STATISTIKA

STATISTICS
AND ECONOMY
JOURNAL

VOL. **95** (1) 2015

CZECH
STATISTICAL
OFFICE

EDITOR-IN-CHIEF

Stanislava Hronová
Prof., Faculty of Informatics and Statistics,
University of Economics, Prague
Prague, Czech Republic

EDITORIAL BOARD

Iva Ritschelová
President, Czech Statistical Office
Prague, Czech Republic

Marie Bohatá
Former President of the Czech Statistical Office
Prague, Czech Republic

Ľudmila Benkovičová
President, Statistical Office of the Slovak Republic
Bratislava, Slovak Republic

Roderich Egeler
President, German Federal Statistical Office
Wiesbaden, Germany

Richard Hindls
Deputy chairman of the Czech Statistical Council
Prof., Faculty of Informatics and Statistics,
University of Economics, Prague
Prague, Czech Republic

Gejza Dohnal
Vice-President of the Czech Statistical Society
Czech Technical University in Prague
Prague, Czech Republic

Štěpán Jurajda
Director, CERGE-EI: Center for Economic Research
and Graduate Education — Economics Institute
Prague, Czech Republic

Vladimír Tomšík
Vice-Governor, Czech National Bank
Prague, Czech Republic

Jana Jurečková
Prof., Department of Probability and Mathematical
Statistics, Charles University in Prague
Prague, Czech Republic

Jaromír Antoch
Prof., Department of Probability and Mathematical
Statistics, Charles University in Prague
Prague, Czech Republic

Martin Mandel
Prof., Department of Monetary Theory and Policy,
University of Economics, Prague
Prague, Czech Republic

František Cvengroš
Head of the Macroeconomic Predictions Unit,
Financial Policy Department,
Ministry of Finance of the Czech Republic
Prague, Czech Republic

Josef Plandor
Department of Analysis and Statistics,
Ministry of Industry and Trade of the Czech Republic
Prague, Czech Republic

Petr Zahradník
ČEZ, a.s.
Prague, Czech Republic

Kamil Janáček
Board Member, Czech National Bank
Prague, Czech Republic

Petr Vojtíšek
Deputy Director, Monetary and Statistics Department,
Czech National Bank
Prague, Czech Republic

Milan Terek
Prof., Department of Statistics,
University of Economics in Bratislava
Bratislava, Slovak Republic

Cesare Costantino
Head, Environmental Accounts and Satellite Accounting,
National Accounts
Italian National Statistical Institute
Rome, Italy

Walenty Ostasiewicz
Head, Department of Statistics,
Wroclaw University of Economics
Wroclaw, Poland

ASSOCIATE EDITORS

Jakub Fischer
Vice-Rector, University of Economics, Prague
Prague, Czech Republic

Luboš Marek
Dean of the Faculty of Informatics and Statistics,
University of Economics, Prague
Prague, Czech Republic

Marek Rojíček
Vice-President,
Czech Statistical Office
Prague, Czech Republic

Hana Řezanková
President of the Czech Statistical Society
Prof., Faculty of Informatics and Statistics,
University of Economics, Prague
Prague, Czech Republic

MANAGING EDITOR

Jiří Novotný
Czech Statistical Office
Prague, Czech Republic

CONTENTS

ANALYSES

INFORMATION

About Statistika

The journal of Statistika has been published by the Czech Statistical Office since 1964. Its aim is to create a platform enabling national statistical and research institutions to present the progress and results of complex analyses in the economic, environmental, and social spheres. Its mission is to promote the official statistics as a tool supporting the decision making at the level of international organizations, central and local authorities, as well as businesses. We contribute to the world debate and efforts in strengthening the bridge between theory and practice of the official statistics. Statistika is a professional double-blind peer reviewed journal included (since 2008) in the List of Czech non-impact peer-reviewed periodicals (updated in 2013). Since 2011 Statistika has been published quarterly in English only.

Publisher

The Czech Statistical Office is an official national statistical institution of the Czech Republic. The Office main goal, as the coordinator of the State Statistical Service, consists in the acquisition of data and the subsequent production of statistical information on social, economic, demographic, and environmental development of the state. Based on the data acquired, the Czech Statistical Office produces a reliable and consistent image of the current society and its developments satisfying various needs of potential users.

Contact us

Journal of Statistika | Czech Statistical Office | Na padesátém 81 | 100 82 Prague 10 | Czech Republic
e-mail: statistika.journal@czso.cz | web: www.czso.cz/statistika_journal

Accounting for Wealth in the Czech Republic[1]

Vítězslav Ondruš[2] | *Czech Statistical Office, Prague, Czech Republic*

Abstract

We often meet with the analysis of household final consumption expenditure, less frequently with analyzes of household wealth. Both categories are important for the characterization of the standard of living, both are provided by the system of national accounts, but each of them is quite a different phenomenon. Household wealth comes, besides capital transfers and other changes, mainly from accumulated part of disposable income not used for current expenditure on final consumption.

Keywords

Gross domestic product, savings, national wealth, assets, liabilities, households

JEL code

E01

INTRODUCTION

We often meet with the analysis of household final consumption expenditure[3], less frequently with analyses of household wealth. Both categories are important for the characterization of the standard of living, both are provided by the system of national accounts, but each of them is quite a different phenomenon. Household wealth comes, besides capital transfers and other changes, mainly from accumulated part of disposable income not used for current expenditure on final consumption.

Dependence of both categories, however, is mutual – the wealth of households comes not only from what was not consumed in the current year (about 11 % of disposable income), but also on the contrary, existing wealth directly affects the final consumption expenditure (essential part of imputed rent) or generates a large part of disposable income by some form of property income, e.g., rents, dividends, interest (about 20% of disposable income).

National accounts presents the economy in a form of a consistent model where flows are followed by the stocks, see Hronova et al. (2009).

1 NATIONAL WEALTH OF THE CZECH REPUBLIC

National wealth or net worth of the total economy and individual institutional sectors have been recorded in the Czech national accounts from the beginning of their compilation, since 1993. The original data simply reflected figures from business accounts, i.e., they matched the methodological principles of

[1] The article is based on the paper presented on the OECD Working Party on Financial Statistics, October 3rd, 2013, Paris. However, all figures used in the article correspond now to mew methodologies SNA2008/ESA2010. They are based on new time series published by the CZSO in October 2014: <http://apl.czso.cz/pll/rocenka/rocenka.indexnu>.

[2] Department of Annual National Accounts, Na padesátém 81, 100 82 Prague 10, Czech Republic. E-mail: vitezslav.ondrus@czso.cz.

[3] The issue of households consumption from long perspective was profoundly elaborated in Sixta et al. (2014).

business accounting and did not cover all of economic units, or in other words, they were not complete and methodologically correct (see National accounts for the Czech Republic, Paris 1998, OECD). Since this introduction of the system, the CZSO seeks for compilation with all methodological principles of the national accounts.

Gradually we are improving the quality and completeness of the quantification of the national wealth, or particular types of assets, and within each main/occasional revision of the national accounts data to perform retrospective estimates to ensure comparable time series. Yet, even after 20 years, after last main revision made in occasion with transition to new methodology SNA 2008/ESA 2010, this process is not completely finished. Some of the most significant weaknesses are further indentified.

The national wealth of the Czech Republic for the year 2012 recorded in the national accounts amounts to 27.9 trillions CZK, of which non-financial assets represent CZK 29.3 trillion and net financial worth shows negative value of CZK –1.4 trillion.

The system of national accounts provides a combined view on the national wealth. Besides the subject structure of the national wealth, the ownership and purpose have the same importance. In the Czech national accounts, there are created in addition to the standard system of sector accounts and input-output tables so called balances of non-financial assets that give three-dimensional views on each group of non-financial assets: asset x sector x industry.

However, an industrial structure of non-financial assets is analytically important for the non-financial corporations sector, while in the household sector it is relevant for sub-sector of entrepreneurs only. But, the article is focused further on the household sector that is why the industrial structure of non-financial assets will be omitted. The overview of subject structure of the net worth of all five institutional sectors for 2012 is provided in the Table 1.

Due to the different role of each sector in the national economy it is quite logical that the structure of the net worth of each sector differs greatly from the structure in the other sectors.

1.1 Non-financial assets

Non-financial assets (see Table 2) are represented mainly by fixed assets. Quantification of all types of fixed capital (except of cultivated assets) is performed using PIM.[4] Here we have the most elaborated procedures, but we are working on other specific improvements, especially in capturing other changes in volume of assets and acquisitions less disposals of used fixed assets.

Inventories represent 7% of non-financial assets, in particular, due to a big value of forests recorded under the heading "Work in progress on cultivated biological assets". Valuation of forests is carried out, however, using the market price of wood for the current year instead of the discounted future income from the sale of wood. More about this issue will be in the second part of the paper.

The value of the valuables recorded in the Czech national accounts is very low, because it involves basically only valuables captured in the business accounts of the corporations and valuables made or traded during the past 20 years. Therefore the Czech national accounts do not include collections in museums, or valuables held by households on a long-term basis.

Non-produced assets recorded in the Czech national accounts represent 33% of the value of non-financial assets. They include, in principle, only land, subsoil assets and the part of patents that corporations record in their business accounts. The value of subsoil assets was newly quantified using quantity x price method for individual type of mineral reserves. This new estimate represents the biggest change in the Czech national accounts made under the last main revision in 2014.

[4] Brief description of PIM (perpetual inventory method) can be found in Sixta (2007). Alternative estimates of stocks and deprecation is presented in Krejci and Sixta (2012).

Table 1 Net Worth by sector and type of assets, Czech Republic, 2012, CZK bill.

Code	Assets	Total economy	Non-financial corporations	Financial corporations	General government	Households	NPISH
		S.1	S.11	S.12	S.13	S.14	S.15
AN	**Non-financial assets**	**29 340**	**9 798**	**256**	**13 231**	**5 945**	**110**
AN.1	Produced assets	19 693	9 138	221	5 428	4 811	95
AN.11	Fixed assets	17 467	7 632	206	5 154	4 382	93
	Dwellings	*4 735*	*497*	*1*	*260*	*3 974*	*3*
	Other buildings and structures	*9 432*	*4 440*	*155*	*4 580*	*176*	*81*
	Machinery and equipment	*2 807*	*2 386*	*21*	*166*	*227*	*7*
AN.12	Inventories	2 066	1 503	14	271	276	2
	Work in progress on culti assets	*1 017*	*608*	*0*	*200*	*208*	*2*
AN.13	Valuables	160	3	1	3	153	0
AN.2	Non-produced assets	9 647	660	35	7803	1134	15
	Land	*2 064*	*529*	*19*	*367*	*1134*	*15*
	Mineral and energy reserves	*7 434*	*5*	*0*	*7429*	*0*	*0*
AF	**Financial assets**	**18 150**	**4 936**	**7 208**	**1 822**	**4 082**	**102**
AF.1	*Monetary gold and SDRs*	*34*	*0*	*34*	*0*	*0*	*0*
AF.2	*Currency and deposits*	*4 864*	*887*	*1 240*	*501*	*2 211*	*25*
AF.3	*Debt securities*	*2 492*	*57*	*2 274*	*36*	*121*	*4*
AF.4	*Loans*	*3 293*	*386*	*2 770*	*120*	*16*	*1*
AF.5	*Equity and shares*	*3 585*	*1 266*	*471*	*818*	*1027*	*3*
AF.6	*Insurance, pension schemes*	*659*	*34*	*55*	*1*	*568*	*1*
AF.7	*Financial derivates*	*193*	*35*	*153*	*3*	*2*	*0*
AF.8	*Other accounts receivable*	*3 030*	*2 271*	*211*	*343*	*137*	*68*
AF	**Liabilities**	**19 594**	**8 516**	**7 262**	**2 426**	**1 385**	**5**
AF.1	*Monetary gold and SDRs*	*23*	*0*	*23*	*0*	*0*	*0*
AF.2	*Currency and deposits*	*4 489*	*0*	*4 485*	*4*	*0*	*0*
AF.3	*Debt securities*	*2 410*	*260*	*267*	*1 883*	*0*	*0*
AF.4	*Loans*	*3 555*	*1 608*	*494*	*197*	*1 253*	*3*
AF.5	*Equity and shares*	*5 088*	*4 095*	*992*	*1*	*0*	*0*
AF.6	*Insurance, pension schemes*	*641*	*0*	*641*	*0*	*0*	*0*
AF.7	*Financial derivates*	*212*	*67*	*126*	*19*	*0*	*0*
AF.8	*Other accounts payable*	*3 176*	*2 486*	*234*	*322*	*132*	*2*
BF.90	**Net Financial Worth**	***–1 444***	***–3 580***	***–54***	***–604***	***2 697***	***97***
B.90	**Net Worth**	**27 896**	**6 218**	**202**	**12 627**	**8 642**	**207**

Source: Own research, CZSO

Table 2 Non-financial assets, Czech Republic, 2012

		CZK billions	%
AN.	**Non-financial assets**	**29 340**	**100%**
AN.1	Produced assets	19 693	67%
AN.11	Fixed assets	17 467	60%
	Dwellings	*4 735*	*16%*
	Other buildings and structures	*9 432*	*32%*
	Machinery and equipment	*2 807*	*10%*
AN.12	Inventories	2 066	7%
	Work in progress on cultivated biological assets	*1 208*	*3%*
AN.13	Valuables	160	1%
AN.2	Non-produced assets	9 647	33%
	Land	*2 064*	*7%*
	Mineral and energy reserves	*7 434*	*25%*

Source: Own research, CZSO

1.2 Financial assets

Financial assets or balance of financial assets less liabilities in total (see Table 3) have for the Czech economy a negative value of CZK –1.4 trillion. This balance as "Net financial worth" of the Czech economy represents the final relation to non-residents.

Table 3 Net financial worth and relation to RoW, Czech Republic, 2012, CZK bill.

		Czech Republic			Non residents		
		Assets	Liabilities	Diff	Assets	Liabilities	Diff
BF.90	**Financial net worth**	**18 150**	**19 594**	**–1 444**	**4 317**	**2 861**	**1 456**
AF.1	Monetary gold and SDRs	34	23	11	23	22	1
AF.2	Currency and deposits	4 864	4 489	375	324	699	–375
AF.3	Debt securities	2 492	2 410	82	720	802	–82
AF.4	Loans	3 293	3 555	–262	535	273	262
AF.5	Equity and shares	3 585	5 088	–1 503	2 113	610	1 503
AF.6	Insurance, pension schemes	659	641	18	13	31	–18
AF.7	Financial derivates	193	212	–19	113	94	19
AF.8	Other acc receivable/payable	3 030	3 176	–146	476	330	146

Source: Own research, CZSO

The most important liability in relation to rest of the world (RoW) are shares of non-residents in corporations representing almost half the assets of non-residents in the Czech Republic. Moreover, according to the experimental valuation at a real market value level the equity owned by non-residents would be doubled. The second most important financial instrument in relation to RoW is "Debt securities" that serves mostly for financing of the debt of governmental institutions. Mainly due to these two types of financial assets the net financial worth of the Czech Republic significantly declined over the past 20 years (see Table 4).

Table 4 Net worth, Czech Republic, CZK bill.

Items	1993 (OS)	1995	2000	2005	2010	2011	2012	2013
Non-financial assets	10 756	14 686	20 879	23 691	28 393	29 148	29 340	29 336
of which fixed assets	*5 005*	*7 499*	*11 679*	*14 145*	*17 283*	*17 477*	*17 467*	*17 516*
% of fixed assets in NW	*46%*	*51%*	*56%*	*62%*	*64%*	*63%*	*63%*	*62%*
Financial assets	5 047	6 779	9 579	12 831	16 600	17 427	18 150	19 573
Liabilities	4 917	6 673	9 759	13 708	18 112	18 931	19 594	20 682
Net financial worth	*130*	*106*	*–180*	*–877*	*–1 512*	*–1 504*	*–1 444*	*–1 109*
Net worth	**10 886**	**14 792**	**20 699**	**22 814**	**26 881**	**27 644**	**27 896**	**28 227**

Source: Own research, CZSO

For each of the financial asset a "whom to whom" matrix is compiled. This technique is used for balancing of assets and liabilities among institutional sectors, subsectors and RoW. However, the CZSO has not yet published these matrices that give complete information about ownership. An example of such matrix for the item "securities other than shares" (AF.3) is shown in the aggregate form in Table 5.

Table 5 "Whom to whom" matrix for closing stocks of debt securities (AF.3), 2012, CZK bill.

		ASSETS						Liabilities total
		S.11	S.12	S.13	S.14	S.15	S.2	
LIABILITIES	S.11	5	38	6	0	1	210	**260**
	S.12	19	193	0	19	0	36	**267**
	S.13	21	1 294	29	62	3	474	**1 883**
	S.14	0	0	0	0	0	0	**0**
	S.15	0	0	0	0	0	0	**0**
	S.2	12	749	1	40	0	0	**802**
Assets, total		**57**	**2 274**	**36**	**121**	**4**	**720**	**3 212**

Source: Own research, CZSO

These "whom to whom" matrices are a key instrument for the estimates of missing, weak or incomplete data for some sectors and subsectors. For each financial asset five matrices are always compiled and balanced, for: (1) opening stock (os), (2) transactions, (3) other changes in volume, (4) revaluation and (5) closing stock. They serve also as key instrument because they allow balancing the whole system of accounts for the total economy and for all sectors.

1.3 Experimental valuation of equity

Compliance with the methodological principles of the national accounts for the valuation of assets (e.g., fixed capital formation, stocks, land) raises the similar needs to approach the valuation of the equity on the liabilities side. If the equities of corporations on liability side are valuated according to data taken from business accounts (e.g. for limited liability companies in the amount of paid-up capital) without any adjustments due to changes in valuation of assets the net worth of corporations will be influenced, instead of the net worth of owners. The net worth then ceases to reflect actual ownership structure.

The shares are evaluated in the Czech national accounts by several ways, according to the types of units: joint-stock companies quoted on the stock exchange, the other joint-stock companies, banks, insurance companies and pension funds, investment funds, limited liability companies, housing associations, and others.

For the estimation of ***listed shares*** a special database MAGNUS is used, which allows to find prices of listed shares at a given moment, as well as their owners. The calculation procedure is therefore based on a comprehensive assessment of the amount of liabilities side of the sector and its subsequent distribution to the holders of this amount on the assets side.

Unquoted shares of non-financial corporations and ancillary and other financial institutions (S.11, S.123 and S.124) are valued in the amount of the book value of equity capital. If the data for equity capital is missing, the value of stockholders' equity is used. Next, estimated equity is allocated to counter-parties on the basis of information from the MAGNUS, SCP and commercial register.

For ***financial intermediaries and insurance companies and pension funds*** (S.122 and S.125), it is based on the value of equity according to the banking statistics or met system for insurance companies and pension funds. The assessed value is divided to counter parties according to information from the MAGNUS, SCP (a governmental institution Centre of securities) and commercial register. ***Mutual funds*** shares are estimated according to the database of the Association for the capital market.

Other equity in ***limited liability companies*** and incorporated partnerships are valued in the amount of the paid-up capital. This approach is not sustainable for future because by new business law the paid-up capital is CZK 1.

Other equities of ***housing cooperatives*** are valued by value of apartments in which the members of cooperatives live. The debate whether this procedure is correct initiated more general discussion about

the valuation of equity and definition of the net worth.[5] If not, then the net equity of housing cooperatives is very high, but real owners of net worth are the members of the cooperative. If yes, then we should analogously evaluate the shares in other corporations.

The above listed approaches show very heterogeneous valuation of equity in the current practice. Both the volume and sectoral structure of equity are therefore distorted, and the same is true for the net worth. For this reason, the CZSO has approached to the experimental valuation of equity and their allocation of institutional sectors of the owners. The starting point is the definition of equity in ESA 2010, mainly in paragraphs §5.141 and § 7.71.

Definition of Equity (F.51) by ESA 2010 §5.141: equity is a financial asset that is a claim on the residual value of a corporation, after all other claims have been met. So, the definition says that equity of a corporation is the residual value, and then what is the net worth?

ESA 2010 § 7.71: Listed shares (AF.511) are valued at their market values. The same value is adopted for both the asset side and the liability side, although shares and other equity are not, legally, a liability of the issuer, but an ownership right to a share in the liquidation value of a corporation, where the liquidation value is not known in advance. So, the definition says that equity should be evaluated as a share in the liquidation value of a corporation.

Both definitions lead us to conclusions that in the first place it is necessary to define net worth of corporations and then equity to valuate as a difference between assets and liabilities. If market value of fixed assets (buildings) increases, the value of equity in a corporation should increase by the same value (revaluation).

From both definitions we also made a conclusion that the net worth of a corporation consists only from a net disposable income in the current year because the net disposable income for all previous financial years have been yet distributed to the owners as dividends or reinvested. Reinvested net disposable income of corporations under domestic control increasing the equity value should be recorded as revaluation.

Quantifying the actual value of equity, however, is only the first stage; more difficult task is an allocation of this adjustment to the sector of owners. Allocation process has not yet been sufficiently developed. Therefore, an experimental estimate was made only for non-financial corporations with provisional assumption:

- the change of the value of equity for the public non-financial corporations has been fully allocated to the general government sector;
- the change of the value of equity for the national private non-financial corporations has been fully allocated to the households sector;
- the change of the value of equity for the foreign controlled non-financial corporations has been fully allocated to the rest of the world sector.

Table 6 shows an impact of this experimental reallocation on the net worth. Total national worth fell by 9% because of decreased net worth of the companies under foreign control. The picture haschanged also inside the national economy. Under the principle "wealth do not belongs to company but to its

[5] The owners of the housing cooperatives are members of cooperatives that usually live in the cooperative flats. Members-tenants, therefore, do not own the flats in which they live directly, but they owned shares of the cooperative. Rentals paid by members-tenants to cooperatives cover their operating costs and amortization/reproduction of their housing fund; however, cooperatives do not make any profit from their core business. So, housing cooperatives can create profit, according to the law, only from some secondary activities or from renting apartments to non-members of the cooperative. The question is how to evaluate the shares of the members in the cooperative. By business accounting they are evaluated in the amount of initial deposits to the cooperative. The actual market value of the flats, however, is much higher due to the general trend of prices of real estate, and, in particular, therefore, that cooperative financed flats from deposits by members, but also by subsidies and loans, or due to privatisation of municipal flats for significantly lower price than their market value. For these reasons, we correct the data taken from accounts of housing cooperatives, and the participation of the members in cooperatives we evaluate by the market value of flats reduced by taken loans.

owners" the households are richer mainly due to ownership of limited liability companies, and, similarly, governmental institutions are richer mainly due to ownership of public/state companies.

Table 6 Experimental reallocation of Net Worth, Czech Republic, 31.12.2012, CZK bill.

Code	Assets	Total economy	Non-financial corporations	Financial corporations	General government	Households	NPISH	Non - residents
		S.1	S.11	S.12	S.13	S.14	S.15	S.2
Officially published figures (Net Worth is attributed to corporations)								
AF.5A	*Equity and shares (assets)*	*3 585*	*1 266*	*471*	*818*	*1 027*	*3*	*2 113*
AF.5L	*Equity and shares (liabilities)*	*5 088*	*4 095*	*992*	*1*	*0*	*0*	*610*
BF.90	***Net Financial Worth***	***–1 444***	***–3 580***	***–54***	***–604***	***2 697***	***97***	***1 453***
B.90	**Net Worth**	**27 896**	**6 218**	**202**	**12 627**	**8 642**	**207**	**1 453**
Experimental figures (Net Worth is attributed to owners)								
AF.5A	*Equity and shares (assets)*	*7 145*	*1 266*	*471*	*2 320*	*3 085*	*3*	*3 615*
AF.5L	*Equity and shares (liabilities)*	*11 248*	*10 255*	*992*	*1*	*0*	*0*	*610*
BF.90	***Net Financial Worth***	***–4 044***	***–9 740***	***–54***	***898***	***4 755***	***97***	***2 955***
B.90	**Net Worth**	**25 296**	**58**	**202**	**14 129**	**10 700**	**207**	**2 955**

Source: Own research, CZSO

2 NET WORTH OF THE HOUSEHOLD SECTOR

Net worth of Czech households recorded in national accounts consists in principle of three types of non-financial assets (dwellings, forests and land) and three types of financial assets (deposits minus loans, equity and insurance and pension schemes), see Table 7. Other assets together represent less than 10% of the net worth of households. Therefore, we focus on these six types of asset, both on methods of estimation and values in the time series since 1993.

Table 7 Households sector – Structure of net worth, Czech Republic, 2012, CZK bill.

Code	Assets	by published NAcz			with revaluated other equity		
		CZK billions	% of total economy	% of NW S.14	CZK billions	% of total economy	% of NW S.14
AN	**Non-financial assets**	**5 945**	**20.3**	**68.8**	**5 945**	**20.3**	**55.6**
	Dwellings	*3 974*	*83.9*	*46.0*	*3 974*	*83.9*	*37.1*
	Forests	*208*	*20.5*	*2.4*	*208*	*20.5*	*1.9*
	Land	*1 134*	*54.9*	*13.1*	*1 134*	*54.9*	*10.6*
BF.90	**Net Financial Worth**	**2 697**	**x**	**31.2**	**4 755**	**x**	**44.4**
	Currency and deposits less loans	*974*	*74.4*	*11.3*	*974*	*74.4*	*9.1*
	Equity and shares	*1027*	*28.6*	*11.9*	*3 085*	*43.2*	*28.8*
	Insurance, pension schemes	*568*	*86.2*	*6.6*	*568*	*86.2*	*5.3*
B.90	**Net Worth**	**8 642**	**31.0**	**100.0**	**10 700**	**42.3**	**100.0**

Source: Own research, CZSO

2.1 Dwellings

Dwellings represent 46% of the net worth of Czech households. Net stock of dwellings is calculated by perpetual inventory method (PIM). The depreciation function is linear, retirement pattern is lognormal derived from average service life of 80 years (by blocks of flats) or 90 years (by family houses).

The original ground of application of PIM on dwellings was an estimation of gross fixed stock of dwellings at the end of year 2000. For this purpose, the following data sources were used:

- Census was used for regional structure, division into family houses and flats, age structure and square meters of living area (Census 2001).
- Compilation and evaluation of dwellings gross stock was done separately for two categories:
  - For municipalities with more than 50,000 inhabitants and regions Prague-west and Prague-east the prices were taken from tax return statistics.
  - For municipalities with less than 50,000 inhabitants, the prices were taken from annual statistical survey on completed houses.
- In cooperation with external Research institute for rationalization in building industry the quality change was implemented into the computation of the value of dwellings from different periods, the quality coefficients used reflect the construction material etc. Also the sewerage, gas pipeline and water supply connections are reflected in the quality coefficients.
- Gross fixed capital formation (GFCF) of dwellings divided to flats and family houses based on surveyed data.

As a result the value of gross stock of dwellings at the end of year 2000 was obtained in an age-and-type (family house x flat) structure. The division into institutional sectors and subsectors was done on the basis of census. The ratios of book keeping gross stocks were used for division into industries.

The gross stock was used for construction of artificial time series of gross fixed capital formation in dwellings before 1995. The gross stock divided into age groups was transformed into artificial GFCF by backward calculation of retirement (from retirement function it is easy to compute the percentage of already retired part). The shape of the series is based on number of newly constructed dwellings and on their age structure. Finally, these time series, data on GFCF and price indices in 2011 are used for PIM calculations in order to compile a balance sheet of dwellings.

In 2014 all the data used for the above described method were updated based on new data from the Census 2011. The results of this census showed that the sectoral / ownership structure of the housing stock compared to the last census changed significantly. The CZSO does not have sufficient quality information on the privatization and liquidation of the housing stock in the years between censuses. That is why the calculation is made firstly for the total economy, and then the results are allocated to subsectors by data from the censuses annually updated by figures about new construction of dwellings from a statistical survey. Finally an adjustment to stock of dwellings owned by non-residents is done: these are taken out of household sector and added to foreign controlled non-financial corporation subsector.

Table 8 Households sector – dwellings, current prices, CZK bill.

	1993 to 1995	1996 to 2000	2001 to 2005	2006 to 2010	2011	2012	2013
Opening stock	1 172.9	1 491.7	2 385.7	2 949.2	3 897.0	4 012.5	3 974.2
GFCF	97.5	314.5	473.7	762.4	140.5	125.6	126.6
CFC	–78.8	–200.3	–262.4	–366.3	–81.1	–80.7	–81.5
Other changes in volume	–192.0	66.4	–11.6	50.4	17.3	0.2	3.7
Revaluation	492.1	713.3	363.9	501.3	38.8	–83.4	–32.3
Closing stock	1 491.7	2 385.7	2 949.2	3 897.0	4 012.5	3 974.2	3 990.7
Share of NW (%)	*41.1*	*45.7*	*47.5*	*46.9*	*47.4*	*46.0*	*45.4*

Source: Own research, CZSO

The new results in very aggregate form for the period 1993 to 2013 can be seen in Table 8. Gross fixed capital formation looks very high (annually in average almost 4% to opening stock), it is mainly due to privatisation of flats – municipality and cooperative flats. Other volume changes cover destroyed dwell-

ings due to catastrophic events, e.g. floods. The accumulated impact of dwellings revaluation is bigger than the amount of gross fixed capital formation in dwellings. Almost the same amount of changes in value of housing stock is the cumulated result of revaluation. Price indices are differing by region and by type of dwellings. Valuation of stock is one of the weakest part of our calculation in replacement or market value because of the danger of double counting of underlying land.

2.2 Forests

Forests are important national resource for the Czech Republic; therefore, nationwide inventory of forest are performed regularly. More than 20% of all forestry land are owned by households. Current estimate of the value of forests is based on the results of the nationwide inventory conducted during 2001–2004. Presently, another national inventory of forests is conducted (2011–2015). It was completed by the end of 2014; the results will be processed, evaluated and published by the end of 2015, so we assume that its results will be used in the national accounts for the year 2016.

An estimate of the value of forests in the Czech national accounts is now carried out by applying the average prices of raw wood on stock of standing timber in cubic metres per kind of timber. Both methodology and current annual valuation are elaborated and processed for the CZSO by two external research institutes. So the CZSO replaced previously used method of valuation based on discounted future proceeds from a timber sale. This method was, in theory, more correct, but more demanding and especially negatively accepted by users. Therefore, we do not plan to reintroduce it.

The calculation is done for the total economy. The estimated values are then allocated proportionally to institutional sectors. Sector structure is derived from the ownership of forest land in hectares. Whereas the structure of standing timber, growth and woodcutting are calculated only for the total economy, it does not reflect the quality or value of the standing timber in the household sector separately.

Table 9 Households sector – forestry, current prices, CZK bill.

	1993 to 1995	1996 to 2000	2001 to 2005	2006 to 2010	2011	2012	2013
Opening stock	134.9	137.9	170.5	142.5	155.9	201.4	207.6
Changes in inventories	1.0	3.5	2.1	3.3	1.0	0.8	1.3
Other changes in volume	0.0	0.0	0.0	0.4	0.0	0.0	0.0
Revaluation	1.9	29.1	–30.1	9.8	44.4	5.4	11.3
Closing stock	137.9	170.5	142.5	155.9	201.4	207.6	220.1
Share of NW (%)	*3.8*	*3.3*	*2.3*	*1.9*	*2.4*	*2.4*	*2.5*

Source: Own research, CZSO

The final data for the period 1993 to 2013 are shown in aggregated form in table 9. Changes in inventories represents the balance of the forest here (growth and woodcutting of standing timber), as well as net sales or purchases of forests by households. Other changes in volume of forests are negligible, however, changes in the valuation of stock play important role, due to changes in prices of wood. The current method of calculation is sensitive to fluctuations in the market prices of raw wood.

2.3 Land

The value of the land recorded in the national accounts represents more than 7% of the national wealth. The land is owned mainly by households (55% of total value of land), by non-financial corporations (26%) and by government institutions (18%). In households sector the value of land represents more than 13% of their net worth. Table 10 provides an overview of the stock and changes in stock of land owned by households during last twenty years period.

Table 10 Households sector – land, CZK bill.

	1993 to 1995	1996 to 2000	2001 to 2005	2006 to 2010	2011	2012	2013
Opening stock	635.3	665.9	670.3	706.7	1 093.7	1 116.7	1 133.7
Net acquisition of land	–0.1	–11.9	–11.3	7.0	1.7	1.7	3.4
Other changes in volume	0.4	15.6	–0.6	6.1	2.2	0.6	1.9
Revaluation	30.3	0.7	48.4	373.8	19.2	14.8	33.9
Closing stock	665.9	670.3	706.7	1 093.7	1 116.7	1 133.7	1 172.9
Share of NW (%)	*18.3*	*12.8*	*11.4*	*13.2*	*13.2*	*13.1*	*13.3*

Source: Own research, CZSO

The value of land is estimated in three stages. The first and second stages are carried out without breakdown by institutional sectors, so acquisitions and disposals of land between sectors and subsectors can be ignored. The allocation of the results to institutional sectors and subsectors and balancing of acquisitions and disposals of land among sectors and subsectors is made at the third stage.

The first stage represents a compilation of the balance for each type of land and regions in hectares. The difference between opening and closing amount of various types of land are regarded as a change in use of land. Also economic appearance and disappearance of land due to refining of total area recorded in State cadastre are there.

The second stage represents a conversion of the balances for each type of land in hectares to value expression by applying average prices for each type of land and also structured by regions, received from price statistics.

The calculation is carried out in basic breakdown of the land as follows:

- agricultural land *x* the average purchases prices of agricultural land,
- non-agricultural land:
  - land underlying buildings and courtyards *x* the average purchases prices of building site area,
  - forestry land *x* the average purchase price of forestry land,
  - surface water *x* the average price of water and other areas (estimated for total economy by data available on the internet),
  - other land *x* the average price of water and other areas (estimated for total economy by data available on the internet).

Changes in the use of land and related changes in land prices are intercepted as the changes in classification of land (K.122). Changes due to refining of total area are also valuated and recorded as economic appearance and disappearance of land (K.3 or K.62). Finally, the difference between closing and opening stock and total other volume changes are interpreted as revaluation (K.11).

The third stage covers the balancing of acquisitions and disposals of land between sectors and subsectors (data are mainly from statistical surveys and tax returns for real estate transfer tax) and the allocation of stock, other volume changes and revaluation for total economy to individual institutional sectors and subsectors.

For the distribution of agricultural land, data from the survey made by agriculture statistics and information from the annual reports of the State Land Fund on an area of land under its records are used. The agriculture statistics survey shows that the majority of agricultural land is used by legal persons, but mostly it is rented out. Major the part of this rented land is owned by natural persons and, therefore, the value of agricultural land is the largest in the households sector.

Forestry land is divided into institutional sectors based on the information received from the Forest Management Institute that controls inter alia the national inventory of forests. According to this institute about 20.6% of the area of forest land belongs to the household sector.

Given the prices the value of the land underlying buildings and courtyards makes more than 70% of total value of all land in the Czech Republic (majority is owned by households because of land under houses). The value of built-up land in households sector is estimated based on the number of dwellings, average built-up area and the average price of a built-up area. The calculation is performed separately for family houses and apartment buildings. The rest of the value is allocated mostly to the non-financial corporations sector and government institutions in accordance with the statistical survey data. Therefore, there is a special project aimed on usage of individual cadastre data providing information about ownership, type of land and the character of its use.

2.4 Currency and deposits less loans (credits)

Currency and deposits reduced by taken loans represent 11% of net worth of households. To estimate the amount of currency held by households, deposits and loans mainly the data from bank statistics and statistical surveys are used, that are finely balanced in the whom to whom matrixes. For each asset, the separate matrix is prepared and balanced. For the estimation of the loans given to households, a special database kept by the Czech National Bank is used.

Table 11 Households sector – currency and deposits less loans, CZK bill.

	1993 to 1995	1996 to 2000	2001 to 2005	2006 to 2010	2011	2012	2013
Opening stock	184.5	353.5	780.9	781.2	891.7	947.3	973.9
Transaction	159.8	409.3	17.7	24.6	37.6	21.0	4.1
Other changes in volume	10.8	7.9	8.0	103.5	5.2	4.9	4.9
Revaluation	–1.7	10.3	–25.4	–17.6	12.7	0.8	7.3
Closing stock	353.5	780.9	781.2	891.7	947.3	973.9	990.1
Share of NW (%)	*9.7*	*15.0*	*12.6*	*10.7*	*11.2*	*11.3*	*11.3*

Source: Own research, CZSO

The share of net deposits (reduced by loans) on net worth of households was very high in the late nineties, and then the Czech households had been less saving and more taking loans in the context of increasing investments to dwellings and rapidly increasing expenditure on final consumption.

Deposits as well as loan stocks of households were also significantly affected by their revaluation and other changes in the volume, see Table 11. Revaluation refers to deposits of or loans to households saved/received in foreign currency as a result of changes in the exchange rate of the Czech Crown to foreign currencies, in particular USD, DEM, EUR and CHF. In particular writing-off or writing-down of bad debts and financial leasing by creditors were included in the other changes in volume, and in 2007, also newly included deposits of Czech households abroad. These deposits are estimated based on information about the taxation of interest and the average interest rates in countries where these deposits are located.

2.5 Equity and investment funds shares

Equity and investment funds shares represent by officially published data about 12% of net worth of households – only 6% of them were allocated in listed shares, 31% in unlisted shares, 46% in other equity and 17% in investment fund shares in 2012. In total, their role in net equity of households has been falling continuously since 1995 in connection with sale of shares acquired in the voucher privatisation, and also as a result of privatisation of cooperative apartments. However, investment fund shares experienced specific development– long-term growth was affected by a large decline during two financial crises in 1997–1999 and 2007–2011.

A significant part of the present value of the equity and investment funds shares held by households is not the result of transactions, but the revaluation and other changes in the volume, see Table 12.

Table 12 Households sector – equity and investment funds shares, Czech Republic, CZK bill.

	1993 to 1995	1996 to 2000	2001 to 2005	2006 to 2010	2011	2012	2013
Opening stock	438.3	658.4	707.5	857.6	1 144.6	958.5	1 027.3
Transaction	282.2	–80.6	132.7	1.4	–23.0	15.3	–17.7
Other changes in volume	–60.7	62.3	1.2	154.9	–184.4	50.8	–4.2
Revaluation	–1.4	67.4	16.2	130.6	21.4	2.7	16.3
Closing stock	658.4	707.5	857.6	1 144.6	958.5	1 027.3	1 021.7
Share of NW (%)	*18.1*	*13.5*	*13.8*	*13.8*	*11.3*	*11.9*	*11.6*

Source: Own research, CZSO

The experimental reallocation of the net worth by application of the principle "wealth does not belong to company but to its owners" shows that the Czech households are actually richer in the amount of revaluated other equity of owned companies. See chapter 1.3.

2.6 Insurance and pension entitlements

Insurance and pension entitlements are quickly and constantly increasing asset in possession of Czech households. Their share in the worth of households increased from 1.6% in beginning of 1993 to the current 6.9%, see Table 13.

Table 13 Households sector – insurance schemes, Czech Republic, CZK bill.

	1993 to 1995	1996 to 2000	2001 to 2005	2006 to 2010	2011	2012	2013
Opening stock	44.4	74.3	153.7	318.3	512.2	539.6	568.1
Transaction	20.7	84.1	172.0	200.5	29.4	34.2	42.5
Other changes in volume	9.3	–4.7	–7.5	–3.3	–2.1	–5.7	–0.6
Revaluation	0.0	0.0	0.0	–3.2	0.0	0.0	0.0
Closing stock	74.3	153.7	318.3	512.2	539.6	568.1	610.0
Share of NW (%)	*2.0*	*2.9*	*5.1*	*6.2*	*6.4*	*6.6*	*6.9*

Source: Own research, CZSO

Our estimation of their stocks, transactions and other changes for households is conducted on the basis of administrative data from insurance companies and pension funds, or based on data from supervision conducted by the Czech National Bank. The share of net equity of households in life insurance entitlements (AF.62) and in pension entitlements (AF.63) is 100%, because the non-resident households are not assumed in any life insurance or participation in pension funds. The share of households in "non-life insurance technical reserves" (AF.61) is 33%. Allocation to sectors is made according to the percentage of premiums received.

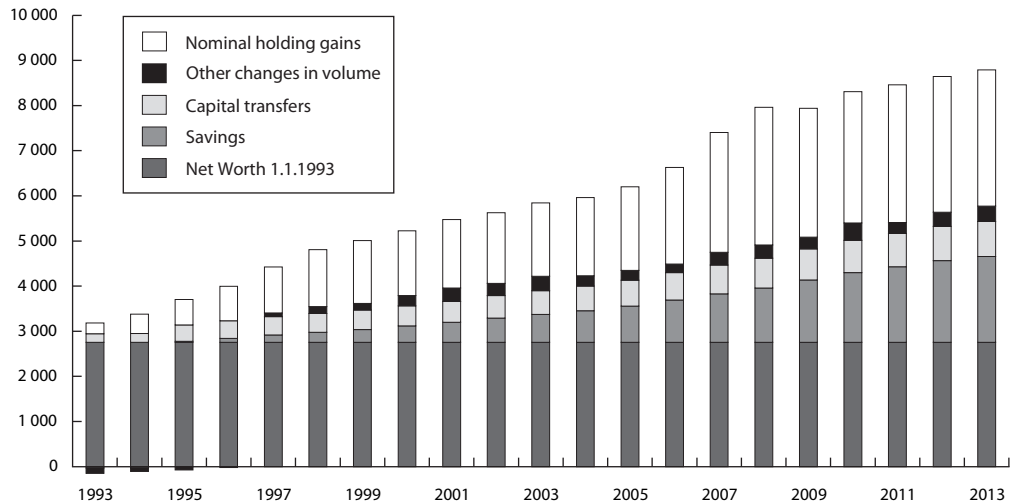
The big other change in the volume in household sector is mainly caused by methodological correction of "non-life insurance technical reserves" – crossed from the net to the gross reserves, i.e. incl. reserves of reinsures. Only one record in the revaluation account was made in 2009 for pension funds due to the revaluation of their assets. In fact, however, this is an extraordinary change, which should be recorded in other changes in volume account.

3 APPRECIATION/DEPRECIATION OF HOUSEHOLD WEALTH

Where-from the households wealth comes? Generally, or in simple terms, the wealth originated from accumulated not consumed disposable income, i.e. from savings. However, there are other factors in re-

ality – capital transfers from other institutional sectors, inflation or changes in prices of different assets and other volume changes in ownership of assets. 31% of the value of household wealth in 2013 came from a time before the formation of the independent Czech Republic in 1993 (see Figure 1).

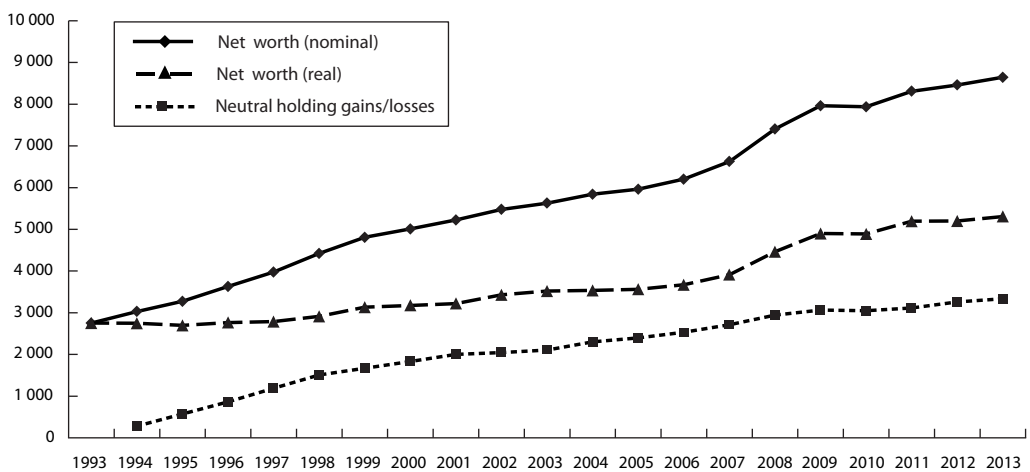
Figure 1 Household wealth comes from, CZK billions



Source: Own research, CZSO

A further 9% has its origin in capital transfers, in particular from voucher privatization in the nineties, from restitution and privatization of cooperative and municipal flats for lower prices than market ones. The small part, 4%, originated from other volume changes – mainly due to consideration of land that was not recorded before at all, and also due to writing-off or writing-down of bad debts.

Figure 2 Net worth of Czech households in nominal and real terms, CZK billions

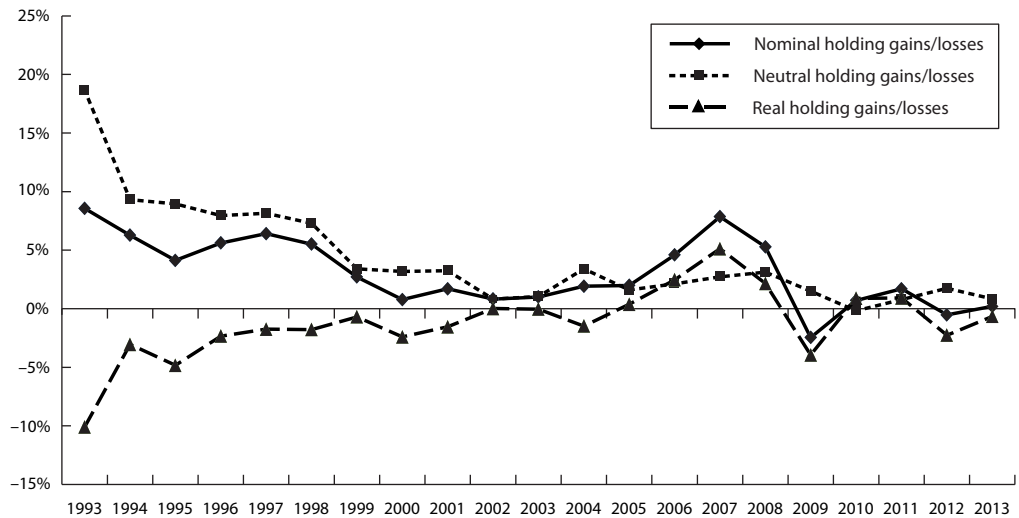


Source: Own research, CZSO

The main part of the value of household wealth in 2013 came from revaluation of existing assets and liabilities held by households. Nominal holding gains accumulated during the period 1993 to 2013 represented 34% of the total value of household wealth. However, these nominal holding gains cover also neutral holding gains caused due to changes in the general price level (measured by the index of the final national uses, excluding changes in inventories). So, the total wealth of Czech households increased in nominal terms 3.2 times, but after deduction of the neutral holding gains the real value of wealth increased only 1.8 times. The trend of nominal and real net worth of Czech households is seen in Figure 2. Or otherwise, increasing the value of the wealth owned by Czech households was covered from 64% by increasing in the general price level.[6]

For analytical evaluation of the development of net wealth for individual institutional sectors it is extremely important to record holding gains for all assets and liabilities and to show the nominal and real appreciation or depreciation. The Figure 3 shows that the real appreciation of the assets of Czech households in total, took place six times only – in the period 2006–2008 and in the years 2010 and 2011. Over the other fifteen years, the value of household wealth declined in relative terms.

Figure 3 Net worth of Czech households in nominal and real terms, CZK billions



Source: Own research, CZSO

CONCLUSIONS

The System of National Accounts provides a wealth of information on the status and development of the household sector. However, comprehensive analyses of the stock and development of the Czech household's wealth are not frequent. This is not only due to preferences in production/consumption analyses, but also because the figures on the stock of the wealth are still of lower quality, international comparability is missing, and most importantly, they lack social dimension.

Improving data quality is promising. In recent years it was caused particularly by improving the valuation of land and houses. The project focused on the use of the Cadastre data in the land continues. In the coming years, the experimental work on the valuation of equity will bring official results – and we

[6] The issue of nominal, neutral and real holding gains was deeply presented in Rybáček (2010).

can expect significant rewriting of currently presented levels of the Czech household's wealth. Projects focused on international comparisons of household wealth are currently organized by OECD and in the Euro area, but so far without the participation of the CZSO.

Social dimension of wealth is the most serious weakness in the Czech National Accounts. Because during the last twenty years former homogenous Czech households have been differentiating, traditional views on social groups are unsatisfactory. Social statistics respond to this important trend of the Czech society inadequately, national accounts not at all. Although in current practice of the CZSO the household sector is divided to two subsectors – households as consumers and households as entrepreneurs, but the original intention of this breakdown was the technique of estimates of some items and accounts. Now, of course, it is also used for analytical reasons, even if it is not too appropriate because one institutional unit (entrepreneurs) is artificially split into businesses and consumers. This breakdown does not provide any information about the development and structure of the wealth of individual social groups of households. That is why, the CZSO has launched the experimental work on the breakdown of the household sector according social and income groups. At present, however, it is still too early to talk about the results.

References

EUROPEAN COMMISSION. *European system of accounts – ESA 2010*. Luxembourg: Publications Office of European Union, 2013, ISBN 978-92-79-31242-7.

HRONOVÁ, S., FISCHER, J., HINDLS, R, SIXTA, J. *Národní účetnictví, nástroj popisu globální ekonomiky.* Prague: C. H. Beck, 2009. ISBN 978-80-7400-153-6.

KREJČÍ, I., SIXTA, J. Využití alternativních metod při odhadech stavů a spotřeby fixního kapitálu. *Politická ekonomie*, No. 6, 2012. ISSN 0032-3233.

ONDRUŠ, V. Compilation of Non-Financial Balances in the Czech Republic. *Statistika: Statistics and Economy Journal*, No. 3, 2011. ISSN 0322-788X.

RYBÁČEK, V. Theory and practice of holding gains and losses: Is the importance of revaluation reflected in national accounts? *Statistika: Statistics and Economy Journal*, No. 6, 2010. ISSN 0322-788X.

SIXTA, J. Odhady spotřeby fixního kapitálu. *Statistika: Statistics and Economy Journal*, No. 2, 2007. ISSN 0322-788X.

SIXTA, J., VLTAVSKÁ, K., HRONOVÁ, S., HINDLS, R. Struktura spotřeby českých domácností 1970–2012. *Politická ekonomie*, No. 6., 2014. ISSN 0032-3233.

SHRESTHA, M., MINK, R. *An Integrated Framework for Financial Flows and positions on a From-Whom-to-Whom Basis*. Paper presented at the IMF-OECD Conference on Strengthening Sectoral Position and Flow Data in the Macroeconomic Accounts, 28 February – 2 March 2011, Washington D. C. Available at: <http://www.imf.org/external/np/seminars/eng/2011/sta/pdf/whom.pdf>.

UNITED NATIONS, EUROPEAN COMMISSION, INTERNATIONAL MONETARY FUND, ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT AND WORLD BANK. *System of National Accounts 2008*. United Nations: New York, 2009. Available at: <http://unstats.un.org/unsd/nationalaccount/docs/SNA2008.pdf>.

Fertility of Czech Females Could Be Lower than Expected: Trends in Future Development of Age-Specific Fertility Rates up to the Year 2050

Ondřej Šimpach[1] | *University of Economics, Prague, Czech Republic*

Abstract

Fertility is an essential aspect of reproduction or population replacement of each country. The challenge for demographers is to model fertility and also to estimate its potential future level for the purposes of population projections. In the case of the Czech Republic we have the population projections provided by the Czech Statistical Office (CZSO) with overlooking of the total fertility rate in low, medium and high variant. These estimates despite being based on expert judgments, seem to be too positive compared to the past development of the time series of age-specific fertility rates. The aim of this paper is to assess the situation of fertility in the Czech Republic, to analyse the past development of the time series of age-specific fertility rates using one-dimensional Box-Jenkins models and multidimensional stochastic Lee-Carter approach. Together with found trend in time series and principal components estimated by Lee-Carter's model a forecasts of age-specific fertility rates up to the year 2050 is constructed. These rates are lower than those provided by CZSO in its three variants of the Czech Republic's population projection, and therefore we discuss the causes at the end of the paper. We would like to point out that the potential future development of Czech females fertility could be lower than which are currently expected.

Keywords

Age-specific fertility rates, ARIMA, Lee-Carter, population projection

JEL code

C22, C32, J13

INTRODUCTION

Mortality and fertility are important parts of the natural population change. Given that the most of populations on Earth started with the dynamic development in recent decades, the standard of living rises and the mortality rates in these countries decline. Due to the fact that the living conditions are better, the forecasting of mortality is not so difficult, because we have common assumptions about the potential

[1] Department of Statistics and Probability, Nám. W. Churchilla 4, 130 67 Prague 3. E-mail: ondrej.simpach@vse.cz, phone: (+420)737665461.

future development, which we simply follow in our predictions (Stauffer, 2002, or Dotlačilová, Šimpach, Langhamrová, 2014). Modelling and estimating future fertility is more complicated, because fertility is influenced by several factors. The population development and improving the living standard in the country is closely related to postponement of first childbirth to the later age and together the decline of live births in total (see e.g. the paper from Rueda, Rodriguez, 2010). This decrease is below the level of simple reproduction of the population (2.08 children per 1 female within the reproduction period) in many populations of developed countries. However, in comparison with mortality, there is still one very important factor that must not be overlooked.

There is a good database in the Czech Republic that will allow us to obtain the age-specific fertility rates of Czech females from 1925 to 2012. During this period, the development of these rates was affected by a wide range of social changes. This was especially the Second World War, the two parts of the consecutive Communist regime, targeted pro-population policies and massive support of young families with a higher number of children, and as well as the downturn of this development during the post-revolutionary period. All these social circumstances brought the consequences of changes in fertility of Czech females, which we are able to justify. It is difficult to explain and to predict as the behaviour is the result of individual decisions in family planning. Neither, the level of fertility can permanently decrease in the future, because there is a value below which the fertility never decreased before. Neither this value can permanently grow in the future, because of health point of view there is a maximum possible value of age that a female cannot exceed (see e.g. Caputo, Nicotra, Gloria-Bottini, 2008, or Myrskylae, Goldstein, Cheng, 2013). The level of fertility varies between its logical lower and upper limits in time, and also depends on the shape of the distribution of age-specific rates. Czech Statistical Office (CZSO) provides regularly updated population projections of the Czech Republic. These projections are constructed by sophisticated cohort-component method, whereby the input attributes and other assumptions are discussed by respected professionals. In the case of the total fertility rate there are currently considered three potential future scenarios, pessimistic (low variant), middle (medium variant) and optimistic (high variant) (CZSO, 2013). Pessimistic scenario consider the same level of total fertility rate in the future as today (1.45 children born to one female during her reproductive period), middle and optimistic consider some increase (see below). The potential future decline is not considered at all, because the past development of the Czech time series showed that e.g. in 1999 there was the total fertility rate 1.13 live birth child per 1 female during her reproductive period and the range of values 1.13–1.18 was in many other cases during the 90s of the last century. It is important to note that the decline of fertility of Czech females at the end of the last century had been mainly caused by rapid changes in reproductive behaviour – postponing of childbirth to the later ages which is normal in the most of Western European countries today. The sharp fertility decline of younger females was partly compensated (with a delay) by fertility increase of females in higher age groups (Langhamrová, Fiala, 2014). Medium variant of CZSO expect a gradual linear increase up to a value of 1.56, high variant even up to a value 1.61. Is it possible that some of these variations will happen? Can we expect, that females, married couples and partners change their views on the family and this increase will occur? It is possible to read and judge from the population structure of the Czech Republic that the strong generations of 70s are already reproductively exhausted and other strong generation which will be able to significantly revived this situation will not appear within next 20–30 years. The population structure enables to see the development of age-specific fertility rates using a statistical approach and together with the founded trend and the main components explaining fertility levels estimate, how these rates could develop in the future.

The aim of this paper is especially to analyse the past trend in the individual time series of age-specific fertility rates using Box-Jenkins methodology (Box, Jenkins, 1970) with Random Walk models and ARIMA. These models are applied to 35 time series (for age range of 15–49 completed years of life), the periodicity of the time series is annual (1925–2012) with sufficient number of observations. We evaluate

the models by diagnostic control (see e.g. Stauffer, 2002) and consequently calculate other predictions of age-specific fertility rates for the period 2013–2050 (different from CZSO approach). The approach of Random Walk models provide one potential prediction, ARIMA slightly different one. At the same time we estimate the principal components that explain fertility from the multidimensional matrix of age-specific fertility rates (see Hyndman, Booth, 2008, or Arltová, 2011). This is performed using the singular value decomposition (SVD), the Lee-Carter model (Lee, Carter, 1992). Estimates are gradually made for the different lengths of the analysed matrix – (I) since 1925 (the beginning of the time series), (II) 1948 (the end of the Second World War, pacification the social situation and the beginning of a new political regime in the country), (III) 1968 (again the restructuring of the society and the beginning of hard normalization), and (IV) 1988 (weakening of the Communist regime in the country and preparing for the new democratic system in our society).[2] Only the results of the model based on data for the period 1925 to 2012 and 1988 to 2012 are presented. It is due to the fact that SDV approach is not appropriate in the case when the multi-dimensional matrix record a wide range of changes in the past and is therefore highly variable. We also calculate the predictions of age-specific fertility rates for the period 2013–2050 on the basis of those two models. All four approaches used for calculation forecasts (Random Walk, ARIMA, LC 1925 and LC 1988) will be compared with each other and with published values of low, medium and high variant of CZSO.

There are other ways to analyse and model fertility in developed populations in order to be able to construct the fertility projections. Peristera and Kostaki (2007) prepared an extensive case study on the United Kingdom, Ireland, France, Greece, Norway, Italy, Denmark, Austria and the United States using the Hadwiger Model, Gamma Model, Beta Model and quadratic Spline Model. Hyndman and Ullah (2010) paid attention to France using robust approach to modelling fertility based on the approach by Lee, Carter (1992). This was used before only on mortality modelling, and its capabilities were extended and used on fertility analysis later. Given that our dataset is suitable for Hyndman and Ullah (2010) approach, (based on studies by Lee and Carter (1992), Lee and Tuljapurkar (1994) and later Rueda, Rodriguez (2010)), we apply this method. The database was also suitable for the application of methodological approach by Box and Jenkins (1970), which is older and which has been used previously in many studies of mortality analysis (Bell, 1997, Stauffer, 2002, or Šimpach, Langhamrová, 2014). It is used in our paper as a comparison of the modern approach of stochastic modelling with principal components and the conventional approach of stochastic modelling with random component. The fertility predictions were calculated in the Czech Republic e.g. by Fiala and Langhamrová (2012), who used a deterministic approach in calculation of the population projections. They calculated with expert judgments of the total fertility rates (available at the CZSO). The age-specific fertility rates were subsequently calculated using the component method.

Using a long series of cross-sectional indicators (fertility rates by age groups) for long-term projections seems, unfortunately, problematic in the Czech Republic – this is the main reason why we do not pay the attention for models LC 1948 and LC 1968. The significant inter-annual fluctuations of fertility rates that do not have the recognizable long-term trend in the cross-sectional point of view are typical for the Czech population (unlike for many developed and Western European countries). It is important for the projection to focus on the relatively recent changes and concentrate on the cohort / generation approach, because the final indicator of fertility in the Czech Republic is long-term stable. (The permanent declining trend begins with generations of females born in 1960 and younger.) This approach was

[2] The various social events did not follow each other exactly at specific 20 years intervals. If we would like to set these dates correctly, we will have to select the 1945, 1968 and 1989. Frequency 20 years between these events was chosen as a compromise variant for the purposes of analysis.

used in article by Myrskylae, Goldstein, Cheng (2013), or in other: Li, Wu (2003), or Morgan, Hagewen (2005). When we analyse the longer time series of variable cross-sectional fertility rates by age groups, we do not improve the projection. The situation is rather the opposite (unlike the natural sciences). This also confirm our results presented at the end of this paper, where the most probable results are provided by the Lee-Carter model for the period 1988–2012, while other models give too high predicted values, which are deflected particularly by high level of fertility of young females between 60s and 80s of the last century and the presented models are not robust for these extreme changes.

Statistically estimated values of the total fertility rates in this article are lower than published values by CZSO. *This paper does not attempt to say in any cases that the estimated future values by CZSO are wrong!* Our predictions were calculated by statistical approach that takes into account the trend and the principal components that make up the main explanatory system. Published values of CZSO take into account the opinions of *demographers* and other experts from the fields of sociology, political science, and medicine. Therefore they are not only enriched by a factor of technological progress, but also affected by subjective thinking that a statistical approach does not have. Another advantage is that statistical approach takes into account in addition the random component (Alders, de Beer, 2004, or Caputo, Nicotra, Gloria-Bottini, 2008). We think that the total fertility rate will be located in 2050 between the level of pessimistic variant of CZSO (1.45) and the results obtained using Random Walk models or ARIMA, (which reached the level of 1.20 in 2050). The values published by CZSO are rather optimistic, the values that provide statistical approaches are rather pessimistic. Let us be more careful in our future expectations and think about whether the Czech Republic has such potential up to the year 2050, in which it could approach their fertility to countries as Belgium, Netherlands or Luxembourg in these days.

1 METHODOLOGY AND DATA

The empirical data from CZSO are used – particularly the number of live-born persons to x-year old mothers in year t ($N_{x,t}$) and the number of midyear female population x-year old in year t ($\overline{S}_{x,t}$), where x = 15–49 completed years of life and t = 1925–2012. This allows us to calculate the age-specific fertility rates as:

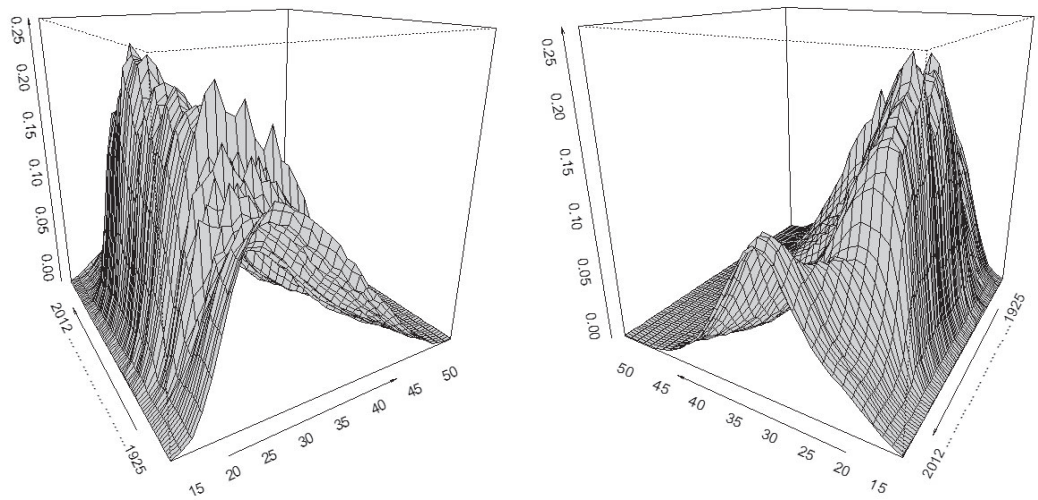
$$f_{x,t} = \frac{N_{x,t}}{\overline{S}_{x,t}}, \tag{1}$$

and after (× 1 000) we interpret the result as the number of live births per 1 000 x-year old females in year t. Sum of age-specific fertility rates is the total fertility rate in year t

$$tfr_t = \sum_{x=15}^{49} f_{x,t}, \tag{2}$$

which is the sum of live births to one female during her reproductive period. In order to be clear, which changes in age-specific fertility rates occurred in the past, their development is shown in Figure 1 as perspective 3D chart. (This technology uses Charpentier, Dutang, 2012, simple presentation by X-Y chart of these empirical rates shows Figure 8 in Annex). We can see the changing of maximum values of the age-specific fertility rates in the past, as well as the moving of modus. It is especially due to the trend of postponing childbirth to the later ages which emerged at the beginning of 90s of the last century. In these days the modal age have exceeded the value of 30 years.

Figure 1 Empirical data of age-specific fertility rates $f_{t,x}$ of the Czech females for the period 1925 to 2012 in perspective 3D charts. See the significant change of mode of this distribution to the advanced ages



Source: CZSO (2013), author's illustration

If we transform the age-specific fertility rates to the logarithms $f_{x,t} \rightarrow \ln(f_{x,t})$, their variability in time is smaller. We may look on each specific age from 15 to 49 completed years of life as on the individual time series of logarithms of age-specific fertility rates with 88 annual observations (from 1925 to 2012). This approach is generally used by authors to model the logarithms of age-specific mortality rates and for the estimation of the coefficients for declining mortality over time (Stauffer, 2002, or Šimpach, Dotlačilová, Langhamrová, 2014). We do not estimate any coefficients for declining fertility rates over time, but we apply Box-Jenkins methodology on the less variable time series of age-specific fertility rates. When we have a larger number of the analysed time series of similar nature, it is preferable to determine the universal structure of model for all of the considered time series.[3] One of the most frequently used approaches that are chosen as a compromise variant is the Random Walk model (RW) with drift. It is more sophisticated than a simple linear deterministic model (which was used in the past), because it takes into account the random component. We denote the random walk model for the logarithms of age-specific fertility rates as:

$$\ln\left(f_{x,t}\right) = c_x + \ln\left(f_{x,t-1}\right) + \varepsilon_{x,t}\,, \tag{3}$$

[3] In the case that we analyse a small number of time series, it is right that each time series should be analysed with maximum precision (Hyndman et al. 2002). Each obtained model should be tested with a wide range of diagnostic tests and we have to insist that all conditions have been fulfilled to the last detail. For the larger matrices of the time series we are not able to satisfy all assumptions and diagnostic controls, so it is the time to find and choose a compromise structure of model, which satisfies the most of the time series from the matrix (Alders, de Beer, 2004). This issue is devoted e.g. by authors Melard, Pasteels (2000) or Hyndman et al. (2002), who developed the approaches for automatic modelling of time series and automatic forecasting. Nowadays these systems are very sophisticated, but their disadvantage is that the models often include too much of parameters to satisfy the most of the evaluation criteria during the evaluation of model. Very often happens that these parameters are mostly statistically insignificant, or from the logical point of view they do not belong to the model. Their estimated values are also very often incorrect, located in senseless intervals according to statistical theory.

where c_x is drift and $\varepsilon_{x,t}$ is the error term with characteristics of white noise. This formula can be modified according to Box, Jenkins (1970). We get ARIMA (0,1,0) model with drift for age-specific fertility rates as:

$$\ln\left(f_{x,t}\right) = \ln\left(f_{x,0}\right) + c_x \cdot t_x + \sum_{t=1}^{T} \varepsilon_{x,t}\,, \tag{4}$$

where $c_x \cdot t_x$ is the deterministic trend. This trend is linear increase / decrease of fertility rates in time. Its most common usage is in the case of modelling age-specific mortality rates, there will be experimentally used in case of fertility. It was further examined by empirical verification which parameters (Auto-Regressive AR or Moving Averages MA) are statistically most important part of the ARIMA model in demographic time series. In the case of modelling mortality is mostly statistically significant component MA(1). Models which contain AR component often do not remove autocorrelation (Melard, Pasteels, 2000). This autocorrelation unfortunately not disappear even if the model includes a drift that often this unpleasant characteristic pulls into itself. Therefore we use component MA and define the model of moving averages without drift according to Box, Jenkins (1970) as:

$$\ln\left(f_{x,t}\right) = \varepsilon_{x,t} - \theta_{x,1}\varepsilon_{x,t-1}\,, \tag{5}$$

with drift respectively as:

$$\ln\left(f_{x,t}\right) = c_x + \varepsilon_{x,t} - \theta_{x,1}\varepsilon_{x,t-1}\,. \tag{6}$$

whereby the provisions of condition $|\theta_{x,1}| < 1$.

The other used approach based on principal component is that the empirical values of age-specific fertility rates can be decomposed (see Lee, Carter, 1992, or Lee, Tuljapurkar, 1994) as:

$$f_{x,t} = a_x + b_x \cdot k_t + \varepsilon_{x,t}, \tag{7}$$

where x = 15–49, t = 1,2, …, T, a_x are the age-specific fertility profiles independent of time, b_x are the additional age-specific components determine how much each age group changes when k_t changes and finally k_t are the time-varying parameters – the fertility indices. ($\varepsilon_{x,t}$ is the error term with characteristics of white noise). The estimation of b_x and k_t is based on Singular Value Decomposition (SVD) of matrix of age-specific fertility rates, presented e.g. by Bell, Monsell (1991), Lee, Carter (1992), or Hyndman, Ullah (2010). The age-specific fertility rates $f_{x,t}$ at age x and time t create 35 × T dimensional matrix

$$\mathbf{F} = \mathbf{A} + \mathbf{B}\mathbf{K}^{\mathrm{T}} + \mathbf{E}, \tag{8}$$

and the identification of Lee-Carter model is ensured by

$$\sum_{x=15}^{49} b_x = 1 \quad \text{and} \quad \sum_{t=1}^{T} k_t = 0\,. \tag{9}$$

Finally,

$$a_x = \frac{\sum_{t=1}^{T} f_{x,t}}{T} \tag{10}$$

is the simple arithmetic average of age-specific fertility rates. For predicting the future age-specific fertility rates it is necessary to forecast the values of parameter k_t only. This forecast is mostly calculated by ARIMA (p,d,q) models with or without drift (Box, Jenkins 1970). The values of the parameters a_x and b_x

are independent of time and the prediction using the Lee-Carter model is therefore purely extrapolative (Lee, Tuljapurkar, 1994).

Czech Republic has, unfortunately, very variable development of data of age-specific fertility rates. Therefore we estimate the parameters a_x, b_x and k_t for complete model based on data 1925–2012 (LC 1925), and also for 3 shortened models, based on data 1948–2012 (LC 1948), 1968–2012 (LC 1968) and 1988–2012 (LC 1988). We present the detailed results provided by LC 1925 and LC 1988 model only. Results from complete model LC 1925 provide misleading predictions, because the average age-specific fertility profile a_x is heavily biased by high level of fertility during the period between 50s and 80s. Therefore, the shorter is the analysed database of fertility, the more realistic results for the Czech population can be expected. The second model interpreted also in detail is the LC 1988. Its results are closest to the expected reality. In the case of fertility analysis by Lee-Carter model it is not a priority to analyse the longest time series, but the most stable ones. Our application of cross-sectional fertility rates by age is primarily intended to identify the parameters of changes for particular time periods. It is better for fertility projection to rely on shorter time series with the newest known development, because in the case of fertility this development is significantly affected by decision of people (opposed to mortality, where the development is influenced by mortality law and other factors). Human decision making is currently more social phenomenon than biological.

2 RESULTS

Firstly we universally look at 35 time series of logarithms of age-specific fertility rates as on the random walk process and estimate 35 drifts. Our second used approach is the ARIMA (0,1,1) process with drift. The drift is included into the model, because there is the higher probability that the model will not involve the autocorrelation. This is made even though there is a risk that many drifts in the model will be statistically insignificant (equal to zero). Estimated drifts for random walk models (in logarithms) are shown in Table 1.

Table 1 Estimated drifts (in logarithms) for individual Random Walk models. Each drift was calculated for individual time series of logarithms of age-specific fertility rates in Statgraphics Centurion XVI

Age	**15**	**16**	**17**	**18**	**19**	**20**	**21**
Drift	–0.001585	0.001479	–0.001347	–0.005822	–0.009456	–0.011595	–0.012407
Age	**22**	**23**	**24**	**25**	**26**	**27**	**28**
Drift	–0.013151	–0.012464	–0.011278	–0.009754	–0.007786	–0.005407	–0.003361
Age	**29**	**30**	**31**	**32**	**33**	**34**	**35**
Drift	–0.002307	–0.001229	–0.001165	–0.001894	–0.002166	–0.003306	–0.004398
Age	**36**	**37**	**38**	**39**	**40**	**41**	**42**
Drift	–0.005540	–0.007668	–0.009417	–0.011651	–0.013910	–0.015735	–0.019497
Age	**43**	**44**	**45**	**46**	**47**	**48**	**49**
Drift	–0.021675	–0.025519	–0.031128	–0.022158	–0.024409	–0.021511	–0.030460

Source: Author's calculation

Estimated components MA(1) for all 35 time series of logarithms of age-specific fertility rates are shown in Table 2. It is clear (grey highlighted values), that 9 of 35 parameters are statistically insignificant at 5% significance level. It is not a bad result for situation, when we selected a universal model for all series. Hyndman et al. (2002) have dealt with situations and issues, where the universal form of model and automatic forecasting was based on a much larger number of broken assumptions. We perform the diagnostics of two approaches on autocorrelation tests. The Box-Pierce test (Box, Pierce, 1970) is implemented in an automated process of automatic forecasting system in Statgraphics Centurion XVI. We test

the null hypothesis: there is no autocorrelation, and the results for the random walk model with drift are shown in Table 3, the results for the ARIMA (0,1,1) model with drift are in Table 4.

Table 2 Estimated parameters for individual ARIMA (0,1,1) models with drift. Each model was calculated for individual time series of logarithms of age-specific fertility rates in Statgraphics Centurion XVI. Most of drifts are statistically insignificant at the 5% significance level – but the drifts are included due to capture autocorrelation

Age 15	**Est.**	**s.e.**	**t-stat**	**P**	**Age 16**	**Est.**	**s.e.**	**t-stat**	**P**	**Age 17**	**Est.**	**s.e.**	**t-stat**	**P**
MA(1)	0.157	0.118	1.325	0.190	MA(1)	–0.240	0.107	–2.250	0.027	MA(1)	–0.216	0.106	–2.040	0.044
C	–0.002	0.017	–0.113	0.910	C	0.002	0.015	0.110	0.913	C	–0.001	0.015	–0.084	0.934
Age 18	**Est.**	**s.e.**	**t-stat**	**P**	**Age 19**	**Est.**	**s.e.**	**t-stat**	**P**	**Age 20**	**Est.**	**s.e.**	**t-stat**	**P**
MA(1)	–0.593	0.091	–6.492	0.000	MA(1)	–0.498	0.091	–5.493	0.000	MA(1)	–0.304	0.104	–2.940	0.004
C	–0.004	0.014	–0.311	0.756	C	–0.009	0.012	–0.766	0.446	C	–0.011	0.010	–1.129	0.262
Age 21	**Est.**	**s.e.**	**t-stat**	**P**	**Age 22**	**Est.**	**s.e.**	**t-stat**	**P**	**Age 23**	**Est.**	**s.e.**	**t-stat**	**P**
MA(1)	–0.285	0.106	–2.681	0.009	MA(1)	–0.287	0.103	–2.773	0.007	MA(1)	–0.321	0.102	–3.143	0.002
C	–0.012	0.011	–1.126	0.263	C	–0.013	0.009	–1.500	0.137	C	–0.012	0.008	–1.470	0.145
Age 24	**Est.**	**s.e.**	**t-stat**	**P**	**Age 25**	**Est.**	**s.e.**	**t-stat**	**P**	**Age 26**	**Est.**	**s.e.**	**t-stat**	**P**
MA(1)	0.010	0.108	0.093	0.926	MA(1)	–0.187	0.108	–1.736	0.086	MA(1)	–0.211	0.106	–1.997	0.049
C	–0.011	0.008	–1.460	0.148	C	–0.010	0.007	–1.361	0.177	C	–0.008	0.007	–1.127	0.263
Age 27	**Est.**	**s.e.**	**t-stat**	**P**	**Age 28**	**Est.**	**s.e.**	**t-stat**	**P**	**Age 29**	**Est.**	**s.e.**	**t-stat**	**P**
MA(1)	–0.143	0.108	–1.324	0.189	MA(1)	–0.123	0.108	–1.147	0.255	MA(1)	–0.282	0.103	–2.737	0.008
C	–0.005	0.008	–0.659	0.512	C	–0.003	0.008	–0.415	0.679	C	–0.002	0.009	–0.256	0.798
Age 30	**Est.**	**s.e.**	**t-stat**	**P**	**Age 31**	**Est.**	**s.e.**	**t-stat**	**P**	**Age 32**	**Est.**	**s.e.**	**t-stat**	**P**
MA(1)	–0.114	0.108	–1.058	0.293	MA(1)	–0.246	0.105	–2.355	0.021	MA(1)	–0.252	0.104	–2.416	0.018
C	–0.001	0.010	–0.125	0.901	C	–0.001	0,011	–0.105	0.916	C	–0.002	0.011	–0.173	0.863
Age 33	**Est.**	**s.e.**	**t-stat**	**P**	**Age 34**	**Est.**	**s.e.**	**t-stat**	**P**	**Age 35**	**Est.**	**s.e.**	**t-stat**	**P**
MA(1)	–0.249	0.104	–2.405	0.018	MA(1)	–0.281	0.103	–2.720	0.008	MA(1)	–0.281	0.102	–2.748	0.007
C	–0.002	0.012	–0.183	0.855	C	–0.003	0.012	–0.256	0.798	C	–0.004	0.013	–0.316	0.753
Age 36	**Est.**	**s.e.**	**t-stat**	**P**	**Age 37**	**Est.**	**s.e.**	**t-stat**	**P**	**Age 38**	**Est.**	**s.e.**	**t-stat**	**P**
MA(1)	–0.299	0.102	–2.927	0.004	MA(1)	–0.366	0.101	–3.619	0.001	MA(1)	–0.174	0.107	–1.618	0.109
C	–0.005	0.013	–0.420	0.676	C	–0.008	0.014	–0.557	0.579	C	–0.009	0.012	–0.745	0.458
Age 39	**Est.**	**s.e.**	**t-stat**	**P**	**Age 40**	**Est.**	**s.e.**	**t-stat**	**P**	**Age 41**	**Est.**	**s.e.**	**t-stat**	**P**
MA(1)	–0.305	0.103	–2.961	0.004	MA(1)	–0.370	0.100	–3.698	0.000	MA(1)	–0.248	0.107	–2.314	0.023
C	–0.011	0.014	–0.842	0.402	C	–0.014	0.014	–0.975	0.332	C	–0.015	0.016	–0.937	0.351
Age 42	**Est.**	**s.e.**	**t-stat**	**P**	**Age 43**	**Est.**	**s.e.**	**t-stat**	**P**	**Age 44**	**Est.**	**s.e.**	**t-stat**	**P**
MA(1)	–0.076	0.109	–0.694	0.489	MA(1)	0.051	0.108	0.473	0.638	MA(1)	0.083	0.108	0.772	0.443
C	–0.019	0.014	–1.351	0.180	C	–0.022	0.016	–1.347	0.182	C	–0.025	0.018	–1.430	0.156
Age 45	**Est.**	**s.e.**	**t-stat**	**P**	**Age 46**	**Est.**	**s.e.**	**t-stat**	**P**	**Age 47**	**Est.**	**s.e.**	**t-stat**	**P**
MA(1)	0.331	0.102	3.251	0.002	MA(1)	0.343	0.105	3.262	0.002	MA(1)	0.418	0.098	4.264	0.000
C	–0.031	0.018	–1.712	0.091	C	–0.024	0.022	–1.135	0.259	C	–0.024	0.023	–1.074	0.286
Age 48	**Est.**	**s.e.**	**t-stat**	**P**	**Age 49**	**Est.**	**s.e.**	**t-stat**	**P**					
MA(1)	0.610	0.087	7.043	0.000	MA(1)	0.368	0.102	3.615	0.001					
C	–0.021	0.032	–0.676	0.501	C	–0.027	0.022	–1.211	0.229					

Source: Author's calculation

Approach of random walk models with drift is clearly worse after the evaluation by autocorrelation tests. The development system was not well explained in 23 cases from 35, there was left too much unexplained variability and the residues are auto-correlated. Therefore the estimates will be statistically distorted and skewed. The situation is much better in ARIMA (0,1,1) model with drift. Although there

are many statistically insignificant drifts at the 5% significance level in this approach, their inclusion into the model fulfilled its goal. Only five models has its residues auto-correlated and therefore there is a real risk of statistical bias only in five cases. The difference between forecasts predicted by relatively bad and by relatively good model will be presented later in Figure 2 and the final comparison of all approaches will be provided in Figure 7.

Table 3 Diagnostic control of individual Random walk models with drift – Box-Pierce serial autocorrelation tests. Null hypothesis: There is no autocorrelation, unfortunately rejected in 23 cases from 35 at 5% significance level (grey highlighted values). This model is not good. (TC = Test Criterion)

Age	15	16	17	18	19	20	21
Box-Pierce TC	25.772	23.460	16.528	38.436	81.986	83.879	44.468
P-value	0.262	0.493	0.868	0.031	0.000	0.000	0.007
Age	22	23	24	25	26	27	28
Box-Pierce TC	73.289	43.349	19.348	16.990	25.259	17.286	26.033
P-value	0.000	0.009	0.733	0.849	0.392	0.836	0.352
Age	29	30	31	32	33	34	35
Box-Pierce TC	40.916	22.742	41.734	45.883	43.472	47.832	45.205
P-value	0.017	0.535	0.014	0.005	0.009	0.003	0.006
Age	36	37	38	39	40	41	42
Box-Pierce TC	47.239	48.553	50.099	49.782	56.903	30.147	56.238
P-value	0.003	0.002	0.001	0.002	0.000	0.180	0.000
Age	43	44	45	46	47	48	49
Box-Pierce TC	23.729	38.620	54.133	43.958	62.288	22.497	105.053
P-value	0.477	0.030	0.000	0.008	0.000	0.550	0.000

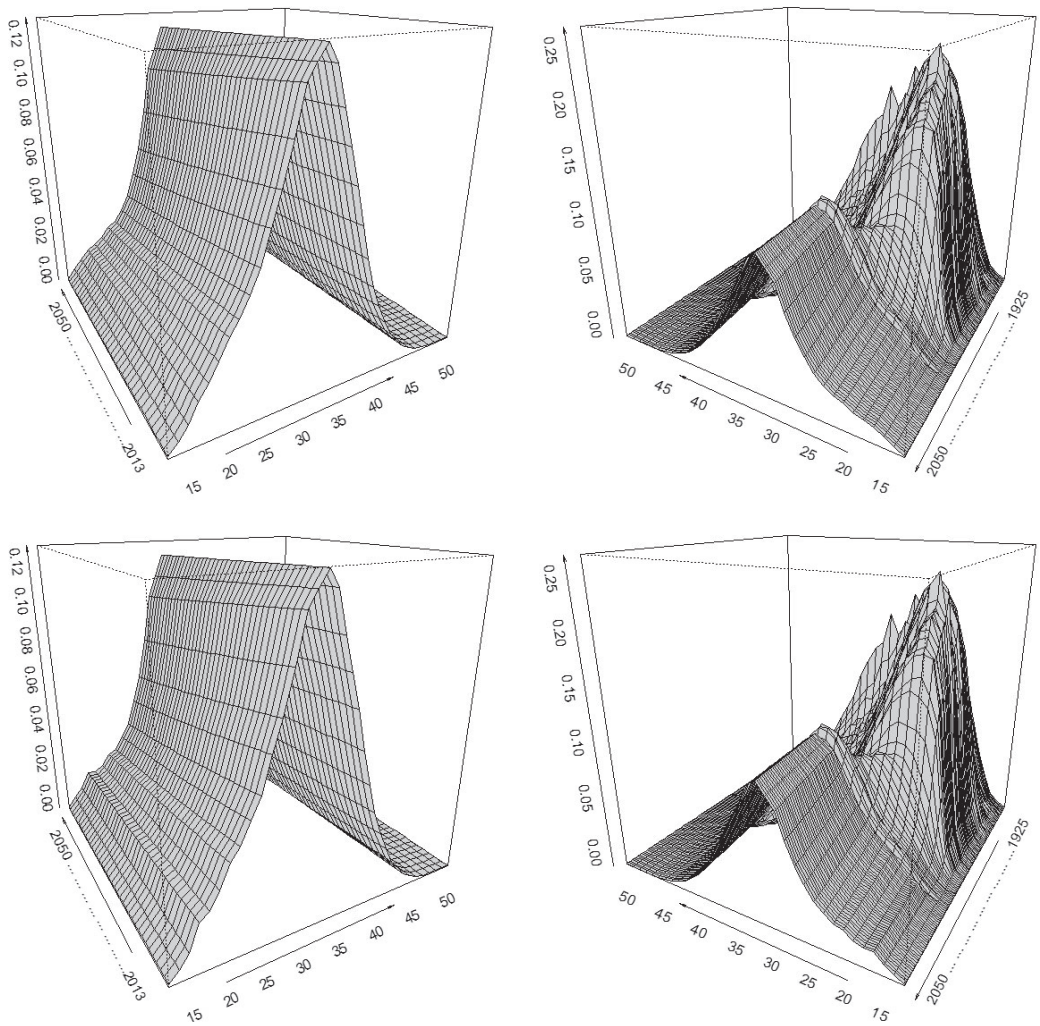
Source: Author's calculation

Table 4 Diagnostic control of individual ARIMA (0,1,1) models with drift – Box-Pierce serial autocorrelation tests. Null hypothesis: There is no autocorrelation, rejected only in 5 cases from 35 at 5% significance level (grey highlighted values). This model is much better than Random walk with drift (Table 3). (TC = Test Criterion)

Age	15	16	17	18	19	20	21
Box-Pierce TC	17.859	15.711	12.936	13.367	27.744	39.360	24.183
P-value	0.658	0.867	0.953	0.944	0.226	0.018	0.394
Age	22	23	24	25	26	27	28
Box-Pierce TC	45.409	24.697	19.289	13.538	21.157	15.633	24.896
P-value	0.004	0.366	0.684	0.939	0.571	0.871	0.356
Age	29	30	31	32	33	34	35
Box-Pierce TC	23.346	20.970	28.974	26.779	23.874	24.790	24.936
P-value	0.441	0.583	0.181	0.265	0.411	0.361	0.354
Age	36	37	38	39	40	41	42
Box-Pierce TC	21.860	24.809	35.622	24.346	28.506	19.886	61.300
P-value	0.529	0.360	0.045	0.385	0.197	0.649	0.000
Age	43	44	45	46	47	48	49
Box-Pierce TC	23.563	34.705	29.705	19.857	23.006	9.162	52.091
P-value	0.428	0.056	0.158	0.651	0.460	0.995	0.000

Source: Author's calculation

Figure 2 Forecasted values of age-specific fertility rates $f_{x,t}$ of Czech females for the period 2013 to 2050 by Random Walk model with drift (top left) and the empirical values of these rates for the period 1925 to 2012 with attached forecasts (top right). Forecasted values of age-specific fertility rates $f_{x,t}$ by ARIMA (0,1,1) models with drift and the empirical values of these rates with attached forecasts are shown bottom left, bottom right respectively

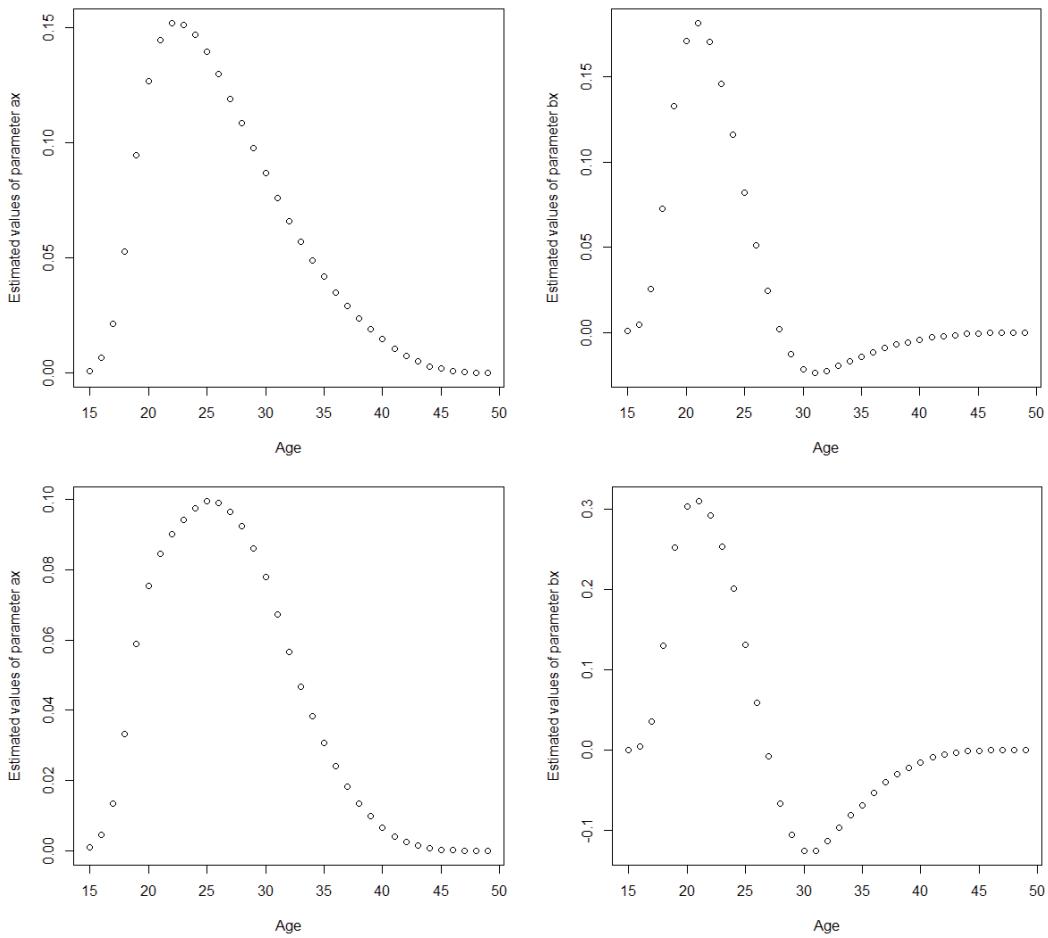


Source: CZSO (2013), author's construction and illustration

We estimate the parameters $\hat{a}_x$ (age-specific fertility profiles independent in time) and $\hat{b}_x$ (additional age-specific components determine how much each age group changes when k_t changes) for 4 Lee-Carter's models (LC 1925, LC 1948, LC 1968 and LC 1988) using the SVD method implemented in the package "demography" (Hyndman, 2012), which is developed for RStudio. We can see the parameters for LC 1925 and LC 1988 in the Figure 3, from which it is also clear the comparison between the different evolutions of these parameters. The age-specific fertility profiles independent of time ($\hat{a}_x$) are lower in the shortened model (LC 1988), because in the considered period there were already the fertility rates of Czech females

lower. Also this profile is deflected to the right (to the highest age groups). Given that the length of the analysed time series is shorter, the variability of the estimated additional age-specific components ($\hat{b}_x$) is higher, especially at the advanced ages. The fertility indices $\hat{k}_t$ (the time-varying parameters) were estimated for the period 1925 to 2012 (LC 1925) and 1988 to 2012 (LC 1988). The estimates are provided in the Figure 4. There were calculated the predictions up to the year 2050 to these estimates based on the methodological approach of ARIMA, (Box, Jenkins, 1970) and ran by "forecast" package in RStudio (Hyndman et al., 2002, Hyndman, 2012). Parameters of ARIMA models are displayed in Table 5.

Figure 3 Comparison of two Lee-Carter's models – The estimates of age-specific fertility profiles independent in time (parameter $\hat{a}_x$, left) and additional age-specific components determine how much each age group changes when $\hat{k}_t$ changes (parameter $\hat{b}_x$, right). Top charts represent the model based on data for the period 1925 to 2012, bottom charts the model based on data for the period 1988 to 2012



Source: Author's construction and illustration

It is clear from these predictions with 95% confidence intervals (which can be seen in Figure 4 too) that the LC 1988 model provides lower values of these estimates (decreasing trend). Confidence intervals are slightly wider at the case of LC 1988 model. Now we evaluate two Lee-Carter models on the basis of

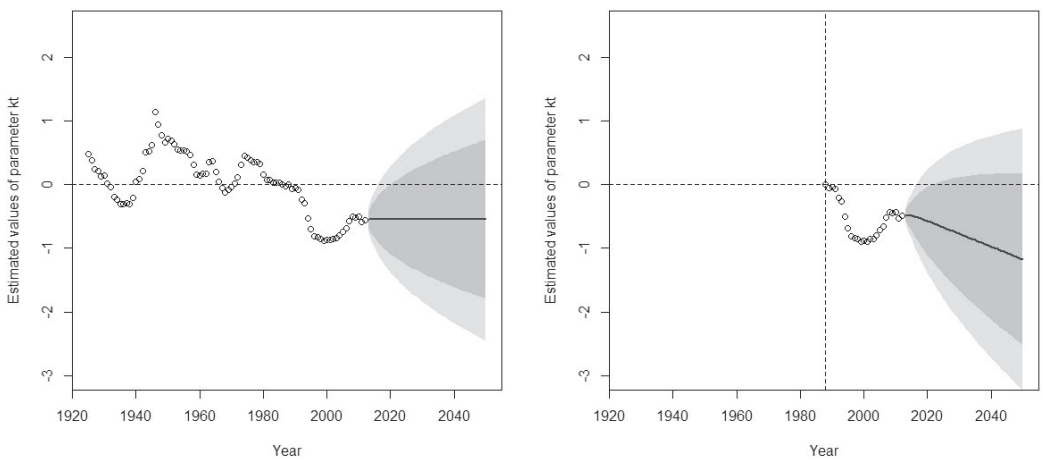
approach, which is presented by Charpentier, Dutang, (2012). Using RStudio we display the Pearson's residuals first for the LC 1925 and then for the LC 1988 model. Each model is evaluated on the basis of the residues by age x and of the residues at time t.

Table 5 Estimated parameters of two ARIMA models for parameter $\hat{k}_t$ of two Lee-Carter's models (LC 1925 and LC 1988)

Parameter k_t, Lee-Carter 1925			**Parameter k_t, Lee-Carter 1988**		
ARIMA (1,1,0) without drift			ARIMA (1,1,0) without drift		
Coefficients:			Coefficients:		
	AR(1)	**Drift**		**AR(1)**	**Drift**
	0.3494	x		0.6094	x
s.e.	0.1000	x	s.e.	0.1577	x
[t-stat]	3.4940	x	[t-stat]	3.8643	x
AIC= –142.98	AICc = –142.84	BIC= –138.05	AIC= –52.44	AICc= –51.24	BIC= –48.91

Source: Author's calculation

Figure 4 Comparison of two Lee-Carter's models – The estimates of the time-varying parameters $\hat{k}_t$ – the fertility indices. On the left side is the model based on data for the period 1925 to 2012, on the right side is the model based on data for the period 1988 to 2012



Source: Author's construction and illustration

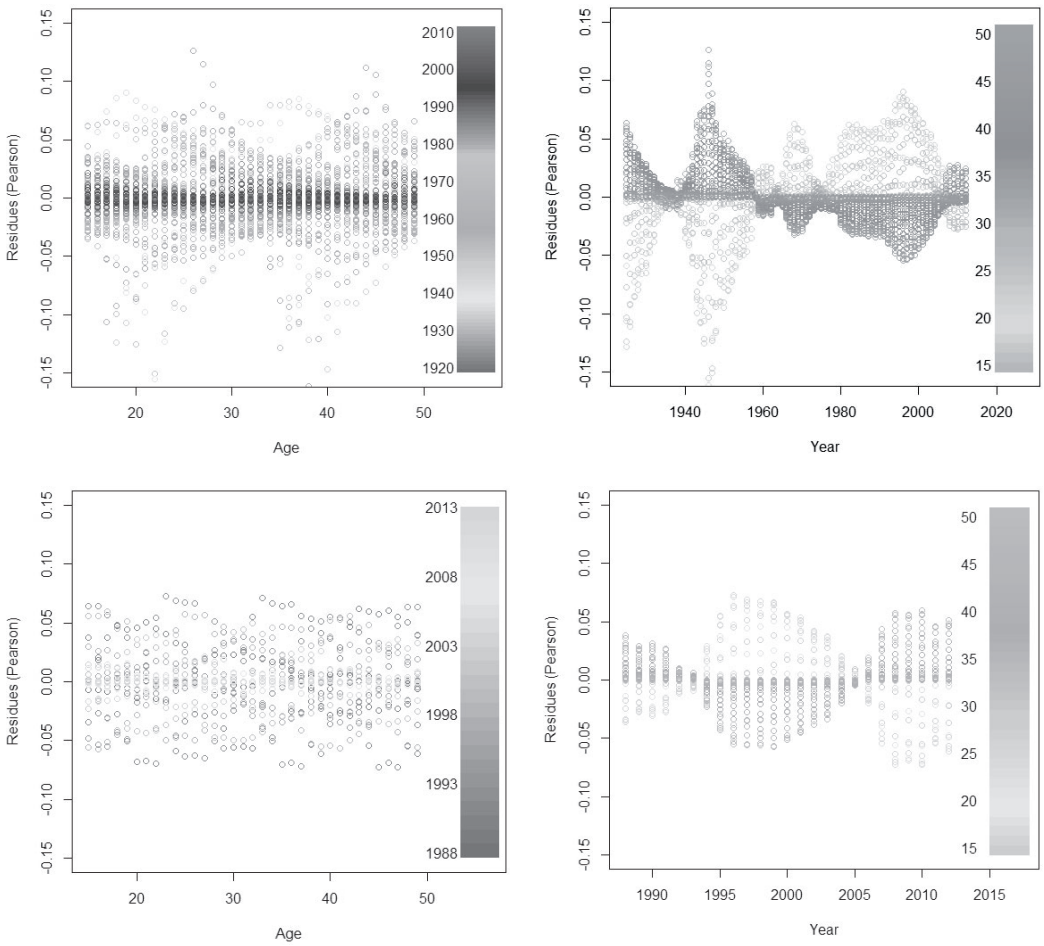
The most residues are concentrated around 0, the more variability is explained by the estimated model. The Pearson's residues for LC 1925 model are shown in the Figure 5 (top), where residues by age x are on the left side and the residues at time t are on the right side. Given that this model also includes the normalisation period, it is understandable that the residues will be much more variable than in the case of shortened model LC 1988. The residues of the shortened model are shown in the Figure 5 (bottom).

Based on the estimated parameters $\hat{a}_x$, $\hat{b}_x$ and $\hat{k}_t$ of two Lee-Carter's models we can estimate the future values of $f_{x,t}$ as:

$$f_{x,t} = \hat{a}_x + \hat{b}_x \cdot \hat{k}_t. \tag{11}$$

Estimated values (left) and the empirical values with the attached estimates (right) of $f_{x,t}$ based on LC 1925 model are displayed in 3D perspective chart in the Figure 6 (top). The estimated values based on the shortened model LC 1988 (left) and then the empirical values with these attached forecasts (right) are displayed below.

Figure 5 Diagnostic control of the Lee-Carter's model – Pearson's residues (model based on data for the period 1925 to 2012 – top charts, for the period 1988 to 2012 – bottom charts respectively)

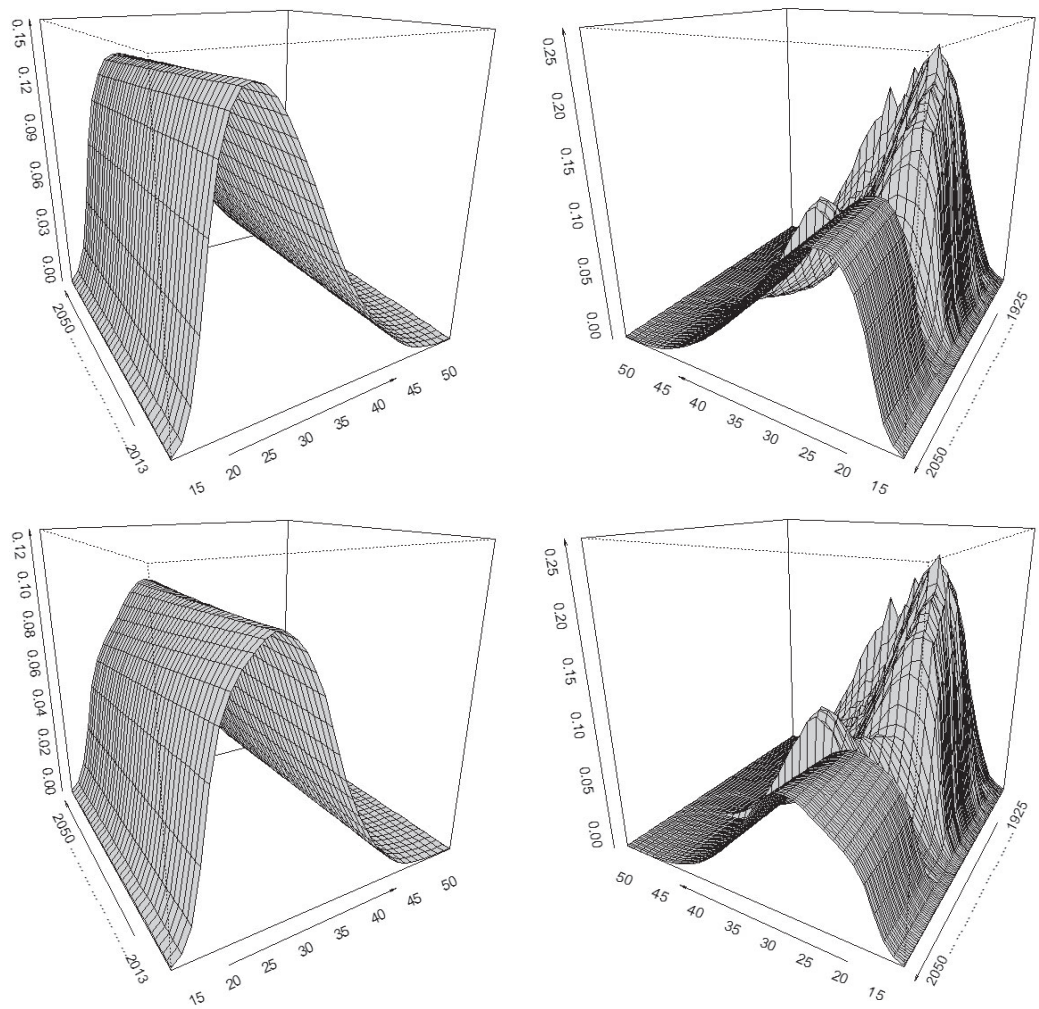


Source: Author's construction and illustration

The predicted values by LC 1925 model are unreasonably high (see Figure 6). This inadequacy is caused by non-robustness of Lee-Carter model for the case of fertility, because the prediction of this model is strongly influenced by the average fertility profile independent of time (parameter a_x). The average profile is deflected by high level of fertility of Czech females in the post-war period and during the pro-population policies implemented under the previous regime (Communist party of Czechoslovakia). Excessively high values of age-specific fertility rates ($f_{x,t}$) are not tied up with the empirical data. There is particularly a significant decline in the case of the distribution's mode, which is sharply and vigorously

returned to the lower ages. The predicted values $f_{x,t}$ by LC 1988 model are much lower. The parameter a_x is not so much affected by the high fertility rates arising in the period of normalization and the further projection looks more realistic. Unfortunately, even in this case there is not a fluent connection of predictions to the empirical data of $f_{x,t}$, because we can see that the mode of this distribution is not retained at its original level, but also moved back into the lower ages.

Figure 6 Forecasted values of age-specific fertility rates $f_{x,t}$ of Czech females for the period 2013 to 2050 by Lee-Carter's model based on full data matrix for the period 1925 to 2012 (top left), by Lee-Carter's model based on shortened data matrix for the period 1988 to 2012 (bottom left) respectively, and the empirical values of these rates for the period 1925 to 2012 with attached forecasts based on full Lee-Carter's model (top right), based on shortened model (bottom right) respectively

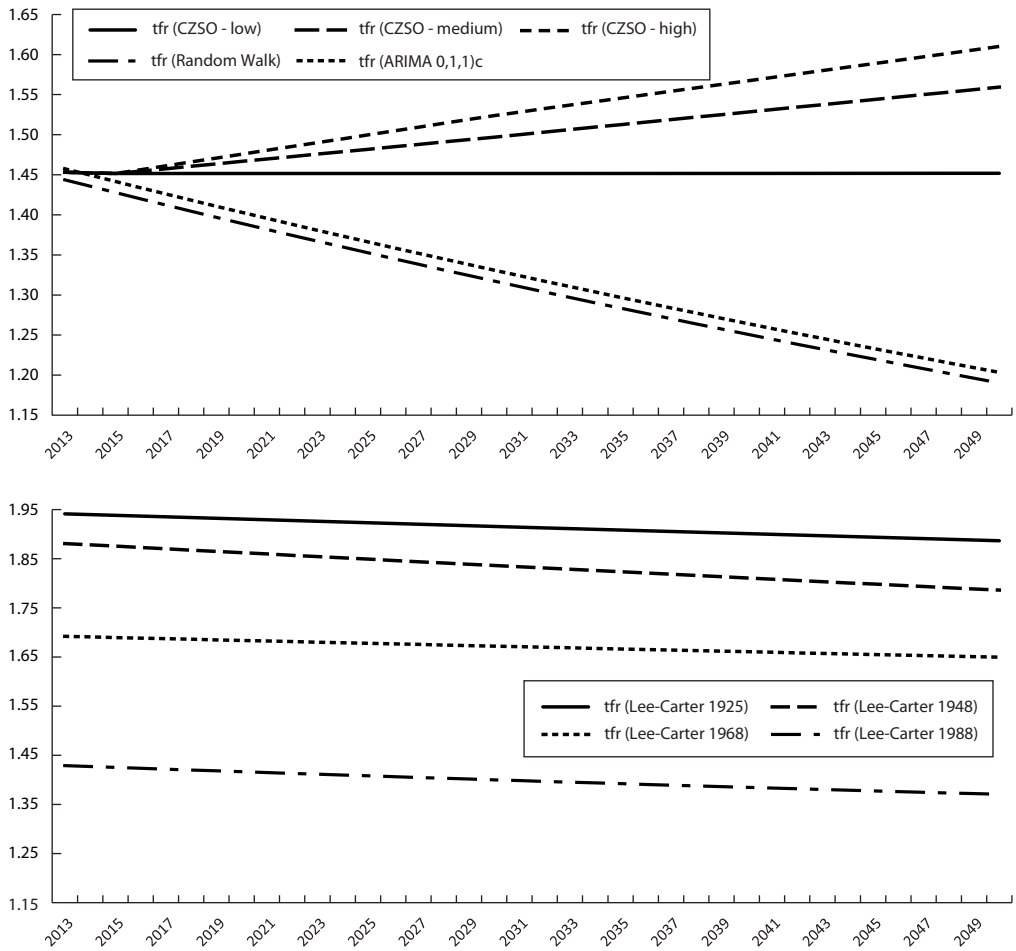


Source: CZSO (2013), author's construction and illustration

Following section compares the approaches of random walk models with drift, ARIMA (0,1,1) models with drift, LC 1925, LC 1948, LC 1968 and LC 1988 models, (regardless of the fact that models LC 1948

and LC 1968 have not been commented in detail). This evaluation is based on the results of total fertility rates calculated according to formula (2). It is clear from Figure 7 (right) that the models LC 1925, LC 1948 and LC 1968 are useless, because they provide the unrealistic values of total fertility for the situation in the Czech Republic. This fertility development does not follow the current fertility level and is quite skewed. Successive reduction of the input base of the Lee-Carter's model was in this situation quite useless, because the model is not able to respond to the dynamically changing fertility development of Czech females. Models of random walks with drift, ARIMA (0,1,1) models with drift, (and also LC 1988) provide more realistic results (please see Figure 7). Beginnings of prediction start at the current fertility level (1.45) and the predicted values for the future slowly decrease up to the values of 1.190, 1.203 and 1.371 respectively.

Figure 7 Forecast of the total fertility rates for Czech females up to the year 2050 based on prediction by the Czech Statistical Office in low, medium and high scenario and by the prediction by Random Walk model with drift (left chart) and four Lee-Carter's models (one based on data for the period 1925 to 2012, the second one for the period 1948 to 2012, the third for the period 1968 to 2012 and the fourth for the period 1988 to 2012 respectively



Source: CZSO (2013), author's construction and illustration

The prediction provided by ARIMA (0,1,1) model with drift seems to be the most acceptable. Model was relatively positively evaluated and its predicted values are probable. The approach of random walk models with drift, which was unfavourably evaluated, provides not so much different the expected future development. Some of the predicted values of $f_{x,t}$ will be probably skewed due to the consequences of poor model, but the difference should be really negligible in the summary.

We can also see in the Figure 7 the predictions that in their low, medium and high variant publishes the CZSO. Medium and high variant has a growing character, the low one has in all the time of horizon the same level as today. It is really questionable whether the future values of the total fertility rates in the Czech Republic actually rise, fall, or be rather constant at +/– 1.45.

DISCUSSION AND CONCLUSION

The aim of this paper was to construct forecasts of age-specific fertility rates $f_{x,t}$ of Czech females for the period 2013–2050 using different approaches. One of them was the Box-Jenkins methodology for modelling of 35 individual time series $f_{x,t}$ (x = 15–49) on an annual basis of 88 observations (t = 1925–2012). The second approach was the Lee-Carter model for identifying the major components explaining the level of fertility. The sensitivity of this model was to evaluate a total of 4 cases, where we gradually analysed and shortened the different length of $x \times T$ dimensional matrix of $f_{x,t}$. We concluded that the Lee-Carter model should be used for the shortest possible time series development as it is strongly influenced by fluctuations of the past. The average age-specific fertility profiles independent of time are affected by the different shape of the distribution $f_{x,t}$, created during the previous regime and the predicted values are due to this affection largely distorted. More realistic forecasts were provided by Box-Jenkins methodological approach. Looking at the shape of the distribution of predicted $f_{x,t}$, in Figure 2, there are no doubts that they are meaningful. These values were obtained using the found trend of development of each individual time series in the past, while the largest weights are set on the newest values. This implies that the estimated prediction describe the best expected future trend. Annex of the paper (Figure 9 to Figure 12) display all 4 calculated predictions in 5-year intervals using simple X-Y charts.

The values of $f_{x,t}$, which are expected by CZSO are optimistic. *We do not want to say that the expected values of future fertility of Czech females according to CZSO forecasts are wrong.* These expectations are based on expert judgments of professionals from different scientific disciplines and have its reasons. We would like to just point to the fact that the statistical trend, which does not take into account the expert discussed expectation is different – declining. Thus, there is a certain degree of probability that the expected future development of fertility will be rather "low" (pessimistic) variant of the CZSO. The goal for the future research is to find and elaborate a robust approach to modelling fertility in the Czech conditions based on literature review. This approach should not be so much affected by high average of fertility from previous regime. We will use the whole data matrix that is available for a better explanation of random effects.

ACKNOWLEDGEMENT

The author gratefully acknowledge to the Czech Science Foundation project No. P402/12/G097 DYME – "Dynamic Models in Economics" for supporting this paper.

References

ARLTOVÁ, M. *Stochastické metody modelování a předpovídání demografických procesů* [habilitation thesis]. Prague: University of Economics Prague, 2011.

ALDERS, M., DE BEER, J. Assumptions on fertility in stochastic population forecasts. *International statistical review*, 2004, Vol. 72, Iss. 1, pp. 65–79.

BELL, W. R., MONSELL, B. Using principal components in time series modelling and forecasting of age-specific mortality rates. In: *Proceedings of the American Statistical Association, Social Statistics Section*, 1991, pp. 154–159.

BELL, W. R. Comparing and assessing time series methods for forecasting age-specific fertility and mortality rates. *Journal of Official Statistics*, 1997, Vol. 13, No. 3, pp. 279–303.

BOX, G. E. P., JENKINS, G. *Time series analysis: Forecasting and control*. San Francisco: Holden-Day, 1970.

BOX, G. E. P., PIERCE, D. A. Distribution of the Autocorrelations in Autoregressive Moving Average Time Series Models. *Journal of the American Statistical Association*, 1970, Vol. 65, pp. 1509–1526.

CAPUTO, M., NICOTRA, M., GLORIA-BOTTINI, E. Fertility Transition: Forecast for Demography. *Human biology*, 2008, Vol. 80, Iss. 4, pp. 359–376.

CHARPENTIER, A., DUTANG, CH. *L'Actuariat avec R*. [working paper]. Paternite-Partage a lindentique 3.0 France de Creative Commons, Decembre 2012.

CZSO. *Projekce obyvatelstva* České *republiky do roku 2100* [online]. 2013. [cit.: 09.09.2014] <http://www.czso.cz/csu/2013edicniplan.nsf/publ/4020-13-n_2013>.

DOTLAČILOVÁ, P., ŠIMPACH, O., LANGHAMROVÁ, J. The Use of Polynomial Functions for Modelling of Mortality at the Advanced Ages. In: *Mathematical Methods in Economics 2014*. Olomouc: Palacký University in Olomouc, 2014, p. 174–179.

FIALA, T., LANGHAMROVÁ, J. What Rate of Fertility and Extent of Migration Would Be Needed for Stable Population Development in the Czech Republic in This Century? *Demografie*, 2012, Vol. 54, No. 4, pp. 382–404.

HYNDMAN, R. J., BOOTH, H. Stochastic population forecasts using functional data models for mortality, fertility and migration. *International Journal of Forecasting*, 2008, Vol. 24, Iss. 3, pp. 323–342.

HYNDMAN, R. J., KOEHLER, A. B., SNYDER, R. D., GROSE, S. A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, 2002, Vol. 18, Iss. 3, pp. 439–454.

HYNDMAN, R. J., ULLAH, MD. SHAHID Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, 2010, Vol. 51, Iss. 10, pp. 4942–4956.

HYNDMAN, R. J. *Demography: Forecasting mortality, fertility, migration and population data* [online]. 2010, R package v. 1.16. <http://robjhyndman.com/software/demography>.

LANGHAMROVÁ, J., FIALA, T. Změny reprodukčního chování v České republice a jeho důsledky. *Fórum sociální politiky*, 2014, Vol. 8, No. 3, pp. 23–25.

LEE, R. D., CARTER, L. R. Modeling and forecasting U. S. mortality. *Journal of the American Statistical Association*, 1992, Vol. 87, pp. 659–675.

LEE, R. D., TULJAPURKAR, S. Stochastic population forecasts for the United States: beyond high, medium, and low. *Journal of the American Statistical Association*, 1994, Vol. 89, pp. 1175–1189.

LI, N., WU, Z. Forecasting cohort incomplete fertility: A method and an application. *Population Studies*, 2003, Vol. 57, No. 3, pp. 303–320.

MELARD, G., PASTEELS, J. M. Automatic ARIMA Modeling Including Intervention, Using Time Series Expert Software. *International Journal of Forecasting*, 2000, Vol. 16, pp. 497–508.

MORGAN, SP., HAGEWEN, K. Is very low fertility inevitable in America? Insights and forecasts from an integrative model of fertility. In: *New Population Problem: Why Families in Developed Countries are Shrinking and What it Means, Penn State University Family Issues Symposia Series*, Mahwah: Lawrence Erlbaum Associates, 2005, pp. 3–28.

MYRSKYLAE, M., GOLDSTEIN, J. R., CHENG, YEN-HSIN. A. New Cohort Fertility Forecasts for the Developed World: Rises, Falls, and Reversals. *Population and Development review*, 2013, Vol. 39, Iss. 1, pp. 31–56.

PERISTERA, P., KOSTAKI, A. Modeling fertility in modern populations. *Demographic Research*, 2007, Vol. 16, No. 6, pp. 141–194.

RUEDA, C., RODRIGUEZ, P. State space models for estimating and forecasting fertility. *International Journal of Forecasting*, 2010, Vol. 26, Iss. 4, pp. 712–724.

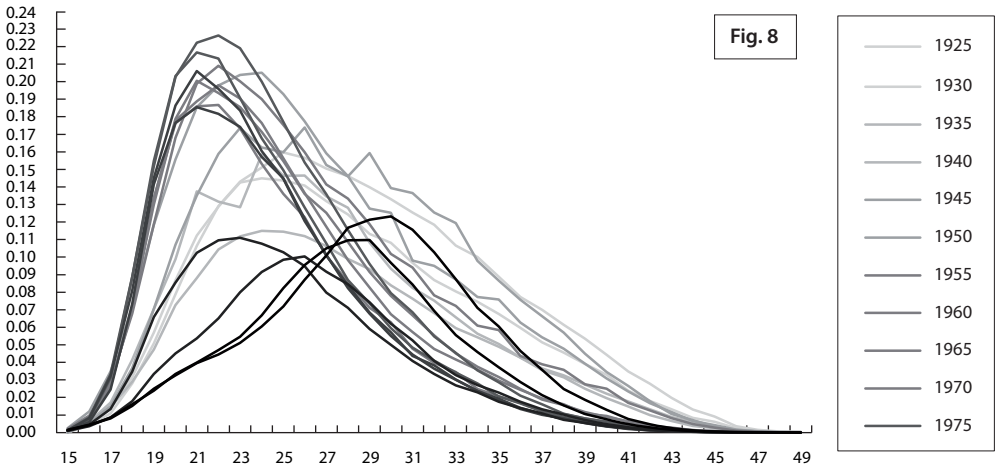
STAUFFER, D. Simple tools for forecasts of population ageing in developed countries based on extrapolations of human mortality, fertility and migration. *Experimental Gerontology*, 2002, Vol. 37, Iss. 8–9, pp. 1131–1136.

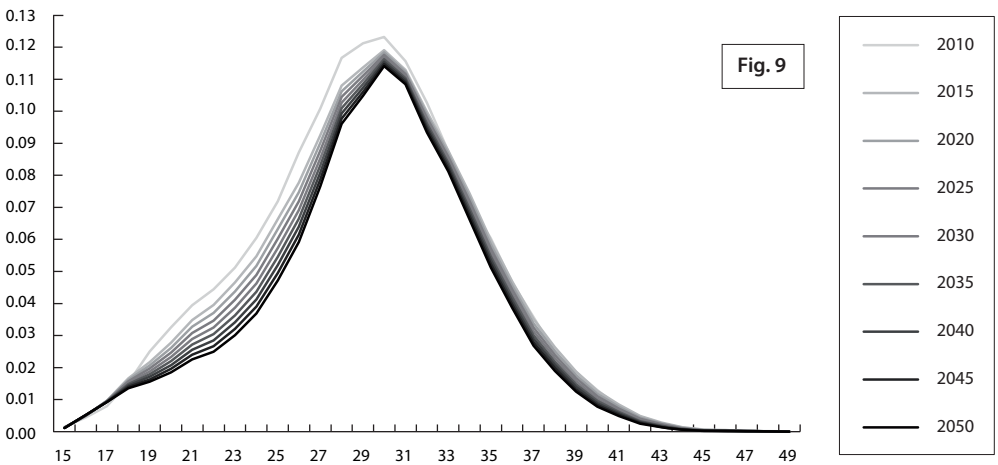
ŠIMPACH, O., DOTLAČILOVÁ, P., LANGHAMROVÁ, J. Effect of the Length and Stability of the Time Series on the Results of Stochastic Mortality Projection: An application of the Lee-Carter model. In: *ITISE 2014*. Granada: University of Granada, 2014, p. 1375–1386.

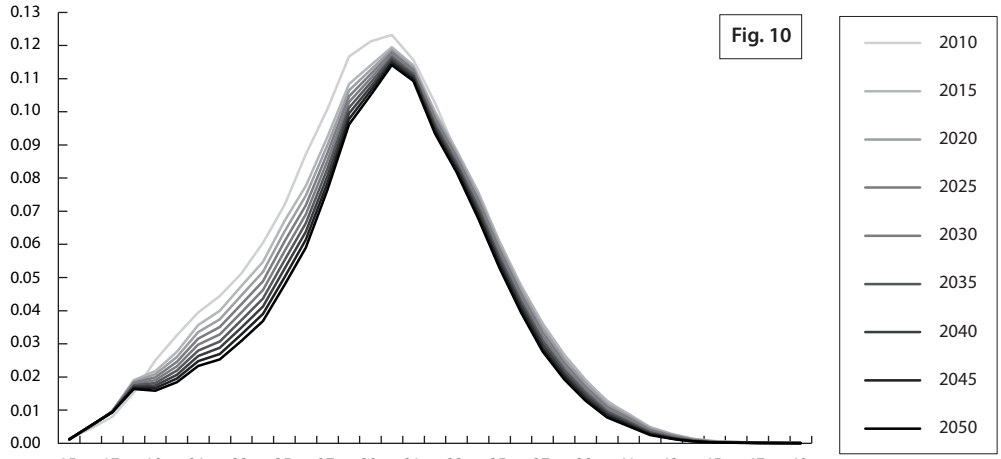
ŠIMPACH, O., LANGHAMROVÁ, J. Stochastic Modelling of Age-specific Mortality Rates for Demographic Projections: Two Different Approaches. In: *Mathematical Methods in Economics 2014*. Olomouc: Palacký University in Olomouc, 2014, pp. 890–895.

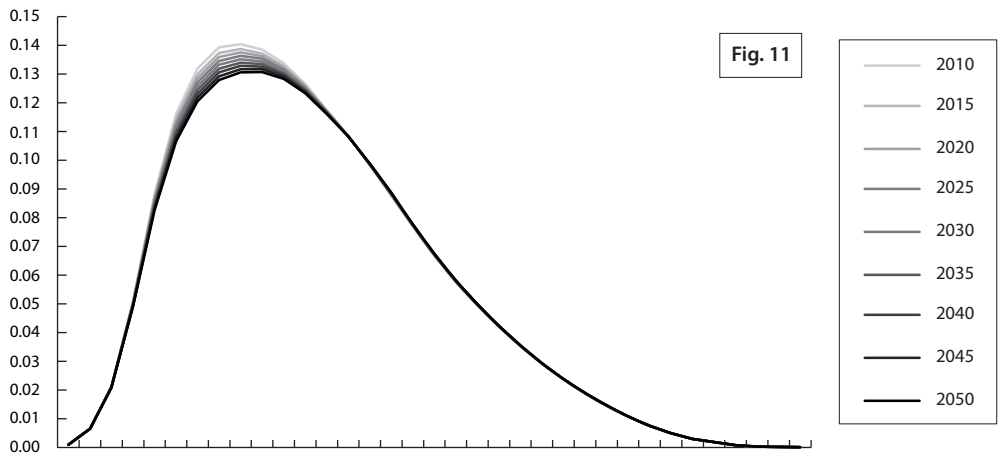
ANNEX

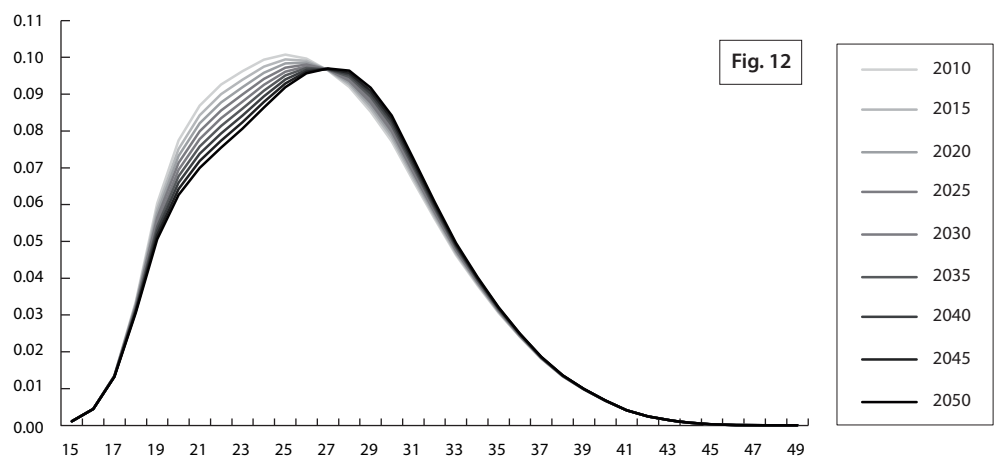
Figure 8–12 Empirical values of age-specific fertility rates $f_{x,t}$ of Czech females in the period 1925–2010 by 5years intervals (Figure 8). Forecasted values of age-specific fertility rates up to the year 2050 in 5years intervals by Random Walk model with drift (Figure 9), by ARIMA (0,1,1) models with drift (Figure 10), by Lee-Carter's model based on full data matrix for the period 1925 to 2012 (Figure 11) and by Lee-Carter's model based on shortened data matrix for the period 1988 to 2012 (Figure 12)











Source: CZSO (2013), author's construction and illustration

Two-Step Classification of Unemployed People in the Czech Republic

Zdeněk Šulc[1] | *University of Economics, Prague, Czech Republic*
Marina Stecenková | *University of Economics, Prague, Czech Republic*
Jiří Vild | *University of Economics, Prague, Czech Republic*

Abstract

The paper analyzes structure and behavior of unemployed people in the Czech Republic by means of latent class analysis (LCA) and CHAID analysis where the output of LCA serves as the input for CHAID. The unemployed are classified in two steps; for each step different characteristics are used. In the first step, respondents are split into latent classes according to their answers to questions concerning ways of searching for a new job. In the second step, CHAID analysis is performed with results obtained from LCA as a dependent variable. In the paper, data from periodical Labor Force Survey conducted in Czech Republic in spring 2011 are used. The results indicate that unemployed people in the Czech Republic can be divided into four segments: Active, Passive, Typical and Specific. A special attention is paid to extreme segments Active and Passive.

Keywords

Double-step classification, unemployment, latent class analysis, CHAID algorithm

JEL code

E24, C38

INTRODUCTION

Unemployment is thoroughly observed socio-economic phenomenon. High unemployment shows negative effects on economic, social and psychic situation of individuals, households, and generally, on a whole society. It is connected with many adverse effects, such as poverty, criminality, increased social (and other) expenditures of a state, problems caused by long-term unemployment and other, see Katrňák, Mareš (2007). Therefore, it is necessary to fully understand all aspects of unemployment and to make all possible arrangements to minimize its negative effects. In this paper, we focus on explaining the structure of unemployed people in the Czech Republic. This knowledge can help to improve the unemployment strategy of the Czech Republic and this procedure can also be applied in other countries.

The unemployment rate in the Czech Republic was influenced by economic crisis with a slight delay. The first signs of rising unemployment became obvious in 2009. Since then the unemployment rate holds the constant level around 7%,[2] in 2014 it dropped even below 6%. In comparison with other EU

[1] Faculty of Informatics and Statistics, Department of Statistics and Probability, Nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. Corresponding author: zdenek.sulc@vse.cz.
[2] Czech Statistical Office <www.czso.cz>.

countries, the Czech unemployment rate was among the lowest. In September 2014, it was lower only in Malta, Austria and Germany.[3] From among the Czech regions, the steadily lowest unemployment rate is in Prague and surrounding Středočeský region. It is just because many companies have their headquarters in these regions. On the other hand, the highest unemployment rate occurs in the Ústecký and Moravskoslezský regions because of high concentration of heavy industry from the past, which was liquidated or significantly reduced in the beginning of the 90's. The situation there improves very slowly. The unemployment rates in the rest of regions can be found in Statistical yearbook of the Czech Republic (2013).

An efficient hybrid methodology, which was introduced by Magidson and Vermunt (2004a), was applied to the data. This methodology combines features of the CHAID algorithm and latent class modelling. Using this methodology, the paper aims to achieve a better understanding of structure and behavior of unemployed people in the Czech Republic. The approach is performed in two steps. In the first step, the unemployed people are divided into several segments (latent classes) according to the way of seeking for a new job. In the second step, CHAID analysis is used to estimate the membership of each object in one of the latent classes, which were created in the first step. The second step is performed in two ways. First, there are used the same indicators, which served for construction of the latent classes. Second, CHAID analysis with five socio-demographic variables is performed with the aim to describe the created latent classes from a different perspective. Thus, we are able to reveal the background of people's attitudes in terms of other social characteristics. There are other interesting approaches to achieve the same goal. For example, Shaunna and Muthen (2009) introduced an approach based on posterior probabilities from LCA. The rest of the paper is organized as follows: Section 1 describes how the unemployment is measured. Section 2 introduces the principles of LCA and in Section 3 the principles of CHAID analysis. Section 4 offers practical application which is performed on data from periodical Labor Force Survey conducted by Eurostat in the Czech Republic in 2011. The results are summarized in Conclusion.

1 UNEMPLOYMENT AND THE LABOR FORCE SURVEY

Indicator of unemployment used in the research comes from Labor Force Survey (LFS). LFS is organized by Eurostat and it provides results which are comparable in all countries of EU and some other countries. It is conducted on a random sample of private households in which all persons are surveyed according to their labor status (employed, unemployed or inactive). The labor status is determined by ILO conditions. According to them, the unemployed are at least 15 years old and have to meet the following conditions during a reference week to be involved in the survey. First, they do not work as paid employees during a reference week. Second, they have been actively looking for a job within a four-week period ending with a reference week. Third, they are available for paid employment within two weeks since the end of a reference week. The indicator of unemployment is computed as the ratio of unemployed people according to ILO conditions to the total labor force.

The data from Eurostat can be broken down by many additional criteria like age, nationality, full-time/part-time employment etc. Thus, they allow for a detailed view at unemployment issues, which suits well for the research performed in this paper.

2 LATENT CLASS ANALYSIS

Latent class analysis (LCA) is a latent variable model in which a categorical latent variable is constructed comprising set of discrete, mutually exclusive latent classes. It is described in detail for example from Haberman (1979) or Vermunt and Magidson (2004). The latent variable is not measured directly but indirectly by means of two or more categorical observed variables. LCA is used to divide objects (respondents, individuals) into homogeneous groups (clusters) according to their characteristics, see Collins

[3] Eurostat <ec.europa.eu/eurostat>.

and Lanza (2010). It is often used in questionnaire surveys, where it helps to identify groups of similar respondents. For each object, probability of belonging (prevalence) to each latent class is computed. Usually, the object is assigned to the latent class with the highest prevalence. When performing LCA, there is a constriction that all input variables have to be mutually independent.

In LCA, two sets of parameters are estimated – the latent class membership probabilities (prevalence), which represent the proportion of the researched population in particular latent class, and the item-response probabilities, which express probability of a particular response to an observed variable, conditional on latent class membership.

2.1 Model fit and model selection

The goodness-of-fit of an estimated latent class model is usually tested by the likelihood-ratio chi-squared statistic which is compared to a critical value of chi-square distribution. To approximate the G2 statistics with the chi-square distribution, it is necessary that each cell of a contingency table has a sufficient number of cases. This situation may not been fulfilled when there are too many observed variables or the observed variables have too many categories compared to the total sample size. One possibility to solve such a problem is to estimate p-values with a bootstrapping technique.

Other methods of evaluating model fit are based on information criteria which penalize models with a higher number of parameters. The most common information criteria are Akaike information criterion (AIC) and Bayesian information criterion (BIC), see Akaike (1973), Schwartz (1978) or Bozdogan (1987). A model with minimum value of AIC or BIC is then selected.

2.2 Classification

When a latent class model is estimated, it can be used for assigning objects to latent classes. The classification is based on their response pattern and posterior probability of membership in each of the latent classes. The classification probabilities are obtained using Bayes rule with estimates of prevalence and item response probabilities. The most common classification rule is modal assignment, which assigns each individual to the latent class with the highest posterior probability. Correctness of the classification can be measured by the entropy, which is measured on a zero to one scale with a value one indicating that the individuals are perfectly classified into latent classes. Generally, higher values indicate better classification of objects.

3 CHAID ANALYSIS

The Chi Square Automatic Interaction Detection Analysis (CHAID), which was originally introduced by Kass (1980), belongs to group of classification methods using unsupervised learning. It provides much better results in comparison with not so appropriate methods for response modelling, such as discriminant analysis or multiple regression, see Madgison (2006). The main aim of the analysis is to find a combination of variables, which best explains the outcome of a given dependent variable. The main output displays relationships among variables in a hierarchical form. CHAID offers two main fields of use. The first one is to determine relationships among variables; the second one to classify objects into classes of dependent variable. It can also be used to find interactions among observed predictors; thus, it can serve to improve results of other data analyses (e. g. classification by neural networks). More detailed view into the CHAID analysis is well described e.g. in Tufféry (2011) or in Madgison (1994). CHAID analysis has several advantages, especially, it does not impose almost any conditions on data. The method is determined for categorical data primarily; however, continuous variables can be categorized into a suitable number of intervals. CHAID visualizes results easily in form of a classification tree. The CHAID algorithm builds multiple classification trees, in which each node can be further divided into two and more nodes. Tree structure allows to reveal interesting interactions among variables.

3.1 Algorithm

The algorithm is performed in three steps. In the first one, each predictor is tested whether all its categories are significantly different in terms of a dependent variable. If not, these categories are merged into so called reduced categories. The algorithm continues iteratively until all pairs of categories are treated as statistically different and the initial contingency table turns into reduced contingency table. In the second step, the algorithm searches for a predictor, which best differentiates values of dependent variable. Criterion for the selection is an adjusted p-value of chi-square test of independence in a reduced contingency table. The adjusted p-value takes into account a number of categories of the newly reduced predictor (Bonferroni adjustment). The predictor with the lowest adjusted p-value is then chosen as a branching variable and each of its categories builds one node of a classification tree. In the first two steps, Pearson's Chi-square test of independence in a contingency table is used. In the third step, the first and the second step are recurring until any of the rules for stopping the algorithm is satisfied, see Kass (1980). The rules for stopping are following: first, there is no other predictor, for which an adjusted p-value would be lower than a given significance level; second, the maximal number of levels of the classification tree was achieved; third, the minimal number of observations cannot be reached in any new node. These criteria must be set before the analysis. There are no strict rules for setting these parameters. Usually, such a combination of parameters securing sufficient interpretation of a classification tree is selected.

3.2 Evaluation of Classification Results

Classification quality indicators are calculated from a confusion matrix. Rows of this matrix represent instances in actual categories, whereas columns correspond to estimated categories by the model. In this article, the average probability of correct classification is denoted as p. Average probability is calculated as the proportion of well-classified objects to all objects.

4 RESULTS

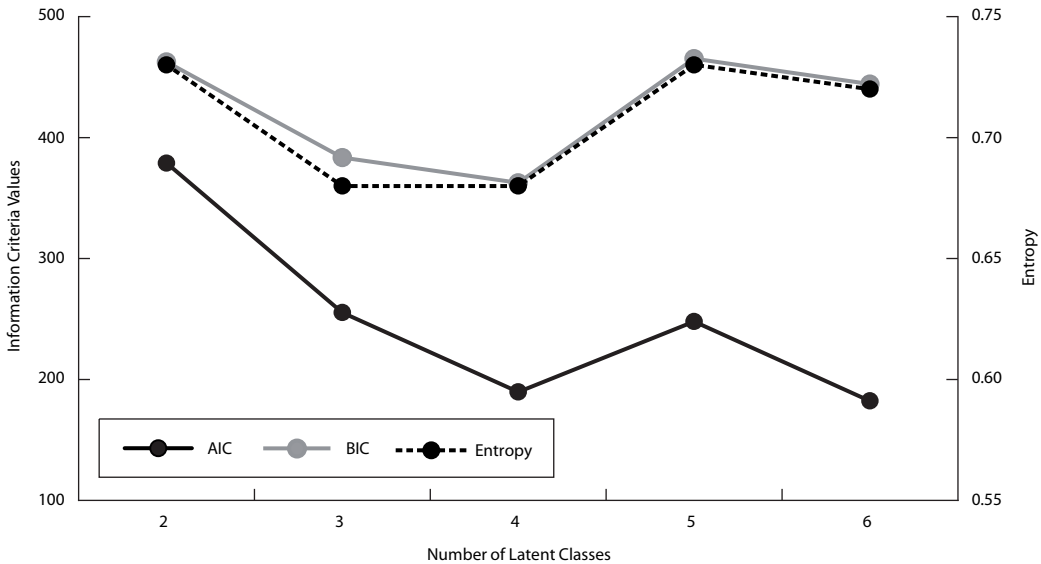
4.1 First step – latent class analysis on survey questions

LCA was performed on seven bivariate questions concerning ways of search for a new job. People specified whether they search for a new job being registered at the Labor Bureau, through a personal agency, by directly contacting a potential employer, getting contacts from acquaintances, answering to advertisements or just browsing through them or by other way. According to information criteria, the most suitable model appears to be the model with four latent classes, even though the entropy favors a higher number of latent classes (Figure 1).

Table 1 shows parameter estimates, both prevalence and item-response probabilities. The latent class *Typical* has the highest prevalence, 0.58. That means there is a 58% probability that a given object is going to be classified into this class. Because of this high probability, we assume the unemployed people in this cluster represent the typical behavior when searching for a job. The remaining three classes, which have prevalence under 0.20, are to be profiled deeper according to item-response probabilities. In this case, the item-response probabilities express conditional probability of *Yes* answer to each of the surveyed questions under the condition of being in a particular latent class. As the fourth latent class (*Active*) reaches maximal values across all questions, it clearly refers to those who search for a job by all possible ways. We can describe the behavior of the unemployed in this latent class as *Active*. On the other side, the third latent class comprises people who only registered at the Labor Bureau and some of them ask their acquaintances. The unemployed in this cluster can be called *Passive*. The unemployed people in the last latent class are very similar to those in the *Typical* class; they are very likely to ask their acquaintances or contact the potential employer directly, but unlikely to the Typical class, they browse less through advertisements and nearly do not respond to them. We will refer to this cluster

of unemployed as to *Specific*. Distribution of respondents into latent classes is following: *Typical* 69%, *Active* 12%, *Specific* 12%, *Passive* 7%.

Figure 1 Information criteria and entropy of the various LCA models



Source: Authors' computation

Table 1 Share of positive answers to job search questions and item-response probabilities

Way of a job search	Share of unemployed	Latent class			
		Typical	Specific	Passive	Active
Labor Bureau	0.90	0.92	0.82	0.85	0.96
Personal Agency	0.17	0.06	0.19	0.03	0.60
Employer	0.81	0.81	0.98	0.02	0.96
Acquaintances	0.91	0.98	0.80	0.40	0.98
Ads – Active	0.35	0.35	0.02	0.01	0.80
Ads – Passive	0.83	1.00	0.42	0.11	0.99
Other	0.37	0.33	0.26	0.14	0.70
Prevalence		0.58	0.16	0.08	0.18

Source: Authors' computation

4.2 Second step – CHAID on latent classes

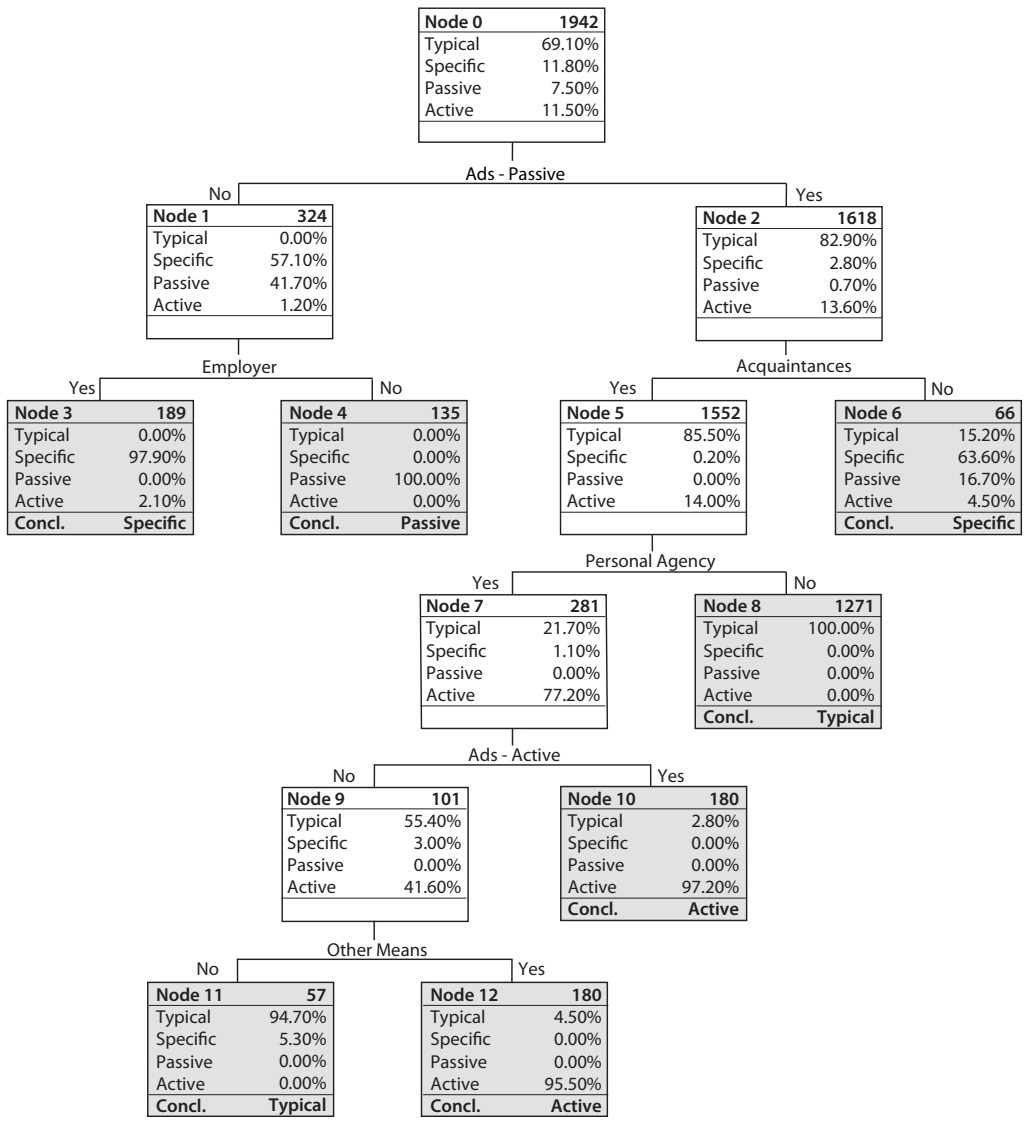
The four latent classes, revealed in LCA, are further examined by CHAID analysis, which is performed in two ways. First, it is performed on the same variables which were used for construction of latent classes, i.e. according to a method of a job search. Second, as its input serve socio-demographic variables, which are described in Table 2.

CHAID analysis based on the first way produces a classification tree which correctly classifies 98% of unemployed persons into one of four defined latent classes.[4] Thus, the four latent classes can be profiled

[4] The model has following setting. Significance level for splitting nodes and merging categories is 0.05, maximal tree depth is 5 levels, a minimal number of cases in parent nodes is 80, and a number of cases in child nodes is 40.

by considering both the information from the classification tree (Figure 2) and item-response probabilities in Table 1. *Typical* individuals do not use services of personal agencies very much. They usually check advertisements and ask people in their surroundings. These statements are true for 95% persons in this group (path to node 10). The *Specific* individuals do not answer to advertisements and their most common way of searching for a job is to contact the potential employer directly (path to node 3). Such behavior is characteristic for 80% of persons in this group.

Figure 2 Classification tree for CHAID model with seven explanatory variables related to the way of a job search for classification into four classes



Source: Authors' computation

An important question is why the *Specific*, unlike *Typical*, do not browse through advertisements and do not answer them. Unfortunately, we are unable to find it out from the data we have available. We can assume two basic hypotheses – they do not want to or they cannot. More probable is the second one. These people could have such education and skills which can be used only in very specific and specialized fields (pilots, craft workers …) where it is not usual to use advertisement. It is also possible that they live in small towns where searching for a job is based more on direct personal contact than on any mediators.

The second approach to CHAID analysis is based on construction of a classification tree with socio-demographic variables. It leads to more accurate profiles of unemployed people in each of latent classes. Due to the fact that the class *Typical* contains 70% of objects, the classification process of four classes is very difficult, because a majority of the unemployed falls very easily into this class. Therefore, only extreme classes *Active* and *Passive*, which are easier to differentiate, will be taken into account for further analysis.

Following variables were chosen as an input for the second approach of CHAID analysis: Actual economic status (labelled *Status* in the analysis), education, usual economic status (*Usual_status*), a number of persons in the responder's household (*Num_of_Person*) and responder's age interval (*Age_interval*). Values of all the variables are in Table 2.

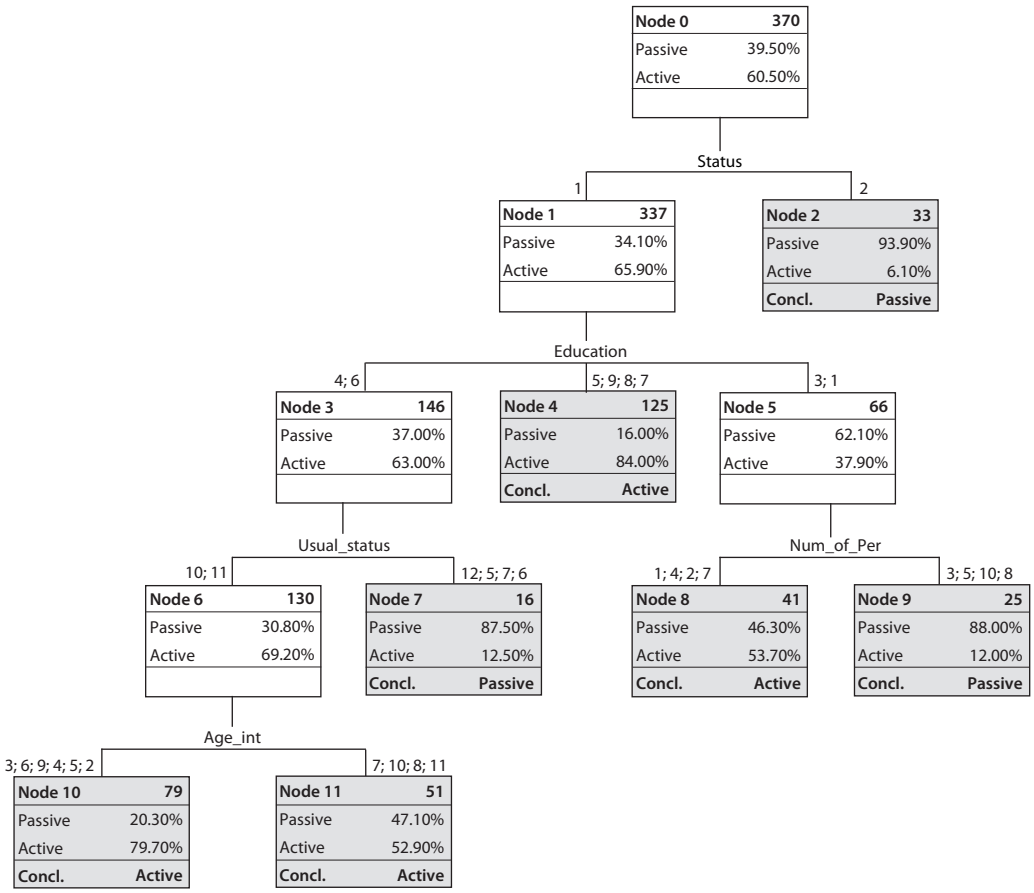
Table 2 Social-demographic variables and their categories

Variables	Categories of variables
Status	**1** – Is seeking any paid job; **2** – S/he has already found job which will start later.
Education	**1** – No education; **2** – Elementary school; **3** – Secondary school; **4** – High school without leaving exam; **5** – High school with leaving exam; **6** – Postsecondary school; **7** – Vocational school of tertiary education and conservatory; **8** – University with bachelor degree; **9** – University with master degree; **10** – University with doctor degree.
Usual_status – main status	**1** – On maternity leave; **2** – In education; **3** – On parental leave; **4** – In early retirement; **5** – In regular retirement; **6** – In retirement due to full disability; **7** – In retirement due to partial disability; **8** – Permanently disabled from healthy reasons; **9** – Works; **10** – Is unemployed; **11** – In the household; **12** – Other.
Num_of_Person	Number of persons (0 – 15).
Age_interval	**1** – "0-14"; **2** – "15-19"; **3** – "20-24"; **4** – "25-29"; **5** – "30-34"; **6** – "35-39"; **7** – "40-44"; **8** – "45-49"; **9** – "50-54"; **10** – "55-59"; **11** – "60-64"; **12** – "65+".

Source: Authors' computation

On the first branching level, persons are divided according to the fact whether they have already found a job, or they still have not found it, see Figure 3. It explains, why 21% of *Passive* have negative attitude towards searching for a job – they have already found it. On the second branching level, the unemployed are divided into three groups according to a level of education. The first group consists of persons with lower or no education, the second group contains persons with secondary education and the third group brings together persons with higher and tertiary education. It is obvious that the higher level of education the more active are the unemployed people in searching for a job. On the third level, the group of low-educated people is further divided into smaller and bigger households. The borders between these groups are not accurate; there are visible rather general tendencies. In smaller households, ratios of *Active* and *Passive* persons are nearly equal, whereas bigger households contain almost 90% of *Passive* persons. This is very typical for gipsy families which have a lot of members and that there is a high unemployment rate in this ethnic group. The group of people with the secondary education is divided according to a usual status of a person. One group is made of people who are usually unemployed or stay in a household. The ratio of *Active* persons is 69% in this group. The second branch mostly consists of people who are in a regular retirement, in a retirement due to partial disability or in retirement due to full disability. This group is dominated by the passive unemployed.

Figure 3 Classification tree for a CHAID model with 5 socio-demographic explanatory variables for classification into two classes



Source: Authors' computation

DISCUSSION AND CONCLUSION

In the paper, attitudes of Czech unemployed persons towards a job search was analyzed. Latent class analysis identified four main attitudes. Besides the *Typical* attitude (prevalence of 58%), which is characteristic by browsing through advertisements, asking acquaintances and directly contacting potential employers, there are groups of *Active* (18%) and *Passive* (8%) unemployed. *Specific* class (16%) consists of people who take nearly typical attitude but they do not use advertisements. After classifying the responders into latent classes, we analyzed them further by CHAID analysis where the constructed latent variables served as an input. First, we built a classification tree with the same variables which were used for the construction of the latent classes. This helped us to characterize the attitudes more precisely. We found out that 95% of people with *Typical* attitude browsed through advertisements and asked people in their surroundings but did not use services of personal agencies, whereas 80% of *Specific* unemployed did not read advertisements but contacted the potential employer directly. Second, we built a classification tree in which the attitude towards a job search was analyzed by the means of socio-demographic variables. Because of high prevalence of the *Typical* group, which would cause low classification perfor-

mance of a classification tree, we decided to look for socio-demographic differences between the two extreme attitudes, i.e. Active and Passive unemployed. The results showed that *Passive* attitude towards a job search can be found mainly among people with lower education, people living in bigger households and retired people.

Further questions arose during performing the analysis in the paper. First, what is the differentiating factor between the *Specific* and the *Typical class*. It would be very useful to determine why the unemployed people from the Specific group do not use advertisements. Another question is stability of these attitudes in time. Next time, we would like to use longitudinal data to perform longitudinal LCA which would follow development of latent classes in time. Last but not least is the question how to adjust the data where one of the latent classes is prevailing (in our case the *Typical*) so that the classification performance would not be worse when incorporating this latent class into the analysis. All these questions are subjects of our next research.

ACKNOWLEDGMENTS

This work was supported by the University of Economics, Prague under Grant IGA F4/104/2014.

References

AKAIKE, H. *Information theory as an extension of the maximum likelihood principle, Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, 1973.

BOZDOGAN, H. Model selection and Akaike´s information criterion (AIC): the general theory and its analytical extensions. *Psychometrika*, 1987, 52 (3), pp. 345–370.

COLLINS, L. M., LANZA, S. T. *Latent class and latent transition analysis for the aocial, behavioral, and health sciences.* New York: Wiley, 2010.

CZSO. *Statistical yearbook of the Czech Republic 2013* [online]. Prague: Czech Statistical Office, 2013. [cit.20.9.2014]. <http://www.czso.cz/csu/2013edicniplan.nsf/engt/0E002418FB/$File/000113.pdf>.

HABERMAN, S. J. *Analysis of qualitative data.* New York: Academic Press, 1979.

KASS, G. V. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics,* 1980, 29, pp. 119–127.

KATRŇÁK, T., MAREŠ, P. Thee employed and the unemployed in the Czech labour market between 1998 and 2004. *Czech Sociological Review,* 2007, 43 (2), pp. 281–304.

MAGIDSON, J. Improved statistical techniques for response modelling. Progression beyond regression. *J. of Direct Marketing,* 1988, 2, pp. 6–18.

MAGIDSON, J. The CHAID approach to segmentation modelling: chi-squared automatic interaction detection. In: BAGOZZI, RICHARD P., eds. *Advanced Methods of Marketing Research.* Oxford: Blackwell, 1994.

MAGIDSON, J., VERMUNT, K. J. *An extension of the CHAID tree-based segmentation algorithm to multiple dependent variables.* In: Proceedings of the 28th Annual Conference of the Gesellschaft für Klassifikation e.V. University of Dortmund, March 9–11, 2004, pp 176–183.

SCHWARTZ, G. Estimating the dimension of a model. *The Annals of Statistics,* 1978, 6 (2), pp. 461–464.

SHAUNNA, C., MUTHEN, B. *Relating latent class analysis results to variables not included in the analysis,* 2009 [online]. [cit.10.5.2014]. <http://statmodel2.com/download/relatinglca.pdf>.

TUFFÉRY, S. *Data mining and statistics for decision making.* United Kingdom: Wiley, 2011.

VERMUNT, K. J., MAGIDSON, J. Latent class analysis. *The Sage Encyclopedia of Social Science Research Methods.* NewBury Park: Sage Publ., Inc., 2004.

Testing the Effectiveness of Some Macroeconomic Variables in Stimulating Foreign Trade in the Czech Republic, Hungary, Poland and Slovakia

Marcin Salamaga[1] | *Cracow University of Economics, Cracow, Poland*

Abstract

Some concepts of contemporary econometrics depart from the arbitrary division of variables into endogenous and exogenous. In the estimation process of the econometric model or in prediction process, it may be important to test weak or strong exogeneity of variables. In the foreign trade modelling, we often deal with variables between which there may be feedback. Thus, the causality of variables in the classic sense is not always obvious and it should be tested to facilitate the proper specification of foreign trade models. This article is aimed at testing the exogeneity of selected macroeconomic variables used in foreign trade models, based on Visegrád Group countries. Exogeneity tests made in this paper are based on the results of the VEC and VAR models, which enabled to explain dynamic relations between variables in foreign trade. The results of this research can be helpful for determining the structure of actual links between variables, estimation of proper models and forecasting of variable values.

Keywords

Foreign trade, exogeneity, Visegrád Group, Granger causality test, VAR model, VEC model

JEL code

C32, F49

INTRODUCTION

The a priori classification of variables as exogenous and endogenous ones may be troublesome in econometric modelling due to the complexity of economic phenomena occurring in the contemporary world and the existence of feedback between economic values. Such "traditional" classifications are claimed to

[1] Department of Statistics, Rakowicka 27, 31-510 Cracow, Poland. E-mail: salamaga@uek.krakow.pl.

be arbitrary, omitting some variables (the Liu critique) or having parameters in multi-equation models dependent on the values of exogenous variables (the Lucas critique) (Maddala, 2006).

Dilemmas appear also at the level of determining the causality of variables. This happens, for example, in the sphere of international trade where causality of variables understood in traditional way is not always obvious. It can be tested to what extent export is a cause of import or to what extent import is a cause of export. Contemporary econometrics, which offers the Granger causality test, faces these dilemmas. The Granger test assumes *a priori* that there is no distinction between exogenous and endogenous variables. When it comes to the properties of the econometric model and the character of variables occurring in it, it is important to distinguish weak exogeneity, strong exogeneity and superexogeneity of variables (Engle, Hendry, Richard, 1983). Weak exogeneity is required for model estimation; strong exogeneity is needed for forecasting and superexogeneity guarantees that model parameters remain unchanged in relation to variables.

Until recently, the modelling of foreign trade was dominated by "traditional" approach compliant with the interpretation of econometrics offered by the Cowles Commission for Research in Economics in the mid-20th century. This approach is based on the assumption that both the causal structure of the model and the division of variables into endogenous and exogenous are determined in advance and do not require testing. This perception of the role of variables is obvious both in classic models of foreign trade and in models proposed in the new theory of foreign trade which is used to explain new trends in the international exchange of goods and services (e.g. intra-industry trade) (Cieślik, 2000).

The dynamics of contemporary economic phenomena and the progressing globalisation cause that the role of some macroeconomic values in mutual cause-effect relations does not need to be always determined strictly. As a result, it may be difficult to keep the assumption that the foreign trade model has a pre-determined cause-effect structure. Thus, it seems that the testing of variable causality in foreign trade models and the testing of their exogeneity is authorised or even necessary.

In the recent decade or so, there have been attempts in the literature to look at the modelling of foreign trade from the perspective of new econometrics. Granger causality of macroeconomic factors in the models of international trade (Liu, Wang, Wei, 2001, Hsiao, Hsiao, 2006; Sharma, Kaur, 2013; Simionescu, 2014) is tested most frequently whereas the exogeneity of variables in such models (especially strong exogeneity and superexogeneity) is examined less often (Strauß 2002, Mehrara, Firouzjaee, 2011).

This article focuses on the examination of weak and strong exogeneity of the most frequent variables in foreign trade models. Exogeneity tests used in this paper are based on the results of the VEC and VAR models, which enabled to explain dynamic relations between variables in foreign trade. The results of this research can be helpful when determining the structure of actual links between variables, estimating proper models and forecasting variable values. Calculations were done separately for data for Poland, Czech Republic, Slovakia and Hungary, so the Visegrád Group countries. These countries were chosen for their similar economic structure, analogous economic potential, comparable social and economic conditions, and similar economic history in at least the last several decades.

Research results included in this article have led to the formulation of methodological conclusions on the structure of models and conclusions on the effectiveness of some macroeconomic instruments in the development of international trade in individual countries.

1 METHODOLOGY

In the traditional approach corresponding to the approach of the Cowles Foundation, the concept of exogeneity was most frequently identified with predeterminedness or strict exogeneity. A predetermined variable is independent of the current and future values of the random component of an econometric equation. But in case of strict exogeneity, the variable in the equation does not depend on current, past and future values of the random component (Charemza, Deadman, 1997). One of the charges brought

against this perception of exogeneity is the fact that it is not stated precisely in relation to what the exogeneity of variables should be considered. Another concept of exogeneity was formulated by Engle, Hendry and Richard (1983), who distinguished weak exogeneity, strong exogeneity and superexogeneity. This article covers weak and strong exogeneity of variables.

Variables are weakly exogenous if they carry all the information necessary for the consistent estimation of the parameters of conditional value in relation to these variables. The function of density $\boldsymbol{f}$, which can be presented as a quotient of conditional probability density function of process $\boldsymbol{f_1}$ and marginal probability density of $\boldsymbol{f_2}$ variable process (Osińska, Kośko, Stempińska, 2007), can prove helpful in defining exogeneity in a more formal way:

$$f(Z_t \mid \mathrm{Z}_{\mathrm{t\text{-}i}};\Theta) = f_1(Y_t \mid \mathrm{Y}_{\mathrm{t\text{-}i}}, \mathrm{X}_{\mathrm{t}};\Theta_1) \cdot f_2(X_t \mid \mathrm{Z}_{\mathrm{t\text{-}i}};\Theta_2), \tag{1}$$

where:

Θ – parameter vector whereas $\Theta = [\Theta_1; \Theta_2]$,

X_t, Y_t, Z_t – variables.

Variable Z_t is weakly exogenous in relation to function $\Psi = \mathrm{h}(\Theta)$ if:

- Ψ is the function of only parameters Θ_1 ($\Psi = \mathrm{h}(\Theta_1)$), so the model related to conditional density f_1 is sufficient for estimating parameters,
- there are no mixed conditions for both parameters Θ_1, Θ_2 simultaneously, which means that they are variation free.

Z_t variable is strongly exogenous with respect to variable Y_t for the function of parameters Ψ if:

- variable Z_t is weakly exogenous for Ψ,
- Y_{t-1} is not a Granger cause of Z_t (Charemza, Deadman, 1997).

Weak exogeneity is tested in a slightly different way in the vector autoregression model (VAR model) and in the vector error correction model (VEC model).

VAR model can be presented as follows (Osińska, Kośko, Stempińska, 2007):

$$X_t = A_0 D_t + \sum_{i=1}^{s} A_i X_{t-i} + \varepsilon_t, \tag{2}$$

where:

$\boldsymbol{X_t}$ – observation vector of current values of analysed processes,

$\boldsymbol{D_t}$ – vector containing determinist components (e.g. trend, seasonality),

$\boldsymbol{A_i}$ – matrix of autoregressive operators of individual processes,

$\boldsymbol{A_0}$ – parameter matrix with vector D_t, components,

ε_t – vector of residual processes,

k – VAR model rank.

The existence of time series cointegration is justified by the application of VEC model which can be written, in general, in the following way (Johansen, 1995; Kusideł, 2000):

$$\Delta Z_t = \Psi_0 D_t + \sum_{i=1}^{s} \Pi_i \Delta Z_{t-i} - \Pi Z_{t-1} + \zeta_t, \tag{3}$$

where:

Π – long-run multiplier matrix, $\Pi = \sum_{i=1}^{k} A_i - I$,

Π_i – short-run multiplier matrix, $\Pi_i = -\sum_{i=j+1}^{k} A_i$,

$\boldsymbol{A_i}$ – parameter matrices of polynomial delay operator,

$\boldsymbol{D}_t$ – vector containing determinist components (e.g. trend, seasonality),
$\boldsymbol{X}_t$ – observation vector of the values of analysed processes,
Ψ_0 – coefficient matrix with determinist components of vector $\boldsymbol{D}_t$,
ξ_t – white noise.

The testing procedure of weak exogeneity requires the estimation of the boundary process for variable X_t and conditional one for variable Y_t represented respectively by equations (4) and (5) (Osińska, Kośko, Stempińska, 2007).

$$X_t = \sum_{i=1}^{g} c_i Y_{t-i} + \sum_{i=1}^{h} d_i X_{t-i} + \varepsilon_t, \tag{4}$$

$$\beta Y_t + \gamma_0 X_t + \sum_{i=1}^{p} \gamma_i Y_{t-i} + \alpha \hat{\varepsilon}_t = u_t, \tag{5}$$

where:
X_t, Y_t –variables,
α, β, γ_0, γ_i, c_i, d_i –parameters,
ε_t, u_t – random terms.

The testing of weak exogeneity of X_t variable with respect to variable Y_t involves model estimation (4) and calculation of residuals $\hat{\varepsilon}$. These residuals are then put into the model (5) as realisations of a new explanatory variable. Statistical significance of the parameter next to the added "residual variable" is tested in the estimated model (5). There are no grounds for rejecting the zero hypothesis that the parameter α in question is equal to 0, which means that variable X_t is weakly exogenous in relation to variable Y_t.

The testing of weak exogeneity of the distinguished variable is also conducted in two stages in VEC models. Firstly, it is examined if boundary processes for all variables in this equation do not contain the same mechanism of error correction in the short-run equation of VEC model. Thus, the parameter next to the error correction term is subjected to significance testing (Charemza, Deadman, 1997). In the case the error correction term occurs with different delays in the VEC model, the total significance test F can be conducted for coefficients standing with lagged variables in the error correction mechanism. When there are no grounds for rejecting the zero hypothesis of the insignificance of error correction term, it is omitted in the relevant VEC model equation and the equation is subject to estimation in this form. The further procedure of exogeneity testing is similar as in the case of VAR model. Residuals from the equation with the removed error correction term are put into a different equation of VEC model as new variables. After estimating the second equation, the parameter next to "residual variable" is tested. The Granger causality test is an extra element of the examination of strong exogeneity of variables.

In this test, the causality of variable X (Y) in relation to variable Y (X) occurs when the total influence of the current and delayed explanatory variable X (Y) on Y (X) is statistically significant. According to the tested zero hypothesis, parameters next to the explanatory variable and its delays are equal to zero in confrontation with the alternative hypothesis which states that these parameters are significantly different from zero. Test statistics take the following form (Osińska, 2007):

$$G = \frac{(S^2(u_t) - S^2(\eta_t))/q}{S^2(\eta_t)/(T-m)}, \tag{6}$$

where:
S^2 (u_t) – residual variance in the model without a variable whose causality is tested,
S^2 (η_t) – residual variance in the model with a variable whose causality is tested,
q – number of lags of explanatory variable,
T – size of the sample,
m – number of parameters in the model with an explanatory variable.

Statistics G with the correct zero hypothesis have the Fisher-Snedecor distribution with q and T-m degrees of freedom, respectively.

The rejection of zero hypothesis proves that the explanatory variable is a Granger cause of the explained variable.

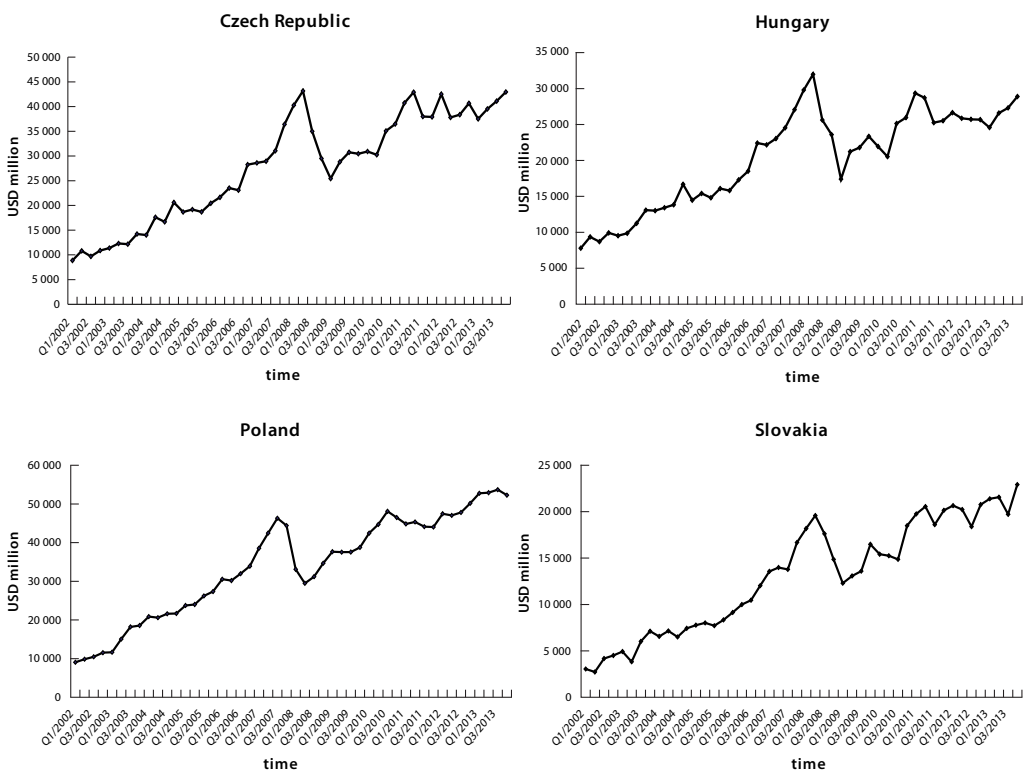
Weak exogeneity of the distinguished variable occurs when in consequence of the conducted test there are no grounds for rejecting the zero hypothesis according to which the parameter of residual variable is insignificant. Strong exogeneity of variables is a reason to use dynamic inference with an estimated model.

2 RESULTS AND DISCUSSION

2.1 Results of stationarity and cointegration test

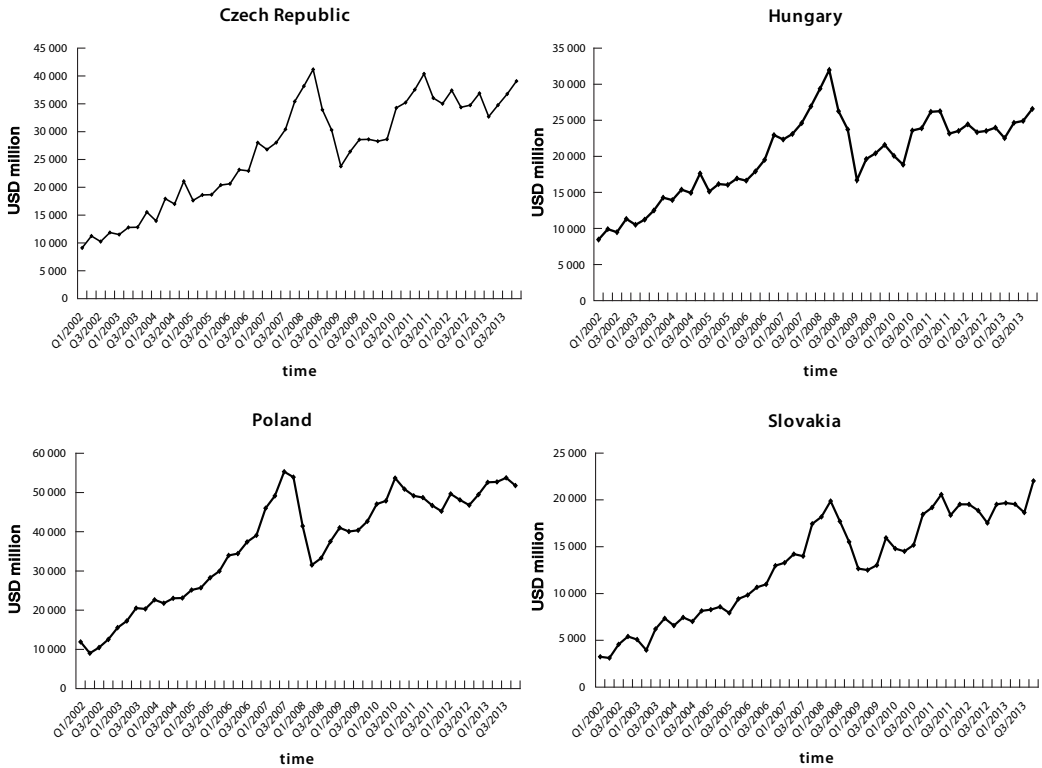
Exogeneity testing considers variables used most frequently in foreign trade models, so export (Ex), import (Im), gross domestic product (GDP), foreign direct investment inflow (FDI), exchange rate (FX). Figure 1 presents time series of export and Figure 2 presents time series of import in Visegrád Group countries. As both figures show, in all countries, the long-term levels of exports and imports are generally increased, although the growth rates are different in each country. We can observe also seasonal and random fluctuations of exports and imports, and the amplitude of fluctuations depends on the period of observation. The strongest export and import fluctuations we can see in Slovakia in 2002–2004, and in Hungary, Czech Republic in 2007–2009.

Figure 1 Export in Visegrád Group countries



Source: Own construction based on data from CEIC database

Figure 2 Import in Visegrád Group countries



Source: Own construction based on data from CEIC database

It is worth noting that the relationships between exports, imports and GDP have been studied in the traditional models of foreign trade. As an example we can mention the gravity model of foreign trade. According to this model the trade flow between countries grows in proportion to the GDP of these countries, and inversely proportional to the square of the distance between them. According to economic theory (Kojima, 1975; Ozawa, 1992) foreign direct investment can enhance export or weaken it. Relations between FDI and exports in economic Ozawa's theory are determined according to the level of economic development of the country. The exchange rate, in turn, affects the competitiveness of exports primarily in short term. As we know, an appreciation of domestic currency can cause the growth of a country's exports. In this research, the domestic currency exchange rate expressed in USD was taken into account. The calculations used data of the Visegrád Group countries from the integrated macroeconomic database CEIC[2] (*A Euromoney Institutional Investor Company)*. The analysis covered time series of variables for the period from Q1 2002 to Q4 2013.[3] VAR or VEC models were treated as a basis for testing the relation between these variables.

[2] <http://www.ceicdata.com>, retrieved: 10/10/2014. CEIC database is a collection of data from, e.g. national statistical offices, central banks, Eurostat, the International Monetary Fund.

[3] Restriction to data for 2002–2013 was caused by the availability of complete quarterly data. Export, import, GDP and FDI values are expressed in USD million.

The choice between these models depended on the results of stationarity and cointegration testsof time series representing individual variables.

The augmented Dickey-Fuller test (ADF test) was used to examine the stationarity of time series (Osińska, Kośko, Stempińska, 2007). The results of the test are presented in Table 1 (*p*-values are given in brackets). The ADF test verifies the null hypothesis, which states that the time series is stationary. Based on the results included in Table 1, it can be stated that time series of original variables were not stationary (with the probability values higher than 0.05) except for FDI time series in Poland and in Slovakia. The first difference of these variables are stationary in all cases. Thus, time series of tested variables are integrated of order 1, except for the two situations mentioned. An optimal variables lag in models is determined according to the Akaike information criterion (AIC) and the Schwarz criterion (BIC). The cointegration of suitable time series was examined and the number of cointegrating vectors was determined with the use of the Johansen test (Johansen, 1991). Table 2 presents the results of information criteria and the Johansen test with the recommendation of the model used to examine the dependence of all pairs of variables in which export is one of the variables. Based on the results in Table 2, it can be stated that only in the Czech Republic the occurrence of a cointegrating vector was detected for each pair of variables, so the VEC model was estimated for these variables. In other countries, the lack of cointegrating vectors was stated for at least one pair of variables and in these situations it entailed the need to estimate the VAR model.

Below is an example of the VEC model built for the import and export of the Czech Republic. Based on the results of a preliminary analysis of dynamic structure of exports and imports time series there is introduced a deterministic linear trend with time variable t and seasonal effects S_i to the VEC model, so $D_t = [S_1, S_2, S_3, t, const]$ i $X_t = [Ex, Im]$. EC1 means the error correction mechanism representing the short-term adjustments process to long-term equilibrium. In brackets, below the parameter estimates, are *p*-values.

$$\Delta Ex = \underset{(0.550)}{-603.57} + \underset{(0.012)}{2.15}\,\Delta Ex\_d - \underset{(0.023)}{2.00}\,\Delta \mathrm{Im}\_d - \underset{(0.007)}{2241.58}\,S_1 - \underset{(0.266)}{2013.69}\,S_2 - \underset{(0.446)}{918.30}\,S_3 + \underset{(0.009)}{335.60}\,t - \underset{(0.014)}{2.16}\,EC1 \tag{7}$$

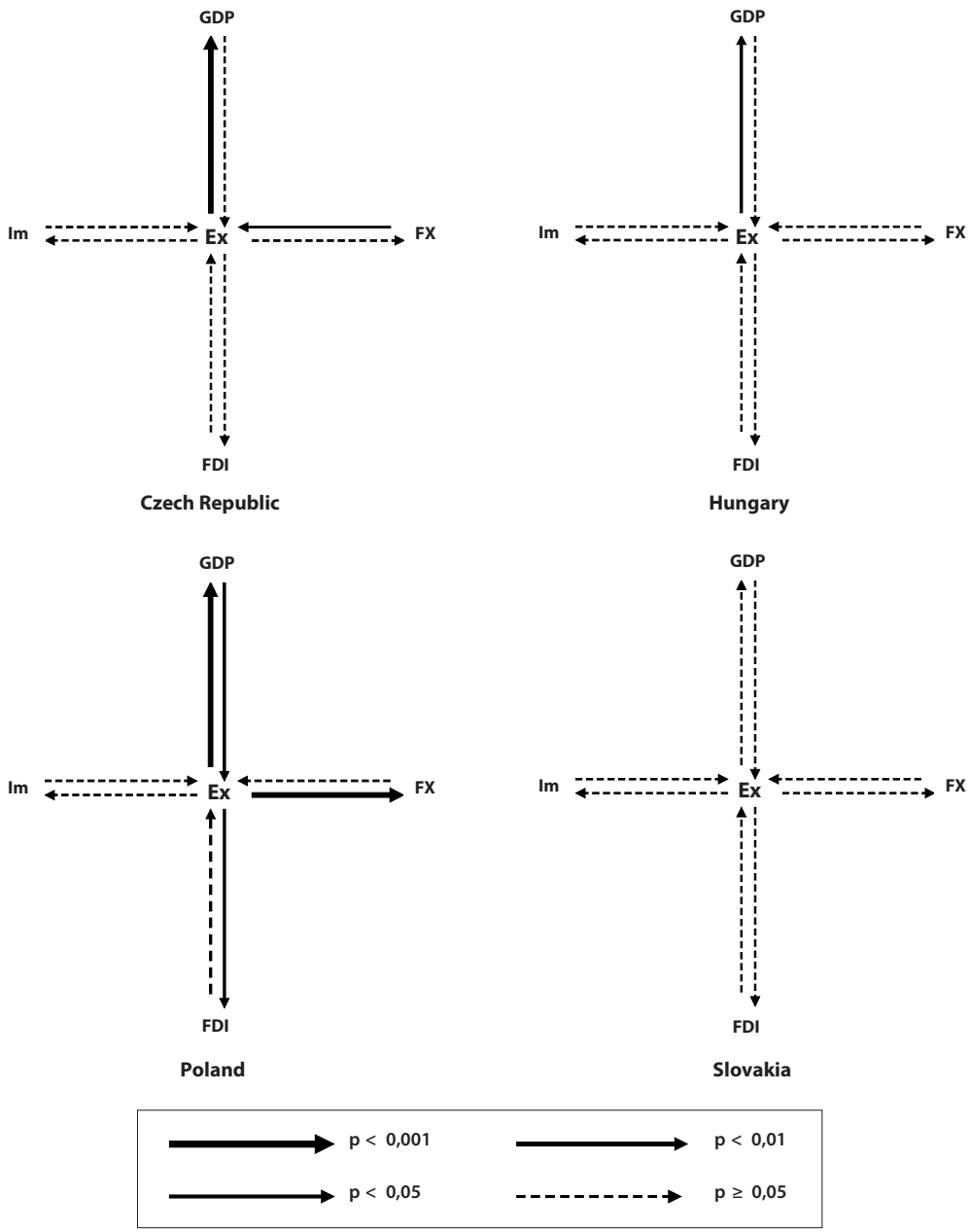
$$\Delta \mathrm{Im} = \underset{(0.769)}{-285.44} + \underset{(0.020)}{1.90}\,\Delta Ex\_d - \underset{(0.038)}{1.76}\,\Delta \mathrm{Im}\_d - \underset{(0.003)}{3663.09}\,S_1 - \underset{(0.308)}{1779.23}\,S_2 - \underset{(0.369)}{1032.25}\,S_3 + \underset{(0.042)}{244.79}\,t - \underset{(0.027)}{1.61}\,EC1 \tag{8}$$

It should be noted that Ex and Im in the previous period, the seasonal component S_1, the time variable t and the indicator EC1 statistically significantly affect the current difference of Ex and Im. Based on the model, we can conclude that an increase of imports in previous period caused a decline of exports and imports in the current period. Similarly, an increase of imports in the previous period causes an increase of exports and imports in the current period. The coefficient of error correction component EC1 is negative in the equations (7), (8), which ensures that a balance through a short-term adjustment process will be achieved. Similarly, VECM and VAR models (as shown in Table 2) were built for other variables and for all countries of the Visegrád Group. On the basis of these models anexogenous or endogenous nature of the variables was identified.

2.2 Results of weak and strong exogeneity test

The confirmation of weak exogeneity of a variable in the macromodel may suggest that it is an effective instrument to influence the foreign trade policy in a given country. The lack of weak exogeneity of a variable means that it is an endogenous part of the equation and it should be modelled in a separate equation. Such a variable is not an effective tool of foreign trade policy and inference based on this variable may be erroneous (Kireyev, 2001). The variable is said to be strongly exogenous if historical changes in the foreign trade structure do not affect the present effectiveness of this variable in the development of international trade.

Figure 3 Granger causal relations for variables used in the modelling of international trade



Source: Author's study

Tables 3 and 4 present the results of testing weak exogeneity of variables, in VAR and VEC models, respectively, in accordance with the procedure described in point 2. Symbol X→Y visible in these tables means the exogeneity of variable X with respect to variable Y whereas in each case export is one of

these variables. Based on Table 3, it can be stated that almost all variables in VAR models were weakly exogenous. Export in Poland was an exception as it was not weakly exogenous in relation to import. In VEC models (Table 4) in turn, none of the variables was weakly exogenous, with the exception of the Czech export which was weakly exogenous in relation to GDP.

The Granger causality test was necessary for testing strong exogeneity. The results of the test are illustrated in Figure 3. This figure demonstrates the power of directional dependencies between the analysed variables in individual Visegrád Group countries.

Based on Figure 3, it can be said that the majority of Granger causal relations, including one feedback and two unidirectional causalities, occur in the Polish foreign trade. The strongest causalities are in the direction from export to GDP and from export to the exchange rate. In case of Slovakia, no significant Granger causality was confirmed. In the Czech Republic, only one significant causal relation from export to GDP was diagnosed. In Hungary, export was a cause for GDP and exchange rate was a significant cause for export. Considering the results of the weak exogeneity test and the Granger causality test, conclusions can be formulated with regard to strong exogeneity of variables.

The results of strong exogeneity testing are presented in Table 5. Strong mutual exogeneity in this model suggests that the relevant time series create a system in which equations may be used directly to forecast macroeconomic values (this information can be found in the last column in Table 5). Table 5 shows that the majority of strongly exogenous variables may be seen in dynamic macroeconomic relations occurring in the Slovak foreign trade. In the case of Slovakia, time series of variables in three out of four models create the system. In the Hungarian trade, in turn, one system of equations created by mutually strong exogenous variables was found. In Poland and in the Czech Republic, there are only cases of strongly exogenous unidirectional relations.

Among comparable countries, only Poland is characterised by strongly exogeneous import with respect to export, which means that import may be an effective tool for developing export and historical structure of export does not affect the current import. In other Visegrád Group countries, import is not a weakly exogenous variable. This means that it is an endogenous part of the export model and cannot be treated as a determinant of export. Export is not a weakly exogenous variable with respect to import in any of the Visegrád Group countries. Therefore, this variable may not be used directly in the modelling of import.

2.3 Implications of exogeneity testing results for foreign trade of Visegrád Group countries

GDP is a weakly exogenous variable in relation to export in Slovakia only. Thus, only in Slovakia the value of the produced goods and services is a significant determinant of export whereas historical values of export do not affect the current economic development of Slovakia measured with GDP. On the other hand, Slovakian export is a strongly exogenous variable with respect to GDP. This means that the trade structure may be an effective tool to support economic growth in that country and historical values of GDP have no influence on the current trade structure. Slovakia turned out to be the only country in which GDP and export form the system – it is possible to forecast the values of both variables directly from equations estimated in the VAR model. In each of the other Visegrád Group countries, export was an endogenous component of the model describing GDP and it should be estimated in separate equations. Foreign direct investments can be considered as an effective tool to shape export in Poland, Slovakia and Hungary whereas historical changes of export structure do not affect the current value of FDI in case of Slovakia and Hungary where strong exogeneity of FDI in relation to export was identified. Since export was a strongly exogenous variable with respect to FDI in these countries, the relevant time series of these variables form systems. This opens an opportunity for dynamic inference concerning FDI and export directly from the relevant equations of the VAR model built for Slovakia and Hungary. The export structure in Poland and in the Czech Republic has a considerable impact on FDI but, additionally, in the Czech Republic the earlier structure of FDI does not affect the current export value. Variables repre-

sented by time series of foreign exchange rate and export exhibit bidirectional strong exogeneity in case of Slovakia only. This means that the exchange rate is a significant determinant of export and vice versa. But it is also justified to forecast these two values directly from VAR model equations. In the Czech Republic and Hungary, the variables in question have not shown any exogeneity in relation to one another, so they should be modelled in separate equations.

Also exchange rate and export do not form a system in Poland whereas the exchange rate is weakly exogenous with respect to export and export is strongly exogenous in relation to the exchange rate. Therefore, dynamic inference is justified here only on the basis of the VAR model equation which describes the exchange rate.

CONCLUSION

This article presents the contemporary concept of the exogeneity of variables in foreign trade models, based on Visegrád Group countries. The obtained results for strong and weak exogeneity have made it possible to formulate methodological conclusions as well as conclusions concerning the effectiveness of various macroeconomic instruments for the development of foreign trade in individual countries. The presented results allow for analysing macroeconomic variables in the context of the homogeneous estimation of the parameters of VAR and VEC models as well as the possible prediction of the values of variables. On the other hand, conclusions enable the recommendation of foreign trade policy tools, for example, in the context of aiming at balance in foreign trade or deficit reduction in trade.

The presented results lead to the conclusion that in case of Slovakia foreign trade is subject to modelling to a greater extent than in other countries, and this modelling takes into account a dynamic structure of time series of the analysed variables. The presented dependencies also suggest that in case there is no cointegration of time series, strongly exogenous variables occur more frequently than in case when such cointegration exists. In the dependencies described with VEC model, it is much more difficult than in case of VAR models to reject the hypothesis that variables are not strongly exogenous. Obviously, the above conclusion is not universal but it refers to the modelling of foreign trade in specific countries. Potential generalisation of conclusions concerning the exogeneity of variables used in foreign trade models requires further research for the respectively higher number of countries. It is worth noting that this article offers only analyses of weak and strong exogeneity of variables. It seems that the continuation of research on the subject should also lead to testing other kinds of exogeneity, e.g. the superexogeneity of variables. This will enable the verification of additional features of variables, for example in the aspect of the stability of foreign trade model parameters.

References

CHAREMZA, W. W., DEADMAN, D. F. *Nowa ekonometria*. Warszawa: PWE, 1997.

ENGLE, R. F., HENDRY, D. F., RICHARD, J-F., Exogeneity. *Econometrica*, 1983, 51, pp. 277–304.

CIEŚLIK, A. *Nowa teoria handlu zagranicznego w świetle badań empirycznych*. Warszawa: Wydawnictwo Naukowe PWN, 2000.

HSIAO, F. S., HSIAO, M. C. W. FDI, exports, and GDP in East and Southeast Asia – Panel data versus time-series causality analyses. *Journal of Asian Economics*, 2006, 17, pp. 1082–1106.

JOHANSEN, S. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica*, 1991, 59, pp. 1551–1581.

KIREYEV, A. Econometric Analysis of Discrete Reforms. *IMF Working Paper*, 01/156, Geneva: Internetional Monetary Fund, 2001.

KOJIMA, K. International Trade and Foreign Investment: Substitutes or Complements. *Hitotsubashi Journal of Economics*, 1975, 16, pp. 1–12.

KUSIDEŁ, E. Modele wektorowo-autoregresyjne VAR. Metodologia i zastosowania. [w:] *Dane panelowe i modelowanie wielowymiarowe w badaniach ekonomicznych*, pod red. B. Sucheckiego, Łódź: ABSOLWENT, 2000.

LIU, X., WANG, C., WEI, Y. Causal links between foreign direct investment and trade in China. *China Economic Review*, 2001, 12, pp. 190–202.

MADDALA, G. S. *Ekonometria*. Warszawa: Wydawnictwo Naukowe PWN, 2006.
MEHRARA, M., FIROUZJAEE, B. A. Granger Causality Relationship between Export Growth and GDP Growth in Developing Countries: Panel Cointegration Approach. *International Journal of Humanities and Social Science*, 2011, 1, pp. 223–231.
OSIŃSKA, M. (ed.), KOŚKO, M., STEMPIŃSKA, J. *Ekonometria współczesna*. Toruń: „Dom Organizatora", 2007.
OZAWA, T. Foreign Direct Investment and Economic Development. *Transnational Corporation*, 1992, 1, pp. 27–54.
SIMIONESCU, M. The Relationship between Trade and Foreign Direct Investment in G7 Countries a Panel Data Approach. *Journal of Economics and Development Studies*, 2014, 2, pp. 447–454.
SHARMA, R., KAUR, M. Causal Links between Foreign Direct Investments and Trade: A Comparative Study of India and China. *Eurasian Journal of Business and Economics*, 2013, 6, pp. 75–91.
STRAUSS, H. Multivariate Cointegration Analysisof Aggregate Exports: Empirical Evidence for the United States, Canada, and Germany. *Kiel Working Paper*, 1101, Kiel Institute for World Economics, 2002.

ANNEX

Table 1 The results of the ADF test for variables used in foreign trade models

Variable	Czech Republic		Hungary		Poland		Slovakia	
	primary variable	first difference	primary variable	first difference	primary variable	first difference	primary variable	first difference
Ex	1.146 (0.9328)	–9.060 (0.0000)	0.8532 (0.8912)	–6.572 (0.000)	–1.565 (0.1095)	–9.060 (0.0000)	–0.1544 (0.6251)	–1.0721 (0.0000)
Im	0.9620 (0.9084)	–6.807 (0.0000)	0.6795 (0.8591)	–6.705 (0.000)	–1.729 (0.0793)	–9.113 (0.0000)	–0.0153 (0.4584)	–1.0912 (0.0000)
GDP	0.6253 (0.8479)	–7.190 (0.0000)	–4.134 (0.0001)	–8.494 (0.000)	0.7279 (0.8687)	–7.560 (0.0000)	0.0166 (0.9872)	–0.5054 (0.0001)
FDI	-0.8265 (0.3526)	–5.413 (0.0000)	0.2716 (0.7605)	–10.63 (0.000)	–2.784 (0.0064)	–10.180 (0.0000)	–0.1244 (0.0344)	–1.0365 (0.0000)
FX	0.4672 (0.8119)	–6.819 (0.0000)	–0.9176 (0.3138)	–7.449 (0.000)	0.1942 (0.7382)	–4.508 (0.0000)	0.0077 (0.8620)	–0.9317 (0.0000)

Source: Own calculations based on data from CEIC database

Table 2 The results for Akaike and Schwarz criteria and the Johansen test for variables used in foreign trade models

Variables	Czech Republic			Hungary		
	Optimal order of lag (AIC, BIC)	Number of cointegrating vectors (Johansen test)	Model	Optimal order of lag (AIC, BIC)	Number of cointegrating vectors (Johansen test)	Model
Ex, Im	2	1	VEC	3	1	VEC
Ex, GDP	4	1	VEC	3	1	VEC
Ex, FDI	4	1	VEC	1	0	VAR
Ex , FX	1	1	VEC	1	1	VEC
Variables	**Poland**			**Slovakia**		
	Optimal order of lag (AIC, BIC)	Number of cointegrating vectors (Johansen test)	Model	Optimal order of lag (AIC, BIC)	Number of cointegrating vectors (Johansen test)	Model
Ex, Im	5	0	VAR	1	1	VEC
Ex, GDP	5	1	VEC	1	0	VAR
Ex, FDI	5	0	VAR	1	0	VAR
Ex , FX	5	0	VAR	5	0	VAR

Source: Own calculations based on data from CEIC database

Table 3 The results of testing weak exogeneity of variables in VAR models

Country	Variables	The result of test *t* for a residual variable	*p*-value	Weak exogeneity
Hungary	FDI→Ex	–1.126	0.2665	YES
	Ex →FDI	0.5044	0.6166	YES
Poland	Im→Ex	–1.843	0.0741	YES
	Ex→Im	5.721	0.000	NO
	FDI→Ex	–1.390	0.1735	YES
	Ex→FDI	0.6776	0.5026	YES
	FX→Ex	0.7447	0.4616	YES
	Ex→FX	1.577	0.1240	YES
Slovakia	GDP→Ex	0.0799	0.9367	YES
	Ex→GDP	1.183	0.2433	YES
	FDI→Ex	0.2183	0.8282	YES
	Ex→FDI	0.3910	0.6978	YES
	FX→Ex	–0.2958	0.7692	YES
	Ex→FX	–0.5054	0.6166	YES

Source: Own calculations based on data from CEIC database

Table 4 The results of testing weak exogeneity of variables in VEC models

Country	Variables	The test result for lagged variables of error correction term		The result of test t for a residual variable[4]		Weak exogeneity
		F	*p*-value	*t*	*p*-value	
Czech Republic	Im→Ex	0.522	0.597	19.950	0.000	NO
	Ex→Im	3.346	0.045	----	----	NO
	GDP→Ex	3.812	0.011	----	----	NO
	Ex→GDP	1.046	0.398	12.190	0.000	YES
	FDI→Ex	9.792	0.000	----	----	NO
	Ex→FDI	0.327	0.858	1.500	0.143	NO
	FX→Ex	0.921	0.343	9.978	0.000	NO
	Ex→FX	0.533	0.469	9.978	0.000	NO
Hungary	Im→Ex	0.959	0.422	32.060	0.000	NO
	Ex→Im	0.549	0.652	32.000	0.000	NO
	GDP→Ex	19.784	0.000	----	----	NO
	Ex→GDP	0.967	0.419	2.338	0.025	NO
	FX→Ex	3.385	0.073	–2.134	0.039	NO
	Ex→FX	2.000	0.164	–8.688	0.000	NO
Poland	GDP→Ex	1.412	0.247	3.085	0.004	NO
	Ex →GDP	232.479	0.000	----	----	NO
Slovakia	Im→Ex	0.100	0.753	36.710	0.000	NO
	Ex→Im	0.135	0.715	36.520	0.000	NO

Source: Own calculations based on data from CEIC database

[4] A significant result of test *F* for delayed variables of error correction component in VEC models automatically implied the lack of weak exogeneity and the resulting lack of strong exogeneity of a variable. In such situations, *t* test was abandoned.

Table 5 Results of strong exogeneity testing of variables

Country	Variables	Strong exogeneity	System
Czech Republic	Ex→Im	NO	NO
	Im→Ex	NO	
	Ex→GDP	NO	NO
	GDP→Ex	NO	
	Ex→FDI	YES	NO
	FDI→Ex	NO	
	Ex→FX	NO	NO
	FX→Ex	NO	
Hungary	Ex→Im	NO	NO
	Im→Ex	NO	
	Ex→GDP	NO	NO
	GDP→Ex	NO	
	Ex→FDI	YES	YES
	FDI→Ex	YES	
	Ex→FX	NO	NO
	FX→Ex	NO	
Poland	Ex→Im	NO	NO
	Im→Ex	YES	
	Ex→GDP	NO	NO
	GDP→Ex	NO	
	EX→FDI	NO	NO
	FDI→Ex	NO	
	Ex→FX	YES	NO
	FX→Ex	NO	
Slovakia	Ex→Im	NO	NO
	Im→Ex	NO	
	Ex→GDP	YES	YES
	GDP→Ex	YES	
	Ex→FDI	YES	YES
	FDI→Ex	YES	
	Ex→FX	YES	YES
	FX→Ex	YES	

Source: Own calculations based on data from CEIC database

Some Practical Issues Related to the Integration of Data from Sample Surveys

Wojciech Roszka[1] | *Poznań University of Economics, Poznań, Poland*

Abstract

The users of official statistics data expect multivariate estimates at a low level of aggregation. However, due to financial restrictions it is impossible to carry out studies on a sample large enough to meet the demand for low-level aggregation of results. At the same time, respondents' burden prevents creation of long questionnaires covering many aspects of socio-economic life.
Statistical methods for data integration seem to provide a solution to such problems. These methods involve fusion of distinct data sources to be available in one set, which enables joint observation of variables from both files and estimations based on the sample size being the sum of sizes of integrated sources.

Keywords

Data fusion, data integration, multiple imputation, quality assessment, statistical matching, sample survey

JEL code

C02, C18, C31, C63, C83

INTRODUCTION

Official statistics institutions conduct many sample surveys in order to respond to the demand for information reported by a number of different public and private institutions. The substantive content of the surveys derives not only from the needs of the recipients, but also from international commitments enabling comparative analysis of different socio-economic phenomena in the European Union. At the same time due to the very high costs a sample size in the studies does not allow for generalization of the results in the detailed cross-sections, while the respondent burden which results in refusals and missing data enforces design of relatively short questionnaires. Hence, a statistical inference for small domains[2] is impossible (due to large sampling error) and none of the studies cover all the aspects of the socio-economic phenomena. For these reasons the current process of modernization of the statistical infrastructure includes increasing the efficiency of reporting systems through the integration of statistical information from available data sources (Leulescu, Agafitei, 2013).

Statistical data integration methods can provide a response to the problems of disjoint observation of variables in the sample surveys, and also allow for the estimation of better quality for small domains. For several years they are considered the subject of public statistics, and Eurostat in particular. The projects

[1] Poznań University of Economics, al. Niepodległości 10, Poznań, Poland. E-mail: wojciech.roszka@ue.poznan.pl.
[2] I.e. demographic cross sections within a small geographical area (i.e. NUTS 4).

like *CENEX-ISAD* (CENEX 2006) or *Data Integration* (ESSnet on Data Integration 2011) improved and disseminated the methodology of the statistical data integration.

Statistical matching (*data fusion*) is a methodological approach that provides a joint observation of variables not jointly observed in two (or more) datasets. Potential benefits of this approach lies in the possibility of increasing the analytical capacity of existing data sources without increasing the cost of research and the burden on respondents.

The scope of this paper is to identify some practical issues related to the integration of data from sample surveys with statistical matching method like in Raessler (2002) and D'Orazio et al. (2006). In first section the statistical matching framework will be described with particular emphasis on combining the microdata sets. The next section will deal with the methodology of merging two sample survey data files with some practical remarks. Especially approaches to harmonizing and concatenation of datasets will be shown as well as methods of the missing data imputation. In the third section integration of data from two sample surveys – Household Budget Survey (HBS) and European Union Statistics on Income and Living Conditions (EU-SILC) – will be presented with particular emphasis on the quality and efficiency of the algorithms used. At the end the general conclusions will be presented.

1 STATISTICAL MATCHING OVERVIEW

Statistical matching is a group of statistical methods for the integration of two (or more) data sources (usually originating from sample surveys) referring to the same target population. The aim of the integration is the joint observation of variables not jointly observed in any of the sources and the possibility to make inference on their joint distribution.

1.1 Matching scheme

In each of the input datasets (labeled *A* and *B*) a vector of variables of the same (or similar) definitions and categories is available. These are so called common variables (labeled as $\boldsymbol{x}$). Dataset *A* contains also a vector of variables which is observed only in this dataset (labeled $\boldsymbol{y}$) and, analogically, dataset *B* contains also a vector of variables which is observed only in this dataset (labeled $\boldsymbol{z}$). Variables $\boldsymbol{y}$ and $\boldsymbol{z}$ are called distinct (or target) variables. Since the probability of selection the same unit to two (or more) samples simultaneously is close to zero, it is assumed that the input datasets are disjoint.

($\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}$) are random variables with the density function f($\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}$). It is assumed that $\boldsymbol{x}=(X_1,\ldots,X_P)'$, $\boldsymbol{y}=(Y_1,\ldots,Y_Q)$, $\boldsymbol{z}=(Z_1,\ldots,Z_R)$ are random variables vectors of a size *P*, *Q* and *R* respectively. It is also assumed that *A* and *B* are two independent samples consisting of n_A and n_B independently drawn units (Di Zio, 2007).

Vector $(x_a^A, y_a^A) = (x_{a1}^A,\ldots, x_{aP}^A; y_a^A,\ldots, y_{aQ}^A)$, $a=1,\ldots,n_a$, consists of the observed values of variables for units in dataset *A*. Analogically, vector $(x_b^B, z_b^B) = (x_{b1}^B,\ldots, x_{bP}^B; z_b^B,\ldots, y_{bR}^B)$, where $b=1,\ldots,n_b$, consists of the observed values of variables for units in dataset *B* (see Scheme 1).

Both *A* and *B* datasets should contain information about the same target population. Hence, the type of statistical/observation unit should also be the same (i.e. person, household etc.). The reference periods also ought to be similar. Should any of mentioned conditions failed, harmonization needs to be performed. If it is impossible to harmonize datasets (different populations, inconsistent unit types etc.), integration is impossible to conduct.

The statistical matching algorithm is initialized with the choice of target variables. These are variables selected from vector of distinct variables $\boldsymbol{y}$ (and $\boldsymbol{z}$) which are going to be merged with data in set *B* (*A*). The dataset to which, in particular integration step, variables are being matched is called *recipient*, while the dataset which variables are being matched from is called *donor*. The choice of the target variables is usually dictated by the information needs, and depending on the nature of the variables used, set of rules and methods of integration is used.

Scheme 1 Initial data in statistical matching

Dataset A

Y_1	...	Y_Q	X_1	.	X_P
y_{11}^A	...	y_{1Q}^A	x_{11}^A	...	x_{1P}^A
...	...	...	...	...	...
y_{a1}^A	...	y_{aQ}^A	x_{a1}^A	...	x_{aP}^A
...	...	...	...	...	...
$y_{n_A1}^A$	...	$y_{n_AQ}^A$	$x_{n_A1}^A$	...	$x_{n_AP}^A$

Dataset B

X_1	.	X_P	Z_1	.	Z_R
x_{11}^B	...	x_{1P}^B	z_{11}^B	...	z_{1R}^B
...	...	...	...	...	...
x_{b1}^B	...	x_{bP}^B	z_{b1}^B	...	z_{bR}^B
...	...	...	...	...	...
$x_{n_B1}^B$	...	$x_{n_BP}^B$	$z_{n_B1}^B$	...	$z_{n_BR}^B$

Source: Di Zio (2007)

In the next step a vector of common variables $\boldsymbol{x}$ is identified. From that vector, according to particular target variables, a set of matching variables is being chosen $\boldsymbol{x}_M \subset \boldsymbol{x}$. Usually variables that explain the most of the variance of the target variable are being chosen. The relationship between common and target variables is usually not one-dimensional. Hence, the matching variables are usually being chosen using multidimensional methods like stepwise regression, cluster analysis or classification and regression trees (CART).

1.2 Conditional Independence Assumption

Since variables $\boldsymbol{y}$ and $\boldsymbol{z}$ are not jointly observed in any sources, in the estimation of the relationship between these characteristics it is usually assumed that $\boldsymbol{y}$ and $\boldsymbol{z}$ are conditionally independent given $\boldsymbol{x}$ (Raessler, 2004, D'Orazio et al., 2006, Moriarity, 2009). It is called *conditional independence assumption* (CIA) and under CIA the density function of ($\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$) has the following property:

$$f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = f_{Y|X}(\boldsymbol{y}|\boldsymbol{x}) f_{Z|X}(\boldsymbol{z}|\boldsymbol{x}) f_X(\boldsymbol{x}), \forall\, \boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y}, \boldsymbol{z} \in \mathcal{Z}, \tag{1}$$

where $\boldsymbol{f}_{(Y|X)}$ is a conditional density function of $\boldsymbol{y}$ given $\boldsymbol{x}$, $\boldsymbol{f}_{Z|X}$ is a conditional density function of $\boldsymbol{z}$ given $\boldsymbol{x}$, and $\boldsymbol{f}_X$ is a marginal density of $\boldsymbol{x}$. When the assumption of conditional independence is true, information about marginal distribution of $\boldsymbol{x}$ and about relationships between $\boldsymbol{x}$ and $\boldsymbol{y}$ as well as $\boldsymbol{x}$ and $\boldsymbol{z}$ is sufficient to estimate (1). That information can be derived from *A* and *B* datasets.

It is worth underlining that the veracity of the CIA cannot be tested using information from $A \cup B$ solely. False assumption may lead to biased estimates. In order to obtain a point density estimate $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$ it is necessary to refer to external sources of information. Singh et al. (1993) determined two types of such sources:

- third database *C* in which ($\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$) or ($\boldsymbol{y}$, $\boldsymbol{z}$) are jointly observed,
- reliable values of unknown relations between ($\boldsymbol{y}$, $\boldsymbol{z}|\boldsymbol{x}$) or ($\boldsymbol{y}$, $\boldsymbol{z}$).

In practice, many problems with the dataset *C* may occur. It may be inconsistent with *A* and *B* in terms of population, definitions or reference period. Conducting a new study in order to obtain joint observation of (***x, y, z***) or (***y, z***) raises problems of economical (cost and time to carry out research) and statistical (new dataset can be characterized by missing data and random and/or non-random errors) nature.

When additional data sources are unavailable, *uncertainty analysis* is being performed (D'Orazio et al. 2006) which is a kind of interval estimation for unknown characteristics, such as correlation matrix of (***y,z***). The narrower the estimated intervals are, the better quality of the integrated data sets characteristics are. The product of application of statistical matching methods using interval estimation are, for microdata sets, family of synthetic datasets created by using a variety of reliable parameters used in the integration model.[3]

In conclusion, data fusion can be performed using (i) conditional independence assumption, (ii) auxiliary (additional) data sources, (iii) uncertainty analysis.

In the works, among others, of Kadane (1978), Paas (1986) and Singh et al. (1993) it is showed that the integration with the conditional independence assumption usually leads to estimates of sufficient quality. The conditional independence assumption is most commonly used because of the ease of application and, as practice shows, a good quality of integration.

2 MATCHING ALGORITHM

2.1 Datasets harmonization

Harmonization is laborious but a necessary initial step in the integration. It allows, among others, comparison of distributions of variables from various sources and subsequent evaluation of the results of the integration. According to van der Laan (2000) 8 steps of harmonization can be distinguished (see also Scanu, 2008):

1. units definition harmonization;
2. reference periods harmonization;
3. population completeness analysis;
4. variables harmonization;
5. variables categories harmonization;
6. measurement error correction;
7. handling missing data;
8. creation of derivative variables.

Without loss of generality, the above mentioned steps can be grouped into 2 categories: (i) compatibility of the population and units (1–3), (ii) harmonization of variables (4–8).

The integration of the two data sources is justified when: (i) reference periods of the surveys are consistent, (ii) populations in the surveys are the same or different but overlapping.

In the case of non-consistent reference periods, they should be corrected (i.e. by performing demographic projections).

If the populations are different but overlapping, in the integrated datasets (labeled as *A* i *B*) subsets *A*1 and *B*1 must be extracted, in such a way that they contain a common part of the population. It has to be verified whether the obtained subsamples are representative for the surveyed population (Scanu, 2008). If the verification is successful, subsets *A*1 and *B*1 can be integrated.

When the two datasets refer to two different (disjoint) populations, none of the methods will be proper.

---

[3] Another approach is solely an estimate of specific relations (e.g. correlation, regression coefficients, contingency table) between vectors of variables ***Y*** and ***Z***, without creating a synthetic microdata set – the so-called *macro approach* (D'Orazio et al., 2006).

Common variables should be fully consistent. It means that both definitions and distributions ought to be at least very similar. In the datasets from different sources meeting both of these conditions in full may be difficult. The most common problems which can be encountered here are the following:

- different definitions of variables and occurrence of different categories,
- missing data,
- distribution of the same variables among populations.

In the case of non-consistent definitions and/or categories of common variables, there are three types of variables:

1. The variables for which there is no possibility for harmonization

Such variables should not be regarded as 'common' and therefore they should not be considered as matching variables at all. This situation happens quite often, especially when the datasets come from different institutions.

2. The variables that can be harmonized by modification of their categories

Qualitative characteristics often contain many variants. Their harmonization is usually done by aggregation in such a way that derivative variants are created. These are consistent in both datasets (i.e. education 'primary' and 'no education' can be aggregated to 'primary or no education'). Aggregation of categories can lead to loss of information, though.

3. The derivative variables

In the absence of appropriate common variables or their insufficient number, new variables can be created by transforming other available variables. If the derived variables meet certain criteria (qualitative and definitional), they can be used as matching variables.

The common variables should also show appropriate quality. It means, among others, that they shouldn't contain missing data. Unit non-response results in the removal of the unit from the dataset. In the case of item non-response, two ways can be distinguished: (i) using variables without missing data only, (ii) impute missing data.

The third issue concerns the compatibility of distributions of variables. This is due to the assumption that input datasets refer to the same population. In situations where the distributions of the common variables are very different, it might be suspected that populations are non-consistent. More frequent situation is that the differences in the distributions of common variables arise from the variation of the sample.

Differences in the distributions can be examined by commonly used statistical tests (i.e. chi-squared test for goodness of fit, Kolmogorov–Smirnov test, etc.). However, for large samples formal statistical tests tend to reject the hypothesis of equality of distributions or fraction even at very small differences. Also, most of the 'classical' statistical tests were constructed for a simple randomized sampling scheme, while the input datasets often come from studies of complex sampling scheme.

Scanu (2008) suggested so called 'empirical approach'. Its essence is to compare the distributions of appropriate variables using visual methods and the use of some simple measures:

- for continuous variables – comparison of histograms;
- for qualitative variables – comparing fraction differences of the particular categories:
  - for 'big' fractions – differences lower than 5% are acceptable,
  - for 'small' fractions – differences lower than 2% are acceptable;
- for both scale and qualitative variables – *total variation distance*:

$$\Delta(w_A, w_B) = \frac{1}{2}\sum_{i=1}^{k}\left|w_{A,i} - w_{B,i}\right|, \tag{2}$$

where $w_{A,i}$ and $w_{B,i}$ are i-th (i=1,…,k) relative frequencies of a particular variable in the integrated datasets. In practice, it is accepted that distributions are "acceptably" compatible when $\Delta \leq 6\%$;

– for scale variables it is possible to compare estimates of population parameters, i.e. $\frac{\hat{\mu}_A}{\hat{\mu}_B}, \frac{\hat{\sigma}_A}{\hat{\sigma}_B}, \frac{\hat{\rho}_A}{\hat{\rho}_B}$.

2.2 Matching methods

Taxonomy of integration methods is described in detail in D'Orazio et al. (2006). For the purpose of integration of microdata sets, three frameworks are distinguished: (1) parametric, (2) non-parametric, (3) mixed. In the parametric framework, generally two techniques are used: (1) regression imputation, (2) stochastic regression.

The regression imputation in statistical matching is a fairly simple approach. Two models $Y(X)$ and $Z(X)$ are being estimated. Then predicted values are being imputed to B and A respectively. This process consists of three steps:

1. Predicted values resulting from a model:

$$\hat{z}_a^{(A)} = \hat{\alpha}_Z + \hat{\beta}_{ZX} x_a, a = 1,2, \dots, n_A, \tag{3}$$

are imputed to A.

2. Predicted values resulting from a model:

$$\hat{y}_b^{(B)} = \hat{\alpha}_Y + \hat{\beta}_{YX} x_b, b = 1,2, \dots, n_B. \tag{4}$$

are imputed to B.

3. Datasets A and B are concatenated: $S = A \cup B$; $n_S = n_A + n_B$.

The advantage of this approach is its simplicity. The disadvantage is the fact that it is a single imputation and the predicted values lie on the regression line.

Little and Rubin (2002) suggested to use a stochastic imputation in the statistical matching. It consists on drawing residual values for regression models obtained in such a way that:

$$\tilde{z}_a^{(A)} = \hat{z}_a^{(A)} + e_a = \hat{\alpha}_Z + \hat{\beta}_{ZX} x_a, \tag{5}$$

where $e_a \sim N(0, \hat{\sigma}_{Z|X})$, and

$$B = \frac{1}{m-1} \sum_{t=1}^{m} (\hat{\theta}^{(t)} - \hat{\theta}_{MI})^2, \tag{6}$$

where $e_b \sim N(0, \hat{\sigma}_{Y|X})$.

Development of the stochastic imputation method is a multiple imputation, suggested by Raessler (2002) to be used in the statistical matching framework. For the purpose of the multiple imputation m models are created. Each of the models is created using the stochastic regression method. Drawing residuals reflects the sample variability and allows to perform point and interval estimation for the unknown values of missing data (which is also a pro-solution for the problem of uncertainty, as described in section 1.2).

The imputation estimator for each of t (t=1,2,…,m) models is $\hat{\theta}^{(t)} = \hat{\theta}(U_{obs}, U_{mis}^{(t)})$, where U_{obs} are observed values, and $U_{mis}^{(t)}$) are imputed missing data (Raessler 2002). The variance of the estimator is formulated as $\widehat{var}(\hat{\theta}^{(t)}) = \widehat{var}(\hat{\theta}(U_{obs}, U_{mis}^{(t)}))$. The point estimate of the multiple imputations is an arithmetic mean:

$$\hat{\theta}_{MI} = \frac{1}{m} \sum_{t=1}^{m} \hat{\theta}^{(t)}. \tag{7}$$

"Between-imputation" variance is estimated by formula:

$$B = \frac{1}{m-1} \sum_{t=1}^{m} (\hat{\theta}^{(t)} - \hat{\theta}_{MI})^2 \tag{8}$$

and "within-imputation" variance is estimated by:

$$W = \frac{1}{m}\sum_{t=1}^{m} \widehat{var}\left(\hat{\theta}^{(t)}\right). \tag{9}$$

Total variance is a sum of between- and within-variance modified by $\frac{m+1}{m}$, to reflect the uncertainty about the true values of imputed missing data:

$$T = W + \frac{m+1}{m}B. \tag{10}$$

Interval estimates are based on t-distribution:

$$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}\sqrt{T} < \theta < \hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}\sqrt{T} \tag{11}$$

with degrees of freedom:

$$v = (m-1)\left(1 + \frac{W}{(1+\frac{1}{m})B}\right)^2. \tag{12}$$

The main advantage of the parametric approach is the 'economy' of the model – a small number of predictors explains a large part of the variance of the target variables. Among the disadvantages a need of model specification can be mentioned. Poorly constructed imputation model can generate results with poor quality. In addition, the imputed values are artificial, i.e. resulting solely from the model, not having their counterparts in reality. This problem is usually solved by the use of a mixed approach.

The non-parametric framework in data fusion is related to *hot deck* imputation methods (Singh et al., 1993). In practice, two groups of methods are most commonly used: (1) random imputation, (2) nearest neighbor matching.

Random imputation includes random draws of values of $Z(Y)$ variables from dataset $B(A)$ to $A(B)$. To maintain maximum distribution compliance of target variables, datasets are divided into many homogeneous groups, on the basis of categories of chosen common variables $\boldsymbol{x_G} \subset \boldsymbol{x}$. Random matching proceeds within the designated groups.

Nearest neighbor method involves choosing for each record in the set $A(B)$ most similar record of the set $B(A)$. 'Similarity' is measured as the distance between the values of matching variables:

$$d_{ab} = (x_{M.a}, x_{M.b}). \tag{13}$$

Hot deck imputation methods are commonly used in practice. Their main advantage is that imputed values are 'life' – they are empirically observed. Also, the non-parametric methods do not need a model specification and are quite simple in use. Main disadvantages is computational burden (each record in one dataset is compared to each record in the other dataset[4]) as well as only single values are being imputed.

The mixed methods combine advantages of parametric and non-parametric methods and alleviate their disadvantages. Most commonly (D'Orazio et al., 2006) mixed methods are described as a two-step algorithm:

1. multiple imputation with draws based on conditional predictive distribution;
2. empirical values with the shortest distance from the imputed values are matched: dab($\tilde{z}_a, z_b$).

Such an approach ensures that the imputed values are real as well as the multiple imputation provides possibility of uncertainty analysis. Commonly used method is *predictive mean matching* (Landerman et. al., 1997).

[4] To avoid that, dataset is divided analogically like in random matching.

2.3 Integration of data from complex sample surveys

In sample surveys carried out by the official statistics most frequently complex (multi-stage) sampling schemes are used. Rubin (1986) proposed a solution taking into account the sampling schemes of integrated studies. The idea is to transform inclusion probabilities of particular units in such a way that integrated repository reflects the size of the population (N).

The inclusion probability of each -th unit in the integrated dataset is the sum of the inclusion probabilities in and surveys minus the probability of selecting the units for both surveys simultaneously:

$$\pi_{A\cup B,i} = \pi_{A,i} + \pi_{B,i} - \pi_{A\cap B,i}\,. \tag{14}$$

Since normally sample size is a very small percentage of the size of the entire population, and, in addition, the institutions carrying out the measurement, ensuring that respondents were not overly burdened with obligations arising from the study, tend not to take into account one unit in several studies over a given period, equation (1) can be simplified as:

$$\pi_{A\cup B,i} \cong \pi_{A,i} + \pi_{B,i}\,. \tag{15}$$

Resulting from the sampling scheme survey weight is the inverse of inclusion probability. In an integrated dataset it will have the form:

$$w_{i_{A\cup B}} = \frac{1}{\pi_{A\cup B,i}}\,. \tag{16}$$

In practice, however, generally the inclusion probability is not available in the final dataset, but it contains computed weights (e.g. calibrated due to missing data). For the synthetic data set corresponded to the size the target population, the transformation of weights by the following formula is made:

$$w'_{i_{A\cup B}} = \frac{w_{i_{A\cup B}}}{\sum_{i=1}^{s} w_{i_{A\cup B}}} N, \tag{17}$$

where:

$w'_{i_{A\cup B}}$ – harmonized analytical weight for i-th unit in the integrated data set,

$w_{i_{A\cup B}}$ – original analytical weight,

N – population size.

Before matching procedure is performed, datasets are concatenated ($S = A \cup B$; $n_S = n_A + n_B$) and an imputation model, which takes into account survey weights is specified.

2.4 Quality assessment

Quality assessment of joint distribution of variables never jointly observed is a non-trivial task. Barr and Turner (1990) as well as Rodgers (1984) suggested relatively simple measures of quality assessment of integrated dataset $S = A \cup B$ – a comparison of basic statistics (mean, standard deviation etc.) in donor and integrated datasets.

Raessler (2002) proposed a more complex way of the quality evaluation, called an 'integration validity'. It consist of a verification of four 'validity levels':

1. A reproduction of true, unknown values of $Z(Y)$ in the recipient file – in result a '*hit ratio*' coefficient can be calculated. When a true value is replicated, a is noted. The coefficient is a ratio between number of 'hits' and the number of imputed values.
2. A joint distribution preservation – a true unknown joint distribution of $(\boldsymbol{x,y,z})$ is preserved in an integrated dataset.
3. A covariance structure $\widetilde{cov}(\boldsymbol{x,y,z}) = cov(\boldsymbol{x,y,z})$ is reflected in the integrated dataset as well as marginal distribution $\tilde{f}_{XY} = f_{XY}$ and $\tilde{f}_{XZ} = f_{XZ}$ are copied.

All above mentioned 'levels' can be evaluated only by the simulation study. Empirical evaluation, in the situation of no joint observation of target variables, is impossible.

4. Marginal distribution of $Z(Y)$ as well as joint distribution of $\boldsymbol{x}$ and $\boldsymbol{z}$ ($\boldsymbol{x}$ and $\boldsymbol{y}$) of the donor file should be similar in the integrated dataset.

In practice, most commonly used is the one suggested by German Association of Media Analysis (Raessler, 2002):

1. comparing the empirical distribution of target variables included in the integrated file with the one in the recipient and the donor files,
2. comparing the joint distribution $f_{X,Z}$ ($f_{X,Y}$) observed in donor file with the joint distribution $\tilde{f}_{X,Z}$ ($\tilde{f}_{X,Y}$) observed in the integrated file.

3. EMPIRICAL STUDY

In this research the issues of empirical verification of selected statistical methods for data integration, evaluation of the quality of a combination of various sources, integrated data quality assessment and compliance and accuracy of the estimation are carried out.

Due to the availability of data, as well as the content, the empirical study was conducted using sets of the Household Budget Survey (2005) and the European Union Statistics on Income and Living Conditions (2006).[5]

Table 1 Basic characteristics of HBS and EU-SILC surveys

Characteristics	HBS	EU-SILC
Measurement period	Whole year 2005	2nd May – 19th June 2006
Population	Households in Poland	Households in Poland
Sampling method	Two-stage, stratified	Two-stage, stratified
Subject of study	– household budget – household equipment – the volume of consumption of products and services	– income situation – household equipment – poverty – various aspects of the living conditions of the population
Assumed population size	13 332 605	13 300 839
Sample size	34 767	14 914

Source: Own study

[5] A dataset of 2006 was used due to the fact that the reference period of households' income in the EU-SILC survey was set in the year preceding the survey. It was assumed that the other variables like household equipment, living conditions and socio-demographic characteristics are less volatile in time than financial categories. In this way, efforts were made to maintain compliance of common variables of EU-SILC with HBS.

For the purposes of empirical study it was decided to merge households expenditures (to EU-SILC) dataset and head of household incomes[6] (to HBS file). The extension of the substantive scope of the estimates contained among others estimation of the unknown correlation coefficient between household expenditure and heads of households income. Hence, the integration leads to the extension of the substantive scope of the estimates. Integration includes information on households (see Table 1).

3.1 Datasets harmonization

A very important aspect is to harmonize the datasets before the integration. In both repositories variables with the same or similar definitions existed. Categories, however, were often divergent and aggregation was required to harmonize variants to the same definition (see Table 2). Both studies were carried out by the same institution, similar aims were guided and measurement was subject to a very similar areas of socio-economic life. It seems, therefore, that the definitions of their variants should be consistent not only for data integration but primarily for comparative purposes. It seems that discrepancies occurring in both studies arise from specific international obligations and the need for comparisons with other analogous studies carried out in other European countries.

Table 2 Harmonization of categories of selected common variables

Variables	HBS categories	EU-SILC categories	Harmonized categories
Type of building	1 'multiple dwelling'	1 'detached house'	1 'multiple dwelling'
	2 'single family terraced house'	2 'terraced house'	2 'single-family house'
	3 'single family detached house'	3 'apartment or flat in a building with less than 10 dwellings'	3 'single family terraced house'
	4 'other'	4 'apartment or flat in a building with 10 or more dwellings'	4 'other'
Number of rooms	Scale variable with min=1 and max=12	1	1
		2	2
		3	3
		4	4
		5	5
		6 '6 and more'	6 '6 and more'

Source: Own study

After performing harmonization, the similarity of distribution of the harmonized common variables in the integrated datasets was done. For qualitative variables of total variation distance (TVD) coefficient was used (see section 2.1; see Table 3). Variables with the value greater than 6% were rejected. For the quantitative variables a ratio between basic distribution parameters was calculated (see Table 4). The closer the value of the coefficient is to one, the more similar the distributions are.

[6] As an example of a personal income of a household member which was not measured in HBS.

Table 3 Distribution compliance assessment for selected qualitative common variables (in %)

Variable	Category	Dataset		*TVD*
		HBS	EU-SILC	
Type of building	multiple dwelling	59.21	55.96	4.03
	single family house	35.11	39.14	
	single family terraced house	5.36	4.61	
	others	0.32	0.29	
Number of rooms	1	15.93	13.41	4.03
	2	36.56	35.34	
	3	29.43	29.14	
	4	9.86	11.07	
	5	4.98	6.19	
	6 and more	3.24	4.85	

Source: Own study

Table 4 Distribution compliance assessment for selected quantitative common variables

Variable	Statistics	Dataset		Ratio of sample parameters
		HBS	EU-SILC	
Disposable income	Mean	2 155.7	2 286.3	0.943
	Variance	3 451 099.9	3 266 838.2	1.056
	Standard deviation	1 857.7	1 807.4	1.028
Equivalised disposable income	Mean	1 222.8	1 288.6	0.949
	Variance	991 775.5	828 935.2	1.196
	Standard deviation	995.9	910.5	1.094

Source: Own study

3.2 Choice of matching variables

Among the common variables, for each of the target variables as the dependent variable, selection of variables was performed using the CART method. Following variables were chosen:

- for variable *household expenditures*: if the household has a private bathroom, if the household has a flushable toilet, if the household has a car, number of rooms, type of building, equivalised disposable income, disposable income, household size;
- for variable *income of head of household*: if the household has a flushable toilet, if the household has a washing machine, if the household has a car, if the household has a TV, number of rooms, type of building, legal title to occupied apartment, equivalised disposable income, disposable income, household size.

3.3 Integration results

One hundred imputations were performed based on a linear regression models. Residuals were drawn using Markov Chain Monte Carlo method (MCMC) (IBM SPSS Missing Values 20 2011). Rubin's approach was used to harmonize analytical weights. Both multiple imputation (MI) and mixed approach were used.

For the selected methods of integration consistent results were achieved (see Table 5 and Table 6). Designated confidence intervals had a low spread. Additionally, through the use of interval estimation

it was possible to estimate the uncertainty of the true unknown values in the integrated dataset (assessment veracity of the CIA in terms of the integrated sets was impossible).

Table 5 Assessment of estimators of the arithmetic mean of variables in an integrated data set

Variable	Statistic	MI	Mixed model
Household expenditures	B	8.14	32.25
	W	8.80	10.06
	T	17.03	42.64
	$\sqrt{T}$	4.13	6.53
	$t_{v,\frac{\alpha}{2}}$	2.2414093	2.2414031
	$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}\sqrt{T}$	1 950.86	2 005.29
	$\hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}\sqrt{T}$	1 969.36	2 034.56
Head of household income	B	35.20	5.23
	W	14.68	11.88
	T	50.23	17.17
	$\sqrt{T}$	7.09	4.14
	$t_{v,\frac{\alpha}{2}}$	2.2414029	2.2414114
	$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}\sqrt{T}$	2 006.58	2 004.91
	$\hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}\sqrt{T}$	2 038.35	2 023.48

Source: Own study

Table 6 Assessment of estimators of the correlation coefficient of not jointly observed variables in an integrated dataset

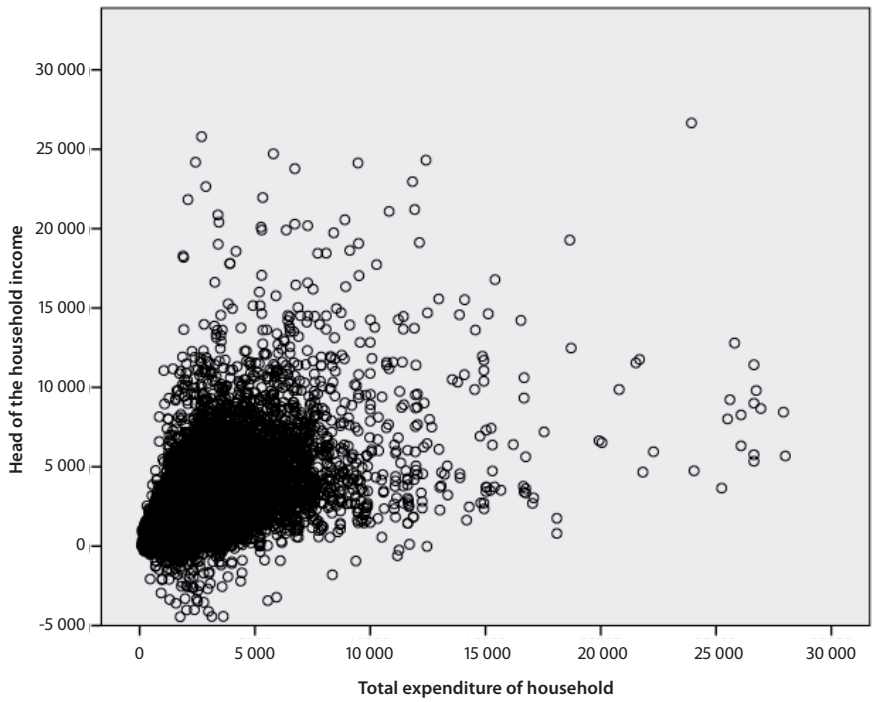
Variable	Statistic	MI	Mixed model
Correlation coefficient $z(\hat{\rho}^{(t)})$	B	0.00006	0.00013
	W	2E–15	2E–15
	T	0.00006	0.00013
	$\sqrt{T}$	0.01	0.01
	$t_{v,\frac{\alpha}{2}}$	2.2760035	2.2760035
	$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}\sqrt{T}$	0.5611	0.5534
	$\hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}\sqrt{T}$	0.5849	0.5884

Note: $z(\hat{\rho}^{(t)})$ – z-transformed ρ estimate: $z\left(\hat{\rho}^{(t)}\right) = \frac{1}{2} ln \frac{1+\rho_{YZ}^{(t)}}{1-\rho_{YZ}^{(t)}}$; $z\left(\hat{\rho}^{(t)}\right)$ has a normal distribution with the constant variance $\frac{1}{n-3}$. The confidence intervals are given for $\hat{\rho}$ (marked grey).

Source: Own study

Also, a joint observation of variables not jointly observed in any of the input databases was achieved (see Figure 1).

Figure 1 Diagram of correlation between variables not jointly observed in input sets



Note: For a convenience of illustration range of variables in the graph was reduced.
Source: Own study

Table 7 Estimation of average values of matched variables [in PLN] by region with an assessment of the precision of estimate

Variable	Region (NUTS1)	HBS			EU-SILC			integrated			$1-\frac{s_{int}(\bar{x})}{s_{emp}(\bar{x})}$
		$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV	
Household expenditures	central	1 916	16.40	0.86	2 130	29.48	1.38	2 024	14.95	0.74	0.088
	south	2 008	16.44	0.82	2 072	27.59	1.33	2 040	14.45	0.71	0.121
	east	1 970	19.27	0.98	2 040	29.12	1.43	2 005	16.19	0.81	0.160
	north-west	1 973	18.20	0.92	2 147	37.72	1.76	2 061	18.11	0.88	0.005
	south-west	2 138	27.10	1.27	2 121	44.00	2.07	2 130	23.27	1.09	0.141
	north	2 058	21.28	1.03	1 994	27.64	1.39	2 026	16.51	0.82	0.224
Head of household income	central	1 831	18.40	1.00	2 270	34.05	1.50	2 052	17.16	0.84	0.168
	south	2 005	21.04	1.05	2 097	42.49	2.03	2 051	20.63	1.01	0.041
	east	1 928	18.61	0.97	2 017	26.84	1.33	1 972	15.27	0.77	0.198
	north-west	2 000	20.86	1.04	2 092	38.55	1.84	2 046	19.31	0.94	0.095
	south-west	2 111	34.24	1.62	2 134	37.40	1.75	2 122	24.96	1.18	0.275
	north	2 108	25.16	1.19	1 990	30.51	1.53	2 049	18.99	0.93	0.223

Note: Note: grey – estimates based on the imputed values, the gain measure is defined as: $1-\frac{s_{int}(\bar{x})}{s_{HBS}(\bar{x})}$ or $1-\frac{s_{int}(\bar{x})}{s_{EU-SILC}(\bar{x})}$.
Source: Own study

Joint observation of variables not jointly observed was not the only gain achieved through data fusion. Thanks to the dataset concatenation approach the integrated dataset contained 49681 units (the sum of sample sizes of input datasets, see Table 1). Enlarged sample size allows to reduce the sampling error (see Table 7). That can be a contribution to the use of statistical data integration methods in small area estimation.

3.4 Quality assessment

The quality of integration was performed using German Association of Media Analysis approach (see section 2.4). The empirical distributions were compared by analysis of characteristics of distribution of integrated variables.

Table 8 Comparison of marginal distributions of appending variables in input and integrated datasets

Variable	Dataset	MI				Mixed			
		Mean	Median	Std. Dev.	Skewness	Mean	Median	Std. Dev.	Skewness
HH expenditures	HBS	1 954.20	1 602.67	1 507.15	5.22	1 954.20	1 602.67	1 507.15	5.22
	EU-SILC (imp.)	1 966.02	1 697.63	1 282.89	10.33	2 085.71	1 825.76	1 162.18	2.45
	Integrated	1 960.11	1 653.46	1 399.59	7.23	2 019.92	1 720.77	1 347.46	4.4
Head of HH income	HBS (imp.)	1 979.49	1 644.79	1 755.34	3.03	1 962.96	1 558.36	1 553.94	4.55
	EU-SILC	2 065.48	1 566.00	1 773.33	3.13	2 065.48	1 566.00	1 773.33	3.13
	Integrated	2 022.46	1 613.19	1 764.87	3.08	2 014.19	1 560.80	1 667.98	3.74

Source: Own study

Analysis of the basic distribution characteristics shows that both methods retain the essential characteristics of the distribution in a good manner.[7] The statistics in the integrated dataset are always located between the values coming from input sets (see Table 8). Also the imputed with described methods values retain similar to the empirical values. It should also be noted that the MI method returns better results when one imputes from smaller to larger dataset and mixed method seems better otherwise.

Table 9 Comparison of joint distribution (correlation coefficient) with selected common variables in input and integrated datasets

Variable	Dataset	MI		Mixed	
		Disposable income	Equivalent income	Disposable income	Equivalent income
HH expenditures	HBS	0.588	0.484	0.588	0.484
	EU-SILC (imp.)	0.855	0.677	0.872	0.675
	Integrated	0.707	0.568	0.706	0.562
Head of HH income	HBS (imp.)	0.965	0.937	0.873	0.834
	EU-SILC	0.887	0.854	0.887	0.854
	Integrated	0.926	0.896	0.877	0.839

Source: Own study

[7] Formal statistical tests, due to the large sample size, always lead to the rejection of the null hypothesis. Comparison of basic characteristics as a measure of the goodness of integration was proposed by D'Orazio et al. (2006).

Analysis of selected joint distributions with chosen common variables indicates that the most consistent results were obtained using the mixed method (see Table 9).

It has to be noted that imputation using mixed model from EU-SILC, which has much smaller sample size, to HBS could disrupt the continuity of the variable since one record can be used more than once. In such case the MI approach seem to be better.

CONCLUSIONS

Conducted empirical study showed that the creation of the integrated data set allows to extend the substantive scope compared to the input datasets. At the same time, estimation accuracy was verified based on the integrated socio-economic dataset. The quality of the integration as well as compliance of joint and marginal distributions of target variables was assessed in the integrated dataset.

The integrated data repository contains information about the joint household financial characteristics. Among others, correlation coefficient between two distinct variables was possible to estimate. Such characteristics was not possible to estimate in any of the input datasets. Concatenation of the input datasets gave the opportunity to create estimates with greater accuracy

At the same time, the results of empirical research enabled the formulation of conclusions of more general nature:

harmonization of definitions and categories of matching variables usually comes down to creation of derivative variables with aggregated (to 'the lowest common denominator') categories, which naturally reduces the amount of information coming from the variable,

- without access to additional information, it is possible to construct high-quality estimators in the integrated dataset using conditional independence assumption,
- each target variable and should be analyzed separately by, among others, selection of appropriate matching variables and integration model.

In final conclusion it should be assumed that the statistical method of data integration will be increasingly used in statistical studies. This is due to two main reasons. First, rising costs of conducting surveys, increasing burden on respondents resulting to missing data, may force the entities to design studies with shorter questionnaires for their subsequent integration. Second, the increase in demand for detailed information on the low level of aggregation will enforce the fusion of information from different sources to increase the effective sample size. Statistical data integration is an opportunity to greatly enrich the methodological workshop of statistical institution in meeting the needs of recipients of information.

ACKNOWLEDGMENT

The author gratefully acknowledges the support of the National Research Council (NCN) through its grant DEC-2011/02/A/HS6/00183.

References

BARR, R. S., TURNER, J. S. *Quality issues and evidence in statistical file merging* [w:] *Data Quality Control: Theory and Pragmatics*. New York: Marcel Dekker, 1990.

CENEX. *Description of the action 2006.* CENEX Statistical Methodology Area "Combination of survey and administrative data".

D'ORAZIO, M., DI ZIO, M., SCANU, M. *Statistical Matching. Theory and Practice*. John Wiley & Sons Ltd., England, 2006.

DI ZIO, M. *What is statistical matching*. Course on Methods for Integration of Surveys and Administrative Data, Budapest, Hungary, 2007.

ESSnet on Data Integration. Report on WP1. State of the art on statistical methodologies for data integration 2011. ESSnet on Data Integration, Rome.

IBM SPSS Missing Values 20. IBM White Papers, 2011.

KADANE, J. B. *Some statistical problems in merging data files* [w:] *Department of Treasury, Compendium of Tax Research*. Washington, DC: US Government Printing Office, 1978.

LANDERMAN, L., LAND, K., PIEPER, C. *An Empirical Evaluation of the Predictive Mean Matching Method for Imputing Missing Values*. Sociological Methods Research August 1997, Vol. 26, No. 1.

LEULESCU, A., AGAFITEI, M. *Statistical matching: a model based approach for data integration* [online]. Eurostat Methodologies and Working Papers, 2013. <http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-13-020/EN/KS-RA-13-020-EN.PDF>.

LITTLE, R. J. A., RUBIN, D. B. *Statistical Analysis with Missing Data*. Wiley, 2002.

MORIARITY, C. *Statistical Properties of Statistical Matching. Data Fusion Algorithm*. Saarbrucken: VDM Verlag Dr. Mueller, 2009.

PAAS, G. *Statistical match: evaluation of existing procedures and improvements by using additional information* [w:] *Microanalytic Simulation Models to Support Social and Financial Policy*. Amsterdam: Elsevier Science, 1986.

RAESSLER, S. *Statistical Matching. A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. New York, USA: Springer, 2002.

RAESSLER, S. Data fusion: identification problems, validity, and multiple imputation. *Austrian Journal of Statistics*, 2004, 33 (1–2).

RODGERS, W. L. An evaluation of statistical matching. *Journal of Business and Economic Statistics*, 1984, 2.

RUBIN, D. B. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, 1986, 4.

SCANU, M. *Some preliminary common aspects for record linkage and matching* [w:] *Report of WP2. Recommendations on the use of methodologies for the integration of surveys and administrative data*. CENEX-ISAD, 2008.

SINGH, A. C., MANTEL, H., KINACK, M., ROWE, G. *Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption*. Survey Methodology 19, 1993.

VAN DER LAAN, P. Integrating administrative registers and household surveys. *Netherlands Official Statistics*, Vol. 15, Summer 2000, Special Issue: Integrating administrative registers and household surveys, Statistics Netherlands, Voorburg/Heerlen.

Developing Improved Predictive Estimator for Finite Population Mean Using Auxiliary Information

Subhash Kumar Yadav[1] | *Dr. Ram Manohar Lohia Avadh University, U.P., India*
Sant Sharan Mishra[2] | *Dr. Ram Manohar Lohia Avadh University, U.P., India*

Abstract

The present paper deals with the estimation of population mean using predictive method of estimation utilizing auxiliary information. In this paper, we have proposed improved ratio cum product type predictive estimators to estimate the population mean. The large sample properties of the proposed estimator have been studied. The expressions for the bias and mean square error (MSE) have been obtained up to the first order of approximation. The minimum value of MSE of proposed estimator is also obtained for the optimum value of the characterizing scalar. A comparison is made with the ratio and product type estimators and the conditions under which the proposed estimator is more efficient than the other mentioned estimators. An empirical study is carried out to justify the theoretical findings.

Keywords

Predictive estimators, auxiliary information, finite population mean, MSE, efficiency

JEL code

C13, C83

INTRODUCTION

Sampling is a method or technique of drawing sample from the population. It is used whenever the population is large and the complete enumeration is very time consuming and costly. Parameters of the population are estimated through their appropriate estimators using the information supplied by the sample and their large sample properties are studied up to a certain order of approximation.

In this paper we have studied the large sample properties of the proposed estimator up to the first order of approximation. The use of auxiliary information prevails since the use of sampling itself. It is a unanimous agreement stating that the proper use of auxiliary information enhances the efficiency of the parameter estimator of the main variable under study. The auxiliary variable which supplies the auxiliary information is highly positively or negatively correlated with the main variable under study.

[1] Assistant Professor, Dept. of Mathematics & Statistics, Faizabad-224001, U.P., India.
[2] Assoc. Professor, Dept. of Mathematics & Statistics, Faizabad-224001, U.P., India. Corresponding author: sant_x2003@yahoo.co.in.

When the auxiliary variable is highly and positively correlated with the main variable then the ratio method of estimation is used in which the ratio type estimators are used for the estimation of parameters. On the other hand, product type estimators are used when the auxiliary variable is highly and negatively correlated with the main variable under study.

In certain situations, the estimation of the population mean of the study variable has received a considerable attention from experts engaged in survey-statistics. For example, in agriculture, the average production of crop is required for further planning or in manufacturing industries and pharmaceutical laboratories, the average life of their products is a necessity for their quality control. Although, in literature, a great variety of techniques have been used mentioning the use of auxiliary information by means of ratio, product and regression methods for estimating population mean and other parameters. However, some efforts in this direction are reported by many authors. Agrawal and Roy (1999) proposed the efficient estimators of population variance using ratio and regression type predictive estimators, Upadhyaya and Singh (1999) used transformed auxiliary variable and proposed the estimator of population mean, Singh (2003), suggested the improved product type estimator of population mean for negative correlated auxiliary variable, Singh and Tailor (2003), utilized the correlation coefficient of auxiliary and main variable and proposed the improved estimator of population mean, Singh *et al.* (2004, 2014), proposed improved estimators using power transformation and the predictive exponential estimators of population mean, respectively, Kadilar and Cingi (2004, 2006), utilized the different parameters of auxiliary variable and proposed improved ration type estimators of population mean, Yan and Tian (2010), suggested the estimators of population mean using coefficient of skewness of auxiliary variable, Yadav (2011), proposed an efficient ratio estimator of population variance using auxiliary variable, Subramani and Kumarapandiyan (2012), suggested the efficient estimator of population mean using coefficient of variation and the median of auxiliary variable, Solanki et al. (2012), proposed an alternative estimator of population mean using variable for improved estimation of population mean, Onyeka (2012), proposed improved estimator of population mean in poststratified random sampling scheme, Jeelani et al. (2013), suggested modified ratio estimators of population mean using linear combination of coefficient of skewness and the quartile deviation of auxiliary variable, Saini (2013), proposed a predictive class of estimators on two stage random sampling, Yadav and Kadilar (2013) proposed the improved class of ratio and product estimators of population mean and Yadav et al. (2014) suggested the improved ratio estimators for population mean based on median using linear combination of population mean and median of an auxiliary variable. They have thoroughly used auxiliary information for improved estimation of population mean through ratio and product type estimators for the main characteristic under study.

In this paper, we attempt to develop an improved estimator for population mean using predictive method of estimation with the help of auxiliary information. The proposed estimator falls under the category of ratio cum product type predictive estimator. The expressions for the bias and mean square error (MSE) have been obtained up to the first order of approximation. Efficiency condition has been attained by using numerical illustration and proved by comparison showing that present estimator is more efficient than previous ones.

1 MATERIAL AND METHODS

In surveys sampling the information on supplementary population is often used at the estimation stage to enhance the precision of estimates of a population mean and other parameters. The use of auxiliary information is common in practice for estimation of the finite population mean of a study character under consideration. A vast variety of approaches is available in literature to construct more and more efficient estimators for the population mean including design based and model based methods. Many authors have discussed the estimation of population mean using auxiliary information on design based methods. Herein, we have considered the use of auxiliary information on model based method also

known as predictive method of estimation of population mean of study variable. The model-based approach or the predictive method of estimation in sampling theory is based on super population models. This approach assumed that the population under consideration was a realization of super-population random variables containing a super population model. Under this super population model the prior information about the population is formalized and used to predict the non-sampled values of the population that is the finite population quantities, mean and other parameters of the study variable. Some of the advantages of super population model approach lie in the fact as: the statistical inferences about the population parameters of the study variable may be drawn using the predictive estimation theory of survey sampling or model based theory. The very well known and popular estimators of population mean in the classical theory are the ratio, regression and other estimators. These estimators can be used as predictors in a general prediction theory under a special model. Many authors have used ratio, product and regression type estimators of population parameters for predictive estimation. It is well established in predictive estimation theory that the use of aforementioned estimators as predictors for population parameters of the unobserved units of the population results in the corresponding estimators of the population parameters for the whole population.

In this paper we propose a predictive estimator of population mean for simple random sampling design using auxiliary variable to construct ratio cum product exponential type estimator by utilizing the prediction criterion.

Let the finite population $U = (U_1, U_2, ..., U_N)$ consists of N distinct and identifiable units. Let the main variable under study be denoted by Y and the auxiliary variable by X. Thus (y_i, x_i), $(i = 1, 2, ..., N)$ denote the ith observations for the main and auxiliary variables respectively. Thus we have $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, the population mean of the main variable to be estimated.

Further, let S denote the set of all possible samples of fixed size from the population U. Let s be a member of S and let $\vartheta(s)$ denote the effective sample size that is the number of distinct elements in s and $\overline{s}$ denote the collection of all those units of U which are not in s. We now denote:

$$\overline{Y}_s = \frac{1}{\vartheta(s)} \sum\nolimits_{i \in s} y_i,$$

$$\overline{Y}_{\overline{s}} = \frac{1}{\left[N - \vartheta(s)\right]} \sum\nolimits_{i \in \overline{s}} y_i.$$

For a given sample $s \in S$, , can be written as:

$$\overline{Y} = \frac{\vartheta(s)}{N} \overline{Y}_s + \frac{\left[N - \vartheta(s)\right]}{N} \overline{Y}_{\overline{s}}. \tag{1}$$

In simple random sampling procedure, the sample mean for the sample of size n (i.e. $\vartheta(s) = n$) is:

$\overline{y} = \frac{1}{n} \sum\nolimits_{i \in s} y_i$ that is $\overline{Y}_s = \overline{y}$.

Thus $\overline{Y}$ in equation (1) can be written as:

$$\overline{Y} = \frac{n}{N} \overline{Y}_s + \frac{N - n}{N} \overline{Y}_{\overline{s}}. \tag{2}$$

In the light of equation (2), an appropriate estimator of population mean $\overline{Y}$ is obtained as:

$$t = \frac{n}{N} \overline{y} + \frac{N - n}{N} T, \tag{3}$$

where T is taken as the predictor of $\overline{Y}_{\overline{s}}$.

Srivastava (1983) proposed the following estimator as the predictor of the population mean $\overline{Y}_{\bar{s}}$ of unobserved units of the population as:

$T = \bar{y} = \frac{1}{n}\sum_{i\in\bar{s}} y_i$, the mean per unit estimator that is the sample mean,

$T = \bar{y}_{\bar{R}} = \overline{X}_{\bar{s}}\left(\frac{\bar{y}}{\bar{x}}\right)$, the classical ratio estimator,

where:

$$\bar{x} = \frac{1}{n}\sum_{i\in s} x_i,\ \overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i \text{ and } \overline{X}_{\bar{s}} = \frac{1}{N-n}\sum_{i\in\bar{s}} x_i = \frac{N\overline{X} - n\bar{x}}{N-n}.$$

Srivastava (1983) has shown that when above estimators are used as predictive estimators of $\overline{Y}_{\bar{s}}$, then the estimator *t* in (3) results in respective classical estimators:

$\bar{y} = \frac{1}{n}\sum_{i\in s} y_i$, the mean per unit estimator that is the sample mean,

$\bar{y}_R = \bar{y}\left(\frac{\overline{X}}{\bar{x}}\right)$, the classical ratio estimator.

However, for product predictive estimator $T = \bar{y}_{\bar{P}} = \bar{y}\left(\frac{\bar{x}}{\overline{X}_s}\right)$ of the population mean $\overline{Y}_{\bar{s}}$ of unobserved units of the population, he has shown that *t* does not result in the classical product estimator $\bar{y}_P = \bar{y}\left(\frac{\bar{x}}{\overline{X}}\right)$ of population mean $\overline{Y}$.

Thus whenever:

$$T = \bar{y}_P = \bar{y}\left(\frac{\bar{x}}{\overline{X}_s}\right),$$

$$t = \frac{n\overline{X} + (N-2n)\bar{x}}{N\overline{X} - n\bar{x}} = t_p. \tag{4}$$

The biases and the mean square errors of the estimators $\bar{y}_R$, $\bar{y}_P$ and t_P, up to the first order of approximations respectively are found as:

$$Bias(\bar{y}_R) = \theta\overline{Y}C_x^2(1-C),$$

$$Bias(\bar{y}_P) = \theta\overline{Y}CC_x^2,$$

$$Bias(t_p) = \theta\overline{Y}C_x^2[C + f(1-f)^{-1}],$$

$$MSE(\bar{y}_R) = \theta\overline{Y}^2[C_y^2 + C_x^2(1-2C)], \tag{5}$$

$$MSE(\bar{y}_P) = MSE(t_p) = \theta\overline{Y}^2[C_y^2 + C_x^2(1+2C)], \tag{6}$$

where:

$$\theta = (1-f)^{-1},\ f = (n/N),\ C_y^2 = (S_y^2/\overline{Y}^2),\ C_x^2 = (S_x^2/\overline{X}^2),\ C = \rho(C_y/C_x),\ \rho = (S_{yx}/S_yS_x),$$

$$S_y^2 = (N-1)^{-1}\sum_{i=1}^{N}(y_i - \overline{Y})^2,\ S_x^2 = (N-1)^{-1}\sum_{i=1}^{N}(x_i - \overline{X})^2 \text{ and } S_{yx} = (N-1)^{-1}\sum_{i=1}^{N}(y_i - \overline{Y})(x_i - \overline{X}).$$

Singh et al. (2014) proposed the following ratio and product type exponential estimators of population mean $\overline{Y}$ using Bahl and Tuteja (1991) ratio and product types exponential estimators of population mean as the predictive estimators of $\overline{Y}_{\bar{s}}$ respectively as:

$$t = t_{\mathrm{Re}} = \left[\frac{n}{N}\bar{y} + \left(\frac{N-n}{N}\right)\bar{y}\exp\left(\frac{\overline{X}_{\bar{s}} - \bar{x}}{\overline{X}_{\bar{s}} + \bar{x}}\right)\right] = \left[\frac{n}{N}\bar{y} + \left(\frac{N-n}{N}\right)\bar{y}\exp\left(\frac{N(\overline{X} - \bar{x})}{N(\overline{X} - \bar{x}) - 2n\bar{x}}\right)\right] \tag{7}$$

$$t = t_{Pe} = \left[\frac{n}{N}\bar{y} + \left(\frac{N-n}{N}\right)\bar{y}\exp\left(\frac{\bar{x} - \overline{X}_{\bar{s}}}{\bar{x} + \overline{X}_{\bar{s}}}\right)\right] = \left[\frac{n}{N}\bar{y} + \left(\frac{N-n}{N}\right)\bar{y}\exp\left(\frac{N(\bar{x} - \overline{X})}{N\overline{X} + (N-2n)\bar{x}}\right)\right] \tag{8}$$

The biases and the mean square errors of above estimators up to the first order of approximations respectively are given as:

$$Bias(t_{\mathrm{Re}}) = \frac{\theta}{8}\overline{Y}C_x^2[3 - 4(C + f)],$$

$$Bias(t_{Pe}) = \frac{\theta}{8}\overline{Y}C_x^2\left[4C - \frac{1}{(1-f)}\right],$$

$$MSE(t_{\mathrm{Re}}) = \theta\overline{Y}^2[C_y^2 + \frac{C_x^2}{4}(1 - 4C)], \tag{9}$$

$$MSE(t_{Pe}) = \theta\overline{Y}^2[C_y^2 + \frac{C_x^2}{4}(1 + 4C)]. \tag{10}$$

Note: The mean square errors of above Singh et.al (2014) estimators are equals to the mean square errors of Bahl and Tuteja (1991) estimators.

2 PROPOSED ESTIMATORS

Motivated by Singh et. al (2014), we have proposed the following improved ratio cum product type exponential estimator in predictive estimation approach as:

$$\tau = \alpha t_{\mathrm{Re}} + (1 - \alpha)t_{Pe}, \tag{11}$$

where, t_{Re} and t_{Pe} are the estimators in (7) and (8) respectively and is the characterizing scalar to be determined such that the mean square error of the proposed estimator τ is minimum.

To study the large sample properties of the proposed estimator τ, we define, $\bar{y} = \overline{Y}(1 + e_0)$ and $\bar{x} = \overline{X}(1 + e_1)$ such that $E(e_0) = E(e_1) = 0$ and up to the first order of approximation:

$$E(e_0^2) = \theta C_y^2,\ E(e_1^2) = \theta C_x^2 \text{ and } E(e_0 e_1) = \theta C C_x^2.$$

Expressing t_{Re} in terms of e_i's, we have:

$$t_{\mathrm{Re}} = \overline{Y}(1 + e_0)\left[\frac{n}{N} + \left(\frac{N-n}{N}\right)\exp\left(\frac{Ne_1}{2(N-n) + (N-2n)e_1}\right)\right],$$

$$= \overline{Y}(1 + e_0)\left[f + (1-f)\exp\left(-\frac{e_1}{2(1-f) + (1-2f)e_1}\right)\right],$$

$$= \bar{Y}(1+e_0)\left[f + (1-f)\exp\left\{ -\frac{e_1}{2(1-f)}\left(1+\frac{(1-2f)}{2(1-f)}e_1\right)^{-1}\right\}\right].$$

Expanding the right hand side of above equation, simplifying after multiplication and retaining the terms up to the first order of approximation, we have:

$$t_{\mathrm{Re}} = \bar{Y}\left\{1+e_0-\frac{e_1}{2}-\frac{e_0e_1}{2}+\frac{e_1^2}{8}(3-4f)\right\}. \tag{12}$$

Similarly expressing t_{Pe} in terms of e_i's, we have:

$$t_{Pe} = \bar{Y}(1+e_0)\left[\frac{n}{N}+\left(\frac{N-n}{N}\right)\exp\left(\frac{Ne_1}{2(N-n)+Ne_1}\right)\right],$$

$$= \bar{Y}(1+e_0)\left[f + (1-f)\exp\left(\frac{e_1}{2(1-f)+e_1}\right)\right],$$

$$= \bar{Y}(1+e_0)\left[f + (1-f)\exp\left\{\frac{e_1}{2(1-f)}\left(1+\frac{e_1}{2(1-f)}\right)^{-1}\right\}\right].$$

Expanding the right hand side of above equation, simplifying after multiplication and retaining the terms up to the first order of approximation, we have:

$$t_{Pe} = \bar{Y}\left\{1+e_0+\frac{e_1}{2}+\frac{e_0e_1}{2}-\frac{e_1^2}{8(1-f)}\right\}. \tag{13}$$

Now expressing τ from (11) in terms of e_i's, using (12) and (13), we have:

$$\tau = \bar{Y}\left[\alpha\left\{1+e_0-\frac{e_1}{2}-\frac{e_0e_1}{2}+\frac{e_1^2}{8}(3-4f)\right\}+(1-\alpha)\left\{1+e_0+\frac{e_1}{2}+\frac{e_0e_1}{2}-\frac{e_1^2}{8(1-f)}\right\}\right].$$

More simplifying, we get:

$$\tau = \bar{Y}\left[1+e_0-(2\alpha-1)\frac{e_1}{2}-(2\alpha-1)\frac{e_0e_1}{2}+\frac{e_1^2}{8}\left\{\frac{4\alpha-1-7f+4f^2}{(1-f)}\right\}\right]. \tag{14}$$

Subtracting $\bar{Y}$ on both sides of (14) and taking expectation on both sides, we get the bias of τ as:

$$Bias(\tau) = \bar{Y}\left[E(e_0)-(2\alpha-1)\frac{E(e_1)}{2}-(2\alpha-1)\frac{E(e_0e_1)}{2}+\frac{E(e_1^2)}{8}\left\{\frac{4\alpha-1-7f+4f^2}{(1-f)}\right\}\right].$$

Putting the values of different expectations, we get:

$$Bias(\tau) = \bar{Y}\left[\frac{1}{8}\left\{\frac{4\alpha-1-7f+4f^2}{(1-f)}\right\}\theta C_x^2-(2\alpha-1)\frac{1}{2}\theta CC_x^2\right]. \tag{15}$$

From (14), we have the mean square error of the estimator τ, up to the first order of approximation as:

$$MSE(\tau) \cong E(\tau - \bar{Y})^2$$

$$= E\left[\bar{Y}\left\{e_0 - (2\alpha - 1)\frac{e_1}{2}\right\}\right]^2,$$

$$= \bar{Y}^2 E\left[e_0 - \alpha_1 \frac{e_1}{2}\right]^2, \text{ where } \alpha_1 = (2\alpha - 1),$$

$$= \bar{Y}^2 E\left[e_0^2 + \alpha_1^2 \frac{e_1^4}{4} - \alpha_1 e_0 e_1\right],$$

$$= \bar{Y}^2 \left[E(e_0^2) + \alpha_1^2 \frac{E(e_1^4)}{4} - \alpha_1 E(e_0 e_1)\right].$$

Putting the values of different expectations, we get:

$$MSE(\tau) = \bar{Y}^2 \left[\theta C_y^2 + \alpha_1^2 \frac{1}{4}\theta C_x^2 - \alpha_1 \theta C C_x^2\right], \tag{16}$$

which is minimal for:

$$\alpha_1 = \frac{2\theta C C_x^2}{\theta C_x^2} = 2C \text{ or } \alpha = \frac{1}{2}(1 + 2C). \tag{17}$$

And the minimum mean square error of τ is:

$$MSE_{\min}(\tau) = \theta \bar{Y}^2 (C_y^2 - C C_x^2). \tag{18}$$

3 EFFICIENCY COMPARISON

Using (5) and (18), we fairly get that:

$$MSE(\bar{y}_R) - MSE(\tau) = \theta \bar{Y}^2 C_x^2 (1 - C) > 0. \tag{19}$$

In view of (6) and (18), we obtain that:

$$MSE(\bar{y}_P) - MSE(\tau) = \theta \bar{Y}^2 C_x^2 (1 + 3C) > 0. \tag{20}$$

Equations (9) and (18) are used to get the following,

$$MSE(t_{\text{Re}}) - MSE(\tau) = \theta \bar{Y}^2 \frac{C_x^2}{4} > 0. \tag{21}$$

And finally equations (10) and (18) produce that:

$$MSE(t_{Pe}) - MSE(\tau) = \theta \bar{Y}^2 C_x^2 \left(\frac{1}{4} + 2C\right) > 0. \tag{22}$$

4 EMPIRICAL STUDY

To justify the theoretical improvements of the proposed estimator τ over the estimators $\bar{y}$, $\bar{y}_R$, $\bar{y}_P$, t_p, t_{Re}, and t_{Pe} of population mean in predictive estimation theory, we have considered the following three natural populations given in the Table 1.

Table 1 The data used in the study

Population	N	n	C_y	C_x	ρ	C
I. Steel and Torrie (1960) y: Log of leaf burn in sec. x: Chlorine percentages	30	6	0.7493	0.7000	0.4996	0.4667
II. Das (1988) y: The number of agricultural laborers for 1961 x: The number of agricultural laborers for 1971	278	60	1.6198	1.4451	0.7213	0.6435
III. Cochran (1977) y: The number of persons per block x: The number of rooms per block	20	8	0.1281	0.1445	0.6500	0.7332

Source: Own construction

We wish to elaborate the tabulated values in various columns of the Table 1. In the first column of the table, various populations namely Steel and Torrie (1960), Das (1988) and Cochran (1977) are given along with their characteristics denoted as x and y respectively. Second and third columns present respective sizes of populations and samples. Coefficients of variations of two characteristics of each population are arranged in columns fourth and fifth, respectively, which evince the variation propensity of x and y characteristics. Correlation coefficients between two variables are put up in sixth column which are evident as positive. Last column is devised as a ratio of coefficient of variations of x, y and product with correlation coefficient to easily simplify the MSE involved in the discussion of result.

Table 2 The PREs of different estimators with respect to $\bar{y}$

Population	PRE($\bar{y}_R$, $\bar{y}$)	PRE($\bar{y}_P$, or t_p, $\bar{y}$)	PRE(t_{Re}, $\bar{y}$)	PRE(t_{Pe}, $\bar{y}$)	PRE(τ, $\bar{y}$)
I	92.9156	31.1004	133.0374	54.9076	**214.94**
II	156.3967	25.8171	197.7846	47.1121	**520.27**
III	157.8695	34.0327	161.2267	56.4111	**235.93**

Source: Own construction

where:

$$PRE(.,\bar{y}) = \frac{MSE(\bar{y})}{MSE(.)}\,100\,.$$

Further, Table 2 is subject to description such that first column of the table presents percentage relative efficiency of ratio estimators of all three populations. Similarly, percentage relative efficiencies of product estimator corresponding to three populations are shown in second column of the Table 2. Furthermore, relative percentage efficiencies of ratio and product exponential corresponding to the populations are fairly accommodated in third and fourth columns of the table. Last column of the table provides relative percentage efficiencies of proposed estimator corresponding to the populations under study which enabled final insight to identify them as the most efficient estimators.

RESULTS AND CONCLUSION

Sample surveys are legitimately considered as cost effective apparatus for estimation of the population parameter. The main aim of statistician is to minimize the mean square error in estimation to ideally infer the parameter of the given population. Before conducting the survey an auxiliary information is used to minimize the error in the estimation so that efficiency of the estimator goes up. To this end, we have made comparisons of desired results with previous researchers.

From the theoretical discussion of efficiency comparisons and the results in Table 2, we infer that the proposed estimator τ is better than the estimators $\bar{y}_R$, $\bar{y}_P$, t_p of Srivastava (1983) and the estimators t_{Re}, t_{Pe}, of Singh et al. (2014) as it shows smaller mean square error than all these estimators. We observe from Table 2 that the percentage relative efficiency (PRE) of the proposed estimator τ with respect to the sample mean is larger than the PREs of estimators $\bar{y}_R$, $\bar{y}_P$, t_p, t_{Re}, and t_{Pe} of population mean in predictive estimation approach. If we numerically compute the percentage increase of efficiencies it comes out to very high corresponding to the populations. For first population, percentage increase of proposed estimator with respect to ratio estimator is about 131, with respect to product estimator is about 594, with respect to ratio exponential and product exponential are 62 and 291, respectively, which are significantly higher for positive decision in favour of proposed estimator. Likewise, we can repeat the same process to get the increasing nature of proposed estimator for remaining populations. Therefore, the proposed estimator τ should be preferred for the estimation of population mean as most efficient estimator in predictive estimation approach.

ACKNOWLEDGMENT

Authors are very much thankful to the Managing Editor of the Statistika journal and two anonymous referees for their valuable and constructive remarks for improving the earlier draft.

References

AGRAWAL, M. C., ROY, D. C. *Efficient estimators of population variance with regression-type and ratio-type predictor-inputs*. Metron, 1999, 57(3), pp. 4–13.

COCHRAN, W. G. *Sampling Techniques*. 3rd ed. New York, USA: John Wiley and Sons, 1977.

DAS, A. K. *Contributions to the Theory of Sampling Strategies Based on Auxiliary Information*. Ph.D. thesis submitted to B. C. K. V. Mohanpur, Nadia, West Bengal, India, 1988.

JEELANI, M. I., MAQBOOL, S., MIR, S. A. Modified Ratio Estimators of Population Mean Using Linear Combination of Coefficient of Skewness and Quartile Deviation. *International Journal of Modern Mathematical Sciences*, 2013, 6, 3, pp. 174–183.

KADILAR, C., CINGI, H. Ratio estimators in simple random sampling. *Applied Mathematics and Computation*, 2004, 151, pp. 893–902.

KADILAR, C., CINGI, H. An improvement in estimating the population mean by using the correlation coefficient. *Hacettepe Journal of Mathematics and Statistics*, Vol. 2006, 35 (1), pp. 103–109.

MURTHY, M. N. *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta, 1967.

ONYEKA, A. C. Estimation of population mean in poststratified sampling using known value of some population parameter(s). *Statistics in Transition-New Series*, 2012, 13, pp. 65–78.

PANDEY, H., YADAV, S. K, SHUKLA, A. K. An improved general class of estimators estimating population mean using auxiliary information. *International Journal of Statistics and Systems*, 2011, 6, pp. 1–7.

SAINI, M. A. Class of predictive estimators in two-stage sampling when auxiliary character is estimated at SSU level. *International Journal of Pure and Applied Mathematics*, 2013, 85 (2), pp. 285–295.

SOLANKI, R. S, SINGH, H. P, RATHOUR, A. An alternative estimator for estimating the finite population mean using auxiliary information in sample surveys. *International Scholarly Research Network: Probability and Statistics*, 2012, pp. 1–14.

SINGH, G. N. On the improvement of product method of estimation in sample surveys. *JISAS*, 2003, 56, pp. 267–275.

SRIVASTAVA, S. K. Predictive estimation of finite population mean using product estimator. *Metrika*, 1983, 30, pp. 93–99.

SINGH, H. P., TAILOR, R. Use of known correlation coefficient in estimating the finite population means. *Statistics in Transition*, 2003, 6 (4), pp. 555–560.

SINGH, H. P., TAILOR, R., TAILOR, R., KAKRAN, M. S. An improved estimator of population mean using power transformation. *Journal of the Indian Society of Agricultural Statistics*, 2004, 58 (2), pp. 223–230.

SINGH, H. P., SOLANKI, R. S., SINGH, A. K. Predictive Estimation of Finite Population Mean Using Exponential Estimators. *Statistika: Statistics and Economy Journal*, 2014, 94 (1), pp. 41–53.

STEEL, R. G. D., TORRIE, J. H. *Principles and Procedures of Statistics*. New York, USA: McGraw, 1960.

SUBRAMANI, J., KUMARAPANDIYAN, G. Estimation of Population Mean Using Coefficient of Variation and Median of an Auxiliary Variable. *International Journal of Probability and Statistics*, 2012, 1 (4), pp. 111–118.

UPADHYAYA, L. N., SINGH, H. P. Use of transformed auxiliary variable in estimating the finite population mean. *Biometrical Journal*, 1999, 41 (5), pp. 627–636.

YADAV, S. K. Efficient estimators for population variance using auxiliary information. *Global Journal of Mathematical Sciences: Theory and Practical*, 2011, 3, pp. 369–376.

YADAV, S. K., KADILAR, C. Improved class of ratio and product estimators. *Applied Mathematics and Computation*, 2013, 219, pp. 10726–10731.

YADAV, S. K., MISHRA, S. S., SHUKLA, A. Improved ratio estimators for population mean based on median using linear combination of population mean and median of an auxiliary variable. *American Journal of Operations Research*, Scientific and Academic Publishing, 2014, 4, 2, pp. 21–27.

YAN, Z., TIAN, B. Ratio method to the mean estimation using coefficient of skewness of auxiliary variable. *ICICA*, Part II, CCIS, 2010, 106, pp. 103–110.

The Dutch Census 2011[1]

Eric Schulte Nordholt[2] | *Statistics Netherlands, The Hague, Netherlands*

Abstract

The Dutch 2011 Census tables were produced by combining existing register and sample survey data. Since the last census based on a complete enumeration was held in 1971, the willingness of the population to participate has fallen sharply. Statistics Netherlands no longer uses census questionnaires and has found an alternative in the register-based census, using only existing data. The register-based census is cheaper and more socially acceptable. The table results of the Netherlands are not only comparable with earlier Dutch censuses, but also with those of the other countries in the 2011 European Census Round.

Keywords

Census, methodology, registers, Statistics Netherlands

JEL code

C18

INTRODUCTION

The 2011 European Census Round was coordinated by Eurostat for all European Union (EU) and European Free Trade Association (EFTA) member states. The EU population and housing censuses have a broad basis: they are covered by four Regulations (European Commission, 2008, 2009, 2010a and 2010b), which have served to harmonise population definitions, census variables and categories, census hypercubes (high-dimensional tables) and metadata within the EU. Moreover, they specify the technical format (SDMX) for data delivery. All EU member states were required to conduct a census for 2011. For most national statistical institutes this was a major operation involving a lot of work and high costs. Each country had to collect census data and validate and protect its census output in the hypercubes. All data had to be transformed to SDMX format and put in the so-called Census Hub. Lastly, in addition to sixty mandatory hypercubes, all countries had to produce a number of quality hypercubes and a metadata file describing the methodology used.

Census experts at Statistics Netherlands started preparations for the 2011 Population and Housing Census in 2008. In 2009 they started work on the data collection procedures required to collect the census information about the 16 655 799 people living in the Netherlands on 1 January 2011.

The 2011 Census in the Netherlands resulted in sixty hypercubes (European Commission, 2010a). Five relate to the Netherlands as a whole, forty contain data at provincial level (NUTS 2), ten at COROP level (NUTS 3) and five at municipal level (LAU 2). The sixty hypercubes fall into three different groups: five are about housing, four relate to commuting and the remainder are demographic tables, concerning economic activity, occupation and level of education, for example.

[1] This article is based on Schulte Nordholt (2014a and 2014b).

[2] Statistics Netherlands, P. O. Box 24500, 2490 HA The Hague, The Netherlands. E-mail: e.schultenordholt@cbs.nl, phone: (+31)703374931.

1 THE DUTCH POPULATION AND HOUSING CENSUS 2011

1.1 A register-based census

Data from different sources were combined to produce the 2011 Census tables. These data were not obtained by interviewing inhabitants in a complete enumeration, as in traditional censuses in most other countries, but by using data from registers and sample surveys that are already available at Statistics Netherlands. This approach has a number of advantages and disadvantages.

One of the advantages of this innovative approach is a much lower census bill for Dutch tax payers. A traditional census in the Netherlands would cost a few hundred million euros, while with this method it costs 'only' around 1.4 million euros. This bill includes the costs for all preparatory work, such as extending the methodology and updating and developing accompanying software, as well as the analyses of the results. It does not include the costs of the registers, as these are not kept for censuses but primarily for other purposes. Also, under Dutch law, Statistics Netherlands may access government registers free of charge. This low-cost census approach is only possible for countries with sufficient register information. By way of example, let us compare the costs of the Dutch register-based census with those of the traditional census held in the United Kingdom in 2011. In the United Kingdom the census cost approximately 565 million euros. In terms of PPP per capita (in 2011 US dollars), the census cost 11.82 in the UK, compared with 0.10 in the Netherlands (United Nations, 2014). A register-based census costing less than 1 percent of a traditional census is not exceptional. Today, the huge costs of traditional censuses are often justified by pointing out the enormous implications of the census results for regional funding distribution. But a register-based census would be impossible in the UK anyway, because of the lack of sufficient register data and access restrictions.

Apart from the financial aspect, there are also other important differences between a traditional census and the register-based census conducted in the Netherlands. A well-known problem with traditional censuses is that participation is limited and selective. In spite of the mandatory character of a traditional census, part of the population will not participate at all (unit non-response) and those who do will not answer all questions (item non-response). Although correcting for non-response by weighting and imputation techniques is worth trying, traditional correction methods are inadequate to obtain reliable results. The last traditional census in the Netherlands, in 1971, met with many privacy objections against the collection of integral information about the population living in the Netherlands. This increased the non-response problem, and non-response was expected to be even higher if another traditional census were to be held in the Netherlands (Corbey, 1994). There are almost no objections to a register-based census in the Netherlands and the non-response problem only plays a role when survey microdata are reused.

Another advantage of the register-based census is the short production time. The register-based census in the Netherlands got off to a later start than traditional censuses in other countries. It would have been pointless to start the production phase of the 2011 census project before all sources were available, and some registers became available relatively late. In spite of this delay, Statistics Netherlands compiled its census tables faster than most other countries in the 2011 European Census Round. In fact, the Netherlands had one of the shortest production times for the complete set of tables required by Eurostat. Statistics Netherlands had the advantage that no incoming census forms had to be checked and corrected.

A disadvantage of the Dutch census is that for some variables only sample information is available, which meant it was impossible to meet the level of detail required in some census hypercubes. At the moment, however, the Netherlands perceives the advantages of the register-based census in terms of cost and non-response problems to amply outweigh the loss of some detail compared with a traditional census.

Statistics Netherlands is not the only country that uses registers to produce census information. Four Nordic countries (Denmark, Finland, Norway and Sweden), Austria and Slovenia have more variables available in registers than the Netherlands, and the problem of insufficient detail in the outcome does not play a major role there. Most of the other register-based countries are in a similar position to the Nether-

lands: not all variables relevant for the census can be found in registers. They are therefore very interested in the Dutch approach of combining registers and existing sample surveys and using modern statistical techniques and accompanying software to compile the hypercubes. Obviously, it is essential that statistical bureaus are permitted to make use of registers that are relevant for the census. For Statistics Netherlands this is laid down in the statistical law that came into force in 2004. Nevertheless, Statistics Netherlands will have to maintain the good contact it has established with register holders over the last 25 years. Timely deliveries with relevant variables for Statistics Netherlands are crucial for official statistics production.

Figure 1 The Kingdom of the Netherlands in Europe

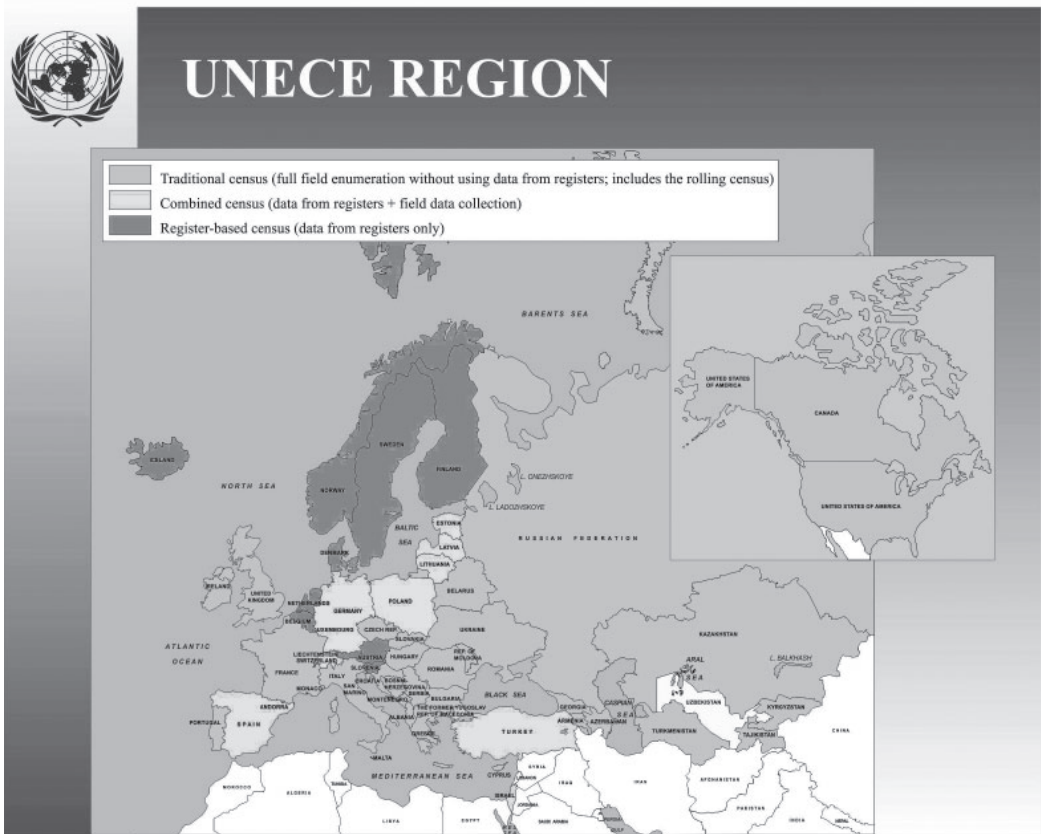
Source: Statistics Netherlands

1.2 The Netherlands

The Kingdom of the Netherlands includes the Netherlands in Europe (see Figure 1) and six islands in the Caribbean. The Kingdom consists of four constituent countries: the Netherlands (consisting of twelve prov-

inces), Aruba, Curaçao and St Maarten. The latter three islands have an independent status as a country within the Kingdom of the Netherlands. The other three Caribbean islands (Bonaire, Saba and St Eustatius) are part of the Netherlands and have had the status of 'special municipality' since 10 October 2010. All four countries produce their own official statistics. Statistics Netherlands has a regional office in the Caribbean responsible for statistics on Bonaire, Saba and St Eustatius (the Caribbean Netherlands). Although the Caribbean Netherlands is part of the Netherlands, statistics on the Netherlands do not include the Caribbean Netherlands. All statistics concerning the Caribbean Netherlands are published separately. The results of the Dutch 2011 Census therefore relate only to the European part of the Netherlands.

Figure 2 Census methods in UNECE countries



Source: UNECE Task Force on Census Methodology (2013)

2 CENSUS METHODS IN THE UNECE REGION

2.1 Census practices in the 2010 Census Round

As mentioned in the previous section, a number of countries conduct register-based censuses. As it is interesting to know how other countries conduct their censuses, in 2013 the United Nations Economic Commission for Europe (UNECE) conducted an online survey among its members to collect information about national census practices in the 2010 Census Round, and about plans for the 2020 round. All EU and EFTA countries are also members of the UNECE, which also includes Canada, the Russian Federation and the United States, among others. Response to the UNECE questionnaire was high and

the results of fifty countries on important methodological issues were analysed (UNECE Task Force on Census Methodology, 2013).

Most countries participating in the online survey had conducted a census in the 2010 Round. Four – Bosnia-Herzegovina, Georgia, Republic of Moldova and Ukraine – had not yet done so, but were still planning to conduct a traditional census during this round. The traditional census in the Former Yugoslav Republic of Macedonia was cancelled in 2011, and there are as yet no firm plans to take another one. This country has therefore been excluded from the following analyses.

As expected the countries used different methods, and some countries reported a different method for the population than for the housing census, often connected with the availability of registers for these domains. Using registers to produce official statistics reduces costs and bypasses the problem of declining survey response rates. Three main types of census method can be distinguished: the traditional census, the combined census, and the register-based census.

2.2 Traditional censuses

The traditional census approach collects basic characteristics from all individuals and housing units (full enumeration) for a specific point in time. More detailed characteristics can be collected either from the whole population or on a sample basis. Collection modes include personal interviews, self-completed paper questionnaires, and data collection by telephone and the internet. Across the world, this is still the most common approach to census-taking. Most UNECE countries with a traditional census use personal face-to-face interviews with paper questionnaires as their main approach. However, in the Czech Republic, France, Ireland, Italy, Luxembourg, Slovakia, the United Kingdom and the United States, the main method is self-completion of paper questionnaires by respondents. In Canada most respondents participate online (CAWI), while in Portugal self-completed paper questionnaires and online response were equally popular.

Just as in the census round of 2000, full field enumeration without register information (traditional census) is still the most popular method in the UNECE region in this census round. Almost two-thirds of countries collected data using 'traditional' methods. But although it is still the most common general approach in the region, it is less so than in the 2000 round, when four-fifths of countries used this approach. A substantial minority (33 percent) of the full field enumeration countries used information from registers only as a frame or control. The United States was alone in using traditional enumeration with yearly updates of characteristics on a sample basis. Another alternative approach to the traditional model was used by France: the rolling census. This is a cumulative continuous survey covering the whole country over a period of time rather than on one particular day.

2.3 Combined censuses

Four countries (Estonia, Latvia, Liechtenstein and Lithuania) used a combination of register data with complete field data collection for selected population census variables, and six countries (Germany, Israel, Poland, Spain, Switzerland and Turkey) used a combination of register data with ad-hoc sample data collection for selected population census variables.

2.4 Register-based censuses

A growing number of EU and EFTA countries have switched to methods without field data collection, relying on registers for their 2011 population and housing Censuses, and skipping census questionnaires completely. Some of these countries 'recycled' information from their labour force surveys, combining it with register data (Belgium, Iceland and the Netherlands). Lastly, some countries used registers only in this census round (Austria, Denmark, Finland, Norway, Slovenia and Sweden). All nine register-based countries collected census information relating to housing entirely from these registers.

2.5 Overview

The map of the UNECE area reveals interesting east-west and north-south tendencies in census methods (see Figure 2). Three main categories are distinguished on the map: traditional (31 countries), combined (10 countries) and register-based (9 countries). Register-based censuses are becoming increasingly popular in northern Europe, combined censuses are more often found in central Europe. Traditional censuses continue to be more popular in English-speaking and Commonwealth of Independent States (CIS) countries. All UNECE countries outside Europe conduct traditional censuses. Only Uzbekistan did not conduct a census in this round, and had no plans to do so.

3 COMPILATION METHODS IN THE NETHERLANDS

The current census results in the Netherlands refer to 2011. The backbone of the Dutch census is the central population register (PR), which combines all the municipal population registers. PR data for 1 January 2011 were used as the basis for the set of hypercubes. The hypercubes focus on frequency counts, not on quantitative information. Data not available or derivable from the PR were taken from other registers. All register variables are now available from Statistics Netherlands' system of social statistical datasets (SSD), and their quality has been improved by applying micro-integration techniques. More information about the SSD can be found in Bakker, Van Rooijen and Van Toor (2014). Micro-integration entails checking the data and adjusting those that are incorrect. It is widely assumed that micro-integrated data provide more reliable results, as they are based on a maximum amount of information. They also provide better coverage of subpopulations: if data are missing in one source, another source can be used.

In the 2011 Census, only two variables were not taken from a register: 'occupation' and 'educational attainment'. Records from the Labour Force Survey (LFS) in a three year period around the enumeration date (1 January 2011) were used to estimate values for these two variables, which are included in 23 of the 60 hypercubes. Table consistency was guaranteed by using repeated weighting for these 23 hypercubes. The method of repeated weighting, described extensively in Houbiers et al. (2003), is based on the repeated application of the regression estimator, generating a new set of weights for each table estimated. The weights of the records in the microdata are adjusted in such a way that a new table estimate is consistent with all earlier table estimates.

We used the latest version of VRD software developed by Statistics Netherlands for this repeated weighting. VRD stands for *Vullen* (= Filling) Reference Database, and the aim of the application is to fill and manage the reference database. The main functions of VRD are estimating tables via repeated weighting, adding these to the reference database, and withdrawing aggregates from the database. Under the condition of small, independent samples, variances of table values can also be estimated. Such estimated variances were used to set publication rules for cells and to calculate variation coefficients for the quality hypercubes, which serve as a quality assessment of the census hypercubes.

To maximise accuracy, all estimates are based on the largest possible number of records. Tables containing only register variables are counted from the registers. Tables with at least one variable from the LFS are estimated from the largest possible combination of register and survey data. Initial weights have to be available for these estimations.

As part of the 2011 Census was compiled on the basis of sample data, margins of inaccuracy have to be taken into account for some results. A rule of thumb was applied for cell values based on a sample from the census population: only estimated table cells based on at least five persons are published. In addition, rare categories have been made confidential to prevent disclosure of individual information.

CONCLUSION

The register-based census has proven to be a successful concept in the Netherlands. It has many advantages compared with traditional censuses: costs are considerably lower, problems with non-response only

play a role when survey microdata are reused, and the production time is much shorter. These advantages more than make up for the loss of some detail in tables based on survey variables. The 2011 Census provides data on the Netherlands that can be compared to results of earlier Dutch censuses and to results of other countries taking part in the 2011 Census Round.

Although most countries in the world still conduct traditional censuses, the Netherlands is not the only country with a register-based census. A number of countries in Europe have switched to combined and register-based censuses. The 2011 Census was the fourth that the Netherlands conducted without census questionnaires.

Just as in the 2001 Census, the repeated weighting technique was used successfully to produce a consistent set of tables for the 2011 Census. A new additional method was introduced for the 23 hypercubes to be estimated. All tables that had to be estimated were based on the largest number of records possible and the resulting hypercubes are mutually consistent. It is important to apply micro-integration of the different sources in the SSD before compiling tables using the estimation techniques. The use of micro-integration and the applied estimation techniques guarantee the consistency between table results from different hypercubes. There is thus no confusion for users of census information, as there is one figure on each socio-economic phenomenon, instead of several figures depending on which sources are used.

References

BAKKER, B. F. M., VAN ROOIJEN, J., VAN TOOR, L. The System of social statistical datasets of Statistics Netherlands: An integral approach to the production of register-based social statistics. *Statistical Journal of the IAOS*, 2014, Vol. 30, pp. 411–424.

CORBEY, P. Exit the population census. *Netherlands Official Statistics*, 1994, Vol. 9, pp. 41–44.

EUROPEAN COMMISSION. Regulation (EC) No 763/2008 of the European Parliament and of the Council of 9 July 2008 on population and housing censuses. *Official Journal of the European Union*, 2008, L218, pp. 14–20.

EUROPEAN COMMISSION. Commission Regulation (EC) No 1201/2009 of 30 November 2009 implementing Regulation (EC) No 763/2008 of the European Parliament and of the Council on population and housing censuses as regards the technical specifications of the topics and of their breakdowns. *Official Journal of the European Union*, 2009, L329, pp. 29–68.

EUROPEAN COMMISSION. Commission Regulation (EU) No 519/2010 of 16 June 2010 adopting the programme of the statistical data and of the metadata for population and housing censuses provided for by Regulation (EC) No 763/2008 of the European Parliament and of the Council. *Official Journal of the European Union*, 2010a, L151, pp. 1–13.

EUROPEAN COMMISSION. Commission Regulation (EU) No 1151/2010 of 8 December 2010 implementing Regulation (EC) No 763/2008 of the European Parliament and of the Council on population and housing censuses, as regards the modalities and structure of the quality reports and the technical format for data transmission. *Official Journal of the European Union*, 2010b, L324, pp. 1–12.

HOUBIERS, M., KNOTTNERUS, P., KROESE, A. H., RENSSEN, R. H., SNIJDERS, V. *Estimating consistent table sets: position paper on repeated weighting* [online]. Discussion paper 03005, Statistics Netherlands, Voorburg/Heerlen, 2003. <http://www.cbs.nl/NR/rdonlyres/6C31D31C-831F-41E5-8A94-7F321297ADB8/0/discussionpaper03005.pdf>.

SCHULTE NORDHOLT, E. Introduction to the Dutch Census 2011. In: SCHULTE NORDHOLT, E., VAN ZEIJL, J, HOEKSMA, L. eds. *Dutch Census 2011, Analysis and Methodology*, Statistics Netherlands, The Hague/Heerlen, November, 2014a, pp. 7–18.

SCHULTE NORDHOLT, E. The Dutch Census 2011. Paper presented at the Conference of European Statistics Stakeholders, Methodologists, Producers and Users of European Statistics (24–25 November 2014, Rome) by Eric Schulte Nordholt, 2014b.

UNECE TASK FORCE ON CENSUS METHODOLOGY. *Census methodology: Key results of the UNECE Survey on National Census Practices, and first proposals about the CES Recommendations for the 2020 census round* [online]. Paper presented at the Fifteenth Meeting of the Group of Experts on Population and Housing Censuses (30 September–3 October 2013, Geneva) by Eric Schulte Nordholt, 2013. <http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2013/census_meeting/3_E_x_15_Aug_WEB_revised_map.pdf>.

UNITED NATIONS. *Measuring Population and Housing. Practices of UNECE countries in the 2010 round of censuses*. New York and Geneva: United Nations, 2014.

Current Problems of National Accounts

Richard Hindls[1] | *University of Economics, Prague, Czech Republic*
Stanislava Hronová[2] | *University of Economics, Prague, Czech Republic*

On 19–21 November 2014, the already 15th ***Conference of the ACN*** (Association de Comptabilité Nationale) was held in Paris, at the Ministry of Finance. This international conference took place, as a rule, under the auspices of the French Statistical Office (INSEE) and University Paris I – Panthéon-Sorbonne. The agenda of the Conference attended by more than 200 experts from ten countries representing statistical offices, universities, research institutes and other national and multinational institutions (Eurostat, International Monetary Fund, OECD), was structured into four relatively independent thematic units. Organizers have precisely set out the subjects of individual sections which subsequently raised a great interest of all participants because of wide range of presented issues and profound related discussion.

The opening ceremony was delivered by Mr. Jean-Luc Tavernier, the INSEE General Manager, and Mr. Philippe Boutry, rector of University Paris I – Pantheon-Sorbonne, who emphasized the significance of regular meetings of experts in the area of national accounts especially in a time of economic turbulences when it is necessary to provide a unified view on economic development in European countries monitored by means of harmonized methodology of the system of national accounts. Unified rules of the system of macroeconomic information now updated by the ESA 2010 standard, represent a condition of mutual understanding of experts in the field of national accounts, economy, statistics, economic policy, and other similar fields. An atmosphere of actual mutual co-operation leads, as a rule, to formal and informal meetings of experts in the above areas. This purpose is also supported by the biennial conference of ACN, an organization associating over 800 experts from 70 countries.

The first day of the conference presided by the ACN honorary chairman and worldwide recognized expert in the field of national accounts, André Vanoli, focused on environmental issues, economy/nature relations, reference framework and works in progress. G. Gagnon from Statistics Canada presented the *System of Environmental – Economic Accounts* (SEEA 2012) as a framework for determination of characteristics of sustainable development and strategies of green economy policy. This system was adopted by the Statistical Commission of the UNO in 2012 as an international standard having the nature of recommendation. Basic framework of SEEA 2012 contains fundamental classification, definitions of indicators and methodological processes in compilation of economic and environmental accounts.

Interesting was also the presentation of the TEEB (*The Economics of Ecosystems and Biodiversity*) project resulted from global initiative aimed at interconnection of research, decision-making centres and economic policy for the purpose of understanding and evaluation of ecosystems, and the French project EFESE (French evaluation of ecosystems and services of ecosystems) which should be a contribution of France to the policy of ecosystems management with respect to its economic, environmental and social impacts.

[1] Faculty of Informatics and Statistics, Department of Statistics and Probability, Nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. E-mail: hindls@vse.cz.
[2] Faculty of Informatics and Statistics, Department of Economic Statistics, Nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. E-mail: hronova@vse.cz.

Other contributions focused on ways of evaluation of unpaid environmental costs, ways of construction of the *Natural Resource Index* within the OECD. Discussion was aimed mainly at the paper of A. Vanoli called *The degradation of the natural assets by economic activities and the central framework of the national accounts*. A. Vanoli has recalled that the interconnection of environmental and national accounts should be approached carefully to avoid non-interpretable characteristics. These systems should be developed in parallel respecting their specifics and differences.

The other subject of the conference was narrowly focussed on issues of *Financial intermediaries and international financial interconnection*. B. Couillot from the Banque de France presented LEI (Legal Entity Identifier), which was born due to the initiative of G20 and Financial Stability Board and which offers a unique possibility of global identification of units acting at financial market. Other contributions dealt with stress tests in network for banks and insurance companies and results of survey of *Co-ordinated Portfolio Investment Survey* organized by the International Monetary Fund. A great attention was naturally paid to the paper of F. Sedillot from Banque de France called *The Who-to-Whom tables in the French financial accounts*, who presented French experience from detailed characteristics of mutual financial relations of sectors displayed by means of transactions and changes in assets and liabilities in national accounts.

The third item of the agenda was aimed at globalization related issues and their reflection in national accounts (measure of world production, creating an integrated statistical framework for understanding global value chains, enterprises, globalisation and the challenge for the national accounts). Revision of French national accounts (transition to EA 2010), stages of its implementation, results and response of professional public and press were presented in the last contribution of the second day of the conference.

The fifth half-day of the conference was focused mainly on the issues of property accounts. J. Ribarsky from OECD explained different ways of evaluation of lands as part of assets of individual countries which were based on the paper prepared by joint working group of Eurostat and OECD. Problems (conceptual and practical) related to evaluation of intellectual property rights in relation to capitalization of the research and development expenses according to ESA 2010 were presented by F. Malherbe from Eurostat. Research, methodology and results of Household Finance and Consumption Survey co-ordinated by the European Central Bank as a complement to current statistical surveys for determination of the structure of household´s assets and their changes were presented in the paper of J. Coffinet and F. Savignac from the Banque de France.

The session of the Conference was concluded by the ACN General Assembly which approved the report on the ACN activities and its economic results and discussed other prospects of future development of this international organization. Let us remind in this respect that membership in the ACN at INSEE is voluntary and free of charge and people interested in becoming members of this association may register at INSEE[3] web-site where all contributions delivered at the Conference[4] are available.

[3] <http://www.insee.fr/en/insee-statistique-publique/formulaire.asp?page=connaitre/colloques/acn/acn-inscription.htm>.
[4] <http://www.insee.fr/en/insee-statistique-publique/default.asp?page=connaitre/colloques/acn/acn15.htm>.

Recent Publications and Events

New publications of the Czech Statistical Office

Foreigners in the Czech Republic 2014. Prague: CZSO, 2014.

KUČERA, L. *Postavení primárního sektoru v ekonomice ČR* (The position of the primary sector in the Czech economy). Prague: CZSO, 2014.

Informační a komunikační technologie v podnikatelském sektoru za rok 2014 (Information and communication technologies in the business sector in 2014). Prague: CZSO, 2014.

Společnost pro etiku v ekonomice. Stručný přehled činnosti v období 1994–2014 (Society for Ethics in Economics. A brief overview of the activities in the period 1994–2014). Prague: CZSO, 2014.

Other selected publications

BALDWIN, R., WYPLOSZ, CH. *Ekonomie evropské integrace* (Economics of European integration). Prague: Grada Publishing, 2013.

BYLOTOV, I., ČAJKA, R., GAJDUŠKOVÁ, K. *Economic Development of the EU New Member States. The impact of the crisis and the role of the single European Currency*. Prague: Oeconomica, 2013.

FIALOVÁ, H., FIALA, J. *Ekonomický slovník s odborným výkladem česky a anglicky* (Economic dictionary with expert commentary Czech and English). Prague: A plus, 2014.

FIALA, P. *Modely a metody rozhodování* (Models and methods of decision making). Prague: Oeconomica, 2013.

HEBÁK, P., KAHOUNOVÁ, J. *Počet pravděpodobnosti v příkladech* (Probability in examples). Prague: Informatorium, 2014.

Lexicon of basic concepts in Polish, Czech and German statistics – methodological notes. Wroclaw, Liberec, Kamenz: Statistical Office in Wroclaw, Regional Office of the Czech Statistical Office in Liberec, Statistical Office of the Free state of Saxony, 2014.

Living conditions in Europe. Luxembourg: Publications Office of the European Union, 2014.

Conferences

The ***12th International Conference on Computational Management Science (CMS) 2015*** will take place between ***27–29 May 2015*** in ***Prague***, organized by the Faculty of Mathematics and Physics, Charles University, Prague. More information available at: *http://cms2015.cuni.cz*.

The ***European Meeting of Statisticians (EMS) 2015*** will be held from ***6th*** **to** ***10th July 2015*** **in** ***Amsterdam*****.** More information available at: *http://www.ems2015.nl*.

The ***60th World Statistics Congress ISI 2015*** will take place between ***26–31 July 2015*** in ***Rio de Janeiro***. The congress will bring together members of the statistical community to present, discuss, promote and disseminate research and best practice in every field of Statistics and its applications. More information available at: *http://www.isi2015.ibge.gov.br*.

The ***18th International Scientific Conference „Applications of Mathematics and Statistics in Economics" (AMSE) 2015***, organized each year by three Faculties of three Universities from three countries (University of Economics, Prague, Czech Republic; Matej Bel University in Banská Bystrica, Slovakia; and Wrocław University of Economics, Poland), will this year be held in the Czech Republic in ***Jindřichův Hradec*** from ***2nd*** to ***6th September 2015***. More information available at: *http://amse2015.cz/conference.*

Papers

We publish articles focused at theoretical and applied statistics, mathematical and statistical methods, conception of official (state) statistics, statistical education, applied economics and econometrics, economic, social and environmental analyses, economic indicators, social and environmental issues in terms of statistics or economics, and regional development issues.

The journal of *Statistika* has the following sections:

The *Analyses* section publishes high quality, complex, and advanced analyses based on the official statistics data focused on economic, environmental, and social spheres. Papers shall have up to 12 000 words or up to twenty (20) 1.5-spaced pages.

The *Discussion* section brings the opportunity to openly discuss the current or more general statistical or economic issues; in short, with what the authors would like to contribute to the scientific debate. Discussions shall have up to 6 000 words or up to 10 1.5-spaced pages.

The *Methodology* section gives space for the discussion on potential approaches to the statistical description of social, economic, and environmental phenomena, development of indicators, estimation issues, etc. Papers shall have up to 12 000 words or up to twenty (20) 1.5-spaced pages.

The *Book Review* section brings further evaluation of selected titles of recent books from the official statistics field (published in the Czech Republic and abroad). Reviews shall have up to 600 words or up to 1 1.5-spaced page.

In the *Information* section we publish informative (descriptive) texts and also information on the latest publications issued (not only from the CZSO) as well as past and upcoming conferences. The maximum range of information is 6 000 words or up to 10 1.5-spaced pages.

Language

The submission language is English only. Authors are expected to refer to a native language speaker in case they are not sure of language quality of their papers.

Recommended Paper Structure

Title (e.g. On Laconic and Informative Titles) — Authors and Contacts — Abstract (max. 160 words) — Keywords (max. 6 words / phrases) — JEL classification code — Introduction — … — Conclusion — Annex — Acknowledgments — References — Tables and Figures

Authors and Contacts

Rudolf Novak*, Institution Name, Street, City, Country
Jonathan Davis, Institution Name, Street, City, Country
* Corresponding author: e-mail: rudolf.novak@domainname.cz, phone: (+420) 111 222 333

Main Text Format

Times 12 (main text), 1.5 spacing between lines. Page numbers in the lower right-hand corner. *Italics* can be used in the text if necessary. *Do not* use **bold** or underline in the text. Paper parts numbering: 1, 1.1, 1.2, etc.

Headings

1 FIRST-LEVEL HEADING (Times New Roman 12, bold)
1.1 Second-level heading (Times New Roman 12, bold)
1.1.1 Third-level heading (Times New Roman 12, bold italic)

Footnotes

Footnotes should be used sparingly. Do not use endnotes. Do not use footnotes for citing references (except headings).

References in the Text

Place reference in the text enclosing authors' names and the year of the reference, e.g. "White (2009) points out that…", "… recent literature (Atkinson et Black, 2010a, 2010b, 2011, Chase et al., 2011, pp. 12–14) conclude…". Note the use of alphabetical order. Include page numbers if appropriate.

List of References

Arrange list of references alphabetically. Use the following reference styles: [for a book] HICKS, J. *Value and Capital: An inquiry into some fundamental principles of economic theory.* Oxford: Clarendon Press, 1939. [for chapter in an edited book] DASGUPTA, P. et al. Intergenerational Equity, Social Discount Rates and Global Warming. In PORTNEY, P., WEYANT, J., eds. *Discounting and Intergenerational Equity.* Washington, D.C.: Resources for the Future, 1999. [for a journal] HRONOVÁ, S., HINDLS, R., ČABLA, A. Conjunctural Evolution of the Czech Economy. *Statistika, Economy and Statistics Journal,* 2011, 3 (September), pp. 4–17. [for an online source] CZECH COAL. *Annual Report and Financial Statement 2007* [online]. Prague: Czech Coal, 2008. [cit. 20.9.2008]. <http://www.czechcoal.cz/cs/ur/zprava/ur2007cz.pdf>.

Tables

Provide each table on a separate page. Indicate position of the table by placing in the text "insert Table 1 about here". Number tables in the order of appearance Table 1, Table 2, etc. Each table should be titled (e.g. Table 1 Self-explanatory title). Refer to tables using their numbers (e.g. see Table 1, Table A1 in the Annex). Try to break one large table into several smaller tables, whenever possible. Separate thousands with a *space* (e.g. 1 528 000) and decimal points with a *dot* (e.g. 1.0). Specify the data source below the tables.

Figures

Figure is any graphical object other than table. Attach each figure as a separate file. Indicate position of the figure by placing in the text "insert Figure 1 about here". Number figures in the order of appearance Figure 1, Figure 2, etc. Each figure should be titled (e.g. Figure 1 Self-explanatory title). Refer to figures using their numbers (e.g. see Figure 1, Figure A1 in the Annex).

Figures should be accompanied by the *.xls, *.xlsx table with the source data. Please provide cartograms in the vector format. Other graphic objects should be provided in *.tif, *.jpg, *.eps formats. Do not supply low-resolution files optimized for the screen use. Specify the source below the figures.

Formulas

Formulas should be prepared in formula editor in the same text format (Times 12) as the main text.

Paper Submission

Please email your papers in *.doc, *.docx or *.pdf formats to statistika.journal@czso.cz. All papers are subject to double-blind peer review procedure. You will be informed by our managing editor about all necessary details and terms.

Contacts

Journal of Statistika | Czech Statistical Office
Na padesátém 81 | 100 82 Prague 10 | Czech Republic
e-mail: statistika.journal@czso.cz
web: www.czso.cz/statistika_journal

Contact us
Managing Editor: Jiří Novotný
phone: (+420) 274 054 299
fax: (+420) 274 052 133
e-mail: statistika.journal@czso.cz
web: www.czso.cz/statistika_journal
address: Czech Statistical Office | Na padesátém 81 | 100 82 Prague 10 | Czech Republic

Subscription price (4 issues yearly)
CZK 372 (incl. postage) for the Czech Republic,
EUR 110 or USD 165 (incl. postage) for other countries.
Printed copies can be bought at the Publications Shop of the Czech Statistical Office (CZK 66 per copy).
address: Na padesátém 81 | 100 82 Prague 10 | Czech Republic

Subscriptions and orders
MYRIS TRADE, s. r. o.
P. O. BOX 2 | 142 01 Prague 4 | Czech Republic
phone: (+420) 234 035 200,
fax: (+420) 234 035 207
e-mail: myris@myris.cz

Design: Toman Design
Layout: Ondřej Pazdera
Typesetting: Josef Neckář
Print: Czech Statistical Office

All views expressed in the journal of Statistika are those of the authors only and do not necessarily represent the views of the Czech Statistical Office, the Editorial Board, the staff, or any associates of the journal of Statistika.

© 2015 by the Czech Statistical Office. All rights reserved.

95th year of the series of professional economy and statistics journals of the State Statistical Service in the Czech Republic: **Statistika** (since 1964), **Statistika a kontrola** (1962–1963), **Statistický obzor** (1931–1961), **Československý statistický věstník** (1920–1930).

Published by the Czech Statistical Office
ISSN 1804-8765 (Online)
ISSN 0322-788X (Print)
Reg. MK CR E 4684