

Volume 32 Number 2 June 2008

ISSN 0350-5596

Informatica

**An International Journal of Computing
and Informatics**

Special Issue:

**Demography and Informatics:
Some Interconnections**

Guest Editors:

Janez Malačič

Janina Józwiak

Alexia Fürnkranz-Prskawetz



1977

EDITORIAL BOARDS, PUBLISHING COUNCIL

Informatica is a journal primarily covering the European computer science and informatics community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor from the Editorial Board can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the list of referees. Each paper bears the name of the editor who appointed the referees. Each editor can propose new members for the Editorial Board or referees. Editors and referees inactive for a longer period can be automatically replaced. Changes in the Editorial Board are confirmed by the Executive Editors.

The coordination necessary is made through the Executive Editors who examine the reviews, sort the accepted articles and maintain appropriate international distribution. The Executive Board is appointed by the Society Informatika. Informatica is partially supported by the Slovenian Ministry of Higher Education, Science and Technology.

Each author is guaranteed to receive the reviews of his article. When accepted, publication in Informatica is guaranteed in less than one year after the Executive Editors receive the corrected version of the article.

Executive Editor – Editor in Chief

Anton P. Železnikar
Volaričeva 8, Ljubljana, Slovenia
s51em@lea.hamradio.si
<http://lea.hamradio.si/~s51em/>

Executive Associate Editor - Managing Editor

Matjaž Gams, Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Phone: +386 1 4773 900, Fax: +386 1 251 93 85
matjaz.gams@ijs.si
<http://dis.ijs.si/mezi/matjaz.html>

Executive Associate Editor - Deputy Managing Editor

Mitja Luštrek, Jožef Stefan Institute
mitja.lustrek@ijs.si

Executive Associate Editor - Technical Editor

Drago Torkar, Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Phone: +386 1 4773 900, Fax: +386 1 251 93 85
drago.torkar@ijs.si

Board of Advisors:

Ivan Bratko, Marko Jagodič,
Tomaž Pisanski, Stanko Strmčnik

Editorial Board

Juan Carlos Augusto (Argentina)
Costin Badica (Romania)
Vladimir Batagelj (Slovenia)
Francesco Bergadano (Italy)
Ranjit Biswas (India)
Marco Botta (Italy)
Pavel Brazdil (Portugal)
Andrej Brodnik (Slovenia)
Ivan Bruha (Canada)
Wray Buntine (Finland)
Hubert L. Dreyfus (USA)
Jozo Dujmović (USA)
Johann Eder (Austria)
Vladimir A. Fomichov (Russia)
Maria Ganzha (Poland)
Janez Grad (Slovenia)
Marjan Gušev (Macedonia)
Dimitris Kanellopoulos (Greece)
Hiroaki Kitano (Japan)
Igor Kononenko (Slovenia)
Miroslav Kubat (USA)
Ante Lauc (Croatia)
Jadran Lenarčič (Slovenia)
Huan Liu (USA)
Suzana Loskovska (Macedonia)
Ramon L. de Mantras (Spain)
Angelo Montanari (Italy)
Pavol Návrat (Slovakia)
Jerzy R. Nawrocki (Poland)
Nadja Nedjah (Brasil)
Franc Novak (Slovenia)
Marcin Paprzycki (USA/Poland)
Gert S. Pedersen (Denmark)
Ivana Podnar Žarko (Croatia)
Karl H. Pribram (USA)
Luc De Raedt (Belgium)
Dejan Raković (Serbia)
Jean Ramaekers (Belgium)
Wilhelm Rossak (Germany)
Ivan Rozman (Slovenia)
Sugata Sanyal (India)
Walter Schempp (Germany)
Johannes Schwinn (Germany)
Zhongzhi Shi (China)
Oliviero Stock (Italy)
Robert Trapp (Austria)
Terry Winograd (USA)
Stefan Wrobel (Germany)
Konrad Wrona (France)
Xindong Wu (USA)

Publishing Council:

Ciril Baškovič, Cene Bavec, Jožko Čuk,
Matjan Krisper, Vladislav Rajkovič, Tatjana Welzer

Editor's Introduction to the Special Issue

Demography and Informatics: Some Interconnections

Our modern time is a period of specialization in almost every field of human activity. The scientific activity is probably the best example of the specialization of scientists in usually very tiny spheres of investigation and research. This kind of approach to the development of the scientific knowledge has been very fruitful in practically all scientific disciplines, also including demography and informatics. However, in science and research work the cross disciplinary cooperation is also very important. From time to time it has provided a real breakthrough in many research areas. Therefore, the interdisciplinary discussions, investigations and collaborations are of crucial importance for the modern science.

In this line of thinking the editors of *Informatica* journal have invited the three of us to edit a special issue with the title "Demography and Informatics: Some Interconnections". This invitation was an extension of the collaboration which started at the 10th International Multiconference Information Society 2007, 8th – 9th October 2007, Ljubljana, Slovenia. Between several other conferences there was also a demographic one titled "Slovenian Demographic Challenges of the 21st Century". The conference gathered professionals from different scientific disciplines to discuss recent demographic situations and problems of Slovenia.

The idea of a special issue of the *Informatica* journal was to upgrade some contributions from the conference and to acquire some additional papers from the broader international scientific community and to produce an issue which would be devoted to (some of) the interconnections between demography and informatics. The time period was short and the number of collected papers limited. In spite of that we have succeeded to prepare this special issue.

It contains six very different papers dealing with numerous demographic and broader informatics topics. Only two of them are devoted to the local Slovenian situation. In the first of these two J. Sambt and M. Čok analyze the economic consequences of the demographic pressure on the public pension system in Slovenia. In the second M. Černič Istenič and A. Kveder study the relationship between fertility decisions of different generations and developmental characteristics of urban and rural areas in Slovenia. This second paper is more sociologically oriented.

The next two papers of J. Malačič and J. Bijak and D. Kupiszewska deal with broader European topics. J. Malačič analyzes recent dynamics of late fertility trends in Europe and concludes that late age – specific fertility will very likely retain more or less marginal shares of total fertility in a modern demographic regime. The paper of J. Bijak and D. Kupiszewska addresses selected computational issues of poor statistics on international migration. A range of computational methods and the

algorithm for choosing the best method for estimating missing data are proposed and illustrated by examples for selected countries.

The last two papers cover the topics which deal with the world perspective. M. Gams and J. Krivec use data mining techniques to discover the determinants of fertility. The decision trees are presented and the iterative use of data mining techniques employed for the finding of complex relations. Finally, S. Korenjak – Černe, N. Kejžar and V. Batagelj present the results of the clustering of the population pyramids. The big number of countries and the counties of the United States of America are examined.

J. Malačič, J. Józwiak and A. Fürnkranz-Prskawetz

Demographic Pressure on the Public Pension System

Jože Sambt

University of Ljubljana, Faculty of Economics
Kardeljeva ploščad 17, SI-1000 Ljubljana, Slovenia
E-mail: joze.sambt@ef.uni-lj.si

Mitja Čok

University of Ljubljana, Faculty of Economics
Kardeljeva ploščad 17, SI-1000 Ljubljana, Slovenia
E-mail: mitja.cok@ef.uni-lj.si

Keywords: models, projections, Slovenia, pensions, intergenerational relations

Received: March 15, 2007

The combination of low fertility, decreasing mortality and the baby-boom generation entering retirement will dramatically increase the share of elderly people in Slovenia in future decades. Without further changes in the pension system this will bring about strong pressure on the public pension system. In the analysis we use a cohort-based model to project the share of public expenditure on pensions in gross domestic product. This model enables us to analyse the long-term effects of the forthcoming demographic changes in connection with the current public pension system. The projected rise in pension expenditure will have to be mitigated at some point in the future and reducing pension benefits is one of the options. The Slovenian pension legislation provides equity among pensioners who retire at different points in time. An equal reduction of pension benefits suggests an equal distribution of burdens arising from the ageing population. However, the model reveals very different effects of this measure in relation to different cohorts. The analysis tackles increasingly relevant topics of intergenerational relations and questions the distribution of fiscal burdens and benefit among cohorts and generations.

Povzetek: Članek predstavi predviden prihodnji pritisk spremenjene starostne strukture prebivalstva Slovenije na javnofinančni pokojninski sistem in učinke njegove blažitve na posamezne kohorte.

1 Introduction

For decades the research community has warned the public via demographic projections concerning forthcoming radical demographic changes. However, this did not actually receive much general attention until developments started to influence current generations and caused problems associated with the long-term sustainability of the public finance system. Resolving population-ageing pressures on the public finance system means elevating the tax burden or cutting benefits to individuals. Of course, these measures do not appeal to the public and politicians are trying to delay them as long as possible. Lately, this is hardly possible any more and population ageing is becoming one of the central issues facing the European Union and many other institutions and countries around the globe.

The pressure on public expenditure stems predominantly from three systems: health care, long-term care and the pension system. In the paper we concentrate on public pension expenditure. We present projections of this expenditure in the future. It is unlikely that Slovenia's present public finance system can absorb such a large increase in pension expenditure. An adjustment in the direction of a sustainable path raises questions about distributing burdens over different cohorts and

generations. The current pension legislation provides horizontal equity between existing and new pensioners. We argue that this seemingly fair arrangement is only one possibility which brings about a different impact on different cohorts when introducing the time dimension to the analysis, as our estimations created by the cohort-based model reveal.

In Section 2, the latest demographic projections for Slovenia are given; presenting forthcoming demographic changes. In Section 3 the cohort-based model used in the analysis is explained. Section 4 includes projections of public pension expenditures in Slovenia expressed as a share of GDP. Section 5 presents the effect of limited pension spending on different cohorts. The conclusions are given in Section 6.

2 Future demographic development in Slovenia

The Slovenian population belongs to the modern demographic regime with low levels of fertility and mortality. In 1981 the total fertility rate¹ ('TFR') dropped

¹ The total fertility rate is the average number of children that a woman gives birth to in her lifetime, assuming: 1) that the prevailing rates remain unchanged; and 2) she will survive from birth through to the end of her reproductive life.

below 2.1, which represents the replacement fertility rate for developed countries. Since then, the TFR has been continually falling and in the last few years it has stabilized at a level somewhat above 1.2. Since 1960 mortality has also been declining. Life expectancy at birth increased in the 1958/59 – 2005/06 period from 65.6 to 74.8 years for males and from 70.7 to 81.9 years for females.

During the 1960s Slovenia transformed from a traditional emigration country to an immigration destination. The most important here was the Balkan South-East to North-West immigration stream. In the 1970-1990 period, all net migration flows between Slovenia and other federal parts of Yugoslavia were positive for the then north-west developed Yugoslav republic [11]. Since 1990 this pattern has not changed in spite of the several new state borders which have emerged after the breakdown of Yugoslavia. In the last decade the net migration has amounted to 2,000 to 3,000 people per year, with higher values being seen in the last two years (6,400 in 2005 and 6,200 in 2006).

These trends of fertility, mortality and net migration formed the basis of the Eurostat demographic projections [6] published in 2005. Figure 1 to Figure 3 present the assumptions about fertility, mortality and migrations on which those projections are based.

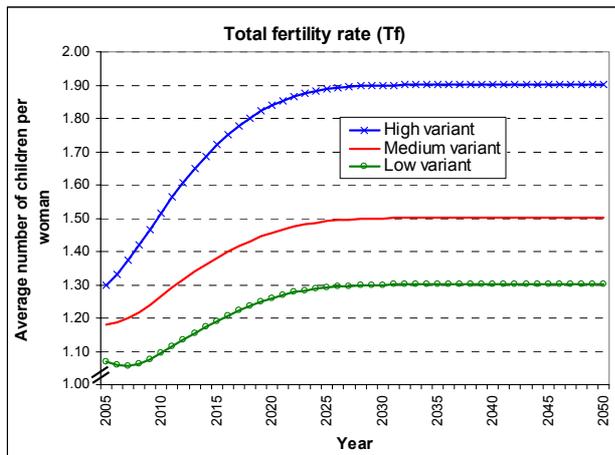


Figure 1: Demographic assumption: fertility

The results for the low, medium and high variants of the demographic projections are summarised in Table 1. According to the medium variant, the size of the population decreases by about 100,000 inhabitants by 2050. Because of an assumed substantial positive net migration, a fertility increase (compared to the current level) and increased longevity the projected drop in the total population is only moderate. However, the share of elderly people (aged 65 years and over) is projected to double in the period up until 2050: from the current 16% to 31%.

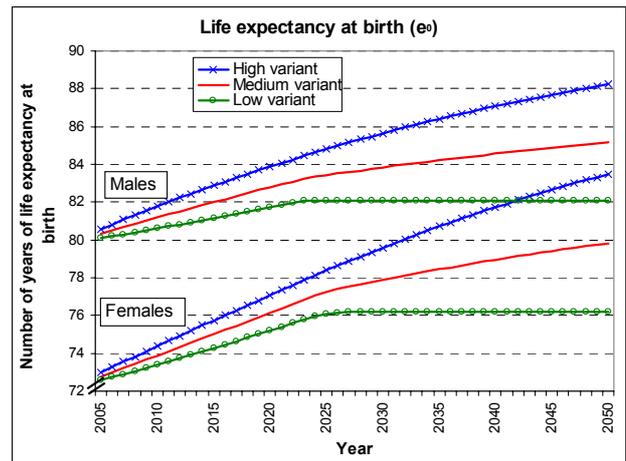
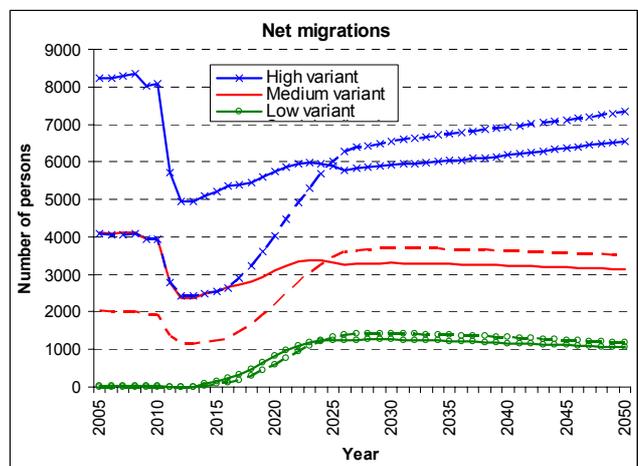


Figure 2: Demographic assumption: mortality



Note: A dashed line denotes values for females while a full line denotes values for males.

Figure 3: Demographic assumption: migration

The high (low) variant projects much higher (lower) number of inhabitants since it combines optimistic (pessimistic) assumptions regarding all three dimensions: fertility, mortality and net migration. Despite the big differences in those two variants compared to the medium variant, the share of people aged 65 years and over is very similar.

To achieve further information, two additional variants are simulated by rearranging the assumptions relating to fertility, mortality and net migration to obtain a range of extremes regarding the share of elderly people (65 years and over). In the ‘favourable’² variant we combine fertility and net migrations from the high variant with the mortality from the low variant, while in the ‘unfavourable’ variant we combine fertility and net migrations from the low variant and mortality from the high variant. However, even with the very optimistic combination of assumptions the projected share of people aged 65 years and over increases from the current 16% to 24% by 2050, while the pessimistic combination of assumptions even yields an increase to 38%.

	1. 1. 2005*	1. 1. 2010	1. 1. 2020	1. 1. 2030	1. 1. 2040	1. 1. 2050
Low variant	1,997,590	1,963,853	1,890,415	1,801,674	1,663,014	1,490,760
Aged 65 years and over (%)	15.3	16.7	20.8	25.8	29.4	32.7
Medium variant	1,997,590	2,014,802	2,016,694	2,005,999	1,965,314	1,900,839
Aged 65 years and over (%)	15.3	16.5	20.4	25.1	28.4	31.1
High variant	1,997,590	2,069,175	2,170,058	2,271,619	2,383,601	2,520,801
Aged 65 years and over (%)	15.3	16.3	19.7	23.8	26.5	28.0
'Favourable' variant	1,997,590	2,063,048	2,143,299	2,216,623	2,276,009	2,355,838
Aged 65 years and over (%)	15.3	16.1	19.0	22.4	23.7	23.9
'Unfavourable' variant	1,997,590	1,969,808	1,916,117	1,853,220	1,760,870	1,634,846
Aged 65 years and over (%)	15.3	16.9	21.6	27.4	32.6	37.7

* Actual number of inhabitants.

Table 1: Eurostat's demographic projections for Slovenia, published in 2005

Besides these assumptions, the age structure of the population also affects the results. The large baby-boom generations born after World War II are approaching their retirement. In the next decade these people are going to shift from employment to retirement status, rapidly expanding the share of the elderly population.

These processes are also reflected in Figure 4 where a population pyramid³ graphically represents the projected demographic development of Slovenia. The pyramid in solid colour is for the year 2020, while the shades are outlines presenting demographic developments in the years 2005 to 2050. Shading in the lower age groups depicts the number of people in those age groups in the time period 2005-2019, while the shading in the higher age groups represents the number of people in those older cohorts for the projected period 2021-2050. The figure presents an intermediate stage (the situation in 2020) i.e. the 'emptying' of the number of people in lower age groups and the 'filling in' of higher age groups during the period of the projections.

Europe and many other countries around the world, especially developed ones, are also facing a similar process of population ageing so Slovenia is no exception in this regard. However, longevity in Slovenia is increasing relatively rapidly compared with other developed countries and fertility is among the lowest in the world and therefore the process is especially intensified.

3 Cohort-based model

The analysis used in this section derives from a cohort-based model which simulates pension expenditures for different cohorts. It is based on a pension profiles matrix, population matrix and a coefficient matrix.

The pension profile matrix includes average pensions by age. It builds on the situation from the base year (2006).

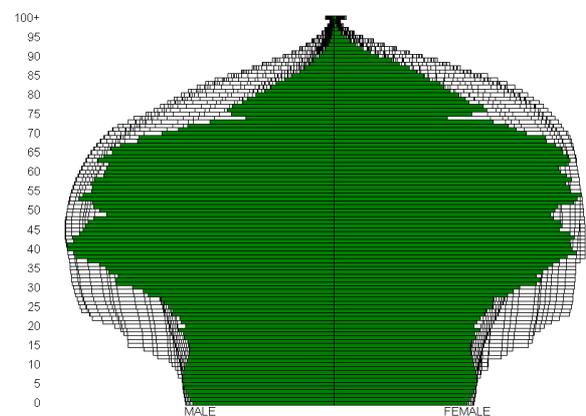


Figure 4: Population pyramid for projected Slovenian population in 2020

The population matrix is based on the Eurostat demographic projections presented earlier. Where a longer time span is required, we extend the medium variant of the projections. We thereby use the same data set and software as Eurostat, holding demographic assumptions at the level for 2050.

The coefficient matrix summarises the effects of the Pension and Disability Insurance Act introduced in 1999 (PDIA-1999) and gradually coming into effect after 2000. The PDIA-1999 will thus be fully effective in 2024. The transition period is taken into account along with further changes to the pension legislation from 2005. With detailed data about individuals retiring before introducing the PDIA-1999 we simulated the retirement behaviour, wage level etc. – amidst the new conditions.

Technically, the matrices have age (a) in their rows and calendar years (t) in their columns. The matrix of pension profiles ($PROF$) has the pension levels in its cells; the population matrix (P) has the number of people in its cells; and the coefficients matrix (C) contains the coefficients of adjustments. Pensions paid to individuals aged k in year t are thus calculated as:

$$PENS_{a,t} = PROF_{a,t} P_{a,t} C_{a,t} G_t \tag{Eq. 1}$$

where G contains coefficients of the cumulative growth of wages from the base year (in our case 2006) to

³ The population pyramid is a graphical presentation of the population age structure in a presented year. On the vertical axis are age groups and on the horizontal axis is the number (sometimes the share) of the population (males on the left-hand side and females on the right-hand side) by those age groups.

time t . Namely, according to the Slovenian pension legislation the growth of pensions is fully indexed in line with the growth of wages (but in practice in the period up until 2024 pensions will grow more slowly due to certain provisions of the pension legislation which are captured by the coefficient matrix (C)). Pension expenditures in year t are calculated as the sum of projected pension expenditures by all age groups:

$$PENS_t = \sum_{a=0}^D PENS_{a,t} \quad (\text{Eq. 2})$$

where index a runs from 0 to D ; with D denoting the maximum length of life (in our model it is the age group 100+).

This pension module is linked to the GDP module. Pension expenditures are namely expressed as a share of GDP. GDP growth is calculated as the sum of the labour productivity growth rate and the labour input growth rate. Further, the labour input growth is defined as the growth of employees in the 16–64 years age group. The same procedure is also used by the European Commission (Ageing Working Group) when projecting public expenditures related to the ageing population [5].

In our model various demographic projections thus affect public pension expenditures expressed as a share of GDP through the pension expenditures and GDP. Labour productivity growth enters into the calculations exogenously, neglecting any possible dependence on the number and age structure of the population (employees).

In our analysis we are interested in the effects on pension benefits of currently living cohorts on the assumption that the government caps pension spending at some point. Depending on the chosen ceiling level (cap) the extent of the pension cuts differs. Thus we calculate the reduction in pension benefits that representatives of different cohorts will receive in their remaining lifetime, i.e. the reduction of their *pension wealth*. For a detailed explanation of the pension wealth definition and empirical results, see, for example [2] and [7]. Technically, pension wealth is obtained by performing a diagonal aggregation of the expected pension benefits in the future, discounted back to the base year t_0 (in our case 2006).

$$PENS_a^W = \sum_{i=a}^D PENS_{i,t_0+i-a} (1+r)^{-(i-a)} \quad (\text{Eq. 3})$$

4 Projecting pension expenditures

Slovenia inherited a PAYG system from former Yugoslavia after gaining its independence in 1991. The transition to a market economy and loss of markets in other Yugoslav republics caused high unemployment and other labour-market problems. Mass early retirements in the early 1990s was used to mitigate them. Consequently, the share of pensions in GDP rose from 9.6% in 1989 to 14.4% in 1994. The Pension and Disability Insurance Act

of 1992 (PDIA-1992) introduced a gradual increase in the retirement age and some other measures to cope with rapidly growing pension expenditures. In 1999 the share of pensions in GDP was 13.4% but the projections simulated a sharp increase in the future if no further measures are introduced.

In 1999 a new Pension and Disability Insurance Act (PDIA-1999) was adopted. It tightens retirement conditions and decreases benefits deriving from the mandatory pension system (for details, see [3] and [12]). The effects of this pension reform have been analysed by several researchers using different models, assumptions and partial simulations of the complex Slovenian pension system. However, all of them concluded that, despite the positive effects of the pension reform starting in 2000, further measures will be required in the future to maintain the system’s long-term fiscal sustainability (see, for example, [3], [4], [13], [15] and [16]).

In the analysis we present results of the projections stemming from the cohort-based model, presented earlier in the text.

Apart from the methodology and assumptions described earlier, we applied assumptions about macroeconomic variables (like productivity growth, activity rates etc.) provided by the European Commission [5]. For linking employment rates with the retirement rates the sub-model of the *Institute of Macroeconomic Analysis and Development* [10] is used. Without going into further details about the assumptions and calculations in Figure 5 we present the results by different demographic variants. In the analysis we excluded some categories of pension expenditure which predominantly or exclusively have a social function (e.g., state pensions).

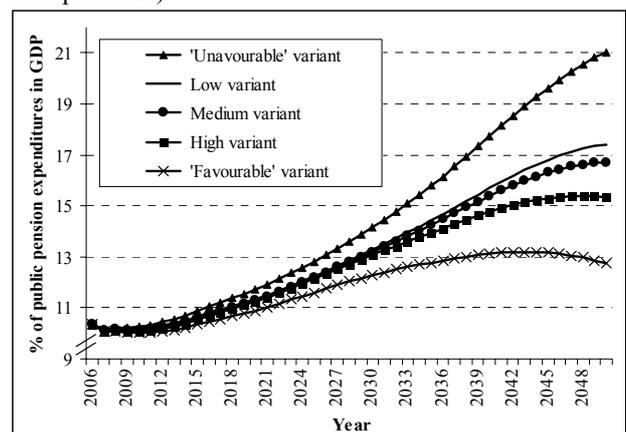


Figure 5: Projections of pension expenditures in GDP

It is projected that, without further measures in the next decade, demographic pressure of increased longevity and low fertility, further enhanced by baby-boom generations entering retirement is transmitted into public pension system.

5 Distributing the fiscal burden

One of the cornerstones of the Slovenian pension legislation is the principle of equal benefits for

individuals with the same pension conditions, regardless of the time they retired. The first item in Article 151 of the latest PDIA (adjusted in 2005) explicitly states that upon the February⁴ adjustment of the growth of pensions in line with the growth of wages an adjustment for existing pensioners relative to new pensioners is also taking place ‘... to assure equal rights for pensioners, who have retired at different time points’. That is to say, someone who will retire 10 years from now will have for the same retiring conditions the same net replacement rate⁵ as someone who is already retired.

According to the current pension legislation, the replacement rate for people entering retirement is decreasing up until 2024, but the pension growth of existing pensioners will also lag behind the growth of wages to keep pace with the conditions for new pensioners. This arrangement suggests fairness in the light of growing questions about the positions of different generations to engage in the problems (or *challenges* if we employ the word used by politicians) of an ageing population. If we ignore payments and benefits into/from the public system that individuals faced in the past, it seems reasonable and fair to distribute future burdens equally across all generations. In the rest of the paper we contrast this view with the results of the cohort-based model.

With the model for each cohort we follow all taxes/benefits that it will pay/receive to/from the public finance system. In this paper we concentrate on public pension benefits only. We calculate pension wealth by cohorts by discounting projected future pensions to the base year, which in our case is 2006. The present value of future pensions is very sensitive to the assumed discount rate. A 5% discount rate was used. This value, for example, is also used as a default value in the generational accounting method for discounting future flows to the base year. However, since we do not analyse absolute values this effect is much smaller as we analyse the *relative* position regarding the present value of future pension benefits (by cohorts). The range from 2 to 7 percent has been tested without having a significant effect on the results and without altering the conclusions of the analysis.

Estimating the effects of the pension legislation on an individual’s pension benefits is undertaken by following the parameters of the pension system and the life expectancy of the individual. For an individual with full pension conditions the scale of the projected net replacement rate is presented in Figure 6 – until 2008 there are actual values, thereafter followed by projected values.

On the other hand, a calculation at the cohort level has to take into consideration the heterogeneity of the

cohort in terms of service years, the future mortality of the cohort members etc.



Figure 6: Net replacement rate (in %) for an individual with full pension conditions

We believe that the government will not allow an increase in pension expenditures as a share of GDP to the levels presented in Section 4. In the analysis we thus assume that at a certain point the government will limit any further rise in public pension expenditure. This could be done in various ways, among which we analyse the option of cutting pensions. We set the tolerated maximum share of pensions in GDP, alternatively, at rates of 11, 12, 13, 14 and 15 percent; i.e. we assume that after reaching this ‘tolerated’ maximum the government would cut all pensions simultaneously in order to achieve this goal. We concentrate only on medium variant of demographic projections.

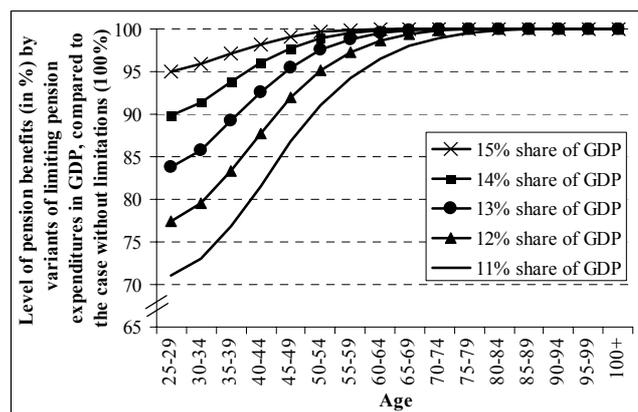


Figure 7: Reduction of pensions by age when limiting pension expenditure

Figure 7 compares: 1) the present values of pensions that representatives of a certain age group would receive in their remaining lifetime when limiting pension expenditure with 2) the case without limitations.

⁴ The growth of pensions is adjusted in line with the growth of wages in February and in November.

⁵ The net replacement rate for pensions is defined as a person's net pension divided by their net wage before retirement. This includes an assumption about non-extreme high or low values which are limited by maximum and minimum values etc.

Age	Limiting share of pensions in GDP to:				
	15%	14%	13%	12%	11%
20-24	94.7	89.4	83.5	77.1	70.7
25-29	95.0	89.8	83.8	77.5	71.1
30-34	95.9	91.4	85.8	79.5	73.0
35-39	97.2	93.8	89.2	83.3	76.8
40-44	98.2	96.0	92.6	87.8	81.5
45-49	99.1	97.8	95.5	92.0	86.8
50-54	99.6	98.9	97.6	95.1	91.1
55-59	99.9	99.6	98.8	97.2	94.3
60-64	100.0	99.9	99.5	98.6	96.6
65-69	100.0	100.0	99.8	99.4	98.0
70-74	100.0	100.0	100.0	99.8	99.0
75-79	100.0	100.0	100.0	99.9	99.6
80-84	100.0	100.0	100.0	100.0	99.9
85-89	100.0	100.0	100.0	100.0	100.0
90-94	100.0	100.0	100.0	100.0	100.0
95-99	100.0	100.0	100.0	100.0	100.0
100+	100.0	100.0	100.0	100.0	100.0

Table 2: Reduction of discounted pension benefits by age when limiting pension expenditure

Figure 7 and Table 2 should be read as follows: if the government were to limit pension expenditures in GDP to 15%, then a representative of the 20-24 years age group (i.e. being on average 22.5 years old) would receive in their remaining life time 94.7% of the amount of pension benefits (discounted to the base year) compared to their remaining lifetime had the government not posed such limitations. This cut would thus reduce the discounted value of expected pension benefits for a person aged 22.5 years by 5.3%. However, the same measure would reduce the discounted value of expected pension benefits of someone aged 52.5 years by just 0.4%.

The key factor driving this part of the results is the timing of a cut in pension benefits. According to the projections, the share of public pension expenditure exceeds the 15% limit in 2039. A cut in pensions would follow thereafter; therefore the 50-54 years cohort would be only slightly affected. It would namely collect pension benefits until then at an unchanged rate and only a few of them would still be living to collect benefits at the reduced rate. Further, people aged 80-84 years in 2006, for example, would not be affected at all since according to demographic projections none of them will still be alive in 2039. On the contrary, those aged 20-24 years in the base year would receive reduced pensions for their entire period when retired.

If the government were to decide on a much tougher limitation of public pension expenditure – e.g. to 11% of GDP, the effect would be much greater. The cutting of pensions would have to start already in 2018 so practically all cohorts except those aged 85 years and more would be affected. But the magnitude of the reduction for different cohorts would be very different. For those aged 70-74 years this measure would reduce the discounted value of their pension benefits collected in their remaining lifetime by only 1%, while for those aged

20-24 years the reduction would be 29.3%. This can be explained by virtue of the fact that at the beginning only minor pension reductions would be required to stay within the 11% limit. This cohort would thus not be strongly affected by this measure. On the other hand, when cohorts currently aged 20-24 years collect pension benefits, a strong cut of pensions will be required to stay within the 11% limit.

The results of the analysis reveal that the timing of measures for mitigating the pressure of an ageing population on pension expenditures decisively determines the distribution of burdens across different cohorts. It is evident that pensioners and people approaching retirement will prefer delaying measures in the form of cutting pensions as long as possible. Ideally for them, they should not be implemented while they are still alive. On the contrary, younger cohort/generations would prefer (or at least they should) prompt actions to distribute the burden over all generations instead of only turning the burden on to them.

These opposite aspirations are confronted in the political field since decisions are made by politicians who are elected by people with a right to vote. Positions in this intergenerational ‘battle’ are thus very unequal. Children do not have voting power; nor do generations that have still to be born, which is especially emphasised by the method of generational accounts (see, for example, [1]), have representatives in these ‘negotiations’. On the contrary, there is a rapidly growing number of older people who have voting power and participate at elections over-proportionally (compared to those aged 18-30 years, for example) and who have very clear and unified criteria – the level of benefits they expect to receive from the government. ‘In democracies, one-issue voters have a disproportionate impact on the political process, since they don’t split their votes because of conflicting interests on other issues’ [14]. Some authors see this as enormous issue in the future, employing expressions like ‘war between generations’ [8] and the ‘coming generational storm’ [9], while some of them even see this as a threat to democracy in the future [14].

6 Conclusions

According to population projections published by Eurostat in 2005, drastic demographic changes are forthcoming. The share of elderly people aged 65 years and over is expected to about double from the current 16% to about 31% by 2050 in Slovenia. Other European countries and many other countries around the world face the same process of rapid population ageing. In Slovenia it is especially emphasised because of the still rapidly increasing longevity and the very low fertility which is among the lowest in the world.

This strong demographic pressure will effect public systems, especially the public pension system which is the focus of this article. It includes simulations of future public pension expenditures as a share of GDP using the cohort-based model. The effects of the pension reform passed in 1999 are expected to almost neutralise

demographic pressure during the next decade. Thereafter, the share would increase rapidly from about 11% to almost 17% in 2050 if no further measures are implemented. Our results are in line with earlier results using different models.

However, the message is not to postpone measures for a decade. On the contrary, the results speak in favour of acting in a timely fashion since the necessary measures will have to be more drastic if they are delayed. Further, as the analysis reveals there is a huge difference in distribution of the burden across cohorts depending on when the measures are implemented. In the pension legislation the proclaimed equity of replacement rates for pensioners retiring at different times thus does not mean equal burdens on all cohorts and generations.

Younger generations would prefer immediate action while older generations would benefit from delaying them for as long as possible. In this 'conflict' the older generations are in a much better position since they have voting power, they over-proportionally participate at elections and their political preferences are much more straightforward and therefore more powerful.

7 References

- [1] Auerbach, A. J., Gokhale, J., Kotlikoff, L. J. (1991). *A Meaningful Alternative to Deficit Accounting*. Working Paper no. 3589. National Bureau of Economic Research.
- [2] Brugiavini, K., Maser, A., Sundén, A. (2005). Measuring Pension Wealth. LWS Workshop "Construction and Usage of Comparable Microdata on Wealth: The LWS". Banca d'Italia, Perugia, Italy, 27–29 (January 2005).
- [3] Čok, M., Sambt, J., Berk, A., Košak, M. (2008). Long-Term Sustainability of the Slovenian Pension System. *Economic and Business Review*, vol. 10, no. 4., forthcoming. University of Ljubljana, Faculty of Economics.
- [4] European Commission (2006). The Impact of Ageing on Public Expenditure: Projections for the EU25 Member States on Pensions, Health Care, Long Term Care, Education and Unemployment Transfers (2004–2050). *European Economy*, Special report no. 1/2006. Office for Official Publications of the European Communities.
- [5] European Commission (2006). The 2005 Projections of Age-Related Expenditure (2004–50) for the EU-25 Member States: Underlying Assumptions and Projection Methodologies. *European Economy*, Special report no. 4/2005. Office for Official Publications of the European Communities.
- [6] Eurostat: Population projections. http://epp.eurostat.ec.europa.eu/portal/page?_pageid=1996,45323734&_dad=portal&_schema=PORTAL&screen=welcomeref&open=/popul/popula/proj/proj_trend/proj_tbp&language=en&product=EU_MAIN_TREE&root=EU_MAIN_TREE&scrollto=358
- [7] Feldstein, M. (1974). Social Security, Induced Retirement, and Aggregate Capital Accumulation. *Journal of Political Economy*, vol. 82, no. 5, 905–926. The University of Chicago Press.
- [8] Gokhale, J., Kotlikoff, J. L. (2001). *Is War Between Generations Inevitable?* NCPA Policy Report no. 246. National Center for Policy Analysis.
- [9] Kotlikoff, J. L., Burns, S. (2004). *The Coming Generational Storm: What You Need to Know about America's Economic Future*. MIT Press.
- [10] Kraigher, T. (2005). *Middle- and long term projection of demographic development of Slovenia and its socio-economic components* (in Slovene). Working Paper no. 10/2005, pp. 78. Insitute of Macroeconomic Analysis and Development.
- [11] Malačič, J. (2000). The Balkan migration stream South-East to North-West. *Studi emigrazione*, vol. 37, no. 139, pp. 581–594. Centro Studi Emigrazione Roma.
- [12] Pension and Disability Insurance Act (PDIA-1999), revised in 2005. http://www.uradni-list.si/_pdf/2005/Ur/u2005104.pdf
- [13] Sambt, J. (2004). *Generational Accounts for Slovenia* (in Slovene). Master's thesis. University of Ljubljana, Faculty of Economics.
- [14] Thurow, C. L. (1999). Generational Equity and the Birth of a Revolutionary Class. In the book: Williamson, B. J., Watts, M. D., Kingson, R. E.: *The Generational Equity Debate*. Columbia University Press.
- [15] Verbič, M. (2007). Varying the Parameters of the Slovenian Pension System: an Analysis with an Overlapping-Generations General Equilibrium Model. *Post-Communist Economies*, vol. 19, no. 4, pp. 449–470. Routledge.
- [16] Verbič, M., Majcen B., Van Nieuwkoop, R. (2006). Sustainability of the Slovenian Pension System: An Analysis with an Overlapping-Generations General Equilibrium Model. *Eastern European Economics*, vol. 44, no. 4, pp. 60–81. M.E. Sharpe Inc.

Urban-Rural Life Setting as the Explanatory Factor of Differences in Fertility Behaviour in Slovenia

Majda Černič Istenič

Sociomedical Institute at Scientific Research Centre of Slovenian Academy of Sciences and Arts,
Novi trg 2, 1000 Ljubljana, Slovenia
University of Ljubljana, Biotechnical Faculty, Agronomy Department,
Jamnikarjeva 101, 1000 Ljubljana, Slovenia,
E-mail: majdaci@zrc-sazu.si

Andrej Kveder

Population Activities Unit,
UNECE, Palais des Nations,
CH-1211 Geneva 10,
E-mail: andrej.kveder@unece.org

Keywords: survey data, statistical data, fertility behaviour, typology of urban-rural settings, Slovenia

Received: March 31, 2008

Contemporary research of fertility behaviour that considers simultaneous longitudinal and prospective inclusion of individual and contextual levels of observation increases its explanatory power of explanation. The only research data of the kind in Slovenian is Fertility behaviour of Slovenians which was a part of international project the Family and Fertility Survey (FFS), carried out in 1995. On the basis of this survey data combined with statistical census data in the form of socio-economic typology of Slovenian countryside the article aims to explain the tendency and reasons of still persistent difference in urban-rural fertility. The study of relationship between fertility decisions of different generations and developmental characteristics of rural and urban areas in Slovenia reveals that this relationship is a complex and dynamic one. Obtained research results call for diversified actions of population policies in Slovenia. In its field of actions not merely family and social policy should be integrated, but also space and regional measures that will consider different every day's life conditions and needs of people, living in particular space setting, and in this way assist them to fulfil their desired number of children.

Povzetek: Narejena je analiza slovenske rodnosti na podeželju in v mestih.

1 Introduction

According to Notestein's theory of demographic transition (1953) a high-level fertility regime of pre-modern societies was replaced with a low-level one owing to the process of modernisation and its accompanying processes: industrialisation and urbanisation. As supposed, the growth of big city agglomerations and mobile urban population disrupted the strength of traditional norms and commitments to a traditional way of life and stimulated individualism and affirmation of personal aspirations. However, as later research revealed (e.g. "The European Fertility Project" (Coale, Watkins, 1986)), the casual relationship between urbanisation and the beginning of demographic transition was not always entirely direct and unproblematic. For instance, the decline in fertility in France "began in the late eighteenth century, long before the appearance of the modern city, while the decline in fertility in England only got underway decades after cities like Birmingham and Manchester become grimy industrial centres" (Sharlin,

1986). In spite of relatively great variety in the beginning of demographic transition and the level of urbanisation causal relations between these two phenomena certainly existed. Research on social-group forerunners of fertility control in Europe (Livi-Bacci, 1986) proved that urban-rural difference in fertility existed even prior to the onset of general decline in fertility. Some groups within the city population (mainly elites: aristocracies and bourgeoisie) were practising effective family limitation which influenced the overall urban fertility level. Other reasons for variation in levels of fertility between urban and rural populations before the demographic transition are also low levels of nuptiality in the cities due to high concentration of servants (it was not easy for them to get married) (Sharlin, 1986). Urban-rural difference in natural fertility was also affected with factors such as infant mortality (higher in the cities) and breast-feeding (more widespread in the country) (van de Walle, 1986). After the completion of the transition from high fertility

to the low one (enforcement of modern demographic behaviour), urban marital fertility remains lower than rural marital fertility. The differences declined in size, but nonetheless continued to exist in most of the countries that experienced demographic transition (Andorka, 1978).

As some recent research (Černič Istenič, Kveder, 2000, Černič Istenič, Kveder, Obersnel Kveder, 2000) indicates, this 'state of affairs' also holds true for Slovenia: today all social strata that live in the cities still have fewer children than those who live in the countryside despite the fact that the difference in the level of fertility among various social strata is indeed diminishing (Javornik, 1999) due to the predominance of two children per family norm (Malačič, 1990, Černič Istenič 1994, Obersnel Kveder et al., 2001). To explicate this issue more precisely in the present article we explore the relationship between fertility decisions of different generations and developmental characteristics of rural and urban areas in Slovenia. On the basis of individual survey data combined with statistical census data in the form of socio-economic typology of Slovenian countryside we intend to discover the tendency of this still persistent difference in urban-rural fertility and its reasons. In this vein, the main traits of Slovenian urbanisation and deagrarianisation process over the last 150 years are firstly briefly sketched. Secondly, the main analytical frame of our analysis is outlined. It follows the pertinent theoretical observations which indicate that due to urbanization social, economic and political differentiation firstly increased elsewhere, but later on, a kind of homogenisation in social behaviour of the population took place due to further impacts of urban life patterns on the countryside. In the third part the main explanations of the selected data and applied methods are presented. In the fourth section, the results are presented according to the outlined hypothesis and in the final part they are discussed and conclusions are outlined.

1.1 Urbanisation and deagrarianisation of Slovenia

The beginning of urbanisation in Slovenia was relatively late. A growth of cities started only after 1848. In that time Ljubljana, the capital and the biggest city of today's Slovenia, had just 17000 residents. Urban-rural difference in fertility level existed before, during and after the demographic transition in Slovenia. The earliest available data for the level of fertility in Slovenia that make the comparison between urban and rural fertility possible pertain to generations born during 1873-1877 and during 1898-1902. This comparison shows that urban women had from 34 to 47 percentage points lower fertility than countryside women did (Šircelj, 1991). The beginning of demographic transition in Slovenia started only in the first quarter of the twenty-century. However, fertility in the cities already began to decline towards the end of the nineteenth-century. Firstly it started to decline in larger cities and then spread onto the smaller ones. At the break of the twentieth century fertility in cities

declined for about 40 percentage points and for about 20 percentage points in the countryside (Šircelj, 1991).

Slovenian society was still weakly structured in the middle of the nineteenth century. The main occupation was farming. Specifically, in 1868 the share of farm population was 81,4 percent. At the break of the 20th century it decreased to 73,2 percent (Klemenčič 2002). The difference between areas with the highest and the lowest share of farm population was 13 percentage points. In 1931 this difference reached 50 percentage points. The share of farm population decreased especially in the neighbourhoods of Ljubljana. At that time no association between the share of farm population and the level of fertility was observed. It appeared only after the Second World War. "Up to that time the level of fertility was much more influenced by the type of settlement where a woman lived than by the share of farm population" (Šircelj, 1991: 244). Since the Second World War farm women have the highest fertility and today they are the only socio-professional group who assures itself biological reproduction (Šircelj, Ilič, Kuhar, Zupančič, 1990, Šircelj 2007). However, fertility of farm women has actually no effect on the national fertility level due to its very low share compared to the whole population¹. In the period 1931-2002 the share of farm population declined from 60 to 6,5 percent, the most rapidly in the period 1948-1981; from 48,9 to 9,4 percent (Klemenčič 2002, Statistical Yearbook 2003). At the same time, the age structure of agrarian countryside deteriorated significantly. Young generation was/is immigrating into the cities whereas older generations remain in the countryside. In some parts of Slovenia (hills, highlands, the Karst and particularly borderlands) the age structure of rural areas is so unfavourable that it causes demographic extinction and stagnation. Namely, only 2 percent of the population lives on the above mentioned areas which represent 20 percent of the surface (Pečan, Ravbar, 1999). On the other hand, countryside has also experienced social strata transformation. Due to abandonment of farming and forestry, moving of provincials into other occupations and owing to permanent or accessional immigration of a part of the urban population into the countryside, its social structure becomes more and more heterogeneous (Barbič, 1991; Kovačič, 1995). In relation to this process the way of life of the countryside population is changing. Achievements of urban society are advancing rapidly into the rural areas; activities and services once typical for urban areas become more and more widespread in rural areas. Thus, rural population is getting similar to the urban one (Pečan, Ravbar, 1999). It can be expected that all these processes are influencing social behaviour of the countryside people and consequently their fertility behaviour as well.

¹ However, the correlation between decrease in total fertility rate in Slovenia and the decline in the share of farm population was high after the Second World War period.

1.2 Theoretical background

In now already classical textbook “Determinants of fertility in advanced societies” by Andorka (1978) a relationship between fertility behaviour and place of residence was characterised as a direct linkage. There is a fairly consistent correlation between urban or rural trait of the place of residence and fertility. The place of residence has a property of natural or man-made environment. In this sense highly populated, densely build-up areas heavily loaded with traffic are defined as urban areas. Correspondingly, there is little space left for parks, private gardens and other places where children can safely spend time outside their homes. On the other hand, the trait of rural area is determined by living predominantly in a one-family house with garden or in a relatively small apartment house. Life there is quieter and safer and children have plentiful space for playing outside their homes. According to Andorka, this ecological characteristic of urban-rural differential is also connected with different monetary costs and efforts necessary for raising and educating children that are much greater in urban areas than in the rural ones. Besides, he presupposed different preferences and different consumption alternatives in these two types of residence.

To understand urban-rural differences in fertility behaviour of advanced societies it is also necessary to take into consideration the social characteristics of these types of environment. Mackensen (1982) believes that one general theory of fertility that could adequately explain fertility behaviour in all societies and at all periods of time is neither possible nor justifiable. He is convinced that for this reason, every explanation, observation and research of fertility behaviour like any other social behaviour should proceed from the concept of specific structural and cultural characteristics of each society, which is the product of certain historical processes. Hoffnan-Nowotny (1987) who asserts that fertility behaviour of an individual is connected with structural and cultural characteristics of his/her micro, mezzo and macro social environment shares similar view. Importance of geographical variations in place or context in understanding fertility decision-making is further stressed by Boyle (2003).

Discussion concerning typical characteristics of urban and rural places is already a long present one. Classics of sociological thought like K. Marx, F. Tonnies, G. Simmel, (European representatives) and L. Wirth, R. Redfield (attached or influenced by the Chicago School) argued that strong distinctions exist between urban and rural societies. E.g. accordingly to Marx’s theorization, an individual living in urban place has universal chances to develop all his/her abilities while in the countryside he/she is bounded by constant reiteration of firmly established patterns of thinking and acting derived from direct dependence of men to nature. Quite on the contrary, Tonnies was convinced that life in the countryside (in *Gemeinschaft*) bounded together by a unity of wills and solidarity with its tradition and social order, gave an individual a supreme opportunity to

harmoniously live with others, while life in urban areas (in *Gesellschaft*), constituted by commodity exchange and rested on “union of rationale wills”, led to undermining genuine attachment between people and community (see Bonner, 1999 for more extensive overview of the urban-rural debate).

Extensive empirical research of the rural society evolved since the 1950s, showed that there is no clear rural-urban distinction. Concepts like community and locality did not wholly prove their justification. Fieldwork investigations (Williams, 1963, Bell and Newby, 1971, Pahl, 1965, 1966, Newby, 1978, 1980) revealed that there are no close, isolated, harmonious, functioning in traditional manners and closely kinship-bounded communities. They also showed that low population density or certain fixed settlement patterns were insufficient basis to distinguish between urban and rural places. Marsden (1998, 1999), focusing on restructuring of agriculture, drew attention to the varying degrees of influence and interaction of agricultural, residential and commercial interests in shaping differentiating rural areas that could be analyzed by modelling of typologies (see Mahon, 2001 for more extensive review of recent comprehension of rurality).

According to Bourdieu (2003) each classification of social world should take into consideration the principle of differentiation in order to theoretically construct empirical reality. Basically, this principle pertains to the distribution of the forms of power or the kinds of capital that vary according to the specific place and moment. This means that the set of agents or institutions which possess sufficient quantum of a specific capital (especially economic and cultural), that enables them to possess a dominant social position, conserve or transform the “exchange rate” between different kinds of capital through more or less administrative measures. Of course, this holds true for relationships between city and country as well. Characteristics of today’s urban and rural societies are undoubtedly strongly related to their specific position and mutual relationship in the last century and a half, which has often taken the framework of city domination over the countryside (Hays 1993). An urbanized society evolved out of the former rural society by exploiting the material and human resources of the countryside intensively, which led to considerable economic, social and political inequalities between these two settings. Along with this, cities changed or gave up thoroughly the culture and the way of life, which long prevailed in the countryside. However in this process all parts of countryside were not equally affected.

Hypothesis 1: on the basis of the above statements, we expected in our analysis that fertility behaviour of individuals is closely linked with economic and social characteristics of their life settings observed through the selected typology of rural areas.

Over the last decades, the social structure and culture of rural areas in Europe and other industrially developed countries has changed significantly. Due to massive abandonment of agriculture by a great part of rural

population and their engagement in other occupations, rural areas became multifunctional and multistructural (Djurfeldt, 1999). Migrations from urban to rural areas contributed to this heterogeneity to some extent as well. With rapid development of new technologies (information and computer sciences) and improved traffic connections among urban and rural areas, the entire societies became increasingly urbanised, "infected" with urban values and the urban way of life. Owing to these changes, it could be supposed that urban-rural difference in fertility behaviour is diminishing or even vanishing. However, changes in social behaviour do not occur quickly, or at the run of one single generation. According to Inglehart's (1989) theory of value changes they occur with the exchange of generations. This theory is based on two hypotheses:

1. deficiency effect: An individual most strongly appreciates the things that are relatively rare in his/her socio-economic setting.
2. socialisation effect: Value priorities of an individual are not direct reflections of his/hers present socio-economic setting but are the reflections of conditions in which he/she grew up.

Therefore it could be supposed that due to increased impact of globalisation (preferring urban life style) every younger rural generation will be more similar in its social and consequently fertility behaviour to urban population than previous rural generations were.

Hypothesis 2: in this respect we expected in our analysis smaller differences in fertility behaviour in younger generation of the respondents from different space settings than in the older ones.

Cohorts who voluntarily limited marital fertility, enforced demographic transition, which took place during the second part of the 19th and the first part of the 20th century in industrialised world, from high to low fertility. Their belief in marriage and family was still very strong, although they practised contraception from preventive reasons according to Van de Kaa (1999). They wanted to give their children a good start in life and stemmed towards limiting the number of children to be correspondent with that goal. They disciplined themselves to remain married even when love was lost. However, these features of "modern demographic behaviour" did not last very long. Since 1965, new demographic changes have been observed in many European countries: decrease of marriages, increase of cohabitations, postponement or abandonment of parenthood, increase in divorces and single parent households. These changes also denoted as the second demographic transition (Van de Kaa, 1987) presumably occur due to a shift of value orientation from a modern trend to post-modern that signifies further enforcement of the individual's free choice principle, which was introduced during the time of Renaissance and Enlightenment centuries ago. The motto of this trend is that individuals can and should make their own choices. "Post-modern demographic behaviour" of contemporary reproductive cohorts presumably corresponds to the

individualistic "lifestyle, where it is understood that sex and marriage/union are no longer closely related, and that contraception is only interrupted to have a self-fulfilling conception" (Van de Kaa, 1999:31). This new pattern of behaviour is seemingly reflected in the changes of the life course of young generations – earlier entry into first sexual intercourse but later achievement of economic and housing autonomy and formation of own families (Iedema et al., 1997; Cordon, 1997; Nave-Herz, 1997).

Hypothesis 3: in this vein in our analysis we supposed considerable differences in the life course, especially during the transition into adulthood between younger and older generations of the respondents from different space settings.

2 Methods

2.1 Data

The analysis was based on the Family and Fertility Survey data collected in Slovenia between December 1994 and December 1995. A representative area sample of the inhabitants of the Republic of Slovenia in their reproductive age period (i.e. aged 15 to 45) of both genders was drawn. Face to face interviewing was used as the data collection method. Realised sample consisted of n=4.559 respondents. The data were weighted according to survey design and population adjustment based on a specified set of socio-demographic variables: gender, age and size of the settlement (see Obersnel Kveder et al., 2001 for more detailed description of the data set). Considering our objective to explore differences between generations, two birth cohorts were emphasized in the present analysis:

- respondents aged 20-24 years of both genders - the "younger" generation at the entrance to their reproductive period and also representing contemporary fertility patterns,
- respondents aged 40-45 years of both genders - the "older" generation at the end of their reproductive period and also presenting the immediate past fertility patterns.

The main interest of the present article was in the variables concerned with the timing of childbearing as well as other events relating to the entry of an individual into adulthood such as sexual debut, partnership history, leaving parental home, education and employment spells. All these events were measured as retrospective event histories. For all strictly reproductive events (i.e. childbearing, use of contraception), the age at the first sexual intercourse was taken as the threshold of becoming at risk, while for all others (i.e. partnership, education, employment, and leaving parental home) the threshold was the respondents' birth. Right censoring was determined by the date of the interview. Besides, the variables describing respondent's preferences, values, attitudes and status characteristics were considered. Some of them: household structure, attitudes towards

abortion, gender roles and marriage were constructed from the set of other variables from the survey data set.

The variable 'Household structure' was derived from the household grid. It has 13 distinct attributive values depicting various household types in which the respondents can live:

- 1 alone
- 2 with a partner
- 3 with a partner and others (e.g. his/hers parents and siblings, partner's parents and siblings, other relatives)*
- 4 with a partner and children
- 5 with a partner, children and others*
- 6 with children
- 7 with children and others*
- 8 with parents and siblings
- 9 with parents, siblings and other relatives
- 10 with one parent
- 11 with non-relatives
- 12 with other relatives
- 13 with other relatives and non-relatives

The variable 'Abortion' was derived by taking into account "approval" answers from the set of the following dichotomous statements:

- An abortion is approved/not approved when mother's life is in danger due to pregnancy.
- An abortion is approved/not approved when the risk of the birth of an abnormal child is great.
- An abortion is approved/not approved when a woman is not married.
- An abortion is approved/not approved when a married couple does not want to have another child.
- An abortion is approved/not approved when a woman does not want to have a child at the time being.

The variable 'Gender roles' was constructed by taking into account answers "strongly agree" and "agree" from the Likert 5- item scale of the following attitudes:

- An employed mother can create as warm and safe relationship with her children as a mother who is not employed.
- Employment is the best way for a woman to achieve independence.
- Being a housewife is as fulfilling for a woman as being employed.
- Both man and woman should equally contribute to their household's budget.
- Preschool child would most probably suffer if his mother was employed.
- It is acceptable for a woman to be employed, but what most of women really want is home and children.

The variable 'Marriage vis-à-vis Cohabitation' was constructed by taking into account responses "very favourable" and "favourable" from the Likert 5- item

scale regarding advantages of cohabitation over marriage to achieve the following aims:

- general happiness,
- economic security,
- friendly relationship with others,
- personal freedom,
- stable relationship,
- having children,
- social acceptance.

Following the aim of this article, fertility behaviour should be put in the perspective of the contextual micro, mezzo and macro level variables. Microenvironment is defined as the immediate living surrounding of the individual varying from their family, household to the neighbourhood. Mezzo environment is by definition broader than the micro and thus can encapsulate a variety of geographical units from the settlement to the municipality and region. Macro socio-economic context is usually associated with the national level indicators. This article focuses on the use of one exemplary indicator measured at mezzo level; the respondent's place of residence determined by the basis of Slovenian Census in 1991. In its essence, the indicator reflects the urban-rural dichotomy and defines 4 possible types of living surroundings: urban, suburban, typical rural and rural depopulation areas (Kovačič et al. 2000, Kovačič, et al. 2002):

- Those places that have an urban management character according to the space planning documents have been classified as *urban* environments. Additionally, the function of centrality of each geographical unit was considered as criteria of demarcation between urban and rural areas. Slovenian geographers classify settlements into seven groups according to their centrality (Vrišer, 1998). Those settlements with the centrality index between 7 and 3 were also defined as *urban* environments. The settlements with the lowest centrality index 3 were additionally bounded by the minimal size of 3,500 inhabitants. The share of population living in this type of settlements is estimated at 39,20 percent.
- All local communities (*slo. krajevne skupnosti*) with the density of the population greater than 200/km² were considered as *suburban*. In addition, the areas with the density lower than 100/mk², with the index of population growth in the period 1981/91 greater than 110, were also included in this type of settlements. The share of population living in this type of settlements is estimated at 14,80 percent.
- All local communities with the long-term (1961/91) and short-term (1981/91) population growth index below 97,5 were considered as *depopulation* areas. In addition, local communities with the non-negative short-term population growth and with the index of population ageing above the absolute demographic threshold (i.e. 72) were also considered as depopulating. The share of population living in this type of settlements is estimated at 14,87 percent.

- All rural areas between suburban and depopulation were considered as *typically rural*. The share of population living in this type of settlements is estimated at 31,22 percent.

Additional analysis which applied the above typology (Perpar, Kovačič, 2002) revealed that these types of rural settlements significantly differ among each other in terms of socio-economic development, infrastructure and natural resources. Taking into account several indicators pertaining to statistical data from 1991, suburban areas are in the most favourable position and depopulation areas are in the least favourable one. Considering the employment structure of population, the highest share of population in suburban and typically rural areas works in secondary sector, whereas in depopulation areas population mostly works in the primary sector². Daily migrations additionally indicate the level of engagement in gainful employment. Its share is again the highest in suburban areas (770 per 1000 inhabitants) and the lowest in depopulation areas (540 per 1000 inhabitants). Further, significant differences among the areas are also observed concerning the education. The highest level of education is reached in suburban areas (43 percent of inhabitants finished at least high school), whereas in typically rural and especially in depopulation areas this share is considerably smaller. The same picture is indicated by the proportion of students per 1000 inhabitants; it is again the most favourable in suburban areas (25 students per 1000 inhabitants), less favourable in typical rural areas (21 to 25 students/1000 inhabitants) and the least favourable in depopulation areas (20 students/1000 inhabitants). Furthermore, indicators pertaining to economic situation confirm already indicated differences. E.g. density of business entities that indicates economic development of the area is the most favourable again in suburban settlements with more than 14 business entities per 1000 inhabitants, less favourable in typical rural areas (12 to 14 business entities per 1000 inhabitants) and the least favourable in depopulation areas (less than 12 business entities per 1000 inhabitants). Considering the infrastructure, the analysis (Perpar, Kovačič, 2002) demonstrated that all areas are relatively well equipped with basic infrastructure, but the best equipped are the suburban ones. In typical and depopulation areas the population is still frequently faced with the problems of drinking water supply, unsettled canalisation and purifying plants and maintenance of local roads. An additional problem of the depopulation areas is abandonment of farming and consequently the forest over-growing.

² The share of farm population in depopulation areas presents 200 per 1000 inhabitants, in typical rural areas this proportion counts 120 per 1000 inhabitants and in suburban areas only 40 per 1000 inhabitants (Perpar, Kovačič, 2002).

Each respondent from the survey was ascribed a settlement type according to his/her residence at the time of the interview. The key for information matching was the local community that could be matched to the survey data as well as the settlement typology specification. The analysed sub-samples were as follows:

Table 1: The size of sub-samples (N).

Settlements type	20-24	40-45
Urban	192	302
Suburban	109	182
Typical rural	204	295
Depopulation rural	75	131

2.2 Analysis

Bivariate associations were analyzed using either the analysis of contingency tables either using comparisons among means. Timings of observed events were analyzed through event history models (Allison, 1995), which enable the estimation of the differences in the individuals' life course. Since only the distribution of the probability of time (T) was taken into account, it was described through cumulative distribution function perspective (survivor function):

$$S(t) = \Pr\{T > t\} = 1 - F(t) \quad (1)$$

The result of the survivor function is a probability that an individual "survives" in the process beyond time t. It is defined on the interval from 0 to 1. Life-table method was used for the estimation of the survivor function. The estimate is calculated using the conditional probability of failure (i.e. the probability for an event within a certain interval, given that an individual made it to the start of interval):

$$S(t_i) = \prod_{j=1}^{j-1} (1 - q_j) \quad (2)$$

where q_j is the conditional estimation of failure and is calculated as the ratio of number failed over the effective sample size. Major events in a life history were depicted by the means of calculating the quartiles of the survivor function.

2.3 Results

In order to evaluate the importance of contextual information in explaining fertility and family behaviour, the major moments of the reproductive life period, structural and cultural characteristics of individuals were compared against the regional typology.

At first, the processes of entry into the first, second and third parenthood were compared among the regions, taking into account both generations together (Figure 1, 2 and 3). The major differences that can be observed at the first birth are related to the differentiation of the urban areas from the rest. As survivor functions show, inhabitants of the urban areas tend to delay the birth of the first child more than the inhabitants of the other regions. There are only 15 percent of women from suburban and rural areas that are still childless 10 years

after their sexual debut, while the proportion of nullipara women in the urban areas is above 30 percent. At the second birth, the differentiation of the urban areas from the rest still prevails; with more than 30 percent of inhabitants that did not experience the birth of the second child 10 years after the entry into the first parenthood. However, in this case the suburban areas also differ significantly from typical rural and depopulation rural areas; approximately 24 percent of its inhabitants did not experience the birth of the second child 10 years after the first birth, whereas in typical rural and depopulation rural areas only 18 and 14 percent of inhabitants did respectively. In the case of the third birth, the difference between areas got the character of polarity. The share of inhabitants that experienced the third birth decreased significantly in all regions in comparison with the share of inhabitants that experienced the second birth, but particularly in urban and suburban ones, where more than 80 percent of dwellers did not experience that event at all. In typical rural and depopulation rural areas the comparable shares are 69 and 66 percent respectively.

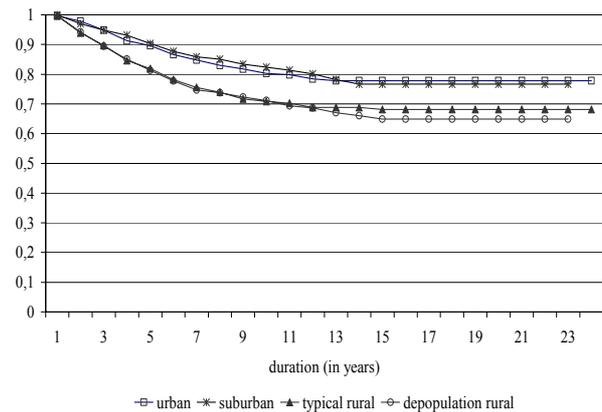


Figure 3: Time to the third birth.

Table 2 shows times in years from the first sexual intercourse to the first, second and third parenthood for both generations separately. The older generation relates a more homogenous picture than the younger one, where the differences among the rural and urban areas are more transparent. The most evidently, the 20-24 urban generation postpones the parenthood in comparison to the urban older one and their rural coevals as well. Among the younger generation in urban areas, 25 percent enters into first parenthood more than 4 years later (7,1 years in comparison to 2,9 years) than the first quarter of the older generation and approximately 3 years later than their counterparts in typical and depopulation rural areas. Furthermore, the postponement of parenthood among younger urban generation is manifested also in the absence of any childbirth at 50th and 75th percentile and consequently any second or third birth at all. Results pertaining to the younger generation also show that the first births reach their maximum level the fastest in typical rural areas.

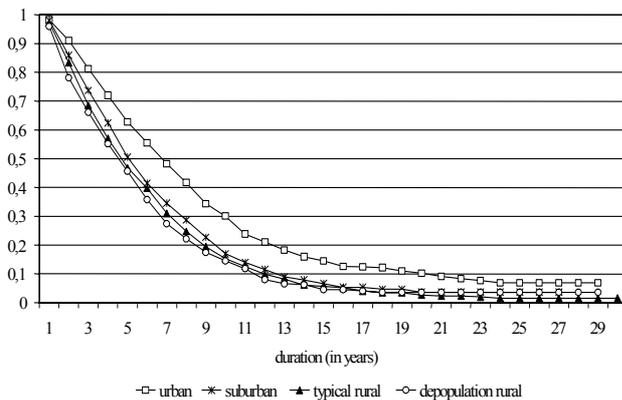


Figure 1: Time to the first birth.

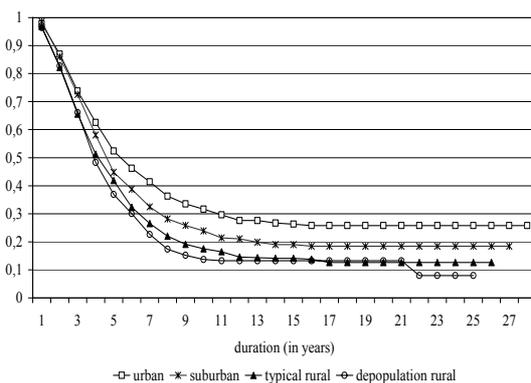


Figure 2: Time to the second birth.

Table 2: Time to parenthood (in years).

	First Birth		Second birth		Third birth	
	20-24	40-45	20-24	40-45	20-24	40-45
25th percentile						
Urban	7.1	2.9	-	3.0	-	-
Suburban	4.3	2.3	3.0	2.3	-	-
Typical rural	3.3	2.2	2.6	2.3	3.2	6.0
Depopulation	3.9	2.4	3.6	2.0	-	3.8
50th percentile						
Urban	-	7.2	-	5.1	-	-
Suburban	11.9	4.3	4.3	4.4	-	-
Typical rural	6.6	6.6	4.2	3.7	-	-
Depopulation	6.8	4.1	-	3.3	-	-
75th percentile						
Urban	-	10.3	-	-	-	-
Suburban	-	7.7	6.9	9.3	-	-
Typical rural	8.8	7.0	-	6.0	-	-
Depopulation		6.5	-	6.4	-	-

Although different patterns of entering into parenthood can be observed within generations, substantial differences in the 40-45 generation can be observed in the 75th percentile values. At this level the second births are rare among urban dwellers. There are also scarcely any families with three or more children. The third births are presented only at the level of 25th percentile among families in typical and depopulation rural areas. In general, the intervals between childbirths are considerably shorter in rural than in urban areas, what has an important implication on the overall number of children in each of the regions.

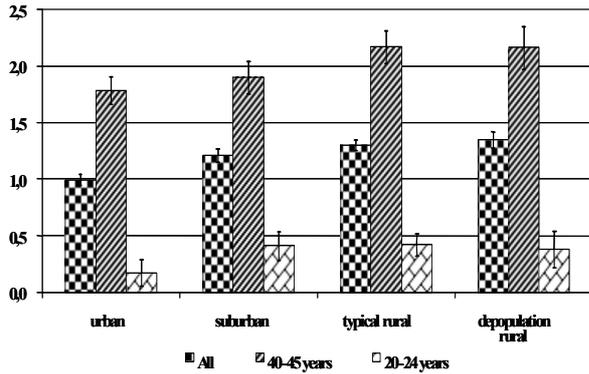


Figure 4: Actual number of children (means with confidence intervals).

As a consequence of these events' and spells' patterns, in general, families in both types of rural areas have statistically significantly ($F=27.595^{***3}$, Bonferroni post-hoc mean difference tests for Urban compared to other types are -0.216^{***} , -0.308^{***} and -0.358^{***} respectively) more children than families in urban areas have (Figure 4). Suburban areas do not differ significantly from urban or the rest of rural areas, however they are more similar to urban than rural areas. In the case of 20-24 generation, the differentiation is significant among urban and typical rural areas ($F=8.562^{***}$, Bonferroni post-hoc mean difference tests for Urban compared to other types are -0.240^{**} , -0.248^{***} and -0.209 respectively). It is also observed that in this generation, the number of children slightly decreases in depopulation areas in comparison with typical rural and suburban areas.

Considering the perception of the ideal number of children per family, there are no great differences among rural and urban areas (Figure 5). The overall perceived ideal family size is above two children, though it is a bit diverse concerning the region type and age. In particular, the members of the younger generation from urban and suburban areas want more children than their older counterparts do. In this regards they are quite similar to their rural counterparts. Of course, more conclusive observations concerning the match between preferred and actual number of children can only be drawn from the older generation, which more or less already finished its reproductive period. It seems that among all older

respondents, those from suburban areas came the most near to their desired number. Among older generation, concerning preferred number of children, urban and suburban areas create one pattern (lower number) and typical and depopulation rural areas create the other one (higher number). This observation is consistent with previous as well as with subsequent findings.

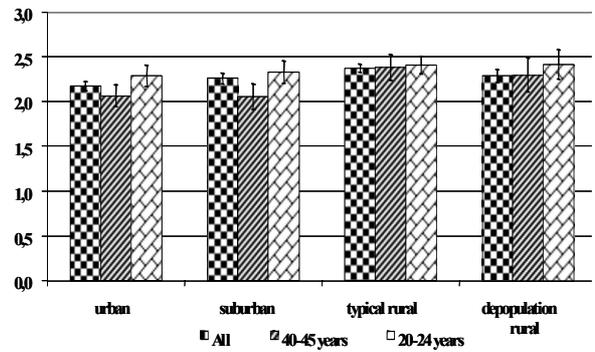


Figure 5: Preferences towards the number of children (means with confidence intervals).

Table 3 distinctly shows extremely interlaced differences in the ages of entering certain life events among the two observed generations and the four area types. The prevailing pattern of sequence of events can be described as starting with the entry into sexual life, finishing education, getting the first job, leaving the parental home and as the last step, entry into partnership union. The entry into partnership union is consistently the last event to be experienced by any observed group. Some resemblances can be observed between both urban and suburban areas, as well as between typical rural and depopulation areas. In rural areas the conclusion of education process tends to precede the sexual debut.

Table 3: Age at life events (in years).

	First sexual intercourse		First partnership		The end of education		Leaving parental home		First employment	
	20-24	40-45	20-24	40-45	20-24	40-45	20-24	40-45	20-24	40-45
25th percentile										
Urban	16.0	17.9	22.3	20.8	17.8	16.9	19.3	18.4	19.0	18.2
Suburban	16.0	17.0	20.5	20.0	17.2	15.8	19.7	18.9	18.5	17.7
Typical rural	17.0	17.0	20.5	20.6	15.9	14.5	19.3	18.3	18.0	17.3
Depopulation	17.0	17.0	20.1	20.3	15.9	14.5	18.9	18.0	17.9	17.2
50th percentile (Median)										
Urban	17.0	18.9	24.8	23.3	20.6	20.5	22.7	20.3	21.3	19.8
Suburban	18.0	18.0	-	22.1	18.3	18.3	24.4	20.8	19.8	18.8
Typical rural	18.0	18.0	22.9	22.5	17.4	17.5	22.8	20.5	18.8	18.5
Depopulation	18.0	18.0	23.3	22.4	17.1	16.7	20.4	19.3	18.6	18.6
75th percentile										
Urban	18.0	20.0	-	26.7	-	25.8	-	24.3	-	22.5
Suburban	18.0	20.0	-	25.4	23.7	24.9	-	24.8	23.9	21.5
Typical rural	19.0	20.0	-	25.3	19.1	22.7	-	24.7	20.2	21.2
Depopulation	19.0	20.0	-	24.8	18.8	18.4	-	24.0	20.2	21.6

The main difference between the two generations irrespectively of the region is in the earlier sexual debut and the postponing of all other events of the younger population, especially setting up a partnership union and an independent life away from parents, which is most evident in the young urban generation. The end of

³ $p < 0.01$ - ***, $0.01 < p < 0.05$ - **, $0.05 < p < 0.10$ - *

education process tends to differentiate most strongly among generations and regions. The older generation in depopulation areas is the first to finish their education, whereas the young urban generation is the last. As in case of entry into parenthood, the intervals between above listed events which are considerably shorter in rural areas than in urban ones, additionally contribute to a different fertility level in observed areas.

Table 3 can be observed in relation to the contraceptive use (Figure 6). Among the younger generation, the most common contraception method used at the time of the interview was hormonal contraception following by condom and withdrawal. The rarest method used was injection. Quite substantial differences can be observed among different regions. Depopulation areas have a very low condom or diaphragm usage (6,7 percent) compared to other regions (14 percent in urban to 23 percent in suburban areas), but on the other hand, they have the highest usage of hormonal contraception (33,3 percent). However, the key finding is higher proportion of non users among typical rural dwellers (40 percent) than in other areas (24 to 33 percent). Thus, the differences among the regions in relation to the entries into certain life events, as well as in the birth of the first child, are reflected in variations of contraceptive use.

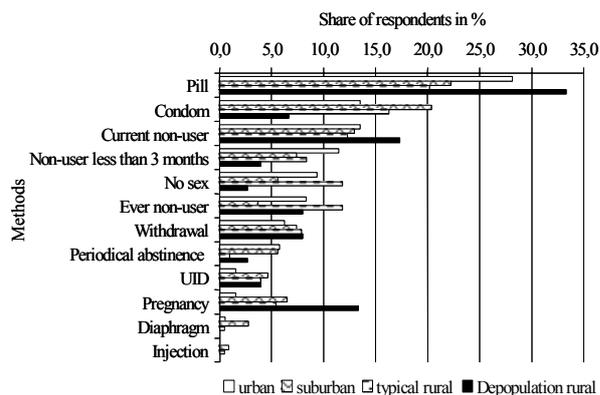


Figure 6: Current usage of contraception (20-25 years).

Explanation for the difference in the number of children can also be indicated in relation with the attitudes towards abortion. The results in Figure 7 show, that abortion is not a very acceptable phenomenon among the respondents. However, there is a small, but a consistent downward trend in acceptability of abortion from urban to depopulation areas. Abortion tends to be more often considered fairly unacceptable in rural areas, whereas in urban areas, the average attitude is more pro-choice oriented. In general, there are no statistically significant differences between the two observed generations, except in typical rural areas where the young show significantly greater ($F=10.370^{***}$, Bonferroni post-hoc mean difference tests for Typical Rural compared to urban is -0.613^{***}) opposition towards the abortion as the free choice act than the older generation does.

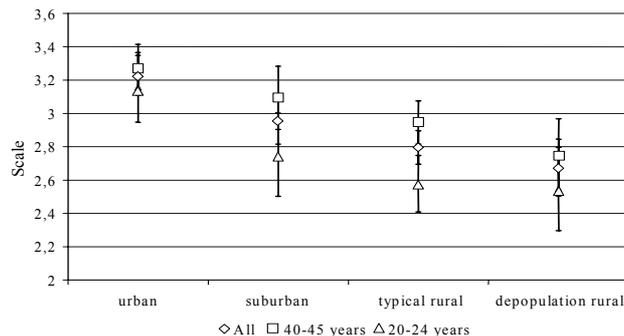


Figure 7: Attitudes towards abortion (means with confidence intervals).

Attitudes towards abortion are in line with the respondents' expressed religiosity. Figure 8 shows quite high and significant differences across observed regions. Only one third of urban dwellers defined themselves as being religious, while more than a half of the inhabitants of the other three regions did. Typically, rural areas have the highest proportion of religious people. Within the younger generation, these differences are even more expressed (χ^2 sig. $p=0.000$; Adjusted Standardized Residuals for Urban 20-24 are -5.6 "Religious" and 6.4 "Not Religious").

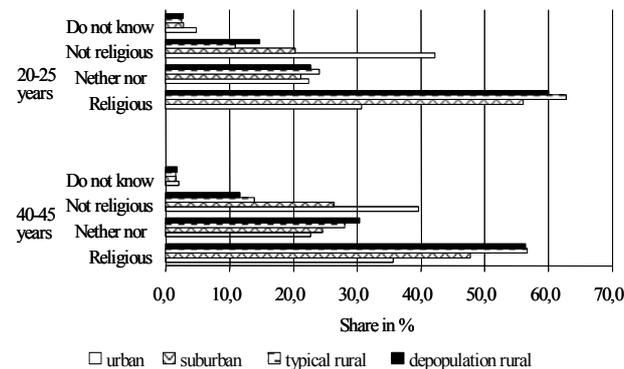


Figure 8: Religiosity.

The results concerning the attitudes towards marriage and gender roles reveal a rather different picture. As Figure 9 shows, irrespective of the region and the age among the respondents, there are no significant differences in the preference of either cohabitation or institution of marriage. Only the 40-45 generation in urban areas shows a significant tendency ($F=8.620^{***}$; Bonferroni post-hoc mean difference tests for Urban compared to Suburban and Depopulation Rural urban are -0.2379^{**} and -2562^{**} respectively) towards favouring the cohabitation over marriage in comparison to other regions, whereas younger generation everywhere is more inclined towards cohabitation than marriage. These results are consistent with the share of respondents living in cohabitation unions: in urban areas the share is 15 percent, in suburban 10,8, in typical rural 12,4 and in depopulation rural 13,7 percent. All respondents also share the same tendency in attitudes towards supporting

the equality of gender roles (Figure 10). There are no major differences between generations or regions.

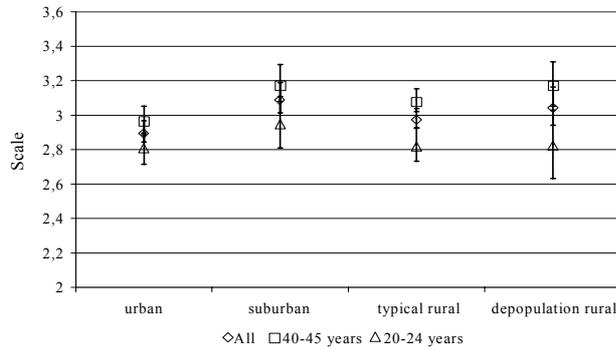


Figure 9: Marriage vis-à-vis cohabitation (means with confidence intervals).

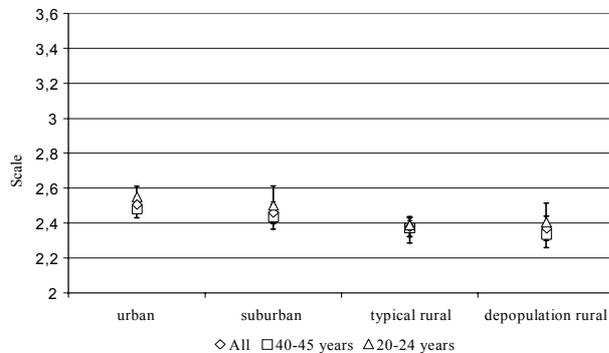


Figure 10: Gender (means with confidence intervals).

Substantive differences among regions (χ^2 sig. $p=0.000$) are revealed in the household composition (Figure 11). The largest differences are related to the rate of nuclear and extended families - of orientation and procreation - which represent two predominant household composition types in all regions and both generations. The extended family in depopulation regions of Slovenia represents one third of households, while in the urban areas, this proportion is less than 10 percent. Nuclear family is the most representative composition type where the majority of the respondents live. However, in this regards the difference among the regions is significant (Adjusted Standardized Residuals for Depopulation Rural are -5.4 “Partner and Children” and -2.5 “Parents and siblings”); depopulation areas are just above the 50 percent mark, whereas in other regions the share of this type of households is around 70 percent, being the highest in suburban areas (over 75 percent). Although one-parent family households are relatively rare, they can only be observed in the urban and suburban areas (6,9 – 4,1 percent), while in the rural areas they are more an exception than a rule (1,9 - 0,9 percent). Very similar proportions among regions hold true for those living alone. Thus, the variation in number of children in various areas is reflected in the household composition and consequently in their life style.

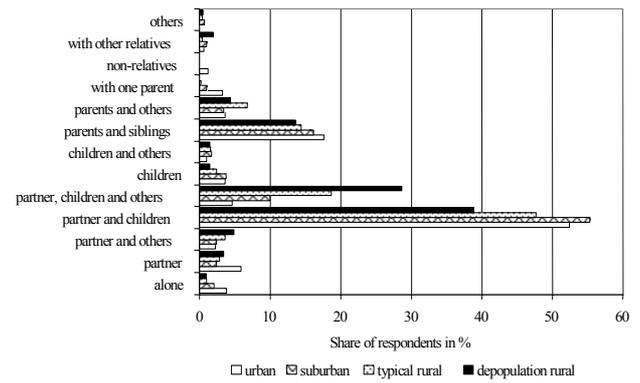


Figure 11: Household composition.

2.4 Conclusions

Based on the presented results, the first and the third hypothesis can fully be confirmed. The analysis revealed that fertility behaviour of Slovenian population is significantly related to the socio-economic context of diverse socio-geographical regions; the more developed is one setting in terms of favourable and diverse economic activities, available infrastructural capacities and social services, the lower is fertility of its dwellers. Nevertheless, the application of regional typology also revealed that relations of urban-rural division are not uniform and static. Population of urban areas with the pattern of postponement of first births considerably differs from the other regions, whereas at the second and even more at the third birth, the population of suburban areas is getting gradually more similar to the urban ones. The typical rural areas are the most stable in relatively high fertility pattern, whereas in depopulation areas, this pattern is no longer very firm, most probably due to less favourable developmental conditions. Changing similarities and differences among the regions are particularly related to behavioural patterns of the younger generation. The results revealed that tempo and sequences of events in entering into adulthood changed significantly between generations. In older generation among all regions, the life events, that normatively precede parenthood, followed each other quicker than in the case of the younger generation. In older generation among the regions the variations in transition to adulthood certainly exist. But contrary to our expectations, formulated in the second hypothesis, these variations are even more pronounced among the younger generation, particularly due to distinctive behaviour of urban youth. According to the theory of second demographic transition, this group shows an indicative pattern of post-modern demographic life course both through their behaviour and expressed attitudes, whereas rural young generation, for the most part those living in typical rural areas, mainly follow the pattern of their older counterparts. As our analysis indicates, variations in fertility behaviour among the regions and generations can be more pertinently explained by social structural factors (education, household type) than values and attitudes pertaining to the post-modern life patterns.

The obtained results call for greater attention to more contextualised approaches in demographic research. Our analysis demonstrated that rural setting as a social space is not a homogeneous category, but a grouping of various structural and cultural characteristics that might specifically determine fertility behaviour. By neglecting this reality, too much information needed for better understandings of variations in fertility behaviour could be lost unduly. The residues of the past urban-rural differences in fertility behaviour are still present today despite the effects of global economic and social trends that carry structural and ideational changes into the countryside and will, as our analysis indicates, also remain in the near future. From this point of view, it was unreasonable that international project Fertility and Family Surveys in ECE countries under supervision of United Nation gave so little attention to urban-rural dimension. Among 22 European countries included in this project, only five of them (Poland, Estonia, Lithuania, Switzerland and Slovenia) took into consideration the respondents' place of residence. In the new Generations and Gender Programme (GGP), which started in the year 2000, the need for more contextual approach is fully considered; all micro, mezzo and macro levels are taken into account. On that basis pertinent results are expected. Hopefully, Slovenian researchers will also have the opportunity to join these endeavours.

The observed differences in fertility behaviour patterns among generations living in various geographical regions with different socio-economic characteristics also call for diversified actions of population policies. In the near past it was anticipated that family and social policy measures would contribute to uninterrupted population reproduction in Slovenia the most successfully. However, as our results indicate, population policy also has to integrate space and regional measures in to its field of actions; i.e. urban and rural developmental programmes that will consider different every day's life conditions and needs of people living in particular space setting. To scientifically support such actions, a new research data, e.g. in the frame of GGP, that will provide the picture of the present generations of reproductive age from various viewing angles, are urgently needed.

References

- [1] Allison, P.D. (1995) *Survival Analysis Using The SAS System. A Practical Guide*. Cary. SAS Institute Inc.
- [2] Andorka, F. (1978) *Determinants of Fertility in Advanced Countries*. Methuen & Co. Ltd. London.
- [3] Barbič, A. (1991) K "novemu" pojmovanju ruralnosti (Towards a "New" Comprehension of Rurality) In: Barbič, A. (ed.) *Prihodnost slovenskega podeželja. Prostor, prebivalci, gospodarske dejavnosti (A future of Slovenian Countryside. Space, Inhabitants and Economic Activities)*. Dolenjska založba. Novo mesto.
- [4] Bell, C., Newby, H. (1971) *Community Studies*. Allen and Unwin Ltd. London.
- [5] Bonner, K.M. (1999) *A Great Place to Raise Kids. Interpretation, Science, and the Urban-Rural Debate*. McGill-Queen's University Press, Montreal & Kingstone · London · Ithach.
- [6] Boyle, P. (2003) Population geography: does geography matter in fertility research? *Progress in Human Geography*, 27(5): 615-626.
- [7] Bourdieu, P. (2003) *Practical Reason. On the Theory of Action*. Polity Press. Cambridge.
- [8] Coale, A. J., Watkins, S. C.(1986) *The Fertility Decline in Europe*. Princeton University Press. Princeton. New Jersey.
- [9] Cordon, J. A. F. (1997) Youth residential independence and autonomy: A comparative study. *Journal of Family Issues*. 18 (6): 576-607.
- [10] Černič Istenič, M. (1994) *Rodost v Sloveniji. (Fertility in Slovenia) Znanstveno in publicistično središče*. Ljubljana.
- [11] Černič Istenič, M., Obersnel-Kveder, D., Kveder, A. (2000) Differences in contraceptive behaviour of men and women in Slovenia regarding their partnership and parenthood history. Paper presented at the Flagship Conference, Brussels, May 29-31.
- [12] Černič Istenič, M., Kveder, A. (2000) Urban-rural differences in fertility behaviour in Slovenia. Paper presented at X. World Conference of Rural Sociologists, Rio de Janeiro; July 30 - August 5.
- [13] Djurfeldt, G. (1999) From CAP to CAR? New Directions in European Rural Sociology. <http://www.soc.lu.se/esrs>.
- [14] Hays, S.P. (1993) From the history of the city to the history of the urbanized society. *Journal of Urban History*. 19(4): 3-23.
- [15] Hoffman-Novotny, H.-J. (1987) *The Future of the Family. European Population Conference. Issues and Prospects. Plenary Séances*. Central Statistical Office of Finland. Helsinki. pp. 113-199.
- [16] Iedema, J., Becker, H. A., Sanders, K. (1997). Transition into independence: A comparison of cohorts born since 1930 in The Netherlands. *European Sociological Review*, 13 (2): 117-137.
- [17] Inglehart, R. (1989). *Kultureller Umbruch*. Campus Verlag. Frankfurt/New York.
- [18] Javornik, J.(1999). *Slovenija v 90-ih. Prebivalstvo (Slovenia in the 1990s)*. In: Hanžek, M. (ed.) *Poročilo o človekovemu razvoju Slovenije 1999 (Human Development Report for Slovenia 1999)*. Urad R Slovenije za makroekonomske analize in razvoj. Ljubljana. pp. 65-67.
- [19] Kaa, Van de, D. (1987) Europe's second demographic transition. *Population Bulletin*, 42(1): 1-47.
- [20] Kaa, Van de, D. (1999) Europe and its population: the long view. In: Kaa, Van de, D., Leridon, H., Gesano, G., Okolski, M. (eds.) *European Populations Unity in Diversity. European Population Studies no.6*. Kluwer Academic Publishers. Dordrecht/Boston / London. pp. 1-49.
- [21] Klemenčič, V. (2002) *Procesi deagrarizacije in urbanizacije slovenskega podeželja (Processes of deagrarisation and urbanisation of Slovenian*

- countryside). Podeželje na prelomu tisočletja (Countryside at the break of the century/ Mednarodna znanstvena geografska konferenca (International scientific geographic conference), Ljubljana, 19.-21. September 2001. Ljubljana. Oddelek za geografijo Filozofske fakultete. Accessible on: (http://www.ff.uni-lj.si/oddelki/geo/publikacije/dela/files/Dela_17/Vladimir_Klemencic.pdf)
- [22] Kovačič, M., Perpar, A., Gosar, L. (2002) Členitev podeželja v Sloveniji (Differentiation of countryside in Slovenia) *Sodobno kmetijstvo*. 35(2): 62-66.
- [23] Kovačič, M., Gosar, L., Fabijan, R., Perpar, A. (2000). Razvojno – tipološka členitev podeželja v Republiki Sloveniji (Developmentally – Typological Differentiation of Rural Areas in the Republic of Slovenia.). Institute of Agricultural Economics, Biotechnical Faculty, University of Ljubljana.
- [24] Kovačič, M. (1995) Idejne in teoretske osnove za razvoj podeželja (Ideal and theoretical basis for the countryside development) In: Kovačič, M. (ed.) *Izhodišča, sestavine in problemi celovitega razvoja podeželja v Sloveniji (Starting-points, components and problems of integral countryside development in Slovenia)* Biotechnical Faculty, University of Ljubljana, pp. 3-20
- [25] Livi-Bacci, M. (1986) Social-Group Forerunners in Fertility Control in Europe. In: Coale, A. J., Watkins, S. C. (ed.): *The Fertility Decline in Europe*. Princeton University Press. Princeton. New Jersey, pp.182-200.
- [26] Mackensen, R. (1982) Social Change and Reproductive Behaviour – on Continuous Transition. In: Hohn, C., Mackensen, R. (eds.) *Determinants of Fertility Trends: Theories Re-examined*. Ordina ed.. Liege. pp. 251-279.
- [27] Mahon, M. (2001) Reconceptualising rural change. A literature review. Paper presented at XIX European Congress for Rural Sociology. Dijon, September 3-7.
- [28] Malačič, J. (1990) Načrtovanje zdravega potomstva – odgovornost posameznika do družbe (The planning of healthy future generations – the individual's responsibility towards society). *Zdravstveni vestnik*. 59(7-8): 349-354.
- [29] Marsden, T. (1998) New rural territories: Regulating the differentiated rural space. *Journal of Rural Studies*. 14(1): 107-117.
- [30] Marsden, T. (1999) Rural futures: The consumption countryside and its regulation. *Sociologia Ruralis*. 39(4): 501-520.
- [31] Nave-Herz, R. (1997) Still in the nest: The family and young adults in Germany. *Journal of Family Issues*. 18(6): 671-689.
- [32] Newby, H. (1978) *Property, Paternalism and Power: Class and Control in Rural England*. Hutchinson. London.
- [33] Newby, H. (1980) *Urbanisation and the Rural Class Structure: Reflections on a Case Study*. In *The Rural Sociology of Advanced Societies: Critical Perspectives*. Croom Helm. London.
- [34] Notestein, F. (1953) Economic problems of population change. In: *Proceedings of the Eight International Conference of Agricultural Economists*. London. Oxford University Press. pp. 13-31.
- [35] Obersnel Kveder, D., Kožuh Novak, M., Černič Istenič, M., Šircelj, V., Vehovar, V., Rojnik, B. (2001) *Fertility and Family Surveys in Countries of the ECE Region – Standard Country Report, Slovenia*. Economic Studies No. 10r, UNECE, UNFPA, New York/Geneva.
- [36] Pahl, R.E. (1965) *Urbs in Rure: The Metropolitan Fringe in Hertfordshire*. Wiedenfeld & Nickelson. London.
- [37] Pahl R.E. (1966) The rural-urban continuum. *Sociologia Ruralis*. 6(3): 299-329.
- [38] Pečan, J., Ravbar, M. (1999) *Regionalni razvoj Slovenije (Regional Development of Slovenia)*. In: Hanžek, M. (ed.) *Poročilo o človekovemu razvoju Slovenije 1999 (Human Development Report for Slovenia 1999)*. Urad R Slovenije za makroekonomske analize in razvoj. Ljubljana. pp. 31-63.
- [39] Perpar, A., Kovačič, M. (2002) Razvojno stanje, značilnosti in problemi podeželskih območij v Sloveniji (Development situation, characteristics and problems of rural areas in Slovenia). *Sodobno kmetijstvo*. 35(2): 52-61.
- [40] Sharlin, A. (1986) Urban-Rural Difference in Fertility in Europe during the Demographic Transition. In: Coale, A. J., Watkins, S. C. (ed.): *The Fertility Decline in Europe*. Princeton University Press. Princeton. New Jersey, pp.234-257.
- [41] *Statistical Yearbook (2003)* Statistical Office of Republic of Slovenia. Ljubljana.
- [42] Šircelj, V. (2007) Fertility and educational attainment in Slovenia. *Anthropological Notebooks*. 13(2): 11-34.
- [43] Šircelj, V., Ilič, M., Kuhar, A., Zupančič, M. (1990) *Projekcije prebivalstva R Slovenije 1986-2006 (Population projection of republic of Slovenia 1986-2006)*. Statistical Office of Republic of Slovenia. Ljubljana.
- [44] Šircelj, V. (1991) *Determinante rodnosti v Sloveniji (Determinants of fertility in Slovenia): Doctoral Dissertation*. Faculty of Arts, University of Ljubljana.
- [45] Vrišer, I. (1998) *Središčna (centralna) naselja (Central (capital) settlements)*. In: *Geografski atlas Slovenije*. Ljubljana, DZS. pp. 308-309.
- [46] Williams, W.M. (1963) *A West Country Village: Ashworthy: Family, Kinship, and Land*. Routledge Kegan Paul. London.

Late Fertility Trends in Europe

Janez Malačič
 University of Ljubljana, Faculty of Economics
 Kardeljeva ploscad 17, Ljubljana
 E-mail: janez.malacic@ef.uni-lj.si

Keywords: modern demographic regime, fertility trends, late age-specific fertility, Europe

Received: March 25, 2008

European continent has been for-runner in the demographic transition process. Age-specific fertility patterns had changed tremendously during this process in Europe. However, it is still mainly concentrated in the female age group 20 – 34 years. Women and couples tend to limit fertility in marginal age groups of female's reproductive period. In this paper the author addresses late age-specific fertility trends in selected European countries. Age – specific fertility is relatively low after the age of 35 in modern demographic regime. These rates show very low fertility in the age group 40 – 44 years and only exceptional childbearing in the age group 45 – 49 years. In contrast, the age group 35 – 39 years is not only dynamic but also still important for the procreation. In general, total late fertility was higher in 1960 than in 2003 in majority of European countries. Further increase in total late fertility in most of Europe is likely, Eastern Europe included. However, late age – specific fertility will very likely retain more or less marginal share of total fertility in modern demographic regime. Late age – specific abortion will considerably contribute to this result.

Povzetek: Prispevek obravnava značilnosti pozne rodnosti v Evropi v razmerah sodobnega načina obnavljanja prebivalstva.

1 Introduction

European continent has been for-runner in demographic development for several centuries. The most advanced European countries were pioneers at the beginning and at the end of the demographic transition. The main result of the process of demographic transition in Europe was modern demographic regime. Demographers differ in views about the detailed characteristics of the modern demographic regime and about the timing of its appearance. However, there is general agreement that low fertility and low mortality levels are the most prominent features of the post-transitional demographic regime in Europe.

Broadly speaking, it is possible to state that the modern demographic regime in Europe emerged in the 1950s. At present, it is about sixty years old. It is hardly enough for the development of all facets of the new way of population or generational replacement. We need much longer period if we want to competently compare modern regime with the traditional or pre-transitional regime of population development.

It is also premature to expect that demographers would have reached general consensus on the theory of modern demographic regime. The demographers simply need longer period and more countries and national populations with finished demographic transitions to be able to study modern demographic processes in different environments. From this point of view the so called

theory of the second demographic transition is more confusing than illuminating. According to my opinion, it is not wise to speak about the second demographic transition before we really know whether below replacement fertility is one of the main characteristics of the modern demographic regime or not. If the answer is yes, then we do not have any transition at all.

In spite of the theoretical and practical dilemmas it is very important for demographers to study different characteristics of the modern demographic regime in Europe. The author of this paper will focus on late age-specific fertility trends in Europe. Age-specific fertility patterns had changed tremendously during the period of the demographic transition in Europe. A uni-modal distribution of age-specific fertility rates considerably slimmed as a consequence of the fertility decline in younger and older ages of the female reproductive period.

Age-specific fertility in advanced European countries with modern demographic regime is concentrated in the female age group 20 – 34 years. Women and couples tend to limit fertility in marginal age groups of female's reproductive period. Adolescent fertility is not only unwanted but more and more unacceptable for an individual and the society. Main reasons are prolonged schooling and late economic emancipation of young people.

Late age-specific fertility is much more in the domain of an individual woman and her family in spite of the fact that physicians advise avoiding a

childbearing after the age of 35 years. Widespread employment of women, low reproductive norms and modern life styles in Europe do not support the childbearing in the period 35 – 49 years of age. Age-specific fertility rates in this period show efficient family planning and the use of modern contraceptives.

In this paper late age-specific fertility is defined by the age group 35-49 years of the females. The author will follow available statistical data on the late age-specific fertility and abortion trends in Europe up to the recent years. He will also try to answer the question how much have late age-specific fertility and abortion changed as the consequence of very low level of fertility and of the postponement of childbearing among young generations. Late fertility and late abortion differentials will also be analysed. The author will devote much attention to the question whether any breakthrough can be anticipated in late age-specific fertility in Europe in the future or should we further maintain the position that childbearing after the woman's age of 35 will remain marginal for still very obvious reasons.

2 Low fertility level in Europe

Generally, fertility level in almost all of Europe is low indeed. All national period total fertility rates in consecutive years 2003 and 2004 were below 2.04 if we exclude Turkey and Albania (RDDE, 2006, p. 78). Particular rates covered wide range between 1.17 in Ukraine in 2003 and 2.04 in Iceland in 2004. Sporadically and regionally the indicator can be even lower. In the lands of former German Democratic Republic it was only 0.84 in 1995. Intrinsic rates of increase of model populations with such fertility levels are negative. If we take for example Slovenian total fertility level of 1.2 in the year 2003, it is possible to calculate the intrinsic rate of -1.9 which means that model population with such a rate would diminish to a half of its original number in only about 37 years.

European fertility decline is also evident in cohort total fertility rates. Completed fertility rates of a female birth cohort born in 1967 were predominantly under 2.0 with the exception of Albania, Armenia, Azerbaijan, Cyprus, Iceland, Ireland (1966), Norway, Serbia and Montenegro and Macedonia (RDDE, 2005, p. 89). Mainland Europe is well below 2.0. Its range was between 1.46 in Germany and 1.99 in France and Slovak Republic in 1967. Below replacement fertility is predominant characteristic of Europe at the beginning of the 21st century.

Low fertility level in Europe is main cause for growing number of countries with negative natural increase. In 1980, only Austria and Federal Republic of Germany had negative natural increase of the population. Ten years later, number of births was lower than number of deaths in three countries

(Bulgaria, Germany and Hungary). At the turn of the century, in 2003, the number was 19 (RDDE, 2006, p. 56) and it will grow in the future.

Mean age of women at birth of first child has been increasing in recent decades in Europe. In spite of some minor problems with missing data it is possible to see considerable increase in the number of countries with the mean age of women at birth of first child of 25 years and more in the period 1970 – 2000. In the years 1970, 1980, 1990 and 2000 there were 4, 8, 16 and 23 such countries respectively in Europe. In 2004, 9 out of 28 countries with published data had the mean age of women at birth of first child of more than 28 years. There were no such countries in 1990 (RDDE, 2006, p. 87).

Trends of the mean age of women at childbearing in Europe in the period 1960 – 2003 are more complex. They have been influenced by the decline of number of births of the higher parity birth orders, which have generally occurred at the higher age of the mother and by the increase of the mean age of women at birth of lower parity birth orders. These two forces have worked in opposite directions. Real trends in the period considered show quite similar pattern in majority of European countries. There was decline in the mean age of women at childbearing till the 1970s or early 1980s. Thereafter, the trend has changed the direction. The mean age of women at childbearing has started to increase. In 2004, there were 21 out of 32 countries with the published data where the mean age of women at childbearing was more than 28 years. In 8 countries the indicator was higher than 30 years. These countries were Andorra, Denmark, Ireland, Liechtenstein, the Netherlands, San Marino, Sweden and Switzerland (RDDE, 2006, p. 88).

Some decades ago G. Calot and C. Blayo wrote about considerable homogenization of fertility in Western Europe (Calot and Blayo, 1982, p. 353). In recent decades the process has spread over the entire continent. However, it is far from universality. In 2000, the difference between the highest (Iceland) and the lowest (Ukraine) total fertility rate in Europe, Albania and Turkey excluded, was still practically one child per woman (RDDE, 2005, p. 76). Therefore, it is possible to conclude that below replacement fertility level is more important characteristic of European fertility than the homogenization.

3 Late age – specific fertility

Late motherhood and postponement of childbearing in Europe are frequent research topics in recent years. Some examples of studies are Van Nimwegen et al, 2002, Sobotka, 2004, Ni Bhrolchain and Toulemon, 2005 and Billari, 2005. It is possible to see different meanings of the terms late motherhood and postponement in these studies. The Dutch authors use the term late motherhood in

the meaning of postponement of the childbearing (Van Nimwegen et al, 2002. p. 10 – 16). Many other authors use the term postponement in two different meanings. The first meaning stresses the possibility of compensation of the fertility decline at younger ages with (at least partial) fertility rise at later ages. The second meaning most often simply refers to an increase in the mean age of first birth or in the mean age at childbearing (Ni Bhrolchain and Toulemon, 2005, p. 86). However, in almost all of these studies very little is said about late age – specific fertility in the last 15 years of female reproductive period.

Late motherhood or late fertility can be understood also from the viewpoint of the late age – specific fertility. The author of this paper studied late age – specific fertility in Europe in the period 1961 – 1985 (Malacic, 1994). Late age – specific fertility considerably decreased in the studied period in Europe. However, toward the end of the period in certain European countries some discontinuities and turnabouts in the prevalent tendency became evident. In spite of that, the author of the paper forecasted in the cited paper that the late age – specific fertility would retain a marginal share in total fertility in Europe in decades to come.

Late age – specific fertility will be analysed in this paper on the basis of age – specific fertility rates in the age groups 35 – 39, 40 – 44 and 45 – 49 years and late total fertility rate in the age group 35 – 49 years which is defined as the sum of five years age – specific fertility rates in the 15 years age range multiplied by five. Late total fertility rate can

be interpreted as the number of childbirths to the hypothetical cohort of women in the age group 35 – 49 years under the condition that registered five years age – specific fertility rates in the age group 35 – 49 years refer to the given cohort of females. Late total fertility rate is not dependent on the age structure of the particular population.

Age – specific fertility is relatively low after the age of 35 for different reasons. They work practically over the whole reproductive age span and can be grouped as follows: 1. intrauterine mortality, postpartum amenorrhea, ovulation without conception; 2. early sterility and possibly higher intrauterine mortality for older women; 3. extension of birth intervals by abstinence or prolonged breast-feeding and 4. birth control by contraception or abortion (Lutz, 1989, p. 7). The one and two groups are more important for natural fertility regime which can be illustrated by Hutterite fertility. Hutterite fertility rates for age groups 35 – 39, 40 – 44 and 45 – 49 are 0.406, 0.222 and 0.061 respectively (Malacic, 2006, p. 55). Contraception and abortion are predominant causes of low levels of late age – specific rates in modern fertility regime of present day Europe. They are the consequence of more or less conscious decisions.

Nine European countries and the period 1960 – 2003 will be analysed in this paper. The countries will be divided in three groups. In the first group are Denmark, Sweden and United Kingdom. Their basic characteristic is higher fertility rate in the age group 35 – 39 in 2003 than in 1960. The second group includes France, Germany and Italy. Their size is the main reason for the selection from the

		Denmark	Sweden	UK ¹	France ²	Germany ³	Italy ³	Poland	Russia	Ukraine ⁴
f ₍₃₅₋₃₉₎	1960	39.2	39.6	44.1	50.6	44.2	61.2	58.4	53.6	39.6
	1983	15.0	26.0	23.0	22.0	16.0	24.8	30.2	22.2	20.2
	1993	33.4	43.4	33.4	32.2	21.6	30.4	25.4	10.8	12.0
	2003	46.6	51.0	46.0	46.6	30.2	42.6	21.0	16.0	11.0
	I _{2003/1983}	310.7	196.2	200.0	211.8	188.7	171.8	69.5	72.1	54.4
	I _{2003/1960}	118.9	128.8	104.3	92.1	68.3	69.6	36.0	29.8	27.8
f ₍₄₀₋₄₄₎	1960	11.2	12.0	12.3	15.6	11.8	22.2	21.5	18.8	10.6
	1983	2.4	4.6	4.4	4.4	2.6	5.2	7.4	4.8	3.6
	1993	5.0	7.6	5.8	6.2	3.4	5.6	6.2	2.4	2.0
	2003	7.6	9.6	9.4	9.2	4.8	8.0	4.8	2.8	2.0
	I _{2003/1983}	316.7	208.7	213.6	209.1	184.6	153.8	64.0	58.3	55.6
	I _{2003/1960}	67.9	80.0	76.4	59.0	40.7	36.0	22.3	14.9	18.9
f ₍₄₅₋₄₉₎	1960	0.8	0.8	0.8	1.0	0.8	1.8	2.3	2.8	1.0
	1983	0.2	0.2	0.4	0.2	0.2	0.4	0.4	0.4	0.4
	1993	0.2	0.2	0.2	0.2	0.2	0.2	0.4	0.2	0.0
	2003	0.2	0.4	0.6	0.4	0.2	0.4	0.2	0.2	0.2
	I _{2003/1983}	100.0	200.0	150.0	200.0	100.0	100.0	50.0	50.0	50.0
	I _{2003/1960}	25.0	50.0	75.0	40.0	25.0	22.2	8.7	7.1	20.0

Table 1: Late age – specific fertility rates (in %) and indexes 2003/1983 and 2003/1960 for selected European countries in 1960, 1983, 1993 and 2003. Source: Recent demographic developments in Europe 2004, Council of Europe, Strasbourg 2005; Demographic Yearbook. Special issue: Historical supplement, United Nations, New York 1979

Notes: ¹ 1959 instead of 1960; ² 2002 instead of 2003; ³ 2001 instead of 2003; ⁴ 1961 instead of 1960.

pool of North – Western, South and Central European countries with typical U shaped f_{35-39} in the studied period. The third group includes Poland, Russia and Ukraine. These countries represent Eastern Europe and to some extent Albania and Turkey. Late age – specific fertility has declined in this group practically over the whole period covered in the analysis.

Late age – specific fertility rates for age groups 35 – 39, 40 – 44 and 45 – 49 and indexes 2003/1983 and 2003/1960 for selected European countries and for the years 1960, 1983, 1993 and 2003 are shown in table 1. Statistical sources do not cover all data demanded in the title of the table 1. Therefore, some minor data substitutions were necessary for particular years and countries. It

should be outlined also that Germany, Russia and Ukraine have not been independent countries in today's state borders since the 1960s. Fortunately, statistical data for these three states in their present size are available for the whole period studied and can be analysed in this paper. Additionally, figures 1 – 3 are included in the paper to show graphical illustration of the late age – specific fertility trends for two rates f_{35-39} and f_{40-44} for the period 1960 – 2003. The trends for the three groups of countries are shown separately.

Late age – specific fertility rates in selected European countries show very low fertility in the age groups 40 – 44 and 45 – 49 years. Childbearing in the age group 45 – 49 years is really exceptional in modern Europe. In contrast, the age group 35 –

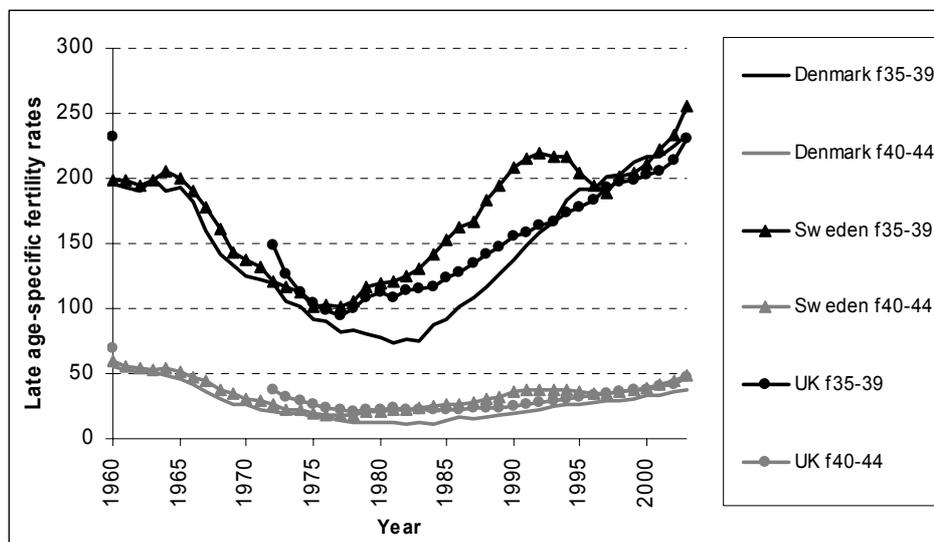


Figure 1: Sum, by five-year age groups, of late age-specific fertility rates f_{35-39} and f_{40-44} for Denmark, Sweden and UK for the period 1960 – 2003. Source: RDDE 2004, Council of Europe, Strasbourg 2005.

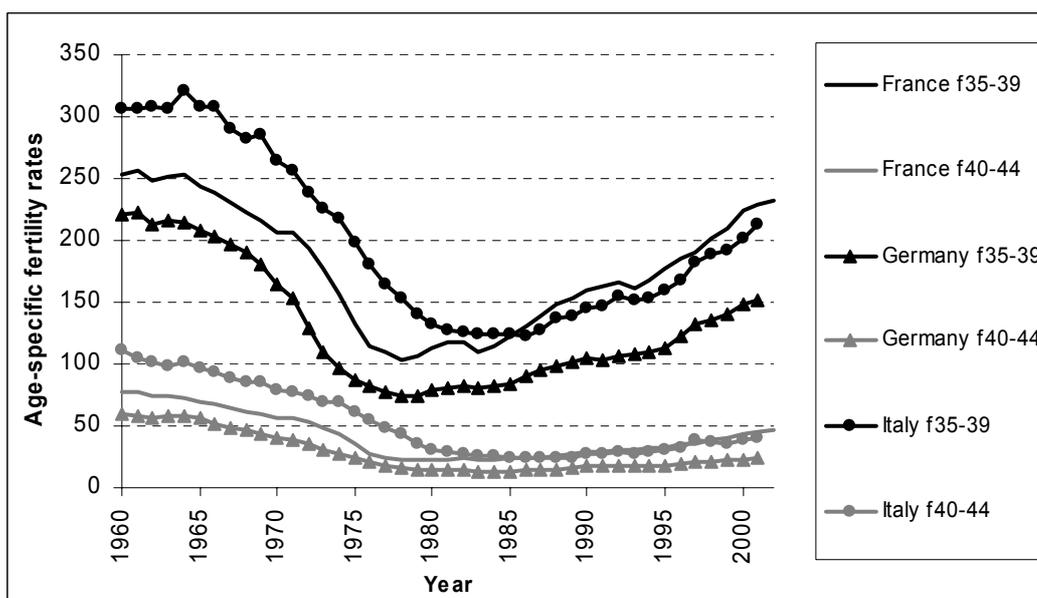


Figure 2: Sum, by five-year age groups, of late age-specific fertility rates f_{35-39} and f_{40-44} for France, Germany and Italy for the period 1960 – 2003. Source: RDDE 2004, Council of Europe, Strasbourg 2005.

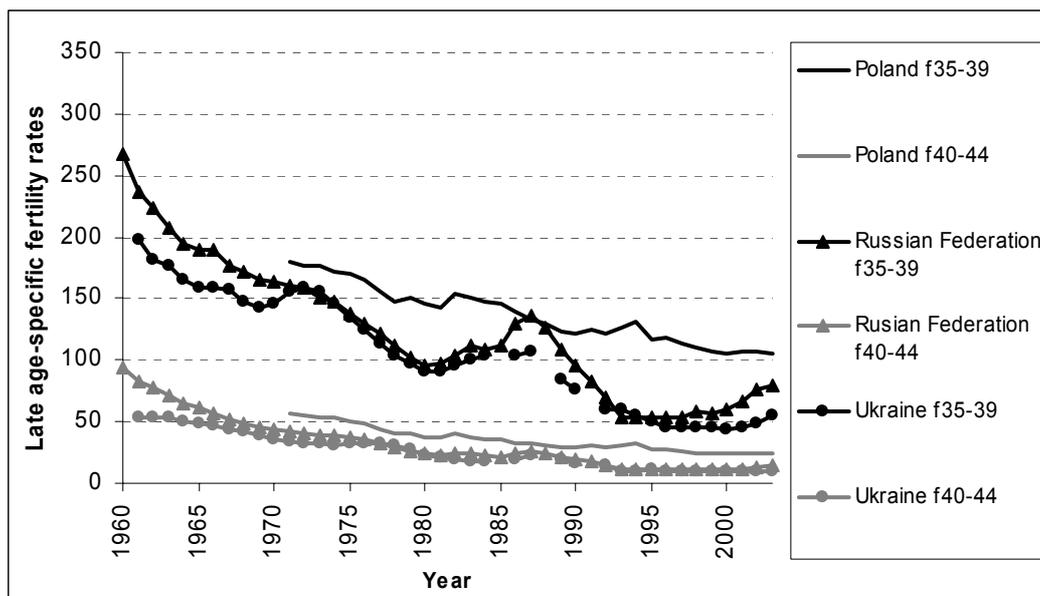


Figure 3: Sum, by five-year age groups, of late age-specific fertility rates f_{35-39} and f_{40-44} for Poland Russian Federation and Ukraine for the period 1960 – 2003. Source: RDDE 2004, Council of Europe, Strasbourg 2005.

39 years is not only dynamic but also still important for the procreation. Our three groups of countries behave differently in this respect. The difference is smaller between for-runners Denmark, Sweden and UK and France, Germany and Italy which represent mainland Europe than between these two groups and the Eastern European group. In the for-runners group we have only three European countries where f_{35-39} was higher in 2003 than in 1960. In other respects for-runners and mainland groups are quite similar.

In all six countries three late age-specific fertility rates declined between 1960 and 1983. Thereafter, two of them, f_{35-39} and f_{40-44} , have increased considerably as indexes 2003/1983 show. The third rate, f_{45-49} , remains more or less at very low level.

The dynamics in the Eastern European group

was very simple in the period considered. All three age-specific fertility rates declined and reached at the end of the period practically the same level of particular rates as other six countries had in 1983. In this respect Poland, Russia and Ukraine are late-comers. In Russia, some signs of turnarounds in 1993 were already visible in f_{35-39} and f_{40-44} .

For more complete elaboration of late age-specific fertility it is necessary to compare late and total fertility. Therefore, we have selected some indicators of late and total fertility for selected European countries for the years 1960, 1983, 1993 and 2003. The indicators are shown in table 2. These indicators are total late fertility rate ($T_{f,35+}$), $T_{f,35+}$ as a percent of T_f , percent change of the $T_{f,35+}$ in the period 1983 – 2003 and percent change of the T_f in the period 1983 – 2003.

For-runners and mainland groups show similar

	Denmark	Sweden	UK ¹	France ²	Germany ³	Italy ³	Poland	Russia	Ukraine ⁴	
$T_{f,35+}$	1960	256	262	286	336	284	426	411	376	256
	1983	88	154	139	133	94	152	190	137	121
	1993	193	256	197	193	126	181	160	67	70
	2003	272	305	280	281	176	255	130	95	66
$T_{f,35+}$ As a % of T_f	1960	10.0	11.9	11.7	12.3	12.0	17.7	13.8	14.7	11.4
	1983	6.4	9.6	7.8	7.5	6.6	10.1	7.9	6.6	5.9
	1993	11.0	12.9	11.2	11.6	9.8	14.2	8.6	4.9	4.5
	2003	15.5	17.8	16.4	14.9	13.0	20.4	10.6	7.2	5.5
$\Delta T_{f,35+}$ in 1983 – 2003 in %	+ 209.1	+ 98.1	+101 .4	+ 45.1	+ 87.2	+ 67.8	- 31.6	- 30.7	- 45.5	
ΔT_f in 1983 – 2003 in %	+ 27.5	+ 6.2	- 3.4	+ 5.6	- 5.6	- 17.2	- 49.0	- 36.8	- 41.7	

Table 2: Selected indicators of late and total fertility for selected European countries in 1960, 1983, 1993 and 2003. Source: RDDE 2004, Council of Europe, Strasbourg 2005; Demographic Yearbook. Special issue: Historical supplement, United Nations, New York 1979.

Notes: ¹ 1959 instead of 1960; ² 2002 instead of 2003; ³ 2001 instead of 2003; ⁴ 1961 instead of 1960.

trends in selected late age-specific fertility indicators in the period 1960 – 2003. Total late fertility rate which is un-comparably lower in modern Europe than in the case of Hutterite fertility declined between 1960 and 1983 and has increased thereafter. Similar dynamics characterised total late fertility as the percentage of the total fertility in six countries of the for-runners and mainland groups. In 2003, however, the total late fertility had higher percentage of the total fertility than in the year 1960. In Italy, 20.4 % of period total fertility occurred in the age group 35 – 49 in the year 2003.

Total late fertility as a percent of total fertility increased considerably in the period 1983 – 2003 in all six countries. The increase was un-comparable to the change of the total fertility rates of these six countries in the same period. In the period studied, total fertility increased in Denmark, Sweden and France and declined in UK, Germany and Italy. It is more than evident that late age-specific fertility trends in the for-runners and mainland groups of European countries indicate certain degree of postponement of childbearing at least in the period 1983 – 2003.

The Eastern European group has had different development in the period studied. Total late and total fertility rates in this group of countries have had declining trends practically to the end of the period. In two decades of the period 1983 – 2003 total fertility declined more than the share of total late in total fertility in Poland and Russia. The opposite was true for Ukraine during the same period. There are no signs of any postponement of childbearing in Poland, Russia and Ukraine in the age groups 35 – 39 or 40 – 44 years yet. Late age-specific fertility differentials

Late age-specific fertility differentials statistical

data are rare in international as well as in national data sources. The data shortage is combined with the differences in periods covered and statistical definitions used in particular countries. The author of this paper uses international data sources mainly. Therefore, very limited picture of the late age-specific fertility differentials can be shown in the paper.

Only three late age-specific fertility differentials will be used on the basis of the rates f_{35-39} , f_{40-44} and f_{45-49} . These three characteristics are birth order, rural-urban residence and marital fertility. Late age-specific fertility rates will be transformed in total late fertility by particular characteristic. The method of calculation and the analytical value of the indicator have been explained in previous section.

In table 3 total late fertility by birth order parity for our selected European countries in the years 1985 and 2001 are shown. Russian Federation and Ukraine are omitted because of the lack of data. The highest birth order parity is 5+ which is appropriate for the selected countries. European females in the age group 35-49 years gave births of different birth order parity in the years 1985 and 2001. Some of them had their first child while others gave birth to the children of higher birth order parity. The highest values of the indicator in table 3 are for the second and the third birth order parity with the exception of France for year 2001.

Parity specific trends in the period 1985 – 2001 followed the division of the selected countries in the three groups, again with the exception of UK, which is closer to mainland group than to the for-runners group.

The for-runners countries Denmark and Sweden show increase practically in all birth order

Country	Year	Parity					
		Total	1	2	3	4	5+
Denmark	1985	105	16	32	33	15	8
	2001	227	47	81	62	23	14
Sweden	1985	181	27	50	58	28	14
	2001	281	66	97	70	28	20
UK ¹	1985	145	26	44	41	21	18
	2001	198	51	75	42	17	13
France ²	1985	181	39	47	41	21	23
	2001	244	93	66	47	20	18
Germany ³	1985	131	35	46	33	14	11
	2001	183	56	71	35	12	9
Italy	1981	153	26	47	42	21	17
	1997	236	66	101	48	14	7
Poland	1985	179	16	37	48	39	39
	2001	115	11	25	31	20	28

Table 3: Parity-specific total late fertility in selected European countries in the years 1985 and 2001 (sum, by each parity in the group 35–49 years of age of mothers). Source: Demographic Yearbook 1986, United Nations, New York, 1988, pp. 791–815 and Eurostat.

Notes: ¹England and Wales for 1985; ²Marital fertility for 1985; ³Marital fertility and FR Germany for 1985.

parities of the late fertility in the period studied. UK and the countries of mainland group experienced increase in the first three parities and decline for the fourth and 5+ parities. The only country with available data from the Eastern European group, Poland, shows decline for all parities in the period 1985 – 2001.

In the for-runners and mainland groups of countries the highest relative increase happened in the case of the first parity. In France the first parity had the highest value of all parities in the year 2001. This trend clearly shows certain postponement of childbearing in these countries.

Late fertility statistical data in table 3 are very good illustration of the low reproductive norms in the modern demographic regime. Parity specific values of late fertility for the 4th and 5+ birth orders are very low and declining. This is very important for population policy. Parity specific population policy measures should be concentrated in younger ages of females (parents) and for the second and third birth order parity.

Rural – urban and marital late fertility in selected European countries and selected years with available data are shown in table 4. Unfortunately, quite a lot of data are not available. Therefore, the table 4 is only partly informative. In Eastern European countries urban and rural late fertility declined in the period studied while the data on marital fertility are not available for Russian Federation and Ukraine. The case of France which is only other country with limited trend data shows increase of both urban and rural late fertility in the

period 1982 – 1990. Similar trends are likely in other 5 selected countries of the for-runners and mainland groups of countries in spite of the fact that we do not have that data. In Eastern Europe rural late fertility is still higher than urban one, while it is impossible to say the same for other two groups of countries.

Total late marital fertility data for countries with available data are rather old. The data are for the late 1970s and early or middle of the 1980s. Total late marital fertility increased in the period covered in all countries with the exception of Italy. However, this evidence is not very important because of considerable transformations of family structures in Europe in the 1970s and the 1980s. Simultaneously, *de iure* marriages lost importance in comparison with *de facto* unions in almost all countries selected.

4 Late fertility and late abortion

In modern demographic regime family planning is very important part of the way of life. Different forms of male and female contraception are widespread. People usually do not rely on contraception when they want to conceive. However, from time to time even modern contraception fails and people are forced to choose between childbirth and abortion

The frequency of legal induced abortion is an indicator of unwanted pregnancies and failed use of contraception. Abortion rate in particular country depends on legal status of induced abortion and political perception of the fertility level in the

Country	Urban-Rural fertility			Marital late fertility	
	Year	Urban	Rural	Year	Marital T _f 35+
Denmark	1969	127	185	1978	30
	1999	n.a.	n.a.	1985	100
Sweden	1985	n.a.	n.a.	1978	109
	1999	n.a.	n.a.	1985	188
UK ¹	1972	188	165	1977	148
	1999	n.a.	n.a.	1985	146
France	1982	178	144	1979	129
	1990	228	178	1985	181
Germany ²	1981	46	65	1978	115
	1999	n.a.	n.a.	1985	143
Italy	1981	n.a.	n.a.	1977	208
	1999	n.a.	n.a.	1981	171
Poland	1985	151	268	1978	168
	1999	110	171	1984	201
Russia	1989	120	212	n.a.	n.a.
	1999	60	90	n.a.	n.a.
Ukraine	1987	114	173	n.a.	n.a.
	1998	47	80	n.a.	n.a.

Table 4: Rural–urban and marital late fertility in selected European countries and selected years with available data (sum, by each characteristic in the group 35–49 years of age of mothers). Source: Demographic Yearbook 1986, United Nations, New York, 1988, pp. 655 – 657 and 876–877. Demographic Yearbooks 1992, 1997, 2000 and 2001, United Nations, New York.

Notes: ¹England and Wales in 1972. ²DR Germany in 1981 and FR Germany in 1978 and 1985.

country. Induced abortion is legal in all selected European countries with the exception of Poland. However, political attitude and legal conditions for the realization of the abortion are quite different in particular countries included in this study.

Selected abortion indicators for seven European countries are shown in table 5. Poland and Ukraine are omitted because of unavailable data. Very limited data are available for Russia which has very high total abortion rate. Abortion indicators are constructed in the same way as fertility indicators. There are six abortion indicators in the table 5. They are three age – specific abortion rates, a_{35-39} , a_{40-44} and a_{45-49} , general abortion rate, a_{15-49} , total late abortion rate, $T_{a,35-49}$, and total abortion rate, T_a . The last column of the table 5 shows a percentage of the total late abortion rate in the total abortion rate. Abortion data in table 5 are directly comparable with fertility data in other tables of the paper.

Total abortion rate increased in Sweden, UK and Germany and declined in Denmark, France and Italy in the period covered. In particular years between 1997 and 2001 which are included in the table 5 the percentages of T_a in T_f were 24.5 in Denmark, 35.9 in Sweden, 27.5 in UK, 22.5 in France, 17.9 in Germany and 26.9 in Italy (RDDE, 2005). In Russia, an index T_a/T_f was 196.3 in 1995 which is incomparably higher than in other countries included in the study. It seems that abortion has replaced modern contraceptives as a dominant mean of family planning in Russia.

Late abortion trends were also different in particular countries included in our study and in the period covered. The changes in UK and Italy were considerable. UK experienced increase while in Italy it was decline. The changes in other countries were modest. It was increase in Germany and decline in Denmark, Sweden and France.

However, late abortion represents still considerable share of total abortion in the countries included in our consideration. The share was somewhere between one fifth and one quarter in most countries as is shown in column 9 in table 5. There are again two exceptions. Russia has the lowest and UK has the highest figure in column 9 of the table 5. UK experienced also the highest increase in the period 1985 – 1999. The percentage of total late abortion in total abortion increased from 12.6 in 1985 to 30.3 in 1999 in UK. It is more than obvious that late abortion is important mean for achieving low level of late age – specific fertility in all countries which are included in table 5.

5 Conclusion

Age – specific fertility distribution has shown considerable changes in modern demographic regime. It seems that postponement of childbearing has been the most prominent feature of these changes in Europe in recent decades. The postponement has influenced at least partly late age – specific fertility as well.

Total late fertility increased considerably in the two groups of countries and declined in Eastern European group in the period 1983 – 2003. However, for – runners in total late fertility increase are still rare in Europe. Generally, total late fertility was higher in 1960 than in 2003 in majority of European countries. Further increase in total late fertility in most of Europe is likely, Eastern Europe included. However, late age – specific fertility will very likely retain more or less marginal share of total fertility in modern demographic regime. Late age – specific abortion will considerably contribute to this result.

Country	Year	a_{35-39}	a_{40-44}	a_{45-49}	a_{15-49}	$T_{a,35-49}$	T_a	7/8 in %
Denmark	1985	13.3	6.7	1.0	15.6	105	546	19.2
	2000	13.1	4.8	0.5	12.4	92	433	21.2
Sweden	1985	15.3	7.3	1.0	15.5	118	542	21.8
	2001	16.0	5.7	0.6	15.8	111	563	19.7
UK ¹	1985	7.0	2.9	0.3	11.6	51	406	12.6
	1999	14.5	10.5	3.3	13.7	141	465	30.3
France	1984	12.0	5.7	0.8	13.4	92	469	19.6
	1997	11.4	4.9	0.5	11.1	84	389	21.6
Germany ²	1985	6.2	2.4	0.4	5.3	45	185	24.3
	2000	6.6	2.7	0.3	6.8	48	247	19.4
Italy	1982	21.9	9.6	1.0	16.9	162	590	27.5
	1998	11.5	5.5	0.5	9.6	87	326	26.7
Russia	-	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
	1995	n.a.	n.a.	n.a.	58.5	390	2630	14.8

Table 5: Selected indicators of late and total abortion for selected European countries and for the years 1985 and 2000 or the nearest year with available data. Source: Demographic Yearbook 1986, United Nations, New York, 1988, pp. 212–219 and 1036–1038; Demographic Yearbooks 1988, 2000 and 2001, United Nations, New York; RDDE, selected issues, Council of Europe, Strasbourg.

Notes: ¹England and Wales in 1985; ²FR Germany in 1985.

References

- [1] Billari, F.C. (2005). *The Transition to Parenthood in European Societies*. Conference Report, European Population Conference 2005, Demographic Challenges for Social Cohesion, Council of Europe, Strasbourg, 7 – 8 April 2005. 49 p.
- [2] Calot, G. and Blayo C. (1982). Recent course of fertility in Western Europe, *Population Studies*, 1982/3, pp. 349 – 372.
- [3] Eurostat home page, November 2005 and 2007.
- [4] Lutz, W. (1989). *Distributional Aspects of Human Fertility*. A global comparative study, Academic Press, London. Xi + 282 p.
- [5] Malacic, J. (1994). Kasni fertilitet i savremena reprodukcija stanovništva (Late fertility and modern reproduction of the population). In: Maksimovic, I., ed. *Prilozi demografskim i ekonomskim naukama* (Contributions to the Demographic and Economic Sciences), SANU (Serbian Academy of Art and Sciences), Monografije, Vol. 103, Beograd. Pp. 189 – 198.
- [6] Malacic, J. (2006). Demografija. Teorija, analiza, metode in modeli (Demography. Theory, analysis, methods and models), 6. izdaja, Ekonomska fakulteta, Ljubljana, viii + 339 p.
- [7] Ni Bhrolchain, M. and Toulemon, L. (2005). Does Postponement Explain the Trend to Later Childbearing in France? In: Vienna Yearbook of Population Research 2005, Vienna Institute of Demography, Austrian Academy of Sciences. Pp. 83 – 107.
- [8] Recent demographic developments in Europe (RDDE), different yearly issues, Council of Europe, Strasbourg.
- [9] Sabotka, T. (2004). Is lowest – low fertility in Europe explained by the postponement of childbearing. *Population and Development Review*, 2004/2, pp. 195 – 220.
- [10] Van Nimwegen, N. et al (2002). Late motherhood in the Netherlands: current trends, attitudes and policies, *Genus*, 2002/2, pp. 9 – 34.
- [11] United Nations (1979). Demographic Yearbook. Special issue: Historical supplement, New York.
- [12] United Nations, *Demographic Yearbook*, different issues, New York

Methodology for the Estimation of Annual Population Stocks by Citizenship Group, Age and Sex in the EU and EFTA Countries

Jakub Bijak and Dorota Kupiszewska

Central European Forum for Migration and Population Research (CEFMR)

ul. Twarda 51/55, 00-818 Warsaw, Poland

E-mail: j.bijak@cefmr.pan.pl, d.kupisz@cefmr.pan.pl

www.cefmr.pan.pl

Keywords: population estimates, stocks by citizenship, Europe, missing data, cohort-wise methods, fitting methods

Received: March 9, 2008

The paper addresses selected computational issues related to the challenge of dealing with poor statistics on international migration. Partial results of the on-going Eurostat-funded project on “Modelling of statistical data on migration and migrant population” (MIMOSA) are presented. The focus is on the data on population stocks by broad group of citizenship, sex and age. After a brief overview of the main problems with data on population by citizenship for 31 European countries (27 European Union countries, Iceland, Liechtenstein, Norway and Switzerland), a range of computational methods is proposed including cohort-wise interpolation, cohort-component projections, cohort-wise weights propagation and proportional fitting methods, as well as other, auxiliary methods. The algorithm for choosing the best method for estimating missing data on population stock by broad citizenship (nationals, foreigners – EU27 citizens, foreigners – non EU27 citizens), five-year age group (up to 85+) and sex on 1st January 2002–2006 is proposed and illustrated by examples of its application for selected countries.

Povzetek: Opisane so različne metode za ocenjevanje demografskih podatkov.

1 Introduction

The deficiencies of statistical information on migration-related variables, such as population flows or stocks, are well-known and widely discussed in the literature [1, 7]. The aim of the paper is to contribute to the works on dealing with these shortcomings and to propose a set of computational methods, as well as an algorithmic procedure of selecting the best one, for the estimation of population stocks as of 1st January in a breakdown by sex, age group and broad citizenship category, for the countries for which information is unavailable or incomplete.

The study was carried out within a Eurostat-funded project on “Modelling of statistical data on migration and migrant population” (MIMOSA). It covers 31 European countries, of which 27 belong to the European Union (as per 1st January 2007), and further four – to the EFTA (Iceland, Liechtenstein, Norway and Switzerland). The period of interest is 2002–2006. The citizenship groups under study are: nationals, European Union (EU27) foreigners and non-EU27 foreigners, while the age groups are five-year, with the last, open-ended category being 85 years or more.

Generally, the proposed estimation methods aim to combine data from different sources (population census, vital statistics, data on acquisition of citizenship, specific surveys, etc.). In principle, the data that are already available are not modified (for example, in order to harmonise definitions, or for any other reason), unless in the case of inconsistencies between the sources. In the latter

cases, the demographic data, provided to Eurostat by national statistical institutes (NSIs), are given priority.

Apart from the Introduction, the paper is structured in four sections. Section 2 contains summary information on the availability and quality of the 2002–2006 data on population stocks for 31 countries under study. In Section 3, the proposed methodology for estimating population stocks by sex, age and citizenship groups is discussed. This section presents such tools as estimation of data in single years of age from five-year age-groups, cohort-wise interpolation of population stocks, cohort-component projections, cohort-wise propagation of weights, proportional fitting, as well as other, auxiliary methods. Subsequently, Section 4 contains recommendations concerning the procedure of selecting appropriate estimation methods for each of the countries under study, presented in the form of a decision algorithm and accompanied by several illustrative examples for selected countries. The discussion is summarised in Section 5.

The study is based on the data available in the Eurostat databases, supplemented by additional information obtained from national statistical institutes, whenever required and feasible. Throughout the paper, the abbreviation ‘NSI’ is used to denote the national statistical institute of the respective country, ‘JMQ’ stands for the Joint Questionnaire on International Migration Statistics (hereafter: Joint Migration Questionnaire) of Eurostat, UN Statistical Division, UN Economic Commission for

Europe, the Council of Europe and the International Labour Office. ‘LFS’ depicts the Labour Force Survey.

2 Availability of the 2002–2006 data on population stocks for 31 European countries

Annual statistics on usually resident population by citizenship, sex and age are collected by Eurostat from the NSIs via the Joint Questionnaire on International Migration, together with migration flow data. Population statistics for 37 European countries, collected through the JMQ are checked and subsequently loaded into Eurostat’s on-line database, NewCronos. The data are located under the *Population and Social conditions* theme, in the *International Migration and Asylum* domain (MIGR), tables *migr_st_popctz* (population by sex and citizenship) and *migr_st_popage* (population by age group, citizenship and sex). The data for 2000–2006 come from the following tables in the 2000–2006 JMQs:

- Table 7a (for 2000–2003, 8a): Usually resident population by citizenship and age, both sexes;
- Table 7b (for 2000–2003, 8b): Usually resident population by citizenship and age, males;
- Table 7c (for 2000–2003, 8c): Usually resident population by citizenship and age, females.

A detailed analysis of statistics on population stocks by citizenship provided by the 31 countries covered the JMQs for the reference period 2002–2006. Selected results of the analysis of the data availability for particular countries are summarised in Table 1, providing an overview of the situation for all 31 countries. The information on the lack of data, marked as ‘not available’ in Table 1, was based on the information provided in the JMQ or on information obtained during the THESIM project¹. Other missing data were marked as ‘not provided to Eurostat’. In addition to missing data, a number of other problems were detected, for example the presence of provisional data, some citizenship categories only, broad age groups, or a different reference date than 1st January.

Data on total population stock on 1st January, not disaggregated by citizenship, are also collected by Eurostat within the framework of the Annual Demographic Statistics data collection. These data, disaggregated by sex and age, are located under the *Population and social conditions* theme, in the *Demography* (DEMO) domain of the database, table *demo_pjan*. The results of the review of the availability of these data for the years 2002–2006 revealed that the data on total population stock by sex and five-year age group (up to 85+) are available for all 31 countries, with the following exceptions: there is no 2006 data by age for Belgium and Italy, while for Romania the highest age group in 2004 data is 80+.

In addition to the annual data, Eurostat also collects and disseminates statistics on population by citizenship, sex and age obtained by the countries during population censuses. Like other statistics, the census data are located under the *Population and social conditions* theme, in the *Census* (CENS) domain of the database, table *cens_nsctz*. Unlike annual population figures, the census data on population by citizenship, sex and age are available for almost all 31 countries, with the notable exceptions of the United Kingdom, Germany and Malta.

A supplementary source of data on population stock by citizenship is the Labour Force Survey. However, the availability of data from the LFS in the Eurostat database is very limited and the reliability of data is probably not high, due to the nature of the data source. By definition, the LFS statistics are estimates and thus bear certain errors, which can be relatively high for disaggregated categories (e.g., for population broken down by age, sex and citizenship groups). However, some use of the LFS data could be considered as an alternative to the proposed methods in the countries where data on total nationals and total foreigners are missing.

In the Eurostat database, the LFS tables are located under the *Population and social conditions* theme, the *Labour market* (LABOUR) domain, in the table with population data containing the nationality dimension (population by sex, age groups, nationality and labour status, table *lfsa_pganws*). However, the table does not contain data on the level of individual countries of citizenship and only data on total population and on nationals could be useful for this project. Estimates of the 2002–2006 stock of the EU27 citizens cannot be prepared using the LFS tables in the Eurostat database. These considerations need to be taken into account when proposing computation methods for the current study.

3 Proposed methods of estimating population stocks by citizenship, sex and age

The current section presents a theoretical background of the methods proposed for the calculations of the missing elements in population stocks by age, sex and citizenship group. After a brief summary of the notation, the following methods are discussed: interpolation of five-year into one-year age groups, regarded as a preparatory method (Section 3.2), followed by cohort-wise interpolation of population stocks (3.3), cohort-component projections, traditionally used in demography (3.4) and cohort-wise weights propagation (3.5). Further, Section 3.6 describes selected proportional fitting methods, which category encompasses three approaches, depending on the availability of information: the proportional adjustment, direct proportional fitting and iterative proportional fitting. Section 3 concludes by presenting some auxiliary methods for dealing with the Unknown categories, and for the estimation of missing elements of age distributions (3.7).

¹ Research project *THESIM: Towards Harmonised European Statistics on International Migration*, funded by the European Commission through the Sixth Framework Programme and executed by a research consortium led by Groupe d'étude de démographie appliquée (GèDAP), Université Catholique de Louvain.

Country	2002	2003	2004	2005	2006
Austria	+	+	+	+	-
Belgium	+	x	-	-	-
Bulgaria	-	-	-	-	-
Cyprus	dref	-	-	-	-
Czech Republic	+	+	+	+	+
Denmark	+	+	+	+	+
Estonia	na	na	na	na	-
Finland	+	+	+	+	+
France	-	-	-	-	-
Germany	for, broad age, ±agesex, i	±age, i	for, i	i	p, i
Greece	-	-	i	for, ±sex	for
Hungary	for	for	+	+	+
Ireland	p, ±ctz, broad age, dref	p, ±ctz, broad age, dref	p, ±ctz, broad age, dref	p, ±ctz, broad age, dref	p, ±ctz, broad age, dref
Italy	dref	-	-	-age	-age
Latvia	-age	+	+	+	+
Lithuania	-	-ctz	-ctz	+	+
Luxembourg	-	-	tot, nat	±ctz, ±age, ±sex	±ctz, ±age, ±sex
Malta	-	-	-	-	-
Netherlands	+	+	+	+	+
Poland	dref	-	-	-ctz	-
Portugal	p, for, -age	p, for	-	-	-
Romania	dref	-	+	+	+
Slovakia	-	-	for	i	i
Slovenia	+	+	+	+	+
Spain	+	-	p	+	+
Sweden	+	+	+	+	+
United Kingdom	-	±ctz, dref	±ctz, dref, a70	±ctz, broad age, dref	-
Iceland	+	+	-	-	-
Lichtenstein	-	-	-	-	-
Norway	+	+	+	+	+
Switzerland	+	+	+	+	+

+ data provided to Eurostat; - data not provided to Eurostat; **-age** no disaggregation by age; **-ctz** no disaggregation by citizenship; **±age** age disaggregation only for a few citizenship categories; **±agesex** disaggregation by age not provided for Males and Females; **±ctz** data provided for a few citizenship categories; **±sex** disaggregation by sex provided for a few citizenship categories only; **a70** age provided only until 70 years, with the open-ended group 70+; **broad age** data disaggregated by broad age groups; **dref** reference date different than 1st January; **for** data provided for foreigners only; **i** data inconsistency problems; **na** data not available; **nat** data provided for nationals; **p** provisional data; **x** problems detected in the data sent by the NSI; **tot** data provided for Total.

Table 1: Availability of data on population stock by citizenship, sex and age in the JMQ, 31 countries, as of 1st January 2002–2006.

3.1 Notation and basic concepts

Throughout the paper, the notation used for population variables follows a common convention presented below. In all cases, the superscript n indicates one of the three broad groups of citizenship: nationals, EU27 foreigners or non-EU27 foreigners, abbreviated as N , EU and nEU , respectively, thus reflecting the composition of the European Union as of 1st January 2007. The non-EU27 group includes also the stateless persons. An abbreviation FOR is used for all foreign population (EU27 and non-EU27 together). For the transparency of presentation, the country index is skipped, as all calculations proposed in the paper are always country-specific. The variables in question are as follows:

Stock variables:

$P^n(x, t)$ - Population in broad citizenship group n , in the age of x years on 1st January, year t .

$P^n(x, c)$ - Population in broad citizenship group n , in the age of x years at the census date c .

Event variables:

$B^n(t)$ - Births during calendar year t in citizenship group n ;

$D^n(x, t)$ - Deaths of persons aged x years, belonging to citizenship group n , during calendar year t ;

$I^n(x, t)$ - Registered immigration of persons in citizenship group n , aged x years, during calendar year t , regardless of the country of origin;

$E^n(x, t)$ - Registered emigration of persons in citizenship group n , aged x years, during calendar year t , regardless of the country of destination;

$R^n(x, t)$ - Outcome of the regularisation of the status of formerly irregular residents (cf. [4]) aged x , in year t , by definition referring only to foreigners, $n \in \{EU, nEU\}$, thus with $R^N(x, t) \equiv 0$;

$S^n(x, t)$ - Statistical adjustment (official correction) of the size of population aged x , in year t , due to the reasons other than regularisations;

$A^n(x, t)$ - Acquisitions of citizenship by the population aged x , in year t , by definition referring only to foreigners, $n \in \{EU, nEU\}$, with $A^N(x, t) \equiv 0$.

Unless noted otherwise, age is reported in years *reached* during a given calendar year, and thus the *events* in question (deaths, migrations, citizenship changes, etc.) correspond to parallelograms with vertical sides on the Lexis diagram. An illustration of the relevant concepts on a Lexis plane is shown in Figure 2, in Section 3.4.

Whenever necessary, the index denoting sex is added as an additional subscript $g \in \{m, f\}$ for males and females, respectively, e.g. $P_g^n(x, t)$ refers to female population stock, and $D_m^n(x, t)$ to deaths among males. In order to distinguish five-year age groups, an additional left-hand side subscript '5' is added. For example, ${}_5P_m^n(x, t)$ refers to male population belonging to broad citizenship group n which was in the age of $[x, x+5)$ years on 1st January of year t . The same principle applies to almost all event variables (D , I , E , R , S and A), with a clear exception of B .

In some instances, for the clarity of presentation, the summation of a particular variable over a given index is indicated by an asterisk in a respective place, e.g. $A^{nEU}(*, t) = \sum_x A^{nEU}(x, t)$ refers to all acquisitions of citizenship by non-EU27 foreigners in year t , regardless of age. Similarly, $I^*(x, t) = \sum_n I^n(x, t)$ denotes all immigrants aged x , in year t , irrespective of their citizenship, and $D^*(*, t) = \sum_n \sum_x D^n(x, t)$ refers to all deaths registered in year t , without respect to nationality or age. It has to be noted that in several cases the summation over n involves only two components, e.g. $n \in \{EU, nEU\}$ for $R^n(x, t)$ and $A^n(x, t)$.

3.2 Interpolation of five-year age groups into one-year groups

Among the preparatory steps for the estimation of missing data, the most frequent problem concerns disaggregation of five-year age groups of population (or events) into single years. This has to be performed in order to enable cohort-wise interpolations, cohort-component projections with yearly steps, or cohort-wise weights propagation, as described in Sections 3.3, 3.4 and 3.5.

If auxiliary information is available from a different source (e.g. from a census, from the previous or next year, etc.), the population size or the number of events can be disaggregated using a 'Prorating' method [11, p. 5-61], whereby the relative distribution from the auxiliary source is imposed on the data in question. The obtained distribution might need to be further adjusted to marginal totals, by means of proportional fitting procedures, described in Section 3.6.

If the data on population stocks by sex, broad citizenship group and five-year age group ${}_5P^n(x, t)$ are available and the stocks by sex and one-year age group $P^*(x, t)$ are also known, then, assuming no other information about the distribution by single years, we can estimate the missing distributions for particular citizenship groups proportionally, that is as: $P^n(x+i, t) = {}_5P^n(x, t) \cdot P^*(x+i, t) / {}_5P^*(x, t)$. This is an example of the application of the direct proportional fitting described in Section 3.6.2.

If none of the above information is available, the proposed methodology is to use the well-known interpolative four-term third-difference solution of Karup and King [11, p. 5-65]. For each five-year group, the disaggregation into fifths is done via applying multiplicative coefficients to the global value of this group and the neighbouring ones. Different multipliers are used for the first group, the middle groups and the last group, as set forth in Table 2. For example, if we want to split a middle five-year group with population N_i into five single-year groups n_1, n_2, n_3, n_4, n_5 , then:

$$\begin{aligned} n_1 &= 0.064 N_{i-1} + 0.152 N_i - 0.016 N_{i+1}, \\ n_2 &= 0.008 N_{i-1} + 0.224 N_i - 0.032 N_{i+1}, \text{ etc.} \end{aligned}$$

When Karup-King multipliers are used, the condition $N_i = n_1 + n_2 + n_3 + n_4 + n_5$ is automatically fulfilled.

	First group, N_0			Middle groups, N_i			Last group, N_k		
	N_0	N_1	N_2	N_{i-1}	N_i	N_{i+1}	N_{k-2}	N_{k-1}	N_k
First fifth	+0.344	-0.208	+0.064	+0.064	+0.152	-0.016	-0.016	+0.112	+0.104
Second fifth	+0.248	-0.056	+0.008	+0.008	+0.224	-0.032	-0.032	+0.104	+0.128
Third fifth	+0.176	+0.048	-0.024	-0.024	+0.248	-0.024	-0.024	+0.048	+0.176
Fourth fifth	+0.128	+0.104	-0.032	-0.032	+0.224	+0.008	+0.008	-0.056	+0.248
Last fifth	+0.104	+0.112	-0.016	-0.016	+0.152	+0.064	+0.064	-0.208	+0.344

Source: [11], Table C-1, p. 5-69.

Table 2: Coefficients for the Karup-King interpolation formula.

As an alternative to the Karup-King interpolation, the six-term fifth-difference interpolative formulae of Sprague or Beers can be applied, which however use information from more surrounding groups. Methodological details can be found in Shryock et al. [11]. In our case, the Karup-King interpolation is recommended for the sake of simplicity.

For variables depicting non-vital events, like migration or citizenship acquisitions, the estimates for particular cohorts can be obtained from two neighbouring period-age estimates yielded by the Karup-King formula, split equally by half. For the first cohort, we can assume that a half of the relevant period-age events concern the cohort in question, while for the last, open-ended cohort, we can add up the period-age estimate for the open-ended group and a half of the events concerning the age group immediately preceding the last one. The underlying rationale is an assumption that non-vital events are equally spread over the period-age squares of the Lexis diagram. In any case, the estimates for the eldest cohorts would be close to zero for all practical migration-related applications.

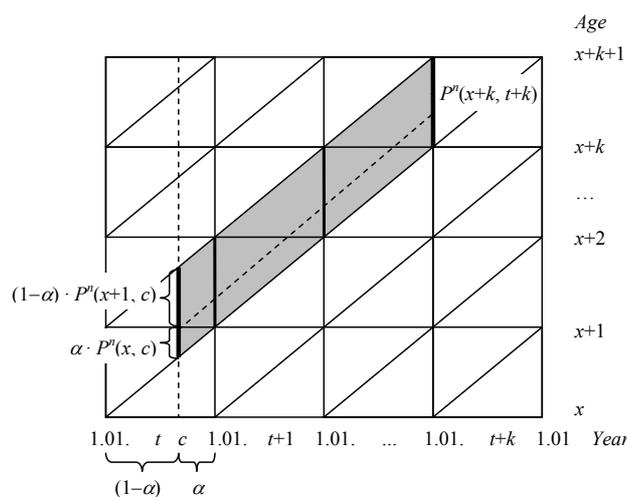
Regardless of the method, if the disaggregation is performed on data broken down by sex or citizenship, the final estimate might need to be obtained by proportional fitting methods (described in Section 3.6), in order to ensure the summation to available marginal totals.

3.3 Cohort-wise interpolation of population stocks

Given the information on the age structures of the population for two non-adjacent moments of time, a simple idea to obtain the missing figures for in-between moments would be to apply interpolation techniques. In this case, we propose cohort-wise interpolation for all cohorts apart from the youngest and oldest one, which are discussed separately. Overall, this method requires much less information on input than the cohort-component projections presented in the next section, but it requires information about population both before and after the moment for which the estimates are to be done. The interpolative approach is recommended for the cases where (a) the span between two points with available data is not wide (say, two-three years), and (b) no information on the distribution of vital and migratory events by citizenship is available.

In practical applications, as the ones described in Section 4, it often happens that the data are available for year t from the census conducted at time c ($t \leq c < t+1$), and for 1st January of the year $t+k$, not immediately following the census. Such situations can be put in a general framework illustrated on a Lexis diagram in Figure 1, where α denotes the fraction of a year remaining after the census until 31st December. Figure 1 and the methodology proposed below cover also the situations when data come from other sources than the census, and the situations when the reference date of the data for year t is 1st January. In the latter case it suffices to set $\alpha = 1$.

For the cohorts already existent at the census date c , the interpolation can follow various patterns, but an arithmetic and geometric pattern of growth [3, 10] will be the most frequent choices. As noted by Rowland [10, p. 50], “under arithmetic growth, successive population totals differ from one another by a constant amount [, while] under geometric growth they differ by a constant ratio”. For short-period interpolations, both approaches should yield similar results, although this is an empirical issue, and there is no convincing argument in favour of either of them. Hence, a selection of appropriate methods should rely on case-specific judgements.



Source: Own elaboration.

Figure 1: Cohort-wise interpolation of population stocks: a general idea.

It has to be noted that the cohort aged x completed years on 1st January $t+k$ was split at the census date between two age groups: the younger one (aged x completed years) and the older (aged $x+1$), as shown in Figure 1. Therefore, the interpolative estimate of $P^n(x, t+i)$ depends on $P^n(x, c)$, $P^n(x+1, c)$ and $P^n(x, t+k)$.

Given the above, the formula for an interpolative estimate of population sizes belonging to a particular age group $x+i$ and citizenship group n , assuming the linear pattern of change, is as follows:

$$P^n(x+i, t+i) = (k-i) / (k-1+\alpha) \cdot [\alpha \cdot P^n(x, c) + (1-\alpha) \cdot P^n(x+1, c)] + (i-1+\alpha) / (k-1+\alpha) \cdot P^n(x+k, t+k), \quad (1a)$$

while for the geometric change:

$$P^n(x+i, t+i) = \{[\alpha \cdot P^n(x, c) + (1-\alpha) \cdot P^n(x+1, c)]^{k-i} \cdot P^n(x+k, t+k)^{i-1+\alpha}\}^{1/(k-1+\alpha)}. \quad (1b)$$

For the youngest and oldest cohorts, for which interpolation as proposed above is not possible, a simplified solution is proposed. In such cases, we suggest to take the average *shares* (proportions) of the sizes of the respective age groups in the total population, calculated from the data available for neighbouring periods, weighted by the distance between the available data points and estimation point.

In order to ensure consistency of the results and summation of the age-specific estimates to the marginal totals by sex or citizenship group, whenever available, the estimates have to be adjusted by the means of proportional fitting, presented in Section 3.6.

The framework presented above can be easily generalised to a much less frequent situation with interpolation between two censuses – in such case, a fraction β of a year between the 1st January of the year of the second census and the second census date, c' , should be additionally accounted for. However, the estimates obtained in such cases would be only very approximate, due to a usually large time span between the censuses.

It should be noted that an identical solution as shown above in (1a), or in (1b) can be used for extrapolating cohort sizes *beyond* the available data points, in whichever direction. In either case, it would suffice to put an appropriate integer $i \leq 0$ for the backward extrapolation (in particular, following the example from Figure 1, set $i = 0$ to obtain values for the beginning of the census year), or $i > k$ for the forward extrapolation.

The methods discussed above resemble to some extent the ones presented in the *Human Mortality Database Methods Protocol* [15], with the exception of the oldest age groups, where the quoted study suggests more sophisticated extinct cohort and survivor ratios approaches. Direct application of the methods proposed by Wilmoth et al. [15] would be, however, difficult. This is not because of computational reasons, but rather due to the lack of yearly estimates of deaths, births and migratory events broken down by citizenship groups, which has been listed at the beginning of the current section as a precondition for selecting cohort-wise interpolation method.

3.4 Cohort-component projections

As concerns projections, let us denote by $X^n(x, t)$ a sum of all event variables *not* related to the natural change of population stocks (i.e. all but births and deaths), thus:

$$X^n(x, t) = I^n(x, t) - E^n(x, t) + S^n(x, t) + \sum_{k \in \{EU, nEU\}} A^k(x, t), \quad \text{for } n = N; \quad (2a)$$

$$X^n(x, t) = I^n(x, t) - E^n(x, t) + S^n(x, t) + R^n(x, t) - A^n(x, t), \quad \text{for } n \neq N. \quad (2b)$$

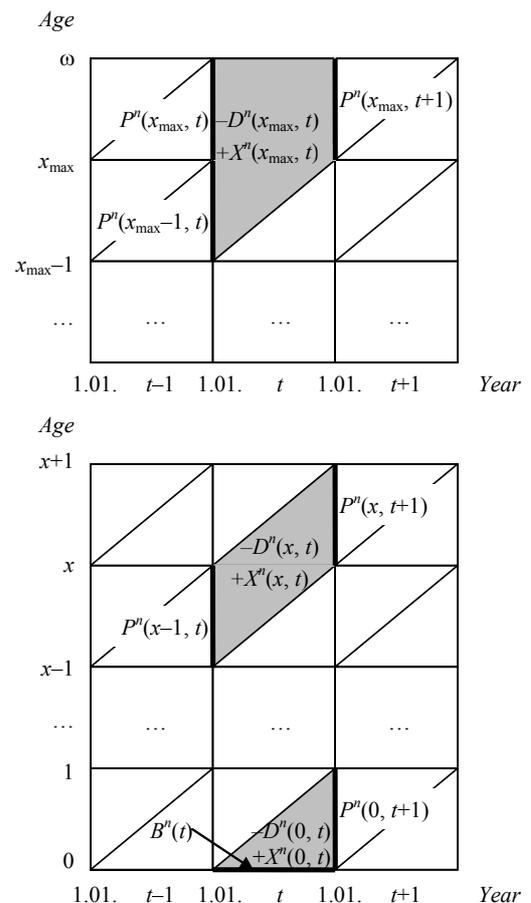
Given (2a) and (2b), the population accounting equations for each broad citizenship group are:

$$P^n(0, t+1) = B^n(t) - D^n(0, t) + X^n(0, t); \quad (3a)$$

$$P^n(x, t+1) = P^n(x-1, t) - D^n(x, t) + X^n(x, t), \quad \text{for } x \in \{1, 2, \dots, x_{\max}-1\}; \quad (3b)$$

$$P^n(x_{\max}, t+1) = [P^n(x_{\max}-1, t) + P^n(x_{\max}, t)] - D^n(x_{\max}, t) + X^n(x_{\max}, t). \quad (3c)$$

In (3c), x_{\max} stands for the highest (open-ended) age group for which information is available. Note also that deaths and other event variables in age group x_{\max} refer to the trapezoid on the Lexis diagram rather than to a parallelogram, while for age group 0 – to a right triangle, as shown in Figure 2.



Source: Own elaboration.

Figure 2: Relationships between population stocks P^n , and events B^n , D^n and X^n on a Lexis diagram.

Under the assumptions presented above, the projection is made following the equations (3a), (3b) and (3c) for consecutive years, on the basis of information available for single-year age groups, decomposed from the five-year groups, if needed.

Note that the default citizenship of a newborn child can differ between the countries, either following the *ius soli* principle, whereby a child acquires the citizenship of the country of birth, or *ius sanguinis*, according to which a child inherits the citizenship of its parent(s), or finally a mixture of those two, for example differentiating between the generations of migrants, taking into account the length of stay in the country, etc. The general rules are as follows:

a) *Ius sanguinis*

If the child gets citizenship of any of the parents, then $B^n(t)$ in equation (3a) may be assumed to be roughly proportional to $P^n(t)$. If the child acquires citizenship of the mother and we have no separate estimate of fertility for nationals and foreigners, then $B^n(t)$ may be assumed to be roughly proportional to $P_f^n(t)$. If the estimates of fertility by broad citizenship and age of mother exist then a better estimate may be obtained using the formula:

$$B^n(t) = B^*(t) \sum_x f^n(x) P_f^n(x, t) / \sum_{k,x} f^k(x) P_f^k(x, t), \quad (4)$$

where $f^n(x)$ denotes age-specific fertility rates for women in age group x , belonging to the group of citizenship n . If the estimates of fertility are available by broad citizenship group, but not by the age of mother, the formula (4) would have to be modified, so as the summation over age reflects only the female population aged 15–49 years.

b) *Ius soli*

If the child automatically acquires the citizenship of a given country, then the balance equation for the youngest age group, (3a), becomes, depending on the citizenship in question:

$$P^N(0, t+1) = B^*(t) - D^N(0, t) + X^N(0, t), \text{ for } n = N; \quad (3a')$$

$$P^n(0, t+1) = X^n(0, t) - D^n(0, t), \text{ for } n \neq N. \quad (3a'')$$

In mixed cases, it is recommended to project one part of births according to formulas for *ius soli* and another part according to the *ius sanguinis* principle.

Note also that losses of citizenship are not accounted for, as they in most instances concern persons in reality either already living abroad, or emigrating (and counted in E). For acquisitions of citizenship, we assume that non-nationals fall in the category of nationals upon naturalization, in order to count the same people only once, regardless of the number of citizenships they have.

If the breakdown by citizenship group of all variables referring to vital and migratory events can be assumed proportional to the citizenship structure of the population at the beginning of each year, then the projection methodology can be often *de facto* simplified to proportional adjustment / decomposition, whereby the citizenship distribution of the considered cohort in the previous year would directly apply to all cohorts except the

first and the last one in each year. In particular, this situation applies if the following four conditions hold:

1. Total population by age, ${}_5P^*(x, t)$, is known for successive years, but the citizenship structure is missing;
2. We may assume that the distribution of deaths and migration flows by broad citizenship is the same as the citizenship composition of the population;
3. Acquisitions of citizenship may be ignored;
4. There was no regularization, or it may be ignored.

In such cases, the projection equation (3b) combined with proportional fitting is equivalent to proportional decomposition of ${}_5P^*(x, t)$ by citizenship group described in Section 3.6.1. The estimations can be performed using the formula:

$${}_5P^n(x, t) = {}_5P^*(x, t) \cdot {}_5P^n(x-1, t-1) / {}_5P^*(x-1, t-1). \quad (5)$$

The first and the last cohort may be disaggregated using the citizenship composition of the first and last age group in the previous year. In such cases, the following formulas apply:

$${}_5P^n(0, t) = {}_5P^*(0, t) \cdot {}_5P^n(0, t-1) / {}_5P^*(0, t-1), \text{ or:} \quad (6a)$$

$${}_5P^n(x_{\max}, t) = {}_5P^*(x_{\max}, t) \cdot {}_5P^n(x_{\max}, t-1) / {}_5P^*(x_{\max}, t-1). \quad (6b)$$

3.5 Cohort-wise weights propagation

In some cases, too much information on the age-sex-citizenship distribution of the components of population change is missing, which renders projections too dubious with respect to the number of assumptions that need to be made. In practice, in such instances the only reliable information comes from the population census and from annual population stocks available in the DEMO domain of the NewCronos database. Hence, the proposed procedure is as follows.

For the census population, apply the structure by citizenship, taken from each five-year age group, to the respective single-year age groups (i.e. from age group 0–4 to single ages 0, 1, ..., 4; from 5–9 to 5, 6, ..., 9 etc.). Let $w^n(x, c) = P^n(x, c) / P(x, c)$ denote the age-specific shares ('weights') of citizenship group n in the census.

Further, set α as a fraction of the calendar year before the census date. It is implicitly assumed that the census population in single-year age groups can be divided between 'older' and 'younger' cohorts using the α and $(1-\alpha)$ partition. For the census date, use the following formula to calculate the share of citizenship group n in the cohort that was aged x years on 1st January of the census year:

$$w^n(x+\alpha, c) = [(1-\alpha) \cdot P^n(x, c) + \alpha \cdot P^n(x+1, c)] / [(1-\alpha) \cdot P^*(x, c) + \alpha \cdot P^*(x+1, c)], \text{ for } x < x_{\max}; \quad (7a)$$

$$w^n(x_{\max}+\alpha, c) = P^n(x_{\max}, c) / P^*(x_{\max}, c). \quad (7b)$$

For the 1st January of the census year assume that the weights $w^n(x, t) = w^n(x+\alpha, c)$. For the 1st January of the year following the census year ($t > c$), assume in turn:

$$w^n(x, t) = w^n(x-1+\alpha, c), \text{ for } 0 < x < x_{\max}; \quad (8a)$$

$$w^n(x_{\max}, t) = [P^n(x_{\max}-1, c) (1-\alpha) + P^n(x_{\max}, c)] / [P^*(x_{\max}-1, c) (1-\alpha) + P^*(x_{\max}, c)]. \quad (8b)$$

For the youngest age group assume $w^n(x, t) = w^n(0, c)$, or alternatively that the shares are the same as the shares of citizenship group n in the births during the census year, so as: $w^n(0, t) = B^n(t-1) / B^*(t-1)$. For consecutive years calculate:

$$w^n(x, t) = w^n(x-1, t-1), \text{ for } x = 1, \dots, x_{\max}-1; \quad (9a)$$

$$w^n(x_{\max}, t) = [P^n(x_{\max}-1, t-1) + P^n(x_{\max}, t-1)] / [P^*(x_{\max}-1, t-1) + P^*(x_{\max}, t-1)]; \quad (9b)$$

$$w^n(0, t) = w^n(0, t-1), \quad (9c)$$

or, as an alternative to (9c): $w^n(0, t) = B^n(t-1) / B^*(t-1)$.

Subsequently, calculate populations for all years using the above shares and total populations (available e.g. from DEMO), as: $P^n(x, t) = P^*(x, t) \cdot w^n(x, t)$. Finally, aggregate single-year age groups into five-year ones.

3.6 Proportional fitting methods

In the proportional fitting methods, the general task is to estimate $P_g^n(x, t)$, i.e. the elements of a three-dimensional cube (with the dimensions being sex, age and citizenship). The choice of a particular method depends on which marginal information (cube's edges or faces) is known and if an initial estimate of the cube elements are available. Below, the examples of frequent situations are presented. The formulas have been given for population by single years of age but the analogical formulas apply to population by five-year age group. For more information on multi-proportional techniques see for example the studies of Willekens [12, 13], Willekens et al. [14], Rees [8] and Norman [6]. Note that proportional fitting methods presented below are known under various names in the scientific literature.

In addition to a potential application as the main estimation method, proportional fitting may be used, in almost all the countries for which estimations are needed, as the final stage of the estimation procedure, in order to adjust the initial estimates to known aggregates or marginal totals. The initial estimate might be obtained for example using interpolation or projection, or assumed to be the same as at some different time (e.g. the same as at the census date). Such an initial estimate has to be subsequently adjusted for example to the known total population size by age and sex.

3.6.1 Proportional adjustment / decomposition

Among the proportional fitting methods, the simplest one can be applied to situations, when a population can be directly disaggregated by a variable (sex, age or citizenship), according to the pattern observed in an auxiliary source. In general, the idea is the same as in the Prorating method [11, p. 5-61] mentioned in Section 3.2.

For example, if the aggregates $P_g^*(x, t)$ and an initial estimate of the citizenship structure $P_g'^n(x, t)$ are known, then the final estimate $P_g''^n(x, t)$ may be obtained as:

$$P_g''^n(x, t) = P_g'^n(x, t) \cdot P_g^*(x, t) / P_g^*(x, t). \quad (10)$$

In particular, if one wants to estimate the breakdown by citizenship using the citizenship structure taken from the census, $P_g^n(x, c)$, then (10) becomes: $P_g^n(x, t) = P_g^*(x, t) \cdot P_g^n(x, c) / P_g^*(x, c)$.

3.6.2 Direct proportional fitting

The estimation problem becomes slightly more complicated, if one wants to estimate $P_g^n(x, t)$, but does not have any initial estimate of it. One possible situation is that at least some fragments of the data cube (faces and/or edges) are available and provide coherent information (sum up to the same totals). In such cases, the most straightforward solution is provided by a direct proportional fitting method, whereby the missing elements (i.e. the inside of the cube) can be obtained by taking simple proportions to all available marginal totals.

For example, let the available data consist of known $P_g^*(x, t)$ and $P^n(*, t)$, i.e. the age-sex face and the citizenship edge of the age-sex-nationality cube. Then, the sought-for $P_g^n(x, t)$ can be estimated as:

$$P_g^n(x, t) = P_g^*(x, t) \cdot P^n(*, t) / P^n(*, t) \quad (11)$$

In practical applications discussed in Section 4, this option was used rather infrequently, because there usually are some initial estimates of the population structures, for example from the census. Willekens et al. [14, p. 97] noted that general formulae of a form akin to (11) for a one face – one edge problem, as well as similar closed-form solutions for the cases with three edges or two faces are the solutions of the entropy-maximisation problems in research tasks aimed at reconstructing the elements of a three-dimensional arrays, given the available marginal totals.

3.6.3 Iterative proportional fitting

In a general case, a closed-form solution (11) may not exist due to possible incoherence between various data at hand. Such problems call for a multi-step iterative proportional fitting (IPF) method, whereby the solutions are sought step-wise, through iterative adjustments of their successive approximations to marginal totals available from the faces or edges of the data cube. In particular, this method can be used for adjusting the preliminary joint distributions to the known marginal distributions.

For example, let the initial estimate of the citizenship structure $P_g'^n(x, t)$ be known, as well as the sex-age face and the citizenship edge of the data cube, respectively $P_g^*(x, t)$ and $P^n(*, t)$. By the IPF algorithm, the initial estimates are iteratively corrected by proportional adjustment. An additional superscript (k) in $P_g^{(k)n}(x, t)$ denotes the iteration step (for $k \geq 1$). The starting value $k = 1$ defines also the initial estimate of the joint sex-age-citizenship distribution, $P_g^{(1)n}(x, t) = P_g'^n(x, t)$. Subsequent steps are computed as follows:

$$P_g^{(2k)n}(x, t) = P_g^{(2k-1)n}(x, t) \cdot P_g^*(x, t) / P_g^{(2k-1)n}(x, t); \quad (12a)$$

$$P_g^{(2k+1)n}(x, t) = P_g^{(2k)n}(x, t) \cdot P^n(*, t) / P^n(*, t). \quad (12b)$$

The procedure defined by (12a) and (12b) is repeated iteratively till some convergence criterion is achieved. For example, the estimates yielded by consecutive steps

should differ by no more than by an arbitrarily-selected small number ε . More details of the method have been discussed by Willekens [13, pp. 69–71], Willekens et al. [14], Rees [8] and Norman [6].

Although the IPF method is purely mechanical, its main advantage is that it does not require any additional information (such as data on vital events or migration) or excessive labour resources, and the obtained results (in terms of joint distributions by all variables under study) are automatically coherent with marginal distributions of particular variables. Moreover, under some general assumptions, the IPF estimates can be interpreted from a statistical viewpoint as joint probability distributions obtained using the maximum likelihood or entropy maximisation methods [2, pp. 83–97; after: 13, p. 70].

3.7 Auxiliary methods

Among the auxiliary methods proposed in the current study, the foremost one is the decomposition of the *Unknown* category wherever it appears (i.e., with respect to age, citizenship, or even sex, as in the case of Greece for 2005). The universal solution proposed in such cases is a proportional disaggregation: population belonging to the *Unknown* category is broken down proportionally to the existing, well defined categories (citizenship groups, age groups, etc.) and the resulting parts are attached to these categories. For example, if total population P consists of n well-defined groups P_1, \dots, P_n , and the *Unknown* category, P_{unk} , such that $P = \sum_i P_i + P_{unk}$, where $i = 1, \dots, n$, then the following corrections apply:

$$P'_j = P_j + P_{unk} \cdot P_j / \sum_i P_i = P_j (1 + P_{unk} / \sum_i P_i), \text{ for all } j, \text{ with } i = 1, \dots, n. \tag{13}$$

If some elements of age structures are missing (e.g. tails of respective age distributions, or a breakdown into five-year groups given the availability of broader ones), we may either use a structure from a different year or fit a mathematical function to available data. For example, we can assume that foreign population stocks are a double-exponential function of age, as originally proposed for the intensity of migration flows by Rogers and Castro [5, 9]. The number of foreign population aged x , $\phi(x)$, would then be given by the following equation:

$$\phi(x) = c + a_1 \cdot \exp(-\alpha_1 \cdot x) + a_2 \cdot \exp\{-\alpha_2 \cdot (x - \mu_2) + \exp[-\lambda_2 \cdot (x - \mu_2)]\}. \tag{14}$$

The parameters $c, a_1, \alpha_1, a_2, \alpha_2, \lambda_2$ and μ_2 can be estimated separately for each sex, for example using the ordinary least squares method (OLS) on the basis of the data for the available age groups (for example, below 65 years of age). Technically, the calculations can be done in a spreadsheet (e.g. MS Excel) using a solver-like tool, controlling for sensitivity of the algorithm to the choice of initial input values. Based on the obtained parameter estimates, formula (14) yields approximations of $\phi(x)$ for the remaining age groups. The last, open-ended group (85+) can be obtained by subtraction of all other figures from the total. To avoid negative numbers in the 85+

category, appropriate constraints should be set during the estimation procedure.

In either case, when adjustment to broader age groups is needed in order to ensure summation to respective totals (e.g. for functional age groups), it can be done via proportional fitting presented in Section 3.6.

4 Estimating population stock for EU27 and EFTA countries

The current section briefly summarises the algorithm for the selection of an appropriate method of computations for a given country (Section 4.1), followed by a brief illustration of the proposed approach employed for the 31 countries under study, and a selection of the results (4.2).

4.1 Procedure for selecting an estimation method

In the light of the overview of data availability presented in Section 2 and the methodological discussion presented in Section 3, it is suggested to inspect the following general options of data availability, in order to apply the relevant data estimation procedures:

Option 1. All the required data are available in the Eurostat database

Whenever all data are available in the Eurostat database, the following five-step procedure is recommended:

1. Organize the data in a database;
2. Verify the data (perform data validation and internal consistency checks);
3. Deal with the *Unknown* categories (if applicable);
4. Calculate the required aggregates;
5. Check the results.

This option includes cases when there is a need for combining data from various parts of the Eurostat database (e.g. in DEMO and in JMQ), and the cases where there is an ‘Unknown’ category, which has to be disaggregated proportionally among the well specified categories, as described in Section 3.7 on ‘Auxilliary methods’.

Option 2. Some of the data missing in the Eurostat database can be obtained from the respective NSI or from other sources

In this case, two situations are possible:

Option 2a. All the missing data may be obtained without contacting the NSI

If all the missing information is publicly available, for example from the NSI webpage, it should be downloaded and combined with the Eurostat data. Such an overall dataset should be then subject to a procedure described under Option 1, points 2. through 5.

Option 2b. Some (or all) missing data are (or are suspected to be) available either from the NSI or from other sources

If some missing information is downloadable from sources like the NSI webpage, it should be collected and merged with the Eurostat data. Nonetheless, there are cases when data are not publicly available but it can be suspected that either some, or even all the missing information is in the possession of the NSI. In such case, the undertaken actions should be as follows:

1. Contact the NSI in order to obtain the missing data. If successful, proceed as in **Option 2a**;
2. For the data that are still unavailable, but can be estimated, proceed as in **Option 3**;
3. For the data that are still not available and cannot be estimated, look at **Option 4**.

Option 2 includes cases when data from various national sources has to be combined, for example aggregated data obtained using the component method and data on citizenship composition from the register of foreigners.

Option 3. Some data are not available anywhere but can be estimated

Even if some auxiliary data can be collected from whichever source, in many instances the available information can be still incomplete. In a vast majority of such cases, the missing information can be still estimated, either following the methodological guidelines and techniques outlined in Section 3, or by means of more straightforward and easy-to-apply solutions. The order of undertaken actions is then as follows:

1. Organize the available data (from Eurostat and other sources);
2. Verify the data (perform data validation and internal consistency checks);
3. Deal with the *Unknown* categories (if applicable);
4. Calculate the required aggregates for the available data;
5. For each year for which data are missing select the best method to estimate missing data;
6. Collect supplementary data needed for the estimations;
7. Estimate the missing data;
8. Check the results.

For the estimations (item 7), various methods can be used, depending on the range and type of the missing information and data availability. In general, five broad groups of methods can be distinguished here, following the outline presented in Section 3:

- a. Proportional fitting methods (Section 3.6);
- b. Cohort-wise weights propagation (Section 3.5);
- c. Cohort-wise interpolation of population stocks (Section 3.3);
- d. Cohort-component projections (Section 3.4);
- e. Other solutions, not listed above, or combined approaches.

The methodology of interpolating the five-year into one-year age groups, presented in Section 3.2, as well as

some methods described in Section 3.7 should be treated as auxiliary to all the remaining ones rather than constituting separate estimation methods *per se*.

Option 4. Data are not available, and no or only very rough estimates can be produced

In principle, this should be a very infrequent option. If no information is available that would enable estimation under Option 3, none or only very rough approximations can be performed, such as for example the 50-50 division of all foreigners into the EU27 and non-EU27 categories.

Under Option 3, in all the cases where several methods could be alternatively applied, preference is given to the more straightforward ones, and definitely to the ones having less judgemental elements, thus less potential sources of error. This approach conforms to the *Occam's razor* principle, stating that “entities are not to be multiplied beyond necessity”², which in this case means that the proposed models should not be more sophisticated than necessary, due to various possible sources of error. For example, given complete data on stocks by age and citizenship group for two moments of time (e.g. from successive population censuses), if the data on flows (*I* and *E*), natural change (*B* and *D*) and citizenship acquisitions (*A*) are not available by citizenship and/or age, and require estimation, then the intermediate values are recommended to be calculated using cohort-wise interpolation rather than projection. In the former case, the only source of possible error is the composition of population as such, whereas in the latter, judgemental assumptions on the relevant distributions of all components of the balance equation are likely to result in higher uncertainty of the ultimate results, which within the deterministic framework of the project is impossible to assess.

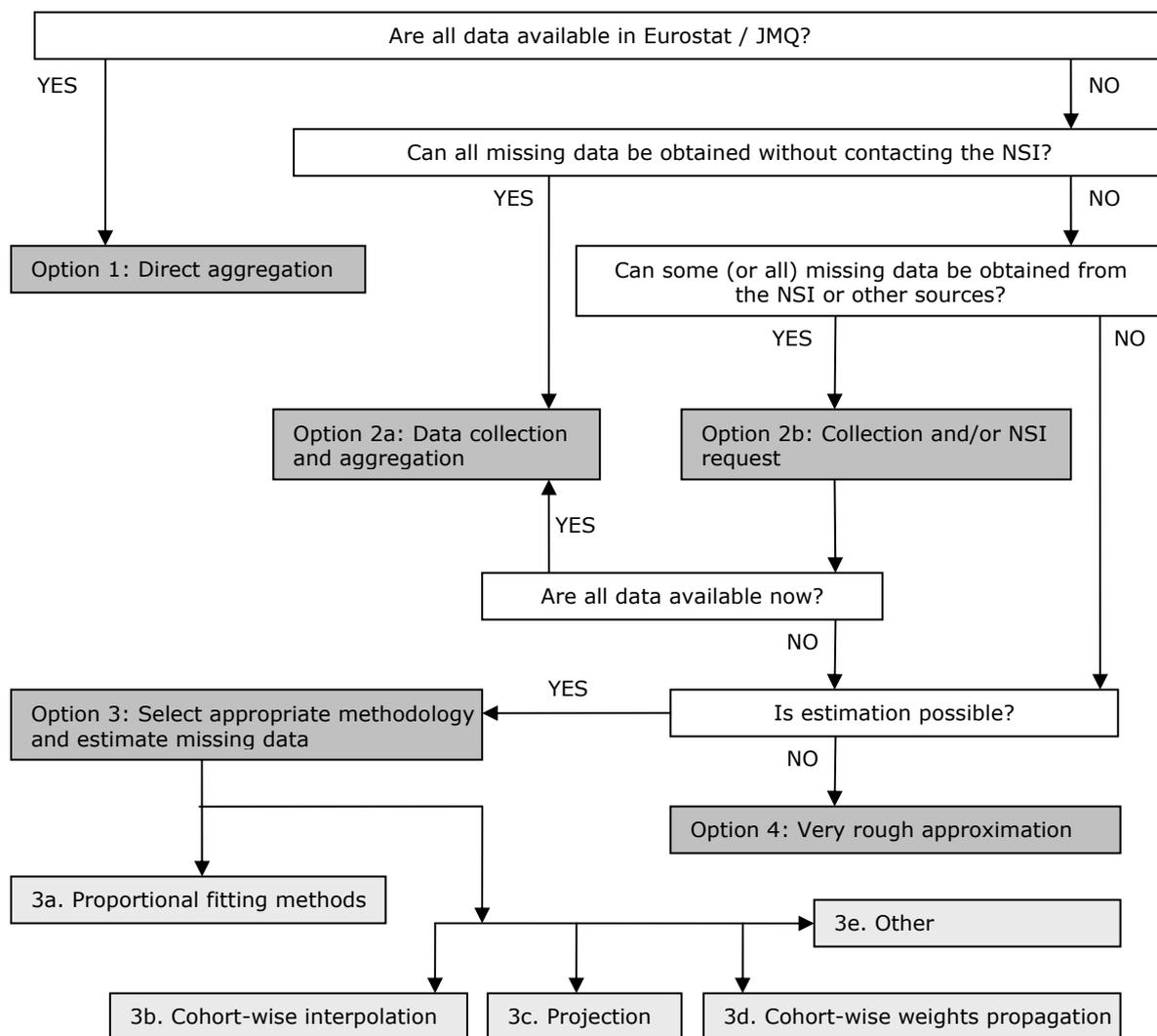
Figure 3 presents a decision tree summarising the procedure for selecting the estimation methodology, taking into account all the above options.

4.2 Application of the methodology, examples, selected results

The decision tree presented in Figure 3 has been used to select the best estimation method for each of the 31 EU and EFTA countries, accounting for the availability of data in the Eurostat database (either on-line or in the JMQs), in the NSI databases, and at other sources. It turned out that complete data needed to estimate population by broad group of citizenship, sex and age on 1st January 2002–2006 were available in the JMQs for nine countries: Austria, the Czech Republic, Denmark, Finland, Hungary³, Norway, Slovenia and Sweden.

² After: ‘Occam’s razor’, in: *Encyclopædia Britannica Online*, <http://www.britannica.com/eb/article-9056716>, accessed on 21st May 2007.

³ For Hungary, data on total population and on the number of Hungarian citizens were not always provided in the JMQ and therefore not available in the migration part of the Eurostat database. However, data on total population were available in the demographic part of the Eurostat database and the number of Hungarian citizens could be calculated directly as a difference between total population and total foreigners, the latter taken from the JMQ.



Source: Own elaboration.

Figure 3: Decision algorithm for obtaining population stocks by broad citizenship group, sex and age.

For additional four countries it was possible either to collect all the missing data from the NSI websites (Belgium and Iceland), or to get them by contacting directly the NSI (Lichtenstein and Switzerland).

For the remaining 18 countries some estimations were necessary. The method that proved to be useful in the largest number of cases was some sort of proportional fitting (one of the three versions presented in Section 3.6). It was used as the main method for estimating population by broad citizenship in Cyprus, France, Germany, Greece, Italy, Latvia, Luxembourg, Malta, Slovakia, Spain and the UK. In all cases the total population was assumed to be as reported by the NSI in their demographic statistics, while the citizenship structure was taken from varied sources, for example the JMQ data for the same year, data taken from the NSI website (Italy), the census data (Cyprus, France), the data for another year (Romania, Spain), the LFS data (Cyprus, France) or the data from the register of foreigners (Germany) (see also examples below).

The cohort-wise interpolation method was used for Ireland, Lithuania and Portugal. For Bulgaria, Estonia and Poland, where only data from the census were available, the cohort-wise weight propagation was applied. For Cyprus, Lithuania, Luxembourg and Poland it was originally planned to use a projection method, however it was decided that it would require too many assumptions that would be difficult to justify, and that the final result would not be reliable enough to justify the additional effort required when using this method.

Estimations done for Romania do not fit any of the above groups. They involved simple combination of data coming from various sources.

Below, more details about the estimation procedures are provided for selected countries. In doing so, we have tried to give an example for each estimation method. The resulting numbers in terms of the estimated citizenship structures of the populations of 18 European countries (all being EU Member States) on 1st January 2006, are presented in Table 3.

Country	Total	Nationals	EU27 foreigners	Non-EU27 foreigners
Bulgaria	7 718 750	7 693 214	3 855	21 681
Cyprus	766 414	678 114	52 217	36 084
Estonia	1 344 684	1 082 605	3 961	258 118
France	61 166 822	58 208 155	1 148 691	1 809 976
Germany	82 437 995	75 148 846	2 448 113	4 841 036
Greece	11 125 179	10 165 903	180 282	778 994
Ireland	4 209 019	3 779 755	295 165	134 099
Italy	58 751 711	56 081 197	538 853	2 131 661
Latvia	2 294 590	1 837 832	5 527	451 231
Lithuania	3 403 284	3 370 422	1 962	30 900
Luxembourg	469 086	280 938	171 876	16 273
Malta	404 962	392 850	7 022	5 090
Poland	38 157 055	38 115 920	18 660	22 476
Portugal	10 569 592	10 293 686	80 039	195 867
Romania	21 610 213	21 584 220	6 058	19 935
Slovakia	5 389 180	5 368 255	12 289	8 636
Spain	43 758 250	39 755 741	1 326 128	2 676 381
United Kingdom	60 393 100	56 990 704	1 365 190	2 036 807

Source: Own calculations based on the Eurostat and NSI data.

Table 3: Estimated population by broad group of citizenship in 18 EU countries, as of 1st January 2006.

As concerns particular examples: in **Germany**, data on foreigners come from two different sources. The component method (*Bevölkerungsfortschreibung*), based on the last traditional German census of 25th May 1987, is used by the NSI to produce annual figures on total population, total nationals and total foreigners, as well as nationals and foreigners by sex and age. The other source is the Central Register on Foreigners which contains data on foreigners by citizenship, sex and age. The total numbers of foreigners and their sex and age structures differ between both sources. In order to obtain a single set of estimates, the total number of German citizens, the total number of foreigners, as well as the age structures of Germans and foreigners were taken, following the NSI procedures, from the *Bevölkerungsfortschreibung* data. The distribution of foreigners into EU27 and non-EU27 foreigners was done in proportion to their shares in respective age groups according to the data from the Central Register of Foreigners. Thus, all in all, the proportional decomposition method was used.

In **Latvia**, no joint distribution of population by citizenship and age was available for 1st January 2002, only the structures by age and by citizenship separately. However, the full joint distribution was available for 2003. The iterative proportional fitting method was selected to deal with this case. The joint distribution by citizenship group and age on 1st January 2003 was taken as the starting point for estimating the 2002 structure of population, which was then iteratively adjusted to the known marginal totals.

Lithuania is an example for the application of a cohort-wise interpolation method. In this country, the joint distribution of population by sex, age and citizenship was available for the Census date (6th April 2001), as well as for 1st January 2005. The cohort-wise interpolation, as

described in Section 3.3, was used to obtain the initial estimates of males and females on 1st January 2002, 2003 and 2004. In the next step those initial estimates were proportionally adjusted to the known numbers of males and females by age, taken from the Eurostat demographic database.

In **Bulgaria**, annual data on population by citizenship were not available. The only information on citizenship structure came from the census of 1st March 2001. There are also annual data on population by age and sex prepared by the NSI using the component method, available from the Eurostat database. The estimates of annual 2002–2006 population by citizenship, sex and age were prepared using the cohort-wise weight propagation method. The census data were used as the starting point for calculating the initial shares (weights) of citizenship groups in each age cohort. These shares were iteratively propagated forward as described in Section 3.5 and the resulting weights were combined with the available data on population by sex and age to calculate the required joint distribution by citizenship, sex and age.

5 Conclusion

As it can be seen from the country-specific overview of problems with data on population stocks by age, sex and citizenship, there is no universal solution for estimating the missing pieces of information in the European countries under study. Nevertheless, depending on the availability of data at hand, either in the Eurostat / JMQ, or in the respective national statistical institutes, several estimation procedures can be proposed and applied, as mentioned in Sections 3 and 4.

The methods and algorithm we proposed for this purpose do not, however, consider the issue of the harmonisation of the data and definitions, as mentioned in Section 1. More work would be needed in order to recalculate the population stocks into a common definition (cf. [1, 7]), and make them consistent with the (also re-estimated) statistics on migration flows. These very important research tasks are still to be performed in the subsequent tasks of the MIMOSA research project, of which the current study forms a part.

Acknowledgement

The paper was prepared within the framework of the research project on “Modelling of statistical data on migration and migrant population” (MIMOSA), commissioned by Eurostat (contract no. 2006/S 100-106607/EN; Lot 2) to the Netherlands Interdisciplinary Demographic Institute (NIDI) and conducted jointly by NIDI, Central European Forum for Migration and Population Research (CEFMR), Groupe d'étude de démographie appliquée, Université Catholique de Louvain (GéDAP UCL) and Southampton Statistical Sciences Research Institute of the University of Southampton (S3RI). We are grateful to our colleagues for their comments and discussions. Special credits go to Peter Ekamper from NIDI for his assistance in the computational part of the current study.

References

- [1] Bilborrow, R., Hugo, G., Oberai, A.S. and Zlotnik, H. (1997). *International migration statistics. Guidelines for improving data collection systems*. International Labour Organisation, Geneva.
- [2] Bishop, Y.M.M., Fienberg, E.F. and Holland, P.W. (1975). *Discrete multivariate analysis*. MIT Press, Cambridge, MA.
- [3] Calot, G. and Sardon, J.-P. (2003). *Methodology for the calculation of Eurostat's demographic indicators*. Eurostat Working Papers and Studies, Population and Social Conditions 3/2003/F/n° 26. Eurostat, Luxembourg.
- [4] Cangiano, A. (2008). Foreign Migrants in Southern European Countries: Evaluation of Recent Data. In: J. Raymer, F. Willekens (eds.), *International Migration in Europe: Data, Models and Estimates*. John Wiley, Chichester, pp. 89–114.
- [5] Castro, L.J. and Rogers, A. (1983). Patterns of Family Migration: Two Methodological Approaches. *Environment and Planning A*, 15 (2), pp. 237–254.
- [6] Norman, P. (1999). *Putting Iterative Proportional Fitting on the Researcher's Desk*. Working Paper 99/03. School of Geography, University of Leeds.
- [7] Poulain, M., Perrin, N. and Singleton, A. (eds.) (2006), *THESIM: Towards Harmonised European Statistics on International Migration*, Presses Universitaires de Louvain, Louvain-la-Neuve.
- [8] Rees, P. (1994). Estimating and projecting the populations of urban communities. *Environment and Planning A*, 26 (11), pp. 1671–1697.
- [9] Rogers, A. and Castro, L.J. (1981). *Model Migration Schedules*. IIASA Report RR-81-30. International Institute for Applied Systems Analysis, Laxenburg.
- [10] Rowland, D.T. (2006). *Demographic methods and concepts*. Oxford University Press, Oxford.
- [11] Shryock, H.S., Siegel, J.S. and Associates (1993). Interpolation: Selected General Methods. In: D.J. Bogue, E.E. Arriaga, D.L. Anderton and G.W. Rumsey (eds.), *Readings in Population Research Methodology. Vol. 1: Basic Tools*. Social Development Center / UN Population Fund, Chicago, pp. 5-48–5-72.
- [12] Willekens, F. (1977). *The Recovery of Detailed Migration Patterns from Aggregate Data: An Entropy Maximizing Approach*. IIASA Report RR-81-30. International Institute for Applied Systems Analysis, Laxenburg.
- [13] Willekens, F. (1982). Multidimensional Population Analysis with Incomplete Data. In: K.C. Land and A. Rogers (eds.), *Multidimensional Mathematical Demography*. Academic Press, New York, pp. 43–111.
- [14] Willekens F.J., Pór A. and Raquillet R. (1981). Entropy, multiproportional, and quadratic techniques for inferring patterns of migration from aggregate data. In: A. Rogers (ed.), *Advances in Multiregional Demography*. IIASA Report RR-81-6. International Institute for Applied Systems Analysis, Laxenburg, pp. 83–124.
- [15] Wilmoth, J.R., Andreev, K., Jdanov, D., Gleit, D.A., with the assistance of C. Boe, M. Bubenheim, D. Philipov, V. Shkolnikov and P. Vachon (2005). *Methods Protocol for the Human Mortality Database, version 4*. Department of Demography, University of California, Berkeley. Available from: <http://www.mortality.org/Public/Docs/MethodsProtocol.pdf> (accessed on 17th May 2007).

Demographic Analysis of Fertility Using Data Mining Tools

Matjaž Gams and Jana Krivec
 Department of intelligent systems
 Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
 E-mail: matjaz.gams@ijs.si, jana.krivec@ijs.si

Keywords: demography, fertility, data mining

Received: March 12, 2008

We used data mining techniques to discover which attributes have the highest impact on country fertility rates. The data was analyzed in various ways; altogether and joined in smaller, meaningful groups, such as sociological, economical, philosophical, biological, etc.. We separately analyzed different groups of countries, current state and fertility trends, and tested several class schemas. Most relevant decision trees are presented and interpreted showing some known and some new conclusions. The iterative use of data mining techniques again proved to be successful in finding complex relations, but still needing expert interpretation as any computer method.

Povzetek: Analizirani so pglavitni razlogi za premajhno oz. preveliko rodnost.

1 Introduction

Populations change through three major processes: fertility, mortality, and migration. A useful way to express the rate at which women have children is the Total Fertility Rate (TFR). TFR is the average number of children that would be born per woman if all women lived to the end of their childbearing years and bore children according to a given set of age-specific fertility rates [21, 3]. If the average woman has approximately 2 children in her lifetime, this is just enough to maintain the population [7].

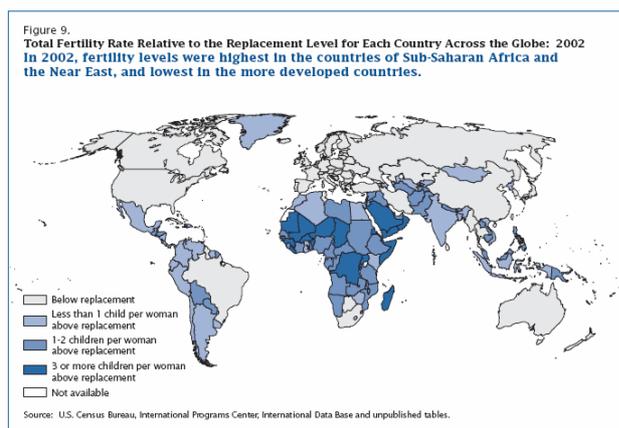


Figure 1: TFR in countries in 2002 [21].

As seen in Figure 1, some countries have high and some low TFR. In most European countries TFR in 2006 was below 1.5 children per women [19], which is far less than desired [15]. Namely, sustained low fertility rates can lead to a rapidly aging population and, in the long-run, may place a burden on the economy and the social security system because the pool of younger workers

responsible for supporting the dependent elderly population is getting smaller. Tracking trends of fertility rates and factors that influence them helps to support effective social planning and the allocation of basic resources across generations [8].

In this article we present a demographic analysis of 147 worldwide countries described by 95 basic attributes that might affect fertility rate. Even though the idea is everything but new, our approach to this problem is. Namely, our research group has decades of experiences in developing and using data mining (DM) and machine learning (ML) systems such as Weka [25] and Orange [4], last being developed in our broader research group. We typically approach a problem domain in a specific manner, usually obtaining similar results than those of the best experts in the field. It was a particular challenge to test our methods on the demographic problem.

We use terms attributes, indicators and factors as synonyms.

2 Related work

So far scientific efforts in demography were devoted mainly to exploration and definition of the process of data collection and qualitative interpretation of the statistical results, consequently not putting emphasis on new data analyzing methods. Data is typically analyzed with event history regression methods, Markov transition models and Optimal matching method using common spread statistical packages like (SPSS, SAS, S-Plus, Stata, R, TDA, etc.) [14]. The hypothesis is that between these typical aggregate descriptions and causal analysis there is a deficit of research on complex relations. Several modern methods, including data mining, offer opportunities to fill this gap.

In the last decades, data mining tools for knowledge discovery from data (KDD) proved successful in various

fields. However, searching through the internet showed that these approaches have received little attention in demographic analyses. There are some publications, e.g. Blockeel et al. [2] showed how mining frequent item sets may be used to detect temporal changes in event sequences frequency from the Austrian FFS data. In Billari et al. [1], three of the authors experienced an induction tree approach for exploring differences in Austrian and Italian life event sequences. Oris et al. [12] initiated social mobility analysis with induction trees. Unlike the statistical modeling approach, the methods make no assumptions about an underlying process generating the data and proceeds mainly heuristically. The approach differs from ours because we study rather static data and do not yet apply sequential rule mining analysis on historical demographic data.

We have not noticed DM analyses on the level of countries, similar to ours.

3 Data mining for demography

Successful data mining is based on various investigations of the data using different methods, parameters, and data to find most meaningful relations.

3.1 Basic data description

Data for machine learning and data mining are most commonly presented in attribute-class form, i.e. in a “learning matrix”, where rows represent examples and columns attributes [22]. In our case, an example corresponds to one country, and a class of the country, presented in the last column, denotes fertility rate. The first attribute is the name of the country. Altogether there are 95 basic attributes and 147 countries. Attributes and their values were partially obtained from the demographic sources such as UN [20], Eurostat [5], and the Slovenian statistical database [16]. Several of the attributes were obtained from the internet, based on the assumption that they might show some interesting demographic relation. We were trying to get as many attributes as possible, nondiscriminatory whether positive or negative in terms of fertility rate.

Attributes in demographic literature are grouped into biological and social [9] since human fertility is a socially formed biological process [18]. Newer literature introduces more and more complex structures, based on detailed grouping of social factors. Malai [10] divided factors that impact fertility rate in six groups: (1) biological, (2) economical, (3) social, (4) cultural, (5) anthropological and (6) psychological. Our 95 basic attributes correspond to these six categories, e.g.: state politics towards maternity leave, homosexuality, religion, suicide, abortion, military etc. Some of our measurements were performed on specific groups like (2) economical, consisting of 12 attributes like unemployment rate, GDP (\$) per habitant, GDP growth (%). Biological factors (1) include 6 attributes (number of habitants, life expectancy rate, number of men per 1000 women to mention a few of them). On top of these six we added a special category “education and R&D

(research and development)” with 38 attributes. There are 11 binary attributes, 2 discrete and the rest numerical.

For the basic class we have chosen Total Fertility Rate (TFR), discretized into two values: high (>2) and low (<2). The branching point 2 was chosen because it represents the replacement level of the population. In reality, replacement level is a bit higher, around 2.1, but this number depends on several other parameters such as mortality rate and immigrations, and furthermore only two countries have fertility rate between 2 and 2.1.

3.2 Data modifications

3.2.1 Attribute modifications

By attribute modifications we denote eliminating some columns in the learning matrix, and adding new columns, i.e. attributes. Subgroups of columns were chosen based on the demographic categories, and by DM methods. There were 5 new attributes added during the process of DM, thus bringing the total number of attributes to 100. Around half of the experiments were performed on 100 attributes.

3.2.2 Class modifications

Besides the basic class discretization into two values, we tried three values of TFR as well: low (<2), middle (2-3) and high (>3).

In another attempt we classified countries according to decrease or increase of TFR. We first calculated average UN predicted TFR for years 2005-2010 and subtracted average TFR for years 2000-2005. The obtained value was discretized into two classes: $\Delta\text{TFR}>0$, $\Delta\text{TFR}\leq 0$; or three classes: decreasing ($\Delta\text{TFR}<0.5$), stable ($-0.5<\Delta\text{TFR}<0.5$), and increasing ($\Delta\text{TFR}>0.5$).

3.2.3 Modifications of learning examples

Learning examples consisted of 147 countries, each represented by a row in the learning matrix. Modifications were performed as eliminating or choosing specific rows to form a new learning matrix. A typical example would be a subgroup “developed countries”, consisting only of countries with high gross domestic product (GDP) or Failed States Index (FSI). GDP is defined as the total market value of all final goods and services produced within a given country or region in a given period of time (usually a calendar year) [24]. FSI on the other hand consists of several attributes, describing the strength of central government, provision of public services, level of corruption and criminality; percentage of refugees and involuntary movement of populations, and an amount of economic decline. Since 2005, the index has been published annually by the United States think-tank, the Fund for Peace and the magazine Foreign Policy [23]. GDP review extracted two groups of countries: well developed countries with GDP above 1000\$ per habitant (39 countries), and developing countries with GDP less than 1000\$ per

habitant (108 countries). Examination of FSI revealed three groups of countries: developed with FSI lower than 39.45 (29 countries), moderately developed with FSI 39.45-61.4 (21 countries), and developing with FSI >61.4 (97 countries).

We also prepared data for analysis based on the geographical region of the country. In correspondence with UN regional classification [21], we grouped our cases in 6 regions: Asia, Africa, Latin America and the Caribbean, Oceania, Europe and Northern America, and 20 sub regions.

3.3 Methods

Machine learning and lately data mining are among the most successful artificial intelligent application areas. Whenever there are lots of learning examples, these systems learn properties of the domain and make predictions about future cases. These systems not only compete with statistical methods in terms of accuracy, they also introduce several new approaches such as cooperation between systems and humans. The constructed knowledge is often in the form of readable, understandable trees, rules and other representations thus enabling further study and fine tuning. Two examples of successful scientific and engineering DM tools are Weka [27] and Orange [4]. Both systems provide tens of DM systems, several data preprocessing and visualization tools. From the ML and DM techniques available in Weka and Orange we have chosen J48, the implementation of C4.5 [25], a method used for induction of classification trees. This method is most commonly used when the emphasis is on transparency of the constructed knowledge. In our case this was indeed so, since the task was to extract most meaningful relation from hundreds of constructed trees.

Most meaningful relations are those most significant to humans with best classification accuracy at the same time. To estimate the accuracy of the trees, we used 10-fold cross-validation, built in the system. The estimated accuracy of a classification tree corresponds to a probability that a new example will be correctly classified.

A short description of decision trees is presented in this paragraph for readers not familiar with classification trees. Classification trees are built in a top-down manner. The first task is to choose the most informative attribute which will be placed at the root of the classification tree. The next step is to add branches according to the values of the attribute. For a discrete attribute, there are as many branches as there are different values. In case of a numeric attribute, there are only two branches, one that represents values less or equal than the border value as proposed by the system, and the other branch with greater values. The set of examples is divided into subsets corresponding to the branches. Now the process can be repeated recursively for each branch, using only those instances at each particular branch. If at any time all instances at a node have the same classification, further branching is stopped and the classification into that class is proclaimed. The splitting process is usually

stopped as soon as sufficient statistical significance is obtained, classifying into the majority class. Classification is performed by starting at the top of the tree and choosing appropriate attribute values to proceed with the chosen branch. At the leaf, the numbers represent all examples and those with different class.

Experiments were performed with various method parameters, mainly changing levels of pruning. However, it turned out that default parameters were most successful.

Unlike in our typical DM session, we did not modify sources of the DM programs. J48 turned out successful enough.

4 Experiments

Tens of trees were created in a systematic way, as presented in Figure 2. First experiments were performed with TFR and ΔTFR, then with all and only developed countries. Finally, several selections of the attributes were tested: all, economical, direct, social, economical, and educational. These tests resulted in 24 basic trees. In addition, various further experiments were performed.

Due to lack of space, only most interesting trees are presented in this paper, those with most meaningful relations to humans and with best classification accuracy at the same time.

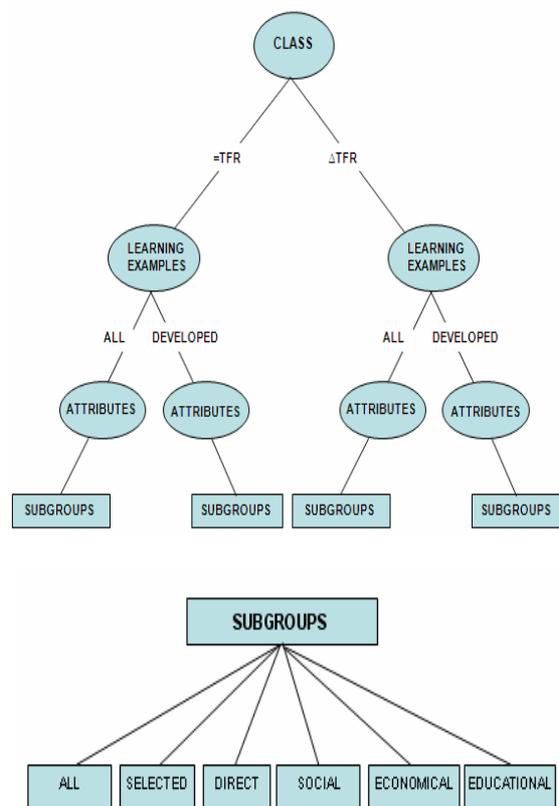


Figure 2: Structure of experiment.

4.1 TFR class

Firstly, the analysis was based on TFR as a class, with 2, 3 or more values. Only experiments with 2 or 3 values were interesting enough to be presented in this paper.

4.1.1 All countries

4.1.1.1 All attributes

In the first fertility rate analysis all 147 countries and all 95 available attributes were taken into consideration. The obtained tree is presented in Figure 3, showing that the most important indicator for high TFR is the number of stillborn children per 1000 births. More than 11.5 stillborn children per 1000 births is a strong supporting factor in favor of high TFR of the country and vice versa. The results are consistent with practically all literature in the demographic field and experts' opinions, who claim that death of newborns is in tight connection with social and economical status of mothers who need to have several children to compensate for those dead. According to experts, higher educated mothers usually have less children and lower newborn mortality, low percentage of stillborns is supposed to be related to the costs of child life-support, different life condition of the urbanized and industrialized society, changes of the attitude towards women, decaying of old patriarchal community etc. as the main reasons for fertility decline. As the tree in Figure 3 shows, these relations are indeed statistically most relevant. However, the tree shows additional relations in a structured way with appropriately weighted leaves, i.e. nodes at the bottom of the tree. For example, the top right leaf "high (104/16)" includes 88 countries with high TFR and 16 with low TFR. The bottom left leaf, on the other hand, encapsulates only 2 countries with high TFR, rendering this information as statistically less important. Therefore, in the tree there is just another statistically strongly confirmed relation: when number of dead born children is less than 11.5 and majority religion is Christianity and there are fewer men than women then TFR is low (35/1). This relation shows another crucial matter regarding interpretations of the tree. Why should Christian majority be negative for fertility rate while Christians give high emphasis on families, strong marriages and devotion to children? Indeed, further analysis show, as pointed out by demographic experts long time ago that population in these countries have high divorce rates etc. meaning that people do not follow church directions, but live according to their own desires. The bottom right part of the tree, starting with low percentage of women in the population is statistically rather meaningless, however, density and number of inhabitants gives some indication that these are among relevant attributes. Therefore, reading and interpreting trees demands some understanding of statistics, trees and demographic literature.

At each Figure title, there is cross-validation accuracy estimate. For Figure 3 it is over 80%, which is a

reasonably good result. Default accuracy obtained by classifying only into the majority class is 89,4%.

In another attempt we divided the class values into three groups: low, moderate and high TFR rate. Immediately it should be noted that the accuracy of such a tree seemingly decreases. Namely, the tree in Figure 3

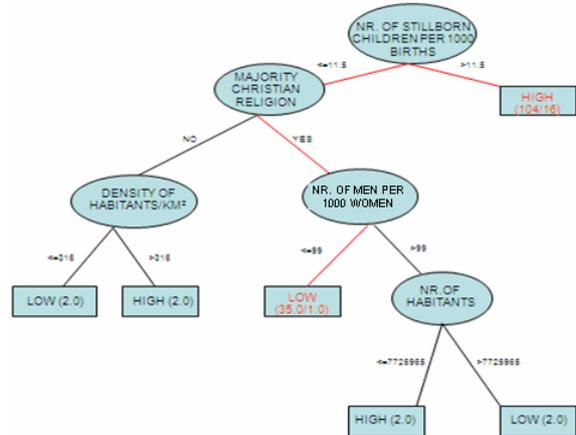


Figure 3: Two-valued TFR classification tree, all attributes (accuracy is 80,3%).

classifies into two classes, therefore, accuracy of blind guessing is 50%. For three classes, blind guessing results in 33% accuracy, and the default accuracy is 36,7% for the majority class. Having these statistics in mind, the obtained classification tree in Figure 4 achieves even better classification accuracy with 74,8%.

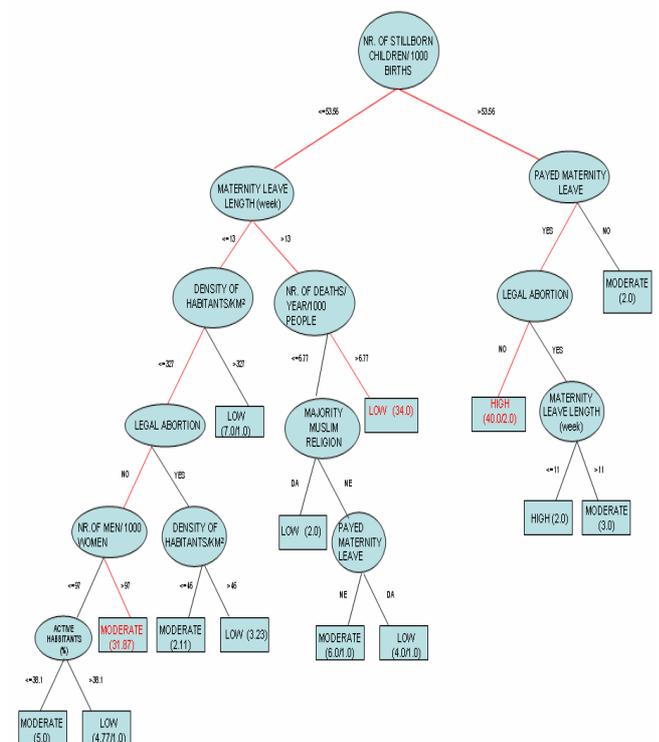


Figure 4: TFR classification tree with three-valued class, considering all attributes (accuracy is 74,8%).

The experiment once again revealed the most important attribute: “number of stillborn children”. However, the branching point leading to high TFR is in this case much higher: 53.56 children per 1000 births. In this tree, there are three major groups all from 30 to 40 countries: high, moderate and low. The major attribute distinguishing between moderate and low TFR countries is the length of the maternity leave. At this point one should be aware that such attributes are semantically potentially misleading - countries with low TFR probably introduced lengthier maternity leave as a consequence and not as cause. The tree therefore shows most important relations without knowing the nature of them.

After obtaining the first tree, in a series of tests seemingly most important attributes are being eliminated in order to test if other attributes can replace them and still obtain similar accuracy. Instead of “number of stillborn children” several attributes can be used: human development index (HDI), life expectancy rate, literacy rate, etc. all denoting the same concept. It is generally accepted that in these, developing countries, TFR is high.

For the maternity leave, the elimination of the attribute results in lower accuracy 68%. Although this attribute is obviously important, we are not able to establish the type of relation. Whatever the case, countries with short maternity leave have moderate TFR, and those with long maternity leave low TFR.

Although the rest of the relations are not so significant, they represent a bigger share than in the previous tree and they seem to have two common denominators: developmental status and value system.

Altogether, analysis so far indicate that the developed countries have low TFR, e.g. most of the European and north American countries, developing countries have high TFR, and moderately developed countries like Botswana, Bolivia, Honduras, Jamaica, etc. have moderate TFR.

4.1.1.2 Selected attributes

We further filtered attributes according to the algorithms in DM tools. Again, as seen in Figure 5, the most distinctive attribute regarding TFR rate appears to be the number of stillborn children per 1000 births. When this number is lower or equal to 11.55, the TFR is low (under 2), with the exception of the countries that do not ensure appropriate delivery treatment and invest most of its educational foundation in a primary sector.

On the other hand, TFR is low despite high number of stillborn children in the case when the human development index (consists of life expectancy rate, literacy rate, educational rate and standard of living) is high, abortion is allowed and unemployment rate is low (under 13.9 %), or if abortion is not allowed, but the country invests most of its educational foundation in a primary sector and has long maternity leave (more than 11 weeks). The discovered relations indicate a meta attribute - developmental status of the country.

4.1.1.3 Direct attributes

The demographic experts classify fertility attributes, i.e. factors, on direct and indirect [10]. Direct factors have direct influence on fertile persons. In this context we built a decision tree including 4 attributes: legality of demanded abortion, number of abortions per 1000 people, percent of married women (between 15 and 49 years old) that use contraception and percent of elders infected with HIV virus or AIDS. The obtained 82,31% accurate tree is presented in Figure 6. Legal abortion associated with low percent of HIV infected elderly relates to low TFR while illegal abortion and lower percent (less then 70) of women using contraception leads to high TFR. These attributes again seems to correlate to the meta attribute - developmental state of the country and to the value system. The other derivation could be that the value system plays an important role. The accuracy is very high indicating that these attributes are meaningful.

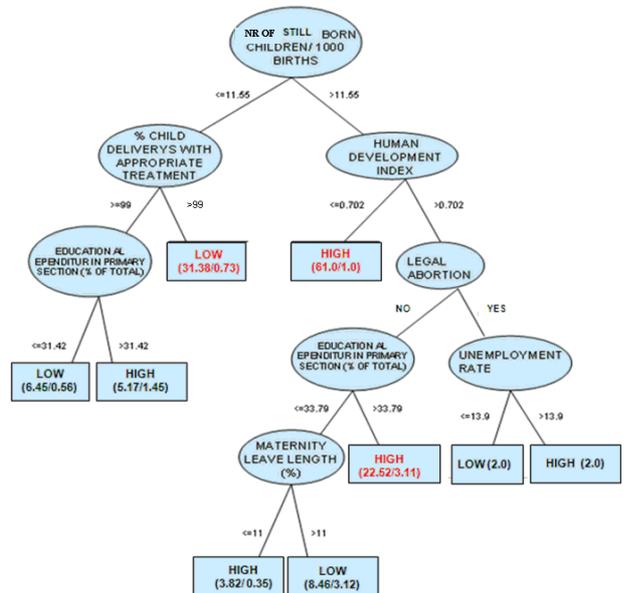


Figure 5: TFR classification tree with two-valued class, considering only automatically selected attributes (accuracy is 81,6%).

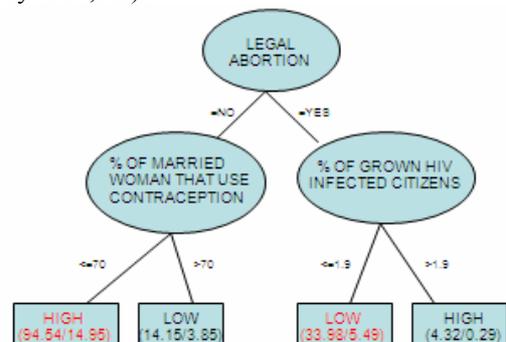


Figure 6: TFR classification tree with two-valued class considering only direct attributes (82,3%).

4.1.1.4 Social attributes

Since many experts in the field agree [10] that only direct factors can not explain the fertility rate determination, we further examined influence of the indirect TFR factors. We analyzed 11 attributes that express the society attitude towards general life questions: legality of homosexuality, legality of homosexual marriages, possibility of adoptions to homosexuals, number of suicides per 10000 persons (men only, women only, altogether), legality of abortion, number of abortions per 1000 people, number of divorces per 1000 persons, percent of women in the parliament.

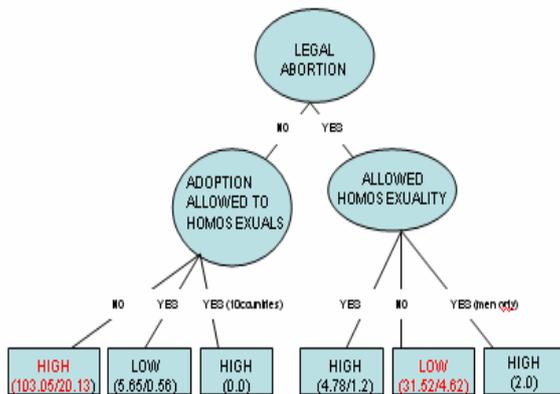


Figure 7: TFR classification tree with two-valued class, considering only social attributes (81,6%).

From Figure 7 one can conclude that TFR is high in more conservative countries that don't allow abortion and adoptions to homosexuals. TFR is high also in the countries that allow abortions but prohibit homosexuality. By this view more liberal countries have low TFR. The accuracy is as high as of the tree constructed on all the attributes, selected by the system. Again it seems that the value system plays an important role.

4.1.1.5 Economical attributes

Experts generally find low TFR strongly related to the economical factors, society modernization and liberalization [19]. We wanted to established the nature of economic relations by extracting 13 economical attributes that refer to the field of unemployment, GDP, public health and social protection expenditure, number of working ours per week and inflation rate. The constructed tree is presented in Figure 8.

The tree indicates that high GDP, low unemployment rate and high inflation GDP deflator relate to low fertility rate, while low GDP per capita usually relates to a high TFR.

As David Heer said [13], economical progress should positively influence fertility rate. Overall statistics significantly disconfirm the hypothesis at least in the modern world where food is not scarce. Our analysis indicates that direct economical attributes are not very relevant for fertility on their own, at least not as other

groups of attributes. For example in figure 8 in some cases high GDP per capita leads to high and in others to low TFR. Becker (1981) [17] presents a plausible explanation of such GDP-TFR relations. He claims that TFR depends on the disposable expenses and expected usefulness of the children. To uphold the thesis he gives an example of the rural family that used to have more children in order to assure help for maintaining the family. Human resources were urgent for working on the fields, in the woods, etc. Nowadays, agriculture has become more and more automated, thus reducing the need for human forces. Consequently, the cost benefit of the children dropped drastically and families began to shrink. Besides, factors like higher educational level, lower child mortality rate, and the desire for career making among young people, pushes TFR even lower. This linkage between income and fertility is typical for developed countries, where despite constant income growth, TFR is continually decreasing. Whereas in developing countries, low income does not influence fertility rate.

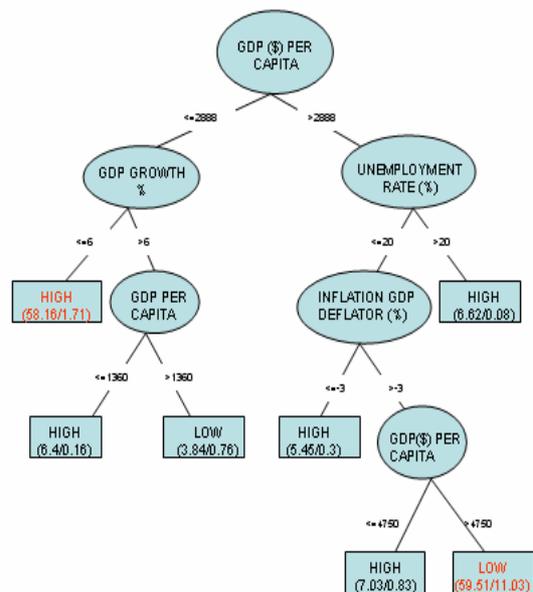


Figure 8: TFR classification tree with two-valued class, considering only economical attributes (78,2%).

In any case, the tree from Figure 8 is only 78,23 % accurate, which is low in comparison with trees based on other attributes. This indicates that direct economical factors are not the main cause for the distinction among countries with low and countries with high fertility rate.

4.1.1.6 Educational attributes

Analyzing the relation between educational factors and TFR resulted in the tree presented in Figure 9. High percentage of enrolment in primary educational level is in general related with high fertility rate, whereas low TFR is more related to enrolment in secondary or tertiary educational factors. As observed by experts before, high education, especially of women, decreases TFR.

4.1.1.7 Developed countries

While developing countries have problems with too high TFR, developed countries, especially in Europe, have problems with low TFR. Mark Steyn, a conservative polemicist, argues that Europe is quickly becoming a barren, ageing, enfeebled place [6]. In the decades after the second world war, rich countries everywhere experienced similar trends. The bonds of traditional family life began to slacken, more women got jobs, people sought enjoyment and satisfaction more and more through individual pursuits rather than in families. This social transformation, which is occurring also in America and East Asia, led to a demographic bonus (a bulge of people working) and to what might be called “the postponement of everything”. People left school later, left home later, married later, had children later, they also died later [6]. Even though these interpretations are not uniformly accepted, they seem to be statistically quite well grounded.

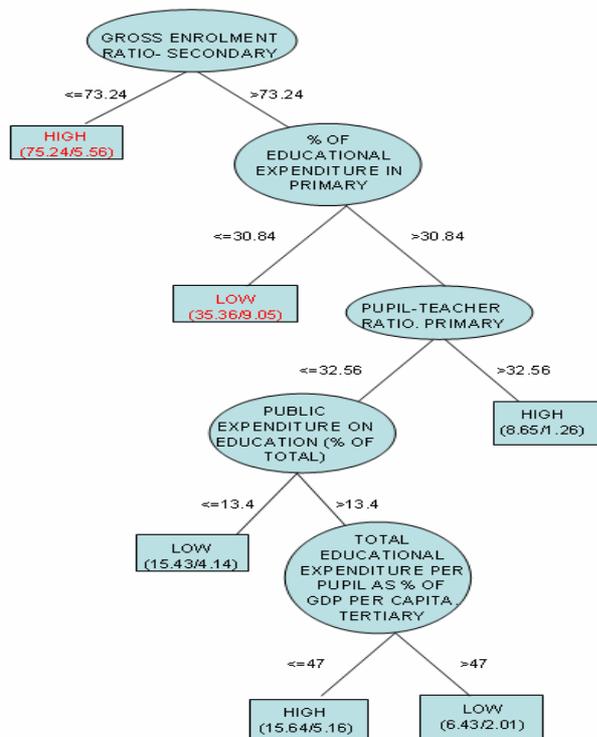


Figure 9: TFR classification tree with two-valued class, considering only educational attributes (78,2%).

Having that in mind, the relevant question is: Why do some rich countries still have high TFR?

In the following experiments we denoted 39 countries with high GDP as rich.

4.1.1.8 Selected attributes

The tree in Figure 10 indicates that exceptions to the low fertility rate have poor education and social system. Further analyses showed that these countries rely on natural resources such as oil.

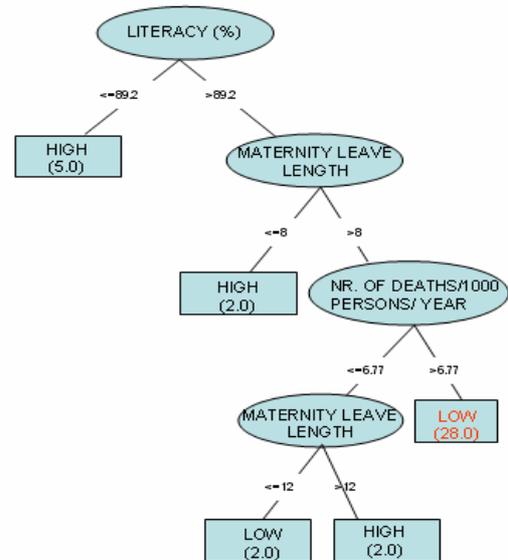


Figure 10: TFR classification tree with two-valued class and automatically selected attributes (78,2%).

4.1.1.9 Social attributes

Analyses of the obtained tree presented in Figure 11 revealed that countries with oil are rich and have Muslim religion. But the relation can be interpreted originally as follows: when Islam is the prevailing religion of the country, then TFR is most likely to be high, while otherwise, TFR decline is the more likely option. Results are consistent with the previously observed relations that TFR is higher in more conservative countries, which Islam countries certainly are.

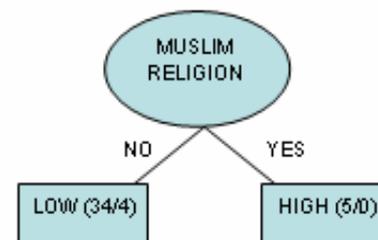


Figure 11. TFR classification tree with two-valued class considering only social attributes (89,7%).

4.2 ΔTFR class

The newest studies of Worldwatch Institute conclude that there is so much variability in fertility rates that we can not know with any confidence how many people the future holds [11]. Indeed, it seems reasonable that ΔTFR analyses are a bit less relevant as those with TFR, since they measure the amount of change and not the obtained situation. Even though, our next attempt was to established factors that might influence TFR growth and decline. In the next section a few of the most interesting and accurate trees are presented.

4.2.1 All countries

4.2.1.1 All attributes

Again, literacy seems to be an important indicator of TFR trends (see Figures 12 and 13). Countries with low percent of literate habitants generally have increase in TFR. Countries with high percent of literate citizens (above 97.9%) and low unemployment rate (below 9.6%) on the other hand have decreasing TFR trend.

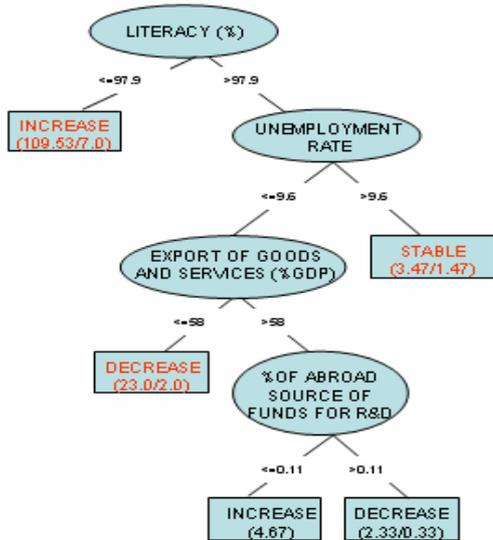


Figure 12: ΔTFR classification tree with three-valued class (81,1%).

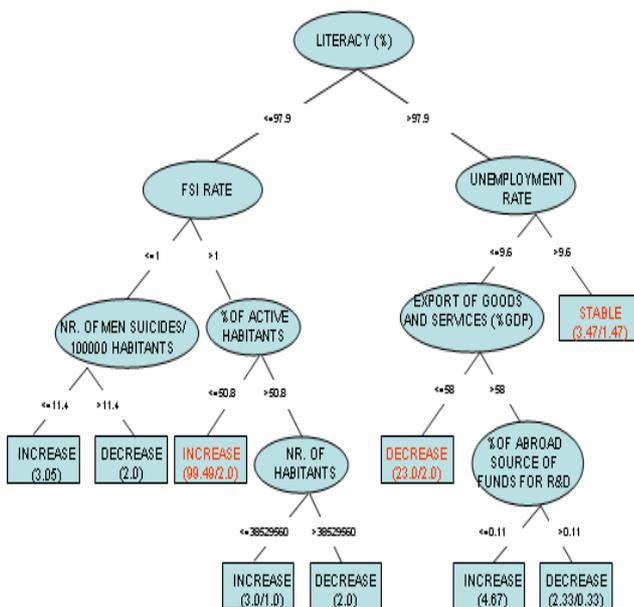


Figure 13: Unpruned ΔTFR classification tree with three-valued class indicator (83,2%).

Similar conclusions can be drawn from the tree on Figure 14, when attributes were automatically selected. This tree has surprising high accuracy.

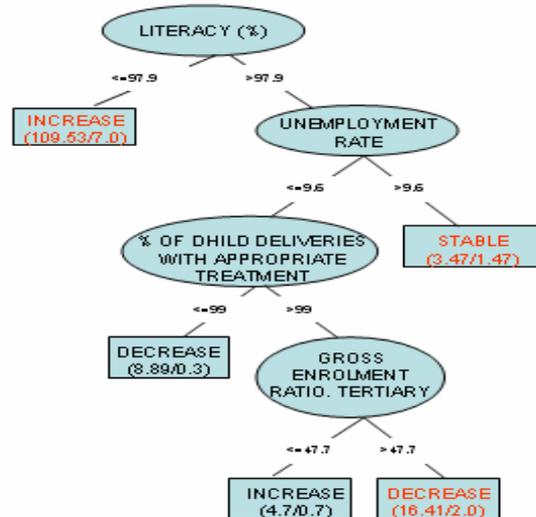


Figure 14: ΔTFR classification tree with three-valued class, automatically selected attributes (85,3%).

4.2.1.2 Social attributes

Considering only social attributes, the same tree as in the case of TFR class appeared (see Figure 7), again exposing the importance of conservative politics of the country for the TFR growth trend. Countries that don't allow abortion and adoption to homosexuals have TFR growth trend, whereas countries that allow abortion and homosexuality have TFR decline trend. Accuracy in this case is 78.32%, much lower than in the tree presented in Figure 14.

4.2.2 Developed countries

In this case our criteria for dividing countries by their developmental status was FSI. A country was classified as well developed if FSI index was less than 39.45, resulting in 27 countries. Analyses were performed on the attributes separately merged in smaller groups.

4.2.2.1 All attributes

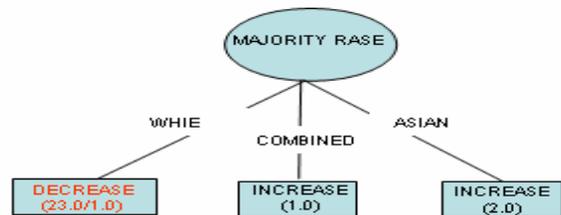


Figure 15: ΔTFR classification tree with two-valued class (accuracy is 84.6 %).

Race appeared to be an important factor of TFR trend (see Figure 15). In nations with prevalent Asian and combined race, TFR is likely to increase, while in countries with a majority of white race, TFR is declining. The nature of this genetic relation is not clear at this point.

4.2.2.2 Economical attributes

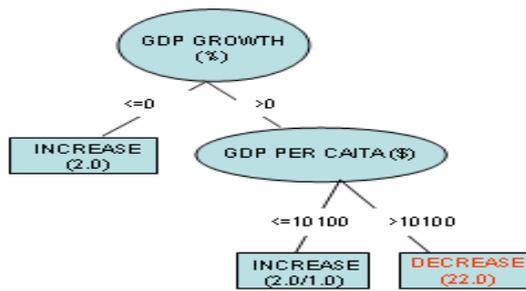


Figure 16: ΔTFR classification tree with two-valued class, considering only economical attributes (84.6 %).

We can see that highly economical developed countries with more than 10100 GDP per capita (\$) have TFR decline. This thesis is for example not in agreement with the Worldwatch Institute study noting that fertility rate is rising in the United States [11]. However, this study is violating the age-old dictum that rich countries do not make lots of babies as well [11]. The tree based on economical attributes is this time quite accurate. Therefore, ΔTFR analyses gave more statistical relevance on economical attributes than analyses with TFR.

4.2.2.3 Social attributes

When selecting only social attributes, the accuracy of ΔTFR classification trees dropped drastically (on 76.9%) what means that these factors are not good indicators for TFR trends.

5 Conclusion and discussion

In fertility analyses, the data mining tools again proved their major asset: the constructed knowledge is in a transparent form, enabling human comprehension of relevant relations in complex forms. In this way, an interactive and interaction process is enabled between computers and humans, exploiting best properties of the two most advanced information machines. Computers fast examine vast search spaces with their advanced speed and accuracy while humans make conclusions and guide search with the advanced cognitive skills.

To readdress the problem, let us restate that the space of all potential hypothesis for 100 binary attributes and a single binary class is $2^{2^{100}}$. This number is far larger than the number of all atoms in our universe, which is according to Wikipedia around 10^{80} , i.e. 2^{266} . Therefore, there is no way humans can analyze any meaningful share of all the hypotheses. But we can examine results of one search, make conclusions and redo the search changing specific details of the search. In this way humans can “mine” for relevant hypothesis.

Regarding the fertility relations, the DM tools enabled rediscovery of major properties. The authors are not experts in the fertility or demographic field, therefore verification of our conclusions by an expert and further analyses of interesting new patterns are a matter of

further research. However, we report our impressions for further discussions:

- Firstly, we were surprised that there are so many distinctive hypothesis, i.e. patterns discriminating countries with low from those of high fertility rate. Rich countries are predominantly white, have good education, women live longer, literacy is high, the predominant race is white, people have no strong religion obligations, they are liberal etc. and vice versa for the developing countries.
- Secondly, according to the constructed trees, it is rather simple to influence the fertility rate – just improve literacy in women or just allow liberalization and decrease the influence of religion. And vice versa – to improve fertility, just e.g. improve moral values and decrease liberalization or decrease literacy or apply any of the remaining 10 or 15 attributes. Some are costly, some are unacceptable, e.g. decreasing literacy. According to some experts, practically all of these attributes are hard to implement in democratic countries. Still, the trees indicates that there are several mechanisms, some of them rather costless, that will change the process in European countries, leading first to economic problems and later to extinction of nations and cultures.
- We did not have time to study each particular attribute in detail, such as the length of maternity leave. While the trees so often show relevance of maternity leave for decreasing fertility, the trees do not show whether this is a cause or a consequence. At first, we thought that it is just a consequence of low fertility, just a mechanism of countries promoting higher fertility. In addition, mothers are generally in favor of longer leaves. But after so many trees, and having in mind that this is one of very costly mechanisms, it is becoming more than a suspicion that longer maternity leave is at least a controversial matter.

Our analyses dealt with countries and not with individuals. Obviously, several of the fertility and demographic matters are open for further investigation with DM techniques, next time with fertility experts.

Acknowledgement

In the analyses performed, contributions of the following computer-science students should be acknowledged: Barbara Tvrđi, Anže Jazbec, Marko Kovki, Domen Muren. We also thank Statistical office of the Republic of Slovenia for the answers and the IS 2007 program committee members for the constructive suggestions.

Reference

[1] F.C. Billari, J. Fürnkranz, and A. Prskawetz (2000). Timing, sequencing, and quantum of life course events: a machine learning approach. Working paper 010, Max-Plank-Institute for Demographic Research, Rostock.

- [2] H. Blockel, J. Fürnkranz, A. Prskawetz, and F. Billari (2001). Detecting temporal change in event sequences: An application to demographic data. In L.D. Raedt and A. Siebes (eds.), *Principles of data Mining and Knowledge discovery: 5th European Conference*, PKDD 2001, Volume LNCS 2168, pp.29-41. Freiburg in Brisgau: Springer.
- [3] M. Christenson, McDevitt, T., and Stanecki, K. (2004). *Global Population Profile: 2002. International Population Reports*. Health Studies Branch, International Programs Center, Washington Plaza II, Room 313A U.S. Census Bureau, Washington, DC 20233-8860.
- [4] J. Demsar, and B. Zupan (2005). From Experimental Machine Learning to Interactive Data Mining, White Paper (www.ailab.si/orange), Faculty of Computer and Information science, University of Ljubljana.
- [5] Eurostat
<http://epp.eurostat.ec.europa.eu/>
- [6] Economist
http://www.economist.com/world/europe/displaystory.cfm?story_id=9334869
- [7] M.Gams (2007). Osnovna demografska gibanja (Basic Demographic Dynamics). In J. Malačič, M. Gams (eds.), *Proceedings of the 10th International Multi-conference Information Society (volume B) Slovenian Demographic Challenges of the 21st Century*. Ljubljana: “Jožef Stefan” Institute, pp. 35-37.
- [8] M. Gams, J. Krivec (2007). Analiza vplivov na rodnost (Analysis of Impacts on Fertility). In J. Malačič, M. Gams (Eds.), *Proceedings of the 10th International Multi-conference Information Society (volume B) Slovenian Demographic Challenges of the 21st Century*. Ljubljana: “Jožef Stefan” Institute, pp. 35-37.
- [9] J. Malačič (2006). *Demografija: teorija, analiza, metode in modeli (Demography: theory, methods and models)*, (EF, Manual). 6. edition. Ljubljana: Faculty of Economy, pp. 339
- [10] J. Malačič (2000). *Demografija– teorija, analiza, metode in modeli (Demography: theory, methods and models)*. 4. edition. Ljubljana.
- [11] New York Times
<http://dotearth.blogs.nytimes.com/2008/03/14/earth-2050-population-unknowable/>
- [12] M. Oris, G. Ritschard, and A. Berchtold (2000). The use of Markow process and induction trees for the study intergenerational social mobility in nineteenth century Geneva. In *Social Science History Association Annual Meeting*, Baltimore.
- [13] M. Potts, Society and Fertility. Estover: McDonald and Evans. 1979. pp. 384
- [14] G. Ritschard, and M. Oris (2005). Dealing with Life Course Data in Demography: Statistical and Data mining Approaches, in R. Lévy, P. Ghisletta, J.-M. Le Goff, D. Spini et E. Widmer (eds.), *Towards an Interdisciplinary Perspective on the Life Course*, Advances in Life Course Research, Vol. 10. Amsterdam: Elsevier, pp. 283-314.
- [15] Slovenija v številkah (Slovenia in Numbers). Ljubljana: Statistical office of the Republic of Slovenia, 2006. pp. 79
- [16] Statistical office of the Republic of Slovenia; Slovenian Statistical Database: <http://www.stat.si/>
- [17] N. Stropnik (1997). *Ekonomski vidik starševstva (Economical side of Parenthood)*. Ljubljana, Scientific and publication center , pp.221.
- [18] M. Širčelj (1991). *Determinante rodnosti v Sloveniji (Fertility Determinates in Slovenia)*. Doctoral dissertation, Faculty of Art, Ljubljana.
- [19] United Nations, Department of Economic and Social Affairs, Population Division (2007). World Population Prospects: The 2006 Review, Highlights, Working Paper No.ESA/P/WP.202.
- [20] UN <http://esa.un.org/unpp/>
- [21] U.S. Census Bureau, International Population Reports WP/02, *Global Population Profile: 2002*, U.S. Government Printing Office, Washington, DC, 2004.
- [22] V. Vidulin, M. Gams (2006). “Vpliv investicij v izobraževanje in R&R na gospodarsko rast,” *Elektroteh. vestn.*, vol. 73, nr. 5, pp. 285-290.
- [23] WIKI-1
http://en.wikipedia.org/wiki/Failed_States_Index
- [24] WIKI-2
http://en.wikipedia.org/wiki/Gross_Domestic_Product
- [25] I. H. Witten, E., Frank (2005). “Data Mining – Practical Machine Learning Tools and Techniques (sec. ed.),” Morgan Kaufmann.

Clustering of Population Pyramids

Simona Korenjak-Černe

University of Ljubljana, Faculty of Economics, Department of Statistics
simona.cerne@ef.uni-lj.si

Nataša Kejžar

University of Ljubljana, Faculty of Social Sciences, Department of Informatics and Methodology
natasa.kejzar@fdv.uni-lj.si

Vladimir Batagelj

University of Ljubljana, Faculty of Mathematics and Physics, Department of Mathematics
vladimir.batagelj@fmf.uni-lj.si

Keywords: clustering, population pyramid, Ward hierarchical method, hierarchical clustering with relational constraint

Received: March 6, 2008

Population pyramid is a very popular presentation of the age-sex distribution of the human population of a particular region. The shape of the pyramid shows many demographic, social, and political characteristics of the time and the region. In the paper results of hierarchical clustering of the world countries based on population pyramids are presented. Special attention is given to the shapes of the pyramids. The changes of the pyramids' shapes, and also changes of the countries inside main clusters are examined for the years 1996, 2001, and 2006.

Also smaller territorial units of a country can be observed through clusters. To illustrate this, clusters of 3111 mainland US counties in the year 2000 obtained using the hierarchical clustering with relational constraint of counties' population pyramids are examined. In the paper, the results for clustering into nine main clusters are presented.

Povzetek: Prikazano je grupiranje demografskih piramid.

1 Introduction

Population pyramid is a very popular presentation of the age-sex distribution of the human population of a particular region. It gives picture of a population's age-sex structure, and can also be used for displaying historical and future trends.

Generally, there are three main pyramids' shapes: expansive, constrictive, and stationary (Figure 1). The *expansive*

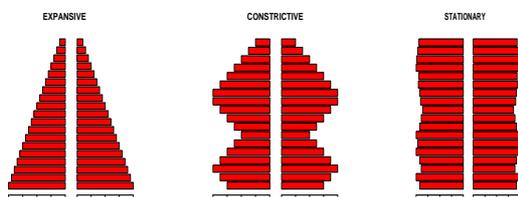


Figure 1: Three main shapes of population pyramids

sive shape is typical for fast-growing populations where each birth cohort (a group of people born in the same year or years period) is larger than the previous one (Latin America, Africa).

Constrictive shape displays lower percentages of younger population (United States).

Stationary shape present somehow similar percentages for almost all age groups. The population pyramids of the

Scandinavian countries tend to fall in this group.

Since the biggest influence on the pyramid's shape have fertility and mortality, the explanation of the pyramids' shapes is often related to the "Demographic Transition Model" (DTM) that describes the population changes over time (Figure 2). It is based on an interpretation that begun in 1929 by the American demographer Warren Thompson, of the observed changes, or transitions, in birth and death rates in industrialized societies over the past two hundred years.

Besides births and mortality, also other processes, depending on social or/and political policy and events (migrations, birth control policy, war, life-style etc.) have strong influence on age-sex structure of the population, that reflect also on the shape of the population pyramid.

Population pyramids are very easily understandable to almost everyone. In the combination with professional knowledge, they offer also many additional explanations about different processes to experts (e.g. demographers, sociologists, politicians, economists, geographers).

Due to these facts we decided to observe clusters of the world countries based on the shapes of their population pyramids. We observed how countries and clusters of them fit with the main pyramids' shapes and how stable are clusters that we got with hierarchical clustering. Selected clustering procedure to determine clusters of the world coun-

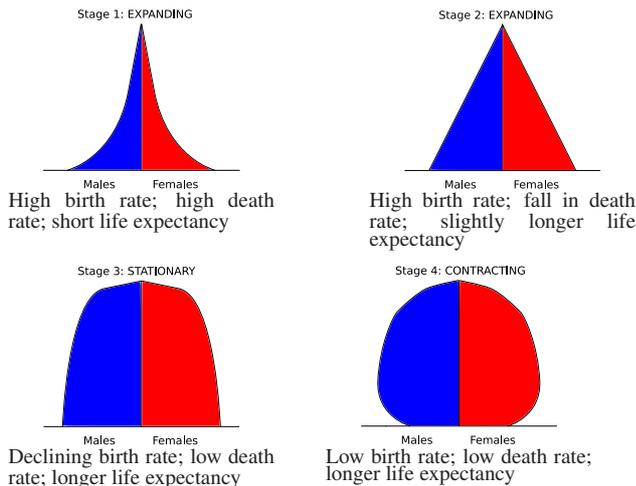


Figure 2: Shapes of population pyramids for 4 stages of the demographic transition model

tries is based on the Ward's hierarchical clustering method.

For each country a clustering can be applied also on smaller regions. In smaller regions the influence of other processes besides births and mortality (for example local migrations caused by schooling, religion, work or health reasons etc.) is even more emphasized, which reflect in differences of pyramids' shapes. Therefore we further examined clusters of 3111 mainland US counties. To determine clusters, hierarchical clustering with relational constraint was used.

Since we are not demographers we focus on the presentation of the methods and results they produce, and we limited the explanation of them to the 'technical' characteristics (results that can be directly seen from the obtained graphs and clusters), hoping that the proposed approaches will catch the attention of researches using pyramids' analysis in their work.

2 Clustering procedure

Data on the population pyramids of the world countries used in our analysis were taken from the web page of the International Data Base (IDB). Age is divided into 17 five-years groups (0-4 years, 5-9 years, 10-14 years, ..., 75-79 years, 80+). In our model, for each country each age group is considered as a separate variable for each sex, so each country is presented with 34 variables: 17 variables for 5-years age groups for men, and 17 variables for 5-years age groups for women. Values are normalized so that they present percentages of the country's population in each age group. Euclidean distance between corresponding vectors is used. Although some objections against the usage of this difference measure can be found (Andreev, 2004), in our opinion in observation of the shapes of the population pyramids, each age-group can be considered as a separate variable.

For clustering procedure, the algorithm for the Ward's

hierarchical method, implemented in a package 'cluster' in the statistical environment R was used. This clustering procedure is implemented based on the description of the method in Kaufman and Rousseeuw, 1990.

Since each country is represented with normalized vector (relative age-sex distribution), the 'centroid' pyramids of the clusters of countries are not real population pyramids describing the whole population in the clusters, but are based on the new vector got with the clustering procedure. So they can be interpreted only in terms of shapes, not as a population pyramids of countries' clusters, because the population size of the whole cluster is not taken into account. But on the other hand such approach enables us to detect even small countries with very different pyramids' shapes based on special countries' characteristics.

3 Clusters of the world's countries over time

We observed clusters of the world countries obtained by the Ward's clustering procedure for each year from 1996 to 2007. Although the time period is rather short for the human life, substantial changes can be seen. In Figure 9, Figure 10, and Figure 11 respectively, hierarchical trees for the years 1996, 2001, and 2006 with the appropriate pyramids' shapes for some of the main clusters are presented.

3.1 Pyramids' shapes of the clusters of world countries

Our first examination is concentrated on the shapes of the population pyramids of the clusters in the hierarchies. For each of the years 1996, 2001, and 2006, some of the interpretations are given. Much more detailed information can be obtained depending on the cutting level in the hierarchy and on the interest of the observer.

3.1.1 Year 1996

For the year 1996, the dataset contains 215 countries. Four main pyramids's shapes are presented in Figure 3. The first cluster has typical expansive pyramid's shape. It includes 77 countries (most of African countries etc.). When comparing these pyramids' shapes with pyramids' shapes of the DTM shown in Figure 2, we can say that the first one corresponds to the Stage 1, the second and the third have characteristics of the Stage 2 and 3, and the last looks between stages 3 and 4. For easier explanation of later observations we will denote them with letters A, B, C, and D respectively.

Clusters at the bottom levels of the hierarchy are more and more similar. When we cut at the level with eight clusters, clusters A and C are each divided into two smaller clusters presented in Figure 4 for cluster A and in Figure 5 for cluster C. Cluster B remains the same also when cutting into eight clusters. First shape from Figure 4 corresponds

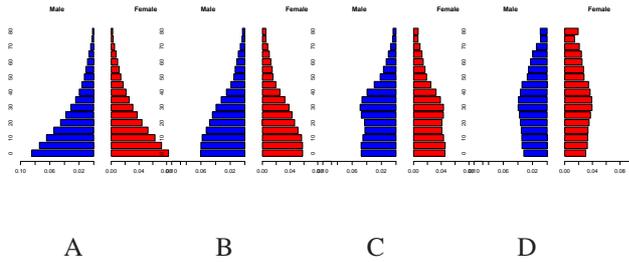


Figure 3: Pyramids’s shapes of four main clusters of the countries for the year 1996

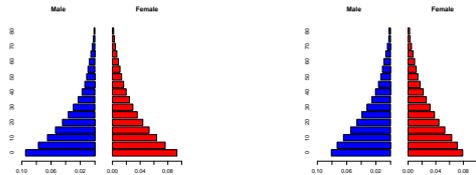


Figure 4: Pyramids’s shapes of the first two (from the left) of eight clusters of the countries for the year 1996

to the Stage 1, but the second one looks closer to Stage 2 of the DTM from Figure 2.

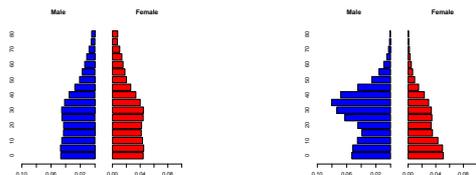


Figure 5: Pyramids’s shapes of the fourth and fifth (from the left) of eight clusters of the countries for the year 1996

The last cluster in Figure 5 includes countries with very big differences between gender’s distributions: Bahrain, Kuwait, Qatar, and United Arab Emirates. Since in all these countries the population of men in the data is much bigger (specially in the middle ages) than of women, this has effect also on the pyramid’s shape of the cluster. We also calculated dissimilarities between gender’s distributions for all countries in 1996, and five of them with the largest differences are: United Arab Emirates, Qatar, Kuwait, Oman, and Bahrain. Four of them are included in the described cluster, detected in the hierarchy.

Cluster D is at lower level divided into three smaller clusters. Their pyramids’s shapes are presented in Figure 6. Slovenia is in the last cluster from the right together with 34 other countries. The most similar country to Slovenia is Croatia, and after it also Belgium, France, Finland, and Gibraltar what can be seen on the dendrogram in Figure 9. The age distributions of both genders in the pyramid’s shape are quite similar, although there are slightly more women, specially those that are older than 70.

Similar and even more detailed descriptions at lower levels can be obtained for each group in the hierarchy. With additional knowledge about countries more detailed explanations of the shapes can be given which also offer many

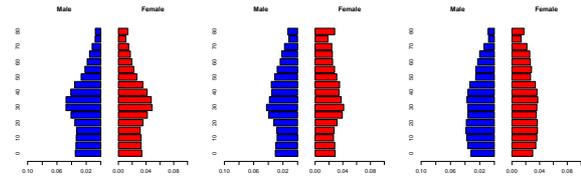


Figure 6: Pyramids’ shapes of the last three (from the right) of eight clusters of the countries for the year 1996

interesting points for discussion about similarities and differences among countries and/or clusters.

3.1.2 Year 2001

Hierarchy on the 222 world countries for the year 2001 is presented in Figure 10. At the upper level of the dendrogram two main shapes of pyramids can be seen. The first pyramid’s shape approximately corresponds to the Stage 2 of the DTM shown in Figure 2, and the second one approximately corresponds to the Stage 4.

One level lower each of two main clusters is divided into two additional clusters. Their pyramids’ shapes are presented in Figure 7. Their shapes are similar to those in the year 1996, therefore we denoted appropriate clusters with the same letters A, B, C and D.

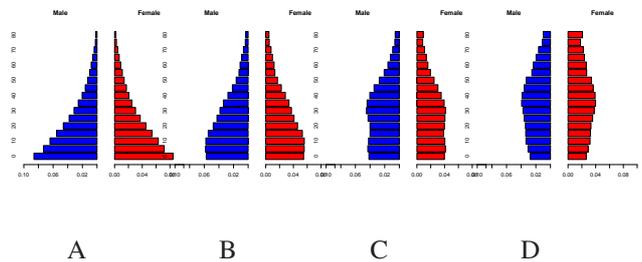


Figure 7: Pyramids’s shapes of four main clusters of the countries for the year 2001

Inspecting lower level of the hierarchy, eight clusters can be detected. First cluster is the same as the first one (A) presented in Figure 7. It includes 60 countries and has typical expanding shape of the population pyramids.

Cluster B is at lower level divided into two smaller clusters. The shapes of the pyramids are presented in Figure 8. They approximately fit to Stage 2 and Stage 3 of the DTM. That shows some progress in countries’ development comparing with the year 1996.

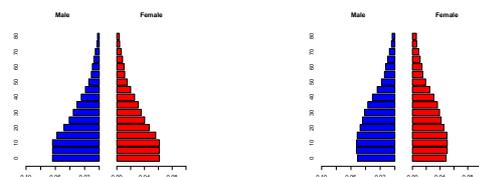


Figure 8: Pyramids’s shapes of the second and third (from left) of eight clusters of the countries for the year 2001

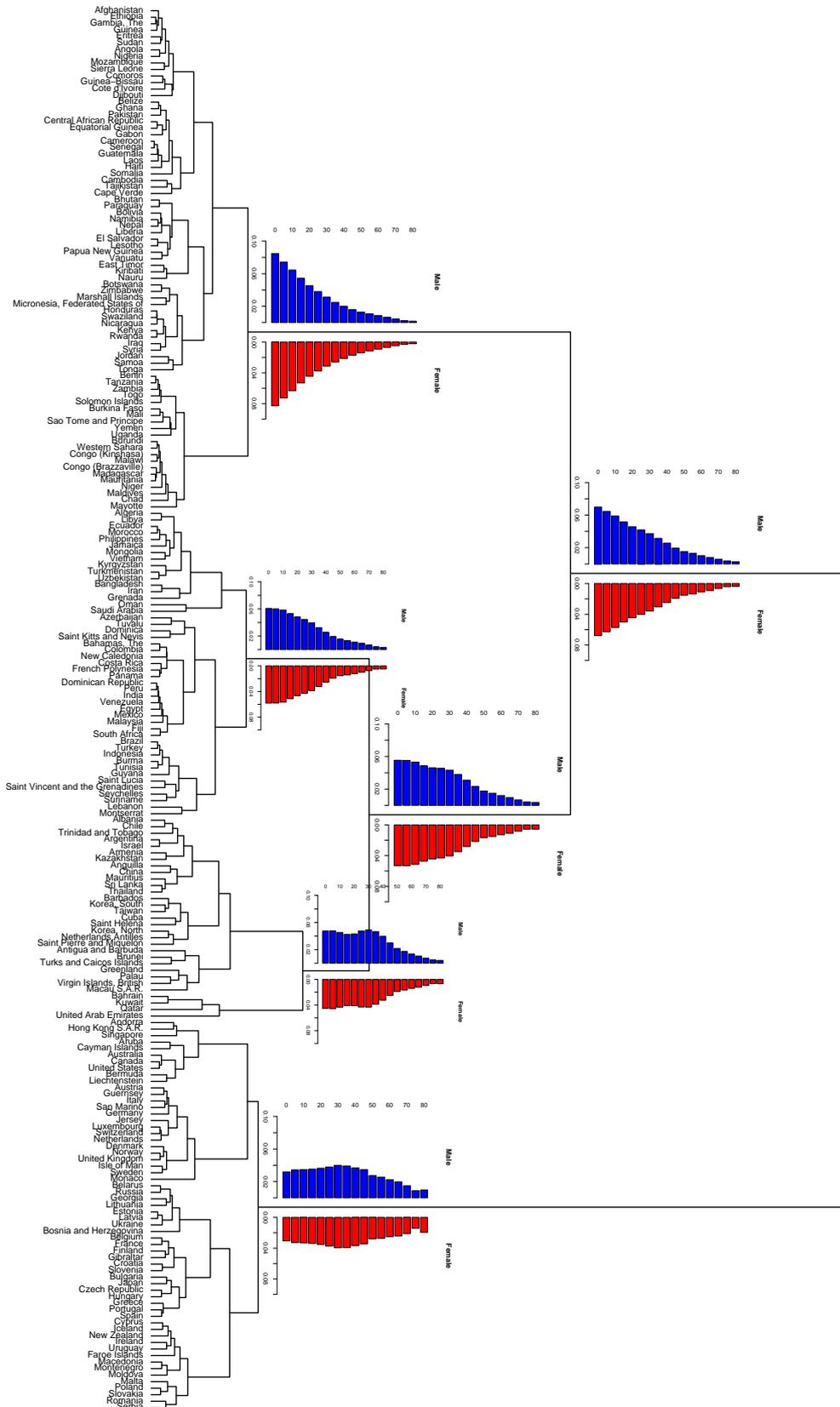


Figure 9: Clusters of the countries and main pyramids' shapes for the year 1996

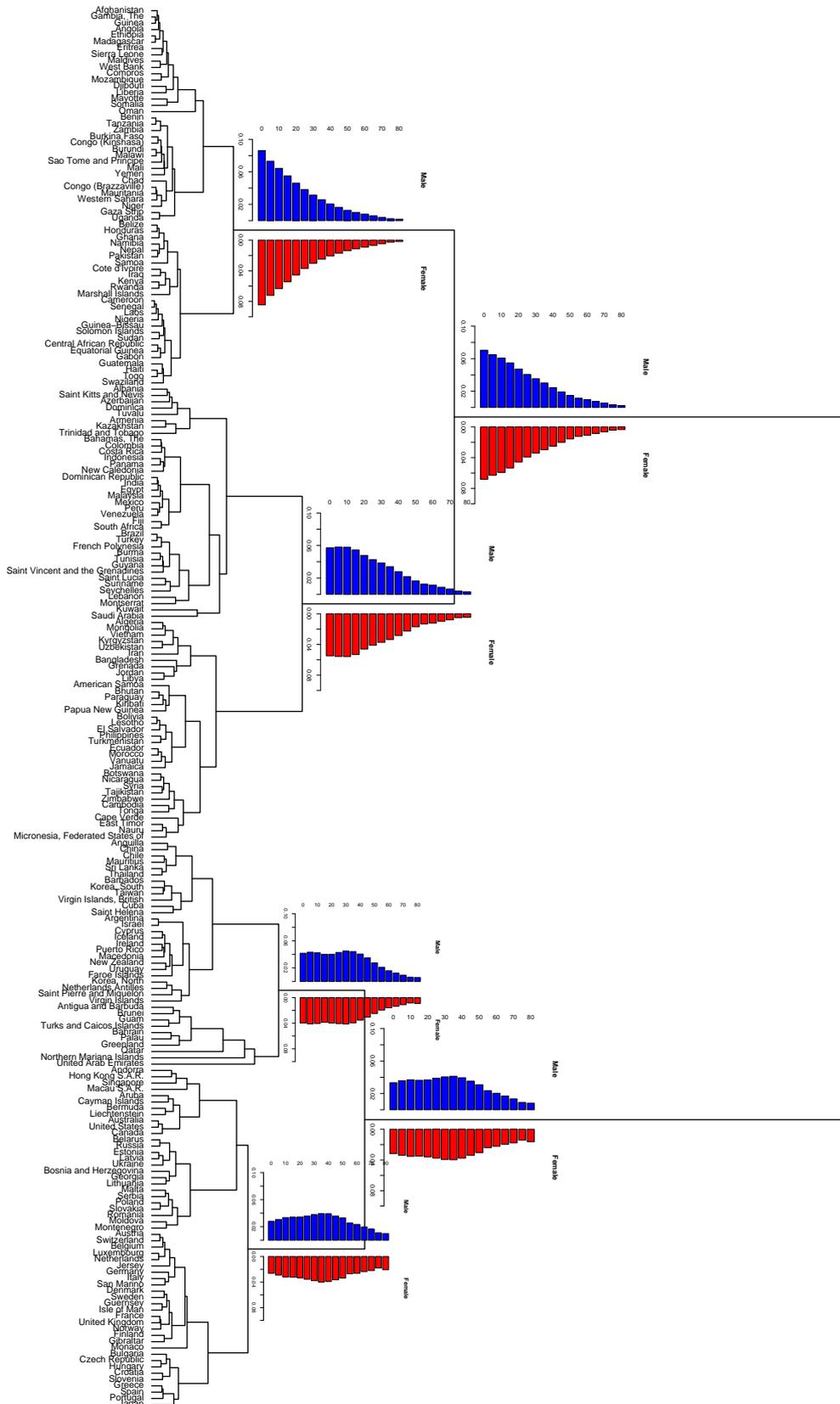


Figure 10: Clusters of the countries and main pyramids' shapes for the year 2001

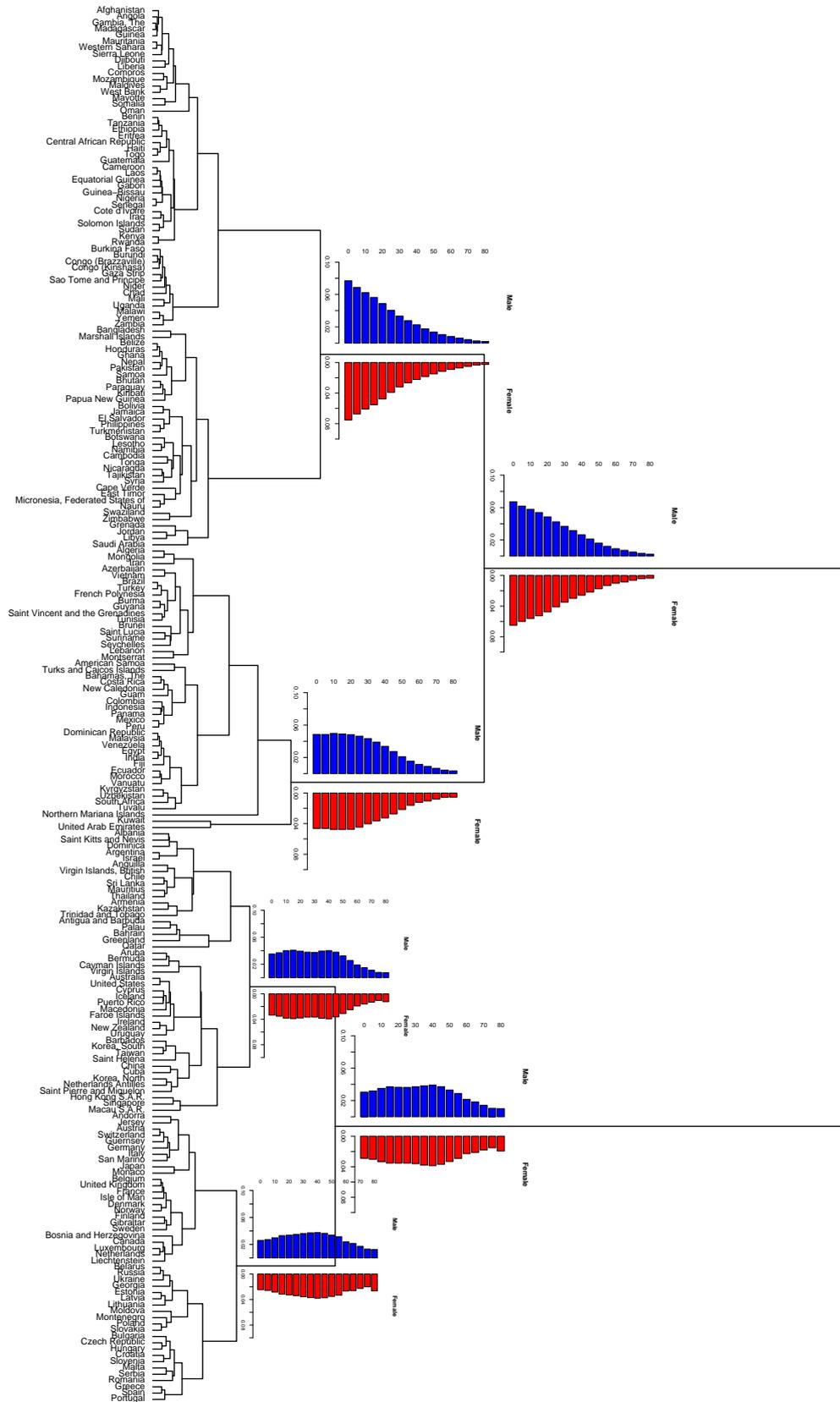


Figure 11: Clusters of the countries and main pyramids' shapes for the year 2006

Cluster C is at the lower level divided into three additional smaller clusters. Their pyramids' shapes are presented in Figure 12. Among eight clusters United Arab Emirates forms separate cluster by itself (the right pyramid in Figure 12).

Its population pyramid's shape shows big differences between gender's distributions in the country. At the upper levels, Northern Mariana Islands, Qatar etc. join United Arab Emirates.

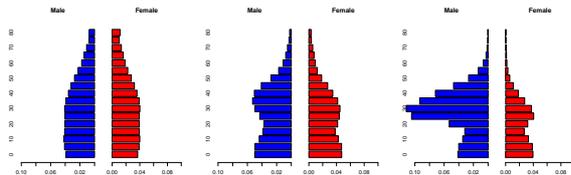


Figure 12: Pyramids's shapes of three of eight clusters of the countries for the year 2001

The last cluster D among four main clusters is at lower level divided into two smaller and more similar clusters of countries. Their pyramids' shapes are presented in Figure 13. Slovenia is in the right cluster in Figure 13 together with 27 other countries. Among them were twelve countries (Belgium, Bulgaria, Croatia, Czech Republic, Finland, France, Gibraltar, Greece, Hungary, Japan, Portugal, and Spain) also in the selected cluster with Slovenia among eight main clusters for the year 1996. The shape of the pyramid in this cluster is rather symmetric, although there are larger values in the women side (specially for older than 75).

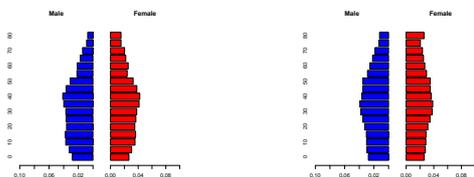


Figure 13: Pyramids's shapes of the last two among eight clusters of the countries for the year 2001

3.1.3 Year 2006

The last hierarchy we present is the hierarchy of 222 countries for the year 2006. It is presented in Figure 11. The shapes of the pyramids of the four main clusters are presented separately in Figure 14. Also these shapes are similar as in the years 1996 and 2001 therefore appropriate clusters are denoted with the same letters. As for the previous two years also for the year 2006 more detailed explanations for each cluster of the hierarchy could be found.

3.2 Stability of the clusters in the hierarchies

In the following section we observe how stable are the main clusters over time. For each of the years 1996, 2001, and

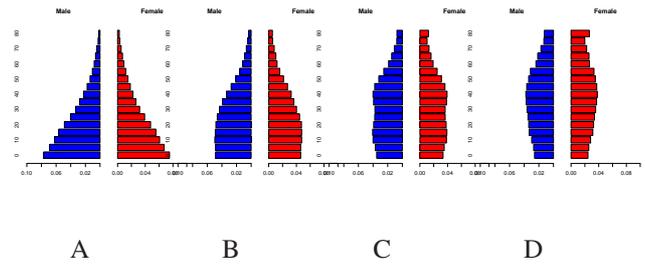


Figure 14: Population pyramids of the main clusters of the countries for the year 2006

2006 we present the results of observation of four main clusters. From Figure 3, Figure 7, and Figure 14 we can see four main rather similar pyramids' shapes, although even these are slightly changing over time.

Observing changes of countries inside each of the clusters A, B, C and D for each of the years 1996, 2001 and 2006, we can conclude the following:

Countries from the cluster A in the year 1996 (Figure 3) are in the year 2001 in clusters A and B (Figure 7). All these countries except one (Vanuatu) are included in the cluster A in the year 2006 (Figure 14).

Cluster B in the year 1996 is included in cluster B in 2001 (the only exception is Oman), and also mostly corresponds to the cluster B in the year 2006.

Cluster C in the year 1996 is mainly presented in clusters B and C in the year 2001, and all except three of the countries from it are included in the cluster C in the year 2006. The remaining three countries (Kuwait, Turks and Caicos Islands, and United Arab Emirates) are in the year 2006 included in the cluster B. This is not surprising because of the differences of gender's distributions in these countries.

The fourth cluster D of four main clusters in the year 1996 is mainly included in cluster D in the year 2001, and most of the countries, precisely 46 from 60 countries from it, are also included in cluster D in the year 2006. The remaining 14 countries from it are in cluster C in 2006.

Similar comparisons were made for the years 2001 and 2006:

The first cluster A of four main clusters in the year 2001 (Figure 7) is included in cluster A in 2006 (Figure 14).

Most of the countries from cluster B in 2001 are included in clusters A and B in 2006, except six countries (Albania, Armenia, Dominica, Kazakhstan, Saint Kitts and Nevis, and Trinidad and Tobago), that are in 2006 included in cluster C.

Five of the countries (Brunei, Guam, Northern Mariana Islands, Turks and Caicos Islands, and United Arab Emirates) from cluster C in 2001 are moved to cluster B in 2006, all the remaining countries are in 2006 included in cluster C.

46 of 54 countries (including Slovenia) from cluster D in 2001 create cluster D in 2006, the remaining eight countries are in 2006 included in cluster C.

In the Figure 15 we present these movements among four

main clusters for the years 1996, 2001 and 2006 with the number of countries.

	A	B	C	D
1996	77	47	31	60
	57 20		5 25 1	
	1	46	7	53
2001	60	72	36	54
	60		5 31	
	26	40	6 8	46
2006	86	45	45	46

Figure 15: Movements presented with the number of countries among four main clusters for the years 1996, 2001 and 2006

Differences among clusters can be observed in greater detail considering the hierarchies in each of the clusters.

4 Hierarchical clustering with relational constraint of US Counties

For US counties age in population pyramids is divided into 18 five-years groups (0-4 years, 5-9 years, 10-14 years, ..., 75-79 years, 80-84 years, 85+). In our model, for each US county each age group is considered as a separate variable for each gender, so each county is presented with 36 variables: 18 variables for 5-years age groups for men, and 18 variables for 5-years age groups for women. Values are normalized so that they present percentages of the county’s population in each age group among the whole county population. Euclidean distance between the corresponding vectors is used. NA values in data for 6 counties were replaced with 0.

For clustering procedure, the algorithm for hierarchical clustering with relational constraints based on the maximum hierarchical method was used. It is implemented in Pajek, the program for analysis and visualization of large networks. The agglomeration of two counties was restricted with the relational constraint based on neighboring counties (Ferligoj, Batagelj, 1983). The maximal method to calculate new dissimilarity between clusters was used. The neighboring relation is symmetric. Therefore the tolerant strategy to determine the relation between the new cluster and other clusters is used (Batagelj, Ferligoj, Mrvar, 2008).

As in the case with the world’s countries, also here the ‘centroid’ pyramids of the clusters of counties are not real population pyramids describing the whole population in the clusters, but are based on the new vector produced by the clustering procedure. So they can be interpreted only in

terms of shapes, not as a population pyramids of counties’ clusters, because the population size of the whole cluster is not taken into account.

The cut of the dendrogram was done at height 0.06, which divided counties in 36 clusters with 155 isolates (counties that are very different from all their neighbors). Out of these 36 clusters were only 14 clusters with more than 10 vertices, therefore we decided to increase the height.

At height 0.1 we obtained 9 clusters and 54 isolated counties. There are 6 clusters of more than 6 counties (precisely with 7, 15, 69, 402, 1152 and 1406 counties) and 3 of them with 2 counties. Clusters (groups of counties) are presented in Figure 16 with different shapes and colors. Group 1 is the largest group situated at eastern part of the USA (light gray circles). Darker gray group with triangles that borders group 1 is group 2, the second largest group of counties. Dark gray circular vertices at the Florida peninsula belong to group 5. 7 dark gray vertices in the middle of group 2 (in the center of the USA) represent group 7. The large white squares in the north and middle of the USA represent group 4, while area with lighter gray circles inside the bottom of group 2 belongs to group 3. Groups 6, 8 and 9 with 2 counties each are in Figure 16 represented with dark gray diamonds (group 6 is in the north-west of the USA, group 8 at the south east of group 3, and group 9 in the middle of the largest group 1).

Further inspection of the pyramids’ shapes of the clusters of counties shows that all three 2-vertex clusters (groups 6, 8 and 9) have average population pyramid with mostly young people in their 20s (Figure 17). We conjecture these are counties with mainly student population (surroundings of larger universities and colleges). More precisely: cluster 6 includes counties with University of Idaho and Washington State University. In cluster 8 are Madison and Walker County, Texas, with median ages 33 and 31 and with male population for more than 50% larger than female population. Cluster 9 includes Montgomery and Radford Counties, Virginia, with West Virginia University Institute of Technology, popularly called WVU Tech, and Radford University, which have strong influence on the age distribution.

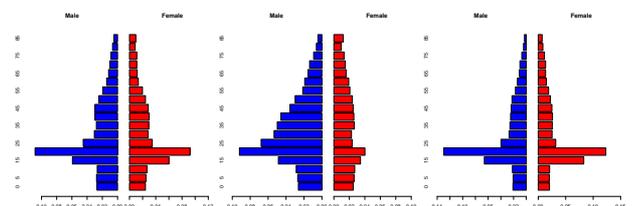


Figure 17: Pyramids’ shapes of three clusters with two counties

Pyramids’ shapes of the two largest clusters show typical all-American population pyramid (Figure 18) with rather typical constrictive shape (Figure 1).

In two of other four clusters’ pyramids’ shapes older



Figure 16: Clustering of US counties in the year 2000 with relational constraints

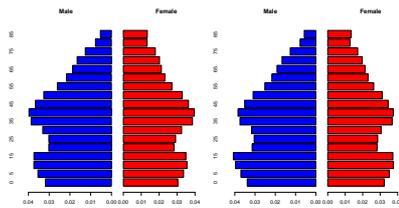


Figure 18: Pyramids' shapes of two largest clusters

population is more pronounced (looking bottom-up the pyramid bars start shrinking later than the overall American population pyramid). The groups are concentrated in Florida (first in Figure 19) and in Missouri (the second one in Figure 19).

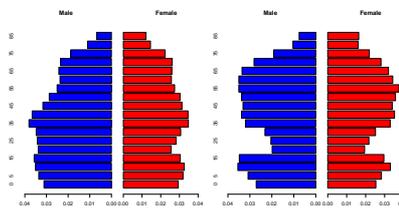


Figure 19: Pyramids' shapes of clusters with older population

First of the last two among nine clusters shows relatively less people older than 30 than the overall American population, while the second one (North and middle of the USA) indicate less people in the 20s (they might be away for study).

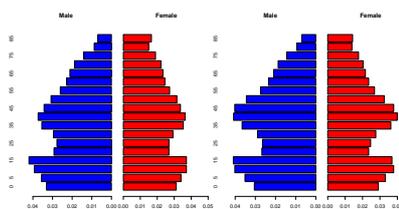


Figure 20: Pyramids' shapes of the two remaining clusters

Because of the relational constraint (regional neighborhood) we got 54 isolated counties at the cutting level 0.1. In Figure 16 they are represented with black circles. They remain isolated because they have different pyramids' shapes than the neighboring counties.

Most of the isolates (29 of them) are due to the proximity of a university. The shape of their population pyramids looks very much like those in Figure 17. 8 other isolates are more gender specific (have mostly more men than women). There are 5 isolates in the state of New York that have slightly different population distribution as the shape of group around them. The other isolates are of two types: they have either considerably less youngsters (or older people) than the surrounding counties or their pyramids look very random due to the small number of inhabitants.

5 Conclusion

Population pyramid is a very popular graphical presentation of the age-sex distribution of the human population of a particular region. Its shape is influenced besides fertility and mortality (usually presented with demographical indicators as birth rates, death rates and growth rates) also by many other social and political policies and events, such as migrations, birth control policy, wars, life-style etc. Population pyramid offers insight into different phenomena in many fields interested in population observations, such as demography, geography, sociology, economy, politics etc.

The aim of the paper was to observe how population pyramids of the world's countries corresponds to the main pyramids' shapes, which are usually related to the "Demographic Transition Model". Although the observation period of 10 years was short for the human life, substantial changes can be seen. Roughly speaking we can conclude from our observations, that the pyramid's shapes of the main clusters correspond to the Stages 1 and Stage between 3 and 4 of the "Demographic Transition Model" in the year 1996, and later are moved to the Stages 2 and even closer to 4 in the 2001 and 2006. The divide between the undeveloped and developed countries is increasing.

Most of the main four clusters are quite stable through observed years. We are aware that for some observers differences are more important than generalization and they can be observed in detail with the separate inspection of the smaller parts of the hierarchy that belong to each cluster.

In the second part, we examine pyramids' shapes of clusters of US counties, because in smaller territorial units the influence of local characteristics is even more emphasized, which reflects also on pyramids' shapes. For clustering 3111 mainland US counties, hierarchical agglomerative clustering procedure with neighborhood constraint was used. The results confirm strong influences of local characteristics (for example universities) on the pyramids' shapes of smaller populations. The clustering procedure exposed some groups of counties with pyramids' shapes which strongly differs from all-American constrictive population pyramid's shape.

Acknowledgement

This work was partially supported by the Slovenian Research Agency, Project J1-6062-0101.

References

- [1] Andreev, L. and Andreev, M. (2004) Analysis of Population Pyramids by a New Method for Intelligent Pattern Recognition, *Matrixreasoning*, Equicom, Inc.
- [2] Andreev, L. (2004) Fusion of Socio-Demographic Variables, *Matrixreasoning*, Equicom, Inc.

<http://www.matrixreasoning.com/publications/1.html>

- [3] Batagelj V., Ferligoj A. and Mrvar A. (2008): Hierarchical clustering in large networks. Presented at Sunbelt XXVIII, 22-27. January 2008, St. Pete Beach, Florida, USA.
- [4] Ferligoj A. and Batagelj V. (1983): Some types of clustering with relational constraints. *Psychometrika*, 48(4), p. 541-552.
- [5] Kaufman, L. and Rousseeuw, P.J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York.
- [6] Pressat, R. (1978) *Statistical Demography* (Translated and adapted by Damien A. Courtney), Methuen, University Press, Cambridge.
- [7] Pressat, R. (1988) *The Dictionary of Demography* (Edited by Wilson, C.), Basil Blackwell.
- [8] International Data Base.
<http://www.census.gov/ipc/www/idbnew.html>
- [9] Mrvar, A. and Batagelj, V. (1996-2008) The Pajek program – home page.
<http://pajek.imfm.si/>
- [10] R Development Core Team (2008) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
<http://www.R-project.org>.

Geometric-Invariant Image Watermarking by Key-Dependent Triangulation

Shiyan Hu
 Department of Electrical and Computer Engineering
 Michigan Technological University
 Houghton, MI 49931, USA
 E-mail: shiyan@mtu.edu

Keywords: image watermarking, geometric invariance, key-dependent triangulation, security, robustness

Received: August 9, 2006

Fast and massive dissemination of image data across the Internet imposes great challenges of protecting images against illegal access and unauthorized reproduction. Image watermarking provides a powerful solution for intellectual protection. In this paper, a new image watermarking approach robust to various geometric distortions is proposed. The new scheme involves detecting image feature points and triangulating them in a secure key-dependent manner. The neighborhood pixel ratio of gray-scale image is investigated in the paper. It is a novel robust image feature which can be seamlessly combined with the proposed key-dependent triangulation scheme. A random pre-warping framework is adopted to make the scheme robust to collusion attack. Our experiments demonstrate that the new scheme is robust to rotation, scaling, noise addition, JPEG compression, StirMark, shearing transformation, collusion, and other common attacks in both spatial and frequency domain.

Povzetek: Predstavljen je nov postopek za varovanje slik na internetu.

1 Introduction

Fast and massive dissemination of image data across the Internet imposes great challenges of protecting images against illegal access and unauthorized reproduction. As an effective and efficient solution, image watermarking superimposes a copyright message into a host image before dissemination and then unauthorized reproduction can be recognized by extracting the copyright information. Numerous image watermarking techniques have been proposed in the literature such as [1, 2, 3, 4, 5, 6, 7, 8]. Along with the rapid growth of novel watermarking schemes, various attacking attempts have also been developed to destroy watermarks. Among these attacks, geometric attacks are very difficult to handle. This is mainly due to the fact that slight geometric manipulation to the marked image, such as scaling or rotation, could significantly reduce the possibility of a successful watermark retrieval, provided that the watermarking extractor has no knowledge of the distortion parameters. In another word, geometric distortion can easily introduce synchronization errors into the watermark extracting process [9, 10]. In recent years, a number of approaches have been proposed to counteract the geometric synchronization attacks. Popular techniques in the literature can be loosely classified into three categories:

1. Geometric invariant domain based watermarking schemes. In these algorithms, Fourier-Mellin transform is incorporated into some watermarking schemes (e.g., [11, 12]) to tackle with geometric attacks such as rotation, scaling and translation. However, these algorithms are computationally inefficient, hard to implement and cannot

survive aspect ratio change [13].

2. Template Matching-based watermarking schemes. In this class of algorithms, a template is embedded into the host image besides the watermark. The affine geometric distortions to the marked image can be reverted using the estimated parameters through detecting the template. After compensating geometric distortions, the watermark can be easily retrieved by the watermarking extractor. The major drawbacks of these techniques are that the template can be easily detected and removed by attackers [10, 14, 15].

3. Content-based watermarking schemes. This class of watermarking schemes achieve recovery from geometric distortion using image content. A particular interesting scheme in this category is proposed in [15] where feature points are used as a content descriptor. The method works as follows. First a set of feature points which are robust to geometric distortion are detected. These points are often near corners or edges of the image. A Delaunay triangulation is computed on the feature points and the watermark is then embedded into the resulting triangles. The above technique has two drawbacks. First, it always computes a Delaunay triangulation on the feature points. Therefore, provided that an attacker can successfully retrieve those feature points, the presence of a watermark can be easily determined and the watermark may be removed or distorted. Since usually well-known feature-point detectors (e.g., Harris detector [16]) are adopted, most feature points can actually be found by attackers. Second, the method is not as robust as expected: feature points are robust (i.e., can be completely retrieved by the watermarking extractor) only against small-degree rotation, and the per-

formance of these techniques is considerably compromised when large-degree rotation occurs. This drawback has also been reported in [15, 17].

In this paper, we propose a new feature point-based image watermarking algorithm which is secure and robust to common attacks in both spatial and frequency domain. The proposed scheme first generates four image feature points using a novel *robust intersection-based feature point detector*. Based on the feature points, a number of additional points are generated and then triangulated in a *key-dependent* manner. Finally, the watermark is embedded into the resulting triangles. The key dependence properties of the proposed technique is motivated by the following results from the computational geometry literature. It is shown in [18] that there exist $\Omega(2.33^n)$ different triangulations for a planar set of n points in general position. Therefore, even if attackers repeat the feature points, they are generally not able to compute the right triangulation. We also consider the application of the proposed watermarking scheme to image fingerprinting. Under this situation, robustness against collusion attacks becomes critical. Therefore, a random pre-warping framework is adopted to make the proposed scheme robust against such attacks.

The performance of the new scheme is substantiated by the extensive experiments. Our experimental results demonstrate that the new scheme is robust to rotation, scaling, noise addition, JPEG compression, StirMark, shearing transformation, collusion, and other common attacks in both spatial and frequency domain.

The rest of the paper is organized as follows: Section 2 describes the robust intersection-based feature point detector. Section 3 describes the key-dependent triangulation-based watermarking scheme. Section 4 presents the experimental results and analysis. A summary of work is given in Section 5.

2 Robust intersection-based feature point detector

The first step is to compute some feature points from an image. To this effect, numerous techniques can be applied, however, even the popular Harris detector [16] cannot guarantee the repeatability of feature points after a large degree rotation [15, 17]. To settle this problem, our strategy is that we first rotate an image by each integer degree, and apply Harris detector to each resulting image. The intersection of the detected points forms the feature point set. Note that smaller degree interval could be applied, however, integer degree interval suffices as indicated in our experiments. The parameters of Harris detector are determined as follows. In principle, we try to find a nice set of parameters such that the intersection of feature points from all images, after *rotated back*, contains only four feature points. Therefore, all that we need to record for the watermarking extractor is the set of these parameters and the key. The latter will be described in Section 3.

For convenience, we use the popular Lena image as an example to illustrate the ideas in this paper. We first rotate the Lena image (of size 512×512) by $1^\circ, 2^\circ, \dots, 359^\circ$ and then apply Harris detector with the same parameter values to each resulting image. Some detection results are shown in Fig. 1. We then rotate each image back (e.g., we rotate the second image in Fig. 1 by 15° in clockwise direction), and compute the intersection of all the feature points after necessary translation. The resulting intersection contains only four points as illustrated in Fig. 2(a). For completeness, some details of our implementation of Harris detector [16] are elaborated as follows.

1. Compute x and y derivatives of image I

$$I_x = G_\sigma^x * I, I_y = G_\sigma^y * I. \quad (1)$$

2. Compute products of derivatives at every pixel

$$I_{x^2} = I_x \cdot I_x, I_{y^2} = I_y \cdot I_y, I_{xy} = I_x \cdot I_y. \quad (2)$$

3. Compute the sums of the products of derivatives at each pixel

$$S_{x^2} = G_{\sigma'} * I_{x^2}, S_{y^2} = G_{\sigma'} * I_{y^2}, S_{xy} = G_{\sigma'} * I_{xy}. \quad (3)$$

4. Define at each pixel (x, y) the matrix

$$H(x, y) = \begin{pmatrix} S_{x^2} & S_{xy} \\ S_{xy} & S_{y^2} \end{pmatrix} \quad (4)$$

5. Compute the response of the detector at each pixel

$$R = \text{Det}(H) - k(\text{Trace}(H))^2. \quad (5)$$

Several parameters are to be determined: the sigma of Gaussian derivatives, the sigma of the Gaussian integration, the k in the computation of “cornerness”, the size of the window for computing the local maximum in R , and finally the threshold for “cornerness”. The parameters used for Fig. 1 are $\sigma = 0.5, \sigma' = 0.8, k = 0.05, \text{Threshold} = 52000$, and window size is set to 30×30 .

The heuristic to determine the parameters reads as follows. σ, σ', k and window size are first set and *Threshold* is changed from larger values to smaller values to obtain the desired effect. It is possible that Harris detector with a set of parameters generate more than four intersection points, while slightly increasing the threshold will lead to less than four intersection points. In this case, we do not increase the threshold, instead, we compute four special points from the obtained intersection points. The following process is also useful for recomputing the intersection points in breaking a tie (see Section 3.1). Suppose that there are k intersection points. We first compute the convex hull [19] on them and three cases follow.

- (1). Exactly four points lie on the convex hull. Then they are returned as the final intersection points.
- (2). More than four points lie on the convex hull. In this case, the hull edges are sorted according to their lengths and points p_i and p_j linking the shortest edge are merged. That is, p_i (resp. p_j) is removed if p_i (resp. p_j) is to the left



Figure 1: Feature points (denoted by +) obtained by Harris detector for Lena after rotation of 0°, 15°, 30°, 60°, 120°, 150°, respectively.



Figure 2: (a) Intersection points: + denotes the intersection of feature points by Harris detector for Lena. (b) New points for Lena: + denotes four feature points, × denotes 30 generated points.

of p_j (resp. p_i) in clockwise direction, and the convex hull is then accordingly updated. The process is repeated until only four points are on the convex hull.

(3). Three points lie on the convex hull. We then arbitrarily pick a feature point inside the hull and return these four points. The newly picked point does not impact the generation of additional points, refer to Section 3.1.

3 Key-dependent triangulation based watermarking

3.1 Generating additional points

Through the above phase, we have four feature points in hand. The following process is carried out to generate N new points in a key-dependent manner. Key-dependent property involves using pseudo-random numbers. Throughout this paper, pseudo-random numbers are generated depending on a secret key, which is stored for the watermarking extractor. The procedure reads as follows.

(1). In the case of four hull vertices, we first compute the longer one of the two *diagonal* segments formed by these vertices. Denote by p_a, p_c the two endpoints and by p_b, p_d other points where p_a, p_b, p_c, p_d are in clock-wise order and $p_a = \operatorname{argmax}_{p_i \in \{p_a, p_c\}} d_2(p_i, p_b) + d_2(p_i, p_d)$, $d_2(\cdot)$ being the Euclidean distance function. In the case of three hull vertices, the longest hull edge is returned as $p_a p_c$ and another hull vertex is p_b such that p_a, p_b, p_c are in clock-wise order. p_d is the point inside the hull. Rotate the image such that $\overrightarrow{p_c p_a}$ is 45° with respect to the horizontal direction. Whenever there is a tie, we choose another set of parameters for Harris detector to recompute the intersection points.

Note that images shown in this paper are first rotated back to its original position for the convenience of illustration.

(2). Compute the bounding box of feature points in the rotated image as follows. Let $A = \{a, b, c, d\}$, and let $\min x = \min_{i \in A} \{x_i\}$, $\max x = \max_{i \in A} \{x_i\}$, $\min y = \min_{i \in A} \{y_i\}$, $\max y = \max_{i \in A} \{y_i\}$. The bounding box is defined by $\{(minx, miny), (maxx, miny), (maxx, maxy), (minx, maxy)\}$.

(3). Generate two uniform deviates λ_1 and λ_2 with our key. A new point is generated as $(minx \cdot \lambda_1 + maxx \cdot (1 - \lambda_1), miny \cdot \lambda_2 + maxy \cdot (1 - \lambda_2))$.

(4). Repeat Step (3) until the number of new points reaches N .

Refer to Fig. 2(b) for 30 newly generated points. Taking the first generated point as the *origin* and the vector formed by the first and the second points as the direction of x -coordinate, we build up a *reference coordinate system* conforming to the right-hand rule. Note that the original four feature points will not be used in the following watermark embedding process. We are now ready to triangulate the N new points in a key-dependent way.

3.2 Computing key-dependent triangulation

Refer to Fig. 3(a) for a reference coordinate system with the origin v_s , where the dotted line with an arrow represents x -axis. Recall that a *triangulation* of a planar point set is a maximal set of non-intersecting straight-line segments connecting points in it [19]. The key-dependent triangulation is computed as follows.

(1). Sort all vertices (i.e., points) by their polar angle in the reference system. The distance to the origin v_s is used



Figure 4: Sample triangulations for Lena with different generated points.

for tie breaking. Denote the resulting set by V^1 . Refer to Fig. 3(b) where the vertices are numbered in the order. (2). The triangulation is built incrementally and involves using pseudo-random numbers. First compute as V_{v_s} the set of all vertices visible to v_s in V . Note that v_a and v_b are visible to each other if v_a, v_b are distinct and the line segment $v_a v_b$ does not intersect with any existing segment in the incomplete triangulation. Of course, we require that $v_a v_b$ itself has not been inserted yet. In Fig. 3(b), there is no inserted segment, so $V_{v_s} = \{v_1, \dots, v_8\}$. Randomly pick a vertex v_{h_1} from V_{v_s} and insert line segment $v_s v_{h_1}$. We then compute $V_{v_{h_1}}$ which contains all the vertices visible to v_{h_1} , randomly select a vertex v_{h_2} from $V_{v_{h_1}}$, and insert the segment $v_{h_1} v_{h_2}$ into the graph. As an illustration, suppose that at a time point, the incomplete triangulation is as Fig. 3(c) and the current vertex is v_7 , then $V_{v_7} = \{v_4, v_6, v_8\}$. The above process is repeated until either the triangulation is completed or no vertex is visible to the current vertex. Suppose that the triangulation proceeds to Fig. 3(d). The last few visited points form the sequence of $v_8 - v_1 - v_2 - v_3 - v_5 - v_4$. Since there is no vertex visible to the current vertex v_4 , we have to find the next vertex through backtracking, i.e., the current vertex is changed to v_5 which is the most recent vertex except v_4 . If yet no visible vertex for v_5 , we will continue the backtracking process. In the case of Fig. 3(d), we need to backtrack to v_3 where $V_{v_3} = \{v_s, v_1, v_8\}$. A complete triangulation is shown in Fig. 3(e). For the Lena image, two sample triangulations are shown in Fig. 4.

3.3 Watermark embedding process

Suppose that the reference coordinate system and the triangulation can be repeated after various attacks to the original image, the present issue is to come up with a robust embedding method for each triangle. For this purpose, we investigate as follows a neighborhood information-based image feature which is robust to various attacks such as noise, JPEG compression and geometric distortions. This feature has been successfully applied to watermarking binary document images in the previous work [20]. However, more work is needed for extending this tool to watermark

¹ $v_s \in V$ may be regarded as the 0-th vertex.

grayscale images.

First note that too small triangles (with the area below a threshold) are first eliminated from consideration. All remaining triangles are then ordered/indexed in the following way. For two triangles defined by vertices v_a, v_b, v_c and v_d, v_e, v_f respectively (without loss of generality, assume that $a < b < c, d < e < f$), the order of them is determined by a and d ; If $a = d$, then compare b and e ; If still tie, then compare c and f . Given N points, all triangles are uniquely indexed from 1 to $N' \leq 2N - k - 2$ where k is the number of vertices on the convex hull of these N points (see [19]). Note that the inequality is due to removal of small triangles. We call each triangle a *partition* of the image. Denote each partition as $par_i, i = 1, \dots, N'$.

We now discuss how to embed the watermark bitstream to a single partition. The *weight* $w(par_i)$ of a partition par_i is defined as follows. A counter, initialized to 0, is associated to par_i . For each pixel g inside par_i , we check for its eight neighbors: if more than three neighbors of g have intensity values larger than a threshold T_L , the counter corresponding to par_i is incremented. $w(par_i)$ is defined as $\frac{counter}{area_i}$ where $area_i$ denotes the area in pixels of par_i . We call this ratio (or *partition weight*) the *neighborhood pixel ratio* (NPR). The NPR ratio for gray-scale image is an extension of the NPR ratio for binary image, which is a robust feature as shown in [20].

For a single partition, noise can be often “filtered out”, e.g., random noise can be “filtered out” from incrementing the counter since we increase it only when at least four neighbors of a pixel have intensity values larger than T_L . Such a computation captures the intrinsic characteristic of a partition to some extent. We illustrate this fact through an example, refer to Fig. 5(a) for a partition from Lena. The involved threshold T_L is set to 120.

The original partition Fig. 5(a) and the noisy partition Fig. 5(b) are visually different, however, their neighborhood pixel ratios are 0.3267 and 0.3140, respectively. Two ratios differ only by 1%! We also test NPR ratio for scaling, rotation, JPEG compression, low pass and median filtering. Refer to Table 1 for the resulting NPR ratios. All ratios are similar to each other. In contrast, the ratio of another partition from Lena shown in Fig. 5(d) computes to 0.2183, which differs from the original’s by 31%! The combinations of the above attacks are also performed, and the NPR ratio is resistant to these attacks. As an example, Fig. 5(c) shows a rotated noisy partition whose NPR ratio computes to 0.3129, very close to 0.3267. The above experiments demonstrate the robustness of the neighborhood pixel ratio. Refer to Section 4 for further experiments on robustness against geometric distortions.

It remains to present the principle for modifying a partition - we only modify pixels whose intensity values are close to T_L . For each such pixel, we may either increase or decrease the intensity value depending on whether the ratio is to be increased or decreased. For example, increasing a pixel’s intensity from $T_L - 2$ to T_L is possible to increase the partition weight. The process is repeated un-

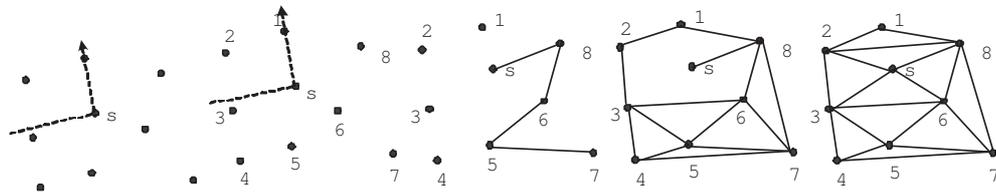


Figure 3: A simple example for key-dependent triangulation, from left to right (a)(b)(c)(d)(e).

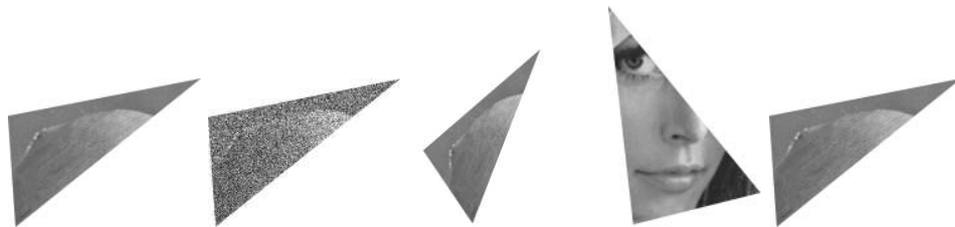


Figure 5: Top to bottom, left to right: (a) the original partition (b) the partition corrupted with synthetic Gaussian noise with $\sigma = 30$, (c) rotating by 30° followed by noise addition (the resulting image is scaled down here due to space limitation) (d) another partition (e) a modified partition of (a).

Table 1: NPR ratio for the attacked image partition.

Attack	A partition with the weight of 0.3267
JPEG with QF of 10%	0.3178
Add. White. Gauss. Noise, $\sigma = 20$	0.3140
Low Pass Filter	0.3126
3×3 Median Filter	0.3098
Scale by Factor of 0.5	0.3248
Scale by Factor of 2.0	0.3263
Rotation by 30°	0.3202
Rotation by 60°	0.3126

til the counter reaches the goal within an error of $4/area$. For instance, refer to Fig. 5(e) for a modified partition (of Fig. 5(a)) whose ratio is 0.2651 compared to the original ratio 0.3267. The modification is not visually perceptible.

Based on the NPR ratio, we are ready to present the watermark embedding process, which is motivated by [21]. Recall that every triangle is indexed. We first randomly select $\lfloor N'/2 \rfloor$ triangles and denote the triangle sequence by $O_1, O_2, \dots, O_{\lfloor N'/2 \rfloor}$. We then randomly select the remaining triangles to form the sequence $Z_1, Z_2, \dots, Z_{\lfloor N'/2 \rfloor}$.

The watermarking extractor works to retrieve the embedded bitstream. It compares $w(O_i)$ to $w(Z_i)$ for each i to decide about the marked bit: if $w(O_i) - w(Z_i) \geq T_J$, then a 1 is embedded; if $w(Z_i) - w(O_i) \geq T_J$, then a 0 is embedded. Therefore, the watermarking embedder needs to accordingly modify the relationship between $w(O_i)$ and $w(Z_i)$ to embed bitstream. To this effect, without loss of generality, assume that we aim to embed a 0, however, presently $w(O_i) > w(Z_i)$. In this case, we modify the partitions to change $w(O_i)$ to $\frac{w(O_i)+w(Z_i)}{2} - \frac{T_J}{2}$ and $w(Z_i)$ to $\frac{w(O_i)+w(Z_i)}{2} + \frac{T_J}{2}$.

3.4 Robustness against collusion attacks

In this section, we discuss the application of the proposed watermarking algorithm to image fingerprinting. The term *fingerprinting* refers to superimposing a unique watermark onto each copy of the distributed data. The embedded watermark can be used to identify the unauthorized copies of the data [22]. Digital watermarking can naturally serve as an effective and efficient approach to fingerprint digital data. However, a main shortcoming in applying conventional image watermarking techniques for fingerprinting is that they are not designed to be robust against *collusion* attacks, which are common attacks to destroy fingerprints. These attacks are performed by a group of colluders with the same digital data containing different fingerprints. A common implementation of such an attack is simply averaging multiple marked versions of an image [23]. Most existing watermarking schemes robust to collusion, such as [24, 25, 26, 27, 28, 29], have shortcomings including compromised watermarking capacity and decreasing effectiveness with the increasing number of colluders [22].

In this paper, we adopt the random pre-warping framework originally proposed in [22] to design a collusion and geometric resistant watermarking scheme. For completeness, some details of the approach in [22] are included as follows. Other than trying to detect collusion and identify traitors by fingerprint, the random pre-warping framework shoots for preventing traitors from obtaining a high-quality copy through collusion. Basically, the method randomly distorts the host image *before* embedding fingerprint to it such that averaging multiple versions will introduce annoying artifacts and only result in a low-quality image with no commercial value. The idea is feasible due to the following reasons as shown in [22]. For additive watermarking procedures, an averaged image from K distinct copies can be

represented as [22]

$$S_a = \frac{1}{K} \sum_{i=1}^K S_i = S + \frac{1}{K} \sum_{i=1}^K W_i, \quad (6)$$

where S is the host image, W_i is the watermark, and $S_i = S + W_i$ is the watermarked image. It is expected that S_a looks similar to S due to the fact that $\frac{1}{K} \sum_{i=1}^K W_i$ should vanish since each watermark W_i can be regarded as a random pattern. In contrast, if we distort S before superimposing W_i onto it, we have [22]

$$S_a = \frac{1}{K} \sum_{i=1}^K S_i = \frac{1}{K} \sum_{i=1}^K \phi_i(S) + \frac{1}{K} \sum_{i=1}^K W_i, \quad (7)$$

where $\phi(\cdot)$ denotes the distortion function. Even if the second term vanishes, we can choose $\phi(\cdot)$ such that $\frac{1}{K} \sum_{i=1}^K \phi_i(S)$ is visually different from S citeCST04. It is shown in [22] that this can be achieved using the standard Stirmark tool [30, 31] which forms the basis of desynchronization attacks on many watermarking schemes.

With the above introduction, a watermarking scheme robust to both geometric attacks and collusion attacks is clear: we first randomly pre-warp the image followed by embedding watermarks as in the previous sections.

The complete process for embedding watermarks onto a host image is summarized in Algorithm 1. To extract a watermark from a possibly modified marked image, we carry out the extracting process as shown in Algorithm 2.

4 Experimental results

We have performed experiments over various gray-scale images. We choose the Lena image to present our results and analysis. The extensive experiments on our whole image set are described in the end of this section (see Table 3).

To evaluate the robustness of the proposed watermarking scheme, common attacks are tested. For a possibly modified watermarked triangulation, we define the watermark strength as follows. Recall that $w(Z_i) - w(O_i) \geq T_J$ denotes embedding 0 while $w(O_i) - w(Z_i) \geq T_J$ denotes embedding 1. When a marked triangle is attacked, a 1 (resp. 0) can be extracted if $w(O_i) - w(Z_i) > 0$ (resp. $w(O_i) - w(Z_i) < 0$). Therefore, our scheme can tolerate up to T_J unit changes in triangle weight. Denote by ξ_1 (resp. ξ_2) the smallest value of $w(O_i) - w(Z_i)$ (resp. $w(Z_i) - w(O_i)$) for all i where a 1 (resp. 0) is actually embedded. The *watermark strength* is defined as $\frac{\min\{\xi_1, \xi_2\}}{T_J}$. Clearly, the watermark becomes more robust with the increasing value of the watermark strength. It follows from Section 3.3 that the maximum possible value of watermark strength is 1². If a watermark strength is negative, the watermark may not be correctly extracted. When this hap-

²Exception occurs when relationship between $w(Z_i)$ and $w(O_i)$ for each i exactly matches the embedding bit sequence. However, it is very unlikely and not observed in the experiments.

Algorithm 1 Watermark Embedding Process

- 1: Use StirMark to randomly pre-warp the image. Only geometric distortions in StirMark is applied.
 - 2: Compute the four feature points as in Section 2, i.e., determine a set of parameters such that the intersection of feature points for all rotated image versions contains only four points. The set of parameters is recorded for the watermarking extractor.
 - 3: Based on the intersection points, generate N new points in a key-dependent manner. The key is recorded for the watermarking extractor.
 - 4: Triangulate the generated points in a key-dependent manner.
 - 5: Remove too small triangles whose areas are below a threshold and index the remaining ones.
 - 6: Use two pseudo-random triangle sequences to embed the watermark as in Section 3.3.
-

Algorithm 2 Watermark Extracting Process

- 1: Use the set of recorded parameters to compute the four intersection feature points, i.e., rotate the possibly modified marked image and the intersection of feature points for all rotated image versions should contain only four points.
 - 2: Based on the computed four feature points, generate N new points using the user's key.
 - 3: Triangulate the generated N points using the user's key.
 - 4: Remove too small triangles and index the remaining ones.
 - 5: Generate two pseudo-random triangle sequences with the key and extract the watermark as follows. Compare $w(O_i)$ with $w(Z_i)$ for each i : if $w(O_i) - w(Z_i) \geq T_J$, then a 1 is embedded; if $w(Z_i) - w(O_i) \geq T_J$, then a 0 is embedded. We set up a small positive value δ for fault-tolerance purpose, i.e., we treat $T_J \approx T_J \pm \delta$.
-

pens, we are yet interested in how many bits can be correctly extracted. We are now ready to present our experimental results.

1. The different image transformations tested are scaling, rotation, and a combination of these transformations. The original Lena image of size 512×512 is shown in Fig. 6(a). The watermarked Lena image is shown in Fig. 6(b). The involved T_L is set to the average intensity value of each image. PSNR for the watermarked image is low (18.82) due to the random pre-warping by StirMark. Without applying random pre-warping, the PSNR for watermarked image is 46.13 (see Fig. 6(c)). In the case of single transformation, the proposed technique is robust to any degree of rotation (without cropping) and scaling with factor of 0.5 and 2, respectively. In the above attacks, all embedded bits are extracted. Note that in scaling attacks, all generated points lie in the bounding box of the rotated convex hull and are dependent on some random ratios (i.e., λ_1, λ_2). Therefore, as long as the scaling transformation is uniform to the image and the index (i.e., the order) of bounding box vertices is repeatable, the generated points and thus the triangulation are repeatable.

In the case of combined transformation, we test the scheme on various sequences of transformations, e.g., Fig. 6(d) shows the resulting image after the sequence of transformations including scaling by factor of 2, 20° rotation and cropping that maintains 60% of the image. All bits are successfully extracted in this class of tests, except those with too much cropping such that feature points have been removed. Since feature points are intrinsic to an image, we consider that such an attack has degraded the quality of the original image to a significant extent and thus the cropped image becomes less useful.

2. For further analysis, we have tested the proposed watermarking scheme for nonlinear geometric attacks through StirMark (with default parameter values), shearing transformation (Gimp software [32] is used to perform this transformation), compression attacks using JPEG (with a wide variety of quality factors ranging from 10% to 90%), and addition of Gaussian noise with $\sigma = 20$. In all cases, successful retrieval of watermark is reported in our experiments. Note that applying StirMark to a watermarked image still gives acceptable quality of image, see Fig. 6(e). Refer to Fig. 6(f) for the image after shearing distortion.

3. To show the robustness of the proposed scheme to re-watermarking, it is necessary to consider the following scenario. We first embed a watermark W_x to an image, then another watermark W_y is embedded to the marked image. We also embed W_y directly to the original image. Two resulting images must be different, and this is the case as demonstrated by our experiment. In addition, we even re-watermark a marked image (1) without applying the StirMark (since it usually causes considerable difference) (2) using the same set of parameters for detecting feature points. That is, two embedding processes only differ in the generated points and thus the triangulations, both of which are purely dependent on the user's key. Refer to



Figure 6: Left to Right, Top to bottom: (a) original Lena image, (b) watermarked Lena with random pre-warping, (c) watermarked Lena without random pre-warping, (d) scaled, rotated and cropped image, (e) StirMarked watermarked image, (f) marked image after shearing distortion, (g) re-watermarking of marked image.

Fig. 6(g) for a re-watermarked Lena over Fig. 6(b). The normalized L_1 distance between Fig. 6(b) and Fig. 6(g) is 4.02, i.e., on average the corresponding intensity values in the two images differ by 4.

4. The next test is to modify the intensity values of the marked image by either increasing or decreasing intensity value by a fixed small amount. Recall that the involved average intensity value T_L is set to the average intensity value of the whole image, thus the watermark is robust to such attacks (refer to Table 2). It is still the case for modifying intensity value by some small random amounts. A more effective attack is to modify the intensity value very close to the average intensity value. However, this attack will not really cause trouble, since instead of setting T_L to the average intensity value, we can set T_L to be the average intensity value multiplied by a key-dependent random number between $[0.5, 1.5]$.

The improved watermarking scheme by key-dependent T_L is tested and with $\delta = T_L/2$, we are usually able to extract all bits. The exception occurs when the attacker correctly guesses T_L and randomly exchanges considerable amount of pixels across T_L . In that case, we are still able to extract more than 80% bits as indicated by our experiments. Assume that the maximum amount for modifying any intensity value is 3, then to effectively defeat the watermarking scheme, one needs to correctly estimate the average intensity value as within the range of $[T_L - 3, T_L + 3]$. For an image with the average intensity value of 120, T_L may be of any value in $[60, 180]$. Thus, an effective guess happens only with a probability of $6/120 = 5\%$. Refer to Table 2 for quantitative results. Refer to Fig. 7 for watermark strength versus trials of attacks. Out of 100 attacks, only 6 attacks effectively defeat the scheme.

5. Spatial domain filtering is also a class of common attacks. In our experiment, 3×3 median filtering and 3×3 mean filtering are considered. A frequency-domain low

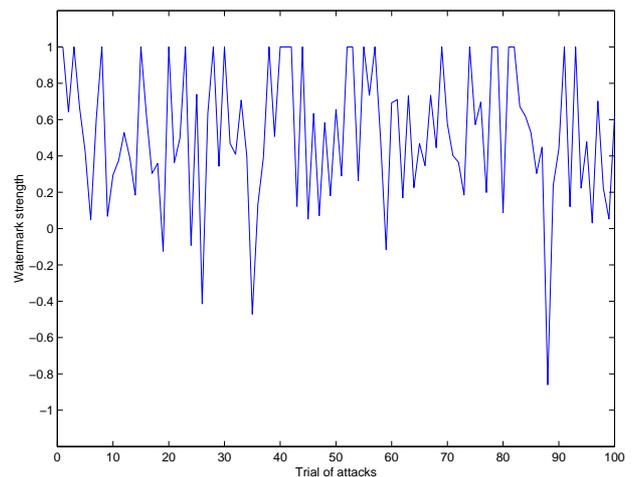


Figure 7: Watermark strength in different trials of attacks through random intensity manipulation in spatial domain.

pass filter is also applied. Our further analysis includes a random change in the image’s frequency domain. The maximum possible change to amplitude of FFT coefficients is $\pm 10\%$ of the original value. Refer to Table 2 for the watermark strength and the ratio of correctly detected bits after these attacks.

6. Printing and Scanning. The print and scan test combines multiple attacks. For example, printing introduces re-quantization and grid apparition while scanning introduces geometric distortions [15]. Our experiments (refer to Table 2) demonstrate that the proposed scheme is robust to the print/scan attack.

7. Collusion Attacks. Due to the random pre-warping framework, the quality of colluded images should be significantly degraded. This is the case as indicated by our experiment. Refer to Fig. 8 for colluded copies of water-

marked images. Both Lenas are significantly blurred and are thus of less commercial value.

In concluding this section, we present the results for carrying out all the above tests on 50 collected images, refer to Table 3. All ratios shown are averaged over 50 images. Note that for the combinational attack with cropping, the watermark strength is not computed since some triangulations are not repeatable due to too much cropping and thus no ratio can be computed in those cases. From Table 3, one sees that the watermark strength is high for all types of attacks and on average, more than 80% marked bits can be correctly extracted even for the most effective attacks. Our experimental results clearly demonstrate the effectiveness of the proposed method.

5 Conclusion

We propose a new content-based image watermarking scheme. The scheme belongs to the class of second generation watermarking schemes whose advantages include automatic re-synchronization and exclusion of unreliable template embedding [15, 33]. The strength of the proposed scheme is demonstrated through successful watermark detection after various common attacks such as geometric distortions, StirMark attacks and shearing transformations. The main contribution of this paper is three-fold. First, a spatial domain key-dependent triangulation framework is proposed. Based on the framework, a highly secure and robust image watermarking scheme is presented. Second, a novel feature for gray-scale images, the neighborhood pixel ratio is investigated in this paper. It is an extension of the binary image NPR ratio presented in our previous work [20]. Third, detecting rotation-invariant feature points through inspecting all rotated images is investigated in this paper. Such an idea may have its own interest as well.

The proposed key-dependent triangulation framework can be easily combined with other watermarking techniques to obtain a highly secure watermarking scheme. For example, NPR ratio-based embedding process (for each triangle) could be substituted by proper existing geometric-resistant techniques. To design and analyze the combination of the key-dependent triangulation framework with the existing watermarking methods would be an interesting future work.

Acknowledgement

The author would like to thank the anonymous reviewers for their helpful comments.

References

[1] I. Cox, J. Kilian, F. Leighton, and T. Shamon, "Secure spread spectrum watermarking for multime-

dia," *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1673 – 1687, 1997.

- [2] J. Hernandez, M. Amado, and F. Perez-Gonzalez, "Dct-domain watermarking techniques for still images: detector performance analysis and a new structure," *IEEE Transactions on Image Processing*, vol. 9, no. 1, pp. 55 – 68, 2000.
- [3] Z.-M. Lu, D.-G. Xu, and S.-H. Sun, "Multipurpose image watermarking algorithm based on multistage vector quantization," *IEEE Transactions on Image Processing*, vol. 14, no. 5, pp. 822 – 831, 2005.
- [4] S.-H. Wang and Y.-P. Lin, "Wavelet tree quantization for copyright protection watermarking," *IEEE Transactions on Image Processing*, vol. 13, no. 2, pp. 154 – 165, 2004.
- [5] C.-H. Chang, Z. Ye, and M. Zhang, "Fuzzy-art based adaptive digital watermarking scheme," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 1, pp. 65 – 81, 2005.
- [6] J. Cannons and P. Moulin, "Design and statistical analysis of a hash-aided image watermarking system," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1393 – 1408, 2004.
- [7] A. Alattar, "Reversible watermark using the difference expansion of a generalized integer transform," *IEEE Transactions on Image Processing*, vol. 13, no. 8, pp. 1147 – 1156, 2004.
- [8] J. Tzeng, W.-L. Hwang, and I.-L. Chern, "An asymmetric subspace watermarking method for copyright protection," *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 784 – 792, 2005.
- [9] M. Alghoniemy and A. Tewfik, "Geometric invariance in image watermarking," *IEEE Transactions on Image Processing*, vol. 13, no. 2, 2004.
- [10] P. Dong and N. Galatsanos, "Affine transformation resistant watermarking based on image normalization," *Proceedings of International Conference on Image Processing (ICIP)*, vol. 3, pp. 489 – 492, 2002.
- [11] J. Ruanaidh and T. Pun, "Rotation, scale and translation invariant digital image watermarking," *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 1997.
- [12] C.-Y. Lin, M. Wu, J. Bloom, M. Miller, I. Cox, and Y.-M. Lui, "Rotation, scale, and translation resilient public watermarking for images," *IEEE Transactions on Image Processing*, vol. 10, no. 5, pp. 767–782, 2001.
- [13] X. Qi and J. Qi, "Improved affine resistant watermarking by using robust templates," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. III, pp. 405–408, 2004.

Table 2: Watermark strength of the attacked Lena image.

Image Attack	Lena	
	Watermark strength	Ratio of cor. det. bits
Scale by Factor 0.5	0.95	100%
Scale by Factor 2.0	0.98	100%
Rotation by 10°	1	100%
Rotation by 20°	0.98	100%
StirMark	0.76	100%
Shearing	0.82	100%
JPEG QF 10%	0.91	100%
Noise	0.83	100%
Low Pass Filter	0.80	100%
3 × 3 Median Filter	0.91	100%
3 × 3 Mean Filter	0.82	100%
Rand. change by same amount	0.91	100%
Worst result in Rand. change in Spa. Domain	-0.25	87%
Worst result in Rand. change in Fre. Domain	0.18	100%
Print and Scan	0.32	100%



Figure 8: Colluded copies of Lena. (a) colluded image using two copies (b) colluded image using five copies.

Table 3: Averaged watermark strength over 50 attacked images.

Attack	Avg. watermark str.	Ratio of cor. det. bits
Scale by Factor 0.5	0.95	100%
Scale by Factor 2.0	0.97	100%
Rotation by 10°	0.99	100%
Rotation by 20°	0.96	100%
Rotation by 60° (no cropping)	0.93	100%
Combinational attack (no cropping)	0.95	100%
Combinational attack (with cropping)	-	83%
StirMark	0.81	98%
Shearing	0.83	98%
JPEG QF 10%	0.89	100%
Noise	0.86	99%
Low Pass Filter	0.81	100%
3 × 3 Median Filter	0.91	100%
3 × 3 Mean Filter	0.85	100%
Rand. change in Spa. Domain	0.58	89%
Rand. change in Fre. Domain	0.70	96%
Print and Scan	0.65	96%

- [14] A. Herrigel, S. Voloshynovskiy, and Y. Rytsar, “The watermark template attack,” *Proceedings of SPIE: Security and Watermarking of Multimedia Contents III*, pp. 394–405, 2001.
- [15] P. Bas, J.-M. Chassery, and B. Macq, “Geometrically invariant watermarking using feature point,” *IEEE Transactions on Image Processing*, vol. 11, no. 9, pp. 1014–1028, 2002.
- [16] C. Harris and M. Stephens, “A combined corner and edge detector,” *Proceedings of 4th Alvey Vision Conference*, pp. 147–151, 1988.
- [17] J. Dittmann, T. Fiebig, and R. Steinmetz, “New approach for transformation-invariant image and video watermarking in the spatial domain: self-spanning patterns (ssp),” *Proceedings of SPIE: Security and Watermarking of Multimedia Contents II*, vol. 3971, pp. 176–185, 2000.
- [18] O. Aichholzer, F. Hurtado, and M. Noy, “A lower bound on the number of triangulations of planar point sets,” *Computational Geometry: Theory and Applications*, vol. 29, no. 2, pp. 135–145, 2004.
- [19] M. de Berg, M. van Kreveld, M. Overmars, and O. Schwarzkopf, *Computational Geometry: Algorithms and Applications*, 2nd ed. Springer-Verlag, 2000.
- [20] S. Hu, “Document image watermarking algorithm based on neighborhood pixel ratio,” *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.
- [21] J. Smith and B. Comiskey, “Modulation and information hiding in images,” *Proceedings of Information Hiding Workshop*, 1996.
- [22] M. Celik, G. Sharma, and A. Tekalp, “Collusion-resistant fingerprinting by random pre-warping,” *IEEE Signal Processing Letters*, vol. 11, no. 10, pp. 831 – 835, 2004.
- [23] M. Wu, W. Trappe, Z. Wang, and K. Liu, “Collusion-resistant fingerprinting for multimedia,” *Signal Processing Magazine, IEEE*, vol. 21, no. 2, pp. 15 – 27, 2004.
- [24] D. Boneh and J. Show, “Collusion-secure fingerprinting for digital data,” *IEEE Transactions on Information Theory*, vol. 44, no. 5, pp. 1897–1905, 1998.
- [25] J. Dittmann, A. Behr, M. Stabenau, P. Schmitt, J. Schwenk, and J. Ueberberg, “Combining digital watermarks and collision secure fingerprints for digital images,” *Proceedings of SPIE: Security and Watermarking of Multimedia Contents*, pp. 171–182, 1999.
- [26] J. Su, J. Eggers, and B. Girod, “Capacity of digital watermarks subjected to an optimal collusion attack,” *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2000.

- [27] J. Domingo-Ferrer and J. Herrera-Joancomarti, “Simple collusion-secure fingerprinting schemes for images,” *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC)*, 2000.
- [28] J. Dittmann, P. Schmitt, E. Saar, J. Schwenk, and J. Ueberberg, “Combining digital watermarks and collusion secure fingerprints for digital images,” *Journal of Electronic Imaging*, vol. 9, 2000.
- [29] Z. Wang, M. Wu, W. Trappe, and K. Liu, “Anti-collusion of group-oriented fingerprinting,” *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2003.
- [30] F. Petitcolas, R. Anderson, and M. Kuhn, “Attacks on copyright marking systems,” *Proceedings of the Second Information Hiding Workshop, LNCS vol. 1525*, Springer-Verlag, New York, pp. 219–239, 1998.
- [31] F. Petitcolas, “Watermarking schemes evaluation,” *IEEE Signal Processing*, vol. 17, no. 5, pp. 58–64, 2000.
- [32] M. Cutts, “An introduction to the gimp,” <http://www.acm.org/crossroads/xrds3-4/gimp.html>, 1997.
- [33] M. Kutter, S. Bhattacharjee, and T. Ebrahimi, “Towards second generation watermarking schemes,” *Proceedings of International Conference on Image Processing (ICIP)*, vol. 1, pp. 320 – 323, 1999.

A Simple Algorithm for the Restoration of Clipped Speech Signal

Abdelhakim Dahimene, Mohamed Nouredine and Aarab Azrar
 Electrical and Electronic Engineering department, Boumerdes University
 Boumerdes, Algeria, 35000
 E-mail: dahimenehakim@yahoo.fr

Keywords: speech signal, clipped speech, restoration, interpolation, linear prediction, least square method, Kalman filter

Received: August 6, 2007

This paper deals with the problem of peak clipped speech. Our basic assumption is that the clipped speech is voiced and can be linearly predicted with a high accuracy. The coefficients of linear prediction are computed using two different algorithms: a least square direct method and a recursive Kalman filter. The speech reconstruction is accomplished using backward prediction.

Povzetek: Predstavljen je algoritem za obnavljanje zvočnega signala.

1 Introduction

Speech acquired by personal computer sound cards is often confronted with two main problems: DC level wandering and peak clipping. While building a data base for our speech recognition project, we have been confronted with both problems. The first one is easily eliminated by simple linear processing but the second one requires more complex algorithms. Peak clipping is fundamentally a non linear distortion. It is characterized by the fact that several successive values of the signal disappear and are replaced by a constant. However, it happens that speech signal is highly predictable. So, in essence, peak clipped speech restoration is a problem of interpolation since we are trying to find missing values by using the properties of the signal itself. There exist several methods of interpolation: polynomial (Lagrange, Newton), spline, etc. In the case of peaked clipped speech, an appropriate method is statistical interpolation [1].

2 Justification of the method

When there is no a priori information on the signal, the classical numerical interpolation methods (polynomial and spline) should be used. Band limited interpolation [2] uses only the fact that the signal is band limited. Statistical interpolation based on linear prediction [2, 4] uses the fact that that speech signal is highly predictable. A speech segment is composed of a sequence of voiced, unvoiced and silence (noise) segments [2]. The type of speech signal that has the greatest probability for being peak clipped is voiced speech [2, 3]. Figure 1 represents a scatter plot of voiced, unvoiced and silence mean magnitude and zero crossing rate of segments of speech. Voiced speech segments are indicated by the letter "V", unvoiced segments by the letter "U" and the silent segments by the letter "S". It shows clearly that the voiced signals cluster at high mean magnitude values.

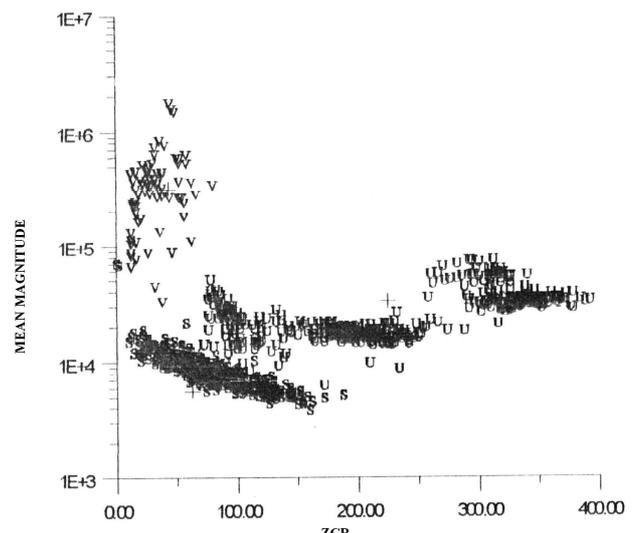


Figure 1: Mean magnitude and ZCR scatter plot [3]

The mean magnitude is defined as:

$$M_n = \sum_{m=-\infty}^{\infty} |x(m)|w(n-m) \quad (1)$$

where $w(n)$ is a rectangular window of length 256 samples and the zero crossing rate (ZCR) is:

$$ZCR(n) = \sum_{m=-\infty}^{+\infty} |\text{sgn}[x(m)] - \text{sgn}[x(m-1)]|w(n-m) \quad (2)$$

Fortunately, voiced speech happens to be quite predictable. Voiced speech follows quite closely the linear prediction equations [4, 5]. Commercial software like DC-6, from Diamond Cut products, use low order linear prediction for clipped audio signal restoration and the problem of audio signal interpolation have also been addressed by Vaseghi [1] who uses linear prediction from adjacent samples and samples one period away (audio signals are assumed to be periodic).

Voiced speech can be considered as a quasi periodic signal. It can be modelled as the output of a linear time invariant system (during few milliseconds, the system can safely be assumed to be time invariant) driven by a periodic train of impulses. In this case, a quite general formulation of the signal will be:

$$x_n = \sum_{k=1}^p a_k x_{n-k} + \sum_{k=0}^p b_k u_{n-k} \quad (3)$$

where the signal u_k is equal to 1 every T seconds and zero otherwise, T being the pitch period. a_k and b_k are respectively the recursive and the non recursive parameters of the above production filter of order p . So, within a pitch period (N_T samples) and after p samples, we can write:

$$x_n = \sum_{k=1}^p a_k x_{n-k} \quad (4)$$

The above equation breaks down in the part of the speech signal that is clipped. So, if we start the time axis at the beginning of a pitch period and if we call N_T the number of samples within the pitch period, we can write:

$$x_n = \sum_{k=1}^p a_k x_{n-k} ; p \leq n \leq N_T \quad (5)$$

for $|x_n| < X_{\max}$

X_{\max} being the saturation value.

3 The proposed restoration algorithm

The proposed algorithm for clipped speech restoration is going to be based on linearly predicting the missing values using equation (4). So, the algorithm consists of two following steps:

- Computation of the prediction coefficients a_k .
- Linear prediction of the missing values.

3.1 Computation of the prediction coefficients

The computation of the prediction coefficients a_k can be accomplished either by using a least square solution or by using a recursive algorithm based on Kalman filtering.

For the least square algorithm, we can use equation (5) and build the following matrix vector equation relating speech samples x_k :

$$\begin{pmatrix} x_{p+1} \\ x_{p+2} \\ \cdot \\ \cdot \\ x_{N_T} \end{pmatrix} = \begin{pmatrix} x_p & x_{p-1} & \cdot & \cdot & x_1 \\ x_{p+1} & x_p & \cdot & \cdot & x_2 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{N_T-1} & x_{N_T-2} & \cdot & \cdot & x_{N_T-p} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ a_p \end{pmatrix} \quad (6)$$

in which all the rows such that $|x_k| = X_{\max}$ are deleted.

Equation (6) can be written as:

$$\mathbf{b} = \mathbf{X} \mathbf{a} \quad (7)$$

and the least square solution of equation (7) can be obtained as [6]:

$$\mathbf{a} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{b} \quad (8)$$

Another approach to the evaluation of the prediction coefficient is the following sequential algorithm (Kalman filter) [7, 8] based on the subsequent set of equations and on an autoregressive model. Consider the next state equation:

$$\mathbf{a}(k+1) = \mathbf{a}(k) + \mathbf{w}(k) \quad (9)$$

where $\mathbf{a}(k) = (a_1, a_2, \dots, a_p)^T$ and $\mathbf{w}(k)$ is a white stationary sequence.

The observation model is:

$$x_n = \sum_{i=1}^p a_i x_{n-i} + b_0 u_n \quad (10)$$

and let us consider $\mathbf{C}(k-1) = [x_{k-1}, x_{k-2}, \dots, x_{k-p}]$

then the observation model becomes:

$$z(k) = x_k = \mathbf{C}(k-1)\mathbf{a}(k) + v(k) \quad (11)$$

if it is taken that: $v(k) = b_0 u_k$

then, starting from an initial estimate $\hat{\mathbf{a}}(0)$, we obtain the following recursion:

$$\hat{\mathbf{a}}(k+1) = \hat{\mathbf{a}}(k) + \mathbf{K}(k+1)[x_{k+1} - \mathbf{C}(k)\hat{\mathbf{a}}(k)] \quad (12)$$

where \mathbf{K} is the Kalman gain given by:

$$\mathbf{K}(k+1) = \frac{\mathbf{V}_{\hat{\mathbf{a}}}(k)\mathbf{C}^T(k)}{b_0^2 + \mathbf{C}(k)\mathbf{V}_{\hat{\mathbf{a}}}(k)\mathbf{C}^T(k)} \quad (13)$$

and the matrix $\mathbf{V}_{\hat{\mathbf{a}}}$ is the variance matrix of the estimator $\hat{\mathbf{a}}$ and is given by the following equation:

$$\mathbf{V}_{\hat{\mathbf{a}}}(k+1) = [\mathbf{I} - \mathbf{K}(k+1)\mathbf{C}(k)]\mathbf{V}_{\hat{\mathbf{a}}}(k) + \mathbf{V}_{\mathbf{w}} \quad (14)$$

where $\mathbf{V}_{\mathbf{w}}$ is the variance matrix of the white noise process $\mathbf{w}(k)$.

The algorithm can be initialized by: $\mathbf{V}_{\mathbf{w}} = \sigma^2 \mathbf{I}$, $\mathbf{V}_{\hat{\mathbf{a}}}(0) = \mathbf{0}$ and $b_0 = 1$ and stopped by using the criterion:

$$\|\hat{\mathbf{a}}(k+1) - \hat{\mathbf{a}}(k)\|^2 \leq \varepsilon \quad (15)$$

The stopping criterion can also be used for pitch detection because it is evident that the above norm will be large while being in a clipped part, since the autoregressive model will not be valid.

3.2 Interpolation of the missing samples

For the computation of the missing samples, equation (4) can be used starting from p previous samples. This interpolation is referred to as forward. The missing samples can also be predicted from p samples that follow the missing part. The first sample can be obtained by solving equation (4) as:

$$x_{n-p} = \sum_{i=1}^p \alpha_i x_{n-p+i} ; p+1 \leq n \leq N_T \quad (16)$$

where the coefficients α_i are computed from the coefficients a_i using:

$$\alpha_1 = -\frac{a_{p-1}}{a_p} ; \dots ; \alpha_{p-1} = -\frac{a_1}{a_p} ; \alpha_p = \frac{1}{a_p} \quad (17)$$

Consequently, the reconstruction is done using backward interpolation.

4 Results

In order to test the previously defined algorithms, we are going to use synthetic and natural speech. The natural speech comes from a very large database of speech samples that were collected for the construction of a speech recognition system in colloquial Algerian Arabic [9]. The pitch frequency is about 100 Hz for male speaker and about 220 Hz for a female one. This corresponds to a pitch period T being between 4.5 ms to 10 ms. So, if a reliable estimation of the prediction parameters is desired, we need a fairly high sampling frequency. For example, a sampling frequency of 10 kHz (sampling period of 100 μ s) will provide between 45 and 100 samples for a pitch period. A sampling frequency of 44.1 kHz (sampling period of 22.73 μ s) is chosen, which provides between 198 and 440 samples for a pitch period, which is quite reasonable. Also, in all of the following tests, the speech signal is normalized to a maximum value of one.

4.1 Synthetic speech

First the algorithm is tested with a synthetic vowel. The choice of synthetic speech is motivated by the fact that it follows exactly the linear prediction model. The vowel /a/ is generated using the following formants [5]:

- The frequencies (F_i) and the bandwidths (BW_i) necessary to specify each formant are shown in the following table.

Formant	F_i (Hz)	BW_i (Hz)
1	730	60
2	1090	100
3	2440	120
4	3500	175
5	4500	281

Table 1: Formant Frequencies and Bandwidths [5]

- The pitch frequency is 120 Hz (male speaker), which corresponds to $N_T = 367$ samples. Figure 2 shows few periods of the synthetic vowel /a/. The prediction order is set to $p = 10$.

This signal is clipped to a level of ± 0.5 and restored using both methods (least square and Kalman filter method).

Figure 3 shows one pitch period of the clipped signal. A window of at least 75 samples following the clipped region is used to compute the predictor coefficients.

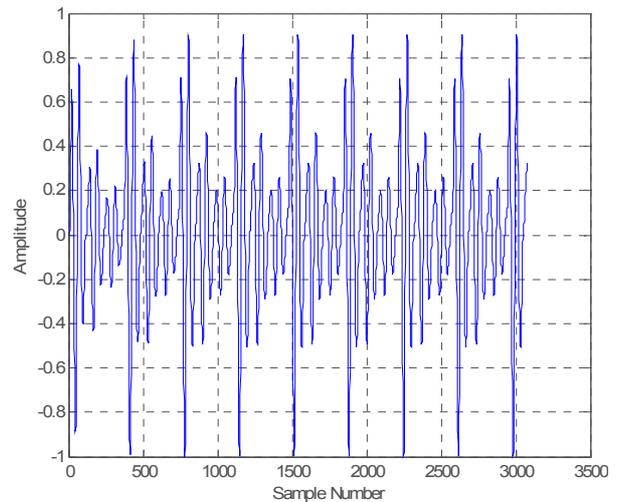


Figure 2 : Synthetic Vowel /a/ Normalized Amplitude Waveform

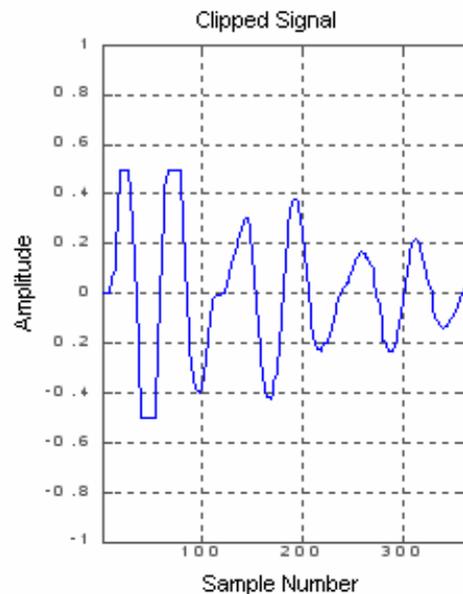


Figure 3 : Clipped artificial vowel

The least square computation of the prediction coefficients along with both forward and backward reconstruction produces the error plots shown in figure 4. It can be seen that the backward error is much smaller than the forward one. Also, the error occurs at the end of the reconstruction. The error can be reduced by performing both reconstructions and averaging the results. However, since the error is essentially a high frequency signal, simple low pass filtering after backward reconstruction yields the same result.

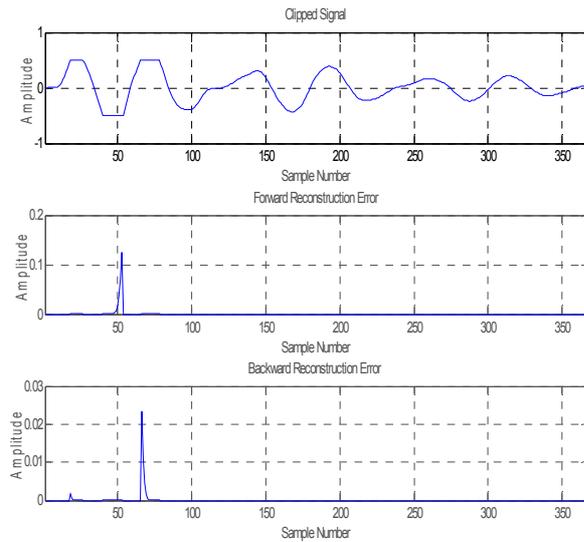


Figure 4 : Forward and Backward Reconstruction Error for Synthetic Speech

Kalman filter is also used for the estimation of the prediction parameters. The stopping criterion of the recursive Kalman algorithm is defined as: $\|a(k+1) - a(k)\|^2 < \varepsilon$, where ε is a small positive number that describes the convergence of the algorithm. From the plot of $\|a(k+1) - a(k)\|^2$ over 04 pitch periods of the signal (under: $\sigma = 0.1$ and $b_0 = 1$) shown in figure 5, it appears that the value $\varepsilon = 0.00008$ is acceptable.

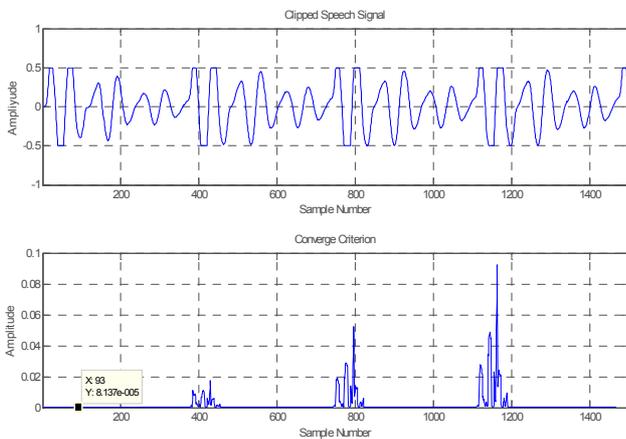


Figure 5 : Kalman Filter Convergence Criterion Plot

As stated before, the above convergence criterion ($\|a(k+1) - a(k)\|^2$) can be used for pitch detection. This is well illustrated in figure 5, the large values occurs at the clipped parts generally located at the beginning of the pitch period. After several tests, the following initial values: $\sigma = 0.1$, $b_0 = 1$, $\hat{a}(0) = \mathbf{0}$ have been selected. After estimation of the prediction parameters and backward prediction, the error is drawn in figure 6.

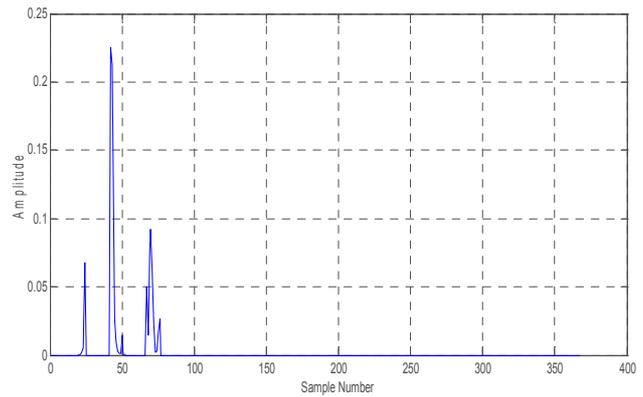


Figure 6 : Kalman Filter Error signal waveform

We observe that the error using the Kalman filter estimation is much larger that the one using the least square method. Another problem that is encountered is the large computation time. So, in the following tests, the results obtained by the least square method are the only ones that will be presented.

4.2 Artificially clipped natural speech

After being applied on a synthetic speech, the proposed technique of interpolation (least square evaluation of the parameters and backward reconstruction) is applied on a voluntarily clipped natural speech. The unclipped signal is taken as a reference when evaluating the reconstruction process precision.

The used recorded speech signal consists on numbers pronounced in Algerian Arabic, sampled at 16 KHz, taken from the database [9]. An audio processing software (Cool Edit Pro 2.1 from Syntrillium Software Corporation) is used to adjust the sampling frequency to 44.1 KHz.

Since speech is a time varying signal (a concatenation of different sounds with different characteristics) and in order to have a good estimation of the prediction parameters, the following method based on the detection of clipped samples is used: after each detection of a clipped sample, an adjacent segment of enough number of successive unclipped samples (ex.: in our case 75 samples) is considered. If this condition is satisfied the reconstruction process that uses the least square algorithm for the estimation of the prediction parameters will be applied. Otherwise, the procedure is repeated. Figure 7 shows the different steps of signal processing. The original speech and the reconstructed one are practically identical. Figure 8 shows the reconstruction error for the natural speech where it can be observed that the error is a high frequency signal with a small peak magnitude. So, a simple low pass filter will eliminate completely the error.

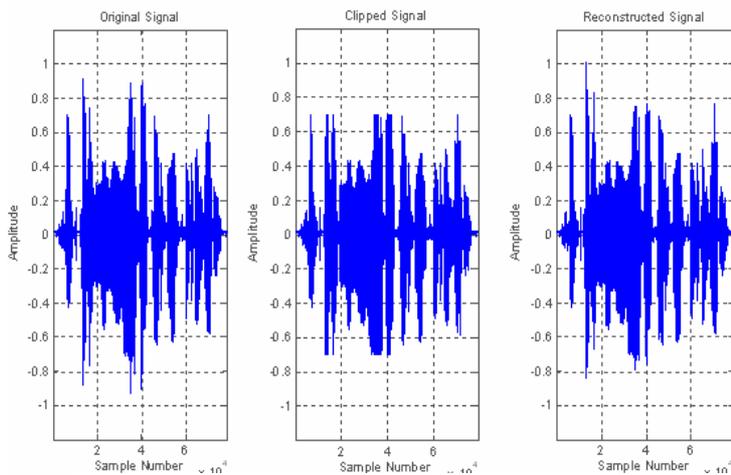


Figure 7 : Artificially Clipped Natural Speech Reconstruction

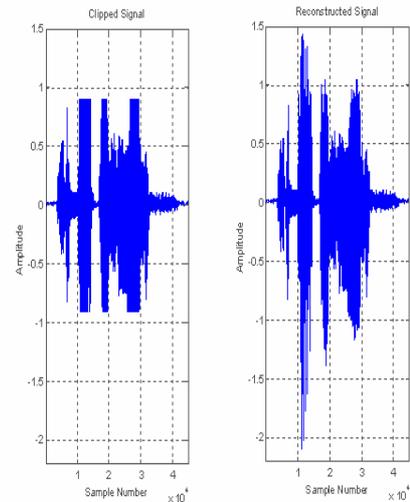


Figure 9 : Clipped Natural Speech Backward Reconstruction using Least Square Estimation

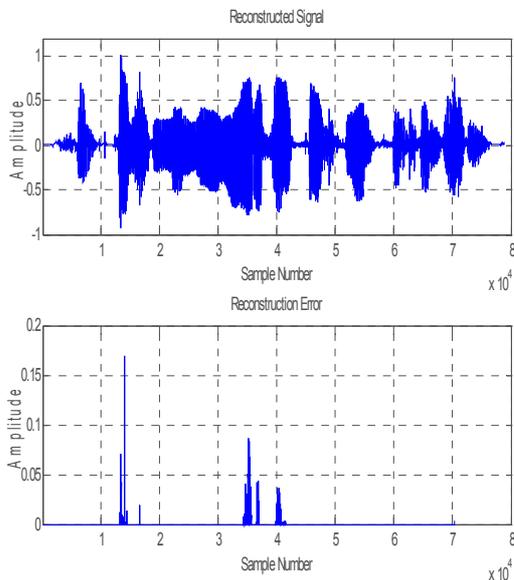


Figure 8 : Artificially Clipped Natural Speech Reconstruction Error

4.3 Clipped natural speech

The final test is performed on clipped natural speech. Figures 9 and 10 show the clipped and the reconstructed signal. It is impossible in this case to present an error plot due to the absence of the original unclipped signal. The only comment that can be made about the above plots is that the reconstructed signal looks like an unclipped signal. Since there is no reference to objectively evaluate the performance of the algorithm, a subjective criterion is used for judging the quality of the restoration. The speech samples (clipped and restored) were presented to several listeners and they were asked to evaluate the intelligibility of the message by giving a grade between zero and five (zero meaning completely unintelligible and five meaning very clear). The result is a great improvement in intelligibility. The clipped signal was given an average grade of about two while the restored signal received a grade that varied between four and five.

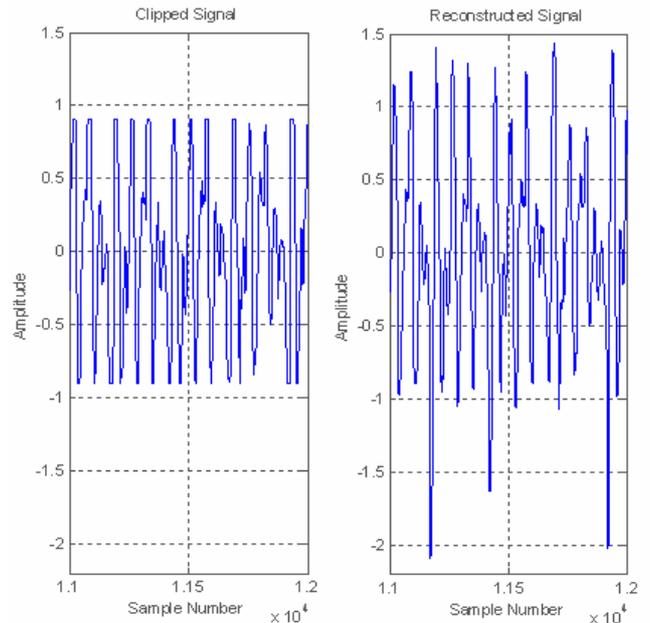


Figure 10 : Zoomed segment of the reconstruction process.

4.4 Discussion

It is quite hard to provide a figure of merit for the method other than the plot of the error signal between the unclipped and the reconstructed speech signal. We can observe from the plots in figure 4 and 8 that two parameters determine the quality of the reconstruction: the amplitude and the duration of the error spikes. We can resume both parameters in the following quality factor:

$$\Gamma = \frac{1}{N_c} \sum_{k=1}^{N_c} |e(k)| \tag{18}$$

where $e(k)$ is the error signal between the unclipped and the reconstructed speech signal and N_c is the number

of clipped samples. Figure 11 shows the quality factor for forward and backward reconstruction as a function of the number of clipped samples per pitch period for synthetic speech. For clipped natural speech, it of course impossible to provide such data. For artificially clipped natural speech, the quality factor curve using backward reconstruction is so close to the one for forward reconstruction for synthetic speech that it is impossible to provide a separate plot. From the different plots (figure 4, 8 and 11), we can conclude that the estimation of parameters using least square followed by a backward reconstruction offers the best results in term of accuracy.

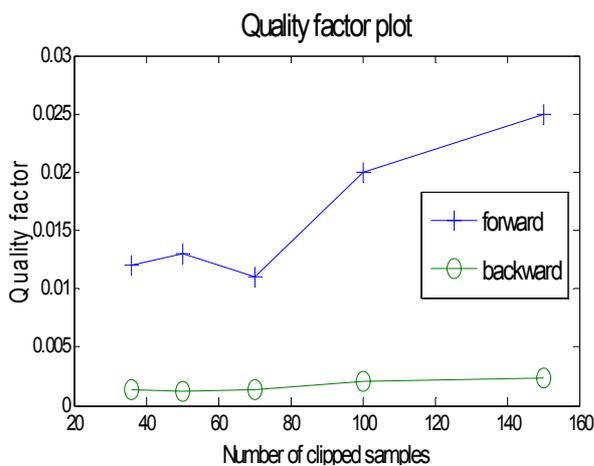


Figure 11 : Plot of quality factor vs. number of clipped samples for synthetic speech.

5 Conclusion

In this paper, an algorithm for clipped speech restoration using linear prediction has been presented and tested. It is able to restore completely the clipped speech. Two different methods for estimating the prediction parameters have been tested. The first one consists on block least square estimation while the second one is a recursive method. It appears that the recursive method is

pitch synchronous but quite inefficient while the block least square is very efficient and very precise. The block least square method followed by backward prediction has been implemented as part of a larger program for speech pre-processing in view of recognition and the results are a great improvement in the recognition rate.

6 References

- [1] Vaseghi, S.V., "Advanced Signal Processing and Digital Noise Reduction", *John Wiley and Teubner* 1996.
- [2] Rabiner, L. R., Shafer, R. W., "Digital Processing of Speech Signals", Prentice Hall, New Jersey, 1978.
- [3] Brakta, N. and Hadibi, M., "Pattern Recognition Techniques Applied to Speech Recognition", Final Year Project, INELEC, 1999.
- [4] Makhoul, J., "Linear Prediction, A Tutorial Review", *Proceeding of the IEEE*, (63), pp. 561-580, April 1976.
- [5] Chandra, S. and Lin, W. C., "Experimental Comparison between Stationary and Non Stationary Formulations of Linear Prediction Applied to Voiced Speech Analysis", *IEEE trans. of ASSP*, (22)6, Dec. 1974.
- [6] Lawson, C.L., and Hanson R.J., "Solving Least Squares Problems", Prentice Hall, 1974.
- [7] Srinath, M.D. and Rajasekaran, P., K., "An Introduction to Statistical Signal Processing with Applications", John Wiley & Sons, 1979.
- [8] Gueguen, C., J., and Carayannis, G., "Analyse de la Parole par Filtrage Optimal de Kalman", *Automatisme*, Tome XVIII, March 1973, pp.99-105.
- [9] Reggab, M., "Continuous Speech Recognition Using Hidden Markov Models, Application to Colloquial Algerian Arabic", Unpublished Magister Thesis, DGEE, FSI, Université M'Hamed Bouguerra, Boumerdes, Algeria, June 2004.

eGovernance: Information and Communication Technology, Knowledge Management and Learning Organisation Culture

Jaro Berce
 Faculty of Social Sciences, University of Ljubljana
 1000 Ljubljana, Slovenia
 phone: +386 1 580 5362, fax: +386 1 580 5101
 E-mail: jaro.berce@fdv.uni-lj.si

Sam Lanfranco
 York University, Toronto, Canada
 phone: +1 416 816 2852 fax: +1 416 534 5268
 E-mail: Lanfran@Yorku.ca

Vasja Vehovar
 Faculty of Social Sciences, University of Ljubljana
 1000 Ljubljana, Slovenia
 phone: +386 1 580 5297, fax: +386 1 580 5101
 E-mail: vasja.vehovar@fdv.uni-lj.si

Keywords: information and communication technology, knowledge management, learning organization culture, eGovernance, eGovernment, iGovernment

Received: May 11, 2007

Changes in information and communication technology (ICT) require continuous modifications in contemporary organizations and in corresponding work processes. The integration of new technologies is highly correlated with the emerging principles of knowledge management (KM) and learning organization (LO) culture. However, the interactions of these two concepts with ICT are rarely studied. This paper addresses this complex issue by first conceptualizing the notions of ICT, knowledge management and the learning organization. It analyses their interrelation and potential impact on eGovernance. Finally, the conceptual model was estimated by normal theory maximum likelihood using the LISREL 8.51 program [28]. It was estimated on data from a survey of Slovenian public administration organizations. The main finding is that – beyond the impact of ICT – the application of knowledge management (KM) principles considerably supports and stimulates eGovernance.

Povzetek: Prikazana raziskovalna naloga vsebuje analizo vplivov "tehnologij informacijske družbe" na razvoj organizacij v javni upravi.

1 Introduction

Notwithstanding the views of Nicholas Carr [39], information and communication technology (ICT) continues to be a powerful force in transforming the ways in which societies and economies operate. The widening use of ICT is having profound impacts on patterns of living, communicating and work. At the same time divisions between an ICT perspective and other disciplinary perspectives (e.g. economics, politics, sociology) are becoming less firm. The impacts of ICT are more than just those of efficiency and effectiveness. The impact of ICT is moving deeper and deeper into substantial economic and social science domains, where it increasingly interacts with them [40]. To understand modern transforming public administration one could not avoid studying the emerging role of ICT and its driving role to improve public administration.

All too often, when the role of ICT is conceptualized, the emphasis is frequently only on its technological aspects and omitting other impacts (e.g. social) [41]. Of course, within public administration, the flow of information society services and applications cannot be understood if detached from an ICT technical prospective. Previous "paper oriented" work activities have now been replaced by modern ICT supported processes, coupled with new organizational principles, where activities and services are provided through flexible and complex multi-organizational structures [3]. ICT advances (networks, computing power, applications, etc.) have enabled disbursed information to be assembled into accessible complex organizational database knowledge repositories. ICT brings new quality to services as it re-organizes knowledge and effectively incorporates intelligent access to and use of the data [18]. The researchers working in the areas of

economic and social policy and public administration increasingly need to incorporate knowledge about the role of ICT in the processes studied.

As previously stated, ICT related technological aspects are only one component of what needs to be understood for proper understanding and effective implementation of ICT. Other aspects that need to be understood relate to the broad array of supporting activities within an organization, and within that organization's societal environment. Within this context, the essential components are (1) knowledge management (KM) and its emerging principles; and (2) learning organization behavior (LO).

Individually, KM and LO, as well as ICT, are already well explored. However, they are rarely studied simultaneously and in an interactive fashion. This paper addresses that challenge: how to conceptualize the interaction of ICT, KM, and LO, to understand their impact on the specific outcome of organizational behaviour. In this study we focus our attention on the governance principles and practices that arise from ICT usage in the public sector, i.e. we focus on eGovernance. The main objective of this study is to build a model of the relations between ICT, KM, and LO, and use that model to test some relationships against Slovenian public service data.

The paper starts with an overview of the basic concepts. Next, a theoretical model is constructed for ICT, KM, LO and their impact on eGovernance. The resulting causal model is applied to survey data on eGovernance practices in Slovenian public administration organizations. The paper concludes with a summary of the results, together with the suggestions for future research.

2 Basic concepts

We first review the terminology and concepts to be used in the work. Technically speaking, **information and communication technology** (ICT) is a mixture of *hardware* (equipment), *software* (operating system, applications, etc.), and *communications facilities* (local area networks, wide area and backbone networks, communication protocols, etc.). Work on the expanding role of ICT on business and administrative processes has been underway since early 1970's, when the process of massive computerization began. However, the corresponding social and economic dimensions were considered only with some delay [24]. The introduction of personal/desktop computers in the 1980's, linked through the Internet in 1990's, further increased the role of ICT, as well as the importance of the accompanying social contexts. This was particularly important for the supportive and innovative aspects of governance in organizations [8]. ICT advances have become a permanent force bringing continuous and sometimes unpredictable changes to organizational structures and processes,

including services delivered, management practices and governance.

Knowledge management (KM) is a relatively new concept, whose emergence is only indirectly linked to the above described ICT developments. It arose within a well-known, and well researched, "Initiative for Managing Knowledge Assets", undertaken in 1989 among a consortium of US companies working on a platform for managing knowledge [42]. Since then, the idea expanded in many directions, from the local and regional level as "Communities of Practice", to within the private sector as a "Corporate Knowledge management", or "Knowledge and Innovation management". Failed KM is even identified as one of the factors playing a major role in the February 2003 Columbia space shuttle disaster.¹

The literature increasingly focuses on how organizations and institutions implement ways of accumulating employees' knowledge in electronic databases so as to use them as repositories of the shared, company/institutional wide "structural capital" [36]. Current conceptions and approaches to KM focus on handling *explicit* knowledge that is "transmittable in formal, systematical language" and can be stored in specifications, reference manuals and institutional handbooks [15]. Knowledge resides in the users and not in the collection. It is how users react to a collection of information that matters [3]. It usually falls short of using knowledge management to create a climate and culture of known practices and assumptions, which are stored in an employee's brain. This can constrain an organization to the limits of the Drucker Theorem of "doing more of the same better and better, however, with diminishing marginal returns" [42].

KM scholarly papers and books began to appear in early 1990's, e.g. Senge's *The Fifth Discipline* [15] and Sakaiya's *The Knowledge Value Revolution* [36]. The first KM textbooks appeared only by the end of the 1990's. The KM concept continues to receive increased, illustrated by a quote from Larry Prusak, the executive director of the IBM Institute for Knowledge-Based Organizations (IKO)²:

"In the emerging economy, a firm's only advantage is its ability to leverage and utilize its knowledge".

Modern ICTs have another consequence in that they can centralize or decentralize teaching and learning. Either way, this can radically broaden the access to learning. This is closely linked to another focus from 1990's, that of the **learning organization** (LO). LO, LO behaviour, and LO culture, relate to the

¹See ComputerWeekly.com:
<http://www.computerweekly.com/Articles/2003/08/27/196769/ReportKnowledgemanagementfailurcesentraltoShuttleDisaster.htm> (August 2003)

² See: <http://www.providersedge.com/kma/> (October 2005)

notion that organizations, and the people within them, continually expand, and access, their collective knowledge and skills in the pursuit of desired outcomes, or organizational deliverables. Within a learning organization culture, new and expansive patterns of thinking are nurtured, collective aspiration is “set free”, and people are encouraged to see challenges and opportunities within “the big picture” [15]. In successful LOs, individual learning is continuous, knowledge is shared, and the organizational culture supports learning. Employees are encouraged to think critically and take risks with new ideas: all employees' contributions are valued [22] [23]. Pedler, Burgoyne, and Boydell define the LO as an organization that facilitates the learning of all its members, in support of the continuous transformation of itself in pursuit of the organizational mission and vision [11].

We now turn to apply these ideas to governance and public administration. Within this work we understand **government** as the system and the organizational process by which a given community is governed. Governments produce huge volumes of information and documents. ICT enabled information networks (e.g. LANs, Intranets, Internet) increasingly remove the boundaries separating internal parts of the government from each other (improving efficiency) and from users (improving effectiveness) [7].

The concept of **eGovernment** is commonly understood as governmental procedures and tasks supported by (ICT enabled) digital means: "eGovernment is the use of information and communication technologies to improve the activities of public sector organizations"³. For purposes of analysis, we propose to decompose government procedures and tasks in two components [1]:

- **iGovernment:** converting existing information processes and paper objects into digital form. This first step focuses on the internal digitalization of documents, and contributes to the efficiency and effectiveness of public administration.
- **eGovernment:** upgrading and building on previous step of iGovernment, it relates to Internet-based digital services offered by the government administration to its non-government clients (i.e. citizens and businesses). eGovernment is converting literal services into virtual services.

Public **governance** is also recognized as the strategic exercise of economical, political and administrative authority to manage the nation's affairs at all levels. Ewalt sees governance as blurring of boundaries and responsibilities for tackling social and

economic issues [3]. Peters and Pierre suggest that governance is about a political theory, and that Public Management is a form of organizational theory [13].

The OECD characterizes effective public governance as helpful for strengthen democracy and human rights, helpful to promote economic prosperity and social cohesion, to reduce poverty, to raise capacity to learn, to enhance environmental protection and deepen confidence in public administration [10]. The International Teledemocracy Centre proposes that eGovernance should support ocommunication between government and civil society [17]. Papers by Clift [2] and Vikas [8] argue that the concept of **eGovernance** relates to strategies of government where ICTs have a substantial role.

For our work we start with the all encompassing OECD role for governance and explore how ICTs enable and deepen KM and LO behaviour. There are many concepts of eGovernance (Malkia, Anttiroiko and Savolainen [19], Reinermann and Lucke [20] etc.). Here, in narrow terms, we will basically understand eGovernance as the use of ICT to improve the effectiveness and the efficiency of all phases of the public institution processes. eGovernance is also closely related to **eDemocracy**, which combines the ICT with increased levels of democratic incentives [2], as well as with promotion of **knowledge society**, characterized by new measures of competitiveness, such as knowledge generation, increased research and development, the availability of knowledge to citizens [8] and their enhanced capacity to learn [6].

The concepts of ICT, KM, LO and eGovernance are relatively new and, within limits, their definitions are constantly changing. It is thus not surprising that measurement and evaluation face serious problems with regard to corresponding definitions.

Within ICT measurement and evaluation are already an inherent and almost endemic problem, spanning early discussions around the “productivity paradox” to contemporary discussions dealing with the return on ICT investments (ROI) and problems with ICT evaluations [27]. The role of ICT is also extensively analyzed in the studies of national economies, productivity growth and the components of product added-value [25].

Similar definition and measurement problems accompany the concept of KM [14]. KM is related to variables and attributes that are hard to standardize and evaluate. These include contents, relations, processes, procedures, infrastructures, networks, institutions, modus operandi, linkages, capacity to learn and evolutionary processes [32]. Though there exists a huge community focusing exactly on measuring challenges e.g.: the Intellectual Capital community with several publications as: Journal of Intellectual Capital, World Congress on Intellectual Capital, European Congress on Intellectual Capital etc.

One can generally distinguish between two different processes of organizational change that are associated with LO [33]:

³ Source: eGovernment for Development, *Basic Definitions Pag*, Richard Heeks, IDPM, University of Manchester, UK, 2004 - <http://www.egov4dev.org/egovdefn.htm>, June 2006

- **Adaptive learning:** internal changes that have been made in reaction to changed external environmental conditions.
- **Proactive learning:** organizational changes that have been made on a more wilful basis. This is learning into action which goes beyond simply reacting to external environmental changes.

A LO, and learning organization behaviour, have to promote information exchange between employees, and create a more knowledgeable workforce. A LO seeks a very flexible organizational structure where people will accept and adapt to new ideas and changes through a shared vision⁴. This brings a new perspective and growing importance to organizational knowledge, which can be responded by the "learning organization" as the challenge of creating a culture of managing knowledge [15].

Several studies have developed methods for assessing knowledge through the LO environment. The include Krebs Valdis: *Knowledge Networks - Mapping and Measuring Knowledge Creation and Re-Use*, 1998⁵; Lethbridge Timothy Christian: *Practical Techniques for Organizing and Measuring Knowledge*, [doctoral thesis] November 1994⁶; and the Gurteen Knowledge Conference 2003⁷. Within this context, Perkmann suggests the measurement of the value of knowledge from two perspectives: the macro view (which quantifies the intangible assets of an organization) and the micro view (where the impacts of individual knowledge projects can be assessed and quantified) [12]. Riley frames KM to include the active creation, transfer, application and re-use of (tacit) individual knowledge, as well as codified (explicit) collective knowledge, supported by new approaches, relationships and technologies [21]. Both aspects are used to increase the speed of innovation, decision-making and responsiveness to organizational objectives and priorities, and can be a basis for the effective implementation of LO culture. Recent literature reminds us that if knowledge is explicit then it can (perhaps) be managed, measured, codified, etc. But, for tacit knowledge, it "can only be learned", but it must shared, fostered etc., to persist and survive

⁴ See: http://www.see.ed.ac.uk/~gerard/MENG/MEAB/learning_organisation/definition.html (August 2006).

⁵ See: <http://www.orgnet.com/IHRIM.html>, Sept. 2003

⁶ See There are many concepts of eGovernance (Malkia, Anttiroiko and Savolainen [17]; Reinermann and Lucke [18]), however we view it as the use of ICT to improve the effectiveness and the efficiency of all phases of public institution processes and its outputs: http://www.csi.uottawa.ca/~tcl/thesis_html/thesis.html, Sept. 2003

⁷ See: <http://www.gurteen.com/gurteen/gurteen.nsf/0/1878458C9E62DB7E80256C9E0062814F/>, Sept. 2003

within the organization. Noordegraff takes the utilitarian view that managing knowledge is "the configuration and control of operational knowledge processes in such a way as to promote the yield and pleasure of knowledge as a factor of production" [9].

New definitions of KM are repeatedly proffered, trying to capture the essence of the concept. See: Drucker: *Management Challenges for 21st Century* [34]; Sveiby: *Managing and Measuring Knowledge-Based Assets* [35]; etc. Each complements the understanding of the KM concept, but observed from different perspectives.

3 Model and hypothesis

Our intent is to construct a model that can assist in analysing here the relationships between ICT, KM, LO, and diverse social, economic and technological factors. More specifically, we examine the relationships between ICT, KM and LO culture within the context of their influence on governance (eGovernance) in public administration. Our ultimate aim is to identify and assess the factors that contribute to the improved eGovernance (Figure 1).

Within this context, we primarily investigate a theoretical model to understand KM, ICT and LO culture can work together to organize government in pursuit of improving **iGovernment** (organized digital and on-line information), **eGovernment** (on-line services) and finally to achieve good **eGovernance** (promote inclusion, democracy, etc.).

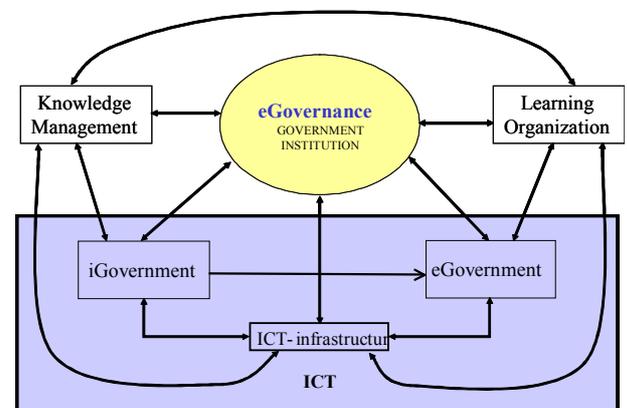


Figure 1: Conceptual Model for Research

Our key hypotheses are as follows:

- H1:** ICT, LO and KM all have strong (and measurable) impacts on **eGovernance**,
- H2:** While ICT impacts LO and KM, we also expect a causal link from KM to LO, because LO culture is stimulated by the KM practices [15].
- H2:** The impact of **eGovernment** on **eGovernance** is high, while **iGovernment** has only indirect impact on **eGovernance** via **eGovernment**.

3.1 Operationalizing the Model

Each of the key concepts discussed above needs to be further operationalized and elaborated upon, so that the model can be empirically tested with survey question data. Here we develop these concepts into measurable components. The labels in the brackets denote the actual name of the corresponding compound variable in table 2 (appendix B). There, the variables are arranged hierarchically according to the final causal model structure (Figure 2), which expands the initial conceptual model from Figure 1.

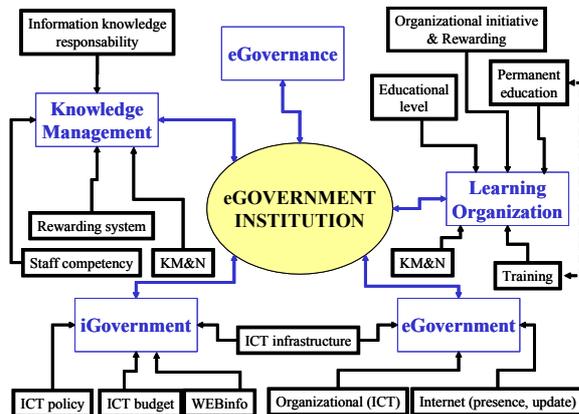


Figure 2: Elaborated relations of Conceptual model for Research

1. We operationalize ICT in the following components:
 - **ICT Infrastructure ("infrastr")**: hardware (computers, ranging from PCs, laptops and personal digital assistance to servers), communication media (networks, cables, routers, switches, etc.) and software (programs, applications, tools). In Table 2 this variable includes 14 elements, ranging from percentage of employees using computers (see: variable "v2_1" from Table 1) to the usage of various ICT services. We expect that this variable has an impact on eGovernment as well as on iGovernment, thus it appears twice on the corresponding left side in Table 2. For this reason it is also shadowed.
 - **eServices ("webinfo")**: intra and inter organization interactions including business to business (B2B), business to government (B2G), internal (G2G), as well as to business to customer (B2C), and government to customer (G2C).
 - **Internet ("webpres")**: communication tools (e-mail, browsers, search engines, blogs, text messaging, etc.). Also includes information, archiving, presentation, marketing, multi-media deployment (e.g. peer2peer, Skype), finance transfer vehicles (e.g. eBanking, eStock-Exchange) and much more.
 - **ICT policy ("policy")**: explicit policy or tacit organizational culture defining working processes, organizational behaviour and the

- processes that produce "outcomes". It also defines how and why and where data, information, transactions, etc. are to be located within electronic (digital) system.
2. Within the model **organizational behavior** is made operational with the following components:
 - **Organizational initiative ("organiza")**: organizational *changes* related to events and environmental factors including technological changes, as well as changes in organizational structure and processes.
 - **Educational level ("educat")** refers to different knowledge and skills, including the array of professional qualification,
 - **Permanent education ("trainb")**: Once a person leaves the formal education system, for a person within a Learning Organization, a life-long learning process becomes necessary and ICT is one specific vehicle for obtaining new knowledge, training and skills (eLearning).
 - **The reward system ("reward")** relates to monetary incentives (salary, options, etc.) for both efficient and innovation work, and it also includes other types of rewards (job satisfaction, prestige, promotion etc.).
3. **Knowledge management and networking** has the following operational elements:
 - **eCapture of staff competency ("eCapture")**: organizational knowledge may be tacit and reside within an employee's head or it may be explicit and clearly documented. Similarly, skills accumulate as a result of learning by doing, if properly identified and motivated, can be effectively accessed and brought into use for the mission of the organization.
 - **Information and knowledge networking ("network")**: A networking "culture" needs to be present and voiced in support of KM. The management must first act as a catalyst or enabler by setting examples, creating trust and inspiring a cohesive and creative knowledge network. A clear communication of vision and scope for an information and knowledge network must be present within an organizational structure. Here, the top management support is crucial for the promotion of effective information and knowledge networking. As seen in Table 2 and in Figure 3 this component contributes to both, the KM and LO outcomes.
 - **Knowledge and information responsibility ("info")**: KM is one of the most important tasks, particularly with relation to "organizational knowledge", "personal knowledge" and employee skills. The information sharing also needs to be based on a system of rewards. Having someone directly responsible for managing these tasks

- gives an organization the ability to carry out effective KM practices.
4. We also used two financial concepts related to **budgeting ("budget1", "budget2")** that measure two different resources devoted to the above listed organizational attributes. With this the "devoted commitment" of top management to could be effectively measured.
 5. The following previously described concepts serve in the model as intervening variables (presented also in second column of Table 2) for impacting on the target notion of **eGovernance**:
 - **iGovernment ("iGovern")**: An activity of converting existing processes and paper objects to digital forms. ICT infrastructure, ICT Policy and ICT Budget are the main building blocks that impact the level of digitalization of the processes.
 - **eGovernment ("eGovrn")** refers to converting existing literal services to virtual services as well as the initiation of new services and new mechanisms.
 - **Knowledge Management ("KMan")**: organizing and optimizing the knowledge embedded within, and used for, services, policies and procedures. The KM literature often refers to a special KM organizational unit, distinct from human resource or ICT unit. The presence of appropriate enablers and rewards is a good indicator of whether or not an organization has properly operationalized the unique properties and circumstances associated with its KM strategy. In this research we measure KM with direct questions about organizational KM practice.
 - **Learning organization ("lear_org")**: behaviour (or culture) that enriches and makes organizational knowledge more effective. It also means that workers need to be formally engaged in corresponding learning processes. Budgetary support for training and other knowledge promotion activities emerge as key elements for creating and sustaining a LO culture.

The ultimate concern in this study is how **eGovernance ("eGovance")**, relates to the use of ICTs, KM and LO to promote the broad objectives of a contemporary democratic government. Mechanisms include digital inclusion, e-participation, computer literacy, and the like. A properly deployed ICT strategy ensures that every local or rural community has access to information and services available on the digital venue. The applications of ICT that support and facilitate government "front-office" services can be divided into three categories: *access to information*, *transaction services* and *citizen participation*. The ICTs, is reducing barriers to access to information, services and communication with institutions and officials [24]. In addition to those front-office

functions the back-office serves two vital and inseparable objectives: to *enhance the performance* and *ensure the conformity* of government.

However, the unique challenge of eGovernance is that, unlike iGovernment and eGovernment, which are mostly technology related and have parallels in the private and business sectors, eGovernance is uniquely linked to the responsibilities of government in a democratic society. This is an extremely important issue and the challenge is for it to lead - rather than borrow from - the private business and civil society institutions with regard to issues such as access, accountability, and participation.

As we see in next section, and in table 2, this concept is measured with an array of elements, from explicit e-mail policy regulation to various levels of digital services. There, based on expert judgments, more advanced levels received increased weights in this compound variable. For example, having on-line payment systems for government services has the highest weight (seven).

The above described concepts are observed graphically in Figure 2, which further elaborates the simple conceptual structure in Figure 1.

3.2 The empirical model

3.2.1 The data collection

The above conceptual scheme was designed and tested on the data obtained from a survey among Slovenian public organizations including: Ministries with constituent office(s), Government offices, Local government offices and Municipality offices. The remaining public administration bodies, e.g. Constitutional Court, Parliament, Ombudsman, Courts of justice etc., were not included, because they perform more narrow specific tasks, and are more heterogeneous.

Information from respondents was obtained by mail questionnaire, targeted to public institution managers and was based on the concepts developed in previous section.

	Included	Replied	Reply percentage
Ministries	14	14(+12*)	100%
Government offices	24	20	83.3%
Local government offices	58	53	91.4%
Municipal offices	192	44	22.9%
TOTAL	288	143	49.7%

Table 1: The response rates by categories

* The "stand-alone" constituents' offices of ministries (e.g. Police) replied separately

A total of 288 Slovenian public organizations were invited to participate in the survey in 2003. Help from two governmental offices was provided (Office for the Organization and Development of Slovenian Administration, and Office for Local Autonomy – through association of municipality secretary). After the initial mailing, which provided 58 responses, email and telephone follow-ups provided a total of 143 responses.

3.2.2 The questionnaire and the variables

The concepts described in previous section (illustrated in figure 1 and figure 2) were made fully operational with data from the explicit survey questions listed in appendix A. As well, some questions from OECD KM survey questionnaire in the public sector⁸ [38] were used. In addition to these OECD question items, further questions were constructed and tested to provide information for the concepts developed in this research.

Both descriptive and attitudinal variables were used. The descriptive variables measured the actual characteristics of the organization, such as ICT profiles, budget spending, number of employees, explicit KM and LO organizational features, etc. Variables of this type dominate in the first two modules, the *Background module* (1) and the *ICT infrastructure module* (2), of the questionnaire (see Appendix A). They also figure prominently in the remaining two modules of the questionnaire, *ICT services and business module* (3) and in *KM and LO module* (4), where in addition, two blocks of variables were also related to attitudes and other features (see: questions 3_56 and 4_01).

To support the measurement scale employed in the research, three types of the survey questions were used. The first type consists of nominal or dichotomous variables (YES=1/NO=2), related to specific phenomena such as whether an organization has its own Internet connection or not. The second group of variables were numeric and on the ratio scale. Typically, they measured the share (percentage) of budget, number of employees, etc. devoted to certain purposes. The last group of variables were attitudinal on 1 to 4 points (strongly disagree – strongly agree) ordinal Lykert scale.

The compound variables in the model were constructed as simple linear combinations of the above elementary variables from the questionnaire. This is described in Table 2 in appendix B. Most typically, these were simple sums, i.e. question items were equally important and given equal weight. The only exception was the concept of eGovernance

(**eGovance**), where weights were attached to the elements with increased importance.

The variables from the questionnaire (Appendix A) were transformed into the variables that correspond to the concepts from figure 2, using the some simple linear transformations described in table 2, appendix B. For some skewed variables (e.g. number of employees) a logarithmic transformation was used. We tested the normality assumptions of all these newly constructed variables in the model and the tests showed that the assumption is acceptable.

3.2.3 The model

The empirical model is built on the concepts described in previous sections, starting with the relations in figure 1 and figure 2. We start with an array of independent operational conceptual variables constructed from the questionnaires, and build variables for our model. Simple linear combinations are used to explain the key output (exogenous) variable (eGovernance). Four intervening (endogenous) variables are also used in the model: iGovernment, eGovernment, KM and LO.

The model was estimated by normal theory maximum likelihood using the LISREL 8.51 program [26]. As mentioned, all the variables in the model were approximately normally distributed. With the remaining discrepancies we refer to several studies indicating high robustness of these models with regard to the assumption of normality [29].

The obtained causal models use the standard diagnostics expressed with χ^2 statistics and corresponding p -values. We obtained here an acceptable value of $p = 0.13$. The p -value is typically requested to be above $p = 0.05$ and preferable closer to $p = 0.5$, because we are seeking the model which cannot be rejected, so that it's fit to our data is accepted. We obtained here an acceptable value of $p = 0.13$ which is within predefined boundaries.

Another key measure to evaluate these models is the **Root Mean Squared Error of Approximation (RMSEA)** value. According to Saurina and Coenders [30] the standardized χ^2 test of the hypothesis of perfect fit to the population covariance matrix (i.e. the p -value) should be given less importance than the measures of the degree of approximation between the model and the population covariance matrix. Values equal to 0.05 or lower are generally considered to be acceptable [31] and we obtained here the value 0.04. We can thus conclude that our final model fits the data according to key benchmarks of linear structural modelling.

The path diagram for our model is in figure 3 and it follows the conceptual model from figure 1 and figure 2. We assume that the causal order flows from basic characteristics of the organization to intervening (endogenous) variables – i.e. eGovernment (**eGovern**), iGovernment (**iGovern**), Knowledge

⁸ The results exclude open-ended questions in the OECD survey. That contains conclusions from the results of the survey of knowledge management practices for ministries/departments/agencies of central government in OECD member countries.

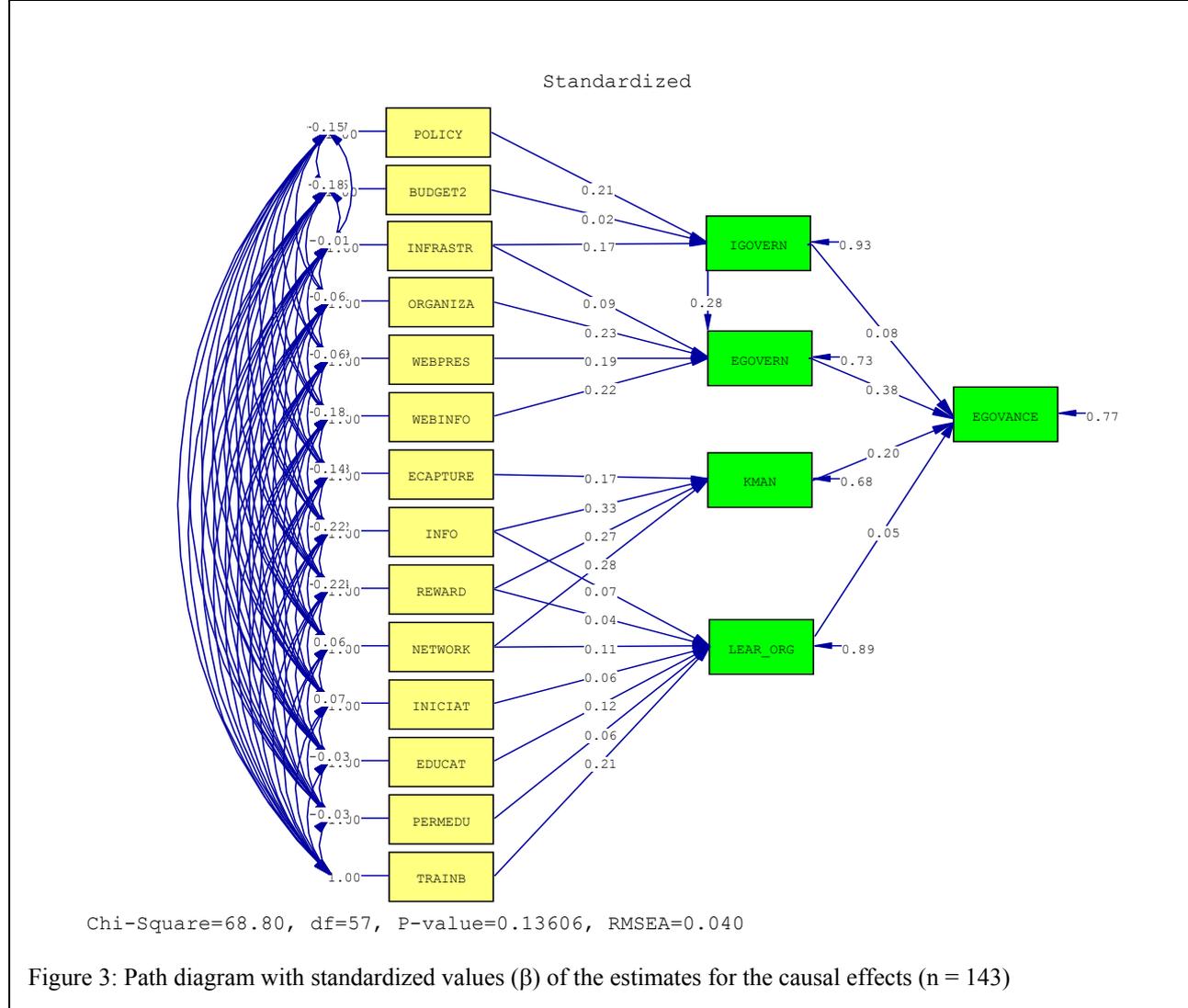


Figure 3: Path diagram with standardized values (β) of the estimates for the causal effects ($n = 143$)

Management (**KMan**), Learning Organization (**lear_org**) – and to eGovernance (**eGovernance**).

The path diagram in figure 3 includes the standardized regression coefficients labeled over the arrows. For example, the value $\beta = 0.20$ in the case of the arrow pointing from **KMan** to **eGovernance** means that the change in one standard deviation of the independent variable **KMan** causes a change/increase that can be expressed as 0.20 of the standard deviation in the dependant variable **eGovernance**. The corresponding t -values, which accompany these coefficients, are not presented in the Figure 2, however, we will explicitly discuss the ones that are larger than $t = 1.96$, as they denote a statistical significant causal relation among the two variables.

The path diagram also shows the amounts of unexplained variances (the short arrows pointing to the variables), which is the share of variability in the exogenous intervening variables that was not explained by the model. Higher values thus denote substantial levels of the variability that remained unexplained. However, the proportion of the unexplained variance for the key target variable (**eGovernance**) has the value of 0.77, what is acceptable. Even more, such a value is much above usual expectations for this type of models, where measures of attitude are involved. On the other hand, the proportions of the unexplained variance for some intervenient variables are much higher. In particular, we have value 0.93 for the iGovernance (**iGovern**),

which was otherwise strongly influenced by the policy regulations (**policy**) with the standardized regression coefficient of $\beta = 0.20$ ($t = 2.35$) and ICT infrastructure (**infrastr**) with $\beta = 0.17$ ($t = 1.82$). The other three intervening variables have much lower values for the share of unexplained variance, **KMan** (0.68), **learn_org** (0.89) and **eGovern** (0.73).

When the proportions of the unexplained variances are higher (particularly in case of LO and iGovernance) we conclude that their values are relatively independent of the measured characteristics. They may be influenced by some other variables, not included in the model, or they may be truly independent of the environment we measured.

We have also tested various other models, particularly the ones, which have the causal links between intervening variables. However, the only relation that was improving the model was the link from **iGovern** to **eGovern**, with a relatively high value of $\beta = 0.28$ ($t = 3.21$). This illustrates the facts that eGovernment is truly built on the foundations of iGovernment. On the other hand, the link between eGovernment and KM and LO did not improve the model, what was also particularly true for the link from LO to KM. A comparison of the primary model and the changed models validated primary one due to results that showed the changed models to perform worst. As well to test some important variables the Cronbach's alpha coefficient of reliability (or consistency) was used.

Let us briefly concentrate on coefficients with values larger than $\beta = 0.2$, which are in this model (given this specific given sample) also the ones that are roughly statistically significant:

- The path diagram (Figure 3) shows that explicit ICT policy (**policy** - $\beta = 0.21$, $t = 2.35$) has significant a effect on iGovernment,
- ICT organization (**organiza** - $\beta = 0.23$, $t = 2.56$) and WEB update of information (**webinfo** - $\beta = 0.22$, $t = 2.48$) are also seen to be significant for eGovernment,
- The responsibility for information and knowledge sharing (**info** - $\beta = 0.33$, $t = 4.15$), rewarding of sharing (**reward** - $\beta = 0.27$, $t = 3.17$), and knowledge networking (**network** - $\beta = 0.28$, $t = 3.43$) were identified as significant factors for good KM process within institutions,
- Training budget (**trainb** - $\beta = 0.21$, $t = 2.10$) standardized values suggest that it is an important factor when developing and sustaining a good LO culture.
- The Information and Knowledge Networking variable (**network**) has a strong impact on both KM (**network** - $\beta = 0.28$, $t = 3.43$) and LO culture (**network** - $\beta = 0.11$, $t = 1.06$) underlining the importance of knowledge networking to best utilize the existing stock, and new acquisition, of employees' knowledge.

With respect to the impact of the four intervening variables on the key target variable (**eGovance**) we can summarize the following:

- As expected, strongly articulated eGovernment characteristics (**eGovern**) also impact the target variable of eGovernance (**eGovance**). We can observe here the strongest causal link in the whole model, with the value $\beta = 0.38$ ($t = 3.93$).
- The impact of iGovernment (**iGovern**) seems to be much lower ($\beta = 0.08$, $t = 0.83$). This is also the consequence of the indirect link via eGovernment, where we already reported high coefficient ($\beta = 0.28$, $t = 3.21$). Another explanation may be the fact that basic activities of iGovernment do not provide advanced services, so we cannot expect strong impact on eGovernance.
- As expected, the KM has relatively high impact on eGovernance ($\beta = 0.2$, $t = 2.13$), and supports our main hypothesis. Nevertheless, it is weaker than the role of eGovernment.
- Relatively low link of LO (**learn_org**) factors ($\beta = 0.05$, $t = 0.57$) is somewhat surprising, as we expect it to be much higher. A close inspection of co-linearity diagnostics showed that there was no major co-linearity between independent input variables. In addition, we also expected an explicit link from LO to the KM component. However, this was shown not to be strong for the model.

With the relation to our initial hypothesis we can thus conclude the following:

- *H1: ICT, LO and KM have all significant impact on eGovernance.* This hypothesis was confirmed for KM and also for ICT, which demonstrated its impact indirectly, particularly thru eGovernment. Contrary to our expectations, the role of LO was found to be relatively weak.
- *H2: LO impact on KM.* This link was not found to be significant, at least not in this formulation of the model, where the target variable is eGovernance.
- *H3: The role of eGovernment on eGovernance was found to be high, while the iGovernment has only indirect impact on eGovernance via eGovernment.* These relations were fully supported by the evidence.

The weak role of LO and the lack of the causal link between LO and KM may be the result of low actual role of LO in the organizations under study. However, it may also arise from how we operationalize these concepts in the study. Recall that eGovernance was measured with regard to the level of ICT sophistication of the governmental services, ranging from Web page presence to on-line payment for governmental services. The LO concept was measured based on the extent of an applicable rewards system, information sharing and networking, and as well by time employee spent in educational activities. On the other hand KM was measured with explicit questions about actual KM practice. It is possible that the LO concept was not measured adequately here. This possibility is supported by the high level of unexplained variance, and by the poor impact on eGovernance and weak role in KM.

4 The research results and Slovenian eGovernance

Although the main focus of this research was to build a credible model linking ICT, KM and LO to eGovernance, our results also provide an opportunity for some substantial comments on the Slovenian eGovernance situation. Gaining its independence in 1991 Slovenia was among the most developed "transition countries" joining EU in 2004. Although still with a substantial lag in GDP/per capita (around 80% of EU15 average in 2006) Slovenia's ICT was highly developed. The comparisons with corresponding results for OECD countries⁹ confirm that Slovenia's ICT deployment in public administration was at high level compared to OECD

⁹ Elsa Pilichowski: The Learning Government: Introduction and Draft Results of the Survey of Knowledge Management Practices in Ministries/agencies/departments of Central Government, 09-Nov-2005
http://www.oecd.org/statisticsdata/0,2643,en_2649_34129_1_119656_1_1_37441,00.html

average. Similarly, there was strong budgetary support for “hard-core” ICT, such as equipment, maintenance, etc. In this respect Slovenian public administration organizations were better equipped than OECD average. This is not surprising and is in line with historical developments of ICT in Slovenia¹⁰. It also matches results from 2006 IDC study¹¹, where the share of ICT spending (around 1%) within the government budget was among the highest in EU, and is in accord with the corresponding Eurostat 2005¹² comparisons for the companies. We observe that Slovenian companies are above the old EU15 member states average in (broadband) Internet access and PC/employee ratio.¹³ However, there appears to be an increasing lag with respect to the advanced use of technologies such as video conferencing, intranets, etc. Similarly, the OECD study comparisons show that Slovenian public administration institutions lag in the application of the advanced features of ICT. A strong orientation towards technology solutions still takes precedence over organizational and knowledge issues. An indicator of this is that, for specialist positions responsible for KM, in Slovenia not a single position of Chief Knowledge Officer (CKO) was found, in contrast to 15% for all institutions in the OECD study. This is in sharp contrast with institutions from the USA, Canada and the United Kingdom where more than half of have a CKO. The system of rewards is another indicator that differentiates Slovenian results – where these systems very rarely exist - from the OECD, where 60% of public administration organizations report that employees are rewarded for sharing knowledge and information, and where 80% of them list knowledge/information sharing as a criterion for the assessment of staff performance.

The survey results suggest that many Slovenian decision-makers still think that KM begins and ends with building sophisticated information technology systems and that no further organizational change is required. As a consequence a certain gap is appearing, one already observed in 2003 SIBIS¹⁴ study. The study reported that, technically speaking, Slovenia provided a rich array of on-line services, but that they were not optimized for users. Citizens wouldn't use them if they are not “user friendly” or they are not even aware of them.

We can observe the persistence of this situation in 2006. There are numerous eGovernment on-line services provided, but they are not optimized. A recent

CapGemini 2006 study thus revealed that with respect to sophistication of on-line services Slovenia ranks a high 7th position among 28 European countries (EU member States and Norway, Iceland and Switzerland) and just ahead of Denmark, Finland, France and Ireland.¹⁵ However, it is noted that everywhere eGovernment on-line services are very often not effectively used. After three years of an on-line income tax filing option, only a small percentage of taxpayers actually use the service. In the 2005 UN Global eGovernment Readiness Report (2005)¹⁶ Slovenian eGovernment services ranked 26th place (index: 0.6762) out of the 191 Member States of the UN researched for an eGovernment Readiness index. In the web index that measure the websites of the governments to determine if they are employing eGovernment to the fullest, Slovenia ranked 36th (index: 0.5923). A 2005 e-participation Index, measuring how relevant and useful the e-participation features of government websites around the world ranked Slovenia 46th (index: 0.2222). The ranking in this last index shows a lack of real access and inclusion, a challenge shared by the majority of countries in the world. In part this can be attributed to poor ICT strategic planning and the failure to maximizing the potential benefits from KM and LO efforts.

Another important finding relates to the factors that affect the introduction of LO principles into public organizations. There, decentralization strategies are highly relevant. We note that 75% of the OECD organizations, but only 20% of Slovenian organizations, report that they have taken initiatives in recent years to decentralize and delegate authority to lower hierarchical levels. In the absence of such decentralization, KM and LO initiatives are difficult to undertake, and when attempted face serious obstacles.

For an effective eGovernance a vision for systemic change throughout the whole organization is called for. This includes the implementation of KM principles and a culture of LO behavior, supported with ICT tools. We can summarize by paraphrasing an introductory statement on OECD eGovernment site:¹⁷

“eGovernance” is much more about “Governance” than about the “e.” The results of this study, for Slovenia in particular, confirm that ICT alone is not enough.

5 Conclusions

¹⁰ Next steps in development Information Society Services in the New Member States. The case of e-Government, Draft, August 2006, <http://ipts.ris.org>.

¹¹ <http://www.gzs.si/DRNivo2.asp?ID=27615&IDpm=511>

¹² <http://epp.eurostat.ec.europa.eu>

¹³ <http://slovenia.ris.org/index.php?fl=2&lact=1&bid=64&menu=0>

¹⁴ <http://slovenia.ris.org/index.php?fl=2&lact=1&bid=58&menu=0>

¹⁵ http://www.de.capgemini.com/m/de/tl/EU_eGovernment-Studie_2006.pdf

¹⁶ UN Global E-government; Readiness Report 2005 - From E-government to E-inclusion; UNPAN/2005/14; see: <http://unpan1.un.org/intradoc/groups/public/documents/un/unpan021888.pdf>, (September 2006).

¹⁷ http://www.oecd.org/topic/0,2686,en_2649_34129_1_1_1_1_37405,00.html

In this research we have investigated the strategic influences of information and communication technology (ICT), knowledge management (KM) and learning organization (LO) principles and culture on eGovernance in public administration organizations. We first developed a conceptual model linking ICT, KM and LO. We then made it operational and empirically tested it on data from Slovenian public administration organizations.

Through our conceptual model, its operationalisation, our researched hypotheses and empirical testing, we are contributing to a better understanding of the principles and practices involved in building the foundations for good eGovernance practice. We demonstrated that building the new ICT infrastructure is not enough, nor is it enough to digitalize the processes, or just provide appropriate knowledge, skills and training to use the technology effectively. Of course, the role of ICT (via iGovernment and eGovernment) was shown to be a strong driver for the eGovernance developments. However, the role of KM was found to be also extremely important. In order to harvest the real benefits of the technology, KM principles need to be applied as an important element for effective eGovernance. The same is likely true for the implementation of a LO culture, although the evidence is weaker.

Our findings about the relationships between ICT, KM, LO and eGovernance are relatively general, and relevant for governmental organizations as well as for other organizations in developed countries. However, they prove to be particularly relevant for the Slovenian situation. On one hand Slovenia is positioned among the leading EU countries with respect to the array and sophistication of its implementation of ICT into eGovernment services. However, on the other hand, the lack of other crucial activities, particularly those related to KM and LO, presents a considerable obstacle for achieving the full benefits of each part in the delivery of eGovernance services.

One of the limitations of our study is that with regard to the effectiveness and efficiency of eGovernance we used no data from users' side. We had no data from citizens with regard to—user's satisfaction (with regard to connectivity, transparency, 'kindness', etc.). We also have no measures of public sector/citizen views on the adequacy eGovernance outcomes (taxation policy, social services, etc.). Our research was focused entirely on the providers' side of the eGovernment process. As well, we note that we actually studied the entire target population, all of the public administration sector, and not just a sample from that sector. Such approach reduces sampling error and bias to zero, and is preferred when feasible.

Another limitation in our research may be the relatively low response from municipal offices. However, with respect to the key variables we found

little difference between them and the larger public administration bodies where we obtained remarkably high response rates. This lack of difference is also true across the various segments of surveyed institutions.

We note that we actually studied the target population, the public administration sector, and not just a sample from that sector. Such approach is preferred, when feasible. In our model we implicitly assume that the behaviour of observed units is just the manifestation of some general principles or causal relationships that we want to identify. As well, in such work we need a critical number of units in order to have statistically significant results.

One additional conclusion from this research suggests that "how the LO principles fit into the model" was not optimally (or adequately) addressed in our research and therefore the results show relatively low link of LO (**learn_org**) factors. This area requires further work, in particular in elaborating the relationship between LO and KM. Within a refined model, much more profound analysis could be performed using advanced tools of causal modeling. Full three-way interactions could be addressed, exploring the extent to which various levels of LO development create different relationships between KM and ICT, and in turn how they affect eGovernance.

ICT, via both computational power and connectivity and the resulting digital venue, is a major player in an organization's pursuit of its mission, vision and mandate. It is reshaping the way organizations communicate within and without, carry out their organizational practices, and deliver services to the public. It is not enough to attempt to improve only one element, ICT, in the ICT, KM, LO triad, and fail to recognize that LO, and particularly KM, are essential for proper targeting eGovernance objectives. Progress depends on both technical and organizational change, and ICT professionals need to work closely with top management, including human resource management, in the deployment of KM and LO strategies.

The speed of technological change, and the fast pace of development of new services and products, raise KM to a crucial role in eGovernment. Workforce reductions, cost cutting measures, and the demands of just-in-time and life-long learning are further factors that raise the importance of KM principles. We suspect that further research in organizations with differential learning organization culture, will also elevate LO culture to the status of a key element in organizational behaviour for efficient and effective eGovernance.

6 References

- [1] Berce Jaro. "The influence of Information Society Technologies on Evolution in the Public Administration", *Merkur Day* 2004, 6th

- Undergraduate and Graduate Students eCommerce Conference & 8th Executive Business, Government, and University Meeting on Cross-border eInvoice in eRegion, Organized by Faculty of Organizational Science, University of Maribor (ISBN 961-232-171-X, CIP 659.2.004(063) (086.2.034) - cd-rom and published on Internet page: <http://ecom.fov.uni-mb.si/MerkurDay2004/program.htm>), Naklo October 22, 2004.
- [2] Clift Steven. "E-Governance to E-Democracy: Progress in Australia and New Zealand toward Information-Age Democracy", March 2002; <http://www.Publicus.Net> (Apr., 2003).
- [3] Churchman C.W.; *The design of inquiring systems*; Basic books; New York, NY, 1971; p.10.
- [4] Ewalt Jo Ann G. "Theories of Governance and New Public Management: Links to Understanding Welfare Policy Implementation", Prepared for presentation at the Annual conference of the American Society for Public Administration, Newark, NJ, March 12, 2001.
- [5] Five Discussion Pieces. "Collaborative Government In The Post-Industrial Age"; Centre for Collaborative Government; http://www.crossingboundaries.ca/reports/ktapublication_may2000.pdf; (May, 2003).
- [6] Mansell Robin, Barnett Andrew. "From moneybags to know-it-all. The World Bank shift its stall", OneWorld News Service, October 7, 1998, <http://www.oneworld.org/guides/wb2.html> (Apr., 2003).
- [7] Marchionini Gary, Samet Hanan, Brand Larry. "Digital Government", *Communication of the ACM* (Vol. 46, No. 1.), January 2003, p. 25-27.
- [8] Nath Vikas. "Concept of Digital Governance", *Digital Governance and KnowNet Initiative*, Inlaks Scholar, London School of Economics, UK, <http://www.cddc.vt.edu/digitalgov/govmenu.html> (Apr., 2003).
- [9] Noordegraff M. "Vigorous knowledge management-Conceptual framework for knowledge management in (central) government", Ministry of the Interior and Kingdom Relations, Public Sector Information Policy Directorate: Utrecht, 2001, p. 28.
- [10] OECD: "Sustainable Development and Governance", http://www.oecd.org/topic/0,2686,en_2649_3740_5_1_1_1_1_37405,00.html (Apr., 2003).
- [11] Pedler, M., Burgoyne, J., T., Boydell. "The Learning Company". A strategy for sustainable development, London: McGraw-Hill 1991, 1996.
- [12] Perkmann Markus. "Measuring knowledge value? Evaluating the impact of knowledge projects"; *Knowledge & Innovation Network*; KIN brief #7 - 26/07/02; <http://www.ki-network.org>; (May, 2003).
- [13] Peters, B. G. and Pierre J. "Governance Without Government? Rethinking Public Administration." *Journal of Public Administration Research and Theory*, Vol. 8, No. 2: 223-243, 1998.
- [14] Pierce John. "Knowledge: What is it?"; *Knowledge Management for beginners*: <http://www.knowledgeboard.com> (Apr., 2003), p.1-8.
- [15] Nonaka Ikujiro, Takeuchi Hirotaka; *The knowledge-creation company*; Oxford University Press; New York, NY; 1995.
- [16] Senge, M. Peter. "The Fifth Discipline". *The art and practice of the learning organization*, London: Random House. 424 + viii pages, 1990.
- [17] The International Teledemocracy Centre. http://itc.napier.ac.uk/ITC_Home/ITC/Research.asp.
- [18] Willet S. & Copeland L.. "Knowledge management key to IBM's enterprise plan"; *Computer reseller news*; July 27, 1998; p. 1-6.
- [19] Malkia Matti, Ari-Veikko Anttiroiko, and Reijo Savolainen, eds.; *eTransformation in Governance - New Directions in Government and Politics*; London: Idea Group; 2004.
- [20] Reinermann Heinrich, and Jörn von Lucke. "Speyerer Definition von Electronic Governance"; *Electronic Government in Deutschland*; Heinrich Reinermann and Jörn von Lucke; Speyer: Forschungsint. für Öffentliche Verwaltung; 2002; pp. 9-19.
- [21] Riley B. Thomas. "Knowledge Management: An Evolving Discipline"; A paper for National Research Council Canada; February 2000, Ottawa; <http://rileyis.indelta.com/publications/NewDocs/KnowledgeManagement/KM-AN-EVOLVING-DISCIPLINE.htm>; (October 2005)
- [22] Marsick J. Victoria and Watkins E. Karen. "The Learning Organization: An Integrative Vision for HRD"; *Human Resource Development Quarterly*, Winter 1994
- [23] Gephart A. Martha, Marsick J. Victoria, Van Buren E. Mark, and Spiro S. Michelle. "Learning Organizations Come Alive"; *Training & Development*; December 1996
- [24] Vehovar, V. "Social Informatics; An Emerging Discipline?" In Berleur J. ed.: "Social Informatics - an Information society for all"; Springer Publisher, September 2006.
- [25] Bučar, M.. "Evolucija Ocenjevanja Razsežnosti In Učinkov Informacijsko-Komunikacijske Tehnologije, Merjenje ITK"; editorial: Metka Stare, Maja Bučar, Fakulteta za družbene vede, ISBN 659.23.004, 2005, p.23-40.
- [26] Baker M.A.. "Policy Bridges for the Digital Divide: Assessing the landscape and Gauging the Dimension"; *First Monday*, volume 6, number 5 (peer-reviewed journal on the Internet - Paper received 9 April 2001; accepted 20 April 2001),

- http://www.firstmonday.org/issues/issue6_5/baker/#author (July 2006)
- [27] Lesjak D., Vehovar V. "Factors affecting evaluation of e-business projects". *Industrial Management & Data Systems* (Vol. 105, No. 4), May 2005, p. 409 – 428.
- [28] Jöreskog, K.G., Sörbom, D. and Du Toit, S.. "LISREL 8: New Statistical features"; Scientific Software International, Lincolnwood, IL et al., 2000).
- [29] Boomsma, A. and Hoogland, J.J.. "The robustness of LISREL modeling revisited", in Cudek, R., Du Toit, S. and Sörbom, D (eds), *Structural Equation models: Present and future. AFestschrift in Honor of Karl Jöreskog*, Scientific Software international, Chicago, IL, pp. 139-68, available at: www.psw.rug.nl/~boomsma/ssi.pdf (accessed September 1st 2004).
- [30] Saurina, C. and Coenders, G.. "Predicting overall service quality. A structural equation modeling approach", in Ferligoj, A. and Mrvar, A. (Eds), *Developments in Social Science Methodology, zbirka Metodološki yveyki*, 18, pp. 217-38, available at: <http://mrvar.fdv.uni-lj.si/pub/mz/mz18/coender2.pdf> (accessed September 1st 2004).
- [31] Browne, M.W. and Cudeck, R.. "Alternative ways of assessing model fit", in Bollen, K.A. and Long, J.S. (Eds), *Testing Structural Equation Models*, Sage, Thousand Oaks, CA, pp. 136-62.
- [32] Lundvall B. and Johnson, B. 'The learning economy', *Journal of Industry Studies*, Vol.1, No.2. 1994.
- [33] Klaus Beck. "Organizational learning"; November 10, 1997; <http://www.sfb504.uni-mannheim.de/glossary/orglearn.htm> (August 2006).
- [34] Druker F. Peter. "Management challenges for the 21st century"; Harper Business; New York, NY, (1999).
- [35] Sveiby Kare-Erik. "The New Organizational Wealth, Managing & Measuring Knowledge-Based Assets"; Berrett - Koehler Publishers, (1997).
- [36] Steward T.A.; *Intellectual capital: The new wealth of organizations*; Doubleday / Currency; New York, NY; 1997.
- [37] Sakaiya Taichi. "Knowledge Value Revolution: Or a History of the Future"; Publisher: Kondansha Intl. (1991).
- [38] OECD - KM survey in the public sector - Draft report on Total results, September 2002; February, 2003; GOV/PUMA /HRM(2002)2 – Unclassified.
- [39] Carr, N. Why IT Doesn't Matter Anymore. *Harvard Business Review*, 81, 5, (May 2003). Source: <http://hbswk.hbs.edu/archive/3520.html> (December 2006).
- [40] Working paper: "Rethinking Globalization after the Collapse of the Financial Bubble: An essay on the challenges of the Third Millennium" (Paper prepared for the First Globelics Conference, Rio, November 2-6, 2003). Source: <http://www.carlotaperez.org/> (December 2006).
- [41] Geldof Marije. "BECOMING AN INFORMATION SOCIETY: THE ROLE OF NEW INFORMATION TECHNOLOGIES IN DEVELOPMENT ", Editorial assistance from Professor Tim Unwin, Report based on Wilton Park Conference 798: 17 – 20 October 2005 on "Becoming an Information Society: The Role of New Information Technologies in Development", Wilton Park Paper, November 2005.
- [42] Drucker Peter F.; *The theory of Business*; Harvard business Review, September-October 1994; p 95-104.
- [43] Barclay O. Rebecca (Managing Editor, Knowledge Praxis). "What is knowledge management?", http://www.providersedge.com/docs/km_articles/What_Is_Knowledge_Management.pdf (May 2007).

APPENDIX A

The QUESTIONNAIRE

1) Background

Please specify on behalf of whom you are filling this questionnaire:

- 1/01 - The Ministry of: _____
 1/02 - The constituent offices: _____
 1/03 - The local government of: _____
 1/04 - The municipality of: _____
 1/05 - The government office of: _____

Please fill the background information on your institution:

- 1/06 - The total budget of your institution in last year:
 1/07 - The total number of employees
 1/08 - (%) of employees spending their entire employment within the public sector

Please indicate in percentage (%) educational structure of your institution:

- 1/09 - The primary school or less _____
 1/10 - The secondary school _____
 1/11 - The tertiary diploma (below university BA/BS one) _____
 1/12 - The university diploma _____
 1/13 - The masters degree _____
 1/14 - The PhD _____

Please define in approximate percentage (%) your customer structure:

- 1/15 - Within public sector: _____
 1/16 - Outside of public sector: _____

2) ICT infrastructure

Approximate what percentage (%) of your staff (or working posts):

- 2/01 - Are equipped with a computer: _____
 2/02 - Are linked to a local area network: _____
 2/03 - Has access to the Internet: _____
 2/04 - Has an e-mail address: _____
 2/05 - Has a "secureID" acces: _____
 2/06 - Has a "CA-sigov" certification key: _____

Your institution has (SEVERAL ANSWERS POSSIBLE)

- 2/07 - An internal document for ICT Security and Protection
 2/08 - A System for Data achieving (e.g.: tape, disk, etc.)
 2/09 - A system for virus protection
 2/10 - A contiguous plan

Your institution has an Internet site:

- 2/11 - YES → which year from _____
 2/12 - NO but we are planning to have in next 5 years /Please precise date: _____/
 2/13 - NO but we are not planning to have an Internet site → *GO TO QUESTION 9.*

Your institution has an information center?

- | | | | |
|---------------------------------|-----|---------------------------------|---|
| | YES | NO | |
| 2/26 - <input type="checkbox"/> | | 2/27 - <input type="checkbox"/> | Our institution has information center |
| 2/28 - <input type="checkbox"/> | | 2/29 - <input type="checkbox"/> | The center representative is attending board meetings |

Please specify the percentage of total budget spent for ICT _____ - 2/30

Has this percentage increased in the last five years? 2/31 - YES 2/32 - NO
If YES, by how much (%): _____ - 2/33

3) ICT services and e-business model

In your institution, initiatives have been/will be taken, such as:

- | | | | |
|---------------------------------|---------------------------------|---------------------------------|---|
| YES, | NO, but we plan | NO, no planning | |
| 3/01 - <input type="checkbox"/> | 3/02 - <input type="checkbox"/> | /03 - <input type="checkbox"/> | Decentralization/delegation of authority to lower hierarchical levels |
| 3/04 - <input type="checkbox"/> | 3/05 - <input type="checkbox"/> | 3/06 - <input type="checkbox"/> | Reorganization of internal process |
| 3/07 - <input type="checkbox"/> | 3/08 - <input type="checkbox"/> | 3/09 - <input type="checkbox"/> | Reorganization of offices |
| 3/10 - <input type="checkbox"/> | 3/11 - <input type="checkbox"/> | 3/12 - <input type="checkbox"/> | Internal network to share information |

Please specify percentages (%) of your employees using information tools:

- 3/13 - Text editor (e.g.: Word): _____
 3/14 - Spreadsheet (e.g.: Excel): _____
 3/15 - Tool for presentation (e.g.: PowerPoint): _____
 3/16 - e-post system (e.g.: SPIS): _____
 3/17 - e-government meetings: _____
 3/18 - Legislation procedures of Parliament: _____
 3/19 - e-budget planning (e.g.: MFERAC): _____
 3/20 - Infoklip: _____
 3/21 - e-Project office: _____
 3/22 - Other /please specify/ _____

Your work procedures are based on:

- | | | |
|--------------------------------------|---------------------------------|--|
| YES | NO | |
| 3/23 - <input type="checkbox"/> | 3/24 - <input type="checkbox"/> | A specific team responds to e-mail on institutional general record |
| 3/25 - <input type="checkbox"/> | 3/26 - <input type="checkbox"/> | Maximum delays for responding to e-mail have been established |
| 3/27 - <input type="checkbox"/> | 3/28 - <input type="checkbox"/> | Clearance mechanisms for correspondence with citizens have been simplified |
| 3/29 - <input type="checkbox"/> | 3/30 - <input type="checkbox"/> | The policy for responding to e-mails is similar to formal letters |
| 3/31 - <input type="checkbox"/> | 3/32 - <input type="checkbox"/> | e-mail notification for meetings is used within our institution |
| 3/33 - <input type="checkbox"/> | 3/34 - <input type="checkbox"/> | e-calendar is used for coordinating meetings within our institution |
| 3/35 - <input type="checkbox"/> | 3/36 - <input type="checkbox"/> | e-addresser is used for our customer base |
| 3/37 - <input type="checkbox"/> | 3/38 - <input type="checkbox"/> | We have internal digital base for advertisement post for employees |
| 3/39 - <input type="checkbox"/> | 3/40 - <input type="checkbox"/> | We are involved in current e-administration projects |
| 3/41 - If YES, please specify: _____ | | |

On your institution’s Internet site, it is possible to:

- | | | |
|---------------------------------|---------------------------------|---|
| YES | NO | |
| 3/42 - <input type="checkbox"/> | 3/43 - <input type="checkbox"/> | Access administrative information |
| 3/44 - <input type="checkbox"/> | 3/45 - <input type="checkbox"/> | Access information about current events related to your institution |
| 3/46 - <input type="checkbox"/> | 3/47 - <input type="checkbox"/> | Download publications, reports from your institution |
| 3/48 - <input type="checkbox"/> | 3/49 - <input type="checkbox"/> | Access to statistical or quantitative data |
| 2/50 - <input type="checkbox"/> | 3/51 - <input type="checkbox"/> | Download administrative forms/applications |
| 3/52 - <input type="checkbox"/> | 3/53 - <input type="checkbox"/> | Fill in administrative forms electronically |
| 3/54 - <input type="checkbox"/> | 3/55 - <input type="checkbox"/> | Make certain payments (taxes, public fees, etc.) |

3/56-3/83 **Please indicates your level of agreement (1 strongly disagree... 4 strongly agree):**

3/56-3/59	Your institution delivers on your Internet site all important documents and information
3/60-3/63	Information delivered on your Internet site is well packed (clear, understandable, etc.)
3/64-3/67	Information delivered on your Internet site is updated on a regular basis
3/68-3/71	Information communication technology is improving working results
3/72-3/75	Information communication technology is improving internal communication (within institution)
3/876-3/79	Information communication technology is improving external communication
3/80-3/83	Information communication technology is improving knowledge sharing

Do you systematically register working processes:

- 3/84 - On paper
 3/85 - With e-Documents
 3/86 - In Database
 3/87 - Within Expert system
 3/88 - Stays with Experts working in your institution

4) Knowledge Management and Learning Organization practices

4/01-4/20 Please indicates your level of agreement (1 strongly disagree... 4 strongly agree):

4/01-4/04	Your institution is using knowledge management practices
4/05-4/08	In your institution employees have time for knowledge management practices
4/09-4/12	Your institution »culture« is encouraging knowledge management sharing
4/13-4/16	Your institution has understanding for knowledge management techniques
4/17-4/20	In your institution organizational processes are designed for knowledge management

In your institution

- | | YES | NO | |
|---------------------------------|---------------------------------|----|--|
| 4/21 - <input type="checkbox"/> | 4/22 - <input type="checkbox"/> | | There is a database of staff competencies |
| 4/23 - <input type="checkbox"/> | 4/24 - <input type="checkbox"/> | | There is a database of presentations and documents for common usage that is systematically updated |

Which of the following groups has the overall responsibility for knowledge and information management and transfer practices in your institution?

- 4/25 - Human resources management team
 4/26 - Information technology team
 4/27 - Special knowledge and information management unit (knowledge officer)
 4/28 - Top management
 4/29 - Other, please specify: _____
 4/30 - None

In your institution, workers are rewarded for knowledge and information sharing:

- 4/31 - YES → *Incentives include (TICK AS MANY BOXES AS YOU DEEM NECESSARY):*
 4/32 - Monetary incentive
 4/33 - Prizes / Rewards
 4/34 - Promotion
 4/35 - Informal encouragement
 4/36 - Other, please specify: _____
 4/37 - NO

How many days of training per staff (on average) are provided each year by your institution?

- 4/38 - None
 4/39 - 1 day: _____
 4/40 - 2 to 5 days: _____
 4/41 - 6 to 10 days: _____
 4/42 - More than 10 days: _____

Is there a special budget allocated to training in your institution?

- 4/43 - YES, please indicate percentage (%) of total budget____, 4/44 - NO

Has percentage increased in last five years? YES 4/46 → how much (%): __ - 4/47 NO 4/45**Please specify the approximate training percentage (%) by type (on day per employee basis):**

- 4/48 - General (e.g. language, etc.): _____
 4/49 - Professional (e.g. learning to get higher degree, etc.): _____
 4/50 - Information technology: _____

APPENDIX B

Compound (Independent) variable name	Compound variable calculation (survey questionnaire questions from APPENDIX A are in the form: “vX Y”)	Second layer endogenous (intervening) variable	Exogenous variable
ICT Infrastructure	$infrastr = (((v2_1+v2_2+v2_3+v2_4+v2_5+v2_6)/6) + ((v3_13+v3_14+v3_15+v3_16+v3_17+v3_18+v3_19+v3_20)/8))/2$	iGovernment iGovern = v3_84+v3_85+v3_86+v3_87+v3_88	eGovernance eGovance = v2_11 +v3_42 + v3_23 +v3_25 +v3_27 +v3_29 +2*v3_44 +3*v3_46 +4*v3_48 +5*v3_50 +6*v3_52 +7*v3_54
ICT Policy	policy = v2_7+v2_8+v2_9+v2_10		
ICT Budget	Budget1=v2_30 Budget2 =v2_31		
Organizational (ICT)	organiza = v2_26+v2_28+v3_10	eGovernment eGovern = v3_31+v3_33+v3_35+v3_37+v3_39	
ICT Infrastructure	$infrastr = (((v2_1+v2_2+v2_3+v2_4+v2_5+v2_6)/6) + ((v3_13+v3_14+v3_15+v3_16+v3_17+v3_18+v3_19+v3_20)/8))/2$		
Internet - Web presence	webpres = v2_11		
Web information Update	webinfo = (v3_56+v3_60+v3_64)/3		
eCapture of Staff Competence	eCapture = v4_21+v4_23	Knowledge Management Kman = (v4_1+v4_5+v4_9+v4_13+v4_17)/5	
Information / Knowledge Responsibility	info = v4_25		
Rewarding System	reward = v4_32+v4_33+v4_34+v4_35+v4_36		
KM Networking	network = (v3_68+v3_72+v3_76+v3_80)/4		
Organizational initiative	iniciat = v3_1+v3_4+v3_7	Learning Organization lear_org = v4_48+v4_49	
Educational Level	educat = v1_12 + v1_13 + v1_14		
Permanent Education	permedu = v4_38		
Training	trainb = v4_43+v4_45		
KM Networking	network = (v3_68+v3_72+v3_76+v3_80)/4		

Table 2: Calculation of compound, endogenous and exogenous variable

Efficient Hierarchical Identity Based Encryption Scheme in the Standard Model

Yanli Ren and Dawu Gu
 Dept. of Computer Science and Engineering
 Shanghai Jiao Tong University
 Shanghai 200240, China
 E-mail: {renyanli1982,dwgu}@situ.edu.cn

Keywords: hierarchical, identity based, standard model

Received: September 12, 2007

Constructing identity based schemes is one of the hot topics of current cryptography. Hierarchical identity based cryptography is a generalization of identity based encryption that mirrors an organizational hierarchy. It allows a root public key generator to distribute the workload by delegating public key generation and identity authentication to lower-level public key generators. Currently, there is no hierarchical identity based encryption scheme that is fully secure in the standard model, with short public parameters and a tight reduction. In this paper, we propose an anonymous hierarchical identity based encryption scheme based on the q -ABDHE problem that is fully secure in the standard model. The ciphertext size is independent of the level of the hierarchy. Moreover, our scheme has short parameters, high efficiency and a tight reduction.

Povzetek: Opisana je kriptografska metoda za hierarhično identifikacijo.

1 Introduction

Identity based (ID-based) cryptosystem [1] is a public key cryptosystem where the public key can be represented as an arbitrary string such as an email address. A private key generator (PKG) uses a master secret key to issue private keys to identities that request them. For an Identity Based Encryption (IBE) scheme, Alice can securely encrypt a message to Bob using an unambiguous name of him, such as email address, as the public key. For an Identity Based Signature (IBS) scheme, Alice can sign a message using her private key that corresponds to Alice's identity. Then anybody can verify the authenticity of the signature from the identity.

The concept was proposed by Shamir in 1984. However, practical IBE schemes were not found until the work of Boneh and Franklin in 2001 [8]. Their scheme is provably secure in the random oracle model. Almost all of the IBE systems since Boneh-Franklin follow the "common strategy" for proving security; consequently, they suffer from long parameters (when security is proven in the standard model) and lossy reductions (in the standard model or the random oracle model). The IBE systems described in [5] have short parameters and achieve a tight reduction, but this is because they are proven secure only against selective-ID attacks. In 2006, Gentry proposed an anonymous IBE scheme [4] that is fully secure in the standard model with a tight reduction. Anonymity means that there is no adversary can distinguish two ciphertexts of same message with two identities in polynomial time.

Hierarchical ID-based cryptography was first proposed in [3] and [9] in 2002. It is a generalization of IBE that mir-

rors an organizational hierarchy. And it allows a root PKG to distribute the workload by delegating private key generation and identity authentication to lower-level PKGs. In a hierarchical ID-based encryption (HIBE) scheme, a root PKG only needs to generate private keys for domain-level PKGs, who in turn generate private keys for their users in the domains of the lower level. To encrypt a message to Bob, Alice only needs to obtain the public parameters of Bob's root PKG and his identity. It is especially useful in large companies or e-government structure where there are hierarchical administrative issues needed to be taken care.

The first construction for HIBE is due to Gentry and Silverberg [3] where security is based on the Bilinear Diffie-Hellman (BDH) assumption in the random oracle model. A subsequent construction due to Boneh and Boyen gives an efficient (selective-ID secure) HIBE based on BDH without random oracles [5]. But the ciphertext length is linear in the depth of the hierarchy. In 2005, they proposed a hierarchical identity based encryption with constant size ciphertext and proved it is selective-ID secure in the standard model [7]. Moreover, the size of public parameters is independent of the number of bit representing an identity, while the size of public parameters of the scheme in [11] grows with a factor of h , where h is the number of block to represent an identity of n bits, with each block using n/h bits. In 2006, Man Ho Au constructed a HIBE scheme that is fully secure in the standard model [10]. However, the scheme can not convert to an IBE scheme, that is to say, it is only valid for a user with identity $ID = (ID_1, ID_2, \dots, ID_i), i \geq 2$. Moreover, the adversary can compute the private key of ID_1 after requesting private key of its children and the

$q - SDH$ problem can not be solved exactly during the reduction. At the same year, an anonymous HIBE [12] is proposed. But the ciphertext size is dependent to the level of the hierarchy. In addition, the scheme has long parameters, large computation and the reduction is not tight. Currently, there is no HIBE scheme that is fully secure in the standard model, with short public parameters and a tight reduction.

Our Contributions. In this paper, we propose a constant size anonymous HIBE scheme that is fully secure in the standard model. The ciphertext size is independent of the level of the hierarchy. Compared to the previous HIBE schemes, our scheme has shorter parameters, higher efficiency and a tighter reduction.

Our scheme is based on Gentry’s IBE scheme, and we convert it to a HIBE scheme. However, the conversion is not straightforward. Several techniques have to suitably combined to obtain the required proof. Moreover, our scheme decreases the redundancy of Gentry’s scheme.

2 Definitions

Before presenting the hierarchical identity based encryption scheme, we introduce some difficult problems and security models of the scheme first.

2.1 Bilinear Map

Let p be a large prime number, G_1, G_2 are two groups of order p , g is a generator of G_1 . $e : G_1 \times G_1 \rightarrow G_2$ is a bilinear map, which satisfies the following properties [2]:

- (1)Bilinearity: For all $u, v \in G_1$ and $a, b \in \mathbb{Z}$, $e(u^a, v^b) = e(u, v)^{ab}$.
- (2)non-degeneracy: $e(g, g) \neq 1$.
- (3)Computability: There exists an efficient algorithm to compute $e(u, v)$, $\forall u, v \in G_1$.

2.2 Complexity Assumption

The security of our scheme is based on a complexity assumption that we call the q -augmented bilinear Diffie-Hellman exponent (ABDHE) problem [4].

q-ABDHE problem

Let g, g' be generators of G_1 . Given $(g, g^\alpha, \dots, g^{\alpha^q}, g', g'^{\alpha^{q+2}}, T) \in G_1^{q+3} \times G_2$, where $\alpha \in \mathbb{Z}_p^*$, decide whether $T = e(g, g')^{\alpha^{q+1}}$.

Since the tuple has not the term $g^{\alpha^{q+1}}, g'^{\alpha^{q+1}}$, the bilinear map does not seem to help decide whether $T = e(g, g')^{\alpha^{q+1}}$. Introducing the additional term $g'^{\alpha^{q+2}}$ still does appear to ease the decision of $e(g, g')^{\alpha^{q+1}}$, since the tuple is missing the term $g^{\alpha^{-1}}$. we say the q -ABDHE problem is (t, ε) -difficult in G_1, G_2 , if no t -time algorithm has advantage at least ε in solving the q -ABDHE problem.

An algorithm A that outputs $b \in \{0, 1\}$ has advantage ε in solving the decision q -ABDHE if

$$|Pr[A(g', g'^{\alpha^{q+2}}, g, g^\alpha, \dots, g^{\alpha^q}, e(g', g)^{\alpha^{q+1}}) = 0] - Pr[A(g', g'^{\alpha^{q+2}}, g, g^\alpha, \dots, g^{\alpha^q}, T) = 0]| \geq \varepsilon,$$

where the probability is over the random choice of generators $g, g' \in G_1, \alpha \in \mathbb{Z}_p^*, T \in G_2$, and the random bits consumed by A . We refer to the distribution on the left as P_{ABDHE} and the distribution on the right as R_{ABDHE} .

We say that the decision (t, ε, q) -ABDHE assumption holds in G_1, G_2 if no t -time algorithm has advantage at least ε in solving the decision q -ABDHE problem in G_1, G_2 .

2.3 Secure Model

IND-ID-CCA2: Boneh and Franklin defined chosen ciphertext security for IBE systems under a chosen ciphertext attack via the following game [6,8].

Setup: The challenger runs Setup, and forwards parameters to the adversary.

Phase 1: Proceeding adaptively, the adversary issues queries q_1, \dots, q_m where q_i is one of the following:

Key generation query $\langle ID_i \rangle$: the challenger runs *KeyGen* on ID_i and forwards the resulting private key to the adversary.

Decryption query $\langle ID_i, c_i \rangle$. The challenger runs *KeyGen* on ID_i , decrypts c_i with the resulting private key, and sends the result to the adversary.

Challenge: The adversary submits two plaintexts m_0, m_1 and an identity ID^* . ID^* or its prefix must not have appeared in any key generation query in Phase 1. The challenger selects a random bit $b \in \{0, 1\}$, sets $c^* = \text{Encrypt}(params, ID^*, m_b)$, and sends c^* to the adversary as its challenge ciphertext.

Phase 2: This is identical to Phase 1, except that the adversary may not request a private key for ID^* or the decryption of (ID^*, c^*) .

Guess: The adversary submits a guess $b' \in \{0, 1\}$. The adversary wins if $b' = b$.

We call an adversary A in the above game an IND-ID-CCA adversary. The advantage of an adversary A in this game is defined as $Pr[b' = b] - \frac{1}{2}$.

Definition 1. An HIBE system is $(t, \varepsilon, q_e, q_d)$ IND-ID-CCA secure if all t -time IND-ID-CCA adversaries making at most q_e key generation queries and at most q_d decryption queries have advantage at most ε in winning the above game.

ANON-ID-CCA2: Informally, we say that an HIBE system is anonymous if an adversary cannot distinguish the public key ID under which a ciphertext was generated. More formally, we define anonymity for HIBE systems under a chosen ciphertext attack via the following game [4].

Setup: As described above.

Phase 1: As described above.

Challenge: The adversary submits two identities ID_0, ID_1 and a message m^* . ID_0, ID_1 or their prefix must not have appeared in any key generation query in Phase 1. The challenger selects a random bit $b \in \{0, 1\}$,

sets $c^* = \text{Encrypt}(\text{params}, ID_b, m^*)$, and sends c^* to the adversary as its challenge ciphertext.

Phase 2: This is identical to Phase 1, except that the adversary may not request a private key for ID_0, ID_1 or the decryption of $(ID_0, c^*), (ID_1, c^*)$.

Guess: The adversary submits a guess $b' \in \{0, 1\}$. The adversary wins if $b' = b$.

We call an adversary A in the above game an ANON-ID-CCA adversary. The advantage of an adversary A in this game is defined as $\Pr[b' = b] - \frac{1}{2}$.

Definition 2. An HIBE system is $(t, \varepsilon, q_e, q_d)$ ANON-ID-CCA secure if all t -time ANON-ID-CCA adversaries making at most q_e key generation queries and at most q_d decryption queries have advantage at most ε in winning the above game.

3 Hierarchical identity based encryption scheme

3.1 Set up

Let p be a large prime number, G_1, G_2 are groups of order p . $e : G_1 \times G_1 \rightarrow G_2$ is a bilinear map, g is a generator of $G_1, g_1 = g^\alpha$, where $\alpha \in Z_p^*$. l is the maximum number of levels in the HIBE. H is a hash function from $G_1^2 \times G_2^2$ to Z_p^* . The PKG randomly chooses $r_0 \in Z_p^*, h_i \in G_1, i = 0, 1, \dots, l$. The public parameters are $(g, g_1, r_0, H, h_i (i = 0, 1, \dots, l)), \alpha$ is the private key of PKG.

3.2 Key generation

To a user U with identity $ID = (ID_1, ID_2, \dots, ID_l) \in Z_p^l$, the PKG randomly chooses $r_i \in Z_p^*$, and computes

$$d_{0,i} = (h_0 g^{-r_0})^{\frac{1}{\alpha}} \cdot (\prod_{k=1}^i h_k^{ID_k})^{r_i}, d_{1,i} = g_1^{r_i},$$

$$d_{i+1,i} = h_{i+1}^{r_i}, \dots, d_{l,i} = h_l^{r_i},$$

so the private key of U is $d = (d_{0,i}, d_{1,i}, d_{i+1,i}, \dots, d_{l,i})$.

The private key can also be generated by its parent $(ID_1, ID_2, \dots, ID_{i-1})$ having the secret key $(d_{0,i-1}, d_{1,i-1}, d_{i,i-1}, \dots, d_{l,i-1})$. It computes:

$$d_{0,i} = d_{0,i-1} \cdot d_i^{ID_i} \cdot (\prod_{k=1}^i h_k^{ID_k})^t,$$

$$d_{1,i} = d_{1,i-1} \cdot g_1^{r_i}, d_{k,i} = d_{k,i-1} \cdot h_k^{r_i} (k = i + 1, \dots, l),$$

where $r_i = r_{i-1} + t$.

3.3 Encryption

To encrypt a message $m \in G_2$ for the user with identity $ID = (ID_1, \dots, ID_l)$, randomly choose $s \in Z_p^*$ and compute

$$c_1 = (\prod_{k=1}^i h_k^{ID_k})^s, c_2 = e(g, g)^s, c_3 = g_1^s,$$

$$c_4 = m \cdot e(g, h_0)^s, c_5 = h_1^s \cdot h_2^{s\beta},$$

where $\beta = H(c_1, c_2, c_3, c_4)$.

The ciphertext is $c = (c_1, c_2, c_3, c_4, c_5)$.

Notice that encryption does not require any pairing computations once $e(g, g), e(g, h_0)$ have been pre-computed.

3.4 Decryption

The receiver computes $\beta = H(c_1, c_2, c_3, c_4)$, and verifies whether $e(g_1, c_5) = e(c_3, h_1 h_2^\beta)$. Then he decrypts $c_4 \cdot \frac{e(d_{1,i}, c_1) c_2^{-r_0}}{e(c_3, d_{0,i})} = m$.

4 Analysis of security

4.1 Correctness

$$(1) e(g_1, c_5) = e(g_1, h_1^s h_2^{s\beta}) = e(c_3, h_1 h_2^\beta)$$

$$(2) c_4 \cdot \frac{e(d_{1,i}, c_1) c_2^{-r_0}}{e(c_3, d_{0,i})}$$

$$= m \cdot e(g, h_0)^s \cdot \frac{e(g_1^{r_i}, \prod_{k=1}^i h_k^{ID_k})^s e(g, g)^{-s r_0}}{e(g_1^{r_i}, (h_0 g^{-r_0})^{\frac{1}{\alpha}} \cdot (\prod_{k=1}^i h_k^{ID_k})^{r_i})}$$

$$= m \cdot e(g, h_0)^s \cdot \frac{1}{e(g^s, h_0)}$$

$$= m$$

4.2 Indistinguishability of ciphertext

Theorem 1 Assume that the q -ABDHE problem is (t', ε') -difficult in group G_1 , then the encryption scheme is $(t, \varepsilon, q_e, q_d)$ -IND-ID-CCA2, where $t = t' - (q_e + q_d)t_{ave}, \varepsilon = \varepsilon' + \frac{1}{2}, t_{ave}$ is the average time of querying oracles.

Proof. Assume A is an IND-ID-CCA adversary, B is a challenger. At the beginning of the game, B is given a tuple $(g, g^\alpha, \dots, g^{\alpha^q}, g', g'^{\alpha^{q+2}}, T)$ to decide whether $T = e(g, g')^{\alpha^{q+1}}$.

Set Up. B randomly chooses $f(x) \in Z_p[x]$ of degree q with $f(0) \neq 0$, and computes $g(x) = \frac{f(x)-f(0)}{x}$. Let $g_1 = g^\alpha, h_0 = g^{f(\alpha)}, r_0 = f(0), h_i = g_1^{a_i} (i = 1, 2, \dots, l), a_i \in Z_p^*$ is a random number. H is a hash function from $G_1^2 \times G_2^2$ to Z_p^* . The public parameters are $(g, g_1, r_0, H, h_0, h_1, \dots, h_l)$.

Phase 1.

Key generation query. A sends identity $ID = (ID_1, ID_2, \dots, ID_l)$ to B .

B randomly chooses $r_i \in Z_p^*$, and computes

$$d_{0,i} = g^{g(\alpha)} \cdot (\prod_{k=1}^i h_k^{ID_k})^{r_i}, d_{1,i} = g_1^{r_i},$$

$$d_{i+1,i} = h_{i+1}^{r_i}, \dots, d_{l,i} = h_l^{r_i}.$$

It is a valid private key, where

$$d_{0,i} = g^{g(\alpha)} \cdot (\prod_{k=1}^i h_k^{ID_k})^{r_i}$$

$$= g^{\frac{f(\alpha)-f(0)}{\alpha}} \cdot (\prod_{k=1}^i h_k^{ID_k})^{r_i}$$

$$= (h_0 g^{-r_0})^{\frac{1}{\alpha}} \cdot (\prod_{k=1}^i h_k^{ID_k})^{r_i}.$$

Decryption query. A sends (ID, c) to B .

B first executes the key generation query to identity ID , then decrypts c with the private key of identity ID .

Challenge. A chooses (ID^*, m_0, m_1) to B , where $ID^* = (ID_1^*, ID_2^*, \dots, ID_l^*)$ and ID^* or its prefix must not have appeared in any key generation query in Phase 1.

B chooses $m_b, b \in \{0, 1\}$, let $s = \log_g g' \cdot \alpha^{q+1}$, and computes

$$c_1^* = \prod_{k=1}^i (g'^{\alpha^{q+2}})^{a_k ID_k}, c_2^* = T,$$

$$c_3^* = g'^{\alpha^{q+2}}, c_4^* = m_b \cdot \frac{e(c_3^*, d_{0,i}^*)}{e(d_{1,i}^*, c_1^*) c_2^{*-r_0}},$$

$c_5^* = g'^{\alpha^{q+2}(a_1+a_2\beta^*)}$, where
 $\beta^* = H(c_1^*, c_2^*, c_3^*, c_4^*)$, d_{0,i^*}, d_{1,i^*} is the private key of ID^* .

If $T = e(g, g')^{\alpha^{q+1}}$, $c^* = (c_1^*, c_2^*, c_3^*, c_4^*, c_5^*)$ is a valid ciphertext. Otherwise, it is an invalid ciphertext.

Phase 2. A executes key generation oracle to ID and decryption oracle as phase 1, except that the adversary may not request a private key for ID^* and its prefix or the decryption of (ID^*, c^*) .

Guess. A submits a guess $b' \in \{0, 1\}$.

Executing the game many times, where q_e, q_d are the number of queries to key generation oracle and decryption oracle respectively. If $Pr(b' = b) = \varepsilon > \frac{1}{2}$, then B has advantage at least ε' in solving the q-ABDHE problem, where $\varepsilon' = \varepsilon - \frac{1}{2}$.

Remark If $T = e(g, g')^{\alpha^{q+1}}$,

$$c_1^* = \prod_{k=1}^i (g'^{\alpha^{q+2}})^{a_k ID_k} = (\prod_{k=1}^i h_k^{ID_k})^s,$$

$$c_2^* = T = e(g, g')^{\alpha^{q+1}} = e(g, g)^s,$$

$$c_3^* = g'^{\alpha^{q+2}} = g_1^s,$$

$$c_4^* = m_b \cdot \frac{e(c_3^*, d_{0,i^*})}{e(d_{1,i^*}, c_1^*) c_2^{*-r_0}} = m_b \cdot e(g, h_0)^s,$$

$$c_5^* = g'^{\alpha^{q+2}(a_1+a_2\beta^*)} = g_1^{s(a_1+a_2\beta^*)} = h_1^s h_2^{s\beta^*},$$

c^* is a valid ciphertext. Otherwise, it is an invalid ciphertext.

4.3 Anonymity of ciphertext

Theorem 2 Assume q-ABDHE problem is (t', ε') -difficult in group G_1 , then the encryption scheme is $(t, \varepsilon, q_e, q_d)$ -ANON-ID-CCA2, where $t = t' - (q_e + q_d)t_{ave}$, $\varepsilon = \varepsilon' + \frac{1}{2}$, t_{ave} is the average time of querying oracles.

Proof. Assume A is an ANON-ID-CCA adversary, B is a simulator. At the beginning of the game, given B a tuple $(g, g^\alpha, \dots, g^{\alpha^q}, g', g'^{\alpha^{q+2}}, T)$ to decide whether $T = e(g, g')^{\alpha^{q+1}}$.

Set Up. As presented in theorem 1.

Phase 1. As presented in theorem 1.

Challenge. A sends (ID_0, ID_1, m^*) to B , where ID_0, ID_1 or their prefix must not have appeared in any key generation query in Phase 1.

B chooses $ID_b, b \in \{0, 1\}$, let $s = \log_g g' \cdot \alpha^{q+1}$,

$$c_1^* = \prod_{k=1}^i (g'^{\alpha^{q+2}})^{a_k ID_{b,k}}, c_2^* = T, c_3^* = g'^{\alpha^{q+2}},$$

$$c_4^* = \frac{m^* \cdot e(c_3^*, d_{0,|ID_b|})}{e(d_{1,|ID_b|}, c_1^*) c_2^{*-r_0}}, c_5^* = g'^{\alpha^{q+2}(a_1+a_2\beta^*)},$$

where $\beta^* = H(c_1^*, c_2^*, c_3^*, c_4^*)$, $d_{0,|ID_b|}, d_{1,|ID_b|}$ is the private key of ID_b .

If $T = e(g, g')^{\alpha^{q+1}}$, $c^* = (c_1^*, c_2^*, c_3^*, c_4^*, c_5^*)$ is a valid ciphertext. Otherwise, it is an invalid ciphertext.

Phase 2. A executes key generation oracle to ID and decryption oracle as phase 1, except that the adversary may not request the private key of ID_0, ID_1 and the decryption of $(c^*, ID_0), (c^*, ID_1)$.

Guess. A submits a guess $b' \in \{0, 1\}$.

Executing the game many times, where q_e, q_d are the number of queries to key generation oracle and decryption

oracle respectively. If $Pr(b' = b) = \varepsilon > \frac{1}{2}$, then B has advantage at least ε' in solving the q-ABDHE problem, where $\varepsilon' = \varepsilon - \frac{1}{2}$.

4.4 Efficiency

In the following table, we compare the efficiency of the known HIBE schemes in the standard model.

Scheme	Security model	Public key size
BB[5]	sID	l+3
BBG[7]	sID	l+3
CS[11]	full	h+l+3
ALYW[10]	wrong	2l+1
BW[12]	full	$2l^2 + 6l + 5$
Our	full	l+5
Private key size	Ciphertext size	Pairing operation
i+1	i+2	i+1
l-i+2	3	2
i+1	i+1	i+1
l-i+2	4	2
$3l^2 + 14l - li$ $-3i + 15$	2l+6	2l+5
l-i+2	5	4

Table 1: Comparison to other HIBE schemes.

In this table, i represents the number of levels of identity on which the operations are performed, l is the maximum number of levels in the HIBE. $\sigma = \max(2q, 2^{i/h})$, where $1 \leq h \leq i, q$ is the number of queries to oracles. "sID, full" denote selective-ID and adaptive-ID model respectively and "wrong" denotes the security proof is wrong.

We conclude that our HIBE scheme has short parameters, small computation and a tight reduction simultaneously from the table.

5 Conclusion

In this paper, we propose a constant size anonymous HIBE scheme that is fully secure in the standard model. The ciphertext size is independent of the level of the hierarchy. Moreover, our scheme has short parameters, high efficiency and a tight reduction. Our scheme is based on the q-ABDHE problem, an interesting problem is to construct an anonymous HIBE scheme that is fully secure based on a more standard assumption.

Acknowledgement

The work described in this paper was supported by the National Science Foundation of China under Grant (No.6057303), and also funded by the Program for New

Century Excellent Talents in University by Ministry of Education (NCET-05-0398). We also thank anonymous referees of this paper for their constructive comments.

References

- [1] A. Shamir.(1984) Identity-Based Cryptosystems and Signature Schemes. In *Advances in Cryptology-CRYPTO 1984*, volume 196 of LNCS, Springer-Verlag, California, USA, pp. 47-53.
- [2] B. Waters.(2005) Efficient Identity-Based Encryption without Random Oracles. In *Advances in Cryptology-Eurocrypt 2005*, volume 3494 of LNCS, Springer-Verlag, Aarhus, Denmark, pp. 114-127.
- [3] C. Gentry and A. Silverberg.(2002) Hierarchical ID-Based Cryptography. In *Advances in Cryptology-ASIACRYPT 2002*, volume 2501 of LNCS, Springer-Verlag, Queenstown, New Zealand, pp. 548-566.
- [4] C. Gentry.(2006) Practical identity-based encryption without random oracles. In *Advances in Cryptology-EUROCRYPT 2006*, volume 4404 of LNCS, Springer-Verlag, Saint Petersburg, Russia, pp. 445-464.
- [5] D. Boneh and X. Boyen.(2004) Efficient Selective-ID Identity Based Encryption without Random Oracles. In *Advances in Cryptology- Eurocrypt 2004*, volume 3027 of LNCS, Springer-Verlag, Interlaken, Switzerland, pp. 223-238.
- [6] D. Boneh, C. Gentry, and B. Waters.(2005) Collusion-Resistant Broadcast Encryption with Short Ciphertexts and Private Keys. In *Advances in Cryptology-Crypto 2005*, volume 3621 of LNCS, Springer-Verlag, California, USA, pp. 258-275, .
- [7] D. Boneh, X. Boyen, E. J. Goh.(2005) Hierarchical Identity Based Encryption with Constant Size Ciphertext. In *Advances in Cryptology-EUROCRYPT 2005*, volume 3493 of LNCS, Springer-Verlag, Aarhus, Denmark, pp. 440-456.
- [8] D.Boneh and M.Franklin.(2001) Identity-Based Encryption from the Weil Pairing. In *Advances in Cryptology-CRYPTO 2001*, volume 2139 of LNCS, Springer-Verlag, California, USA, pp. 213-229, .
- [9] J. Horwitz and B. Lynn.(2002) Toward Hierarchical Identity-Based Encryption. In *Advances in Cryptology-EUROCRYPT 2002*, volume 2332 of LNCS, Springer-Verlag, Amsterdam, The Netherlands, pp. 466-481.
- [10] M. H. Au, J. K. Liu, T. H. Yuen, and D. S. Wong.(2006) Practical Hierarchical Identity Based Encryption and Signature schemes Without Random Oracles. <http://eprint.iacr.org/2006/368>
- [11] S. Chatterjee, P. Sarker.(2006) On Hierarchical Identity Based Encryption Protocols with Short Public Parameters. <http://eprint.iacr.org/2006/279>
- [12] X. Boyen and B. Waters.(2006) Anonymous Hierarchical Identity-Based Encryption (Without Random Oracles). In *Advances in Cryptology-CRYPTO 2006*, volume 4117 of LNCS, Springer-Verlag, California, USA, pp. 290-307.

An Intelligent System for Classification of the Communication Formats Using PSO

Ata E. Zadeh Shermeh and Reza Ghaderi
 Faculty of Electrical and Computer Engineering,
 Babol Noshirvani University of Technology, Iran
 E-mail: ataebrahim@yahoo.com, r.ghaderi@nit.ac.ir

Keywords: classification of communication signals, particle swarm, higher order statistics, supervised learning

Received: May 23, 2007

Text Automatic identification of digital signal types is of interest for both civil and military applications. This paper presents an efficient signal type identifier that includes a variety of digital signals. In this method, a combination of higher order moments (HOM) and higher order cumulants (HOC) are used as the features. A multi-layer perceptron neural network with SASS learning algorithm is proposed to determine the membership of the received signal. We have used swarm intelligence (SI) for feature selection in order to reduce the complexity of the classifier. Simulations results show that the proposed method has high performance for identification of different kinds of digital signal even at very low SNRs. This high efficiency is achieved with only seven features, which have been selected using particle swarm optimizer.

Povzetek: Opisana je metoda za identifikacijo digitalnih signalov.

1 Introduction

Automatic signals type recognition is an important topic for both the civil and military domain. Signal type recognition classification is also believed to play an important part in future 4G software radios [1]. The general idea behind the software radio architecture is to perform a considerable amount of the signal processing in software instead of it being defined in hardware. This would enable the radio to adapt to a changing environment and user requirements by simply updating the software or by using adaptable software systems. In such scenarios, a broadcaster could for example change to appropriate modulation schemes according to the capacity of the channel. A receiver incorporating automatic modulation recognition could then handle this in real times.

Automatic signal classification methods, usually, divided two principle techniques. One is the decision theoretic approach and the other is pattern recognition. Decision theoretic approaches use probabilistic and hypothesis testing arguments to formulate the recognition problem [2-3]. These methods suffer from their very high computational complexity, difficulty to implementation and lack of robustness to model mismatch [7]. Pattern recognition approaches, however, do not need such careful treatment. They are easy to implement. They can be further divided into two subsystems: the feature extraction subsystem and the classifier subsystem. The former extracts prominent characteristics from the raw data, which are called features, and the latter is a classifier [4-14].

In [9], for the first time, Ghani and Lamontagne proposed using the multi-layer perceptron (MLP) neural

network with back-propagation (BP) learning algorithm for automatic signal type identification. They showed that neural network classifier outperforms other classifiers such as K-Nearest Neighbor (KNN). In [10], power spectral density (PSD) measurements were used in conjunction with neural networks to identify the signal's type. The approach was to transform the signal into its power spectrum and then to use that power spectrum as input to a backpropagation neural network. This approach worked well for signals of interest whose power content distinctively varied with changes in frequency. It did not work as well with signal types like PSK. In [11], the authors introduced two classifiers: neural network classifier and fixed threshold classifier, for analog and digital modulation recognition. They showed that the neural network classifier has a higher performance than the threshold classifier. The overall success rate is over 96% at the SNR of 15 dB. In [12], the authors proposed an identifier for the identification of PSK2, PSK4, PSK8, OQPSK, MSK, QAM16, QAM32, FSK2 and FSK4 signal types. The features chosen to characterize the signal types are the mean and the next three moments of the instantaneous characteristics. They used different classifiers and showed that the artificial neural network has better performance than K-Nearest Neighbor (KNN) classifier and the well known binary decision trees. They reported a success rate of 90% with SNR ranges 15-25 dB. However, the performance for lower SNRs is reported to be less than 80%. In [13], the authors proposed an identifier based on cyclic spectral features for identification of AM, USB, LSB, FM, ASK, FSK, BPSK, QPSK and SQPSK. It was claimed that

cyclic spectrum posses more advantage than power spectrum in signal type recognition. A full-connected backpropagation neural network is used for classification in that research. The success rate of this identifier is reported around 90% with SNR ranges 5-25 dB. In [14], the authors have used a combination of the symmetry, fourth order cumulants and power moments of the received signals as the features for identification of PSK2, PSK4, PSK8, QAM16 and QAM32. The classifier was a modified MLP neural network (with few output nodes). They reported a success rate about 92% at SNR of 8dB. In [15], the authors have introduced an identifier based on discrete wavelet decompositions and adaptive network based fuzzy interference system for recognition of ASK8, FSK8, PSK8 and QASK8.

Most of the methods can only recognize a few types of digital signal and/or lower order of digital signal types. They usually need high SNRs in order to achieve the minimum acceptable performance (80%). Basically, this is due to the classifier and the features that are used. Finding the suitable features is a very important step for recognition of these digital signals. From the published works it can be found that the methods, which use the higher order statistical features, are able to include the digital signal types such as QAM and higher orders of digital signals [6-12]. In this paper we have used a combination of the higher order moments and higher order cumulants (up to eighth) as the effective features for recognition of the considered digital signals. The identifiers that use artificial neural networks (ANNs) as the classifier have high performances [12-14]. In this paper we have used a multi-layer perceptron (MLP) neural network with self adaptive step-size (SASS) learning algorithm as the classifier [15]. We have used particle swarm optimization, in order to reduce the complexity of the proposed identifier.

The paper is organizes as follows. Section 2 describes the feature extraction module. Section 3 presents the classifier. Optimization module is introduced in Section 4. Section 5 shows some simulation results. Finally, Section 6 concludes the paper.

2 Feature extraction

A typical pattern recognition system after doing some pre-processing operations, often reduce the size of a raw data set by extracting some distinct attributes called features. The need for feature extraction comes to scene due to the possible inability to use the raw data. These features define a particular pattern. In the signal recognition area, choosing the good features, not only enable the classifier to distinguish more and higher digital signals, but also help to reduce the complexity of the classifier.

In this paper we have considered the following digital signals: ASK2, ASK4, ASK8, PSK2, PSK4, PSK8, QAM16, QAM32, QAM64, Star-QAM8, and V32. For simplifying the indication, the signals ASK2, ASK4, ASK8, PSK2, PSK4, PSK8, QAM16, QAM32, QAM64, Star-QAM8, and V32 are substituted with $P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8, P_9, P_{10},$ and P_{11} respectively. Different

types of the digital signal have different properties; therefore finding the proper features in order to identify them (especially in case of higher order and/or non-square) is a difficult task. Based on our researches, a combination of the higher order moments and higher order cumulants up to eighth make provide a fine way for discrimination of the considered digital signal types.

Probability distribution moments are a generalization of concept of the expected value, and can be used to define the characteristics of a probability density function [16]. The auto-moment of the random variable may be defined as follows:

$$M_{pq} = E[s^{p-q} (s^*)^q] \quad (1)$$

where p called moment order and s^* stands for complex conjugation. Now, consider a zero-mean discrete based-band signal sequence of the form $s_k = a_k + jb_k$. Using the definition of the auto-moment, the expressions for different order may be easily derived.

Consider a scalar zero mean random variable s . The symbolism for p^{th} order of cumulant is:

$$C_{pq} = Cum[\underbrace{s, \dots, s}_{(p-q)\text{terms}}, \underbrace{s^*, \dots, s^*}_{(q)\text{terms}}] \quad (2)$$

3 Classifier

We have used a MLP neural network as the classifier. A MLP neural network consists of an input layer of source nodes, one or more hidden layers of computation nodes (neurons) and an output layer [17]. The number of nodes in the input and the output layers depend on the number of input and output variables, respectively. In this paper a single hidden layer MLP neural network was chosen as the classifier. The issue of learning algorithm is very important for MLP. Backpropagation (BP) learning algorithm is still one of the most popular algorithms. In BP a simple gradient descent algorithm updates the weight values. However under certain conditions, the BP network classifier can produce non-robust results and easily converge to local minimum. Moreover it is time consuming in training phase. New algorithms have been proposed so far in order to network training. However, some algorithms require much computing power to achieve good training, especially when dealing with a large training set.

In this paper, SASS learning algorithm is used [17]. SASS is an adaptive step-size method. it is based on the bisection method for minimization in one dimension, in which the minimum of a valley is found by taking a step in the descent direction of half the previous step. The method has been adapted to allow the step-size to both increase and decrease. It uses the same update rule resilient but updates Δ_{ij} differently. It uses two previous signs and has an increment factor of 2.

Table2: Correct matrix of the identifier (without feature selection) at SNR=10dB.

	P ₁	P ₂	P ₃	P ₄	P ₅	P ₆	P ₇	P ₈	P ₉	P ₁₀	P ₁₁
P ₁	99										
P ₂		100									
P ₃			98.8								
P ₄				100							
P ₅					98						
P ₆						100					
P ₇							98.6				
P ₈								98			
P ₉									98.4		
P ₁₀										98.2	
P ₁₁											98.4

Table3: The performances of the identifier (without feature selection) at different SNR values.

SNR	Training	Testing
-5	85.45	83.76
0	89.85	88.52
5	96.75	96.55
10	98.90	98.85
15	99.25	99.14

Table4: The performances of identifier (without feature selection) at SNR=50dB.

	P ₁	P ₂	P ₃	P ₄	P ₅	P ₆	P ₇	P ₈	P ₉	P ₁₀	P ₁₁
P ₁	100										
P ₂		100									
P ₃			100								
P ₄				99.6							
P ₅					100						
P ₆						100					
P ₇							100				
P ₈								99.6			
P ₉									100		
P ₁₀										99.8	
P ₁₁											99.4

As mentioned the speed of learning algorithm is an important issue for MLP. The efficiency of the learning algorithm affects the amount of experimentation that can be done. To indicate the effectiveness of chosen learning algorithm (SASS), we have compared it with the SUPERSAB learning algorithm that is an adaptive learning rate algorithm [20]. In SUPERSAB learning algorithm each weight w_{ij} , connecting node j with node i , has its own learning rate that can vary in response to error surface. The system that uses MLP with SUPERSAB learning algorithm as the classifier is named as TECH2. Based our experiments, any number in the vicinity of 40 neurons seems to be adequate for reasonable classification. Other simulations setup is the

same. Table 5 shows the performance of TECH2. Table 8 presents the comparison between SPROI (that uses MLP with SASS learning algorithm as the classifier) and TECH2. It can be seen that the performance of the SPROI (that uses MLP with SUPERSAB learning algorithm as the classifier). It can be seen that SPROI has better performance than TECH2.

Table7: The performance of TECH2 at different levels of SNR.

SNR	Training	Testing
-5	73.46	72.35
0	82.42	80.24
5	86.70	85.42
10	94.64	92.45
15	96.25	94.65

Table6: Comparison between SPROI and TECH2: the testing performance and the number of epochs that are needed.

SNR	SASS		SUPERSAB	
	Passed Epochs	Testing	Passed Epochs	Testing
-5	725	83.76	1245	72.35
0	324	88.52	685	80.24
5	565	96.55	1575	85.42
10	364	98.85	978	92.45
15	485	99.14	1245	94.65

5.2 Performance of the identifier with using the optimizer

We have tested the identifier using several features selected using PSO. The selection of the PSO parameters plays an important role in the optimization.

Table 7 shows the performances of the identifier using four features selected by PSO. Table 8 shows this performance using seven features selected by PSO. Table 9 show the performance of the identifier using eighteen features selected by PSO. Also in these tables the performances of identifier without feature selection is showed. It can be seen that in Table 8 the identifier records a performance degradation less than 1% only at SNR= -5dB. For other levels of SNR, the difference is negligible. Thus it can be said that the proposed method achieves high performance on most SNR values with only seven features that have been selected using PSO. Tables 10-11 show the correct matrices of identifier at SNR= 5 dB and SNR=10 with seven features that have been selected by PSO. It is found that proposed is able to identify the different types of digital signal with high accuracy only seven selected features using PSO.

Table7: Testing performance (TP) of the identifier with five features selected using PSO.

SNR	TP with applying PSO	TP without PSO
-5	80.15	83.76
0	83.25	88.85
5	91.20	96.55
10	94.68	98.85
15	97.51	99.14

Table8: Testing performance (TP) of the identifier with seven features selected using PSO.

SNR	TP with applying PSO	TP without PSO
-5	82.90	83.76
0	88.25	88.85
5	96.18	96.55
10	98.50	98.85
15	99.02	99.14

Table9: Testing performance (TP) of the identifier with twenty features selected using PSO.

SNR	TP with applying PSO	TP without PSO
-5	83.45	83.76
0	88.60	88.85
5	96.25	96.55
10	98.65	98.85
15	99.08	99.14

Table12: Correct matrix of PROI with only seven features selected using PSO at SNR=5dB.

	P ₁	P ₂	P ₃	P ₄	P ₅	P ₆	P ₇	P ₈	P ₉	P ₁₀	P ₁₁
P ₁	97										
P ₂		97									
P ₃			96.5								
P ₄				100							
P ₅					95						
P ₆						97					
P ₇							92.5				
P ₈								97			
P ₉									94		
P ₁₀										97.5	
P ₁₁											94.5

Table13: Correct matrix of PROI with only seven features selected using PSO at SNR=10dB.

	P ₁	P ₂	P ₃	P ₄	P ₅	P ₆	P ₇	P ₈	P ₉	P ₁₀	P ₁₁
P ₁	99										
P ₂		100									
P ₃			99								
P ₄				100							
P ₅					97						
P ₆						100					
P ₇							98				
P ₈								98			
P ₉									97.5		
P ₁₀										97	
P ₁₁											98

As mentioned the features play a vital role in signal identification. In order to indicate the effectiveness of the chosen features, we have used the features that have been introduced in [7]. We name this identifier as THECH3. Other simulations setup is the same. Figure1 show a

comparison between TECH3 and our proposed identifier (PROI) at different SNRs. Results imply that our chosen features have effective properties in signal representation.

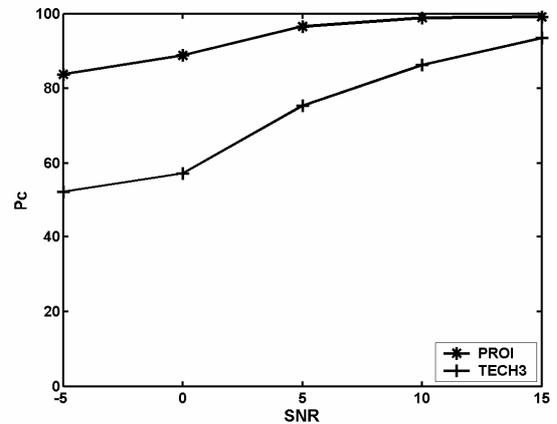


Figure1: Performances comparison between PROI and TECH3.

6 Conclusions

Automatic signal type identification is an important issue for both the civilian and military domain. Most of the techniques can only identify a few kinds of digital signal and/or lower orders of digital signals. They usually need high SNRs for identification of the considered digital signals. These problems are mainly due to two facts: the features and the classifier that are used in the identifiers. In this paper we have proposed a number of the features, i.e. a combination of the higher order moments and the higher order cumulants (up to eighth), which have high ability in representing of the digital signal types. As the classifier, we have proposed a MLP neural network with SASS learning algorithm as the classifier. This classifier has high classification ability. By using the mentioned features and classifier, we have presented a highly efficient identifier. This identifier discriminates a lot of digital signal types with high accuracy even at very low SNRs. But there are a lot of features are used for this identification. In order to reduction the complexity of the proposed identifier we have used an optimizer, i.e. PSO. Using this idea reduces the number of features without trading off the generalization ability and accuracy. The optimized identifier also has high performance for identification of the considered different kinds of digital signal even all SNRs. This high efficiency is achieved with only seven features, which have been selected using particle swarm optimizer. One the advantage of the swarmed features selection is its simple implementation. Furthermore in PSO to switch from one approach to another approach is also easy. For future works, we can use the PSO in order to construct of the classifier as well as the features selection. Also we can consider another set of digital signal types and evaluate this technique for identification of them. We can select the proper features that introduced by others and combine them with the features that are proposed in this paper in order to have suitable

features set for identification the different types of digital signal.

References

- [1] K.E. Nolan, L. Doyle, P. Mackenzie, D. O. Mahony (2002). Modulation Scheme Classification for 4G Software Radio Wireless Network. *Proc. IASTED*.
- [2] W. Wei and J. M. Mendel, "Maximum-likelihood classification for digital amplitude-phase modulations," *IEEE Trans. Commun.*, vol. 48, pp. 189-193, 2000.
- [3] P. Panagiotou, and A. Polydoros, "Likelihood ratio tests for modulation classifications," *Proc. MILCOM*, pp. 670-674, 2000.
- [4] S. Z. Hsue, and S. S. Soliman, Automatic modulation classification using Zero-Crossing, *IEEE Proc. Radar, Sonar and Navigation*, vol. 137, 1990, pp. 459-464.
- [5] B. G. Mobasseri, "Digital modulation classification using constellation shape," *Signal Processing*, vol. 80, pp. 251–277, 2000.
- [6] J. Lopatka, and P. Macrej, "Automatic modulation classification using statistical moments and a fuzzy classifier," *Proc. ICSP*, pp.121-127, 2000.
- [7] A. Swami, and B. M. Sadler, "Hierarchical digital modulation classification using cumulants," *IEEE Trans. Comm.*, vol. 48, pp. 416–429, 2000.
- [8] C. M. Spooner, "Classification of cochannel communication signals using cyclic cumulants," *Proc. ASILOMAR*, pp. 531-536, 1995.
- [9] O. A. Dobre, Y. Bar-Ness, and W. Su, "Robust QAM modulation classification algorithm based on cyclic cumulants," *Proc WCNC*, pp. 745-748, 2004.
- [10] A. E. Shermeh, D. Movahedi, "Signal identification using an efficient method," *ICS 2007*.
- [11] C. M. Spooner, W. A. Brown, and G. K. Yeung, "Automatic radio-frequency environment analysis," *Proc. ASILOMAR*, pp. 1181-1186, 2000.
- [12] A. E. Shermeh, D. Movahedi, "Automatic digital modulation identification using an intelligent method," *ICS 2007*.
- [13] C. L. P. Sehier, "Automatic modulation recognition with a hierarchical neural network," *Proc. MILCOM*, pp. 111–115, 1993.
- [14] A.K. Nandi, E.E. Azzouz, "Algorithms for automatic modulation recognition of communication signals," *IEEE Trans. Comm.*, vol. 46, pp. 431–436, 1998.
- [15] E. Avci, D. Hanbay and A. Varol, "An expert disceret wavelet adaptive network based fuzzy inference system for digital modulation recognition," *Expert Systems with Applications*, 2006.
- [16] J. Hannan, M. bishop, "A class of fast artificial NN learning algorithms", Tech. Report, JMH-JMB 0/96, Dep. of Cybernetics, University of Reading, 1996.
- [17] C. L. Nikias and A. P. Petropulu, *Higher-Order Spectra Analysis: A Nonlinear Signal Processing Framework*. Englewood Cliffs, NJ: PTR Prentice-Hall, 1993.
- [18] S. Haykin, *Neural Networks: A Comprehensive Foundation*. New York: MacMillan, 1994.
- [19] R. C. Eberhart, and J. Kennedy, A new optimizer using particle swarm theory, *Proc. ISMMH S*, 1995, pp. 39-43.
- [20] Y. Shi and R. C. Eberhart, A Modified Particle Swarm Optimizer, *in Proc. IJCNN*, 1999, pp. 69-73.
- [21] T. Tollenaere, Supersab: fast adaptive bp with good scaling properties, *Neural Networks*, Vol.3, 1990, pp. 561-573.

A Global k -means Approach for Autonomous Cluster Initialization of Probabilistic Neural Network

Roy Kwang Yang Chang, Chu Kiong Loo and M.V.C. Rao
Multimedia University, Jalan Ayer Keroh Lama, 75450 Melaka, Malaysia
E-mail: kychang@mmu.edu.my

Keywords: probabilistic neural network, global k -means, condition monitoring

Received: May 14, 2007

This paper focuses on the statistical based Probabilistic Neural Network (PNN) for pattern classification problems with Expectation – Maximization (EM) chosen as the training algorithm. This brings about the problem of random initialization, which means, the user has to predefine the number of clusters through trial and error. Global k -means is used to solve this and to provide a deterministic number of clusters using a selection criterion. On top of that, Fast Global k -means was tested as a substitute for Global k -means, to reduce the computational time taken. Tests were done on both homoscedastic and heteroscedastic PNNs using benchmark medical datasets and also vibration data obtained from a U.S. Navy CH-46E helicopter aft gearbox (Westland).

Povzetek: Opisana je metoda nevronskih mrež.

1 Introduction

The proposed model in this paper uses PNN as our choice of neural network for pattern classification problems. The Probabilistic Neural Network was first introduced in 1990 by Specht [1] and puts the statistical kernel estimator [2] into the framework of radial basis function networks. [3] We then used EM to train the PNN for the simple fact that it can help reduce the number of neurons that were committed in the network. The proposed model can be used in the field of condition monitoring which is garnering more attention due to its perks of time and cost savings. That is the reason why more focus should be spent on the creation of a more error tolerant and accurate yet fast diagnostic model.

The EM method used as the training algorithm for the network has its advantages and disadvantages. In general it is hard to initialize and the quality of the final solution depends heavily on the quality of the initial solution. [4] Initialization of the number of clusters needed has to be done randomly by the user in a series of trial and error values. This brings about an unwanted stochastic nature in the model. Therefore, in order to build an autonomous and deterministic neural network, we opted to use Global k -means to help automatically find the optimum number of clusters based on minimizing the clustering error.

In section 2, the PNN model is briefly discussed followed by section 3 where the E-step and the M-step of the EM method is showed together with the flaws of EM. Section 4 details cluster initialization with a brief discussion on two methods of cluster determination, which is Global k -means and its variant, Fast Global k -means. Experiments on Westland and benchmark medical datasets were done in section 5 to compare

results between Global k -means and random initialization together with Global k -means and Fast Global k -means. Section 6 will conclude the paper.

2 Probabilistic neural network

Probabilistic Neural Network was introduced by Donald Specht in a series of two papers, namely “Probabilistic Neural Networks for Classification, Mapping or Associative Memory” in 1988 [5] and “Probabilistic Neural Networks” in 1990 [1]. This statistical based neural network uses Bayes theory and Parzen Estimators to solve pattern classification problems. The basic idea behind Bayes theory is that it will make use of relative likelihood of events and also a priori information, which in our case would be inter-class mixing coefficients. As for Parzen Estimators, it is a classical probability density function estimator.

Let us assume the dataset, X , will be partitioned into K number of subsets where $X = X_1 \cup X_2 \cup \dots \cup X_K$ and each subset having N_k number of sample size, it would also mean $\sum_{k=1}^K N_k = N$ where N is the size of our sample. This four-layer, feed forward, supervised learning neural network reserves the lowest layer as input neurons and accepts d -dimensional input vectors. Each dimension of the input vector is passed to its corresponding input neuron.

The second layer of the PNN calculates the Gaussian basis functions (GBFs). It takes the form of

$$\rho_{m,k}(X) = \frac{1}{(2\pi\sigma_{m,k}^2)^{\frac{d}{2}}} \exp\left(-\frac{\|X - \nu_{m,k}\|^2}{2\sigma_{m,k}^2}\right) \quad (1)$$

and this specifies the GBF for m -th cluster in the k -th class where $\sigma_{m,k}^2$ is the variance, $\nu_{m,k}$ is the cluster centroid and d represents the dimension of the input vector.

The third layer of the PNN is where the class conditional probability density function is estimated, given by the formula

$$f_k(X) = \sum_{m=1}^{M_k} \beta_{m,k} \rho_{m,k}(X) \quad (2)$$

where M_k is the number of clusters for class k and $\beta_{m,k}$ is the intra-class mixing coefficient that can be defined as below.

$$\sum_{m=1}^{M_k} \beta_{m,k} = 1 \quad (3)$$

The PNN model has a fourth layer which is used as a decision layer to choose the class with the highest probability. An inter-class mixing coefficient, α_k , will be used to increase the accuracy of the result. α_k is obtained by the inverse of its sample size, N_k . Therefore the summation of all α_k shall be bound to 1. o_k depicts the probability of the input vector being class k .

$$o_k = \alpha_k f_k(x) \quad (4)$$

$$decision = \arg(\max_{1 \leq k \leq K} o_k) \quad (5)$$

The advantage that PNN has is that it interprets the network's structure in probability density functions due to its statistical nature. On the downside, the number of nodes that is committed in the PNN can be extremely huge if the training dataset is large. This is because one neuron is created for each training pattern. This makes the PNN simply infeasible for large datasets. Therefore another training method that does not commit every training pattern as a node in the neural network should be used. And for this purpose, we have selected the Expectation-Maximization (EM) method.

3 Learning algorithm

In the learning algorithm, two parameters of the model are adjusted to obtain better results in classification. In each E-step and M-step, the mean and the variance parameter is constantly tweaked until the log posterior likelihood function shows minimal difference.

To calculate the new mean and variance values, EM deploys a weight parameter which is also adjusted after each step.

3.1 Expectation-Maximization

Expectation-Maximization (EM) [6] by Dempster et al. in 1977 is a powerful iterative procedure which converges to an ML estimate. Basically the EM method consists of two steps, namely the E-step and the M-step. Both steps will be iterated until the change in the log posterior likelihood function is minimal.

$$\log L_f = \sum_{k=1}^K \log f_k(X) \quad (6)$$

In the E-step, the missing or hidden data is estimated given the observed data and the current parameter estimate. It will use the PDF estimated in the second layer of the PNN as defined in Equation 1 together with intra-class mixing coefficient to estimate the weight parameter.

$$W_{m,k} = \frac{\beta_{m,k} \rho_{m,k}(X)}{\sum_{i=1}^{M_k} \beta_{i,k} \rho_{i,k}(X)} \quad (7)$$

Next comes the M-step that uses the data estimated in the E-step, the weight parameter, $W_{m,k}$, to form a likelihood function and determine the ML estimate of the parameter. It calculates the new values of the cluster centroid, $\nu_{m,k}$, the variance, $\sigma_{m,k}^2$, and the intra-class mixing coefficients, $\beta_{m,k}$, using the weight calculated from the E-step. The equations for the parameter updates are given as below.

$$\nu_{m,k} = \frac{\sum_{n=1}^{N_k} W_{m,k}(X) X}{\sum_{n=1}^{N_k} W_{m,k}(X)} \quad (8)$$

$$\sigma_{m,k}^2 = \frac{\sum_{n=1}^{N_k} W_{m,k}(X) \|X - \nu_{m,k}\|^2}{d \sum_{n=1}^{N_k} W_{m,k}(X)} \quad (9)$$

$$\beta_{m,k} = \frac{1}{N_k} \sum_{n=1}^{N_k} W_{m,k}(X) \quad (10)$$

The EM algorithm is guaranteed to converge to an ML estimate [7, 8], and the convergence rate of the EM algorithm is usually quite fast. [9] EM also produces lesser neurons than the traditional PNN by Donald Specht. Also another plus side to it is that it does not require computations of gradients or Hessians, thus reducing the computational complexity of the network. Though EM is a good choice for a training method, it is not autonomous. This is attributed to the fact that EM

requires initialization in the form of, number of clusters to be expected of the neural network. The initialization quality severely affects the final outcome of the network. In order to aid in this matter, a method called Global k -means will be chosen as a precursor to find out how many clusters are needed for a certain dataset before being fed into the PNN with EM for training.

4 Cluster initialization

Part of the problems faced by the model is determining the number of clusters needed prior to learning. This is usually inputted by the user through a series of trial and error values. Also the usage of random initialization does not provide deterministic results. Global k -means and Fast Global k -means can overcome these problems.

4.1 Global k -means

Introduced by A. Likas, N. Vlasis and J.J. Verbeek in the paper entitled “The Global k -means clustering algorithm” in 2003, the concept of clustering with Global k -means is partitioning the given dataset into M clusters so that a clustering criterion is optimized. The common clustering criterion is the sum of squared Euclidean distances between each data point and the cluster centroid.

$$E(M_1, \dots, M_k) = \sum_{i=1}^{N_k} \sum_{m=1}^{M_k} \|X_i - v_m\|^2 \quad (11)$$

Global k -means deploys the k -means algorithm to find locally optimal solutions by trying to keep the clustering error to a minimum. The k -means algorithm starts by placing the cluster center arbitrarily and at each step moves the cluster center with the aim to minimize the clustering error. The down side to this algorithm is that it is sensitive to the initial position of the cluster centers. To overcome this, k -means can be scheduled to run several times and each time with a different starting point. The gist of Global k -means is that instead of trying to find all cluster centers at once, it proceeds in an incremental fashion. Incremental in the sense that one cluster center is found at a time.

Assume a K -clustering problem is to be solved; the algorithm starts by solving for a 1-clustering problem and the placement of the cluster center in this instance would equal the centroid of the given dataset. The next step would be to add another cluster center at its optimal position, given, the first cluster center has already been found. To do this, N -executions of k -means algorithm will be executed with the initial positions of the cluster centers being the first cluster which was found when solving for a 1-clustering problem and the second cluster’s starting position will be at x_n where $1 \leq n \leq N$. The final answer for a 2-clustering problem will be the best solution from the N -executions of k -means algorithm. Let $(c_1(k), \dots, c_k(k))$ denote the final solution for the k -clustering problem. We will solve it iteratively which means solving a 1-clustering problem,

then a 2-clustering problem, until a $(k-1)$ -clustering problem and the solution of k -clustering problem can be solved by performing N -executions of k -means algorithm with starting positions of $(c_1(k-1), \dots, c_{(k-1)}(k-1), X_n)$. A simple pseudo code of it will be

Problem: to solve k -clustering problem for dataset, X

```

For  $i=1$  to  $k$ 
{
  If  $i = 1$  then
     $c_i =$  centroid of dataset,  $X$ 
  Else
    For  $j=1$  to  $N$ 
      Run  $k$ -means with initial values of
         $\{c_i, \dots, c_{i-1}, X_j\}$ 
    }
}
    
```

With the final solution, $(c_1(k), \dots, c_k(k))$, Global k -means has actually found solutions of all k -cluster problem where $k=1, \dots, K$ without needing any further computations. This assumption seems very natural: we expect that the solution of a k -clustering problem to be reachable (through local search) from the solution of a $(k-1)$ -clustering problem, once the additional center is placed at an appropriate position within the data set. [10] Alas, the downside is that the computational time of Global k -means can be rather long.

4.2 Fast Global k -means

Using this method will help reduce the computational time taken by the Global k -means algorithm. The core difference is that, Fast Global k -means does not perform N -executions of k -means algorithm with starting positions of $(c_1(k-1), \dots, c_{(k-1)}(k-1), X_n)$. Instead, what the algorithm does is to calculate the upper bound $E_n \leq E - b_n$ on the resulting error, E_n , for every instances of X_n . We define E as the error value of $(k-1)$ -clustering problem and b_n as

$$b_n = \sum_{j=1}^N \max(d_{k-1}^j - \|x_n - x_j\|^2, 0) \quad (12)$$

with d_{k-1}^j as the squared Euclidean distance between x_j and the cluster centroid which it belongs to. After obtaining the value of b_n , select the x_i that maximizes b_n and make it the new cluster centroid that will be added. This is because by maximizing the value of b_n , we are at the same time minimizing the E_n value which as stated is our error. The new cluster centroid, x_n , will allocate all data points which have a smaller squared Euclidean distance from x_n rather than from their previous cluster centroid d_{k-1}^j . In view of that, the reduced clustering error for all those reassigned data points is $d_{k-1}^j - \|x_n - x_j\|^2$. Then we execute the k -means algorithm to find the solution for k -clustering problem.

Since the k -means algorithm is guaranteed to decrease the clustering error at each step, $E - b_n$ upper bounds the error measure that will be obtained if we run the algorithm until convergence after inserting the new center at x_n (this is the error measure used in the Global k -means algorithm). [10]

5 Experiments and results

5.1 General description

First, a test is conducted using EM-based PNN with two types of initialization, random and Global k -means. The benchmark medical datasets together with the Iris dataset was used for this purpose. Then, a test between EM-based PNN with initialization from Global k -means and Fast Global k -means was done to observe the computational time and also the difference in classification performance. The benchmark medical datasets were used. Next were tests done on Westland vibration dataset using EM-based PNN with Global k -means and also tests between Global k -means and Fast Global k -means to observe its accuracy and computational time.

5.2 Comparative tests between random initialization and Global k -means

Tests on the benchmark medical datasets and the Iris dataset [11] were conducted to observe the effects of random initialization and using Global k -means to initialize the parameter values in EM. The medical datasets consist of data from Cancer, Dermatology, Hepato, Heart and Pima.

The Iris dataset consists of 150 samples and 4 input features. It was tested on PNN trained by EM with randomly initialized cluster centroids and EM with Global k -means initialization. Both the methods were executed in heteroscedastic PNN and in homoscedastic PNN. A ten-fold validation was used. The Iris dataset was set as a 10-clustering problem for Global k -means and the number of cluster centroids returned was based on minimizing the squared Euclidean distance between each data point in a cluster and its centroid. This was then used to set the cluster parameter for random initialization to help it get a better result and assume under similar conditions as the Global k -means.

The mean accuracy of the homoscedastic with random initialization is 96.29% whilst the heteroscedastic version reports 95.36% accuracy. But in both cases, they were outdone by the accuracy of EM with Global k -means initialization, whose mean accuracy was 97.86% and 95.71% respectively, for homoscedastic and heteroscedastic PNN. Although random initialization was fed with the number of clusters needed by Global k -means, Global k -means still had the better classification rate.

Table 1: Correct classification rates for the Iris dataset.

	Random initialization		Global k -means	
	Homo	Hetero	Homo	Hetero
Accuracy				
Min	95.71	94.29	-	-
Mean	96.29	95.36	97.86	95.71
Max	96.43	95.71	-	-

The Cancer dataset contains 569 samples with a 30 dimension size, whilst the Dermatology dataset contains 358 samples with a 34 dimension size and the Hepato dataset contains 536 samples with a 9 dimension size. The Heart dataset contains 270 samples with a 13 dimension size and two output labels, which are “0” for absence of heart disease and “1” for presence of heart disease. Pima data set is available from machines learning database at UCI [12]. The Pima dataset contains 768 samples with an eight dimension size and has two classes which are diabetes positive and diabetes negative. A ten-fold validation was employed. When running using all the above datasets, Global k -means was set with a higher than required clustering problem to solve and in every case it returns a lower number of clusters which is optimum to the clustering criterion. This was also fed into EM for random initialization.

Table 2: Correct classification rates for the medical datasets by using homoscedastic PNN.

Dataset	Random initialization			Global k -means
	Min	Mean	Max	Mean
Cancer	90.00	90.63	90.96	91.92
Dermatology	60.76	64.28	65.50	69.31
Hepato	37.35	38.51	39.18	39.39
Heart	62.40	63.52	64.40	58.80
Pima	70.29	71.07	71.43	71.29

Table 3: Correct classification rates for the medical datasets by using heteroscedastic PNN.

Dataset	Random initialization			Global k -means
	Min	Mean	Max	Mean
Cancer	94.23	94.52	94.62	95.38
Dermatology	86.87	88.05	89.08	89.54
Hepato	51.22	52.47	53.27	58.57
Heart	75.60	78.00	78.80	82.80
Pima	66.86	68.17	68.86	69.00

The Medical datasets showed improved performances of EM with Global k -means initialization, in both homoscedastic and heteroscedastic PNNs, over the usage of random initialization. Although in practice both were fed with the same number of clusters required, in most cases of the datasets, even the maximum accuracy from

the EM with random initialization is not higher than the mean of EM with initialization from Global k -means.

5.3 Comparative tests between Global k -means and Fast Global k -means

In order to minimize computational time without sacrificing the classification performance, we opted for the Fast Global k -means implementation. Below is a comparison between Global k -means and Fast Global k -means using both heteroscedastic and homoscedastic PNNs which were trained by the EM method. Tests were conducted on the medical datasets using a ten-fold validation and as usual, Global k -means was set to solve a higher clustering problem than required.

Table 4: Comparison of correct classification rates.

Dataset	Fast Global k -means		Global k -means	
	homo	hetero	homo	hetero
Cancer	92.69	94.23	91.92	95.38
Dermatology	68.70	93.51	69.31	89.54
Heart	68.80	79.60	58.80	82.80
Hepato	47.76	59.59	50.00	59.59
Pima	70.29	71.86	71.29	69.00

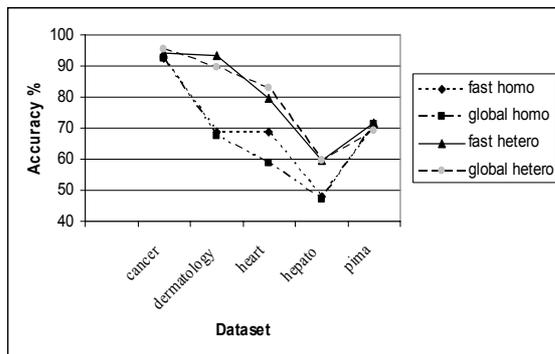


Figure 1: Comparison of correct classification rates.

Table 5: Comparison of computational times in seconds.

Dataset	Fast Global k -means		Global k -means	
	homo	hetero	homo	hetero
Cancer	5.80	14.83	563.20	622.95
Dermatology	11.69	20.95	849.34	950.77
Heart	2.03	5.69	71.08	93.00
Hepato	3.97	4.05	153.88	148.14
Pima	29.55	43.47	3299.11	3427.41

As shown in the results of Table 4 and Figure 1, Fast Global k -means provides a comparable correct classification rate on the benchmark medical datasets. On top of that, it still manages to accomplish its purpose which was to cut down on computational time. And Table 5 clearly supports this matter.

5.4 Westland Vibration Dataset

A real world case study was done; to test the EM trained PNN with initialization parameters obtained from the execution of Global k -means, using the popular benchmark dataset Westland [13]. This dataset consists of vibration time-series data which is gathered from an aft main power transmission of a U.S. Navy CH-46E helicopter by placing eight accelerometers at the known fault sensitive locations of the helicopter gearbox. The data was recorded for various faults including a no-defect case.

Table 6: Westland helicopter gearbox data description.

Fault type	Description
2	Planetary Bearing Corrosion
3	Input Pinion Bearing Corrosion
4	Spiral Bevel Input Pinion Spalling
5	Helical Input Pinion Chipping
6	Helical Idler Gear Crack Propagation
7	Collector Gear Crack Propagation
8	Quill Shaft Crack Propagation
9	No Defect

This dataset consists of 9 torque levels but for our experiment purposes, only the 100% torque level on sensors 1 to 4 is used. As the number of features from this dataset is quite substantial, feature reduction was needed. Wavelet packet feature extraction [14] was used to reduce the dimension of the input vectors without sacrificing too much of the classification performance.

Wavelet packets, a generalization of wavelet bases, are alternative bases that are formed by taking linear combinations of the usual wavelet functions. [15][16] These bases inherit properties such as orthonormality and time-frequency localization from their corresponding wavelet functions. [14] Wavelet packet functions can be defined as

$$W_{j,k}^n(t) = 2^{j/2} W^n(2^j t - k) \tag{13}$$

where n is the modulation or oscillation parameter, j is the index scale and k is the translation.

For a function, f , the wavelet packet coefficients can be calculated as below

$$w_{j,n,k} = \langle f, W_{j,k}^n \rangle = \int f(t) W_{j,k}^n(t) dt \tag{14}$$

In brief, the steps are; firstly, decompose the vibration signal using Wavelet Packet Transform (WPT) to extract out the time-frequency-dependant information. For each vibration signal segment, full decomposition is done up to the seventh level. This will produce a group of $2^{r+1} - 2$ sets of coefficients where r is the resolution level. Therefore, in our case it shall produce a group of

254 sets of coefficients where each set corresponds to a wavelet packet node. For the coefficients of every wavelet packet node, the wavelet packet node energy, $e_{j,n}$, is computed and this acts as the extracted feature.

$$e_{j,n} = \sum_k W_{j,n,k}^2 \tag{15}$$

After that, apply a statistical based feature selection criterion to help identify the features that provide the most discrimination amongst the classes of the dataset in focus, Westland. The Fisher’s criterion was used. [17] As a result, the number of features for Westland was reduced to eight and this modified dataset was fed into our model to test for classification rate by usage of Global k -means.

Table 7: Correct classification rates for Westland using homoscedastic and heteroscedastic PNNs.

Sensor	Accuracy	
	Hetero	Homo
1	96.06	86.06
2	94.51	88.45
3	95.92	87.89
4	95.21	91.41

The performance obtained by the proposed system on the 8-feature, 776-sample Westland dataset strengthens the positive performance that was marked in testing done on benchmark medical datasets.

We then performed further testing on the Westland dataset using Global k -means and its variant, Fast Global k -means. It was tested on both homoscedastic and heteroscedastic PNNs and again ten-fold validation was applied. As can be seen in Figure 2, the performance in terms of accuracy is comparable between the two methods. Not much accuracy degradation is shown by Fast Global k -means on the Westland sensor 1 to 4 data. Though comparable in terms of accuracy, the time taken by both methods is very different. Global k -means is a far slower method in comparison to the computational time of Fast Global k -means. This justifies our proposal of using Fast Global k -means with our model because though admittedly classification performance degrades, but it is by an acceptable margin and the time reduction is significant.

Table 8: Comparison of correct classification rates for Westland dataset.

Sensor	Fast Global k -means		Global k -means	
	homo	hetero	homo	hetero
1	84.93	96.48	86.06	96.06
2	87.46	93.38	88.45	94.51
3	86.76	95.77	87.89	95.92
4	90.99	94.79	91.41	95.21

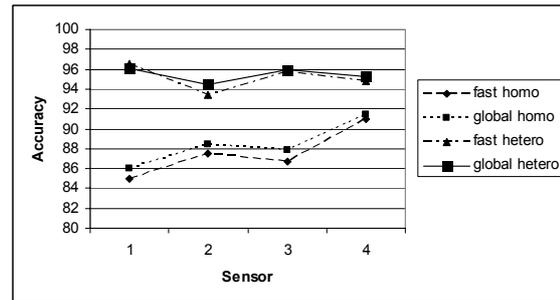


Figure 2: Comparison of correct classification rates.

Table 9: Comparison of computational times in seconds.

Sensor	Fast Global k -means		Global k -means	
	homo	hetero	homo	hetero
1	6.38	7.99	237.30	234.41
2	5.84	6.63	228.89	226.28
3	5.97	8.42	230.64	234.51
4	6.27	7.55	236.02	229.42

6 Conclusion

Although EM is a good choice to be used alongside PNN as its training algorithm, it has its downside. To rid of the stochastic nature that EM brings into our model, the Global k -means algorithm was used prior to EM to deterministically find the number of clusters based upon minimizing the clustering error. With this, the random trial and error values that the user was suppose to provide EM can be eliminated. Comparative test results indicated that even when set with the same number of clusters as Global k -means, EM with random initialization still had a poorer performance. This shows that EM with Global k -means initialization will help instil in the PNN model, autonomous and deterministic traits. We further try to improve the model by doing comparative tests between Fast Global k -means and Global k -means to observe their correct classification rates and their computational times. The results were favourable to Fast Global k -means as it provided relatively close accuracy and yet much improved computational time. Then EM-based PNN with Global k -means initialization was tested on Westland with positive results. Also tested on Westland was Fast Global k -means and Global k -means to determine the accuracy and timing differences. Results further justified the usage of Fast Global k -means in our model. The model presented in this paper is a pattern classifier that is both autonomous and deterministic. Possible application of it is as a diagnosis model that can be used in the business industry to monitor the condition of assets, such as machines, and to classify them into their fault modes based on the input vectors received from sensors placed on the machine.

References

- [1] Specht, D.F.: Probabilistic Neural Network. *Neural Networks* 3 (1990) 109-118.
- [2] E. Parzen.: On the estimation of a probability density function. *Annals of Mathematical Statistics* 3 (1962) 1065-1076.
- [3] Berthold, M.R., Diamond, J.: Constructive Training of Probabilistic Neural Networks. *Neurocomputing* 19 (1998) 167-183.
- [4] Ordonez, C., Omiecinski, E.: FREM: Fast and Robust EM Clustering for Large Data Sets. *CIKM* (2002).
- [5] Specht, D.F.: Probabilistic neural network for classification, mapping, or associative memory. *Proceedings of the IEEE International Conference on Neural Networks* 1 (1998) 525-532.
- [6] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society (B)* 39 (1977) 1-38.
- [7] Wu, C.: On the Convergence Properties of the EM Algorithm. *Annals Statistics* 11 (1983) 95-103.
- [8] Xu, L., Jordan, M.I.: On Convergence Properties of the EM Algorithm for Gaussian Mixtures. *Neural Computation* 8 (1996) 129-151.
- [9] Yang, Z.R., Chen, S.: Robust Maximum Likelihood Training of Heteroscedastic Probabilistic Neural Networks. *Neural Networks* 11 (1998) 739-747.
- [10] Likas, A., Vlassis, N., Verbeek, J.J.: The Global k -means Clustering Algorithm. *Pattern Recognition* 36 (2003) 451-461.
- [11] C. L. Blake, C. J. Merz: *UCI Repository of Machine Learning Databases*, University of California, Irvine, Department of Information and Computer Sciences (1998).
- [12] F. Zarndt: *A Comprehensive Case Study: An Examination of Machine Learning and Connectionist Algorithms*, MSc Thesis. Dept. of Computer Science, Brigham Young University (1995).
- [13] B. G. Cameron: *Final Report on CH-46 Aft Transmission Seeded Fault Testing*, Westland Helicopters, Ltd., U.K., Res. Paper RP907, 1993.
- [14] Yen, G.G., Lin, K.C.: Wavelet packet feature extraction for vibration monitoring. *IEEE Transactions on Industrial Electronics* 47 No. 3 (2000).
- [15] Coifman, R.R., Wickerhauser, M.V.: Entropy based algorithms for best basis selection. *IEEE Transactions on Information Theory* 38 (1992) 713-718.
- [16] Wickerhauser, M.V.: *Adapted wavelet analysis from theory to software*. Natick, MA: Wellesley (1994).
- [17] Fukunaga, K.: *Introduction to statistical pattern recognition*. New York: Academic (1992).

JOŽEF STEFAN INSTITUTE

Jožef Stefan (1835-1893) was one of the most prominent physicists of the 19th century. Born to Slovene parents, he obtained his Ph.D. at Vienna University, where he was later Director of the Physics Institute, Vice-President of the Vienna Academy of Sciences and a member of several scientific institutions in Europe. Stefan explored many areas in hydrodynamics, optics, acoustics, electricity, magnetism and the kinetic theory of gases. Among other things, he originated the law that the total radiation from a black body is proportional to the 4th power of its absolute temperature, known as the Stefan-Boltzmann law.

The Jožef Stefan Institute (JSI) is the leading independent scientific research institution in Slovenia, covering a broad spectrum of fundamental and applied research in the fields of physics, chemistry and biochemistry, electronics and information science, nuclear science technology, energy research and environmental science.

The Jožef Stefan Institute (JSI) is a research organisation for pure and applied research in the natural sciences and technology. Both are closely interconnected in research departments composed of different task teams. Emphasis in basic research is given to the development and education of young scientists, while applied research and development serve for the transfer of advanced knowledge, contributing to the development of the national economy and society in general.

At present the Institute, with a total of about 800 staff, has 600 researchers, about 250 of whom are postgraduates, nearly 400 of whom have doctorates (Ph.D.), and around 200 of whom have permanent professorships or temporary teaching assignments at the Universities.

In view of its activities and status, the JSI plays the role of a national institute, complementing the role of the universities and bridging the gap between basic science and applications.

Research at the JSI includes the following major fields: physics; chemistry; electronics, informatics and computer sciences; biochemistry; ecology; reactor technology; applied mathematics. Most of the activities are more or less closely connected to information sciences, in particular computer sciences, artificial intelligence, language and speech technologies, computer-aided design, computer architectures, biocybernetics and robotics, computer automation and control, professional electronics, digital communications and networks, and applied mathematics.

The Institute is located in Ljubljana, the capital of the independent state of Slovenia (or S^olnia). The capital today is considered a crossroad between East, West and Mediter-

anean Europe, offering excellent productive capabilities and solid business opportunities, with strong international connections. Ljubljana is connected to important centers such as Prague, Budapest, Vienna, Zagreb, Milan, Rome, Monaco, Nice, Bern and Munich, all within a radius of 600 km.

From the Jožef Stefan Institute, the Technology park "Ljubljana" has been proposed as part of the national strategy for technological development to foster synergies between research and industry, to promote joint ventures between university bodies, research institutes and innovative industry, to act as an incubator for high-tech initiatives and to accelerate the development cycle of innovative products.

Part of the Institute was reorganized into several high-tech units supported by and connected within the Technology park at the Jožef Stefan Institute, established as the beginning of a regional Technology park "Ljubljana". The project was developed at a particularly historical moment, characterized by the process of state reorganisation, privatisation and private initiative. The national Technology Park is a shareholding company hosting an independent venture-capital institution.

The promoters and operational entities of the project are the Republic of Slovenia, Ministry of Higher Education, Science and Technology and the Jožef Stefan Institute. The framework of the operation also includes the University of Ljubljana, the National Institute of Chemistry, the Institute for Electronics and Vacuum Technology and the Institute for Materials and Construction Research among others. In addition, the project is supported by the Ministry of the Economy, the National Chamber of Economy and the City of Ljubljana.

Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel.: +386 1 4773 900, Fax.: +386 1 251 93 85
WWW: <http://www.ijs.si>
E-mail: matjaz.gams@ijs.si
Public relations: Polona Strnad

INFORMATICA
AN INTERNATIONAL JOURNAL OF COMPUTING AND INFORMATICS
INVITATION, COOPERATION

Submissions and Refereeing

Please submit an email with the manuscript to one of the editors from the Editorial Board or to the Managing Editor. At least two referees outside the author's country will examine it, and they are invited to make as many remarks as possible from typing errors to global philosophical disagreements. The chosen editor will send the author the obtained reviews. If the paper is accepted, the editor will also send an email to the managing editor. The executive board will inform the author that the paper has been accepted, and the author will send the paper to the managing editor. The paper will be published within one year of receipt of email with the text in Informatica MS Word format or Informatica L^AT_EX format and figures in .eps format. Style and examples of papers can be obtained from <http://www.informatica.si>. Opinions, news, calls for conferences, calls for papers, etc. should be sent directly to the managing editor.

QUESTIONNAIRE

- Send Informatica free of charge
- Yes, we subscribe

Please, complete the order form and send it to Dr. Drago Torkar, Informatica, Institut Jožef Stefan, Jamova 39, 1000 Ljubljana, Slovenia. E-mail: drago.torkar@ijs.si

Since 1977, Informatica has been a major Slovenian scientific journal of computing and informatics, including telecommunications, automation and other related areas. In its 16th year (more than ten years ago) it became truly international, although it still remains connected to Central Europe. The basic aim of Informatica is to impose intellectual values (science, engineering) in a distributed organisation.

Informatica is a journal primarily covering the European computer science and informatics community - scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the Refereeing Board.

Informatica is free of charge for major scientific, educational and governmental institutions. Others should subscribe (see the last page of Informatica).

ORDER FORM – INFORMATICA

Name:	Office Address and Telephone (optional):
Title and Profession (optional):
.....	E-mail Address (optional):
Home Address and Telephone (optional):
.....	Signature and Date:

Informatica WWW:

<http://www.informatica.si/>

Referees:

Witold Abramowicz, David Abramson, Adel Adi, Kenneth Aizawa, Suad Alagić, Mohamad Alam, Dia Ali, Alan Aliu, Richard Amoroso, John Anderson, Hans-Jurgen Appelrath, Iván Araujo, Vladimir Bajič, Michel Barbeau, Grzegorz Bartoszewicz, Catriel Beerli, Daniel Beech, Fevzi Belli, Simon Beloglavec, Sondes Bennisri, Francesco Bergadano, Istvan Berkeley, Azer Bestavros, Andraž Bežek, Balaji Bharadwaj, Ralph Bisland, Jacek Blazewicz, Laszlo Boeszoermenyi, Damjan Bojadžijev, Jeff Bone, Ivan Bratko, Pavel Brazdil, Bostjan Brumen, Jerzy Brzezinski, Marian Bubak, Davide Bugali, Troy Bull, Sabin Corneliu Buraga, Leslie Burkholder, Frada Burstein, Wojciech Buszkowski, Rajkumar Bvyya, Giacomo Cabri, Netiva Caftori, Patricia Carando, Robert Cattral, Jason Ceddia, Ryszard Choras, Wojciech Cellary, Wojciech Chybowski, Andrzej Ciepiewski, Vic Ciesielski, Mel Ó Cinnéide, David Cliff, Maria Cobb, Jean-Pierre Corriveau, Travis Craig, Noel Craske, Matthew Crocker, Tadeusz Czachorski, Milan Češka, Honghua Dai, Bart de Decker, Deborah Dent, Andrej Dobnikar, Sait Dogru, Peter Dolog, Georg Dorfner, Ludoslaw Drelichowski, Matija Drobnič, Maciej Drozdowski, Marek Druzdzel, Marjan Družovec, Jozo Dujmović, Pavol Ďuriš, Amnon Eden, Johann Eder, Hesham El-Rewini, Darrell Ferguson, Warren Fergusson, David Flater, Pierre Flener, Wojciech Fliegner, Vladimir A. Fomichov, Terrence Forgarty, Hans Fraaije, Stan Franklin, Violetta Galant, Hugo de Garis, Eugeniusz Gatnar, Grant Gayed, James Geller, Michael Georgiopolus, Michael Gertz, Jan Goliński, Janusz Gorski, Georg Gottlob, David Green, Herbert Groiss, Jozsef Gyorkos, Marten Haglind, Abdelwahab Hamou-Lhadj, Inman Harvey, Jaak Henno, Marjan Hericko, Henry Hexmoor, Elke Hochmueller, Jack Hodges, John-Paul Hosom, Doug Howe, Rod Howell, Tomáš Hruška, Don Huch, Simone Fischer-Huebner, Zbigniew Huzar, Alexey Ippa, Hannu Jaakkola, Sushil Jajodia, Ryszard Jakubowski, Piotr Jedrzejowicz, A. Milton Jenkins, Eric Johnson, Polina Jordanova, Djani Juričič, Marko Juvancic, Sabhash Kak, Li-Shan Kang, Ivan Kapustok, Orlando Karam, Roland Kaschek, Jacek Kierzenka, Jan Kniat, Stavros Kokkotos, Fabio Kon, Kevin Korb, Gilad Koren, Andrej Krajnc, Henryk Krawczyk, Ben Kroese, Zbyszko Krolikowski, Benjamin Kuipers, Matjaž Kukar, Aarre Laakso, Sofiane Labidi, Les Labuschagne, Ivan Lah, Phil Laplante, Bud Lawson, Herbert Leitold, Ulrike Leopold-Wildburger, Timothy C. Lethbridge, Joseph Y-T. Leung, Barry Levine, Xuefeng Li, Alexander Linkevich, Raymond Lister, Doug Locke, Peter Lockeman, Vincenzo Loia, Matija Lokar, Jason Lowder, Kim Teng Lua, Ann Macintosh, Bernardo Magnini, Andrzej Małachowski, Peter Marcer, Andrzej Marciniak, Witold Marciszewski, Vladimir Marik, Jacek Martinek, Tomasz Maruszewski, Florian Matthes, Daniel Memmi, Timothy Menzies, Dieter Merkl, Zbigniew Michalewicz, Armin R. Mikler, Gautam Mitra, Roland Mittermeir, Madhav Moganti, Reinhard Moller, Tadeusz Morzy, Daniel Mossé, John Mueller, Jari Multisilta, Hari Narayanan, Jerzy Nawrocki, Rance Necaie, Elzbieta Niedzielska, Marian Niedq' zwiędziński, Jaroslav Nieplocha, Oscar Nierstrasz, Roumen Nikolov, Mark Nissen, Jerzy Nogiec, Stefano Nolfi, Franc Novak, Antoni Nowakowski, Adam Nowicki, Tadeusz Nowicki, Daniel Olejar, Hubert Österle, Wojciech Olejniczak, Jerzy Olszewski, Cherry Owen, Mieczyslaw Owoc, Tadeusz Pankowski, Jens Penberg, William C. Perkins, Warren Persons, Mitja Peruš, Fred Petry, Stephen Pike, Niki Pissinou, Aleksander Pivk, Ullin Place, Peter Planinšec, Gabika Polčicová, Gustav Pomberger, James Pomykalski, Tomas E. Potok, Dimithu Prasanna, Gary Preckshot, Dejan Rakovič, Cveta Razdevšek Pučko, Ke Qiu, Michael Quinn, Gerald Quirchmayer, Vojislav D. Radonjic, Luc de Raedt, Ewaryst Rafajlowicz, Sita Ramakrishnan, Kai Rannenberg, Wolf Rauch, Peter Rechenberg, Felix Redmill, James Edward Ries, David Robertson, Marko Robnik, Colette Rolland, Wilhelm Rossak, Ingrid Russel, A.S.M. Sajeev, Kimmo Salmenjoki, Pierangela Samarati, Bo Sanden, P. G. Sarang, Vivek Sarin, Iztok Savnik, Ichiro Satoh, Walter Schempp, Wolfgang Schreiner, Guenter Schmidt, Heinz Schmidt, Dennis Sewer, Zhongzhi Shi, Mária Smolárová, Carine Souveyet, William Spears, Hartmut Stadtler, Stanislaw Stanek, Olivero Stock, Janusz Stokłosa, Przemysław Stpicyński, Andrej Stritar, Maciej Stroinski, Leon Strous, Ron Sun, Tomasz Szmuc, Zdzislaw Szyjewski, Jure Šilc, Metod Škarja, Jiří Šlechta, Chew Lim Tan, Zahir Tari, Jurij Tasič, Gheorge Tecuci, Piotr Teczynski, Stephanie Teufel, Ken Tindell, A Min Tjoa, Drago Torkar, Vladimir Tomic, Wieslaw Traczyk, Denis Trček, Roman Trobec, Marek Tudruj, Andrej Ule, Amjad Umar, Andrzej Urbanski, Marko Uršič, Tadeusz Usowicz, Romana Vajde Horvat, Elisabeth Valentine, Kanonkluk Vanapipat, Alexander P. Vazhenin, Jan Verschuren, Zygmunt Vetulani, Olivier de Vel, Didier Vojtisek, Valentino Vranić, Jozef Vyskoc, Eugene Wallingford, Matthew Warren, John Weckert, Michael Weiss, Tatjana Welzer, Lee White, Gerhard Widmer, Stefan Wrobel, Stanislaw Wrycza, Tatyana Yakhno, Janusz Zalewski, Damir Zazula, Yanchun Zhang, Ales Zivkovic, Zonling Zhou, Robert Zorc, Anton P. Železnikar

Informatica

An International Journal of Computing and Informatics

Web edition of Informatica may be accessed at: <http://www.informatica.si>.

Subscription Information Informatica (ISSN 0350-5596) is published four times a year in Spring, Summer, Autumn, and Winter (4 issues per year) by the Slovene Society Informatika, Vožarski pot 12, 1000 Ljubljana, Slovenia.

The subscription rate for 2008 (Volume 32) is

- 60 EUR for institutions,
- 30 EUR for individuals, and
- 15 EUR for students

Claims for missing issues will be honored free of charge within six months after the publication date of the issue.

Typesetting: Borut Žnidar.

Printing: Dikplast Kregar Ivan s.p., Kotna ulica 5, 3000 Celje.

Orders may be placed by email (drago.torkar@ijs.si), telephone (+386 1 477 3900) or fax (+386 1 251 93 85). The payment should be made to our bank account no.: 02083-0013014662 at NLB d.d., 1520 Ljubljana, Trg republike 2, Slovenija, IBAN no.: SI56020830013014662, SWIFT Code: LJBASI2X.

Informatica is published by Slovene Society Informatika (president Niko Schlamberger) in cooperation with the following societies (and contact persons):

Robotics Society of Slovenia (Jadran Lenarčič)

Slovene Society for Pattern Recognition (Franjo Pernuš)

Slovenian Artificial Intelligence Society; Cognitive Science Society (Matjaž Gams)

Slovenian Society of Mathematicians, Physicists and Astronomers (Bojan Mohar)

Automatic Control Society of Slovenia (Borut Zupančič)

Slovenian Association of Technical and Natural Sciences / Engineering Academy of Slovenia (Igor Grabec)

ACM Slovenia (Dunja Mladenič)

Informatica is surveyed by: Citeseer, COBISS, Compendex, Computer & Information Systems Abstracts, Computer Database, Computer Science Index, Current Mathematical Publications, DBLP Computer Science Bibliography, Directory of Open Access Journals, InfoTrac OneFile, Inspec, Linguistic and Language Behaviour Abstracts, Mathematical Reviews, MatSciNet, MatSci on SilverPlatter, Scopus, Zentralblatt Math
--

The issuing of the Informatica journal is financially supported by the Ministry of Higher Education, Science and Technology, Trg OF 13, 1000 Ljubljana, Slovenia.

Informatica

An International Journal of Computing and Informatics

Introduction	J. Malačič, J. Józwiak, A. Fürnkranz-Prskawetz	101
Demographic Pressure on the Public Pension System	J. Sambt, M. Čok	103
Urban-Rural Life Setting as the Explanatory Factor of Differences in Fertility Behaviour in Slovenia	M. Černič Istenič, A. Kveder	111
Late Fertility Trends in Europe	J. Malacic	123
Methodology for the Estimation of Annual population stocks by citi-zenship group, age and sex in the EU and EFTA countries	J. Bijak, D. Kupiszewska	133
Demographic Analysis of Fertility Using Data Mining Tools	M. Gams, J. Krivec	147
Clustering of Population Pyramids	S. Korenjak-Černe, N. Kejžar, V. Batagelj	157
<hr/> End of Special Issue / Start of normal papers <hr/>		
Geometric-Invariant Image Watermarking by Key-Dependent Triangulation	S. Hu	169
A Simple Algorithm for the Restoration of Clipped Speech Signal	A. Dahimene, M. Noureddine, A. Azrar	183
eGovernance: Information and Communication Technology, Knowledge Management and Learning Organisation Culture	J. Berce, S. Lanfranco, V. Vehovar	189
Efficient Hierarchical Identity Based Encryption Scheme in the Standard Model	Y. Ren, D. Gu	207
An Intelligent System for Classification of the Communication formats Using PSO	A.E.Z. Shermeh, R. Ghaderi	213
A Global k-means Approach for Autonomous Cluster Initialization of Probabilistic Neural Network	R.K.Y. Chang, C.K. Loo, M.V.C. Rao	219

