Variations in the Decline of Marital Fertility in Late Nineteenth-Century London.* Geografiska Annaler. Series B, Human Geography, Vol. 66, No. 1: 29-38 (1984).
12. M.C. de Castro *Spatial Demography: An Opportunity to Improve Policy Making at Diverse Decision Levels*. Population Resource Policy Review 26: 477–509 (2007).
13. G. McNicoll *Institutional Determinants of Fertility Change.* Population and Development. Review, Vol. 6, No. 3: 441-462 (1980).

14. B. Waldorf, Franklin R. *Spatial dimension of the Easterlin hypothesis: fertility variations in Italy.* Journal of regional science, vol. 42, No. 3: 549-578 (2002).
15. D.J. Macunovich *Fertility and the Easterlin hypothesis: An assessment of the literature*. Population Economy 11:53–111 (1998).
16. B. Waldorf, Byun P. *Meta-analysis of the impact of age structure on fertility.* Journal of Population Economy 18:15–40 (2005).
17. M.A. Artzrouni, Easterlin R.A. *Birth History, Age Structure, and Post World War II Fertility in Ten Developed Countries: An Exploratory Empirical Analysis*. Genus 38 (3–4): 81–99 (1982).
18. A.J. Baird *A Note on the Easterlin Model of Fertility in Northwestern Europe and the United States: 1950–1981*. International Journal of Comparative Sociology 23 (1–2) 57–68 (1987).
19. F.C. Pampel *Relative Cohort Size and Fertility: The Socio-Political Context of the Easterlin Effect*. American Sociological Review 58 (August):496–514 (1993).
20. R.E. Wright *The Easterlin Hypothesis and European Fertility Rates.* Population and Development Review, Vol. 15, No. 1: 107-122 (1989).
21. J.P. Sevilla *Fertility and relative cohort size*. Arbetsrapport. Institutet för Framtidsstudier; 2007:11 (2007).
22. G. Andersson *Childbearing Trends in Sweden 1961–1997*. European Journal of Population 15: 1–24 (1999).
23. B. Hoem *Entry into Motherhood in Sweden: the Influence of Economic Factors on the Rise and Fall in Fertility, 1986-1997*. Demographic Research, vol. 2, art. 4 (2000).
24. G.S. Becker *Fertility and the Economy.* Journal of Population Economics, Vol. 5 (3):185–201 (1992).
25. G. Andersson, Duvander, A., and Hank, K. *Do Child Care Characteristics Influence Continued Childbearing in Sweden? An Investigation of the Quantity, Quality, and Price Dimension*. MPIDR WORKING PAPER WP 2003-013 (2003).
26. T. Westerberg *More Work, Less Kids – The Relationship Between Market Experience and Number of Children.* Umeå Economic Studies. No. 683 (2006).
27. J.M. Hoem, Neyer, G. and Andersson G. *Educational attainment and ultimate fertility among Swedish women born in 1955-59.* MPIDR WORKING PAPER WP 2006-004 (2006).
28. D. Berinde *Pathways to a Third Child in Sweden.* European Journal of Population 15: 349–378 (1999).
29. J.R. Silvester *Determinants of Block Matrices*. The Mathematical Gazette. Vol. 84, No. 501, Nov., 2000: 460-467 (2000).

Appendix

Table A1. List of description of variables

Total fertility rate (TFR)	Total fertility rate by region, sex and period: The data is based on the Historical population register and the Multi-Generation Register. The Multi-Generation Register is updated continuously with links between mother/child and father/child. The link between father child is registered with some time delay compared to the link between mother and child. This implies that the numbers are uncertain for fathers in 2007. The statistics could differ a little from official statistics depending on yearly corrections of the registers. Thus, the whole series is updated each year for the complete period.
Relative cohort size (R)	The number of men of age 35 to 64 year divided by the number of men of age 15 to 34 smoothed in 3-years
The share of women 16-49 with education • **women 16-49 with education less than 9 years (ISCED97 1)** • **post-graduate education (ISCED97 6)** • **post secondary education more than 3 years (ISCED97 5A)**	Women 16-49 with given level of education / number of women in their 16-49 age. Level of educational attainment by region, gender, age, income class and time: The classification by level of educational attainment is according to the Swedish national educational classification (SUN)
Income	Total income, average income for residents in Sweden on 31/12, thousand by region, gender, age, income class and time
Income_CPI	Income adjusted by CPI to the basic level in 1981
In-migrated and out-migrated women	In-migrated and out-migrated women by region, age, sex and the period: When calculating age-specific rates per 1 000 of the mean population, the mean population for the year of birth should be used age: Age refers to age attained by the end of the year, i.e. in principle, an account for the year of birth.
The share of balance out-/in-migrated women 16-49 age	(In-migrated women in their 16-49 age – out-migrated women in their 16-49 age)/number of women in their 16-49 age
The share of flow out-/in-migrated women 16-49 age	(In-migrated women in their 16-49 age + out-migrated women in their 16-49 age)/number of women in their 16-49 age
Mean age of woman at birth of the first child (AAFB)	Mean age at birth of the first child by region, sex and period: The data is based on the Historical population register and the Multi-Generation Register. The Multi-Generation Register is updated continuously with links between mother/child and father/child. The link between father child is registered with some time delay compared to the link between mother and child. This implies that the numbers are uncertain for fathers in 2007. The statistics could differ a little from official statistics depending on yearly corrections of the registers. Thus, the whole series is updated each year for the complete period.
Population size (100 thou.)	Mean population (by year of birth) by region, age, sex and period: When summing up the mean population (for instance into 10-year

	groups), any rounded figures are accumulated. This may cause that the accumulated figure becomes somewhat higher than the totals. The mean population refers to the mean value of, for instance, the number of 5-years old at the end of year n and the number of 6-years old at the end of year n+1. The mean population should be used together with vital data that have occurred during the year and distributed by age at the end of the year.
Gross domestic product (GDP) in constant prices	Annual percentages of constant price GDP are year-on-year changes; the base year is 1990.

Table A2: Descriptive statistics for 276 municipalities during 1970-2008

Variable	Obs	Mean	Std. Dev.	Min	Max
Total fertility rate (TFR)	7452	1.929	0.301	0.91	3.31
Mean age at birth of the first child, female (AAFB)	7452	26.596	1.525	22.33	33.07
Mean age at birth of the first child, male (AAFB)	40	29.113	1.566	26.66	31.46
Income to CPI, 1981=100%	4968	58.189	9.427	39.770	180.705
The share of balance out/inmigrated women 16-49 age	7422	-0.001	0.006	-0.022	0.042
The share of flow of out/inmigrated women 16-49 age	7422	0.061	0.019	0.023	0.205
Relative cohort (R)	7452	1.572	0.249	0.952	2.369
The share of women 16-49 with education less than 9 years	7452	0.044	0.046	0.101	0.231
The share of women with post-graduate education	7452	0.002	0.003	0.004	0.047
The share of women 16-49 with post secondary education more than 3 years	7452	0.094	0.064	0.030	0.456
Population size (100 thou.)	7452	0.312	0.566	0.028	8.026
GDP in constant price[1]	28	2.375	1.825	-2.058	4.66

[1] the data for the period 1981-2008

Appendix A3. The municipality restructuring 1992

In 1992 municipality reform was performed in Sweden after that the number and boundaries of municipalities had been changed. We excluded those municipalities, which have been established or annulled due to the reform. Thus, the total number of municipalities is 276. Weight matrices with using spherical distances were not changed due to the reform because of the same geographical coordinates of the municipal center. Weight matrices with using contiguity were changed, but we assume transformation of boundaries was not large for misleading of results. After the reform population changed more than 2% only in 9 municipalities (only in Vaxholm and Värmdö population changed more than 3%), but the total change in population was not greater 5%. Used variables have been merged by SCB.

Table A4: GMM estimates for *TFR*. The model with time specific fixed effects for the period 1981-2008.

Variables	No spatial effects				SAR				SAR with exogenous spatial interaction effect							
	1		2		3		4		5		6		7		8	
	Coeff.	St. dev.	Coeff.	St. dev.	Coeff.	St. dev.	Coeff.	St. dev.	Coeff.	St. dev.	Coeff.	St. dev.	Coeff.	St. dev.	Coeff.	St. dev.
Number of observations	6596		6322		6596		6322		6596		6596		6322		6322	
Wald χ^2	8406		9007		9963		9612		9936		9388		9007		9615	
TFR $_{t-1}$	0.095****	0.015	0.044**	0.016	0.033*	0.015	0.033*	0.015	0.033*	0.015	0.037*	0.015	0.044**	0.016	0.032*	0.015
TFR $_{t-2}$	0.148****	0.014	0.100****	0.015	0.084****	0.014	0.086****	0.014	0.084****	0.014	0.092****	0.015	0.100****	0.015	0.086****	0.014
***Weighted* TFR $_{t-1}$**					0.422****	0.032	0.431****	0.032	0.423****	0.032					0.430****	0.032
Ln(I_{t-1})	0.961****	0.100	0.120	0.130	-0.708	0.194	-0.220*	0.130	-0.759****	0.200	-0.848****	0.203	-0.017	0.156	-0.319*	0.154
***Weighted* Ln(I_{t-1})**									0.306	0.269	0.247	0.274	0.398	0.279	0.328	0.272
R_{t-1}	-0.225****	0.063	-0.014	0.066	-0.215*	0.059	-0.050	0.065	-0.198***	0.061	-0.251****	0.062	-0.011	0.067	-0.042	0.065
***Weighted* R_{t-1}**									-0.660	0.597	-0.458	0.608	-0.639	0.625	-0.708	0.610
The share of flow of out/inmigrated women 16-49 age $_{t-1}$	2.398****	0.550	1.889****	0.537	1.435**	0.518	1.320**	0.526	1.444**	0.518	1.902****	0.528	1.893****	0.537	1.328**	0.526
The share of women 16-49 with postsecondary			1.677****	0.265			1.078****	0.263					1.500****	0.284	0.978****	0.281

education more than 3 years$_{t-1}$																
The share of women 16-49 with education less than 9 years$_{t-1}$			1.498****	0.272			1.095****	0.268					1.428****	0.279	1.067****	0.274
Intercept	-2.076****	0.349	1.105*	0.472	4.126****	0.828	1.740****	0.463	4.185****	0.991	5.299****	1.005	0.946*	0.551	1.798***	0.541

**** means p-value less than 0.0001, *** - 0.001, ** - 0.01, * - 0.1.

Appendix 5
The long-run effect of an explanatory variable in the model SAR (2,1)
Consider the model SAR (2,1) with a second-order autoregressive disturbance, a first-order spatial disturbance weighted by a spatial weights matrix **W** of size $N \times N$ and a first-order lagged common factor weighted by some spatial matrix **V** of size $N \times N$.

$$\mathbf{y}_t = \mu\mathbf{i} + \alpha_1\mathbf{y}_{t-1} + \alpha_2\mathbf{y}_{t-2} + \beta\mathbf{W}\mathbf{y}_{t-1} + \gamma\mathbf{x}_{t-1} + \delta\mathbf{V}\mathbf{x}_{t-1} + \boldsymbol{\varepsilon}_t,$$

where ε_t is a "white noise" vector of size $N \times 1$, $\mathbf{y}_t$ is a vector $N \times 1$ of observations

of endogenous variable for every spatial unit at time t, x_t is a vector explanatory variable at time t, $\mathbf{i}$ is a vector of units, α, β, γ, δ are parameters and μ is an intercept. If y_t converges to equilibrium, the equilibrium condition follows from the next equation:

$((1-\alpha_1-\alpha_2)\mathbf{I} - \beta\mathbf{W})\mathbf{y} = \mu\mathbf{i} + (\gamma\mathbf{I} + \delta\mathbf{V})\mathbf{x}$, where $\mathbf{I}$ is the identity matrix $N \times N$.

Thus, the long-run impact of x on y yields

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \left[((1-\alpha_1-\alpha_2)\mathbf{I} - \beta\mathbf{W})^{-1}(\gamma\mathbf{I} + \delta\mathbf{V}) \right]^{/}.$$

Appendix 6
Stationary condition for the model SAR(2,1)
Consider stationarity condition in a second-order serial and a first-order spatial autoregressive distributed lag model SAR(2,1).

$$\mathbf{y}_t = \alpha_0\mathbf{i} + \sum_{k=1}^{2} \alpha_k \mathbf{y}_{t-k} + \beta\mathbf{W}\mathbf{y}_{t-1} + \gamma\mathbf{x}_{t-1} + \boldsymbol{\varepsilon}_t$$

where $\mathbf{W}$ is a spatial weights matrix of size $N \times N$, ε_t is a "white noise" vector of

size $N \times 1$, $\mathbf{y}_t$ is a vector $N \times 1$ of observations of endogenous variable for every

spatial unit at time t, x_t is a vector explanatory variable at time t, $\mathbf{i}$ is a vector of units, α_1, α_2, β, γ are parameters and α_0 is an intercept.
The vector auto-regression form (VAR) for this model yields:

$$\begin{pmatrix} \mathbf{y}_t \\ \mathbf{y}_{t-1} \end{pmatrix} = \begin{pmatrix} \alpha_1\mathbf{I} + \beta\mathbf{W} & \alpha_2\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \end{pmatrix} + \begin{pmatrix} \gamma\mathbf{x}_{t-1} + \alpha_0\mathbf{i} \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_t \\ \mathbf{0} \end{pmatrix}.$$

Denote $\mathbf{A}=\begin{pmatrix}\alpha_1\mathbf{I}+\beta\mathbf{W} & \alpha_2\mathbf{I}\\ \mathbf{I} & \mathbf{0}\end{pmatrix}$. Stationarity condition implies $|\lambda_i|<1$ where λ_i, $i=1..2n$ is a characteristic root of the matrix A or a root of a characteristic equation $|\mathbf{A}-\lambda\mathbf{I}|=0$, which is a polynomial in λ.

The matrix $\mathbf{A}-\lambda\mathbf{I}=\begin{pmatrix}\alpha_1\mathbf{I}+\beta\mathbf{W}-\lambda\mathbf{I} & \alpha_2\mathbf{I}\\ \mathbf{I} & -\lambda\mathbf{I}\end{pmatrix}$ has a property of a block matrix, described by Silvester (2000, p. 463). According to that property

$$|\mathbf{A}-\lambda\mathbf{I}|=\begin{vmatrix}\alpha_1\mathbf{I}+\beta\mathbf{W}-\lambda\mathbf{I} & \alpha_2\mathbf{I}\\ \mathbf{I} & -\lambda\mathbf{I}\end{vmatrix}=|(\alpha_1\mathbf{I}+\beta\mathbf{W}-\lambda\mathbf{I})(-\lambda\mathbf{I})-\alpha_2\mathbf{I}|=$$

$$|\lambda^2\mathbf{I}-(\alpha_1\mathbf{I}+\beta\mathbf{W})\lambda-\alpha_2\mathbf{I}|$$

Let **M** is a matrix of characteristic roots $\{\mu_i\}$, $i=1..N$ of the weight matrix **W** of size $N\times N$. By the basic property $|\mathbf{M}|=|\mathbf{W}|$ and there exists a nonsingular matrix **R** of characteristic vectors of size $n\times n$ such that $\mathbf{RMR}^{/}=\mathbf{W}$, $\mathbf{RR}^{/}=\mathbf{I}$. With using this decomposition and given properties the characteristic equation for **A** transforms into the next form:

$$|\lambda^2\mathbf{I}-(\alpha_1\mathbf{I}+\beta\mathbf{W})\lambda-\alpha_2\mathbf{I}|=|\lambda^2\mathbf{RIR}^{-1}-(\alpha_1\mathbf{RIR}^{-1}+\beta\mathbf{RMR}^{-1})\lambda-\alpha_2\mathbf{RIR}^{-1}|=$$

$$|\mathbf{R}[\lambda^2\mathbf{I}-(\alpha_1\mathbf{I}+\beta\mathbf{M})\lambda-\alpha_2\mathbf{I}]\mathbf{R}^{-1}|=0$$

While **R** is a nonsingular matrix and $\lambda^2\mathbf{I}-(\alpha_1\mathbf{I}+\beta\mathbf{M})\lambda-\alpha_2\mathbf{I}$ is a diagonal matrix, determinant is equal to zero, when λ is a root of the polynomial $\lambda^2-(\alpha_1+\beta\mu_i)\lambda-\alpha_2$.

Thus, $\lambda_j=\dfrac{\alpha_1+\beta\mu_i\pm\sqrt{(\alpha_1+\beta\mu_i)^2+4\alpha_2}}{2}$, $j=1..2n$.

Technology oversight and classification from topic modeling of patent data

Duncan Kushnir

Environmental Systems Analysis, Chalmers University of Technology
Göteborg, Sweden
Email: kushnir@chalmers.se

Abstract: This paper demonstrates some practical applications of topic modeling to technology and environmental foresight through patent data. Topic modeling methodology considerations are given along with implementation needs. An analysis of patents as a whole and carbon nanotechnology forms the backdrop to support the discussion.
Keywords: Topic Modeling, Patents, Technology, Assessment, Environment, Foresight

1 Introduction

The rapid proliferation of new materials and technologies presents a massive challenge to societal foresight efforts. As a consequence, the time lag between the initial industrial use of a new technology or material and regulatory awareness of its existence or environmental implications is often measured in years, greatly reducing the effectiveness of any policy response. The rapid identification of priority applications and materials needing some form of precautionary assessment is the key bottleneck. Computational approaches to analysis that can aid and prioritize precious human analyst time are therefore of critical importance.

The databases of the international patent organizations are a vast and interesting public resource that is constantly updated with new innovations. There are a number of data sources other than patents that should be viable for analyzing the conception and spread of technology. Scientific papers, news items and even science fiction are all other candidates that could be analyzed on their own or in parallel. Patents are chosen because ideally they represent a practical incarnation of a technology and are thus one of the first publicly observable sources for what could be 'real'.

A wealth of patent analysis literature exists mostly centered around innovation, competitive advantage and forecasting. The analytical challenge here is slightly different: needs for foresight include detection, flagging patents by some arbitrary criteria and clustering/classification to simplify an overwhelming mass of data. The key methodological problems are therefore similarity functions and grouping algorithms. This paper and presentation discusses the utility of statistical techniques such as topic modeling for these needs. This is illustrated here through analyses of both USPTO data and a smaller subset representing carbon nanotechnology.

2 Data set and preprocessing

Patent data is drawn from a custom built database incorporating all USPTO data from 1976 onwards, which at present is freely available via Google in compressed archives totaling some 100Gb. The carbon nanotech corpus comprises 2612 unique patent records with referential links to a background of an additional 11532 patents which are useful as a training corpus, and will be available via email. The corpus is large enough for statistics and small enough that every document in the core set can be read by an expert, thus providing an avenue for human verification of the methods. The goal of current work is to produce a multi-method topic-based network, clustering and search engine comprising all roughly 4.2 million patents in the data set. At present, reference networks, tf-idf and several lda models are implemented for the entire USPTO post-1976.

The title, text and brief description fields along with metadata were extracted and form the basis for modeling. As a quick observation regarding preparation, normal stemming should be used with care on patent data. For instance compos*: composing, composed and composite all have very different meanings in patents. If stemming is to be used, then a customized stemmer exempting important key words should be implemented. Simple stemming for plurals, lemmatizers or word labelers (noun, etc.) are all viable alternatives.

3 The need for new patent network analysis methods

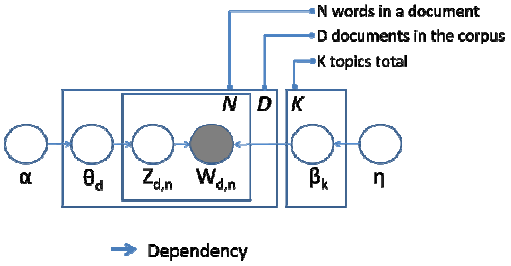
The most commonly used similarity metrics for patent network analysis, as distinguished from simple counts and classes, are the bibliographic approaches: bibliographic coupling, or inferring a relation when two patents cite the same antecedent (Kessler 1963; Marshakova 1973; Small 1973) and co-citation analysis, or inferring a relationship when two patents are referenced together (Glänzel and Czerwon 1996; Small 1998). Both are in widespread use for network analysis in innovation systems but also have serious deficiencies for technology detection and classification (e.g. ; Hicks 1987; Leydesdorff 2005; Barberá-Tomás et al. 2011). Examples of such issues include producing sparse matrices, lack of consensual citing and the lack of the time lag (e.g. currently just under two years per generation for the CNT data set and bit more than two for patents as a whole) that is necessary to build citations. Hicks first noted that these deficiencies have a more marked effect on emerging field detection because of a lack of ‘critical mass’ producing ‘weakly and spuriously connected’ graphs thus greatly discounting the practical value of network approaches with bibliographic metrics. Despite these issues, connectivity analysis of reference networks is of value when it can be validated externally (Verspagen 2007; Barberá-Tomás et al. 2011). Furthermore, such metrics have uses other than technology detection and classification however, as referencing may be a valuable indicator of patent commercial value although even this is now being convincingly questioned (Gambardella et al. 2008).

Text based approaches provide a solution to these issues: in principle, a similarity

score can be calculated between any two nodes, allowing networks of essentially arbitrary richness or connectivity. Furthermore the similarity score does not have any unknown potential disqualifiers as could be the case for spurious references. Completing the case, text methods allow for search, retrieval and training functions which are of immense practical value.

4 Topic modeling for similarity metrics

Topic modeling is one possible approach to derive an indexing structure from text within documents. Other possibilities could include TF-IDF, log-entropy models, random projections, etc. Implemented in parallel, it is likely that they can be used to correct each other's deficiencies. The basis of topic modeling is a probabilistic model of how data is generated. The models employed here are based on Latent Dirichlet Allocation, or LDA (Blei et al. 2003). The core assumption in LDA is that some set of topics, each of which is a distribution of words over some vocabulary, exists for the full document set (the corpus). Each document is then assumed to be have been generated by first picking a distribution of topics contained in the document, and then by sequentially picking each word 'randomly' by first picking a topic for that word from the topic distribution of the document and then by picking from the word distribution of that topic. This process and chain of dependencies is summarized below:



Symbol	Description
α	Dirichlet prior distribution of θ
θ_d	Distribution of topics in document d
$Z_{d,n}$	Specific topic for document d, word n
$W_{d,n}$	Word n of document d
β_k	Word distribution for topic k
η	Dirichlet prior distribution of β

Figure 1: Variable dependencies in the basic LDA generative model

The challenge is to invert this process and to use the corpus (the full set of documents) to infer this set of hidden variables which will ideally correspond to uncorrelated indexing dimensions. The conditional distribution of the hidden variables given the observed words is equal to the joint distribution of all variables divided by the marginal probability of observing a given corpus using all possible topic models (Blei et al. 2003). Using the notation from Figure 1:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} \mid w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})} \quad (1)$$

Eq. 1 will be recognized as an instance of Bayes' Theorem. While the numerator is easy to calculate for each possible setting of the hidden variables, the denominator

(number of possible observable corpuses) grows rapidly with the corpus length and vocabulary and is in practice intractable to compute for any interesting corpus.

4 Implementation choices

While there are many possible approaches to compute an approximation of the latent structure, application to large groups of patents primarily requires computational speed. For this reason, the online LDA algorithm (Hoffman et al. 2010) was selected as the baseline for purposes of expandability and comparability and was used as-implemented in the open-source gensim framework (see software refs). Better alternatives for smaller networks would be the Correlated Topic Model or CTM which has assumptions more appropriate to patent data (Blei and Lafferty 2006) but which will not scale so well. An online implementation of the heirarchical dirichlet process (HDP) model is now available and could possible supercede oLDA as the baseline model in the future for reasons discussed below.

Patent data has features that strain some of the implicit assumptions (Blei et al. 2003) of the generic LDA model. The most relevant are:

1) The appropriate number of topics for a model is difficult to know and is presumably not static. The oLDA algorithm is static and overfitting data is certainly a potential issue. At present, this was resolved by heuristic approximating measures (Arun et al. 2010). In the future the online implementation of the HDP model which determines topic numbers dynamically could be a solution.

2) The word distribution of ideal topics presumably changes over time: technology is evolutionary and patents are ordered in time, thus topics are expected to drift. Ideally models should account for this, but at present the bias of online methods towards initial documents was sidestepped by feeding the training documents in in random order, producing a ‘timeless overview’ topic model.

3) There is evidence (M.Wallach et al. 2009) that choosing an asymmetric prior (α) for the distribution of topics in a document (θ) should be a superior choice. This is intuitively reasonable because for patent data there is no reason to suppose that all topic distributions are equally likely, most particularly in some chosen subset. This is one key rationale behind the CTM, but while the CTM model is computable in reasonable time for the CNT set, current implementations are certainly not viable at present for a million+ document corpus. The specific choice of any symmetric α seems to have no observable effect on the accuracy or convergence of the results with the online implementation. Using asymmetric prior vectors with oLDA caused a twofold increase in processing time but seems to have some potential value worth investigating for sharpening the resulting topics.

5 Practical applications for technology oversight

Broad trends in patenting
By indexing model topics into time slices, the develoment of broad trends in patent language can be observed as per the figures below. Trends sliced by any metadata such as country, company, etc. can reveal information about whether the relevant set is following (or leading) broader trends in patent language.

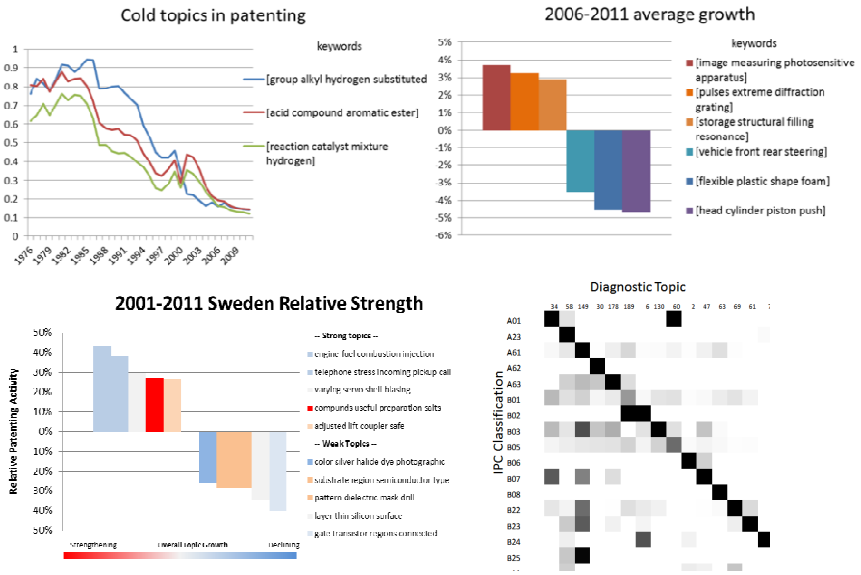


Figure 2: Topic trends, growth, national position and comparison to IPC classification codes

While topics alone can give useful insight, the overall coarseness of topic models is plain to see for those familiar with the patents themselves. For instance, the relative weakness in Sweden's patenting of electronics is a known result, while the [compounds useful preparation salts] segment of the above picture is not immediately comprehensible (it refers to a certain sector of Sweden's pharmaceutical industry). Topic models on data of this scope therefore need additional methods to produce more specific insights. Broad results can be further refined by attribution of diagnostic topics to other metadata categories (Griffiths and Steyvers 2004) as in the figures above, or by separating 'background topics' from noise (AlSumait et al. 2009).

Similarity queries
After producing the topic model, it is a trivial task to convert new documents of any form into the vector space and compare them against the corpus. Thus the

topic model can serve as either an effective search engine, or for using similarity to return patents corresponding to abstract criteria. This application of indexing has potentially wide use in patent intelligence. As one possible example, nanoparticles suspended in liquids have been shown to likely be 'riskier' than other forms, e.g. bound in composites (Hansen et al. 2008). The description of the category reads:

'*The category IIIb consists of systems where nanoparticles are suspended in a liquid. An example of this is nanoparticles inside T cells, used for targeted drug delivery (Mortensen et al. 2006), or TiO2 in cosmetics or paint (Wolfe 2005; Miller et al. 2006). Most biological nanosystems belong to category IIIb'*

Converting this paragraph into an LDA vector search and comparing to the LDA vectors of the rest of the CNT set ranks related patents by topic closeness. The top of the list consists of patents for permanently stabilized nanoparticles in 'fracking' fluid, methods for producing stabilized nanoparticles out of 'household ingredients', methods for solubilizing all types of carbon nanoparticle, and a cosmetics patent: all are applications with known and peer-reviewed environmental risk that are not yet the target of precautionary investigation or public knowledge.

Networks, grouping and cluster classification
A document's similarity to other documents can function as network edges or be transformed into a spatial distance usable for clustering algorithms. In smaller sets, it is possible to compute all edges, but the number grows exponentially with corpus size. For four million patents, there are trillions of edges. Without advanced strategies that are out of scope here, the unfortunate reality is that calculating the most appropriate connections for each node requires computing similarities against the whole corpus and sorting; with the current implementation (i7 2600k@5GHz), this incurs a cost of about 200ms per query against the full set, or roughly 230 hours computation time to calculate the edge list on a fast consumer computer. Using CUDA acceleration shows promise of decreasing processing time by an order of magnitude. Luckily, the task runs in constant memory (about 10Gb required for the whole set) and is distributable; thus disk space is the only limit on how many preferred edges can be saved.

Even for smaller sets such as the CNT example, the display of the full data is visually uninformative and therefore clustering the data helps simplify and display structure. The info-map algorithm (Rosvall and Bergstrom 2008) is used for the results here, it is an efficient choice for both finding community structure in large data sets and for finding changes in community structure over time (Rosvall and Bergstrom 2010). An application to the full edge network for the CNT set produced a set of clusters that can be tracked through time as shown below. From such an approach the structure of a technology can be calculated for visualization nearly instantly. Combining community detection with search can allow simple queries to bring back a large number of 'local' patents and to parse them into comprehensible structures.

The clustered structure for the CNT set corresponds quite neatly to application areas of carbon nanotechnology (labeled by hand, but exhibiting high cohesion and modularity in topic groups) and seen over time it quickly raises interesting questions about where CNT is heading (catalysts, organic complexes, logic structures, graphene).

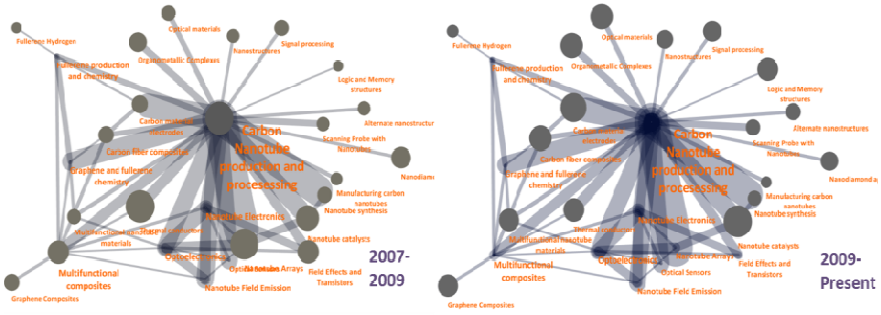


Figure 3: Clustering of CNT set at two different time intervals

Novelty detection

The implementation here offers many potential avenues for detecting 'novel' applications: dictionary additions, new topics being spawned in a HDP model, and network community fission/migration are all directly viable approaches. It is hoped that future work will bring these aspects more into focus.

6 Conclusions

This paper has presented the idea that topic modeling is useful to extract structural similarities from patent data. Topic modeling is a viable method for a number of important tasks that may not be possible with bibliometric analysis or artificially imposed categorizations such as classification codes. The capability of computing similarity is valuable in its own right, but can be extended to evaluate reference networks or to serve as the basis of automatic clustering and evaluating change over time. The properties of networks based on topic model similiarities are not yet well studied or understood but appear to present a practical and robust tool.

With present and widely available computing capacity, topic modeling is tractable for tasks on the scale of millions of documents. As such, topic modeling and the statistical theory behind it is sufficiently ready to be widely employed as one tool to extend our capability of monitoring change in the language of science as a whole as represented by patents or by papers or other media. The implementation presented for consideration here owes much to decades of theoretical work by the statistics, mathematic, computer science and topological communities.

References

Basic software references:

Python: http://www.python.org/
Gensim: http://radimrehurek.com/gensim/
USPTO dataset: http://www.google.com/googlebooks/uspto.html

Papers:

AlSumait, L., D. Barbara, et al. (2009). Topic Signicance Ranking of LDA Generative Models. Fairfax, VA, Department of Computer Science, George Mason University.

Arun, R., V. Suresh, et al. (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations

Advances in Knowledge Discovery and Data Mining. M. Zaki, J. Yu, B. Ravindran and V. Pudi, Springer Berlin / Heidelberg. 6118: 391-402.

Barberá-Tomás, D., F. Jiménez-Sáez, et al. (2011). "Mapping the importance of the real world: The validity of connectivity analysis of patent citations networks." Research Policy 40(3): 473-486.

Blei, D. and J. Lafferty (2006). Correlated Topic Models, Carnegie Mellon University.

Blei, D. M., A. Y. Ng, et al. (2003). "Latent dirichlet allocation." J. Mach. Learn. Res. 3: 993-1022.

Gambardella, A., D. Harhoff, et al. (2008). "The value of European patents." European Management Review 5(2): 69-84.

Glänzel, W. and H. Czerwon (1996). "A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level." Scientometrics 37(2): 195-221.

Griffiths, T. L. and M. Steyvers (2004). "Finding scientific topics." Proceedings of the National Academy of Sciences of the United States of America 101(Suppl 1): 5228-5235.

Hansen, S., E. Michelson, et al. (2008). "Categorization framework to aid exposure assessment of nanomaterials in consumer products." Ecotoxicology 17(5): 438-447.

Hicks, D. (1987). "Limitations of Co-Citation Analysis as a Tool for Science Policy." Social Studies of Science 17(2): 295-316.

Hoffman, M., D. M. Blei, et al. (2010). Online Learning for Latent Dirichlet Allocation. Princeton, NJ, Princeton University.

Kessler, M. M. (1963). "Bibliographic coupling between scientific papers." American Documentation 14(1): 10-25.

Leydesdorff, L. (2005). "Similarity measures, author cocitation analysis, and information theory: Brief Communication." J. Am. Soc. Inf. Sci. Technol. 56(7): 769-772.

M.Wallach, H., D. Mimno, et al. (2009). Rethinking LDA: Why Priors Matter, University of Massachussets Amherst.

Marshakova, I. V. (1973). "Document coupling system based on references taken from ScienceCitation Index (in Russian)." Nauchno-TeknicheskayaInformatsiya 2(6.3).

Rosvall, M. and C. T. Bergstrom (2008). "Maps of random walks on complex networks reveal community structure." Proceedings of the National Academy of Sciences 105(4): 1118-1123.

Rosvall, M. and C. T. Bergstrom (2010). "Mapping Change in Large Networks." PLoS ONE 5(1): e8694.

Small, H. (1973). "Co-citation in the scientific literature: A new measure of the relationship between two documents." Journal of the American Society for Information Science 24(4): 265-269.

Small, H. (1998). "Citations and consilience in science." Scientometrics 43(1): 143-148.

Verspagen, B. (2007). "Mapping technological trajectories as patent citation networks: a study on the history of fuel cell research." Advances in Complex Systems 10(1).

A new method for Batch Learning Self-Organizing Maps based on DC Programming and DCA

H.A. LE THI, M.C. NGUYEN, B.CONAN-GUEZ, and T.PHAM DINH

Laboratory of Theoretical and Applied Computer Science
UFR MIM, University of Lorraine, Ile du Saulcy, 57045 Metz, France
(e-mail: lethi@univ-metz.fr, manhcuong.nguyen@univ-metz.fr, brieuc.conan-guez@univ-metz.fr)
Laboratory of Mathematics. National Institute for Applied Sciences - Rouen
76801 Saint-Etienne-du-Rouvray Cedex, France
(e-mail: pham@insa-rouen.fr)

Abstract. The Self-Organizing Map (SOM), an artificial neural network model, is based on a nonlinear projection mapping from a high-dimensional data space into a low-dimensional representation space. In this paper, an efficient technique for the Batch SOM - one of two main approaches for the SOM learning process, is discussed. We consider the model of Batch SOM which corresponds to a non-smooth, nonconvex optimization problem and investigated the DC (Difference of Convex functions) Programming and a DCA (DC Algorithm) to effectively solve this problem. A training based DCA version for the Batch SOM is also proposed. The numerical results on real-world data sets show the efficiency of the DCA on both quality of solution and representation.
Keywords: Self-Organizing Maps, Batch SOM, DC Programming, DCA.

1 Introduction

The Self-Organizing Maps are often known as low-dimensional grids which represent the distribution of high-dimensional input data sets ([7], [8]). SOM is not only useful in data visualization but also can be applied on many problems in data analysis and clustering ([2]). A SOM is a grid of neurons. Each neuron is represented by a d-dimensional weight vector where d is equal to the dimension of the observations of the input set.

For the SOM learning process, two main approaches are often used: sequential or batch. In both approaches, SOM is trained iteratively. The sequential approach uses a single vector of data set to adjust the weight vectors of the best matching neuron, a winner, and its neighboring neurons while in the batch approach, the whole data set is presented to the map before any adjustments are made.

The neighboring relationship between the neurons is defined according to the Euclidean distance and a neighborhood function which depends on a temperature parameter t. Actually, t is decreased according to a cooling schedule and a temperature variation interval $[T_{min}, T_{max}]$.

In this paper, we consider the Batch learning model and investigate an efficient nonconvex programming approach to solve this problem. Our method is based on Difference of Convex functions (DC) programming and DC Algorithms (DCA) that were introduced by Pham Dinh Tao in their preliminary form in 1985 and have been extensively developed since 1994 by Le Thi Hoai An and Pham Dinh Tao and become now classic and more and more popular (see, e.g. [1], [4], [5], [6], and references therein). Our walk is motivated by the fact that DCA has been successfully applied to many (smooth or nonsmooth) large-scale nonconvex programs in various domains of applied sciences, in particular in Machine Learning for which they provide quite often a global solution and proved to be more robust and efficient than the standard methods. Basing on DCA for the Batch SOM model, we propose a training DCA version for the SOM with an efficient cooling schedule. The training based DCA significantly reduces the computation time and ensure the quality of solution.

The remainder of the paper is organized as follows: Section 2 is devoted to the mathematical formulation of the Batch SOM model. DC programming and DCA for solving the resulting this model (DCASOM) are developed in Section 3. A training version of DCASOM is discussed in Section 4. Finally, numerical experiments are reported in Section 5.

2 The Batch Self-Organizing Maps model

Let $A \in \mathbb{R}^{n*d}$ be a matrix whose ith row is equal to the vector $a^i \in \mathbb{R}^d$, for $i = 1, ..., n$ and $X := \{x^1, ..., x^m\}$ of m weight vectors in $\mathbb{R}^d$. A is the data set of SOM and $x^j (j \in [1..m])$ are the weight vectors associate with map nodes. Let $\phi^t(l, k)$ be a neighborhood function that specifies the influence of node l to node k. This function is computed in term of a temperature parameter t. A measure of compatibility between the observation a^i and the weight vector x^l is defined as follows:

$$\delta^t(a^i, l) = \sum_k \phi^t(l, k)\|x^k - a^i\|^2. \tag{1}$$

Let d^i be a distance from observation a^i to the best matching neuron, so $d^i = \min_l \delta^t(a^i, l)$. When t is kept constant, the Batch SOM problem consists of minimizing $F^t(X)$ which is the sum of all the distances d^i:

$$min \left\{ F^t(X) := \frac{1}{2} \sum_i \min_l \delta^t(a^i, l) = \frac{1}{2} \sum_i \min_l \sum_k \phi^t(l, k)\|x^k - a^i\|^2 \right\}. \tag{2}$$

Actually, SOM is trained iteratively with a monotonous decrease of t. A cooling schedule is used to indicate how t decreases in time. At the k^{th} iteration, a new temperature value of t is often computed according to the

cooling schedule

$$t = T_{max}\left(\frac{T_{min}}{T_{max}}\right)^{\frac{k}{N_{iter}-1}}, \quad (3)$$

with N_{iter} is the number of iterations and $[T_{min}, T_{max}]$ is the temperature variation interval.

3 Solving the Batch SOM model by DC Programming and DCA

- **Outline of DC Programming and DCA**

DC programming and DCA constitute the backbone of smooth/nonsmooth nonconvex programming and global optimization. They address the problem of minimizing a function f which is the difference of two convex functions g and h on the whole space $\mathbb{R}^d$ or on a convex set $C \subset \mathbb{R}^d$.

The idea of DCA is simple: each iteration l of DCA approximates the concave part $-h$ by its affine majorization (that corresponds to taking $y^l \in \partial h(x^l)$) and minimizes the resulting convex function (that is equivalent to determining a point $x^{l+1} \in \partial g^*(y^l)$ with g^* is the conjugate function of the convex function g). The DCA scheme ([1]) includes three main steps as follows:

Initialization: Let $x^0 \in \mathbb{R}^d$ be a best guess, $l \leftarrow 0$.

Repeat

- Calculate $y^l \in \partial h(x^l)$.
- Calculate $x^{l+1} \in arg\min\{g(x) - h(x^l) - \langle x - x^l, y^l\rangle\} : x \in \mathbb{R}^d$.
- $l \leftarrow l + 1$.

Until convergence of $\{x^l\}$.

A description in detail about the DC Programming and DCA and many applications can be found in [1], [4], [5], [6] and http://lita.sciences.univ-metz.fr/∼lethi/.

- **Formulation of (2) as a DC program**

The distance from the observation a^i to the neuron l (equivalent to the weight vector x^l) is

$$\delta^t(a^i, l) = \sum_r \delta^t(a^i, r) - \sum_{r \neq l} \delta^t(a^i, r). \quad (4)$$

Since $\min_l \delta^t(a^i, l) = \sum_r \delta^t(a^i, r) - \max_l \sum_{r \neq l} \delta^t(a^i, r)$ we have

$$F^t(X) = \frac{1}{2}\sum_i \sum_r \delta^t(a^i, r) - \frac{1}{2}\sum_i \max_l \sum_{r \neq l} \delta^t(a^i, r). \quad (5)$$

Hence the following DC decomposition of $F^t(X)$ seems to be natural:

$$F^t(X) = G^t(X) - H^t(X) \quad (6)$$

where

$$G^t(X) = \frac{1}{2}\sum_i \sum_r \delta^t(a^i, r) \tag{7}$$

and

$$H^t(X) = \frac{1}{2}\sum_i \max_l \sum_{r\neq l} \delta^t(a^i, r). \tag{8}$$

According to the convexity of the functions $\sum_r \delta^t(a^i, r)$ and $\max_l \sum_{r\neq l} \delta^t(a^i, r)$, $G^t(X)$ and $H^t(X)$ are clearly convex functions. It is interesting to note that the function $G^t(X)$ is a strictly convex quadratic form.

Let β be a vector in $\mathbb{R}^m$ and $D(\beta)$ is the corresponding diagonal matrix. The Euclidean structure of the matrix vector space $\mathbb{R}^{m*d}$ is defined with the help of the usual scalar product:

$$\langle X, Y\rangle = Tr(X^T Y) = \sum_{i=1}^{m} \langle x^i, y^i\rangle, \tag{9}$$

and its Euclidean norm $\|X\|^2 = \sum_{i=1}^m \langle x^i, x^i\rangle = \sum_{i=1}^m \|x^i\|^2$. So,

$$\sum_{i=1}^{m} \beta_i \langle x^i, y^i\rangle = \sum_{i=1}^{m} \langle x^i, \beta_i y^i\rangle = \langle X, D(\beta)Y\rangle,$$

and

$$\sum_{i=1}^{m} \beta_i \langle x^i, x^i\rangle = \langle X, D(\beta)X\rangle.$$

Let ϕ_+^t be a vector in $\mathbb{R}^m$ such that $\phi_{+k}^t = \sum_r \phi^t(r, k)$. We have, after simple calculations:

$$G^t(X) = \frac{n}{2}\langle X, D(\phi_+^t)X\rangle - n\langle X, D(\phi_+^t)\overline{A}\rangle + \frac{1}{2}\left(\sum_k \phi_{+k}^t\right)\|A\|^2 \tag{10}$$

where $\overline{A} \in \mathbb{R}^{m*d}$ is given by $\overline{A}_i := \overline{a} = \sum_{i=1}^n a^i$.

Determining the DCA scheme applied to (6) amounts to computing the two sequences $\{X^k\}$ and $\{Y^k\}$ in $\mathbb{R}^{m*d}$ such that

$$Y^k \in \partial H^t(X^k), X^{k+1} \in \partial G^*(Y^k).$$

We shall present below the computation of $\partial H^t(X)$ and $\partial G^*(Y)$.

- **Calculation of $\partial H^t(X)$**

We can write, for $i = 1, ..., n$:

$H_i^t(X) = \max_l H_{i,l}^t(X)$ where $H_{i,l}^t(X) = \sum_{r\neq l} \delta^t(a^i, r)$.

We first express the convex function $H_{i,l}^t(X)$ by

$$H_{i,l}^t(X) := \sum_k \left(\sum_{r \neq l} \phi^t(r,k) \right) \|x^k - a^i\|^2 = \sum_k \overline{\phi_l^t} \|x^k - a^i\|^2 \tag{11}$$

where $\overline{\phi_l^t} \in \mathbb{R}^m$ is given by $\overline{\phi_{lk}^t} = \sum_{r \neq l} \phi^t(r,k) = \phi_{+k}^t - \phi^t(l,k)$. Hence,

$$\nabla H_{i,l}^t(X) = 2D(\overline{\phi_l^t})(X - A^{[i]}) \tag{12}$$

where $A^{[i]} \in \mathbb{R}^{m*d}$ is the matrix whose rows are all equal to a^i.

Let $K_i(X) = \{l = 1, ..., m | H_{i,l}^t(X) = H_i^t(X)\}$ and $Y^{[i]} \in \partial H_i^t(X)$ be a combination convex of $\{\nabla H_{i,l}^t | l \in K_i(X)\}$, so

$$Y^{[i]} = \sum_{l \in K_i(X)} \lambda_l^i \nabla H_{i,l}^t(X) \tag{13}$$

with $\lambda_l^i \geq 0$ and $\sum_{l \in K_i(X)} \lambda_l^i = 1$.

We have $Y \in \partial H^t(X) \Leftrightarrow Y = \frac{1}{2} \sum_i Y^{[i]}$ and in particular we can chose

$$Y = \sum_i D(\overline{\phi_{l_i}^t})(X - A^{[i]}). \tag{14}$$

- **Calculation of $\partial G^*(Y)$**

Since $G^t(X)$ is a strictly convex quadratic funtion and its conjugate $(G^t)^*$ is differentiable, we have $X = \nabla (G^t)^*(Y) \Leftrightarrow Y = \nabla G^t(X) = nD(\phi_+^t)(X - \overline{A})$. and therefore

$$X = D\left(\frac{(\phi_+^t)^{-1}}{n} \right) Y + \overline{A}. \tag{15}$$

We are now in a position to describe the DCA for solving problem (2) via the DC decomposition (6).

- **DCA for Batch SOM model**

Initializations: Let $\epsilon > 0$ be given, X^0 be an initial point in $\mathbb{R}^{m*d}$ and set $k = 0$.

Repeat:

1. Calculate Y^k from X^k via (14).
2. Calculate X^{k+1} from Y^k via (15).
3. $k = k + 1$.

Until $\| X^{k+1} - X^k \| \leq \epsilon \| X^k \|$.

Remark 3.1 This DCA requires only elementary operations on matrices (the sum and the scalar multiplication of matrices). It's is very simple and can so handle large-scale Batch SOM problem.

4 A training version for DCASOM

In this section, we propose a training version for DCASOM based on a efficient cooling schedule of t. In this version, t is decreased from T_{max} to T_{min} with N_{iter} iterations. In the kth iteration, t is computed by

$$t = T_{max}\left(\frac{T_{min}}{T_{max}}\right)^{\frac{k}{N_{iter}-1}} - T^r, \tag{16}$$

with T^r is a sufficiently small positive constant. A training version for DCASOM is proposed as follows:

Initializations: Let $\epsilon > 0$ be given, X^0 be an initial point in $\mathbb{R}^{m*d}$. Set $k = 0; T = 0$ and select T_{max}, T_{min}, T^r and N_{iter}.

Repeat: (iteration T)

a. Calculate t via (16).

b. **Repeat:**

1. Calculate Y^k from X^k via (14).
2. Calculate X^{k+1} from Y^k via (15).
3. $k = k + 1$.

Until $\parallel X^{k+1} - X^k \parallel \leq \epsilon \parallel X^k \parallel$. (STOP condition)

c. $T = T + 1$.

Until $T = N_{iter}$.

In this algorithm, N_{iter} may be set to a very small value. We propose an efficient value $N_{iter} = 3$. The STOP condition at each iteration T may be difference. At the first and second iterations, this condition is relaxed. For more details, we can sequentially set $\epsilon = 10^{-2}, 10^{-6}$ and 10^{-8} for each iteration.

The set of weight vectors X^0 is initiated depending on the center vector $\overline{a}$ of A. X^0 is created around $\overline{a}$ with a sufficiently small variance.

5 Numerical Experiments

We evaluate the quality of DCASOM training version on the objective value and the Kaski measure which was introduced by Kaski and Lagus in [3]. The results of DCASOM are compared with the standard Batch SOM algorithm (BSOM) that was introduced by Kohonen and Heskes in [7, 8].

We tested on the data sets which are obtained from the internet. The tiny data sets named Spaeth Cluster Analysis can be found at the address http://people.sc.fsu.edu/∼jburkardt/datasets/spaeth. The sets named A-Sets can be found at http://cs.joensuu.fi/sipu/datasets. Glass and Inosphere data sets are very popular in Machine Learning. Letter Recognition (Letter) and Color Moments of Corel Image Features (ColorM) are two data sets of the UCI's web site.

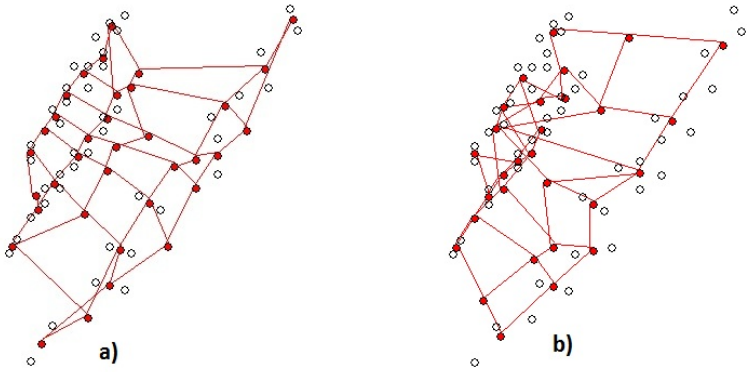


Fig. 1. Topology map (5 x 7) of Spaeth_05; $T_{max} = 10$; $T_{min} = 0.2$; a) DCASOM: $N_{iter} = 3$, Obj. = 135.454, Kaski = 7.672; b) BSOM: $N_{iter} = 1000$, Obj. = 310.829, Kaski = 10.763.

Datasets			Objective values		Kaski measure		Running time (s)	
Name	#Ele.	#Att.	(1)	(2)	(1)	(2)	(1)	(2)
Spaeth_05	55	2	**598.34**	665.77	**20.48**	20.58	**0.36**	1.21
Spaeth_06	50	2	**375.54**	422.53	**16.60**	19.12	**0.30**	1.11
Spaeth_07	52	2	**596.86**	663.73	**22.84**	24.23	**0.40**	1.51
Spaeth_08	80	2	**1,060.96**	1,085.91	**26.02**	28.96	**0.56**	1.70
Spaeth_10	49	56	**134.26**	158.77	**4.92**	6.18	**2.11**	4.49
Spaeth_13	96	5	**1,573.59**	1,578.95	**29.86**	31.31	**0.66**	2.32
Spaeth_09	112	18	**1,025.58**	1,073.89	**12.60**	14.01	**2.15**	4.86
Glass	214	10	**192.11**	217.52	**2.56**	3.36	**1.97**	5.92
Inosphere	351	9	**1,445.02**	1486.10	**6.34**	6.41	**9.18**	19.14
A1_Set	3,000	2	**338.61**	353.27	1.38	**1.36**	**9.7**	60.54
A2_Set	5,250	2	1,099.55	**1,089.11**	**1.47**	1.52	**29.87**	102.23
A3_Set	7,500	2	**2,544.79**	2,602.56	**1.80**	2.38	**43.58**	146.82
Letter	20,000	16	**748,054**	768,276	**49.37**	56.37	**237.31**	699.85
ColorM	68,040	9	256,708	**254,557**	8.33	**7.90**	**722.21**	1,855.18

Table 1. The preliminary results of DCASOM (1) and BSOM (2).

Fig. 1 shows the maps which are provided by DCASOM training version and BSOM on the data set Spaeth_05. With $N_{iter} = 3$, DCASOM can give better on both objective value and topology map.

In the table 1, we present the preliminary results of DCASOM training version and BSOM. The maps with the size (3 x 5) and the Gaussian neighborhood function with learning coefficient $\alpha(t) = 1$ are used. DCASOM training version is performed 3 times with the starting value of T^r is 0.2 and to increase 0.1 for each time. Whereas, BSOM is tested with 5000 iterations.

The temperature parameter t decreases from $T_{max} = 10$ to $T_{min} = 0.2$ for both DCASOM and BSOM.

This table shows that DCASOM training version often give the better quality of solutions. Specially, the maps which are provided by DCASOM often have good quality. It is very important for the data visualization target of SOM. Furthermore, this proves that DCASOM is less sensitive to the cooling schedule of t.

6 Conclusion

We have proposed a DC program and DC Algorithm for the Batch Learning SOM model. The nonsmoth, nonconvex optimization model of Batch SOM is reformulated as a difference of two convex functions which allow us to find a appropriate DCA to solve it.

A training version for DCASOM with an efficient cooling schedule to decrease the value of the temperature parameter t is proposed. Further, a simple but effective strategy to create the initial points for DCASOM is also mentioned. The numeric results show that DCASOM seem to do well and give better results than the standard batch SOM method.

The efficiency of DCA for this problem suggests us to extend on several directions of Self-Organizing Maps. Works in these extensions are in progress.

References

1. H.A. Le Thi, M. Tayeb Belghiti, T. Pham Dinh, "A new efficient algorithm based on DC programming and DCA for clustering", *J Glob Optim*, 37:593-608 (2007).
2. Jean-Claude Fort, Patrick Letremy, Marie Cottrell, "Advantages and drawbacks of the Batch Kohonen algorithm", *In ICNC*, 223–230 (2002).
3. S. Kaski, K. Lagus "Comparing Self-Organizing Maps", *Lecture Notes in Computer Science, Vol. 1112*, 809–814 (1996).
4. H.A. Le Thi, T. Pham Dinh, "DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems", *Ann. Oper. Res.Springer-Verlag*, 133, 23-46 (2005).
5. T. Pham Dinh, H.A. Le Thi, "DC optimization algorithms for solving the trust region sub-problem". *SIAM J. Optim.* 8, 476-505 (1998).
6. H.A. Le Thi, T. Pham Dinh, "Minimum Sum-of-Squares Clustering by DC Programming and DCA". *Lecture Notes in Computer Science, Vol. 5755/2009*, 327–340 (2009).
7. T. Kohonen, "Self-Organization Maps", *Springer, Heidelberg* (1997).
8. T. Heskes, "Self-Organization Maps, vector quantization, and mixture modeling", *IEEE transactions on neural networks, Vol. 12, No. 6* (2001).

Using aggregation functions for Markov-switching models

Jana Lenčuchová

Slovak University of Technology in Bratislava
Faculty of Civil Engineering, Bratislava, Slovakia
(e-mail: lencuchova@math.sk)

Abstract. The Markov-switching (MSW) models were used to model an original and a corresponding aggregated time series. Data are aggregated by the common aggregation functions such as the arithmetic mean, minimum and maximum. We compare these models in term of description and prediction properties. The numbers of inputs to the aggregate function were determined following BDS test of independence.
Keywords: Markov-switching model, aggregation function, BDS test of independence.

1 Introduction

In last two decades, the regime-switching models became very popular in the fields such as finance, economy but also hydrology and so on. Since 1989, when Hamilton published the paper [**?**], Markov-switching models have experienced great success in many applications mainly because of their excellent description properties.

So far we have not encountered the idea of using Markov-switching model applied to an aggregated time series and comparing such model with the one for original data.

The paper is organized as follows. Second, third and fourth sections are devoted to a brief theory of the Markov-switching model, the aggregation functions and the BDS test. The fifth section includes application part of the paper. Finally, a conclusion is made.

2 Markov-switching model

The Markov-switching models belong to the class of nonlinear models, especially to the class of regime-switching models which have their *regimes* or *states* determined by unobservable variables. This means that we can not identify the given regime for sure, but only with some probability. At this point, we have to define a stochastic process $\{s_t\}$ for specifying in which regime the process is situated.

Assuming that the random variable s_t can attain only values from the set $\{1, 2, ..., N\}$, the basic Markov-switching model, we use in applications, has the following form:

$$y_t = \begin{cases} \phi_{0,1} + \phi_{1,1} y_{t-1} + ... + \phi_{q,1} y_{t-q} + \epsilon_t, & s_t = 1 \\ \phi_{0,2} + \phi_{1,2} y_{t-1} + ... + \phi_{q,2} y_{t-q} + \epsilon_t, & s_t = 2 \\ \vdots & \\ \phi_{0,N} + \phi_{1,N} y_{t-1} + ... + \phi_{q,N} y_{t-q} + \epsilon_t, & s_t = N, \end{cases} \tag{1}$$

where the model is described in particular regimes by an autoregressive model AR(q) and $t = q + 1, ..., T$; T is the length of time series. We assume that the random variable ϵ_t is identically, independently and normally distributed with zero mean, that is $\epsilon_t \sim N(0, \sigma_\epsilon^2)$. This notation can be shortened to the next one row notation:

$$y_t = \phi_{0,s_t} + \phi_{1,s_t} y_{t-1} + ... + \phi_{q,s_t} y_{t-q} + \epsilon_t, \qquad s_t = 1, ..., N. \tag{2}$$

To see how the process switches among regimes it is necessary to specify the stochastic process $\{s_t\}$. In 1989, Hamilton [**?**] suggested to define this stochastic process as a first-order Markov process, which is also called *the Markov chain*.

A Markov chain is a sequence of random values $s_1, s_2, ..., s_t, s_{t-1}, ...$ with the *first-order Markov property*:

$$Pr(s_t = j | s_{t-1} = i, s_{t-2} = k, ...) = Pr(s_t = j | s_{t-1} = i) = p_{ij}, \tag{3}$$

where $i, j, k = 1, \ldots, N$ and N is the number of regimes. This means the probability that the process will be in regime j in time t, is dependent only on the previous regime i in time $t-1$. These kinds of probabilities $\{p_{ij}\}_{i,j=1,...,N}$ are called *transition probabilities* and they create a matrix called *a transition matrix*:

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{21} & \cdots & p_{N1} \\ p_{12} & p_{22} & \cdots & p_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1N} & p_{2N} & \cdots & p_{NN} \end{pmatrix}. \tag{4}$$

Naturally, the following equality with transition probabilities must be true:

$$p_{i1} + p_{i2} + ... + p_{iN} = 1, \qquad i = 1, ..., N. \tag{5}$$

For our purpose we need the Markov chain to be *ergodic*; that is, it needs to be possible to get from one regime to another (not necessarily in one step). Thus it must satisfy for all $i, j : p_{ij} > 0$, where $i, j = 1, ..., N$.

3 Aggregation functions

In this section we define briefly the basic concept of the aggregation functions. For thorough study of the aggregation functions, we recommend the book by Grabish *et. al.* [**?**].

Definition 1 Let $\mathbb{I}$ be a nonempty real interval, m an integer as the number of variables and $\mathbf{y} = (y_1, ..., y_m) \in \mathbb{I}^m$. An aggregation function in $\mathbb{I}^m$ is a function $\mathrm{A}^{(m)} : \mathbb{I}^m \to \mathbb{I}$ that

(i) is nondecreasing in each variable
(ii) fulfills the boundary conditions

$$\inf_{\mathbf{y}\in\mathbb{I}^m} \mathrm{A}^{(m)}(\mathbf{y}) = \inf \mathbb{I} \quad and \quad \sup_{\mathbf{y}\in\mathbb{I}^m} \mathrm{A}^{(m)}(\mathbf{y}) = \sup \mathbb{I}.$$

The most well-known aggregation functions are:

- the arithmetic mean:
$$\mathrm{AM}(\mathbf{y}) = \frac{1}{m}\sum_{i=1}^{m} y_i$$
- the minimim:
$$\mathrm{Min}(\mathbf{y}) = \min\{y_1, ..., y_m\}$$
- the maximum:
$$\mathrm{Max}(\mathbf{y}) = \max\{y_1, ..., y_m\}$$

4 BDS test

The number of inputs in the aggregation function were determined according to BDS test of independence in residuals $\{\hat{e}_t\}$. Also, the BDS test is a powerful tool for detecting serial dependence in time series. It tests the null hypothesis, that a time series sample comes from i.i.d. process against an unspecified alternative. This test based on the correlation integral was presented in the paper by Brock, Dechert, Scheinkman, and Le Baron [**?**].

Let us define the correlation integral for some $m \in N$ and $\varepsilon > 0$:

$$C_{m,\varepsilon} = \frac{\sum\sum_{q+1\leqq\tau\leqq T_m} \mathbf{1}\left(\|\widehat{\mathbf{e}}_{\mathbf{t},\mathbf{m}} - \widehat{\mathbf{e}}_{\tau,\mathbf{m}}\| < \varepsilon\right)}{2[(T_m - 1)T_m]} \tag{6}$$

where $T_m = T - m + 1$, $\hat{\mathbf{e}}_{t,m} = (\hat{e}_t, \ldots, \hat{e}_{t+m-1})$, $\mathbf{1}(\mathrm{A})$ is the indicator function of the event A and $\|\cdot\|$ denotes the maximum norm in $\Re^d$ ($\|z\| = \max_{1<i<d} |z_i|$ for $z = (z_1, \ldots, z_d)'$). The BDS statistic is

$$\Lambda_{\mathrm{BDS}} = [(T - q)/V_{m,\varepsilon}]^{1/2}(C_{m,\varepsilon} - C_{1,\varepsilon}^m)$$

where

$$V_{m,\varepsilon} = 4K_\varepsilon^m + 4(m-1)^2 C_\varepsilon^{2m} - 4m^2 K_\varepsilon C_\varepsilon^{2(m-1)} + 8\sum_{j=1}^{m-1} K_\varepsilon^{m-1} C_\varepsilon^{2j},$$

$$K_\varepsilon = \frac{1}{(T-q)^3} \sum_{K=q+1}^{T} \sum_{\tau=q+1}^{T} \sum_{t=q+1}^{T} \mathbf{1}\left(|\hat{e}_\kappa - \hat{e}_\tau| < \varepsilon\right) \mathbf{1}\left(|\hat{e}_\tau - \hat{e}_t| < \varepsilon\right),$$

$$C_\varepsilon = \frac{1}{(T-q)^2} \sum_{\tau=q+1}^{T} \sum_{t=q+1}^{T} \mathbf{1}\left(|\hat{e}_\tau - \hat{e}_t| < \varepsilon\right),$$

T is the length of time series and q is a model order. "Embedding dimension" of the correlation integral is denoted as m (the number of the time-lagged elements) and "metric bound" is set $\varepsilon = \tilde{\sigma}$, where $\tilde{\sigma}^2 = 1/(T-q-1)\sum_{t=q+1}^{T} \hat{e}_t^2$. Brock et al. [**?**] recommendation is that ε should be in the following region $\frac{1}{2}\tilde{\sigma} \leq \varepsilon \leq \frac{3}{2}\tilde{\sigma}$.

The BDS test statistic Λ_{BDS} has N(0,1) asymptotic distribution when $\{e_t\}$ are i.i.d.

5 Application

The main goal of this paper is to compare Markov-switching model of an original time series and an aggregated one from two points of view - description and prediction properties.

First of all, we need to know how many inputs we should aggregate. This is calculated by the BDS test [**?**]. Actually, the pairs $\{(y_t, y_{t+m}), t \in \{1, ..., T-m\}\}$ are tested. But one correction of this set is made. We choose only every sth pair from this set of pairs due to dependence between them. Now we assume that this kind of dependence dissappeared and dependence between dimensions only remains. Thus s should be large enough but also preserving a sufficient number of pairs for the testing. The new thinned set of time points is denoted as $\mathcal{T} = \{1, s+1, 2s+1, ..., rs+1 <= T-m\}$, where $|\mathcal{T}| = r+1$ is the new thinned sample size. Then we test pairs $\{(y_t, y_{t+m}), t \in \mathcal{T}\}$.

As mentioned above, the number of inputs m is determined by the BDS test. We run the testing while the null hypothes is rejected. The first non-rejected null hypothesis is in order $(m+1)th$.

The data were aggregated as follows: x_t =AF$(y_t, ..., y_{t+m-1})$ for $t = 1, ..., T-m$, where AF(.) is some m-ary aggregation function.

First results can be seen in Table **??**. We compare here the residual dispersions $\hat{\sigma}_\epsilon^2$ and mean square errors (MSE= $\sum_{i=1}^{p}(y_i - \hat{y}_i)$, where p is the number of predictions) of Markov-switching model for original time series

and for time series aggregated by the arithmetic mean, the minimum and the maximum. As regards description properties of the Markov-switching model it is shown in the Table **??** that Markov-switching model for time series Dow Jones Euro Stoxx 50 Price Index aggregated by mean gave the least residual dispersions for all model orders $q = 1, ..., 5$ (highlighted by red font). That is because the aggregation by mean causes some kind of time series smoothing. Some smothing is also obvious when one compares graphs in Figure **??**. The best Markov-switching model for original nonaggregated time series ($q = 4$) and the data are drawn in the left graph of Figure **??** and the best Markov-switching model for time series aggregated by mean (the number of inputs $m = 13$ and model order $q = 5$) with data aggregated by mean are in the right graph.

Considering prediction properties, the lowest MSEs are highlighted by blue color. They all belong to aggregation by maximum function. The comparison of Markov-switching predictions of original data and aggregated by maximum can be seen in Figure **??**.

y_t		
q	$\hat{\sigma}_\epsilon^2$	MSE
1	0.1185	0.0194
2	0.1207	0.0409
3	0.1015	0.0094
4	0.0979	0.0121
5	0.0984	0.1164

AM(**y**)		
q	$\hat{\sigma}_\epsilon^2$	MSE
1	**0.0264**	0.0344
2	**0.0134**	0.0336
3	**0.0120**	0.0554
4	**0.0119**	0.0541
5	**0.0118**	0.0460

Min(**y**)		
q	$\hat{\sigma}_\epsilon^2$	MSE
1	0.0736	0.0040
2	0.0640	0.0064
3	0.0624	0.0126
4	0.0599	0.0008
5	0.0567	0.0007

Max(**y**)		
q	$\hat{\sigma}_\epsilon^2$	MSE
1	0.0476	**0.0002**
2	0.0456	**0.0001**
3	0.0425	**0.0002**
4	0.0427	**0.0002**
5	0.0653	**0.0006**

Table 1. Dow Jones Euro Stoxx 50 Price Index January 2000 - February 2012

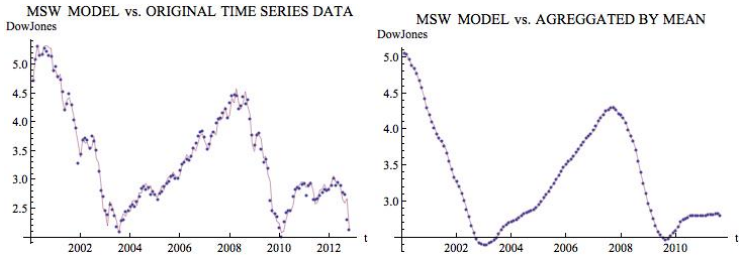


Fig. 1. Markov-switching models for Dow Jones Euro Stoxx 50 Price Index

Relative improvement results of Markov-switching model for aggregated Dow Jones time series can be seen in Table **??**. Actually, improvements of $\hat{\sigma}_\epsilon^2$ and MSE of Markov-switching models for aggregated time series against

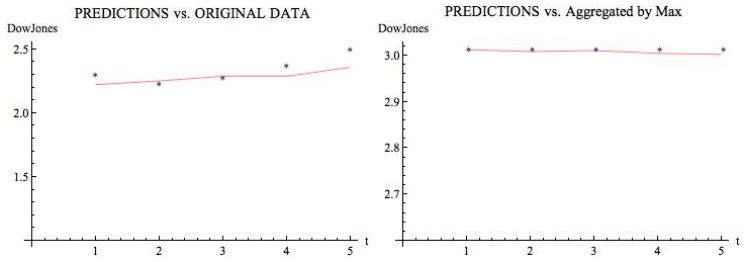


Fig. 2. Predictions of Dow Jones Euro Stoxx 50 Price Index

the $\hat{\sigma}_\epsilon^2$ and MSE of Markov-switching models for original time series are calculated in this table. According to this indicator the best improvements of the $\hat{\sigma}_\epsilon^2$ were noted by Markov-switching model for time series aggregated by arithmetic mean and the best improvements of MSE belong to the Markov-switching model for time series aggregated by maximum function. Thus, it is in aggrement with results before. Some negative improvements are also the other side of time series aggregation. Reader could have noticed that aggregation by arithmetic mean was excellent choice for the improvement of $\hat{\sigma}_\epsilon^2$ but the worst choice for model predictions, because of the MSE comedown. The similar idea can be applied for the aggregation by maximum.

For this analysis, we chose 15 economic and financial time series. Table **??** shows the best relative improvements for all considering model orders $q = 1, ..., 5$ and the average value of improvements distintively for aggregation by the arithmetic mean, the minimum and the maximum. The residual dispersion improvement for time series aggregated by the arithmetic mean is 82% in average and the mean square error is 26,8%, but this value is influenced significantly by the negative improvements of MSE for oil prices and service confidence indicator EU. The other time series noticed really nice improvements in MSE. The same story about average of MSE improvements is in case of aggregation by the minimum and the maximum. But there can be still found massive improvements. Residual dispersions are lower for the minimum and the maximum case in average than it was for the arithmetic mean. It could be interesting to consider some other types of aggregation functions too.

6 Conclusion

In this paper, the Markov-switching models were applied to the aggregated time series. The well-known aggregation functions such as the arithmetic mean, the minimum and the maximum are employed. We followed the results of the BDS test of independence to identify the number of inputs in aggregation. Thereafter the description and prediction properties via the

Relative improvement %									
AM			MIN			MAX		
q	$\hat{\sigma}_\epsilon^2$	MSE	q	$\hat{\sigma}_\epsilon^2$	MSE	q	$\hat{\sigma}_\epsilon^2$	MSE
1	77,70%	-77,30%	1	37,92%	79,45%	1	59,85%	99,03%
2	88,87%	17,98%	2	47,05%	84,39%	2	62,25%	99,77%
3	88,18%	-490,09%	3	38,51%	-33,91%	3	58,11%	96,96%
4	87,86%	-346,15%	4	38,84%	93,10%	4	56,41%	97,62%
5	88,05%	-294,76%	5	42,38%	93,63%	5	33,61%	94,58%

Table 2. Relative improvements of the Markov-switching model for original Dow Jones Euro Stoxx 50 Price Index January 2000 - February 2012

Relative improvement %	AM		MIN		MAX	
	$\hat{\sigma}_\epsilon^2$	MSE	$\hat{\sigma}_\epsilon^2$	MSE	$\hat{\sigma}_\epsilon^2$	MSE
Inflation Slovakia	69,3%	79,5%	45,8%	93,2%	44,7%	87,5%
Dow Jones Euro Stoxx 50 Index	88,1%	18,0%	42,4%	93,6%	58,1%	99,8%
Euribor 1-month	62,5%	95,2%	34,9%	51,7%	-0,2%	99,0%
Current incomes Slovakia	84,0%	95,0%	38,6%	98,0%	64,8%	88,5%
Debit transactions EU	87,1%	90,0%	98,2%	100,0%	-23,5%	-227,4%
Credit transactions EU	88,2%	95,0%	92,2%	94,2%	69,6%	99,9%
Base money EU	95,1%	79,2%	84,9%	12,7%	45,9%	97,1%
Brazilian real	94,3%	99,9%	71,3%	99,8%	64,3%	99,2%
Commodity index- nonenergy EU	93,8%	90,4%	76,1%	100,0%	73,0%	85,7%
Commodity index- nonfood EU	86,8%	80,3%	44,0%	96,0%	57,2%	4,0%
Negotiate wage rates EU	78,4%	36,9%	74,6%	98,6%	57,6%	79,4%
Nikkei 225	86,8%	97,2%	46,5%	99,9%	69,1%	-332,1%
Nomimal wages Slovakia	81,0%	87,6%	53,4%	65,4%	61,3%	95,0%
Oil prices	71,1%	-587,3%	34,4%	-1910,7%	24,0%	92,3%
Service confidence indicator EU	64,4%	-55,5%	38,7%	-1146,0%	28,4%	-19,8%
AVERAGE	**82,0%**	**26,8%**	**58,4%**	**-130,2%**	**46,3%**	**29,9%**

Table 3. Selected relative improvements of the best Markov-switching models according to the residual dispersions and the mean square errors against the corresponding Markov-switching models for original time series

residual dispersions and the mean square errors were compared. So far, there has not been done such study according to our knowledge.

Concerning the residual dispersion - in all cases except two a positive improvement by aggregations occured. The Markov-switching models for time series aggregated by the arithmetic mean noticed the best average improvement. Improvements of MSEs are negative more often than residual dispersion and sometimes these values are dramatically high. This is also the reason of lower averages in all aggregation functions. The winner in this sample is the maximum function. After all, we are able to choose one model from aggregated time series which is the most suitable nearly in all cases

when considering our needs. It means if we are interested more in model description or in prediction. Our intention in the future research is to use other types of aggregation functions such as WAM (Weigted Arithmetic Mean), OWA (Ordered Weighted Average) and so on.

ACKNOWLEDGEMENT: This work was partially supported by the grants APVV No. LPP-0111-09 and VEGA 1/0143/11.

References

1. Brock, W. A., Dechert, W. D., Scheinkman, J. A., and Le Baron, B., "A test for independence based on the correlation dimension", *Econometric Reviews*, 15, 197–235 (1996).
2. Brock, W. A., Hsieh, D. A., and LeBaron, B., *Nonlinear Dynamics, Chaos, and Instability: Statistical Theory and Economic Evidence*, Cambridge, Mass. : MIT Press (1991).
3. Grabisch, M., Marichal, J.-L., Mesiar, R., and Pap, E., *Aggregation Functions*, Cambridge University Press, Encyclopedia of Mathematics and its Applications, No 127 (2009).
4. Hamilton, J.D., "A new approach to the economic analysis of nonstacionary time series and the business cycle", *Econometrica*, 57, 357–384 (1989).
5. Hamilton, J. D., *Time series analysis*, Princeton University Press, Princeton (1994).

Solution of Problem of Stable Extraction of Features with Using of Measure of Informativeness

Alexander E. Lepskiy[1]

National Research University - Higher School of Economics
Moscow, Russia
(e-mail: alex.lepskiy@gmail.com)

Abstract. In the paper we present a new notion of stochastic monotone measure and its application to image processing. By definition, a stochastic monotone measure is a random value with values in the set of monotone measures and it can describe a choice of random features in image processing. In this case, a monotone measure describes uncertainty in the problem of choosing the set of features with the highest value of informativeness and its stochastic behaviour is explained by a noise that can corrupt images.
Keywords: Image processing, Features extraction, Stochastic monotone measure.

1 Introduction

Let $\Omega = \{\omega_i\}_{i=1}^n$ be a set of features that correspond to an image X. To achieve the highest productivity and stable working of a pattern recognition system, it is necessary to choose a sufficiently small subset of features in Ω with the highest information value. There are some very well-known methods that can give us features with the highest information value based on the method of principal components, discriminant analysis and so on [2], [3]. But these methods based on linear algebra fail to take into account structural characteristics that are inherent to geometrical information. In this situation, information measures can be used. By definition, an information measure μ is a set function defined on the power set 2^Ω of Ω that for each $A \in 2^\Omega$ shows an information value of features in A.

As a rule the features used to describe the pattern are defined with varying degrees of imprecision. The nature of the imprecision may be different. The sets of features are an elements of some a probability space in the classical setting. For example, if the pattern is a discrete plane curve that extracted on the image and features are some characteristics of curve points (eg, feature is a estimation of curvature in given point of discrete curve [4]) then a random character of features (eg, curvature) will be due to noise of image. In this work we will consider monotonous measures of informativeness that are defined on the all subsets of the set of random features. Then monotonous measure will be a random variable $\mathrm{M}(A)$ for each fixed set of random features A. In this case the expectation $\mathbf{E}\,[\mathrm{M}(A)]$ be characterize the level of informativeness

of representation A and the variance $\sigma^2[\mathrm{M}(A)]$ be characterize the level of stability of representation to noise pattern. Then there is the problem of finding the most stable and informative representation A of the pattern X. In this article mentioned task will be discussed and resolved in the case when any random feature depends from some other features. This task is investigated for the most popular measures of informativeness for contour image – measure of informativeness by length.

2 Monotone Geometrical Measure of Informativeness

Measures of informativeness can be effectively used in image processing as shown in [1]. In image processing, the most important features are that do not depend on illumination of a scene and orthogonal transformations. Contours of images and their characteristics, for example, curvature of smooth curves can be such features. However, in reality, we have digitized curves that are given by some ordered sets of points. These curves can be corrupted by noise. This means that we can use only some statistical estimates of curvature [4]. A problem of choosing an optimal polygonal representation of a contour consist in finding such a representation that preserves geometrical characteristics of contour and will be statistically efficient. This choice can be produced by using measure of informativeness that are axiomatically defined as follows.

Let X be an initial closed contour given by an ordered finite set points, i.e. $X = \{x_1, ..., x_n\}$, where $x_i \in R^2$, $i = 1, ..., n$. We identify with any nonempty subset $B = \{x_{i_1}, ..., x_{i_m}\}$ a contour generated by connecting points with straight lines starting from points x_{i_1}, x_{i_2} and ending by points x_{i_m}, x_{i_1}.

A geometrical measure of informativeness $\mu : 2^X \to [0, 1]$ is a set function that has to obey the following properties:

1. $\mu(\emptyset) = 0$, $\mu(X) = 1$;
2. $A, B \in 2^X$ and $A \subseteq B$ implies $\mu(A) \leq \mu(B)$;
3. let $B = \left\{x_{i_1}, ..., x_{i_{k-1}}, x_{i_k}, x_{i_{k+1}}, ..., x_{i_m}\right\} \in X$ and neighbouring points $x_{i_{k-1}}, x_{i_k}, x_{i_{k+1}}$ belong to a straight line in the plane, then $\mu(B) = \mu\left(B \backslash \{x_{i_n}\}\right)$;
4. μ is invariant w.r.t. affine transformations in the plane such as parallel transferring, rotation and scaling.

Emphasize that axioms 1, 2 have been introduced by Sugeno for fuzzy measures (see [7]). Consider several ways for defining geometrical measures of informativeness [1].

a) Suppose that the length of an original contour is not equal to zero and a function $L(A)$ gives us the length of subcontour $A \in 2^X$. Then an measure of informativeness defined by contour length is $\mu_L(A) = \frac{L(A)}{L(X)}$.

b) Suppose that the domain limited by an original contour is convex, and a function $S(B)$ determines the area bounded by an subcontour $A \in 2^X$. Then an measure of informativeness defined by contour area is $\mu_S(A) = \frac{S(A)}{S(X)}$.

c) Let $w(x, A)$ be an estimate of information value of the part of a contour in a neighbourhood of point $x \in A$ in a subcontour $A \in 2^X$. Then an average measure of informativeness is defined by $\mu(A) = \frac{\sum_{x \in A} w(x,A)}{\sum_{x \in X} w(x,X)}$, where $w(x, A)$ has to be defined for any non-empty contour $A \in 2^X$ and $\mu(\emptyset) = 0$ by definition. It is easy to see that the introduced geometrical measure of informativeness μ_L and μ_S can be considered as average measure of informativeness. For example, for μ_L function $w(x, A)$ can be defined by $w(x, A) = |x - y|$, where y is the next neighbouring point in contour A; in case of μ_S function $w(x, A)$ can be defined by $w(x, A) = S(O, x, y)$, where O is the centroid of area, bounded by contour A, and $S(O, x, y)$ is the area of triangle with vertices in points O, x, y.

As it was mentioned above, points consisting of the original contour accumulate the influence of several random factors, and they can be therefore considered as random values. This means that the observed measure of informativeness can be viewed as a random monotone measure. This opens new directions of research.

3 Stochastic Average Monotone Measure of Informativeness

In real situations, values $w(x, A)$ can be considered as random values, because an original contour is corrupted by a noise. To stress this, we denote these values by capital letters as $W(x, A)$. In this case we have the measure of informativeness $\mathrm{M}(A) = \frac{\sum_{x \in A} W(x,A)}{\sum_{x \in X} W(x,A)}$. Then there is the problem of finding the most stable and informative representation $A \in 2^X$ of the pattern X for which the expectation $\mathbf{E}\left[\mathrm{M}(A)\right]$ will be most and the the variance $\sigma^2\left[\mathrm{M}(A)\right]$ will be least. If $W(x, A) = W(x)$ and random values $W(x)$, $x \in X$, are independent random variables then the measure of informativeness $\mathrm{M}(A)$ is additive and its values are random. For example, this probabilistic model can be used if values $W(x)$ can be conceived as estimates of absolute value of curvature and this estimation is produced by disjoint neighbourhoods of points $x \in X$ [4]. We emphasize that stochastic additive measures have been already investigated in the literature (see e.g. [6]). In this additive case the problem of finding the most stable and informative representation is investigated in [5].

Now we will be investigated the important case when the value $W(x, A)$ depends on two neighbouring points. For example, the geometrical measures of informativeness μ_L and μ_S are satisfied this condition.

Let $X = \{x_1, ..., x_n\}$ be an original contour and let vertices be ordered by their indices. So if we consider any subcontour $A \in 2^X$, then the order defined on A is assumed to be generated by the order on X and given by indices in the representation $A = \{x_{i_1}, ..., x_{i_m}\}$, where $i_1 < ... < i_m$. So for any

$A \in 2^X$ we can identify its elements by their indices and write $x_k(A) = x_{i_k}$ if $k \in \{1, ..., m\}$ and $A = \{x_{i_1}, ..., x_{i_m}\}$, where $i_1 < ... < i_m$. We can also consider any integer index k assuming that $x_k(A) = x_l(A)$ if $l \equiv k(\mathrm{mod} m)$. To work with such indices, we use a mapping π defined by $x_k(A) = x_{\pi_A(k)}$. We suppose that $W(x_k(A), A) = W(x_k(A), x_{k+1}(A))$, $k = 1, ..., |A|$, i.e. the value $W(x_k(A), A)$ depends on two neighbouring points $x_k(A), x_{k+1}(A)$. Further, for simplicity reasons, we denote $W(x_k(A), x_{k+1}(A)) = W_{k,k+1}(A)$. Then an average monotone measure and stochastic average monotone measure have a view

$$\mu(A) = \frac{\sum_{k=1}^{|A|} w_{k,k+1}(A)}{\sum_{j=1}^{|X|} w_{k,k+1}(X)}, \mathrm{M}(A) = \frac{\sum_{k=1}^{|A|} W_{k,k+1}(A)}{\sum_{j=1}^{|X|} W_{k,k+1}(X)} \tag{1}$$

correspondingly. We call M a stochastic monotone measure of informativeness if $W_{k,k+1}(A)$, $A \in 2^X$ are random variables. In this case M has random values. In this section we find estimates of numerical characteristics of M assuming that random variables $W_{k,k+1}(A)$, $W_{l,l+1}(A)$ are independent if $|l - k| > 1$. This situation appears if we suppose that x_k, $k = 1, ..., n$, are also independent random variables.

We see that $\mathrm{M}(A) = \frac{\xi}{\eta}$, where $\xi = \sum_{k=1}^{|A|} W_{k,k+1}(A)$ and $\eta = \sum_{j=1}^{|X|} W_{k,k+1}(X)$. The following lemma is used for estimating $\mathbf{E}[\mathrm{M}(A)]$ and $\sigma^2[\mathrm{M}(A)]$.

Lemma 1. *Let ξ and η be random variables that taking values in the intervals l_ξ, l_η respectively on positive semiaxis and $l_\eta \subseteq ((1-\delta)\mathbf{E}[\eta], (1+\delta)\mathbf{E}[\eta])$, $l_\xi \subseteq (\mathbf{E}[\xi] - \delta\mathbf{E}[\eta], \mathbf{E}[\xi] + \delta\mathbf{E}[\eta])$. Then it is valid the following formulas for mean and variance of distribution of $\frac{\xi}{\eta}$ respectively*

$$\mathbf{E}\left[\frac{\xi}{\eta}\right] = \frac{\mathbf{E}[\xi]}{\mathbf{E}[\eta]} + \frac{\mathbf{E}[\xi]}{\mathbf{E}^3[\eta]}\sigma^2[\eta] + \frac{1}{\mathbf{E}^2[\eta]}\mathbf{Cov}[\xi, \eta] + r_1, \tag{2}$$

$$\sigma^2\left[\frac{\xi}{\eta}\right] = \frac{1}{\mathbf{E}^2[\eta]}\sigma^2[\xi] + \frac{\mathbf{E}^2[\xi]}{\mathbf{E}^4[\eta]}\sigma^2[\eta] - \frac{2\mathbf{E}[\xi]}{\mathbf{E}^3[\eta]}\mathbf{Cov}[\xi, \eta] + r_2, \tag{3}$$

where $\mathbf{Cov}[\xi, \eta]$ is a covariation of random variables ξ and η, i.e. $\mathbf{Cov}[\xi, \eta] = \mathbf{E}[(\xi - \mathbf{E}[\xi])(\eta - \mathbf{E}[\eta])]$; r_1, r_2 are the residuals those depends on numerical characteristics of ξ and η. It being known that $|r_1| \le \frac{\delta}{1-\delta} \cdot \frac{\mathbf{E}[\xi]+\mathbf{E}[\eta]}{\mathbf{E}^3[\eta]}\sigma^2[\eta] \le \frac{\mathbf{E}[\xi]+\mathbf{E}[\eta]}{(1-\delta)\mathbf{E}[\eta]}\delta^3$, $|r_2| \le C\delta^3$.

This lemma is proved with help of expanding the function $\phi(x, y) = \frac{x}{y}$ into a Taylor series at the point $(\mathbf{E}[\xi], \mathbf{E}[\eta])$.

We will use formulas (2) and (3) without their residuals. Respective values $\tilde{\mathbf{E}}[\mathrm{M}(A)] = \mathbf{E}[\mathrm{M}(A)] - r_1$, $\tilde{\sigma}^2[\mathrm{M}(A)] = \sigma^2[\mathrm{M}(A)] - r_2$ we will call by estimations of numerical characteristics.

Introduce the following notation: $S(A) = \sum_{i=1}^{|A|} \mathbf{E}[W_{i,i+1}(A)]$, $K(A, X) = \sum_{i=1}^{|A|} k_i^X(A)$, where $k_i^X(A) = \sum_{j=1}^{|X|} \mathbf{Cov}[W_{i,i+1}(A), W_{j,j+1}(X)]$, $A \in 2^X$.

Then the formulas for $\tilde{\mathbf{E}}[\mathrm{M}(A)]$ and $\tilde{\sigma}^2[\mathrm{M}(A)]$ based on (2) and (3) can be written in the form

$$\tilde{\mathbf{E}}[\mathrm{M}(A)] = \frac{S(A)}{S(X)} + \frac{S(A)}{S^3(X)}K(X,X) - \frac{1}{S^2(X)}K(A,X), \quad (4)$$

$$\tilde{\sigma}^2[\mathrm{M}(A)] = \frac{1}{S^2(X)}K(A,A) + \frac{S^2(A)}{S^4(X)}K(X,X) - \frac{2S(A)}{S^3(X)}K(A,X). \quad (5)$$

4 Stochastic Measure of Informativeness by Contour Length

Assume that an original contour is corrupted by noise. In this case, $X = \{x_k + \mathbf{n}_k\}_{k=1}^m$, $x_k \in R^2$ and $\mathbf{n}_k = (\xi_k, \eta_k)$ are random variables. Suppose also that ξ_k, η_k, $k = 1, ..., m$, are independent, normally distributed and such that $\mathbf{E}[\xi_k] = \mathbf{E}[\eta_k] = 0$, $\sigma^2[\xi_k] = \sigma^2[\eta_k] = \sigma^2$, $k = 1, ..., m$. In this section we consider a monotone stochastic measure M defined by contour length. This measure has the view of (1), where $W_{k,k+1}(A) = |x_{k+1}(A) + \mathbf{n}_{k+1}(A) - x_k(A) - \mathbf{n}_k(A)|$. We investigate such characteristics of $\mathrm{M}(A)$ as $b[\mathrm{M}(A)] = \tilde{\mathbf{E}}[\mathrm{M}(A)] - \mu(A)$ and $\tilde{\sigma}^2[\mathrm{M}(A)]$. In general, the random variable $\sum_{k=1}^{|A|} W_{k,k+1}(A)$ is not satisfied to conditions of Lemma 1. However the probability of large deviations of random length of noisy polygonal line from non-noisy length will be small if the variance of noise is small. Therefore we assume that the random length satisfied approximately to conditions of Lemma 1. Suppose that $W_{k,k+1}(X)$, $k = 1, ..., m$, are independent random variables. This requirement can be satisfied by the choice of some subcontour (basic contour) from the initial contour.

4.1 Numerical Characteristics of Random Variable $W_{k,k+1}(A)$

We give asymptotic formulas for $\mathbf{E}[W_{k,k+1}(A)]$ and $\sigma^2[W_{k,k+1}(A)]$. These formulas proved with help expansion of the random variable $W_{k,k+1}(A)$ by Taylor formula.

Proposition 1. *The following asymptotic equalities are valid*

$$\mathbf{E}[W_{k,k+1}(A)] = l_k\left(1 + \frac{\sigma^2}{l_k^2} + \frac{\sigma^4}{2l_k^4} + O\left(\frac{\sigma^6}{l_k^6}\right)\right),$$

$$\sigma^2[W_{k,k+1}(A)] = 2\sigma^2\left(1 - \frac{\sigma^2}{l_k^2} + O\left(\frac{\sigma^4}{l_k^4}\right)\right),$$

where $l_k = |x_{k+1}(A) - x_k(A)|$.

Corollary 1. *It is true the equality*

$$S(A) = \sum_{k=1}^{|A|}\mathbf{E}[W_{k,k+1}(A)] = L(A) + \sigma^2\sum_{k=1}^{|A|}|x_{k+1}(A) - x_k(A)|^{-1} + \sigma O\left(\frac{\sigma^3}{l^3}\right),$$

where $L(A)$ *is the length of contour* A *without an influence of noise, i.e.* $L(A) = \sum_{k=1}^{|A|}|x_{k+1}(A) - x_k(A)|$, *and* $l = \min\limits_k |x_{k+1}(A) - x_k(A)|$.

Next we compute the covariance $\mathbf{Cov}\left[W_{k-1,k}(A), W_{k,k+1}(A)\right]$ between random variables $W_{k-1,k}(A)$.

Let $\mathbf{l}_i = \mathbf{l}_i(A) = x_{i+1}(A) - x_i(A)$ be a i-s segment-vector of polygon A and $\alpha_i = \alpha(x_i) = \left(\overset{\wedge}{\mathbf{l}_{i-1}, \mathbf{l}_i}\right)$.

Proposition 2. *We have*

$$\mathbf{Cov}\left[W_{k-1,k}(A), W_{k,k+1}(A)\right] =$$

$$= -\sigma^2 \cos\alpha_k \left(1 - \left(\frac{1}{l_{k-1}^2} + \frac{\cos\alpha_k}{2l_{k-1}l_k} + \frac{1}{l_k^2}\right)\sigma^2 + o\left(\frac{\sigma^2}{l^2}\right)\right),$$

where $l_k = l(x_k) = |x_{k+1}(A) - x_k(A)|$, $l = \min\{l_{k-1}, l_k\}$.

Calculate the covariance $K(A,X) = \sum_i k_i^X(A)$ between the all segments of polygon A and all segments of basic polygon X with help of last corollary. Let $\alpha(x)$ ($\beta(x)$) be an inner angle of polygon A (polygon X) in vertex x, $\gamma(x)$ be an angle between the vectors $x_{+1}(A) - x$, $x_{+1}(X) - x$, where $x_{+1}(A)$ ($x_{+1}(X)$) is the next point w.r.t. x in the contour A (contour X).

Corollary 2. *We have*

$$K(A,X) = 4\sigma^2 \sum_{x\in A} \cos\frac{\alpha(x)}{2}\cos\frac{\beta(x)}{2}\cos\left(\gamma(x) + \frac{\alpha(x)-\beta(x)}{2}\right) + \sigma^2 o\left(\frac{\sigma}{\Delta(A)}\right)$$

for $A \in 2^X$, *where* $\Delta(A) = \min\limits_i \{\Delta_i(A)\}$.

4.2 The Numerical Characteristics of Stochastic Measure of Informativeness by Length

We will find numerical characteristics of stochastic measure of informativeness by length using the results of the previous item. We will consider the application of calculated characteristics to solution of task of finding polygonal representation of curve that is stable to noise.

At first we formulate the theorem about expectation of stochastic measure of informativeness by length $\mathbf{E}\left[\mathrm{M}(\cdot)\right]$ on 2^X that follows from equality (4), Corollaries 1, 2.

Theorem 1. *The asymptotic equality*

$$\tilde{\mathbf{E}}\left[\mathrm{M}(A)\right] = \frac{L(A)}{L(X)} + \frac{C_1(A)}{L^2(X)}\sigma^2 + o\left(\frac{\sigma^2}{\Delta^2(A)}\right), A \in 2^X \tag{6}$$

is true, where

$$C_1(A) = -L(A)\sum_{x\in X}|\mathbf{l}_x|^{-1} + L(X)\sum_{x\in A}|\mathbf{l}_x|^{-1} + 4\frac{L(A)}{L(X)}\sum_{x\in X}\cos^2\frac{\beta(x)}{2} -$$

$$-4\sum_{x\in A}\cos\frac{\alpha(x)}{2}\cos\frac{\beta(x)}{2}\cos\left(\gamma(x) + \frac{1}{2}\alpha(x) - \frac{1}{2}\beta(x)\right).$$

Notice that the asymptotic formula for bias $b\left[\mathrm{M}(A)\right] = \tilde{\mathbf{E}}\left[\mathrm{M}(A)\right] - \mu(A)$ of stochastic measure of informativeness by length with noise follows from (6):

$$b\left[\mathrm{M}(A)\right] = \frac{C_1(A)}{L^2(X)}\sigma^2 + o\left(\frac{\sigma^2}{\Delta^2(A)}\right), \quad A \in 2^X,$$

and also $C_1(X) = 0$.

Similarly, we will find the asymptotic formula for variance of stochastic informational measure by length with help of formula (5), Corollaries 1, 2.

Theorem 2. *The asymptotic equality*

$$\tilde{\sigma}^2\left[\mathrm{M}(A)\right] = 4\frac{C_2(A)}{L^2(X)}\sigma^2 + o\left(\frac{\sigma^2}{\Delta^2(A)}\right), \quad A \in 2^X$$

is true, where

$$C_2(A) = \sum_{x\in A}\cos^2\frac{\alpha(x)}{2} + \frac{L^2(A)}{L^2(X)}\sum_{x\in X}\cos^2\frac{\beta(x)}{2} -$$

$$-2\frac{L(A)}{L(X)}\sum_{x\in A}\cos\frac{\alpha(x)}{2}\cos\frac{\beta(x)}{2}\cos\left(\gamma(x)+\frac{1}{2}\alpha(x)-\frac{1}{2}\beta(x)\right).$$

The value of random error (the variance of stochastic informational measure) characterizes the degree of stability of informational measure of curve with respect to level of curve noise. We can put the task about finding of polygonal representation of fixed cardinality $A \in 2^X$, $|A| = k$, which minimized the value of variance of stochastic informational measure by length. As can be seen from Theorem 2 the polygonal representation

$$A = \arg\min_{A\in 2^X,\ |A|=k} C_2(A)$$

is a solution of indicated task for not great level of curve noise σ.

Example 1. Let $X = \{x_1, ..., x_6\}$ be an ordered set of vertexes of regular 6-gon with length of segment is equal 1. Calculate the value $C_2(A)$ for various polygonal representations A of cardinality $|A| = 3$: $A_1 = \{x_1, x_3, x_5\}$, $A_2 = \{x_1, x_2, x_4\}$, $A_3 = \{x_1, x_2, x_3\}$ (Figure 1). Since $\beta(x) = \frac{2\pi}{3}$, $x \in X$, $L(X) = 6$, $\sum_{x\in X}\cos^2\frac{\beta(x)}{2} = 1.5$, then

$$C_2(A) = \sum_{x\in A}\cos^2\frac{\alpha(x)}{2} + \frac{L^2(A)}{24} - \frac{L(A)}{6}\sum_{x\in A}\cos\frac{\alpha(x)}{2}\cos\left(\gamma(x) + \frac{\alpha(x)}{2} - \frac{\pi}{3}\right).$$

Therefore $C_2(A_1) = 1.125$, $C_2(A_2) = 1.25$, $C_2(A_3) = \frac{56+22\sqrt{3}-5\sqrt{2}-3\sqrt{6}}{48} \approx 1.66$. Thus the contour A_1 is a most stable contour to noise pollution with respect to informational measure by length among of contours of cardinality is equal 3.

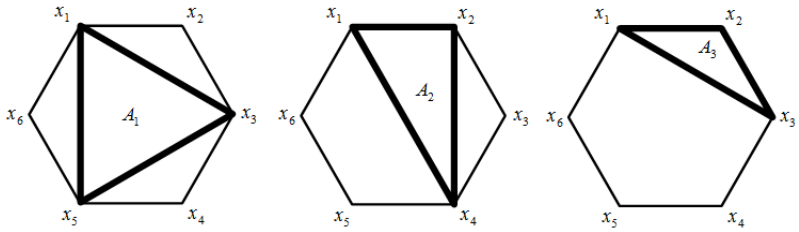


Fig. 1. Various polygonal representations A of X of cardinality $|A| = 3$.

5 Summary and Conclusions

The problem of finding of optimal stable pattern representation was discussed in this paper. We got geometrical conditions for finding a most stable contour to noise pollution with respect to measure of informativeness by length among of contours of equal cardinality. That geometrical conditions may be used to formulation and solution of others problems of finding of optimal stable pattern representation.

Acknowledgements

I would like to thank Andrew Bronevich for his helpful and stimulating comments on the manuscript of my paper. The study was implemented in the framework of The Basic Research Program of the Higher School of Economics in 2012. This work was supported by the grants 11-07-00591 and 10-07-00135 of RFBR (Russian Foundation for Basic Research).

References

1. Bronevich, A., and Lepskiy, A., "Geometrical fuzzy measures in image processing and pattern recognition", *Proc. of the 10th IFSA World Congress*. Istanbul, Turkey, 151–154 (2003).
2. Duda, R.O., Hart, P.E., and Stork, D.G., *Pattern Classification and Scene Analysis: Part I Pattern Classification*, John Wiley & Sons (1998).
3. Jolliffe, I.T., *Principal Component Analysis*, Series: Springer Series in Statistics. 2nd ed., Springer, NY (2002).
4. Lepskii, A.E., "On Stability of the Center of Masses of the Vector Representation in One Probabilistic Model of Noiseness of an Image Contour", *Automation and Remote Control*, 68, 75–84 (2007).
5. Lepskiy, A.E., "Stable Feature Extraction with the Help of Stochastic Information Measure", *Lecture Notes in Computer Science(LNCS)*, vol. 6744, Berlin / Heidelberg: Springer Verlag, 54–59 (2011).
6. Shiryaev, A.N., *Probability*. Graduate Texts in Mathematics, Springer Pub. (1995).
7. Wang, Z., and Klir G.J., *Generalized Measure Theory*, IFSR International Series on Systems Science and Engineering. 25, Springer (2009).

Marriage Dynamics Modeling: a Physical Approach

Valery A Masyukov

Russian Academy of Sciences, Ishlinsky Institute for Problems in Mechanics, Moscow, Russia
E-mail: masyuk@ipmnet.ru

Abstract: A deterministic age-structured two-sex dynamic model for marital union formation is proposed. Unlike common mathematical models with marriage function satisfying hypothetic condition of homogeneity, the proposed model is based on substantially non-homogeneous (quadratic) law of mass action. There are shown: a) the adequacy of the proposed physical model with age-structured variables, and b) non-applicability of the law of mass action when variables are integral over age. There is also developed a numerical technique for processing marriage tables, which affirms the applicability of the model for age 25-35 and allows determination of basic constants. Simple averaging of obtained analytical solution over age gives a satisfactory agreement with statistics for age 16-25 as well.
Key words: marriage function, the law of mass action, marriage table processing.

Introduction

Modeling marriage kinetics is important for understanding many dynamical demographic processes. One can distinguish between two approaches to such modeling: physical and mathematical. Both approaches use mathematics but physicists try to first simplify the problem following the ancient principle, a wonderful phrase of which is attributed to A. Einstein: “Things should be made as simple as possible. But not simpler” [1]. Mathematicians prefer solving the problem in the most general and complex form what often makes it difficult for comprehension.

The simplest physical model of pare marriage dynamics could have been based on quadratic law of mass action (LMA) applied to values integral over age (all designations are given in the end of the article):

$$M = \int_A F dA \sim \left(\int_A N dA \right)^2 = P^2 \qquad (1)$$

However, all known data indicate that M is roughly proportional to P:

$$M \sim P \quad \text{or} \quad M / P \equiv m \approx const\,, \qquad (2a)\ (2b)$$

where m weakly depends not only on P, but also on other factors. According to Demoscope [2], m ranges from 2.9 to 12.5 $^{0}/_{00}$ per year in 1950-2005 for 40 industrialized countries with P = 1-300 million.

The apparent contradiction between (1) and (2) has led many researchers to complete denial of applicability of LMA to nuptiality. Mathematicians have imposed on MF F the hypothetical condition of homogeneity:

$$F(nS_M, nS_W) = nF(S_M, S_W) \qquad (3)$$

A brief history of this issue can be found e.g. in the book by Ianelli et al [3], and in the paper by Hadeler et al [4]. One can easily accept the inapplicability of integral quadratic expressions (1) because it exceeds the reasonable limit of simplification. Indeed, the number P includes people either of premarital age or married. Both of them are not directly involved in formation of MUs, but in sum constitute a large part of the population. Besides, MUs are usually formed between people in a fairly narrow age range. On the contrary, expression (1) allows marriage formation with equal probability between men and women with any difference in age, positive or negative.
The above quote tells an alternative way to resolve the contradiction between (1) and (2): one has to one-step complicate the problem, moving from integral form of LMA (1) to differential (age-structured) one:

(4) $$F(A,T) = \sigma S_W(A,T) S_M(A,T)$$

In fairness it should be noted that type (4) MF is sometimes used *a priori* in more complex models (e.g., Heesterbeek and Metz [5], and Masyukov [6]).
The purpose of this work is detailed studying the LMA (4) and proving its applicability in some age limits with σ being constant in every generation.

Background and problem statement

To examine the kinetics of MU formation in the most "pure" form, the following additional simplifications have been made:
The model is homogeneous, i.e. its parameters are constant over territory. So, both integral and local values can be used. The latter (density values) are preferred because their absolute meanings do not differ very much from 1.
There are mostly considered real generations, i.e. cohorts of men and women born within a relatively short period of time (1-5 years). These generations correspond to straight lines (characteristics) 1-2 on the plane (A,T) (Fig.1). Demographers call such an approach the longitudinal analysis while mathematicians and physicists call it the method of characteristics. It greatly simplifies the problem by solving an ordinary differential equation (8) instead of equation in partial derivatives on T and A.
Inside each generation, ages 16-35 are considered. 16 years is the common average age of nuptiality onset. Below 35 years the rate of MU dissolution can be neglected compared with that of MU formation. In the range indicated, actual MF monotonically grows from zero to a maximum at age 20-22 and then monotonically decreases (Fig.2). The growth is due to stochastic processes. Its accurate account, as e.g. in the works by Diekmann [7], and Kaneko [8], would considerably complicate the picture of marriage kinetics. Therefore, we assume that, in each real generation, all people begin entering marital relations at the same efficient age (A_{0W} for women and A_{0M} for men, Fig. 1-2). Usually A_{0M} exceeds A_{0W} by 1–2 year.

This simplification allows replacing the continuous initial plot of real MF by a jump (step) of idealized MF what dramatically simplifies the mathematical side of the case. As shown in Fig.2, one can easily switch back from idealized to real MF by simple averaging the first model.

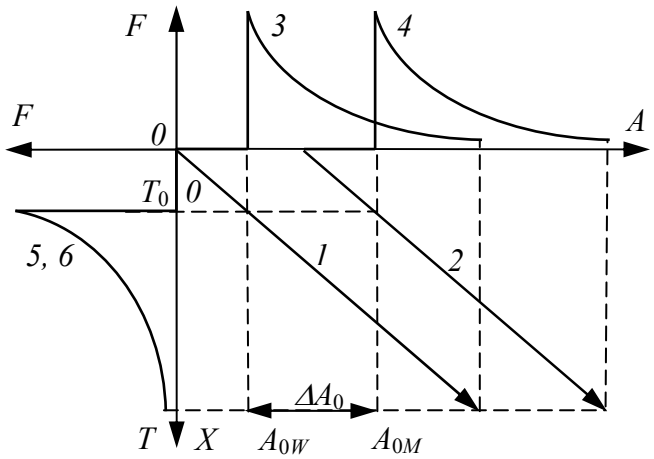


Fig.1. Idealized marriage function F in 3 projections. A - age; X - marital age; T -time; A_{0W}, A_{0M} - effective ages of marriage relation onset for women and men respectively with a difference ΔA_0 between them; T_0 - effective time of marriage relation onset; 1,2 - F projections onto the plane (A,T); 3,4 - F projections onto the plane (A,F); 5,6 - F projections onto the planes (T,F) and (X,F); 1, 3, 5 - women; 2, 4, 6 - men.

Mortality and migration are neglected in the considered age limits. Therefore in every real generation, densities N_M and N_W are supposed constant. Moreover, as a first approximation, we assume $N_M = N_W = N$.

MUs are monogamous and formed only between real generations, i.e. in a relatively small age difference. It is quite opposite to integral condition (1) when each generation of one sex can be equally likely to form unions with any generation of the other sex.

The basic independent variable of the model is marital age X and its dimensionless analogue x:

(5a) $X = A - A_{0W} = A - A_{0M}$; $x = \sigma N X = \mu X$ (5b)

i.e. the age of each marital partner, measured not from his or her YB but from the age of entry into marital relations. This allows usage of one variable (5) for both interacting generations (Fig.1). It also means that MF $F(X)$ is the same for each pair of interacting generations (Fig.1).

It follows from the condition of monogamy that in every marital age X the number of married women M_W in one of the two interacting generations is

equal to the number of married men N_M in other generation and is equal to the number U of MU. So the number of free (single) women S_W and men S_M equals the total number of women and men N minus U:

(6) $$S_{W,M} = N - M_{W,M} = N - U$$

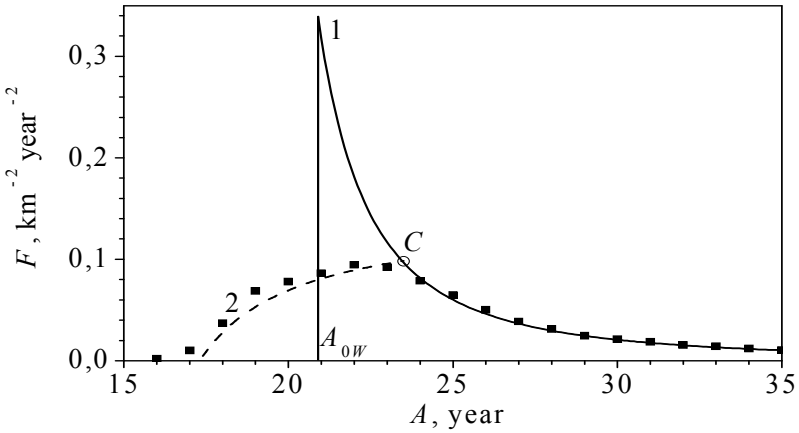


Fig.2. Marriage function F depending on age A for real woman generation born in 1850 in England and Wales. 1 –calculation on idealized formula (10a) with account of (5a) for $\sigma = 0.332$ km^2, N = 1.00 year^{-1} km^{-2}, $A_{0W} = 20.9$ year; 2 (dash line) – line 1 averaged in limits from A - 3.5 to A + 3.5 year; dots are longitudinal data obtained on the basis of marriage tables [9]; C - the intersection point of curves 1 and 2.

MU is understood in the broadest (generalized) sense: it can be marriage, and cohabitation, and non-legal union, and any other form of actual monogamous heterosexual partnership, which can be characterized by some common average characteristics.

The basic starting point of this work is that MF F obeys the LMA (4) and in view of (6) can be expressed as follows:

(7a) $$F = \sigma S_W S_M = \sigma (N-U)^2; \qquad f = (1-u)^2 \qquad (7b)$$

As can be seen, F and f by no means satisfy the condition of homogeneity (3). In general, F can depend on many variables, but thanks to the above simplifications, all values in (7), except constant σ, are functions of the only variable X (or x), F being equal to 0 for $X < 0$ in all generations.

Until recently, MF F has been a well measured characteristic because almost all marriages have been registered in many countries and formed the basis of so called marriage tables. In recent years in some countries significant portion of MUs is not registered what introduces certain difficulties in the

comparison of theory and experiment, as the marriage market of registered and unregistered unions are equivalent in terms of this model.
Thus, in view of the above suppositions, the problem statement can be briefly formulated as follows: to find MU density $U(X)$ and other related functions for each real generation and compare results with statistical data.

Problem solution

MF F (7) is nothing but intensity of MU formation. Therefore, taking into account the proposed lack of MU dissolution, for $U(X)$ and $u(x)$ we can write the following ordinary differential nonlinear equations of the first order with zero initial condition $U = u = 0$ for $X = x = 0$:

(8a) $dU / dX = F = \sigma(N - U)^2$; $du / dx = f = (1 - u)^2$ (8b)

These equations are the same both for women of one generation and for men of other generation interacting with the first one. It is worth underlying that this equation is "bound" to the real generation of women or men and, as physicists and mathematicians say, is a substantial (total) derivative along the characteristics of the corresponding equation in partial derivatives on variables A and T. Demographers and sociologists call this approach a longitudinal analysis. It is one-dimensionality of equations (8) which makes advantage of this analysis. We follow the number of MU over time (age) as if we were inside the generation.
After variable separation and use of table integrals one can get the following solutions of equations (8):

(9a) $U = \sigma N^2 X / (1 + \sigma N X)$; $u = x / (1 + x)$ (9b)

(10a) $F = \sigma N^2 / (1 + \sigma N X)^2$; $f = 1 / (1 + x)^2$ (10b)

(11a) $F^{-1/2} = N^{-1}\sigma^{-1/2} + \sigma^{1/2} X$; $f^{-1/2} = 1 + x$ (11b)

(12a) $UF^{-1/2} = N\sigma^{1/2}(A - A_0)$; $uf^{-1/2} = x$ (12b)

The function $F^{-1/2}$ (11a) is introduced for easy comparison of calculation and experiment, because it happens to be proportional to the age X (or A) and should be represented by a straight line whose slope is determined only by σ.
Even more convenient is the expression (12a) for $UF^{-1/2}$, which determines not only the combination of parameters $N\sigma^{1/2}$ as the angle coefficient of the line (12a), but also the effective age A_0 as zero of this function (Fig.3). Dimensionless formulas (9b-12b) are useful in that they are the same not only for all real generations, but consequently for all conditional generations in the so-called transversal analysis. Therefore, it can be stated that in every conditional generation, union density u is close to 1 for sufficiently large x. This agrees with (2), according to which the total annual number of marriages M is roughly proportional to the total population P.

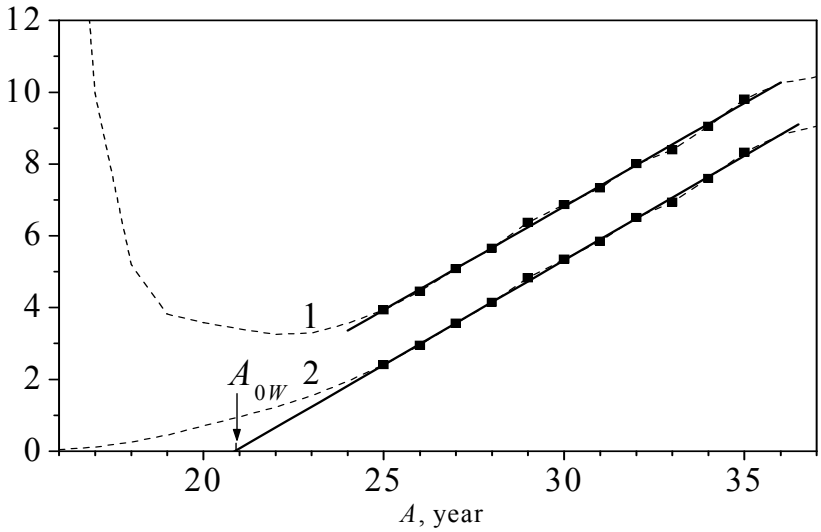


Fig. 3. 1 - function $F^{-1/2}$, year km (11a), 2 – function $UF^{-1/2}$, km $^{-1}$ (12a), A - age of women born in 1850 in England and Wales [9]. A_{0W} - effective age of marital relation onset. Dots are experimental data subjected to linear regression (solid lines).

Determination of parameters μ, N, A_0 from marriage tables

Numerical processing of marriage tables [9] was carried out on the Matlab platform. First of all, in initial tables, total numbers of marriages in five-year intervals were replaced by the mean value of MF in these intervals, which were then attributed to the median values of given intervals (22, 27, 32, and 37 year). These data, together with one-year data for age 16-19 were interpolated to a uniform scale of age 16-37 with one-year step for 1851-2003. After dividing the data by the territory area (151200 km^2), the matrix of discrete MF values having size 153x22 and one-year increment on both axes was obtained. Next, all of 131 complete diagonals of this matrix were found, each diagonal presenting longitudinal MF in a real generation from 1835 to 1963 YB. Discrete MF F for each of these generations was subjected to the following mathematical procedures:

1) A discrete function $F^{-1/2}$ was found, which points fitted well a straight line for ages 25-35 in full agreement with (11a) (Fig.3). The angle coefficient $\sigma^{1/2}$ of this line was defined by a standard method of linear regression. Raising the latter to the second power gave σ values for each generation.

2) By cumulative summation of F, the function U (9a) as well as the function $UF^{-1/2}$ (12a) were found, the latter dots fitting well a straight line in full accordance with (12a) (Fig.3). Similarly there was defined angular coefficient $N\sigma^{1/2}$ of these lines according to (12a). The intersection of these lines with

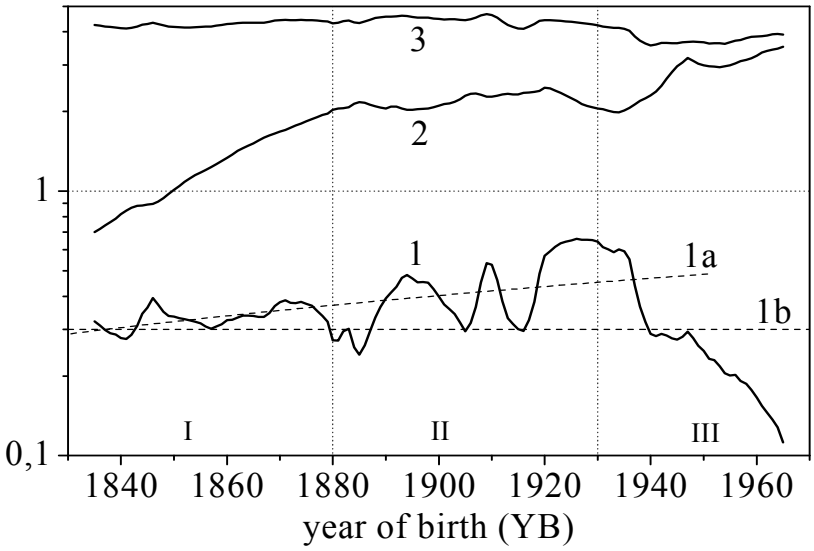


Fig.4. Basic parameters of real woman generations [9] depending on YB.
1 - marriage frequency μ, year^{-1}; 1a - linear approximation of μ; 1b – minimal (undistorted) level of μ; 2 - average density N at age 25-35, year $^{-1}$km $^{-2}$; 3 – effective age A_{0W} /5 of entry in marital market, year.

horizontal axis defined effective age A_{0W} according to (12a). Obtained values of $\sigma^{1/2}$ and $N\sigma^{1/2}$ allowed calculation of $\mu = \sigma N$.

3) All obtained dimensional curves similar to those in Fig.3 were put in a dimensionless form using (11b, 12b) and values μ and N for each curve. All dimensionless data, being plotted on a single graph, fitted well theoretical lines (11b, 12b) for dimensionless age x = 1-4 with scatter no more than ±10%. This confirmed the adequacy of the proposed quadratic model.

Obtained in this way, the parameters μ, N, and A_0 for real generations are represented in Fig.4. The horizontal axis (YB) can be divided into three roughly equal ranges. In the range I for generations of 1830-1880 YB, who have married in the quiet Victorian era, there is a relatively smooth behavior of all parameters. Thereby μ is within 0.3-0.4 year $^{-1}$, A_0 - at level of 21 year and N is growing almost exponentially from 0.7 to 2 year $^{-1}$km $^{-2}$. The range II (1880-1930) refers to the time when people aged 25-35 experienced the two world wars and the beginning of the demographic transition. This range is characterized by relatively large (up to 3 times) fluctuations, the stabilization of N at level about 2 year $^{-1}$km $^{-2}$ and weak reduction of A_{0W}. Finally, in the range III (1930-1970) people experienced the so-called demographic transition. From this work point of view, the most significant

feature of this transition is a sharp decrease of μ, which can be explained by an increased percentage of non-registered MUs.

The behavior of the main characteristic - the marriage frequency μ - on average over bands I and II can be characterized in two ways. First, μ grows over time at a rate of about 1.5% per year, starting from entry-level 0.3 year $^{-1}$ and fluctuating then with increasing amplitude around the average line 1a. Second, μ remains at a constant level 0.3 year $^{-1}$ (line 1b) under stable development and occasionally climbs up under non-stable development. Lines 1a and 1b can be extrapolated onwards to the period III and thus provide an estimate of average and minimum level of actual marriage frequency in recent years.

Conclusion

There is proposed an age-structured model for marriage kinetics, based on marital interaction between two generations according to the quadratic law of mass action. A special case is considered when age population densities in interacting generations are equal. In this case, simple analytical expressions are obtained, describing the basic characteristics of nuptiality in real generations.

The model approbation is fulfilled on large marriage table base of England and Wales for 1851 - 2003. Two special functions are built that are linear with age and can be easily compared with marriage tables using developed numerical technique. As a result, the adequacy of the proposed quadratic model is shown, and basic parameters are calculated for 131 generations.

The well known contradiction between hypothetical quadratic dependence of the total annual marriage number on total population, on the one hand, and approximately linear actual dependence between these values, on the other hand, is considered. This contradiction is shown to be based on incorrectness of the above hypothetical assumption and is overcome through transition from integral age characteristics to age-structured ones.

Abbreviations and designations

const – a constant
d – sign of differential
LMA – the law of mass action
MF – marriage function F
MU – marital union
YB – year of birth
$^{0}/_{00}$ – promille, ppm (10^{-3} or 0,1 %)
A, year – age
A_{0M}, A_{0W}, A_0, year – effective age of marital relation onset for men, women, and both respectively

$\Delta A_0 = A_{0M} - A_{0W}$
F, year $^{-2}$ km $^{-2}$ (MF) – intensity of MU formation
$f = F/N/\mu$ – dimensionless MF
M, year $^{-1}$ – total annual marriage number
$m = M / P$, year $^{-1}$ – total (crude) marriage rate
M_M, M_W, year $^{-1}$ km $^{-2}$ – density of married men or women per unit age and unit territory
$N_M = N_W = N$, year $^{-1}$ km $^{-2}$ – total density of men or women per unit age and unit territory
n – positive number
P – total population number
S_M, S_W, year $^{-1}$ km $^{-2}$ – density of free (single) men or women per unit age and unit territory
T, year – calendar time
U, year $^{-1}$ km $^{-2}$ – MU density per unit age and unit territory
$u = U/N$ – dimensionless MU density
$X = A - A_0$, year – marital age
$x = \mu X$ – dimensionless marital age
$\mu = \sigma N$, year $^{-1}$ – marriage frequency
σ, km^2 – marriage square

References

1. http://quoteinvestigator.com/2011/05/13/einstein-simple/
2. http://demoscope.ru/weekly/app/app4081.php
3. M. Ianelli, M. Martcheva, F .A. Milner, Gender-Structured Population Modeling: Mathematical Methods, Numerics, and Simulations. SIAM (2005).
4. K. P. Hadeler, R. Waldstatter, and A. Worz-Busekros, Models for pair formation in bisexual populations, *J. Math. Biol.* 26, 635-649 (1988).
5. J. A. P. Heesterbeek, J. A .J. Metz, The saturating contact rate in marriage and epidemic models. *J. Math. Biol.* 31, 529-539 (1993).
6. V. A. Masyukov, Stationary Population Model Based on Marital Kinetics, *Preprint IPMech RAS* # 878, Moscow (2008) (in Russian).
7. A. Diekmann, Diffusion and Survival Models for the Process of Entry into Marriage, *J.Math.Siciol.* 14, 31-44 (1989).
8. R. Kaneko, Elaboration of the Coale-McNeil Nuptiality Model as the Generalized Log Gamma Distribution: A New Identity and Empirical Enhancements, *Demographic Research* 9, 223-262 (2003).
9. UK National Statistics. England and Wales, Marriages: 1846-2003, Females. URL: http://www.statistics.gov.uk/

STOCHASTIC MODELING TECHNIQUES FOR GESTURE-SOUND MAPPING

Vassilios Matsoukas[1], Sotirios Manitsaris[2], Athanasios Manitsaris[3]

[1,3]Multimedia Technology and Computer Graphics laboratory (MTCG), Department of Applied Informatics University of Macedonia, Thessaloniki, Greece
Email: vmats@uom.gr, manits@uom.gr
[2]SIGnal processing and Machine Learning laboratory (SIGMA), Ecole Nationale Supérieure de Physique et de Chimie Industrielles de la Ville de Paris (ESPCI ParisTech), France
Email: sotiris.manitsaris@espci.fr

Abstract: In this paper we propose an approach to achieve real-time and continuous gesture following and recognition, focusing on the use of finger gestures to control the sound. We aim at the development of a system that performs real-time gesture control of sound by using stochastic modeling techniques. Our system relies on PianOrasis, (associating the terms Piano and Orasis - meaning «vision» in greek).. We use a simple web camera to capture the gestures of all five fingers of the hand with no use of sensors. After storing the input data we need to train our system during a learning procedure that uses machine learning techniques and in particular HMMs. In order to achieve real-time gesture recognition we apply the "gesture following" method, where Hidden Markov Models are used to update "continuously" the gesture descriptors and choose temporal mapping of high-level features of finger motions to low-level audio descriptors rather than spatial.
Keywords: Gesture recognition, feature extraction, gesture–sound mapping, stochastic modeling, machine learning

1 Introduction

The aim of this research is to propose a computer vision methodology for real-time gesture control of music without using any musical instrument. A simple web camera is used to capture finger musical gestures performed in space. Image analysis techniques are applied in order to detect and identify the fingertips on a video. Finger gesture recognition and prediction is based on the stochastic modeling of the extracted high-level features with the help of Hidden Markov Models and Gaussian Mixture Models. We also propose a methodology for real-time performance of the system

Most of the recent developments in the area of gestural control of musical performance focuses on the analysis and recognition of human gestures like in [1], [2] and [3] and data acquisition, where the use of expensive, complicated and non-portable equipment makes it impractical for musical performances. Also, most of the already existing systems recognize postures or hand gestures and do not focus on real-time performance.

We propose the use of finger gestures for controlling musical performance and we are developing a system which will recognize each finger individually and predict the next gesture in real-time by using stochastic modeling techniques. Our system

will also be simple cost-effective and easily transportable since we propose a standard video camera to capture finger gestures.
For the real-time performance of our system we are adopting the "gesture following" method proposed in [4], where HMMs are used for a system, which "continuously" updates parameters characterizing the performance of a gesture.
After recognition and feature extraction of the gestures we use the MnM mapping toolbox developed in [5] to map the high-level features extracted from the finger gestures to the low-level parameters of the sound.

2 Methodology

2.1 System architecture

The proposed system uses an installation of a camera in front of a piano or on a table, so as to fully record finger motions in two dimensions as proposed in [6]. The prerecorded gestures along with their corresponding sound segments are being stored in a database. The following feature vectors are extracted from the video frame according to the method proposed in [3] (Figure 1): (a) the differences of the ordinates between the fingers and the centroid, (b) the abscissa of the fingers and (c) differences between the abscissas of adjacent fingers.

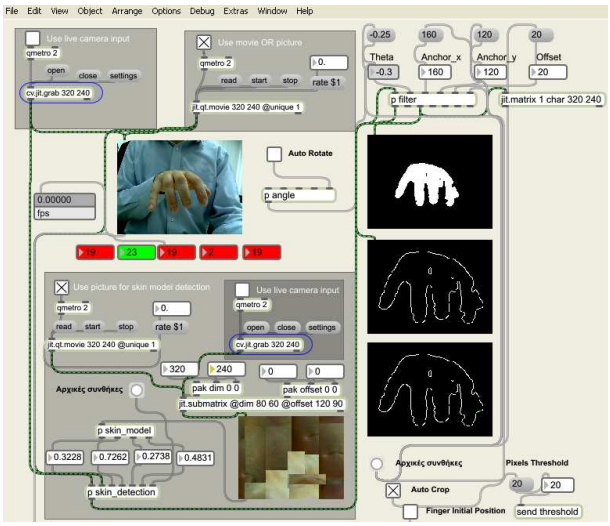


Fig. 1. High-level feature extraction in Pianorasis:max Interface

In [14], the video signal is imported into the computer and is processed using MATLAB Image Acquisition Toolbox (PianOrasis). In our system we are currently implementing this procedure using Max/MSP in order to recognize finger gestures in real-time (PianOrasis:max). Currently PianOrasis performs static and dynamic finger gesture recognition while in PianOrasis:max only static recognition has been implemented and dynamic recognition is under development.

In order to recognize and timely "follow" the gesture we will use the method proposed in [4], which relies on the modeling of multi-dimensional temporal curves using HMMs.. There are 14 features extracted and can be feed to the "Gesture follower system" proposed in [4], which works with any type of regularly sampled multidimensional data flow. At this stage we will also use the sonification feedback technique proposed by [16]. According to this work the sonification feedback will be an audio display, of the changing probabilities and observed states within the HMMs. A Gaussian function of the distances will be used to penalize large distances between gesture models.
The final step is the mapping between gesture and sound. We concatenate high-level video features extracted from PianOrasis with low-level audio features. Some of the major principles describing sound characteristics are log of energy, spectrum balance, peak structure match, spectral centroid, RMS, RollOff, flux etc. We intend to use some of these major principles for our system. We will also record natural and instrumental sounds and perform noise reduction, sound-processing and audio feature extraction as proposed in [15]
The architecture of the proposed system is presented in Figure 2:

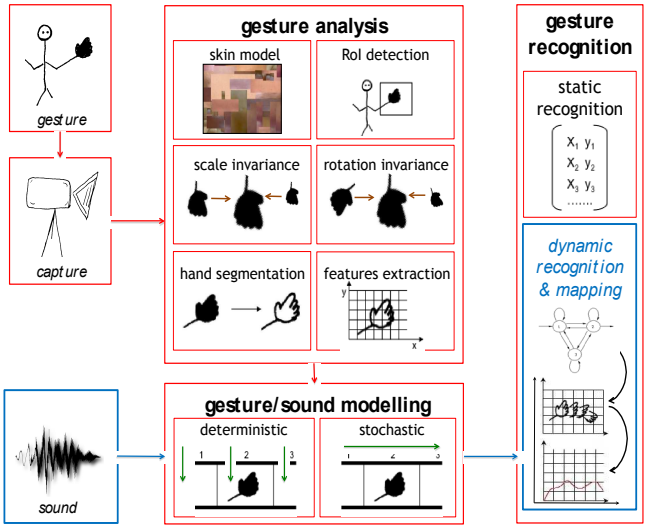


Fig. 2. System architecture.
In red: already implemented; in blue: not yet

2.2 Dynamic recognition

The dynamic recognition of the finger gestures is achieved via the PianOrasis system. As described in [3] the system uses a web cam to capture axes X, Y of the finger movements. X gives us the information about the location of the suitable region of keys and Y refers to detecting preparation and key push. Image segmentation is based on the human skin. Afterwards the hand contour is extracted and the centroid of the hand is calculated for every frame. The local maxima are

detected which give us five points matching to the five fingertips. Finally the feature vectors are extracted for each frame by calculating a) the differences of the ordinates between the fingers and the centroid, (b) the abscissa of the fingers and (c) the differences between the abscissas of adjacent fingers.

2.3 Gesture-sound mapping techniques

"Gesture-sound mapping" is called the procedure which correlates the gesture input data with the sound control parameters. Several approaches of gesture-sound mapping have been proposed in [7], [8].
A technique for geometric control is presented in [9] in which a mixture of Gaussian kernels maps a two dimensional control space into an M-dimensional sound space. Also [10] presents a technique for mapping from two dimensional control space (on-screen mouse position) into M parameters of sound processing.
In order to achieve real-time gesture-sound mapping the user should be able to dynamically modify the control parameters in order to imprint any corresponding time behavior of the sound evolution as proposed in [11].
The idea is to describe mapping procedures as a combination of relatively simple matrix operations. Consider a matrix X containing n gesture parameters and matrix Y containing m sound parameters. A simple mapping operation corresponds to a matrix multiplication with a matrix A of dimension (mxn).

$$Y = A * X \qquad (1)$$

In case of layered mapping as described in [17] we can consider for example three matrices and the mapping can be easily defined as a series of matrix multiplications:

$$Y = \left(A * B * C\right) * X \qquad (2)$$

In [5] there are also other matrix operations considered like element by element multiplication or combining a matrix with a function f applied to each element of the matrix which can be combined to design non-linear mappings.
In the work of [12] it is assumed that even though the relationship between gesture and sound might be more complex, at least part of it can be revealed and quantified by linear multivariate regression applied to the motion capture data and the audio descriptors extracted from the sound. Previous work on the quantitative analysis of gesture-sound relationship often deal with statistical methods like principal correlation analysis (PCA) or canonical correlation analysis (CCA). With CCA we can express linear relationships between two variable groups and analyze the many-to-many mapping strategy. However both methods (PCA and CCA) return a global static similarity measure without considering intrinsic dynamic changes and therefore cannot be used for temporal modeling.

2.4 Machine Learning

Our system needs to continuously update the parameters characterizing the performance of a gesture and therefore we will use the "gesture follower" approach presented in [4] and [11], where HMMs are used to achieve this goal.
The input stream consists of 14 high-level features extracted from each video frame. These features describe fully the movements of the 5 fingers. After storing the input data we need to train our system during a learning procedure that uses machine learning techniques and in particular HMMs. Each gesture template is associated to a state-based structure: each data sample represents a "state". A probability density function is associated to each state, setting the observation probability of the data. This structure can then be associated to an HMM.
For the online processing [11] suggests calculating 2 parameters to be used for the mapping procedure: a) time progression index and b) likelihood value. The first parameter is defined by computing the time warping during the performance. The likelihood value is calculated by measuring the similarity between the gesture being performed and the templates as proposed in [13].
For the temporal modeling of a sequence of incoming events, HMM-based methods have been used in audio speech recognition, gesture recognition and multimodal audio-visual speech recognition [15]. Further the authors of this paper propose to use this method to further define a statistical distance between two time profiles, typically called a divergence measure. They report that this HMM-based divergence measure has properties, induced by its underlying Markov process, that makes it suitable to study the time evolution of the similarity between gesture parameters and sound descriptors
The proposed method will be used in our system, where the gesture is described as a set of gesture parameters and the sound as a set of sound parameters like audio energy, timbre and pitch. For each gesture the model calculates a divergence measure between the gesture and the corresponding sound, namely it determines quantitatively the similarity or dissimilarity between gesture and sound.
The considered sample-based learning method trains an HMM that closely models the time evolution of the signal. Moreover, from forward decoding we can find at each time which model state emits the considered observation. Thus, at each time step the model can inform us on the close relation between both signals in terms of time evolution and amplitudes. This aims to an explicit temporal evolution of the divergence measure.

3 Implementation

We implemented our system in 2 environments. Matlab has been used to implement PianOrasis where, as mentioned in 3.1. of this paper, feature vectors are extracted and static and dynamic recognition is available. We also used the Max/MSP environment to implement our system which currently only performs static gesture recognition, namely PianOrasis:max.
In order to implement complex mappings the MnM patch has been developed in [5] and will be used in our system to model our mapping approach. This toolbox is

a series of Max/MSP externals and abstractions based on the FTM patch, which enables matrix handling. It consists of a set of modules providing basic linear algebra, mapping and statistical modeling algorithms such as PCA, GMMs and HMMs. The implemented classes include matrices, dictionaries, sequences, break point functions and tuples.
The construction of the mapping procedure is performed using both basic matrix operations from the FTM library and using the dedicated MnM set of externals and abstractions. Mapping can be thus built in a modular way.

4 Conclusions

The proposed system allows the recognition and prediction of finger musical gestures using machine-learning techniques following specific aims. It is vision-oriented to the image of the hand, without preliminary analysis, allowing the artist to feel free, with no need of special equipment and cost-effective, allowing its large-scale use. It can be used for creating melodies and composing music without a musical instrument. Other possible applications, like virtual conducting without sophisticated sensor-based equipment, can also be considered as a further step going beyond finger control of sound and using the whole hand of the performer.

5 Acknowledgments

We would like to thank Frédéric Bevilacqua from IRCAM for giving us the new beta version of GF/MuBu for Max/MSP for testing and for the interesting discussions on real-time following and recognition methods.

References

1. N. Rasamimanana, D. Bernardin, M. Wanderley, F. Bevilacqua, “String Bowing Gestures at Varying Bow Stroke Frequencies: A Case Study”, In Lecture Notes in Computer Science, Springer Verlag, vol. 5085, pp. 216–226, (2008)
2. A. Hadjakos, E. Aitenbichler, B. Wetz, M. Muehlhaeuser, “Computer Analysis of the Indirect Piano Touch: Analysis Methods and Results”, 3rd International Conference on Automated Production of Cross Media Content for Multi-channel Distribution, pp. 33-38, Firenze University Press, (2007)
3. S. Manitsaris, “Gesture Recognition in Music Interaction via Computer Vision”, 9th IASTED International Conference on Visualization, Imaging and Image Processing (VIIP 2009), Campridge, United Kingdom, (2009)
4. F. Bevilacqua, B. Zamborlin, A. Sypniewski, N. Schnell, F. Guédy, N. Rasamimanana, “Continuous realtime gesture following and recognition”, In Embodied Communication and Human-Computer Interaction, Springer Verlag, Berlin Heidelberg, vol. 5934 of Lecture Notes in Computer Science, pp. 73–84, (2010)

5. F. Bevilacqua, R. Muller, N. Schnell, “MnM: a Max/MSP mapping toolbox”, In Proc. of the International Conference on New Interfaces for Musical Expression, Vancouver, BC, (2005)
6. S. Manitsaris, “Computer vision for the gesture recognition: gesture analysis and stochastic modelling in music interaction”, Phd Thesis, University of Macedonia-Greece, (2010)
7. D. Arfib, J. Couturier, L. Kessous, V. Verfaille: “Strategies of mapping between gesture data and synthesis model parameters using perceptual spaces” Organized Sound 7(2), 127–144 (2002)
8. D. Levitin, S. McAdams, R. Adams: “Control parameters for musical instruments: a foundation for new mappings of gesture to sound”. Organised Sound 7(2), 171–189 (2002)
9. A. Momeni and D. Wessel, “Characterizing and Controlling Musical Material Intuitively with Geometric Models,” in Proc. of the 2003 conference on New interfaces for Musical Expression (NIME 03), pp. 54–62, (2003)
10. R. Bencina, “The Metasurface: Applying Natural Neighbor Interpolation to Two-to-Many Mappings,” Conference on New Interfaces for Musical Expression (NIME 05), pp. 101–104, (2005)
11. F. Bevilacqua, N. Schnell, N. Rasamimanana, B. Zamborlin, F. Guedy, “Online Gesture Analysis and Control of Audio Processing”, Musical Robots and Interactive Multimodal Systems, Springer Tracts in Advanced Robotics, Volume 74, Springer Verlag, (2011)
12. B. Caramiaux, F. Bevilacqua, N. Schnell. "Towards a Gesture-Sound Cross-Modal Analysis", Lectures Notes in Computer Science, Springer-Verlag, 2009
13. B. Caramiaux, F. Bevilacqua, N. Schnell. "Analysing Gesture and Sound Similarities with a HMM-based Divergence Measure" Sound and Music Computing (SMC'10), Barcelona, Spain (2010)
14. Manitsaris S. «Vision par ordinateur pour la reconnaissance des gestes musicaux des doigts», Revue Francophone d'Informatique Musicale, [En ligne] mis à jour le : 23/09/2011, URL:http://revues.mshparisnord.org/rfim/index.php?id=107 (2011)
15. A. Kapur, G. Tzanetakis, and P. F. Driessen, “Audio-based gesture extraction on the esitar controller” International Conference on Digital Audio Effects, (2004)
16. J. Williamson and R. Murray-Smith, “Sonification of probabilistic feedback through granular synthesis,” IEEE Multimedia, vol. 12, no. 2, pp. 45–52, 2005
17. M. Wanderley, P. Depalle: Gestural Control of Sound Synthesis, Invited paper, Proc of the IEEE, Vol. 92, No. 4, (2004)

Point and interval estimation for the Flexible Dirichlet model

Sonia Migliorati[1], Gianna S. Monti[1], and Andrea Ongaro[1]

Department of Statistics
University of Milano-Bicocca, Italy
(e-mail: sonia.migliorati@unimib.it)

Abstract. The Flexible Dirichlet distribution ([12]) allows to preserve many mathematical and compositional properties of the Dirichlet without inheriting its lack of flexibility ([1]) in modeling the various independence concepts appropriate for compositional data. The present paper addresses some inferential aspects of the model. The attention is mainly focused on maximum likelihood estimation of the parameters. This can be handled by means of the E–M algorithm which can be suitably adapted in order to take advantage of the peculiar finite mixture structure of the model. Yet, the estimation of the (asymptotic) variance-covariance matrix of the estimators is quite a challenging issue. Various alternatives have been considered and it is shown that the most feasible and efficient relies on a bootstrap evaluation of a suitable (complete-data based) expression for the observed information matrix. A simulation study is carried out to evaluate the accuracy of the proposed procedures. An application of the model to a real data set highlights its potential with respect to alternative models.
Keywords: Simplex distribution, Dirichlet mixture, E–M algorithm, Bootstrap, Compositional data.

1 Introduction

Compositional data consist of vectors of proportions, thus they are subject to a unit sum constraint, and they are prevalent in many disciplines (e.g. geology, medicine, economics, psychology, etc.). The Dirichlet model, though mathematically tractable, is unsatisfactory for modeling such data as it embodies extreme forms of simplicial independence. Many other models are present in the literature. Some of them are generalizations of the Dirichlet (see [3], [2], [6], [13], [14]). One of the most widespread alternative proposal is based on mapping the variables from the simplex to a Euclidean space by means of log-ratio transformations as proposed by Aitchison (see [1]). Though, it is still an open problem to find a tractable generalization of the Dirichlet allowing for rich dependence as well as independence properties. Recently, a new distribution on the simplex has been proposed: the Flexible Dirichlet (FD) (see [12]). Such distribution enables to model a variety of different independence concepts appropriate for compositional data, thus overcoming the extreme independence structure implied by the Dirichlet. At the same time, the FD preserves mathematical and compositional properties

of the Dirichlet to a large extent. Nevertheless, some inferential issues have not been fully tackled yet. In the present paper, after briefly recalling the definition of the FD (Section 2) we discuss the identifiability issue of the model (Section 3). Then, we propose an appropriate algorithm for computing the maximum likelihood estimator of the parameter vector (Section 4) and we derive reliable and efficient estimates of its (asymptotic) variance-covariance matrix (Section 5). A simulation study to evaluate the accuracy of the proposed procedures is given in Section 6, while Section 7 is devoted to a real data application.

2 The Flexible Dirichlet distribution

The FD distribution derives from the normalization of a basis of positive dependent random variables obtained by starting from the usual basis of independent equally scaled gamma random variables (i.e. the Dirichlet basis) and randomly allocating to the i^{th} component a further independent gamma random variable.

Its density function can be expressed as

$$f_{FD}(\mathbf{x};\boldsymbol{\alpha},\mathbf{p},\tau)=\frac{\Gamma\left(\alpha_{+}+\tau\right)}{\prod_{h=1}^{D}\Gamma(\alpha_{h})}\left(\prod_{h=1}^{D}x_{h}^{\alpha_{h}-1}\right)\sum_{i=1}^{D}p_{i}\frac{\Gamma(\alpha_{i})}{\Gamma(\alpha_{i}+\tau)}x_{i}^{\tau} \qquad (1)$$

with $\mathbf{x}\in\mathcal{S}^{D}=\left\{\mathbf{x}:\,x_{i}>0\,,i=1,\ldots,D,\,\sum_{i=1}^{D}x_{i}=1\right\}$, $\alpha_{i}>0$, $\alpha_{+}=\sum_{i=1}^{D}\alpha_{i}$, $\tau>0$, $0\leq p_{i}<1$ and $\sum_{i=1}^{D}p_{i}=1$. The FD contains the Dirichlet distribution as the inner point $\tau=1$ and $p_{i}=\alpha_{i}/\alpha_{+}$, $\forall i=1,\ldots,D$.

A key feature of the FD is that its distribution function $FD^{D}(\mathbf{x};\boldsymbol{\alpha},\mathbf{p},\tau)$ can be written as a finite mixture of Dirichlet distributions $\mathcal{D}^{D}(\mathbf{x};\boldsymbol{\alpha}+\tau\mathbf{e}_{i})$, i.e.:

$$FD^{D}(\mathbf{x};\boldsymbol{\alpha},\mathbf{p},\tau)=\sum_{i=1}^{D}p_{i}\mathcal{D}^{D}(\mathbf{x};\boldsymbol{\alpha}+\tau\mathbf{e}_{i}) \qquad (2)$$

where $\mathbf{e}_{i}$ is a vector whose elements are all equal to zero except for the i^{th} element which is one. Such mixture representation, among other aspects, allows for a variety of different shapes for the density, including multi-modality.

Many relevant properties of the FD can be found in [12], namely the distribution of marginals, conditionals and subcompositions, some fruitful representations, closure under permutation and amalgamation, expressions of joint and conditional moments. A further and particularly relevant feature of the FD model is its ability of modeling most types of simplicial independence such as subcompositional independence, left and right neutrality, subcompositional invariance, partition independence as well as complete and high order partition independences (for a definition and interpretation of all such properties see, for example, [1]).

Here we simply recall the first two moments and the conditional mean which will be useful in the following. The first two moments of the FD have the form:

$$\mathrm{E}(X_i) = \frac{\alpha_i + p_i\tau}{\alpha_+ + \tau} \tag{3}$$

$$\mathrm{E}(X_i^2) = \frac{\mathrm{E}(X_i)(1-\mathrm{E}(X_i))}{(\alpha_+ + \tau + 1)} + \frac{\tau^2 p_i(1-p_i)}{(\alpha_+ + \tau)(\alpha_+ + \tau + 1)} + \mathrm{E}(X_i)^2 \tag{4}$$

$(i = 1, \ldots, D)$.

Moreover, consider the order 1 partition $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ where $\mathbf{X}_1 = (X_1, \ldots, X_k)$ and $\mathbf{X}_2 = (X_{k+1}, \ldots, X_D)$ and denote by $\boldsymbol{\alpha}_1 = (\alpha_1, \ldots, \alpha_k)$, $\mathbf{p}_1 = (p_1, \ldots, p_k)$ and by α_+^1 and p_+^1 the corresponding totals. Then, the conditional mean of the subcomposition $\mathbf{S}_1 = \mathbf{X}_1/(X_1 + \ldots + X_k)$ is:

$$\mathrm{E}\left(\mathbf{S}_1 | \mathbf{X}_2 = \mathbf{x}_2\right) = \frac{\boldsymbol{\alpha}_1}{\alpha_+^1} + \frac{\tau}{(\tau + \alpha_+^1)}\left(\frac{\mathbf{p}_1}{p_+^1} - \frac{\boldsymbol{\alpha}_1}{\alpha_+^1}\right)\frac{p_+^1}{p_+^1 + q(\mathbf{x}_2)} \tag{5}$$

where $q(\mathbf{x}_2)$ has the form

$$q(\mathbf{x}_2) = \frac{\Gamma(\alpha_+^1 + \tau)}{\Gamma(\alpha_+^1)(1 - x_{k+1} - \ldots - x_D)^\tau} \sum_{i=k+1}^{D} p_i \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i + \tau) x_i^\tau}. \tag{6}$$

Notice that such conditional mean can accommodate both for non increasing and non decreasing behaviors.

3 Identifiability of the model

The finite mixture structure (2) allows the FD distribution to inherit the flexibility of mixture models which guarantees many useful properties (see [12]). At the same time, the FD does not share one of the typical drawbacks of such models, namely non-identifiability. Indeed, very often mixture distributions are not identifiable since they are invariant under permutations of the component labels. Identifiability of the FD model derives from the special parametric structure: each Dirichlet component of the mixture is characterized by the parameter vector $\boldsymbol{\alpha}_i = (\boldsymbol{\alpha} + \tau\mathbf{e}_i)$, which differs from the other components only in terms of the position of the parameter τ. Such aspect plays a key role in the estimation context with respect both to identifiability and to the initialization of the E–M algorithm. The former property can be proved as follows.

Proposition 1 (Identifiability of the FD) *Let* $\mathbf{X} \sim FD^D(\boldsymbol{\theta})$ *and* $\mathbf{X}' \sim FD^D(\boldsymbol{\theta'})$, *where* $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \mathbf{p}, \tau)$ *and* $\boldsymbol{\theta'} = (\boldsymbol{\alpha'}, \mathbf{p}', \tau')$. *Then* $\mathbf{X} \sim \mathbf{X}'$ *if and only if* $\boldsymbol{\theta} = \boldsymbol{\theta'}$.

Proof. We need to show that if $\mathbf{X} \sim \mathbf{X}'$ then $\boldsymbol{\theta} = \boldsymbol{\theta'}$, the converse being obvious. Clearly, if $\mathbf{X} \sim \mathbf{X}'$ then $X_i \sim X_i'$, $(i = 1, \ldots, D)$. By closure under marginalization of the FD model (see [12]), X_i has density

$$g_i(x;\boldsymbol{\theta}) = x^{\alpha_i-1}(1-x)^{\alpha_+-\alpha_i-1}\left(p_i c_i(\boldsymbol{\alpha},\tau)x^{\tau} + (1-p_i)d_i(\boldsymbol{\alpha},\tau)(1-x)^{\tau}\right)$$

for $x \in (0,1)$, where

$$c_i(\boldsymbol{\alpha},\tau) = \frac{\Gamma(\alpha_+ + \tau)}{\Gamma(\alpha_+ - \alpha_i)\Gamma(\alpha_i + \tau)}, \qquad d_i(\boldsymbol{\alpha},\tau) = \frac{\Gamma(\alpha_+ + \tau)}{\Gamma(\alpha_i)\Gamma(\alpha_+ - \alpha_i + \tau)}.$$

It follows that $g_i(x;\boldsymbol{\theta}) = g_i(x;\boldsymbol{\theta'})$ a.s., $(i = 1,\dots,D)$. As the two densities are continuous on $(0,1)$, equality must hold identically for any $x \in (0,1)$.

For any $i = 1,\dots,D$, we have that $g_i(x;\boldsymbol{\theta})x^{1-\alpha_i}$ tends to $(1-p_i)d_i(\boldsymbol{\alpha},\tau)$ when $x \to 0^+$. Therefore $g_i(x;\boldsymbol{\theta'})x^{1-\alpha_i}$ must tend to the same quantity for $x \to 0^+$. As $p_i < 1$, it is easy to check that this can happen only if $\alpha_i = \alpha_i'$. Furthermore, it must be $(1-p_i)d_i(\boldsymbol{\alpha},\tau) = (1-p_i')d_i(\boldsymbol{\alpha'},\tau')$.

If we introduce such constraints, holding for $i = 1,\dots,D$, in the equality $g_i(x;\boldsymbol{\theta}) = g_i(x;\boldsymbol{\theta'})$, for any $x \in (0,1)$ and $i = 1,\dots,D$ we obtain:

$$p_i c_i(\boldsymbol{\alpha},\tau)x^{\tau} + (1-p_i)d_i(\boldsymbol{\alpha},\tau)(1-x)^{\tau} = p_i' c_i(\boldsymbol{\alpha},\tau')x^{\tau'} + (1-p_i)d_i(\boldsymbol{\alpha},\tau)(1-x)^{\tau'}.$$

By taking the limit as $x \to 1^-$ on both sides of the above equation we have $p_i c_i(\boldsymbol{\alpha},\tau) = p_i' c_i(\boldsymbol{\alpha},\tau')$, $i = 1,\dots,D$. Then, by computing both sides in $x = 0.5$ the implication $\tau = \tau'$ is obtained, which directly leads to equality of p_i and p_i' for all i. □

4 Maximum likelihood estimation via E–M algorithm

Standard methods for likelihood maximization fail to give a solution in the present setup. Yet, the finite mixture structure of the model allows to treat the estimation issue as an incomplete data problem, thus the E–M algorithm can be suitably adapted (see [9]). In particular, given n independent observations from (1) $\mathbf{x}_j$, $j = 1,\dots,n$, the complete–data vector $\mathbf{x}_c$ is given by:

$$\mathbf{x}_c = (\mathbf{x},\mathbf{z}) = (\mathbf{x}_1,\dots,\mathbf{x}_n,\mathbf{z}_1,\dots,\mathbf{z}_n) \tag{7}$$

where the component–label D-dimensional vectors $\mathbf{z}_1,\dots,\mathbf{z}_n$ represent the missing data, z_{ji} being 1 or 0 according to whether $\mathbf{x}_j$ has arisen from the i^{th} component of the mixture model or not $(j = 1,\dots,n;\ i = 1,\dots,D)$.

The true log–likelihood can be thought of as originated from the following complete log–likelihood:

$$\log L_c(\boldsymbol{\theta}) = \sum_{j=1}^{n}\sum_{i=1}^{D} z_{ji}\left\{\log p_i + \log f_{\mathcal{D}}(\mathbf{x}_j;\boldsymbol{\alpha}+\tau\boldsymbol{e}_i)\right\} \tag{8}$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha},\boldsymbol{p},\tau)$ and

$$f_{\mathcal{D}}(\mathbf{x}_j;\boldsymbol{\alpha}+\tau\mathbf{e}_i) = \frac{\Gamma(\alpha_+ + \tau)\Gamma(\alpha_i)}{\Gamma(\alpha_i + \tau)}x_{ji}^{\tau}\prod_{h=1}^{D}\frac{x_{jh}^{\alpha_h-1}}{\Gamma(\alpha_h)}$$

is the density of a Dirichlet with parameter $(\boldsymbol{\alpha} + \tau \mathbf{e}_i)$.

The $k+1$ step of the E–M algorithm can be described as follows.
E–step given the current parameter estimates $\boldsymbol{\theta}^{(k)} = (\boldsymbol{\alpha}^{(k)}, \mathbf{p}^{(k)}, \tau^{(k)})$, calculate the conditional expectation of the complete log-likelihood (8) given $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ as:

$$\sum_{i=1}^{D} \sum_{j=1}^{n} t_i \left(\mathbf{x}_j ; \boldsymbol{\theta}^{(k)} \right) \left\{ \log p_i^{(k)} + \log f_i \left(\mathbf{x}_j ; \boldsymbol{\alpha}^{(k)} + \tau^{(k)} \mathbf{e}_i \right) \right\} \tag{9}$$

where

$$t_i \left(\mathbf{x}_j ; \boldsymbol{\theta}^{(k)} \right) \propto p_i^{(k)} f_{\mathcal{D}} \left(\mathbf{x}_j ; \boldsymbol{\alpha}^{(k)} + \tau^{(k)} \mathbf{e}_i \right) = p_i^{(k)} \frac{\frac{\Gamma(\alpha_i^{(k)})}{\Gamma(\alpha_+^{(k)} + \tau^{(k)})} x_{ji}^{\tau^{(k)}}}{\sum_{h=1}^{D} p_h^{(k)} \frac{\Gamma(\alpha_h^{(k)})}{\Gamma(\alpha_h^{(k)} + \tau^{(k)})} x_{jh}^{\tau^{(k)}}}$$

represents the "posterior" probability that $\mathbf{x}_j$ belongs to the i^{th} component of the mixture.
M–step maximize (9) to obtain the maximum likelihood estimates of the parameters. In particular, we have $\hat{p}_i^{(k+1)} = \frac{1}{n} \sum_{j=1}^{n} t_i \left(\mathbf{x}_j ; \boldsymbol{\theta}^{(k)} \right)$, $(i = 1, \ldots, D-1)$, whereas $\hat{\boldsymbol{\alpha}}^{(k+1)}$ and $\hat{\tau}^{(k+1)}$ can be computed by implementing a Newton–Raphson scheme.

The choice of the starting values for the E–M algorithm is a critical one. We devised a solution which takes advantage of the particular relationships among the parameters of the mixture components. In particular, we applied the k-means clustering algorithm to obtain the initial value $p_i^{(0)}$ as proportion of observations belonging to the i^{th} group $(i = 1, \ldots, D)$. To allocate the groups to the components of the FD an ad hoc algorithm has been developed. This exploits the fact that the i^{th} component of the FD is Dirichlet distributed with parameter $\boldsymbol{\alpha}_i = \boldsymbol{\alpha} + \tau \mathbf{e}_i$. Therefore, it is characterized by an i^{th} element with larger expected value than the i^{th} elements of the other components.

Once the groups are labeled, the Newton Raphson method allows to obtain the estimates of $\boldsymbol{\alpha}_i$ for each component. Estimators of the parameter τ can then be obtained by taking the absolute values of the differences between the components in the i^{th} and the h^{th} position of the estimated $\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_h$ $(i, h = 1, \ldots, D; i \neq h)$. The median of all such estimates is a robust initial value $\tau^{(0)}$ for τ.

Finally, to initialize the vector $\boldsymbol{\alpha}$, given $\mathbf{p}^{(0)}$ and $\tau^{(0)}$, we implemented a robust method of moments. More precisely, given the expressions (3) and (4) of the first two moments, first we estimated α_+ as the median of all estimates

deriving from the second moments, i.e.:

$$\begin{aligned}\alpha_{+}^{(0)} =&\text{median}_i\Big\{\frac{-m_{1i}+m_{1i}^2-2m_{1i}^2\tau+2m_{2i}\tau}{2\left(m_{1i}^2-m_{2i}\right)}\cdot\\ &-\Big(m_{1i}^2-2m_{1i}m_{2i}+m_{2i}^2-4m_{1i}^2p_i\tau^2+\\ &+4m_{2i}p_i\tau^2-4m_{2i}p_i^2\tau^2+4m_{1i}^2p_i^2\tau^2\Big)^{\frac{1}{2}}\Big\}\end{aligned}$$

where $m_{1i}=\frac{1}{n}\sum_{j=1}^{n}x_{ji}$ and $m_{2i}=\frac{1}{n}\sum_{j=1}^{n}x_{ji}^2$, $(i=1,\ldots,D)$. Then, we estimated the α_i's on the basis of the first moments, i.e.:

$$\alpha_i^{(0)}=\left(\alpha_{+}^{(0)}+\tau^{(0)}\right)m_{1i}-p_i^{(0)}\tau^{(0)}.$$

5 Variance-covariance matrix of MLE

The asymptotic variance-covariance matrix of the MLE $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ can be approximated by the inverse of the observed information matrix $\mathbf{I}(\hat{\boldsymbol{\theta}};\boldsymbol{x})$. However, the direct evaluation of the second-order derivatives of the (incomplete) log-likelihood derived from (1) is excessively cumbersome. Some methods for approximating the observed information matrix such as the empirical information-based method (see [10]) or the supplemented E–M algorithm proposed by [11] are feasible but will often underestimate the standard errors (see [5]).

A further approximation of the variance-covariance matrix of the MLE can be achieved via parametric bootstrap. Once the MLE $\hat{\boldsymbol{\theta}}$ is obtained, B' bootstrap samples are drawn from $FD^D(\hat{\boldsymbol{\theta}})$ and the E–M algorithm is applied to each sample. This allows to obtain the (Monte Carlo) distribution of the MLE and, therefore, the sample variance-covariance matrix. Some simulations have been performed showing that the parametric bootstrap produces satisfactory results. However, its implementation is extremely demanding in terms of computational burden as each replication requires an E–M cycle.

An interesting indirect but exact evaluation of $\mathbf{I}(\hat{\boldsymbol{\theta}};\boldsymbol{x})$ can be obtained via complete–data log-likelihood (see [8]) as the following equality holds:

$$\mathbf{I}\left(\hat{\boldsymbol{\theta}};\mathbf{x}\right)=\left[\mathrm{E}_{\boldsymbol{\theta}}\left\{\mathbf{I}_c\left(\boldsymbol{\theta};\mathbf{X}_c\right)|\mathbf{x}\right\}\right]_{\theta=\hat{\theta}}-\left[\mathrm{E}_{\boldsymbol{\theta}}\left\{\mathbf{S}_c\left(\boldsymbol{\theta};\mathbf{X}_c\right)\mathbf{S}_c^T\left(\boldsymbol{\theta};\mathbf{X}_c\right)|\mathbf{x}\right\}\right]_{\theta=\hat{\theta}} \quad (10)$$

where $\mathbf{S}_c\left(\boldsymbol{\theta};\mathbf{X}_c\right)$ denotes the complete–data score statistics, $\mathbf{I}_c\left(\boldsymbol{\theta};\mathbf{X}_c\right)$ is the negative of the Hessian of the complete–data log likelihood (8) and $\mathrm{E}_{\boldsymbol{\theta}}[\cdot|\boldsymbol{x}]$ denotes the conditional expectation, given the observed data $\mathbf{x}$, using the parameter vector $\boldsymbol{\theta}$.

The elements of the score statistic $\mathbf{S}_c(\boldsymbol{\theta}; \mathbf{X}_c)$ can be easily derived from (8):

$$\begin{aligned}
\frac{\partial \log L_c(\boldsymbol{\theta})}{\partial p_i} =& \frac{z_{.i}}{p_i} - \frac{z_{.D}}{p_D} \qquad (i = 1, \ldots, D-1) \\
\frac{\partial \log L_c(\boldsymbol{\theta})}{\partial \alpha_i} =& n\psi(\alpha_+ + \tau) - z_{.i}\left[\psi(\alpha_i + \tau) - \psi(\alpha_i)\right] - n\psi(\alpha_i) + \sum_{j=1}^{n} \log x_{ji} \\
& (i = 1, \ldots, D) \\
\frac{\partial \log L_c(\boldsymbol{\theta})}{\partial \tau} =& n\psi(\alpha_+ + \tau) - \sum_{i=1}^{D} z_{.i}\psi(\alpha_i + \tau) + \sum_{i=1}^{D}\sum_{j=1}^{n} z_{ji} \log x_{ji}
\end{aligned} \tag{11}$$

where $z_{.i} = \sum_{j=1}^{n} z_{ji}$, $(i = 1, \ldots, D)$ and $\psi(\cdot)$ denotes the digamma function.

The elements of the $2D \times 2D$ matrix $\mathbf{I}_c(\boldsymbol{\theta}; \mathbf{X}_c)$ assume the following form:

$$\begin{aligned}
\frac{\partial^2 \log L_c(\boldsymbol{\theta})}{\partial p_i^2} =& \frac{z_{.i}}{p_i^2} + \frac{z_{.D}}{p_D^2} \quad (i = 1, \ldots, D-1) \\
\frac{\partial^2 \log L_c(\boldsymbol{\theta})}{\partial p_i \partial p_h} =& \frac{z_{.D}}{p_D^2} \quad (i \neq h; \;\; i, h = 1, \ldots, D-1) \\
\frac{\partial^2 \log L_c(\boldsymbol{\theta})}{\partial p_i \partial \alpha_h} =& 0 \quad (i = 1, \ldots, D-1) \quad (h = 1, \ldots, D) \\
\frac{\partial^2 \log L_c(\boldsymbol{\theta})}{\partial p_i \partial \tau} =& 0 \quad (i = 1, \ldots, D-1) \\
\frac{\partial^2 \log L_c(\boldsymbol{\theta})}{\partial \alpha_i^2} =& -n\psi'(\alpha_+ + \tau) + z_{.i}\left[\psi'(\alpha_i + \tau) - \psi'(\alpha_i)\right] + n\psi'(\alpha_i) \\
& (i = 1, \ldots, D) \\
\frac{\partial^2 \log L_c(\boldsymbol{\theta})}{\partial \alpha_i \partial \alpha_h} =& -n\psi'(\alpha_+ + \tau) \quad (i \neq h; \;\; i, h = 1, \ldots, D) \\
\frac{\partial^2 \log L_c(\boldsymbol{\theta})}{\partial \alpha_i \partial \tau} =& -n\psi'(\alpha_+ + \tau) + z_{.i}\psi'(\alpha_i + \tau) \quad (i = 1, \ldots, D) \\
\frac{\partial^2 \log L_c(\boldsymbol{\theta})}{\partial \tau^2} =& -n\psi'(\alpha_+ + \tau) + \sum_{i=1}^{D} z_{.i}\psi'(\alpha_i + \tau)
\end{aligned} \tag{12}$$

where $\psi'(\cdot)$ denotes the trigamma function.

The calculation of the conditional expectations required by (10) is feasible but very long winded. The key point is to notice that, conditionally on $\mathbf{x}$, the component-label random vectors $\mathbf{Z}_j$ $(j = 1, \ldots, n)$ are independently distributed as multinomials. In particular:

$$\mathbf{Z}_j | \mathbf{x} = (Z_{j1}, \ldots, Z_{jD}) \, | \mathbf{x} \sim Mu(1, \mathbf{p}_j^*)$$

where $\mathbf{p}_j^* = (p_{j1}^*, \ldots, p_{jD}^*)$ with

$$p_{ji}^* = p_i \frac{\frac{\Gamma(\alpha_i)}{\Gamma(\alpha_+ + \tau)} x_{ji}^\tau}{\sum_{h=1}^D p_h \frac{\Gamma(\alpha_h)}{\Gamma(\alpha_h + \tau)} x_{jh}^\tau} \qquad i = 1, \ldots, D. \tag{13}$$

Here we shall propose and study a simpler but equally accurate evaluation of the conditional expectation in (10) based on conditional bootstrap ([4]). This is obtained by averaging over B independent bootstrap samples $(\mathbf{z}_{1b}^*, \ldots, \mathbf{z}_{nb}^*)$, $(b = 1, \ldots, B)$ where $\mathbf{z}_{jb}^*$ $(j = 1, \ldots, n)$ are independent draws from $\mathbf{Z}_{jb}^* \sim Mu(1, \hat{\mathbf{p}}_j^*)$ and the elements of $\hat{\mathbf{p}}_j^*$ are given by (13) with $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

6 Simulation results

In this section we present the results of a simulation study of the above conditional bootstrap evaluation of the asymptotic MLE variance-covariance matrix. We implemented $K = 1,000$ replications of the conditional bootstrap, each one based on $B = 3,000$ independent bootstrap samples. For each replication we computed the parameter estimates, the standard error estimates derived from the conditional bootstrap and the confidence interval based on the (asymptotic) normal distribution of the estimators.

Several parameters configurations and sample sizes have been investigated. For space constraints we report only two parameters configurations (Table 1 and 2 respectively) and three sample sizes ($n = 100; 500; 1,000$). In each table, rows 1 and 2 report the mean and the standard deviation of the Monte Carlo (MC) distribution based on $K = 1,000$ replications of the MLE estimators of $\boldsymbol{\theta}$. Row 3 reports the mean of the MC distribution of the conditional bootstrap based standard error estimators and row 4 reports the Absolute Relative Bias, i.e. the mean of the absolute deviations between such estimates of the standard errors and the MC standard deviation (row 2) divided by this last quantity. Finally, row 5 gives the simulated confidence levels against a 95% nominal one.

The performance of the MLE estimators in terms of mean and standard deviation appear rather satisfactory. Furthermore, from a graphical inspection of their simulated distributions no evident deviations from normality appear already for $n = 100$. The only exception is the estimator of τ which displays a skewed distribution for small sample sizes.

The standard error estimates provided by the proposed conditional bootstrap approach do also exhibit a satisfactory behavior and are not computationally very demanding. They are approximately correct and generally display low absolute relative error, although in some cases less accurate results are obtained for the parameter τ. Such results require further investigation and might be improved by adopting a suitable re-parametrization.

Finally, the good performance of the simulated confidence levels for all models and even for small sample sizes confirms the accuracy of the proposed method.

Table 1. Simulation results with $\mathbf{X} \sim FD^3(\boldsymbol{\alpha} = (1,1,1), \mathbf{p} = (1,1,1)/3, \tau = 10)$.

		p_1	p_2	α_1	α_2	α_3	τ
$n = 100$	1. MLE mean	0.334	0.332	1.023	1.027	1.027	10.336
	2. MLE s.d.	0.050	0.049	0.112	0.116	0.115	0.985
	3. S.E. mean	0.047	0.047	0.122	0.122	0.122	1.267
	4. ARB	0.068	0.044	0.116	0.100	0.102	0.286
	5. CI coverage	0.940	0.946	0.968	0.966	0.970	0.986
$n = 500$	1. MLE mean	0.333	0.334	1.010	1.008	1.011	10.141
	2. MLE s.d.	0.021	0.022	0.052	0.050	0.053	0.457
	3. S.E. mean	0.021	0.021	0.054	0.054	0.054	0.564
	4. ARB	0.015	0.024	0.053	0.083	0.048	0.233
	5. CI coverage	0.949	0.936	0.954	0.971	0.953	0.976
$n = 1,000$	1. MLE mean	0.333	0.334	1.005	1.005	1.005	10.067
	2. MLE s.d.	0.015	0.015	0.037	0.037	0.037	0.340
	3. S.E. mean	0.015	0.015	0.038	0.038	0.038	0.389
	4. ARB	0.011	0.031	0.038	0.037	0.033	0.146
	5. CI coverage	0.946	0.936	0.948	0.954	0.951	0.964

Table 2. Simulation results with $\mathbf{X} \sim FD^3(\boldsymbol{\alpha} = (1,2,5), \mathbf{p} = (0.3,0.6,0.1), \tau = 3)$.

		p_1	p_2	α_1	α_2	α_3	τ
$n = 100$	1. MLE mean	0.305	0.590	1.028	2.112	5.201	3.161
	2. MLE s.d.	0.063	0.078	0.138	0.368	0.664	0.562
	3. S.E. mean	0.063	0.086	0.143	0.407	0.731	0.676
	4. ARB	0.130	0.273	0.138	0.228	0.149	0.237
	5. CI coverage	0.938	0.931	0.954	0.959	0.964	0.962
$n = 500$	1. MLE mean	0.302	0.594	1.011	2.045	5.087	3.077
	2. MLE s.d.	0.027	0.035	0.058	0.160	0.280	0.257
	3. S.E. mean	0.028	0.035	0.062	0.168	0.318	0.291
	4. ARB	0.058	0.086	0.073	0.084	0.138	0.133
	5. CI coverage	0.959	0.943	0.956	0.955	0.970	0.973
$n = 1000$	1. MLE mean	0.301	0.596	1.008	2.032	5.064	3.057
	2. MLE s.d.	0.020	0.025	0.043	0.115	0.225	0.203
	3. S.E. mean	0.020	0.025	0.043	0.118	0.225	0.206
	4. ARB	0.043	0.068	0.038	0.061	0.034	0.041
	5. CI coverage	0.950	0.945	0.956	0.949	0.941	0.942

7 An application

The applicative potential of the FD can be better highlighted through a real data set analysis. To this end we shall focus on the composition of glacial tills ([1], p. 376) sampled in the Mexico and Kasoag quadrangles (New York) (see [7] for a petrological interpretation of data). The variables involved are the proportions of four lithologic groups, i.e. red sandstone, gray sandstone, crystalline and miscellaneous of 92 sampled tills. The last two components present a relevant number of zero values. This can be dealt with by amalgamating such components as they both present very low values and can be interpreted as a sort of residual category with respect to the first two. Yet, since 6 out of 92 tills still have the last component null, such sampling units have been neglected. This does not substantially alter the results of the analysis as such units follows a pattern not significantly different from the remaining ones, as it is confirmed, for example, by computing the first two moments with and without them.

Some descriptive summaries are reported in Table 4 (means, variances and correlation coefficients). These latter quantities highlight a negative (significant) correlation in the first two cases and a positive (but not significant) one for the pair (X_2, X_3).

In a comparative perspective, three models have been considered: the FD, the Dirichlet and the additive logistic normal (ALN) which is one of the most frequently employed. The latter model is obtained by assigning a multivariate Normal distribution to the variables $Y_i = \log(X_i/X_j)$ $(i \neq j)$ (see [1]). For such model the variable X_j to be used as denominator (fill-up value) has been chosen in all three possible ways giving similar results. For space reasons we present the case $X_j = X_3$.

The implementation in R language of the adaptation of the E–M algorithm described in Section 4 allowed to find the maximum likelihood estimates of the FD parameters. They are reported in Table 3 together with the conditional bootstrap based standard error estimates and the 95% level confidence intervals.

Table 3. Estimates of parameters, standard errors and confidence intervals for the FD model

	$\hat{p}_1$	$\hat{p}_2$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\tau}$
MLE	0.664	0.336	2.885	2.291	0.695	7.766
SE	0.053	0.053	0.424	0.290	0.076	1.036
CI	[0.56, 0.77]	[0.23, 0.44]	[2.06, 3.72]	[1.72, 2.86]	[0.54, 0.83]	[5.76, 9.82]

The estimates of the Dirichlet model parameters are $\hat{\boldsymbol{\alpha}} = (1.841, 1.140, 0.374)$. The ALN parameters have been estimated through sample means and variance-

covariance matrix of the log-ratios, i.e. $\hat{\mu} = (3.073, 2.391)$ and

$$\hat{\Sigma} = \begin{bmatrix} 1.842 & 0.301 \\ 0.301 & 1.699 \end{bmatrix}.$$

The contours of the three estimated models are shown in Figure 1 (top right, bottom left and bottom right panels) together with a plot of the pair (X_1, X_3) (top left panel). The other pairs obviously convey the same information and fit given the unit-sum constraint, but the (X_1, X_3) plot is more readable. Notice that, unlike the Dirichlet, the ALN and the FD pro-

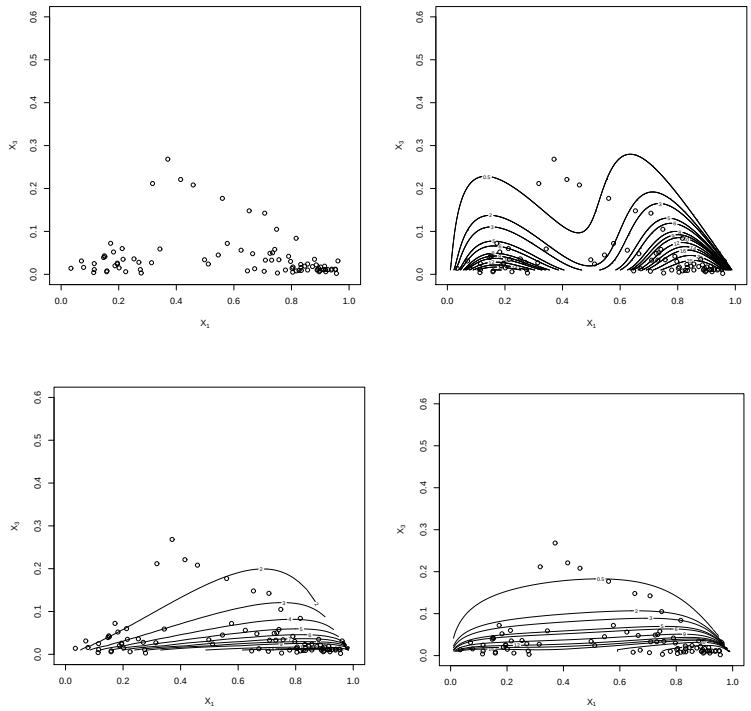


Fig. 1. Plot of X_1 and X_3 (top left) and estimated contours for FD model (top right), Dirichlet model (bottom left) and additive logistic normal (bottom right)

duce a fairly accurate fit. In particular, it is remarkable the ability of the latter to identify the presence of two clusters, characterized by high values of X_1 and X_2 respectively, whose presence is well justifiable from a geologic/glaciological point of view. Indeed, the two predominant components (red and gray sandstone) characterize a particular geographical area each, but

they have been "contaminated" with the remaining lithologic groups because of ice movement.

None of the models captures the few observations with relative high values of X_3. Such observations cannot apparently be given an uncontroversial geological explanation and deserves further investigation.

The estimates of the moments under the different models admit an explicit expression for the FD and the Dirichlet, vice versa they must be approximated via numerical integration for the ALN. Their values are reported in in Table 4 and, once again, highlight a better behavior of the FD and of the ALN models. In particular, it is evident the poor performance of the Dirichlet caused by its inadequate dependence structure.

Table 4. Sample and estimated means, variances and correlation coefficients

	Sample	Dirichlet	FD	ALN
$\mathrm{E}(X_1)$	0.595	0.549	0.590	0.589
$\mathrm{E}(X_2)$	0.366	0.340	0.359	0.373
$\mathrm{E}(X_3)$	0.039	0.111	0.051	0.038
$\mathrm{Var}(X_1)$	0.092	0.057	0.084	0.077
$\mathrm{Var}(X_2)$	0.089	0.052	0.083	0.074
$\mathrm{Var}(X_3)$	0.003	0.023	0.003	0.002
ρ_{X_1,X_2}	-0.985	-0.791	-0.980	-0.987
ρ_{X_1,X_3}	-0.177	-0.391	-0.122	-0.205
ρ_{X_2,X_3}	0.005	-0.254	-0.075	0.033

Interesting considerations can been drawn from the analysis of the one-dimensional marginal plots as well. Two of them (namely X_1 and X_2) clearly suggest the presence of bimodality, which is captured fairly well by the FD model in contrast with both the Dirichlet and the ALN ones. On the contrary the unimodal histogram of X_3 is well approximated by all three models (see Figure 2).

The study of the neutrality concept shows interesting features in the present context. Recall that a component X_i is said neutral if it is independent of the subcomposition $S_{jk} = X_j/(X_j + X_k)$ $(i \neq j \neq k)$. The correlation coefficients between each X_i and the subcomposition S_{jk} suggest neutrality of X_3 (correlation equal to -0.079) and non neutrality for the other two variables (correlations equal to -0.350 and 0.495).

The Dirichlet is unable to model non neutrality while the ALN does not allow to incorporate neutrality. On the contrary, the FD allows for both concepts to be contemplated and modeled. The plots of each pair X_i and S_{jk} are reported in Figure 3 together with the estimates of the FD conditional mean (6) and, as a benchmark, a Normal kernel nonparametric regression. Such conditional means are quite consistent with the pattern of the data. In particular, the FD captures neutrality of X_3, meaning that it is irrelevant

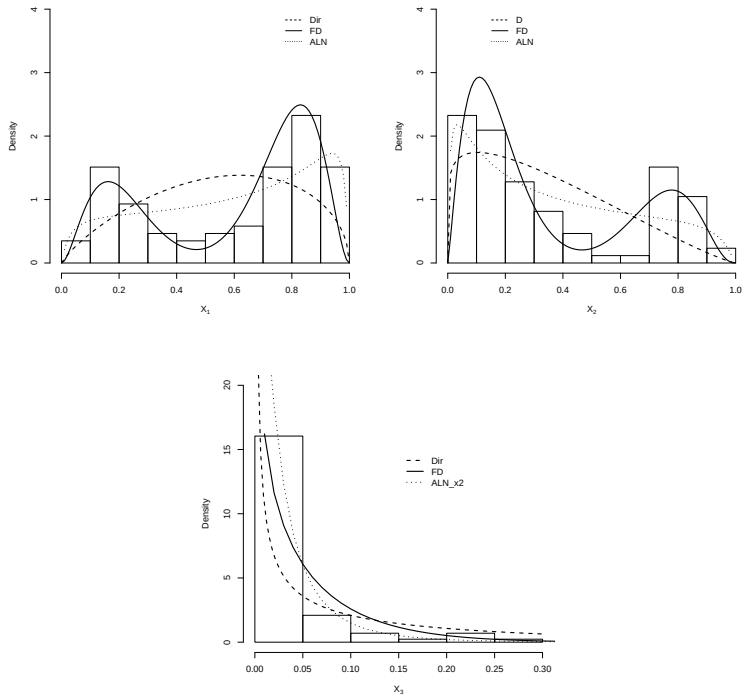


Fig. 2. Histogram and estimated densities of X_1 (top left), X_2 (top right) and X_3 (bottom)

in explaining the relative importance of X_1 and X_2. On the other hand, increasing values of X_1 (respectively of X_2) are associated with a larger relative weight of X_3 with respect to X_2 (X_1).

In conclusion, the FD and the ALN models exhibit a substantially better fit than the Dirichlet. The FD appears to be more helpful in data interpretation as it recognizes bi-modality and it gives a satisfactory picture of the dependencies between subcompositions and residual variables (neutrality/non neutrality).

References

1. Aitchison, J., *The Statistical Analysis of Compositional Data*, The Blackburn Press, London (2003).
2. Barndorff-Nielsen, O.E., Jorgensen, B., "Some parametric models on the simplex", *Journal of Multivariate Analysis* 39, 106–116 (1991).

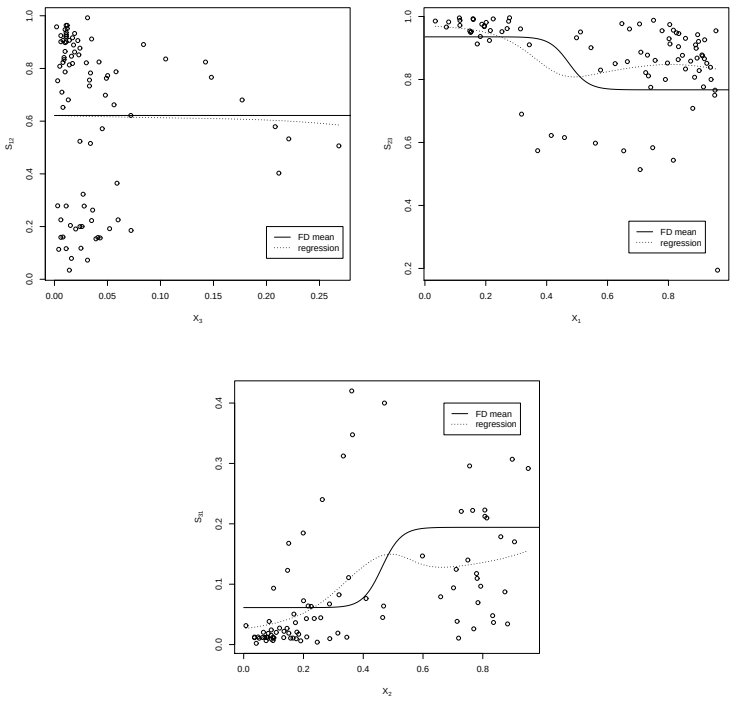


Fig. 3. Conditional mean and nonparametric regression for subcomposition S_{12} (top left), S_{23} (top right) and S_{31} (bottom)

3. Connor, J.R., Mosimann, J.E., "Concepts of independence for proportions with a generalization of the Dirichlet distribution", *Journal of the American Statistical Association* 64, 194–206 (1969).
4. Diebolt, J., Ip, E.H.S., "Stochastic EM: method and application", in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson, and D.J. Spiegelhalter (Eds.), London: Chapman & Hall, pp. 259–273 (1996).
5. Efron, B., "Missing data, imputation and the bootstrap", *Journal of the American Statistical Association* 89, 463-479 (1994).
6. Gupta, R.D., Richards, D.St.P., "The history of the Dirichlet and Liouville distributions", *International Statistical Review* 69, 433–446 (2001).
7. Kaiser, R.F., "Composition and Origin of Glacial Till, Mexico and Kasoag Quadrangles, New York", *Journal of Sedimentary Research* 32, 502-513 (1962).
8. Louis, T.A., "Finding the observed information matrix when using the EM algorithm", *Journal of the Royal Statistical Society B* 44, 226-233 (1982).
9. McLachlan, G., Peel, D., *Finite mixture models*, John Wiley & Sons, New York (2000).
10. Meilijson, I., "A fast improvement to the EM algorithm on its own terms", *Journal of the Royal Statistical Society B* 51, 127-138 (1989).

11. Meng, X.L., Rubin, D.B., "Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm", *Journal of the Americal Statistical Association* 86, 899-909 (1991).
12. Ongaro, A., Migliorati, S., Monti, G.S., "A new distribution on the simplex containing the Dirichlet family", *Proceedings of CODAWORK'08, The 3rd Compositional Data Analysis Workshop, University of Girona, Spain* (2008).
13. Rayens, W.S., Srinivasan, C., "Dependence properties of generalized Liouville distributions on the simplex", *Journal of the American Statistical Association* 89, 1465–1470 (1994).
14. Smith, B., Rayens, W.S., "Conditional generalized Liouville distributions on the simplex", *Statistics* 36, 185–194 (2002).

Descriptive tools for Bayesian Multidimensional Item Response Theory Models

Alvaro Montenegro[1]

Universidad Nacional de Colombia, Sede Bogotá, Departamento de Estadística
(e-mail: ammontenegrod@unal.edu.co)

Abstract. In this paper, we discuss some tools to explore and confirm the dimension of a class of multidimensional item response theory models, called linear latent structure (LSMIRT). It is shown that, the augmented variables used in a data augmented Gibbs sampler (DAGS) algorithm to estimate the parameters of the model, can be interpreted as latent variables that govern the response process, so, they can be used to estimate tetrachoric correlations matrix of the items.
Keywords: LSMIRT model, item response theory, principal component analysis, factor analysis, tetrachoric correlations.

1 Introduction

The LSMIRT models were proposed by Montenegro[6], to be used in large-scale assessment tests designed explicitly to measure more than one latent trait. Those tests, are usually split into subtests where each subtest is designed to measure mainly a unique unidimensional latent trait, called *main latent trait*. Admission tests of some universities are typical examples of such type of tests. The following are the assumptions of the LSMIRT model.

1. The test is split into m subtests. It is assumed that each subtest is essentially unidimensional. Hence, each subtest attempt to measure only one main latent trait. Each subtest has K_v items, so the test has $K = K_1 + K_2 + \cdots K_m$ items.
2. It is assumed that the *basic latent traits* of the examinees correspond to a random sample drawn from a normal multivariate distribution. $N_d(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a correlation matrix, and $d \leq m$.
3. The main latent traits of the examinees being measured by each subtest are composites (linear combinations) of basic latent trait vectors.
4. The link function used to link the linear predictor with the probability of a correct response is the standard normal ogive, denoted $\boldsymbol{\Phi}(\cdot)$

It is assumed that, the basic latent traits are vectors of an Euclidean space of dimension d, where $d \leq m$. This space is called the *latent trait space*. The jth item of subtest v, will be called item vj. The LSMIRT model is specified by the probability of success of examinee i to item vj given by

$$P(Y_{vij} = 1|\alpha_{vj}, \gamma_{vj}, \boldsymbol{\beta}_v, \boldsymbol{\theta_i}) = \Phi(\alpha_{vj}\boldsymbol{\beta}_v^t\boldsymbol{\theta_i} - \gamma_{vj}), \quad (1)$$

where α_{vj} and γ_{vj} will be called respectively the slope (the discrimination) parameter and the intercept parameter of item vj. Vector $\boldsymbol{\beta}_v = (\beta_{v1}, \cdots, \beta_{vd})^t$ is a unitary vector in the latent trait space called the direction of subtest v and $\boldsymbol{\theta}_i = (\theta_{i1}, \cdots, \theta_{id})^t$ represents the vector of the basic latent traits of examinee i.

There exist a parametrization of the LSMIRT model in which d of the main latent traits can be identified with coordinate axes of the latent traits space. The direction vectors $\boldsymbol{\beta}_v$ of the main latent traits identified with coordinate axes are called *main directions* of the test. Consequently, before to intent fit a LSMIRT model to the data the following two questions must be solved.

1. What is the dimension of the latent trait space?
2. Which are the main latent traits that may be identified with the coordinate axis of the latent trait space?

2 Data description

The data used along this paper are from the admission test in the Universidad Nacional de Colombia, applied for the second semester of 2009. The sample size was N=5096. The test size was K = 113 with 5 subtests. The subtests were: textual analysis (Textual) with $K_1 = 15$ items, mathematics (Math) with $K_2 = 26$ items, natural sciences (Science) with $K_3 = 29$ items, social sciences (Social) with $K_4 = 29$ and image analysis (Image) with $K_5 = 14$ items.

3 Detecting the dimension of the latent trait space

To obtain an identifiable LSMIRT model, the main directions are associate with the coordinate axes of the latent trait space. Intuitively, the main directions represent the main constructs that are measured by the test. The other constructs can be represented by some directions in the latent trait space. The number of main directions of the test is defined as the dimension of the latent trait space. So, if we detect the dimension of the latent trait space, the number of main directions is determined. Different techniques and softwares can be used to determine the dimension of the latent trait space. In this section, we describe three known approaches.

3.1 DETECT index

For the case where clusters of items that define subtests are not available, Zhang and Stout [10] proposed the DETECT index, to detect the dimension

of the latent trait space. They implement a procedure also called DETECT to find the clusters of homogeneous items. The number of clusters is proposed as the dimension of the latent trait space. Hence, in this case an approximated simple structure model is obtained. In the tests where the LSMIRT model can be useful, each item is associated to a unique subtest. Thus, each subtest predefines a cluster of items.

3.2 Tetrachoric correlations

A second approach to determine the number of dimensions is based on the first eigenvalues of the tetrachoric correlations. The concept of tetrachoric correlations was introduced by Pearson[7]. The tetrachoric correlation between two dichotomous items estimates the Pearson correlation one would obtain if the two constructs were measured continuously, Drasgow,[5], Olson[8]. The tetrachoric correlation for manifest variables Y_i and Y_j, which we denote by ρ^*_{ij}, is equal to the Pearson correlation between the corresponding latent continuous variables Z_i and Z_j. That is: $\rho^*_{ij} = \rho(Z_i, Z_j)$.

The tetrachoric correlations can be easily understood as the correlation between the latent sample variables $Z_j = (Z_{1j}, \cdots, Z_{Nj})^t$ associated with each item. These latent variables have been used along this work. Let $\boldsymbol{Z} = [Z_{ij}]_{N\times K}$, and consider the standardized matrix $\boldsymbol{Z}^*$, obtained from $\boldsymbol{Z}$ to have unit variances. Then, to determine the dimension of the latent trait space, the eigenvalues of the matrix $(\boldsymbol{Z}^*)^t\boldsymbol{Z}^*$ are computed.

In practice, the variables Z_{ij} are not available. However, some computational methods have been developed to estimate the tetrachoric correlations. For example, a subroutine can be downloaded from the Applied Statistics section of StatLib, Brown [4]. A more recent function to estimate the tetrachoric correlations can be found in package *polycor* for **R** [**?**].

3.3 Principal Component Analysis

In this work we did a principal component analysis (PCA) of each of the subtests and of the full test. PCA was used only as a descriptive tool, so, there is no problem in that variables are binary. The PCA analysis were ran with the *FactorMineR* package of **R**. Firstly, each one of the subtests was fitted with a UIRT model. The estimations were ran with the *ltm*-package written for **R**. Secondly, a PCA was done with each one of the $N \times K_v$, $v = 1, \cdots 5$ binary tables of the subtests. Each PCA was ran with the *FactoMineR* package written for **R**. The correlation between the latent trait and the respective first principal component were greater than 0.99 in all the cases. Thirdly, A PCA was done with the $N \times K$ table of the full test. The univariate latent traits computed in the first step were used as quantitative supplementary variables.

Figure 1 shows the bar plots of first the eigenvalues of the standardized matrix obtained from $\boldsymbol{Y}^t\boldsymbol{Y}$ in the PCA of the full test. The plot and the

values support the decision to select $d = 3$ as the dimension of the latent trait space.

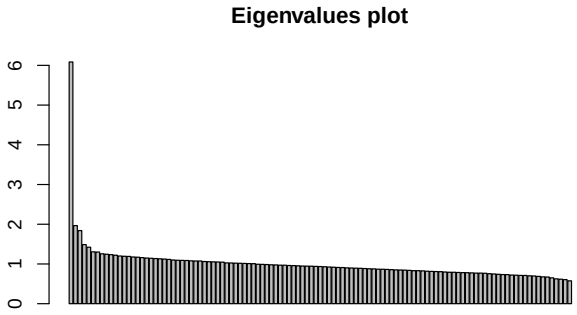


Fig. 1. Eigenvalues bar plots the PCA of the binary responses tables

Figure 2 shows the correlation circle of the variables in the plane 2,3. In the plot, the latent traits obtained from the UIRT model previously were projected as quantitative supplementary variables. Additionally, table 1 shows the square cosine between the latent traits and the coordinate axes. The plot and the table support the decision to select the reference directions of Math, Textual and Image subtests as the main directions.

Subtest	Dim. 1	Dim. 2	Dim. 3
Math	0,5953	0,0896	0,0895
Science	0,5691	0,0271	0,0036
Social	0,5155	0,0078	0,0490
Textual	0,5010	0,0060	0,1761
Image	0,4179	0,4524	0,0773

Table 1. Square cosine of the subtest latent traits with respect to the first three axes in the PCA. The latent traits were projected as supplementary variables

4 Confirmatory Analysis

Before the modern developments of the IRT, the terachoric correlations were used as the basis of factor analysis of binary data. However, some problems arose with this strategy. First, the computation of the tetrachoric correlations is problematic in the cases of missing, data. Second, the estimated matrix of

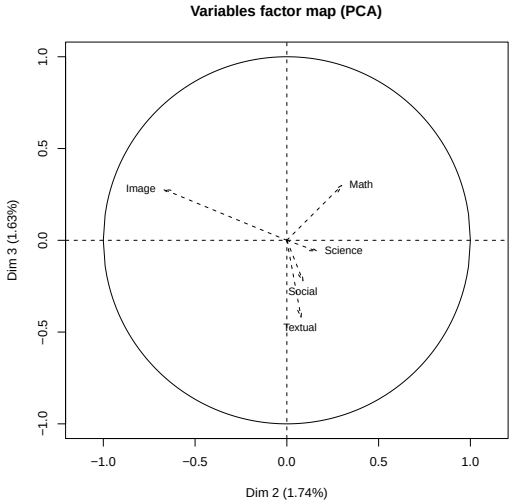


Fig. 2. Plot of the PCA with the subtest latent traits projected as supplementary variables.

the sample tetrachoric correlation obtained from the classical algorithms is often non positive definite, Bock, *et al.*[3].

Takane, and J. Leeuw[9], formally proved that marginal likelihood of the two-parameter normal ogive model in the IRT and factor analysis of dichotomized variables are equivalent. Then, an alternative way to investigate the dimension of the latent trait space is to find directly the first eigenvalues of the standardized matrix obtained from $\boldsymbol{Y}^t\boldsymbol{Y}$. The item factor analysis of the LSMIRT model is stated as follows. Let Z_{vj} be the underline latent variable that governs the response process for item vj. According to Montenegro[6], let $\boldsymbol{\alpha}$ be the matrix given by

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & 0 & \cdots & 0 \\ 0 & \alpha_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \alpha_m \end{bmatrix}_{K\times m}, \tag{2}$$

where $\boldsymbol{\alpha_1} = (\alpha_{11}, \alpha_{12}, \cdots, \alpha_{1K_1})^t, \cdots \boldsymbol{\alpha_m} = (\alpha_{m1}, \alpha_{m2}, \cdots, \alpha_{mK_m})^t$. Let $\boldsymbol{\beta}$ the matrix whose rows are the direction vectors $\boldsymbol{\beta}_v$, $v = 1, \cdots, m$. Let $\boldsymbol{A}$ be the matrix defined as $\boldsymbol{A} = \boldsymbol{\alpha} \times \boldsymbol{\beta}$. By facility, the j-th rows of $\boldsymbol{A}$ are denoted $\boldsymbol{a}_j^t$. Let $\boldsymbol{\theta}$ be a random vector distribute as $N_d(\boldsymbol{0}, \boldsymbol{\Sigma})$. The latent traits of examines are concrete values of vector $\boldsymbol{\theta}$. Let $\boldsymbol{e}$ be a random vector distributed as $N_K(\boldsymbol{0}, \boldsymbol{I}_K)$, where $\boldsymbol{I}_K$ represents the identity matrix of size K. It is assumed that $\boldsymbol{A\theta}$ and $\boldsymbol{e}$ are independent. Let $\boldsymbol{\gamma}$ be the vector of intercepts in the LSMIRT model. Let $\boldsymbol{Z}$ be the random vector defined as

$$\boldsymbol{Z} = \boldsymbol{A\theta} - \boldsymbol{\gamma} + \boldsymbol{e}. \tag{3}$$

Then,

$$\boldsymbol{Z} \sim N(-\boldsymbol{\gamma}, \boldsymbol{A\Sigma A}^t + \boldsymbol{I}_K) \tag{4}$$

and

$$\boldsymbol{Z}|\boldsymbol{A\theta} \sim N(\boldsymbol{A\theta} - \boldsymbol{\gamma}, \boldsymbol{I}_K). \tag{5}$$

Let $\boldsymbol{Z} = (Z_{11}, \cdots, Z_{mK_m})^t$, and let $\boldsymbol{y} = (y_{11}, \cdots, y_{mK_m})^t$ be the random vector representing any response pattern, then it is easy to see that

$$p\left\{[Z_{vj}|\boldsymbol{a}_{vj}^t\boldsymbol{\theta}] > 0\right\} = P(y_{vj} = 1|\boldsymbol{\theta}, \boldsymbol{\beta}_v, \alpha_{vj}, \gamma_{vj}). \tag{6}$$

We can conclude that the random latent variables Z_{vj} govern the response process of item vj. The tetrachoric correlation matrix is just the correlation matrix of $\boldsymbol{Z}$. On the other hand, if we define

$$Y_{vij} = \begin{cases} 1, & \text{if } Z_{vij} > 0 \\ 0, & \text{if } Z_{vij} \leq 0 \end{cases}, \tag{7}$$

it can be shown that $P(Y_{vij} = 1 \mid \boldsymbol{\theta}_i, \boldsymbol{\beta}_v, \boldsymbol{\xi}_{vj}) = \Phi(\eta_{vij})$. In the DAGS algorithms designed for item response theory models, the variables Z_{vij} are just the augmented variables, Albert[1], Montenegro[6]. In each step of the DAGS algorithm after convergence, the values Z_{vij} are samples from the posterior distribution of $Z_{vij}|y_{vij}$.

We have shown a common framework for the IRT, item factor analysis and the augmented variables technique used in the DAGS algorithm. Equation (3) is a classical factorial model, where $\boldsymbol{A}$ is the loading matrix, and $\boldsymbol{e}$ is the perturbation vector. In this case, the theoretical covariance matrix of $\boldsymbol{e}$ is the identity matrix. This covariance matrix could be estimated as follows. The specific perturbations are given by

$$e_{ij} = Z_{ij} - \boldsymbol{a}_j\boldsymbol{\theta_i} + \gamma_j. \tag{8}$$

These specific perturbations e_{ij} and their covariance matrix can be estimated from the posterior distribution of e_{ij}. The values e_{ij} are the Bayesian latent residuals defined by Albert and Chib[2]. If the model is well fitted the bar plot based on the eigenvalues of the correlation matrix of $\boldsymbol{Z}$ (the tetrachoric correlations) and the initial bar plot based on the correlation matrix of $\boldsymbol{y}$ must reveal the same dimension of the latent trait space. Additionally, the bar plot based on the eigenvalues of the correlation matrix of $\boldsymbol{e}$, will not show additional dimensions. That is, this bar plot must show only noise. Figure 3 shows the bar plots of the eigenvalues of $cor(Z)$ and $cor(e)$ for the real case, when data were fitted with the tridimensional latent trait.

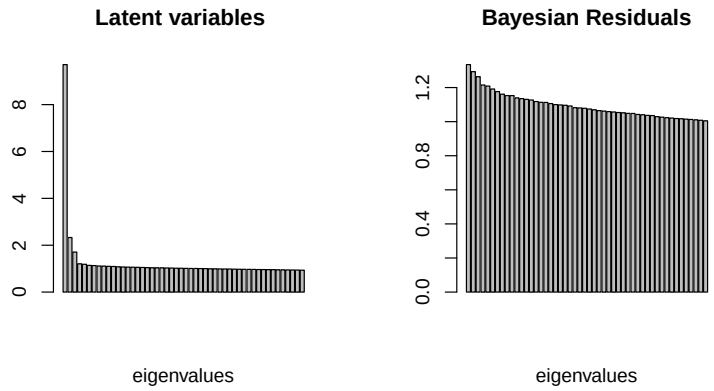


Fig. 3. Eigenvalues bar plot of the correlation matrices of $\boldsymbol{Z}$ and $\boldsymbol{e}$, modeled with three factors.

5 Conclusions

In this paper, it was shown that the dimension of the latent trait space can be initially determined form a PCA analysis of the data matrix $\boldsymbol{y}$. The unidimensional latent traits obtained from an unidimensional IRT modeling of the subtests, were used as supplementary variables, to determine the main directions. Furthermore, it was shown that the LSMIRT model is equivalent to a factor analysis of binary variables, where the latent continuous variables that govern the response process of the items are the augmented variables used in the DAGS algorithms to estimate the parameters in item response theory models. The latent traits are the factors, the matrix $\boldsymbol{A}$ is the loading matrix and the perturbation vector coincides with the theoretical Bayesian latent errors. These facts suggest to use the correlation matrices of $\boldsymbol{Z}$ and $\boldsymbol{e}$ as confirmatory graphical tools.

A new way to estimate the tetrachoric correlations of the items was shown. These correlations can be estimated from the DAGS algorithm used to estimate the parameters of the model. Even though these results can be easily extended to more general multidimensional item response theory models, it is probable, that the case of the LSMIRT models could be one of the more important.

References

1. J. Albert, Bayesian Estimation of Normal Item Response Curves Using Gibbs Sampling, *Journal of Educational Statistics*, 17, **3**, 251–269, 1992.
2. J. Albert and S. Chib, Bayesian Residual Analysis for Binary Response Regression Models, *Biometrika*,82, **3**, 747–759, 1995.

3. R. Bock and R. Gibbons, and E. Muraki, Full-Information Item Factor Analysis, *Applied Psychological Measurement*, 12, 261–280, 1995.
4. M. Brown, The Tetrachoric Correlation and its Asymptotic Standard Error, *Applied Statistics*, 26, **3**, 343-351, 1977.
5. F. Drasgow, Polychoric and polyserial correlations, *Encyclopedia of statistical sciences*, New York Wiley, vol. 7, 69–74,1988.
6. A. Montenegro. Multidimensional Item Response Theory Models where the Ability has a Linear Latent Structure, *Unpublished PhD dissertation*, Universidad Nacional de Colombia, 2011.
7. K. Pearson, Mathematical contributions to the theory of evolution. VII, *Philosophical Transactions of the Royal Society of London, Series A*, 195, 1–45, 1900.
8. U. Olson, Maximum likelihood estimation of the polychoric correlation coefficient, *Psychometrika*, 44, 433–460, 1979.
9. Y. Takane, and J. Leeuw, J, On the relationship between Item Response Theory and Factor Analysis of Discretized Variables, *Psychometrika*, 52, **3**, 393-408, 1987.
10. J. Zhang and W. Stout, The theoretical DETECT index of dimensionality and its applications to approximate simple structure, *Psychometrika*, 64, 213–249, 1999.

Proceedings, 2nd Stochastic Modeling Techniques and Data Analysis International Conference
5 - 8 June 2012 Chania Crete Greece

Maximum Likelihood Estimation of the VAR(1) Model Parameters with Missing Observations

Helena Mouriño[1] and M. Isabel Barão[2]

[1] Faculdade de Ciências - Universidade de Lisboa
Edifício C6, Piso 4 - Campo Grande, Lisboa, Portugal
(e-mail: mhnunes@fc.ul.pt)
[2] Faculdade de Ciências - Universidade de Lisboa
Edifício C6, Piso 4 - Campo Grande, Lisboa, Portugal
(e-mail: mibarao@fc.ul.pt)

Abstract. Missing-data problems are very common in practice. This feature of the data has to be taken into account to achieve reliable inferential results. Suppose that the univariate data set under analysis has missing observations. This paper examines the impact of selecting an auxiliary complete data set - whose underlying stochastic process is to some extent interdependent with the former - to improve the efficiency of the estimators for the relevant parameters of the model. The Vector AutoRegressive (VAR) Model has revealed to be a very useful tool in capturing the dynamics of bivariate time series. We propose maximum likelihood estimators for the parameters of the VAR(1) Model based on monotone missing data pattern. Estimators' precision is also derived. Afterwards, we compare the bivariate modelling scheme with its univariate counterpart, that is, modelling the univariate data set with missing observations by an AutoRegressive Moving Average ARMA(2,1) Model. Due to its practical importance in time series modelling, the behaviour of the AutoRegressive Model of order one, AR(1), is also analysed. We focus on the mean value of the main stochastic process. By simulation studies we conclude the estimator based on the VAR(1) Model is preferable to those derived from the univariate context.
Keywords: Missing Data, VAR Models, Maximum Likelihood Estimation, Hessian Matrix, First-Order Taylor Expansion.

1 Introduction

Literature on missing data has expanded in the last decades focusing mainly on univariate time series models (Jones[5], Kohn and Ansley[6], Pourahmadi[9], Gomez and Maravall[2], and Gomez *et al.*[3]). Nonetheless, if we find an auxiliary stochastic process - with no missing values in the data set - that is related with the process whose data set has missing observations, then we can swap information between both processes. Information transfer from the data set with no missing data to the one with missing values is emphasised. As a consequence, there will be an increase in knowledge about the

main stochastic process, which will certainly improve the overall inferential procedure. In this paper we analyse the veracity of this conjecture.

The above strategy deals with bivariate time series models. The skill of this procedure is to find out a suitable auxiliary stochastic process, denoted by $\{X_t\}_{t\in\mathbb{Z}}$, which is to some extent interdependent with the main stochastic process, designated by $\{Y_t\}_{t\in\mathbb{Z}}$. Due to its practical importance in time series modelling, we consider data generated by a first-order Vector AutoRegressive (VAR(1)) Model. Also, we assume that the incomplete data set has a monotone missing data pattern. We follow a likelihood-based approach to estimate the parameters of the model. It is worth pointing out that likelihood-based estimation is largely used in literature to handle the problem of missing data (Little and Rubin[7], and Gomez and Maravall[2]). The precision of the maximum likelihood estimators is also derived. It is important to emphasise that, even numerically, it is rather awkward to compute the precision of the estimators.

Whether the introduction of an auxiliary variable for estimating the model's parameters increases the accuracy of the estimators is worth of further investigation. To attain this goal, we compare the precision of the estimators just cited with those obtained from modelling the dynamics of the univariate stochastic process $\{Y_t\}_{t\in\mathbb{Z}}$ by an AutoRegressive Moving Average ARMA(2,1) Model, which corresponds to the marginal model of the bivariate VAR(1) Model (Heij *et al.*[4], and Tsay[10]). Due to its practical importance in time series modelling, the behaviour of the AutoRegressive Model of order one, AR(1), is also analysed. Simulation studies allow us to assess the relative efficiency of the different approaches. Special attention is paid to the estimator for the mean value of the stochastic process about which information available is scarce. This is a natural decision given the importance of the mean function of a stochastic process in understanding the behaviour of the time series under consideration.

The paper is organised as follows. In Section 2 we review the VAR(1) Model and highlight a few statistical properties that will be used in the remaining sections. In Section 3 we establish the monotone pattern of missing data and factorise the likelihood function of the VAR(1) Model. The maximum likelihood estimators (m.l.e.'s) of the parameters are obtained in Section 4. Their precision is also deduced. Section 5 reports the simulation studies in evaluating different approaches to estimate the mean value of the stochastic process $\{Y_t\}_{t\in\mathbb{Z}}$. The main conclusions are summarised in Section 6.

2 Brief Description of the VAR(1) Model

In this section, a few properties of the Vectorial Autoregressive Model of order one are analysed. The stochastic process underlying the complete data set is denoted by $\{X_t\}_{t\in\mathbb{Z}}$, while the other one is represented by $\{Y_t\}_{t\in\mathbb{Z}}$. The

VAR(1) Model under consideration takes the form

$$\begin{cases} X_t = \alpha_0 + \alpha_1\, X_{t-1} + \epsilon_t \\ Y_t = \beta_0 + \beta_1\, Y_{t-1} + \beta_2\, X_{t-1} + \xi_t \end{cases},\ t = 0,\, \pm 1,\, \pm 2,\, \ldots \qquad (1)$$

where ϵ_t and ξ_t are Gaussian white noise processes with zero mean and variances σ_ϵ^2 and σ_ξ^2, respectively. The structure of correlation between the error terms is different from zero only at the same date t, i.e., $\mathrm{Cov}(\epsilon_{t-i},\, \xi_{t-j}) = \sigma_{\epsilon\,\xi},$ for $i = j; \mathrm{Cov}(\epsilon_{t-i},\, \xi_{t-j}) = 0,$ for $i \neq j,\ i,\, j \in \mathbb{Z}$. Swapping information between both time series might introduce some noise in the overall process. Therefore, transfer of information from the smallest series to the largest one is not allowed here.

We have to introduce the restrictions $|\alpha_1| < 1$ and $|\beta_1| < 1$. They ensure not only that the underlying processes are ergodic for the respective means, but also that the stochastic processes are covariance stationary (see Nunes[8]). Hereafter, we assume that these restrictions are satisfied.

By writing out the stochastic system of equations (1) in matrix notation, the bivariate stochastic process $\mathbf{Z}_t = [X_t \quad Y_t]'$ can be expressed as

$$\mathbf{Z}_t = \begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix} + \begin{bmatrix} \alpha_1 & 0 \\ \beta_2 & \beta_1 \end{bmatrix} \begin{bmatrix} X_{t-1} \\ Y_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_t \\ \xi_t \end{bmatrix} = \mathbf{c} + \mathbf{\Phi}_1\, \mathbf{Z}_{t-1} + \epsilon_{\mathbf{t}}, \quad \mathbf{t} \in \mathbb{Z},$$

where $\epsilon_{\mathbf{t}} = [\epsilon_{\mathbf{t}} \quad \xi_{\mathbf{t}}]'$ is the 2-dimensional Gaussian white noise random vector.

Straightforward computations lead us to the following factoring of the probability density function of $\mathbf{Z}_t$ conditional to $\mathbf{Z}_{t-1} = \mathbf{z}_{t-1}$:

$$f_{\mathbf{Z}_t|\mathbf{z}_{t-1}}(\mathbf{z}_t|\mathbf{z}_{t-1}) = f_{X_t|\mathbf{z}_{t-1}}(x_t|\mathbf{z}_{t-1}) \times f_{Y_t|x_t,\mathbf{z}_{t-1}}(y_t|x_t, \mathbf{z}_{t-1}).$$

Thus, the joint distribution of the pair X_t and Y_t conditional to the values of the process at the previous date $t-1$, $\mathbf{Z}_{t-1}$, can be decomposed into the product of the marginal distribution of $X_t|_{\mathbf{Z}_{t-1}}$ and the conditional distribution of $Y_t|_{X_t,\mathbf{Z}_{t-1}}$. Both densities follow univariate Gaussian probability laws:

$$X_t\big|_{\mathbf{Z}_{t-1}=\mathbf{z}_{t-1}} \frown \mathcal{N}\big(\alpha_0 + \alpha_1\, x_{t-1},\, \sigma_\epsilon^2\big), \quad \text{for each date } t,\ t \in \mathbb{Z}. \qquad (2)$$

Also, $Y_t\big|_{X_t=x_t,\mathbf{Z}_{t-1}=\mathbf{z}_{t-1}}$ follows a Gaussian distribution with

$$\begin{aligned} &E\Big(Y_t\big|_{X_t=x_t,\mathbf{Z}_{t-1}=\mathbf{z}_{t-1}}\Big) = \\ &= \beta_0 + \beta_1\, y_{t-1} + \beta_2\, x_{t-1} + \tfrac{\sigma_{\epsilon\xi}}{\sigma_\epsilon^2}\ (x_t - \alpha_0 - \alpha_1\, x_{t-1}) = \\ &= \psi_0 + \psi_1\, x_t + \psi_2\, x_{t-1} + \beta_1\, y_{t-1}\,, \end{aligned} \qquad (3)$$

where $\psi_1 = \frac{\sigma_{\epsilon\xi}}{\sigma_\epsilon^2}$ or, for interpretive purposes, $\psi_1 = \frac{\sigma_\xi}{\sigma_\epsilon}\,\rho_{\epsilon\xi}$. The parameter ψ_1 describes, thus, a weighted correlation between the error terms ϵ_t and ξ_t.

The weight corresponds to the ratio of their standard deviations. Moreover, $\psi_0 = \beta_0 - \psi_1\,\alpha_0$; $\psi_2 = \beta_2 - \psi_1\,\alpha_1$. The variance has the following structure,

$$\mathrm{Var}\Big(Y_t \big| X_t = x_t, \mathbf{Z}_{t-1} = \mathbf{z}_{t-1}\Big) = \sigma_\xi^2 - \frac{\sigma_{\epsilon\xi}^2}{\sigma_\epsilon^2} = \sigma_\xi^2\left(1 - \rho_{\epsilon\xi}^2\right) \equiv \psi_3. \quad (4)$$

The conditional distribution of $Y_t|_{X_t,\mathbf{Z}_{t-1}}$ can be interpreted as a straight-line relationship between Y_t and X_t, X_{t-1}, Y_{t-1}. Additionally, it is worth mentioning that if $\rho_{\epsilon\xi} = \pm 1$ or $\sigma_\xi^2 = 0$, the above conditional distribution degenerates into its mean value. Henceforth, we will discard these particular cases, which means that $\psi_3 \neq 0$.

3 Factoring the Likelihood Based on Monotone Missing Data Pattern

We focus here our attention on theoretical background for factoring the likelihood function from the VAR(1) Model when there are missing values in the data. Suppose that we have the following monotone pattern of missing data,

x_0	x_2	$\ldots$	x_{m-1}	x_m	$\ldots$	x_{n-1}
y_0	y_2	$\ldots$	y_{m-1}			

That is, there are n observations available from the stochastic process $\{X_t\}_{t\in\mathbb{Z}}$, whereas due to some uncontrolled factors it was only possible to record $m\,(m < n)$ observations from the stochastic process $\{Y_t\}_{t\in\mathbb{Z}}$. In other words, there are $n - m$ missing observations from Y_t.

Let the observed bivariate sample of size n with missing values,

$$\{(x_0, y_0),\, (x_1,\, y_1),\, \ldots,\, (x_{m-1}, y_{m-1}),\, x_m,\, \ldots,\, x_{n-1}\},$$

denote a realisation of the random process $\mathbf{Z}_t = [X_t \quad Y_t]^{'}$, $t \in \mathbb{Z}$, which follows a vectorial autoregressive model of order one.

The conditional likelihood function is given by

$$L(\boldsymbol{\Upsilon}) = \prod_{\mathbf{t=1}}^{\mathbf{n-1}} \mathbf{f}_{\mathbf{X_t}|\mathbf{x_{t-1}}}(\mathbf{x_t}) \prod_{\mathbf{t=1}}^{\mathbf{m-1}} \mathbf{f}_{\mathbf{Y_t}|\mathbf{x_t},(\mathbf{x_{t-1}},\mathbf{Y_{t-1}})}(\mathbf{y_t}), \quad (5)$$

where $\boldsymbol{\Upsilon} = [\alpha_{\mathbf{0}} \quad \alpha_{\mathbf{1}} \quad \sigma_\epsilon^{\mathbf{2}} \quad \beta_{\mathbf{0}} \quad \beta_{\mathbf{1}} \quad \beta_{\mathbf{2}} \quad \sigma_\xi^{\mathbf{2}} \quad \sigma_{\epsilon\xi}]^{'}$ is the 8-dimensional vector of population parameters.

Based on both the Orthogonal Decomposition Theorem and the Approximation Theorem (Apostol[1]), we can state that the estimation subspaces associated with the conditional distributions $X_t|_{X_{t-1}}$ and $Y_t|_{X_t,X_{t-1},Y_{t-1}}$ are, by construction, orthogonal to each other. These arguments guarantee that the decomposition of the joint likelihood in two components can be carried

out with no loss of information for the whole estimation procedure. From (5) we can, thus, decompose the conditional loglikelihood function as follows,

$$l \equiv l(\boldsymbol{\Upsilon}) = \log \mathbf{L}(\boldsymbol{\Upsilon}) = \\ = \sum_{t=1}^{n-1} \log f_{X_t|X_{t-1}}(x_t) + \sum_{t=1}^{m-1} \log f_{Y_t|X_t,(X_{t-1},Y_{t-1})}(y_t) = l_1 + l_2. \quad (6)$$

The components l_1 and l_2 of the equation (6) will be maximised separately in Section 4.1.

4 Maximum Likelihood Estimators of the Parameters

4.1 Analytical Expressions

Theoretical developments carried out in this section rely on solving the loglikelihood equations obtained from the factored loglikelihood given by equation (6). Before proceeding with theoretical matters, we introduce some relevant notation in the ensuing paragraphs.

Let $\overline{X}_k^{(l)} = \frac{1}{k}\sum_{t=1}^{k} X_{t-l}$ represent the sample mean lagged l time units, $l = 0,\, 1$. The subscript $k,\ k = 1, \ldots, n-1$, allows us to identify the number of observations that take part in the computation of the sample mean. A similar notation is used for denoting the sample mean of the random sample $Y_0, \ldots, Y_k$, for $k = 1, \ldots, m-1$, $\overline{Y}_k^{(l)}$. According to this new definition, the sample variance of each univariate random variable based on k observations and lagged l time units is denoted by $\hat{\gamma}_{X,k}^{(l)}$ and $\hat{\gamma}_{Y,k}^{(l)}$, respectively for X and Y. Let $\hat{\gamma}_{X,k}^{*}(1) = \frac{1}{k}\sum_{t=1}^{k}\left(X_t - \overline{X}_k^{(0)}\right)\left(X_{t-1} - \overline{X}_k^{(1)}\right)$ describe the sample autocovariance coefficient at lag one for the stochastic process X_t, based on k observations. Its counterpart for the stochastic process Y_t, $\hat{\gamma}_{Y,k}^{*}(1)$, is obtained by changing notation accordingly. The sample autocorrelation coefficient of the random process X_t at lag one is denoted by $\hat{\rho}_{X,k}^{(1)} = \frac{\hat{\gamma}_{X,k}^{*}(1)}{\hat{\gamma}_{X,k}^{(0)}\,\hat{\gamma}_{X,k}^{(1)}}$.

Similarly, the empirical covariance between the random processes X_t and Y_t lagged one time unit is represented by $\hat{\gamma}_{XY}^{*}(1)$, for lagged values on Y, and $\hat{\gamma}_{YX}^{*}(1)$, for lagged values on X. The sample covariance coefficient of X_t and Y_t computed from l time units lag for each series, is denoted by $\hat{\gamma}_{XY}^{(l)}$.

- *Maximising the loglikelihood function l_1*
  Using the results (2) and (6), we readily find the following m.l.e.'s,

$$\hat{\alpha_0} = \overline{X}_{n-1}^{(0)} - \hat{\alpha_1}\,\overline{X}_{n-1}^{(1)}; \quad \hat{\alpha_1} = \frac{\hat{\gamma}_{X,n-1}^{*}(1)}{\hat{\gamma}_{X,n-1}^{(1)}}; \quad \hat{\sigma_\epsilon}^{\,2} = \frac{SS_R}{n-1},$$

  where SS_R is given by $SS_R = \sum_{t=1}^{n-1}\left(x_t - \hat{\alpha}_0 - \hat{\alpha}_1\, x_{t-1}\right)^2$.

- *Maximising the loglikelihood function l_2*
  Using the results reported in Section 2, we readily work out the maximum likelihood estimators for the parameters under study:

$$\begin{cases} \hat{\psi}_0 = \overline{Y}^{(0)}_{m-1} - \hat{\psi}_1\,\overline{X}^{(0)}_{m-1} - \hat{\psi}_2\,\overline{X}^{(1)}_{m-1} - \hat{\beta}_1\,\overline{Y}^{(1)}_{m-1} \\ \hat{\psi}_1 = \dfrac{1}{\hat{\gamma}^{(0)}_{X,m-1}}\left\{\hat{\gamma}^{(0)}_{XY} - \hat{\psi}_2\,\hat{\gamma}^{*}_{X,m-1}(1) - \hat{\beta}_1\,\hat{\gamma}^{*}_{XY}(1)\right\} \\ \hat{\psi}_2 = \dfrac{1}{\left(1-\left(\hat{\rho}^{(1)}_{X,m-1}\right)^2\right)\,\hat{\gamma}^{(1)}_{X,m-1}}\left\{\hat{\gamma}^{*}_{YX}(1) - \dfrac{\hat{\gamma}^{(0)}_{XY}\,\hat{\gamma}^{*}_{X,m-1}(1)}{\hat{\gamma}^{(0)}_{X,m-1}} - \right. \\ \qquad \left. - \hat{\beta}_1\,\hat{\gamma}^{(1)}_{XY} + \dfrac{\hat{\beta}_1\,\hat{\gamma}^{*}_{XY}(1)\,\hat{\gamma}^{*}_{X,m-1}(1)}{\hat{\gamma}^{(0)}_{X,m-1}}\right\}; \quad \hat{\psi}_3 = \dfrac{SS^{*}_{R}}{m-1} \\ \hat{\beta}_1 = \dfrac{1}{\left(\hat{\gamma}^{(1)}_{Y,m-1} + \left(\overline{Y}^{(1)}_{m-1}\right)^2\right)}\left\{\hat{\gamma}^{*}_{Y,m-1}(1) + \overline{Y}^{(0)}_{m-1}\overline{Y}^{(1)}_{m-1} - \hat{\psi}_0\,\overline{Y}^{(1)}_{m-1} - \right. \\ \qquad \left. - \hat{\psi}_1\,\hat{\gamma}^{*}_{XY}(1) - \hat{\psi}_1\,\overline{X}^{(0)}_{m-1}\,\overline{Y}^{(1)}_{m-1} - \hat{\psi}_2\left(\hat{\gamma}^{(1)}_{XY} - \overline{X}^{(1)}_{m-1}\,\overline{Y}^{(1)}_{m-1}\right)\right\} \end{cases}$$

  where $SS^{*}_{R} = \sum_{t=1}^{n-1}\left(y_t - \hat{\psi}_0 - \hat{\psi}_1\,x_t - \hat{\psi}_2\,x_{t-1} - \hat{\beta}_1\,y_{t-1}\right)^2$.

Straightforward computations lead us to the estimators of the mean values, variances and covariances at lag zero of the VAR(1) Model. The analytical expression of the m.l.e. of the mean value of $\{Y_t\}_{t\in\mathbb{Z}}$ is given by

$$\hat{\mu}_Y = \frac{\hat{\alpha}_0\,\hat{\beta}_2 + \hat{\beta}_0\,(1-\hat{\alpha}_1)}{(1-\hat{\alpha}_1)\,(1-\hat{\beta}_1)}. \tag{7}$$

It is worth stressing that this estimator will play a central role in the following sections.

4.2 Precision of the Estimators

In this section, the precision of the maximum likelihood estimators of the relevant parameters of the model are derived. The main idea is the following: first, we study the statistical properties of the vector $\hat{\boldsymbol{\Theta}}$, where $\hat{\boldsymbol{\Theta}} = \left[\hat{\boldsymbol{\Theta}}_1 \quad \hat{\boldsymbol{\Theta}}_2\right]'$, with $\hat{\boldsymbol{\Theta}}_1 = \left[\hat{\alpha}_0 \quad \hat{\alpha}_1 \quad \hat{\sigma}^2_{\epsilon}\right]'$ and $\hat{\boldsymbol{\Theta}}_2 = \left[\hat{\psi}_0 \quad \hat{\psi}_1 \quad \hat{\psi}_2 \quad \hat{\beta}_1 \quad \hat{\psi}_3\right]'$. For notation consistency, the unknown parameter β_1 is either denoted by β_1 or ψ_4. That is, $\psi_4 \equiv \beta_1$. Secondly, we derive the precision of the m.l.e.'s of the original parameters of the VAR(1) model (see equation 1), that is, $\boldsymbol{\Upsilon} = [\alpha_{\mathbf{0}} \quad \alpha_{\mathbf{1}} \quad \sigma^{\mathbf{2}}_{\boldsymbol{\epsilon}} \quad \beta_{\mathbf{0}} \quad \beta_{\mathbf{1}} \quad \beta_{\mathbf{2}} \quad \sigma^{\mathbf{2}}_{\boldsymbol{\xi}} \quad \sigma_{\epsilon\xi}]'$. The approximate variance-covariance matrix of the m.l.e.'s for the vector of parameters $\boldsymbol{\Upsilon}$ is obtained

by the first-order Taylor expansion at $\boldsymbol{\Upsilon}$. We also use the chain rule for derivatives of vector fields (for details, see Apostol[1]). Using the same methodologies, we finally obtain the approximate variance-covariance matrix of the m.l.e.'s for the mean vector and the variance-covariance matrix at lag zero of the VAR(1) model with a monotone pattern of missingness, represented by $\boldsymbol{\Xi_2} = [\mu_{\mathbf{X}} \quad \mu_{\mathbf{Y}} \quad \sigma^{\mathbf{2}}_{\mathbf{X}} \quad \sigma^{\mathbf{2}}_{\mathbf{Y}} \quad \sigma_{\mathbf{XY}}]^{'}$, where $\boldsymbol{\Xi_2}$ is a sub-matrix of $\boldsymbol{\Xi}$, $\boldsymbol{\Xi} = \left[\boldsymbol{\Xi_1} \quad \boldsymbol{\Xi_2}\right]^{'}$, with $\boldsymbol{\Xi_1} = [\alpha_{\mathbf{0}} \quad \alpha_{\mathbf{1}} \quad \sigma^{\mathbf{2}}_{\epsilon}]^{'}$.

It can be shown that the approximate variance-covariance matrix of the m.l.e.'s of $\boldsymbol{\Xi_2}$, denoted by $\boldsymbol{\Sigma}$, can be written in the following partitioned form

$$\boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma_1} & \boldsymbol{\Sigma_2} \\ \hline \boldsymbol{\Sigma_3} & \boldsymbol{\Sigma_4} \end{array}\right],$$

with

$$\begin{aligned}
\boldsymbol{\Sigma_1} =& \hat{\sigma}^2_{\epsilon}\, \mathbf{G}_{11} \left(\mathbf{U}^{'}_{R}\mathbf{U}_{R}\right)^{-1} \mathbf{G}^{'}_{11} + \frac{2\,\hat{\sigma}^4_{\epsilon}}{n-1}\, \mathbf{G}_{12}\, \mathbf{G}^{'}_{12} + \psi_1\, \mathbf{H}_{11} \left(\mathbf{U}^{'}\mathbf{U}\right)^{-1} \mathbf{H}^{'}_{11} + \\
&+ \frac{2\,\psi^2_3}{m-1}\, \mathbf{H}_{12}\, \mathbf{H}^{'}_{12} \\
\boldsymbol{\Sigma_2} =& \hat{\sigma}^2_{\epsilon}\, \mathbf{G}_{11} \left(\mathbf{U}^{'}_{R}\mathbf{U}_{R}\right)^{-1} \mathbf{G}^{'}_{21} + \frac{2\,\hat{\sigma}^4_{\epsilon}}{n-1}\, \mathbf{G}_{12}\, \mathbf{G}^{'}_{22} + \psi_3\, \mathbf{H}_{11} \left(\mathbf{U}^{'}\mathbf{U}\right)^{-1} \mathbf{H}^{'}_{21} \\
\boldsymbol{\Sigma_3} =& \hat{\sigma}^2_{\epsilon}\, \mathbf{G}_{21} \left(\mathbf{U}^{'}_{R}\mathbf{U}_{R}\right)^{-1} \mathbf{G}^{'}_{11} + \frac{2\,\hat{\sigma}^4_{\epsilon}}{n-1}\, \mathbf{G}_{22}\, \mathbf{G}^{'}_{12} + \psi_3\, \mathbf{H}_{21} \left(\mathbf{U}^{'}\mathbf{U}\right)^{-1} \mathbf{H}^{'}_{11} \\
\boldsymbol{\Sigma_4} =& \hat{\sigma}^2_{\epsilon}\, \mathbf{G}_{21} \left(\mathbf{U}^{'}_{R}\mathbf{U}_{R}\right)^{-1} \mathbf{G}^{'}_{21} + \frac{2\,\hat{\sigma}^4_{\epsilon}}{n-1}\, \mathbf{G}_{22}\, \mathbf{G}^{'}_{22} + \psi_3\, \mathbf{H}_{21} \left(\mathbf{U}^{'}\mathbf{U}\right)^{-1} \mathbf{H}^{'}_{21},
\end{aligned}$$

where the matrices $\mathbf{U}$ and $\mathbf{U}_R$ are, respectively, defined by

$$\mathbf{U} = \begin{bmatrix} 1 & X_1 & X_0 & Y_0 \\ 1 & X_2 & X_1 & Y_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{m-1} & X_{m-2} & Y_{m-2} \end{bmatrix}, \quad \mathbf{U}_R = \begin{bmatrix} 1 & X_0 \\ 1 & X_1 \\ \vdots & \vdots \\ 1 & X_{m-2} \end{bmatrix}.$$

The matrix $\mathbf{G}$ corresponds to the first order partial derivatives from the composite functions that relate μ_X, μ_Y, σ^2_X, σ^2_Y, σ_{XY} with the vector of parameters $\boldsymbol{\Theta_1}$. On the other hand, the matrix $\mathbf{H}$ corresponds to the first order partial derivatives of the vector $\boldsymbol{\Xi_2}$ with respect to the vector $\boldsymbol{\Theta_2}$.

We cannot write down explicit expressions for those variances and covariances. The limitation arises from the inability to invert the matrix product $\mathbf{U}^{'}\mathbf{U}$ in analytical terms. Hence, its inverse can only be accomplished by numerical techniques using the observed sampled data. This point will be pursued further in Section 5.

5 Simulation Studies

In this section, we analyse the effects of using different strategies to estimate the mean value of the stochastic process $\{Y_t,\, t \in \mathbb{Z}\}$, denoted by μ_Y. More

precisely, the bivariate modelling scheme and its univariate counterparts are compared. Simulation studies are carried out to evaluate the relative efficiency of the estimators with interest.

The m.l.e. of the mean value of the stochastic process $\{Y_t,\ t \in \mathbb{Z}\}$ based on the VAR(1) Model is given by (7). We need to compare this estimator to those obtained by considering the univariate stochastic process $\{Y_t,\ t \in \mathbb{Z}\}$ itself. More precisely, having in mind that we are handling a bivariate VAR(1) Model, the corresponding marginal model is the ARMA(2,1) (Heij *et al.*[4], Tsay[10]). On the other hand, the AR(1) Model is one of the most popular models due to its practical importance in time series modelling. Therefore, the behaviour of the AR(1) Model is also evaluated.

The relative efficiency of $\hat{\mu}_{\text{VAR}}$ with respect to $\hat{\mu}_{\text{ARMA}}$ is quantified by the Gain Index, GI, expressed as percentage,

$$GI = \frac{\text{Var}(\hat{\mu}_{\text{ARMA}}) - \text{Var}(\hat{\mu}_{\text{VAR}})}{\text{Var}(\hat{\mu}_{\text{ARMA}})} \times 100,$$

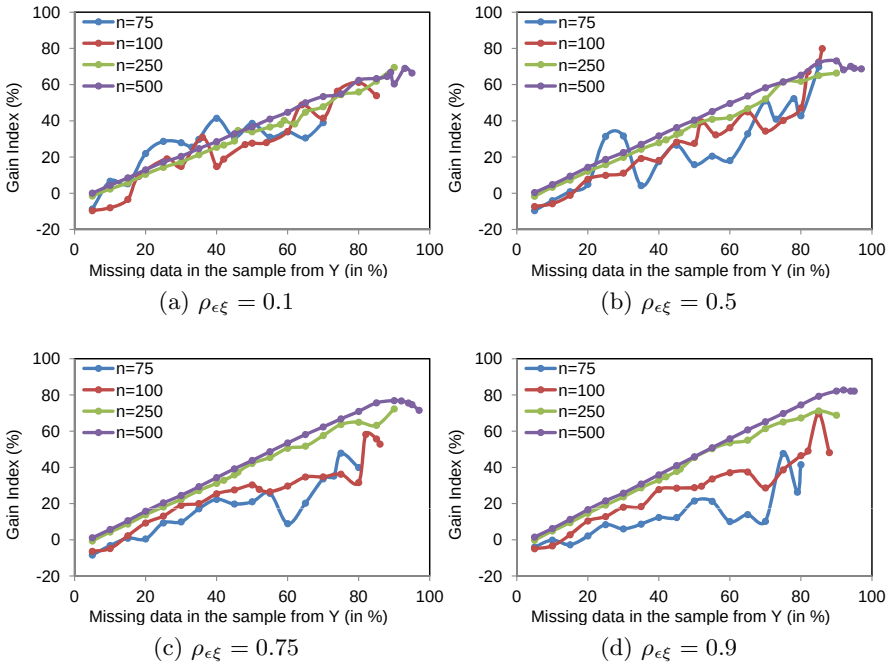
where $\hat{\mu}_{\text{VAR}}$ represents the m.l.e. estimator based on the VAR(1) Model and $\hat{\mu}_{\text{ARMA}}$ comes from the ARMA(2,1) Model. A similar strategy for comparing the estimators $\hat{\mu}_{\text{ARMA}}$ and $\hat{\mu}_{\text{AR}}$ was developed, with $\hat{\mu}_{\text{AR}}$ representing the m.l.e. estimator based on the AR(1) Model.

The data were generated by the VAR(1) Model (system of equations 1). In order to make comparisons on the same basis, a few assumptions to the parameters of the VAR(1) Model are made. We consider that $\mu_X = \mu_Y = 0$. These restrictions have no influence on the results because they are equivalent to $\alpha_0 = \beta_0 = 0$, that is, the constant terms of the VAR(1) Model are equal to zero (system of equations 1). Additionally, we introduce the restriction $\sigma_\epsilon^2 = \sigma_\xi^2 = 1$. Since the correlation coefficient regulates the supply of information between the stochastic processes $\{X_t\}_{t\in\mathbb{Z}}$ and $\{Y_t\}_{t\in\mathbb{Z}}$, special emphasis is given to this parameter. We used the grid of points $\rho_{\epsilon\xi} = 0.1, 0.5, 0.75, 0.9$. It is worth stressing that the value $\rho_{\epsilon\xi} = 1$ is not allowable in this context (see Section 2 for the details).

We analyse the performance of the estimators based on different sample sizes, $n = 75, 100, 250$ and 500. Taking as a reference the dimension of the sampled data from the auxiliary random process $\{X_t\}_{t\in\mathbb{Z}}$, simulations reported next are based on different percentages of missing observations. Simulation runs for each combination of the parameters are based on 1000 replicates. After each simulation run, the Gain Index is computed.

The estimators $\hat{\mu}_{\text{VAR}}$ and $\hat{\mu}_{\text{ARMA}}$ are compared in Figure 1. For each combination of the parameters of the model, we represent graphically the gain indexes as functions of the percentage of missing data in the sampled data from the stochastic process $\{Y_t\}_{t\in\mathbb{Z}}$.

Figure 1 shows that the plot of the gain index against the percentage of missing data in the sample from the stochastic process $\{Y_t\}_{t\in\mathbb{Z}}$ behave roughly as a linear function, regardless of the combination of the parameters.



(a) $\rho_{\epsilon\xi} = 0.1$ (b) $\rho_{\epsilon\xi} = 0.5$

(c) $\rho_{\epsilon\xi} = 0.75$ (d) $\rho_{\epsilon\xi} = 0.9$

Fig. 1. Graphical representation of GI. The data were obtained from a VAR(1) Model, with $\alpha_0 = \beta_0 = 0,\ \alpha_1 = 0.6,\ \beta_1 = 0.7,\ \beta_2 = 0.8$.

In outline, the more the percentage of missing values in the sampled data, the more precise is the estimator $\hat{\mu}_{\text{VAR}}$ when compared with $\hat{\mu}_{\text{ARMA}}$. Comparison between $\hat{\mu}_{\text{VAR}}$ and $\hat{\mu}_{\text{AR}}$ reveals a similar behaviour and provides even higher gains than those obtained from the VAR(1) *versus* AR(1) modelling scheme (the graphics are not shown here).

Further, the gain in precision by using the sophisticated estimator $\hat{\mu}_{\text{VAR}}$ rather than $\hat{\mu}_{\text{ARMA}}$ increases as the strength of the linear relationship between the processes $\{X_t\}_{t\in\mathbb{Z}}$ and $\{Y_t\}_{t\in\mathbb{Z}}$ (described by the correlation coefficient) rises from $\rho = 0.1$ to $\rho = 0.9$ (see Figure 1). This statement has been revealed to be true for both the ARMA(2,1) and AR(1) modelling schemes.

Summing up, the estimator $\hat{\mu}_{\text{VAR}}$ is preferable to those explored in the univariate context, that is, either $\hat{\mu}_{\text{ARMA}}$ or $\hat{\mu}_{\text{AR}}$.

6 Conclusions

This article deals with the problem of missing data in a univariate sample. We have considered an auxiliary complete data set, whose underlying stochastic process is serially correlated with the former by the VAR(1) Model structure.

We have proposed maximum likelihood estimators for the relevant parameters of the model based on a monotone missing data pattern. The precision of the estimators have also been derived. Special emphasis has been given to the estimator for the mean value of the stochastic process whose sampled data has missing values, μ_Y. We have compared the performance of the estimator for μ_Y based on the VAR(1) Model with a monotone pattern of missing data with those obtained from both the ARMA(2,1) Model and the AR(1) Model. By simulation studies, we have showed that the estimator derived in this article based on the VAR(1) Model performs better than those derived from the univariate context. More precisely, for large sample sizes and high percentage of missing observations on the sample from $\{Y_t\}_{t\in\mathbb{Z}}$ considerable gain is obtained. Moreover, the more the linear strength between both stochastic processes, the more the gain in precision obtained from using the sophisticated estimator from the VAR(1) context rather than those derived from the univariate context.

References

1. Apostol, T. M., *Calculus - Second Edition*, John Wiley & Sons, Inc., Singapore (1969).
2. Gómez, V. and Maravall, A., "Estimation, Prediction and Interpolation for Nonstationary Series With the Kalman Filter", *Journal of the American Statistical Association* 89, 611–624 (1994).
3. Gómez, V., Maravall, A., and Peña, D., "Missing observations in ARIMA models: Skipping approach versus additive outlier approach", in *Journal of Econometrics*, 88, 341–363 (1999).
4. Heij, C., De Boer, P., Franses, P. H., Kloek, T., and van Dijk, H. K., *Econometric Methods with Applications in Business and Economics*, Oxford University Press, New York (2004).
5. Jones, R. H., "Maximum Likelihood Fitting of ARMA Models to Time Series With Missing Observations", *Technometrics* 22, 389–395 (1980).
6. Kohn, R. and Ansley, C. F., "Estimation, Prediction and Interpolation for ARIMA Models With Missing Data", *Journal of the American Statistical Association* 81, 751–761 (1986).
7. Little, R. J. A. and Rubin, D. B., *Statistical Analysis with Missing Data*, John Wiley & Sons, Inc., New York (1987).
8. Nunes, M .H., *Dynamics Relating Phytoplankton Abundance With Upwelling Events. An Approach to the Problem of Missing Data in the Gaussian Context.*, PhD Thesis, University of Lisbon, Lisboa (2006).
9. Pourahmadi, M., "Estimation and Interpolation of Missing Values of a Stationary Time Series", *Technometrics* 10, 149–169 (1989).
10. Tsay, R. S., *Analysis of Financial Time Series - Third Edition*, John Wiley & Sons, Inc., New Jersey (2010).

Stability approach : Stationary Versus Stable Population Model

Barun Kumar Mukhopadhyay

Retired Scientist, Population Studies Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata-700108, India.

Email: barun_mukhopadhyay@yahoo.com

Abstract

An attempt is made to testify whether life table stationary population model may be used in demographic study on stability matter, when there is some difficulty to avail of stable population model. The paper simply tries to apply both the models in a demographic problem to estimate very young population through some multiple regression equation using other two sections of age distributions. The testing of the two models has been done through comparing birth rates estimated from the estimated 0-4 population in the present attempt and the same from the published data of Sample Registration System. Further, in order to test the two models applied on the same kind of data ten years apart, an index of dissimilarity has been used. The result shows that either of the two has almost same applicability. The present study is a kind of analysis through which demographers, in particular may get some help in handling their huge data to a certain extent of relief, if at all the vulnerability of the present paper is acceptable.

Introduction

It is already known due to medical and public health technology arriving from Europe and North America and within the country development of new kind of antibiotics, pesticides and other live saving drugs etc., with new economic policies due to globalization, the death rates in India already declined rapidly in the long past and at present almost stable. With fertility declining slowly the Indian population age structures, hence may be assumed to reach at a stable form. In many demographic problems, the stable age distributions are of immense use. In practical sense usually stable age data are procured from secondary source but requiring some inputs. Otherwise life table age data are easily available as many countries have their own life tables. As a matter of fact, in the present attempt the study on utilizing two models, stationary as well as stable separately is accomplished on an empirical basis in order to test whether life table stationary age distribution may be a tool replacing stable population age distribution. As such in the paper a section of population age distribution which is very controversial has been chosen in order to establish the comparability and suitability between the two. The very young population of age group 0-4 years as far as census enumeration is concerned particularly in developing countries is the target of the present paper. To quote, particularly under enumeration, to a great extent, has been usually observed in this particular section of population apart from misreporting and other kind of error. As a matter of fact Jain (1954), as census actuary,

unadjusted the population aged 0-4 years. Because of omissions in enumeration to a greater extent in this particular group, any smoothing process is difficult to be applied with confidence as observed by him. The paper as such tries to estimate the 0-4 population to fill the gap of error which is very useful in terms of adjusted figure required for overall adjustment of the entire age distribution if 5+ population is adjusted through other techniques such as Zelnik (1961), UN (1967) etc. Use of two models, stationary and stable is made separately by Mukhopadhyay (2006; 1986) in estimating the population 0-4 assuming a relationship between fertility and mortality on the one side and on the other the two age groups 5-14 and 60 + which are supposed to indicate either constant or very slowly declining status of recent fertility by the former and the mortality by the latter. In this connection, it may be mentioned that a number of researchers in the past used age groups in some form or other on the one hand and on the other demographic variables to establish some relationships (Thompson,1942; Swanson & Palmore, 1976; Carrier & Hobcraft, 1971). Very recently a paper (Unisa et al, 2009) tried to find a relationship between Age not stated and Whipple's index apart from other detailed analysis on ANS. The present paper used the stationary population model for the time point of census of India, 1981 and the stable population model is for 1991 time point of census of India, 1991. In both the cases west model life table stationary and stable population age distributions have been obtained using the appropriate parameters with linear interpolation as and when necessary (Coale and Demeny, 1966).

The Model

At the time of forming the two models, Migration factor is ignored because of age selective nature of migration. The two parameters of e_0^0 and growth rates have been taken from Indian censuses. Before using the proportional (per cent) figures of the three age groups a multiple linear regression equation is formed. Now it is necessary to transform the variables from binomial to normal in order to facilitate using normal regression. This transformation has been done using $\mathrm{Sin}^{-1}\sqrt{p}$ (Arc Sin) where p stands for proportions. The standard regression equation for both the cases is given below,

$$y = b_{y1.2}\, x_1 + b_{y2.1}\, x_2 \ ,$$

where $b_{y1.2}$ and $b_{y2.1}$ are partial regression coefficients. A table of values of $\mathrm{Sin}^{-1}\sqrt{p}$ (Snedecor and Cochran, 1959) have been used with linear interpolation, if necessary. The percentage distribution from the stationary and stable age distributions for the three age groups 0-4, 5-14 and 60+ for 1981 and 1991 periods have been prepared from which the Arc Sin values have been transformed and the estimated Arc Sin values of 0-4 are given below.

Table 1 : Arc Sin values of stationary and stable population age data with estimated values of 0-4 years

Name of the Major States in India	Arc Sin values for 1981 Stationary Population (model tables)			Estimated Arc Sin Values of age-gr. 0-4 population (Stationary)	Arc Sin values for 1991 Stable Population (model tables)			Estimated Arc Sin Values of age-gr. 0-4 population (Stable)
	0-4	5-14	60+		0-4	5-14	60+	
Andhra Pd.	14.463	23.134	23.404	16.486	16.110	22.710	23.810	14.810
Assam				NA	16.640	23.260	22.140	15.150
Bihar	16.795	23.460	22.654	16.810	16.430	22.870	23.420	15.040
Gujarat	16.342	22.974	23.695	16.334	16.110	22.630	23.970	13.910
Haryana	16.375	23.038	23.612	16.392	15.680	22.140	26.060	15.330
Himachal Pd.	NA	NA	NA	NA	15.680	22.140	25.990	12.930
J & K	NA	NA	NA	NA	14.540	20.530	22.870	NA
Karnataka	16.773	23.444	22.638	16.779	15.790	22.300	25.480	13.760
Kerala	16.630	23.300	23.022	16.651	15.340	21.810	26.420	10.480
Madhya Pd.	16.441	23.110	23.460	16.462	16.430	22.710	23.890	16.250
Maharashtra	17.900	24.540	20.234	17.878	15.790	22.380	24.730	15.040
Orissa	16.784	23.444	22.686	16.795	16.320	22.790	23.810	13.770
Punjab	16.375	23.038	23.612	16.392	15.450	21.810	27.560	13.010
Rajasthan	16.290	22.950	23.810	16.305	16.220	22.550	24.270	16.450
Tamil Nadu	16.630	23.300	23.022	16.651	16.000	22.630	24.120	11.700
Utter Pradesh	17.350	24.005	21.454	17.347	16.430	22.790	23.890	15.760
West Bengal	16.930	23.588	24.160	16.776	15.890	22.460	24.650	14.920

The least square estimates are obtained from normal equations in both the cases. The estimated equation for stationary population model is given below,

$$y = \underset{(0.0665)}{0.7871}\, x_1 - \underset{(0.0284)}{0.7899}\, x_2 ; \qquad R^2=0.988$$

and the same for stable population model is given as,

$$y = \underset{(0.0809)}{0.7990}\, x_1 - \underset{(0.0722)}{0.1939}\, x_2 ; \qquad R^2=0.982$$

From the values of the above two equations, it is obvious that there is little gap between the stationary and stable population model after using the respective

Applications in actual raw census data

At the outset one important point to mention here is that when these model equations are actually applied on the raw census data on the proportional values of age groups, 5-14 and 60+ to get estimates or adjusted figures of proportion of population in the age group of 0-4, the question of some error in the two independent raw age groups comes and ultimately there may be some effect on the estimated values of 0-4 years. But as these two groups are much broader, the errors in those groups may be much less and secondly if at all errors present these may be less effective since both of them have almost the similar kind of error with some variation, if at all and ultimately the effect on the overall estimate may be ignored. Moreover applying the above models the Arc Sin transformed values
of the per cent values of 5-14 and 60+ which are not separately given are used.

Observed and estimated figures of 0-4 population

In the following paragraph the comparison has been made between the raw census figures (observed) and the estimated figures obtained from the regression model based on two standard models namely life table stationary age distribution and stable age distribution. it may be said that population in very young age group of 0-4 years where undercounting is a major problem, in addition to other errors, could, however be adjusted from other age groups suitably in either stationary or stable population model, when it is usually taken that Indian age distribution has arrived at a stable stage from the trend of fertility and mortality rates. Moreover, the adjusted figures of population in this age group seem to be quite consistent with the pattern of recent trend in the rates of fertility and mortality indicators. The following table shows the two things at a time with an error component and facilitates the extent of a parity between the two models. From the table below it may usually be found that observed values are less than the estimated figures both for the 1981 and 1991 censuses. From the figures no such significant variations are observed because of time gap of only 10 years. It is almost sure to say that Indian census count under enumeration, to a great extent, is observed for very young population particularly because of several social and cultural taboos prevailing

prevailing in rural set up. Although the present paper attempted to apply the two models, stationary and stable the

$$Overall\ survival\ ratio(SR) = \frac{105 \times SR(0-4)\ for\ male + 100 \times SR(0-4)\ for\ female}{205}$$

In order to obtain the survival factors for age group 0-4 from the model tables, e_0^0 for each state is required for the stationary model. Secondly e_0^0 and intrinsic growth rates are required for the stable model. Then on the basis of actual values which are not mentioned here, survival factors have been obtained for the two models. Using survival ratios for the different states of India, estimated average number of births during last five years have been calculated by reverse survival method. The corresponding populations for the period two and half years before have been calculated using exponential growth rate in both the cases. As the main purpose of the paper is to study the comparability between the stationary and stable population models in actual field of some demographic variables every detail in this section is deliberately avoided. After all these inputs and justifications, eventually birth rates for the two cases separately have been prepared for the period two and a half years before the each census date of 1981 and 1991 were obtained. And these have then be compared with the same from SRS figures for the corresponding period. The following table gives the figures as mentioned in the above paragraph.

Table 2 : Comparison of estimated birth rates and SRS rates for the two models

Name of the Major States in India	Estimated birth rates from 0-4 population		Sample Registration System (SRS) birth rates	
	1981 (1979)	1991 (1989)	During 1976-'81	During 1987-'91
Andhra Pd.	35.8	34.4	32.4	27.4
Bihar	36.5	38.3	34.0	37.3
Gujarat	35.1	33.2	35.7	29.5
Haryana	38.1	35.5	35.5	33.8
Karnataka	36.4	33.5	28.2	28.7
Kerala	32.1	25.9	26.0	20.3
Madhya Pd.	41.2	35.9	37.8	37.0
Maharashtra	35.1	32.2	27.6	29.4
Orissa	37.9	33.4	32.0	31.9
Punjab	31.9	30.7	30.0	28.5
Rajasthan	38.8	38.4	35.6	33.3
Tamil Nadu	32.5	28.2	28.9	22.7
Uttar Pradesh	40.4	35.6	39.9	37.1
West Bengal	37.5	34.9	31.3	28.4

Source for SRS figures : Family welfare programme in India, Year Book, for different years, New Delhi.

While comparing estimated birth rates with those of SRS (Sample registration system of India) more or less during the same period, one may notice that the estimated figures give in general some higher values than the SRS figures. The same nature is seen

the cases inconsistencies are also found in some states where there are some more differences. However it may be said that validity of the estimated figures from the very young population on the basis of some model as far as adjusted age data, here it is 0-4 population is concerned.

An Index of dissimilarity

The differential is measured through 1991and 1981 ratio of each state under study by dissimilarity coefficient (DC). An Index of Dissimilarity (ID) is developed to examine the differentials over the states as a single measure. This concept is obtained from two studies conducted by Mukhopadhyay (2011) and Mukhopadhyay and Majumdar (2011).

$$\text{Index of Dissimilarity (ID)} = \frac{1}{n}\sum\left|R_x - 1\right|$$

where, R_x stands for 1991/1981 at x, x is for states, R_x is assumed to be Dissimilarity Coefficient (DC) in this paper at individual level, n = total number of observations. The minimum value of the index of dissimilarity (ID) is 'zero' indicating no difference between. The value above 'zero' indicates the extent of differences in absolute term.

Table 4 : Distribution of values of DCs and ABS (DC-1) with the overall ID

Name of the Major States in India	Ratio of Estimated Errors 1991 and 1981(DCs)	ABS (DC-1)
Andhra Pd.	1.3590	0.3590
Bihar	1.0858	0.0858
Gujarat	1.1482	0.1482
Haryana	0.8824	0.1176
Karnataka	0.4834	0.5166
Kerala	1.0390	0.0390
Madhya Pd.	0.8219	0.1781
Maharashtra	0.7900	0.2100
Orissa	0.8721	0.1278
Punjab	1.0506	0.0506
Rajasthan	1.2505	0.2505
Tamil Nadu	1.4262	0.4262
Utter Pradesh	0.9231	0.0768
West Bengal	1.0029	0.0030

ID= 0.009682

The values of DCs are observed to be sometimes more than unity indicating higher proportion of values of 0-4 in 1991 as compared to 1981 for states like, Andhra Pradesh, Bihar, Gujarat, Kerala, Punjab, Rajasthan, Tamil Nadu and West Bengal. Among them the values are very close among 1991 and 1981 censuses in states like West Bengal, in particular and Kerala, Punjab and Bihar. There may be several reasons, like non fulfillment of desired family size, sex biases non use of contraceptive methods etc. However the values are less than unity in rest of the states indicating positive result due to fertility decline during 1981 to 1991. Now to sum up, from the third column it is found that the figures seem to be very less except for some cases. And the overall ID value is still very less indicating that there may be less differences in the result whether stationary or stable population models are used in this particular type of study.

Acknowledgment

I'm thankful to my daughter and wife to help me in matters related to typing and other computer layout.

References

1. Carrier, N and Hobcraft. J., *Demographic estimations for developing societies*, Population Investigation Committee, London School of Economics (1971).

2. Coale, A.J. and Demeny. P., *Regional model life tables and stable population*, Princeton University Press, Princeton, New Jersey (1966).

3. Cochran, G.W., *Statistical methods applied to experiments in agriculture and biology*, USA (1959).

4. Jain, S.P., *Age tables***,** Paper No. **3** of 1954, census of India, 1951 (1954).

5. Mukhopadhyay,B.K.,Gender *Issues of Younger population in India*, paper sent to European Population Conference at Sweden (2011).

6. Mukhopadhyay, B.K. and Majumdar P.K. Status of Gender-Differentials and Trends in India : Population, Health, on and Employment in Pal, M ; Pathak, P ; Bharati, P;Ghosh,B;Majumder,A.(ed.), *Gender Issues and Empowerment women,*(accepted) Nova Science Publishers, INC., New York (2011).

7. Mukhopadhyay, B.K Statistical analysis in evaluating the demographic data with reference to certain Special characteristics, *PhD Thesis*, T.M.Bhagalpur University, Bhagalpur (2006).

8. Mukhopadhyay, B.K., An age adjustment of very young children of India, 1981 and reappraisal of fertility and mortality rates : A model appraoch, *Genus*, Vol. XLII-n.**3-4**, Roma, Italy (1986).

9. Snedecor, W.G and Cochran,G.W., Statistical methods applied to experiments in agriculture and biology, USA (1959).

10. Swanson, D.A and Palmore, J.A.,Two parameter regression estimates of current life expectancy at birth, *Asian and Pacific News Letter*, Part I, 3 (2) (1976).

11. Thompson, W.S., *Population Problems*, M_c Graw-Hill (1942).

12. Unisa, S. Dwivedi, L K. Reshmi, R.S. Kumar,K.., ,*Age reporting in Indian census: An insight*, paper of the 26th Conference of IUSSP, Marrakech, Morocco (2009).

13. United Nations., Methods of estimating basic demographic measures from incomplete data Manual IV, *Population Studies*, **42**, New York (1967).

14. Zelnik, M., Age heaping in the United States Census, *Milbank Memorial Fund Quarterly,* 39 (1961).