STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

Volume 13, Number 2, Summer 2012

Editorial Office **ISSN 1234-7655**
Marek Cierpiał-Wolan, Ph.D.: Scientific Secretary
m.wolan@stat.gov.pl
Beata Witek: Secretary
b.witek@stat.gov.pl. Phone number 00 48 22 — 608 33 66
Rajmund Litkowiec: Technical Assistant
Address for correspondence
SiTns, Editorial Office, GUS, al. Niepodległości 208, 00-925 Warsaw, POLAND, Tel./fax:00 48 22 — 825 03 95

STATISTICS IN TRANSITION-new series, Summer 2012
Vol. 13, No. 2, pp. 219—222

FROM THE EDITOR

This issue is composed of two unequal in size and different in character parts. The first part, and the main one, contains a selection of papers based on presentations at the Congress of Polish Statistics that was held - under the honorary patronage of the President of the Republic of Poland, Mr Bronisław Komorowski - in April 18-20 2012, in the city of Poznań, one of the country's leading academic centers, to commemorate the one hundredth anniversary of founding of the Polish Statistical Society. The other part - which 'represents' a regular type of papers published in this journal - is composed of three articles, from China, India and Poland.

While providing an excellent opportunity to the international community of statisticians to meet and exchange ideas, assess achievements and discuss developments in practically all areas of statistics - as an academic discipline and as a field of institutional activity - the Congress gave also a chance to honour several distinguished scholars with the newly established Jerzy Spława-Neyman Medal by the Polish Statistical Society. The following eminent scientists were awarded: Sir Anthony Atkinson, Tadeusz Caliński, Malay Ghosh, Zdzisław Hellwig, Jana Jurečková, Wojciech Krzanowski, Kazimierz Zając, Ryszard Zieliński. A short presentation of the laureates is given in first congress' paper.

The selected congress' papers come from different sessions and can be organized into two groups which, to some extent, remind the standard format of the journal's contents: 'sampling and estimation' and 'other articles' with prevalence of topics related to measurement and analysis.

The first group of articles - related to the sampling and estimation issues - is opened by **M. Ghosh's** paper ***Finite Population Sampling: A Model-Design Synthesis*** that is devoted to a general class of Bayes estimators for estimating the finite population mean which also achieve design consistency. In the next paper, ***Application of Rotation Methods in Sample Surveys in Poland,*** **J. Kordos** discusses designs of the surveys across time taking into account different objectives while focusing on partial rotation of sub-samples, and on problems of estimation and data quality in general, with special attention given to the recent research on rotation sampling in Polish literature.

J. Kubacki discusses the problem of ***Estimation of Parameters for Small Areas Using Hierarchical Bayes Method in the Case of Known Model***

Hyperparameters which he confronts with subjective model parameters selection showing advantages of the former and illustrating it for estimation of average per capita income from Polish Household Budget Survey at the level of county (NUTS4/*powiat*), as well as demonstrating the efficiency of hierarchical Bayes estimation compared with other small area methods for HB and EBLUP technique. **J. Wywiał** discusses the problems of ***Application of Order Statistics of Auxiliary Variable to Estimation of the Population Mean*** in a finite population by means of sampling strategies, given that an auxiliary variable is highly correlated with a variable under study (observations of which are the values of the concomitant of the order statistic), and derives the expected value and the variance of the estimator, accuracy of which is assessed on the basis of simulation analysis. **G. Dehnel** presents the issue of ***Estimation for short term statistics*** with intention to exploit the possibility of using administrative registers for the purpose of short term business statistics and to use the opportunity being given by participation of Poland in the MEETS program to take steps towards reforming the country's business statistics.

Second group starts with **M. Krzyśko** and **Ł. Waszak's** paper ***Methods of Representation for Kernel Canonical Correlation Analysis*** focused on new way of finding linear combinations of the original variables having maximal correlation through representing the problem as the generalized eigenvalue and constructing nonlinear canonical correlation analysis in reproducing kernel Hilbert spaces; the results obtained by classical and kernel canonical correlation analysis are compared - in conclusion, Q-KCCA is being recommended as one of the best methods in nonlinear canonical correlation analysis. In paper devoted to ***Independence Analysis of Nominal Data with the Use of Log Linear Models in R*** by **J. Brzezińska** an idea of application of log-linear models to variables measured at the level below interval scale is discussed for the case of using **R** software with the loglm function in MASS library and glm function in Stats library, along with illustration of this usage for datasets in economic area. Another example of analysis of qualitative variables is presented by **H. Dudek** and **J. Landmesser's** paper ***Income Satisfaction and Relative Deprivation*** in which income situation of households in relative terms is analyzed for various socio-demographic groups of households using partially generalized ordered logit models, using data from Household Budget Survey 2009.

An idea of employing ***Ensemble approach for clustering of interval-valued symbolic data*** is discussed by **M. Pełka** for the case of cluster analysis of symbolic data using simulation techniques (based on artificial data sets with known cluster structure) to ensemble clustering based on co-occurrence matrix for

symbolic interval-valued data, compared with single clustering method (according to corrected Rand index). In paper ***Taxonomic Analysis of The Polish Public Health in Comparison With Selected European Countries*** by **A. Wierzbicka,** public health in Poland is being analyzed in relation to the selected European countries using taxonomic methods and identifying countries with the highest level of public health based on medical, economic and social indicators (from the EUROSTAT database). In the last paper of this group, entitled ***Statistics and Sociology: The Mutually-Supportive Development from the Perspective of Interdisciplinarization of Social Research*** by **W. Okrasa,** a methodological issue of the confluence of developments in these two disciplines, statistics and sociology, is discussed with a succinct overview of this interaction over time until the recent advances in counterfactual causal modeling that enhance methodology of social science research in general.

Each of the three papers that constitute a regular type of journal's articles concerns issue of different kinds. **P. Gurgul** and **P. Zając** in paper ***Forecasting of Migration Matrices in Business Demography*** demonstrate that the forecast of migration matrices can be conducted by means of updating procedures, well-known in the I-O theory. Using some of its popular version (RAS and some non-biproportional approaches) the authors calculate measures of the ex-post error of predictions. A ranking of forecasting methods of migration matrices (forecast horizon one) is established taking into account the measures of distance between two matrices and problems with some forecasting methods (with respect to one-step ex-post forecasts of migration matrices) are discussed.

D. K., Shangodoyin, J. F., Ojo, J. O., Olaomi, A. O. Adebile are presenting ***Time series model for predicting the mean death rate of a disease*** for either an emerging disease or re-emerging disease with a bilinear induced model. The estimated death rate converges rapidly to the true parameter value for a given mean death at time t. The derived model could be used in predicting the m-step future death rate value of a given disease. The new concept is illustrated with real life data.

Z. Xing and **L. Chu** refer results of ***Research on Constructing Composite Index of Objective Well-Being From China Mainland*** using the Quality of Life (QOL) index that was initiated in China during 1980s. The concept of QOL is defined in terms of the quality of people's life and is used to analyze both the people's living conditions and their subjective well-being. An analytical system of Chinese people's well-being covers economic well-being, health and basic survival well-being, social well-being, cultural well-being, political well-being

and environmental well-being. The composite indicator of objective well-being was employed to evaluate well-being of people in 30 Chinese provinces.

The final item of this issue is a conference report. The 30th Anniversary International Conference on ***Multivariate Statistical Analysis*** was held on November 7th-9th, 2011 in Łódź, Poland, being dedicated to Professor of mathematics Antoni Łomnicki, Professor of statistics and administrative law Józef Kleczyński, and Professor of mathematics and history of science, and a champion of science, Samuel Dickstein.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION-new series, Summer 2012
Vol. 13, No. 2, pp. 223

SUBMISSION INFORMATION FOR AUTHORS

Statistics in Transition – new series (SiT) is an international journal published jointly by the Polish Statistical Association (PTS) and the Central Statistical Office of Poland, on a quarterly basis (during 1993–2006 it was issued twice and since 2006 three times a year). Also, it has extended its scope of interest beyond its originally primary focus on statistical issues pertinent to transition from centrally planned to a market-oriented economy through embracing questions related to systemic transformations of and within the national statistical systems, world-wide.

The SiT-*ns* seeks contributors that address the full range of problems involved in data production, data dissemination and utilization, providing international community of statisticians and users – including researchers, teachers, policy makers and the general public – with a platform for exchange of ideas and for sharing best practices in all areas of the development of statistics.

Accordingly, articles dealing with any topics of statistics and its advancement – as either a scientific domain (new research and data analysis methods) or as a domain of informational infrastructure of the economy, society and the state – are appropriate for *Statistics in Transition new series.*

Demonstration of the role played by statistical research and data in economic growth and social progress (both locally and globally), including better-informed decisions and greater participation of citizens, are of particular interest.

Each paper submitted by prospective authors are peer reviewed by internationally recognized experts, who are guided in their decisions about the publication by criteria of originality and overall quality, including its content and form, and of potential interest to readers (esp. professionals).

Manuscript should be submitted electronically to the Editor:
sit@stat.gov.pl., followed by a hard copy addressed to
Prof. Wlodzimierz Okrasa,
GUS / Central Statistical Office
Al. Niepodległości 208, R. 287, 00-925 Warsaw, Poland

It is assumed, that the submitted manuscript has not been published previously and that it is not under review elsewhere. It should include an abstract (of not more than 1600 characters, including spaces). Inquiries concerning the submitted manuscript, its current status etc., should be directed to the Editor by email, address above, or w.okrasa@stat.gov.pl.

For other aspects of editorial policies and procedures see the SiT Guidelines on its Web site: http://www.stat.gov.pl/pts/15_ENG_HTML.htm

STATISTICS IN TRANSITION-new series, Summer 2012
Vol. 13, No. 2, pp. 225—234

LAUREATES OF THE JERZY SPŁAWA - NEYMAN MEDAL

Czesław Domański

The Congress of Polish Statistics took place in Poznań between 18th and 20th of April, 2012, under the honorary patronage of the President of the Republic of Poland, Mr Bronisław Komorowski. It was held in order to commemorate the one hundredth anniversary of founding of the Polish Statistical Society. The Congress offered an excellent opportunity to present and assess achievements of Polish and foreign statisticians, the development of the Polish statistical thought and the process of establishing institutions of public statistics.

During the Congress the Jerzy Spława-Neyman[1] Medal, awarded by the Chapter of the Polish Statistical Society, was presented to honour the following eminent Professors:

Sir Anthony Atkinson,

Tadeusz Caliński,

Malay Ghosh,

Zdzisław Hellwig,

Jana Jurečková,

Wojciech Krzanowski,

Kazimierz Zając,

Ryszard Zieliński.

A short presentation of laureates of the Jerzy Spława-Neyman Medal is given below.

[1] Biographical entry of Jerzy Spława-Neyman by T. Ledwina in Statistics in Transition-New Series, March 2012.

Sir Anthony Barnes Atkinson, a British economist renowned for his contributions to the economics of inequality and public sector economics, with special interest in income distribution and related issues. The measure of inequality he has proposed (in his groundbreaking paper in the *Journal of Economic Theory* almost 45 years ago) is known as the Atkinson index or Atkinson's class of inequality measures. He is a Fellow of the British Academy and has been President of the Royal Economic Society, of the Econometric Society, of the European Economic Association, and of the International Economic Association; Honorary Member of the American Economic Association and of the American Academy of Arts and Sciences. He was knighted in 2000 for services to economics and is Chevalier de la Legion d'Honneur (2001).

Professor Atkinson studied at Cambridge University where he earned his master's degree in 1966. In the years 1967-1971 he was the Fellow of St. John's College at Cambridge and then the Professor of economics at the University of Essex (1971-1976). Subsequently, he held the post of the Professor of economics at the University College in London (1976-1979) and of the Tooke Professor of Economic Science and Statistics, London School of Economics (1980-1992) serving also as the Chairman of Suntory-Toyota International Centre for Economics and Related Disciplines (1981-1987); of the Professor of economics at Cambridge University and the Fellow in Churchill College (1992-1994). In the following years he acted as the Warden at Nuffield College in Oxford (1994-2005). Over the decades, Sir Anthony Atkinson was invited to serve as a member or chairman of various scientific and government committees, and of international organizations, including the European Union.

Sir Anthony Atkinson is the author and editor of numerous monographs. He also published over 140 papers in renowned economic and statistical periodicals.

- *Poverty in Britain and the Reform of Social Security*, 1969
- *Unequal Shares – Wealth in Britain*, 1972
- *Economics of Inequality*, 1975, 1983
- *The Distribution of Personal Wealth in Britain* (with A. J. Harrison), 1978
- *Lectures on Public Economics* (with J. E. Stiglitz), 1980
- *Social Justice and Public Policy*, 1982
- *Parents and Children* (with A. K. Maynard and C. G. Trinder), 1983
- *Poverty and Social Security*, 1989
- *Empirical Studies of Earnings Mobility* (with F. Bourguignon and C. Morrisson), 1992
- *Economic Transformation in Eastern Europe and the Distribution of Income* (with J. Micklewright), 1992
- *Public Economics in Action*, 1995

- *Income Distribution in OECD Countries* (with L. Rainwater and T. Smeeding), 1995
- *Incomes and the Welfare State*, 1996
- *Poverty in Europe*, 1998
- *The Economic Consequences of Rolling Back the Welfare State*, 1999
- *Social Indicators* (with B. Cantillon, E. Marlier and B. Nolan), 2002.

Professor Tadeusz Caliński started his studies in agriculture in 1948, having completed secondary education in Karol Marcinkowski Grammar School in Poznań. In 1953 he graduated from the Faculty of Agriculture and Forestry of the Higher School of Agriculture in Poznań and obtained a master's degree. He undertook a three-year course of Mathematics at the Faculty of Mathematics, Physics and Chemistry of the Adam Mickiewicz University in Poznań. In 1961 he was conferred a doctor's degree by the Council of the Faculty of Agriculture and Forestry of the Higher School of Agriculture in Poznań, after he had written his doctoral dissertation under the supervision of Professor Stefan Barbacki. He earned his postdoctoral degree in the field of experimental agricultural science and biometrics in 1966 on the basis of the habilitation thesis published in the Journal of the Royal Statistical Society.

In the years 1953-1988 he worked at the Mathematical Statistics Section of the Higher School of Agriculture in Poznań (the name was changed into the Agricultural University in 1972), and for almost two decades (1968-1984) he held the Chair of the Section. In 1988 Professor Caliński retired; currently he holds the post of Professor Emeritus of the University of Natural Sciences in Poznań.

In 1998 he was conferred the title of Honoris Causa Doctor of the August Cieszkowski Agricultural University in Poznań, and in 2007 the Honoris Causa Doctor of Warsaw University of Life Sciences in Warsaw.

The list of achievements of Professor Tadeusz Caliński is long, comprising, among others, such activities as:

- Founding the school of mathematical statistics and biometrics that became well-known in the country and abroad,
- Acting as a supervisor of 24 doctors, out of whom 11 subsequently became Professors,
- Publishing papers in renowned Polish and foreign periodicals, as well as 2 books published by the leading scientific publishing house Springer Verlag, New York,
- Promoting the idea of mathematical statistics and biometrics among mathematicians by setting up and heading the Commission of Mathematical Statistics, a body within the Committee of Mathematical Sciences of the Polish Academy of Science,
- Promoting statistical and biometric methods among students and academics from different fields of science through lectures and consultations,

- Initiating and actively supporting international cooperation between Polish statisticians and their colleagues from different parts of the world, especially from: Holland, France, Italy, Britain, Germany, Japan and Portugal.

Professor Caliński is the author or co-author of 149 scientific publications, including 72 studies and dissertations, 30 papers, 5 monographs, 25 algorithms with computer programmes (the well-known SERGEN package was also implemented abroad), 4 course books (including 2 monographs published by Springer in their Lecture Notes in Statistics Series). In his scientific output in the field of mathematical statistics, biometrics and experimental science, Professor Caliński is definitely a remarkable continuator of the scientific ideas originated by Jerzy Spława-Neyman.

Professor Malay Ghosh graduated from Calcutta University, India in statistics and obtained his bachelor's and master's degree in 1962 and 1964, respectively. He completed the doctoral studies in the field of statistics at the University of North Carolina, USA in 1969.

At present he works for the Department of Statistics, University of Florida, USA, where he holds the post of the distinguished Professor.

Professor Ghosh is widely known in the world community of statisticians as a specialist in statistics and particularly in the sub-field of Bayes inference called also empirical Bayes. His interests also focus on the so-called small area estimation.

The scientific and research output of Professor Malay Ghosh encompasses several books and around 300 articles. It is worth stressing that his work is characterized by highly original approach and high scientific standards.

Professor Ghosh has been the supervisor of around 40 scientific dissertations. In addition to his work for the University of Florida he has also acted as visiting Professor at many prestigious universities around the world. He has been heading many scientific and research grants related to both the theory and application of statistics.

Professor Malay Ghosh is a member of editorial boards of numerous reputable scientific magazines and a member of honour of many significant organizations and professional associations.

The list of Professor Ghosh's most important publications includes:

- Small Area Estimation: An Appraisal (with J.N.K. Rao) (1994). *Statistical Science*, V9, No. 1,
- Second Order Probability Matching Priors (with R. Mukerjee) (1997). *Biometrika*, V84, No. 4,

- Generalized Linear Models for Small Area Estimation (with K. Natrarajan, T.W.F. Stroud and B. Carlin) (1998). *Journal of the American Statistical Association*, V93, No. 1.

Professor Zdzisław Hellwig was the soldier of the 4th Lancer Brigade of the Home Army during World War II. In 1946 he took up economic studies at the College of Commerce and he continued studies at the Main School of Planning and Statistics in Warsaw which he graduated from in 1952. He was appointed a full Professor in 1972. Over the years Professor Hellwig held many responsible posts: he was held the Chair of Statistics in the College of Economics and the Academy of Economics in Wrocław in the years 1962-1995, and served as the director of the Institute and held the Chair of Statistics and Economic Cybernetics in the years 1969-1995, the Vice-President of the Academy of Economics in 1962-1965, 1968-1970, 1979-1981, and also the UNESCO expert in Paris in the years 1968-1974. Professor Hellwig gave lectures at the Ibadan University in Nigeria in the academic year of 1965/1966, and Tohoku University in Japan in 1997. He is best known as the founder of the school of scientific statistics, econometrics and cybernetics.

Professor Hellwig acted as the supervisor of 34 doctors, out of whom 15 later became Professors. His effort in educating research and academic staff was rewarded several times by the Minister of Education. In 1996 he was presented with a prestigious Prime Minister's Award for outstanding scientific achievement.

Zdzisław Hellwig received numerous awards and medals, e.g. Commander's Cross of the Order of Poland's Revival in 1990. In recognition of his merits in the field of statistics he was conferred honorary doctorates by the Academy of Economics in Krakow (1985) and Charles University in Prague – for his exceptional merits in development of economics and international scientific contacts (1994).

In the field of statistics he developed the Hellwig Method also known as the optimal selection of predicants, the method of coefficients of information capacity – a formal method of selection of explanatory variables for a statistical model (especially for an econometric model). The variables selected for the model should be strongly correlated with the explanatory variable yet weakly correlated among themselves. In addition to the above a numerical criterion is used – the so-called integral capacity of data carrier combinations. In this case all explanatory variables are data carriers.

Hellwig is the author of hundreds of publications. The most important ones include:

- Elements of Probability Calculus and Mathematical Statistics (1959),
- Linear Regression and its Application in Economics (1960),
- Stochastic Approximation (1965),
- Elements of Economic Calculus (1976).

Professor Jana Jurečková earned her master's degree at Charles University in 1962. She went on to obtain her doctor's degree in statistics in 1967 at the Czechoslovakian Academy of Sciences.

Professor Jurečková started working for the Chair of Probability and Statistics at the Department of Mathematics and Physics of Charles University in Prague in 1964. In 1984 she was conferred the degree of Dr Sc., and in 1992 she was awarded a professorship by the President of the Czechoslovakia. At present she works at Jaroslav H'ajek Centre for Theoretical and Applied Statistics.

Professor Jana Jurečková is the author of over 120 scientific papers which mostly appeared in periodicals on statistics and probability theory, and a co-author of a few monographs.

Her research interests encompass a wide range of problems of statistical inference:

- statistical procedures based on series,
- detailed statistical procedures based on so-called M-statistics and L-statistics,
- statistical procedures based on extreme sample observations,
- tail behaviour and its application in statistics,
- asymptotic methods of mathematical statistics ,
- behaviour of finite sample of estimations and analyses.

Achievements of Professor Jurečkova have vastly contributed to the development of statistics and probability theory.

Professional activities :

- since 2003 a member of the Scientific Society of the Czech Republic,
- a member of ISI (International Statistical Institute),
- cooperation with IMS (Institute of Mathematical Statistics),
- Associate Editor, Annals of Statistics, 1979 -1980 and 1992-1997,
- Associate Editor, Journal of American Statistical Association, 2005-2008,
- Associate Editor, Sankhya, 2006-2011.

Selected publications of Professor Jurečkova:

- Robust Statistical Procedures: Asymptotics and Interrelations, J. Wiley, New York 1996,
- Adaptive Regression, Springer, New York 2000,
- Robust Statistical Methods with R, Chapman & Hall/CRC, Boca Raton 2006,
- Methodology in Robust and Nonparametric Statistics, Chapman & Hall/CRC, Boca Raton, London 2012.

Professor Wojciech Janusz Krzanowski obtained his bachelor's degree (B.Sc.) from University of Leeds in 1967 and a Diploma in Mathematical Statistics from University of Cambridge in 1968. He was conferred a Ph.D. at the University of Reading in 1974.

Professor Krzanowski conducted his research and scientific activity in the following research centres:

- Rothamsted Experimental Station, Harpenden, United Kingdom – Scientific Officer 1968-1971,
- RAF Institute of Aviation Medicine, Farnborough, United Kingdom – Senior Research Fellow 1971-1974,
- University of Reading, United Kingdom – Lecturer 1974-1980, Senior Lecturer 1980-1987, Reader 1987-1990,
- University of Exeter, Department of Mathematical Sciences, Exeter, United Kingdom – Professor, Dean of Department of Mathematics in Exeter 1998-2002,
- Editor Journal of the Royal Statistical Society, Series C, 1991-1995,
- Member of editorial board, "Journal of Classification" (1984-present),
- Member of editorial board "Journal of the Royal Statistical Society, Series B" (1990-1992)
- Membership in: Biometric Society, Classification Society, International Statistical Institute, Polish Biometric Society , Royal Statistical Society.

Professor Wojciech Krzanowski is the author of five books and over 100 articles in scientific magazines, and chapters in books:

- "Principles of Multivariate Analysis: a User's Perspective", Oxford University Press, 1988, 2000,
- "Multivariate Analysis Part 1: Distributions, Ordination and Inference",
- "Multivariate Analysis Part 2: Classification, Covariance Structures and Repeated Measurements", [co-author.], Edward Arnold 1994/95,
- "Recent Advances in Descriptive Multivariate Analysis", University Press, Oxford 1995,
- "An Introduction to Statistical Modelling, with Edward Arnold, 1998.

Professor Kazimierz Zając during World War II conducted underground teaching activities and, working for the Polish Red Cross, he organized help for thousands of Polish officers staying in P.O.W. camps. He was sworn into the Home Army, Krosno District, under the name of „Konrad" and he acted as aide-de-camp of the commander of Centrum Krosno outpost, second lieutenant Moskal (the pseudonym "Zrąb").

In the years 1938/39 and after the war in 1945/46 he studied at the Academy of Commerce in Cracow where he took a course in diplomacy. He continued his education at the Main School of Commerce in Warsaw where in the years 1947-1948 he was doing his doctoral studies, attending a seminar in economics given by Professor Edward Lipiński. The degree of the Candidate of Science, which was at the time the equivalent of the Doctor of Economic Science, was conferred on him on 26 of June, 1958. In 1965 Professor Zając became the head of the Chair of Statistics which he managed successfully for over 30 years until his retirement in 1986. He was awarded professorship in 1971, and he became full Professor in 1976.

Professor Zając is a pioneer of research on population income and spending and pay in Poland and the main research results were presented in the following books: "Econometric Methods of Consumption Area Determination" (with B. Podolec, 1978), „Methods of Examination of Market Services", (ed. by K. Zając, 1982).

Another important area of Professor Zając's scientific interests covered demography. The problems of both socio-economic and demographic nature are interwoven in the monograph entitled "Demographic Progress and Economic Development" (with: A. Sokołowski, 1987), „Taxonomy Methods in Socio-Economic Research" (with: J. Pociecha, B. Podolec, A. Sokołowski; 1988).

In the field of statistical methods of quality control Professor Zając wrote a number of papers and a book „Statistical Methods of Quality Control" (with: J. Cyran, J. Steczkowski; 1973).

In the years 1965-1995, as a member of the Scientific Council of the Central Statistical Office, he greatly contributed to setting direction of statistical research and development of methodology and practice of public statistics in Poland. Also, for many years he held a seat in the Scientific Council of the "Statistical News".

In the period of 1980-1986 he was the dean of the Department of Economics of Turnover of the Academy of Economic in Cracow.

For several consecutive terms he was a member of the Main Council of the Polish Statistical Society (PSS) and the President of the Cracow Section of the PSS. In 2000 he was awarded the Honorary Membership of the PSS.

He supervised 38 doctoral dissertations and 22 postdoctoral dissertations.

Professor Zając is the author or co-author of numerous course books and textbooks in statistics. His most outstanding and popular textbook entitled „*The Outline of Statistical Methods"*(PWE, Warsaw) has had 5 editions and has been used by generations of students in Cracow and other Polish academic centres.

In recognition of his achievement Professor Kazimierz Zając was awarded with the Commander's Cross and the Knight's Cross of the Order of Poland's Revival.

Professor Kazimierz Zając died on May 6, 2012.

Professor Ryszard Zieliński earned the master's degree from the Main School of Planning and Statistics, Department of Finance and Statistics in Warsaw, in 1955. He then undertook studies in mathematics at Warsaw University, which he completed in 1964. He was appointed the Professor of mathematical sciences in 1988.

His wide scope of scientific interests encompassed different aspects of statistics, which is well reflected in his output. He wrote mainly on mathematical statistics but also on statistical quality control, asymptotic statistics, robust statistical procedures, non-parametric statistics and generators of random numbers. One of his most remarkable achievements is a definition of statistic robustness which he proposed for the first time in a study of 1983 (*Robust statistical procedures: a general approach, Lecture Notes in Mathematics 982, "Stability problems for stochastic models" Eds. V. V. Kalashnikov and V. M. Zolotarev, Springer-Verlag 1983, 283-295*). His innovative approach gave an impulse to the development of robust methods of estimation and verification of statistical hypotheses.

Professor Zieliński interests also focused on non-parametric estimation of quantiles of probability distribution and widely understood probability of success. His studies have also been successfully applied to problems related to estimation of value at risk (V@R).

Throughout his busy lifetime he published over 100 original scientific studies (the first one appeared in 1956, and the last one in 2012).

In several dozens of his works he presented statistical methods and their applications in a highly approachable way. His textbook of statistics for secondary schools (*Probability Calculus with Elements of Mathematical Statistics, Państwowe Zakłady Wydawnictw Szkolnych, Warsaw, 1973*) has had

many editions and is still used in secondary education. Professor Zielinski also wrote a highly acclaimed textbook in advanced mathematical statistics (*Seven Introductory Lectures in Mathematical Statistics. Mathematic Library Vol. 72. PWN, Warsaw, 1990)* Moreover, he was a gifted translator and translated several books from English, Russian and German, e.g. *C. R. Rao, Liner Models of Mathematical Statistics,, PWN, Warsaw, 1982.*

Professor Zielinski is the author of a few monographs devoted to Monte Carlo methods and generators of random numbers, some of which have been translated into German. He also compiled very popular and widely used *Statistical Tables* (PWN 1972, 1990).

Professor Ryszard Zieliński passed away suddenly on April 30, 2012.

STATISTICS IN TRANSITION new series, Summer 2012
Vol.13, No. 2, pp. 235-242

FINITE POPULATION SAMPLING: A MODEL-DESIGN SYNTHESIS

Malay Ghosh[1]

ABSTRACT

The paper considers a general class of Bayes estimators for estimating the finite population mean which also achieve design consistency. Some exact results are given where Bayes estimators agree with the Horvitz-Thompson or ratio estimators. For a wider class of priors, asymptotic mathematical equivalence of Bayes estimators with the above estimators is provided.

1. Introduction

There are primarily two basic approaches towards inference from sample survey data. The first, the design-based approach, finds estimators of population quantities based on probability distributions generated by a given selection mechanism. In contrast, a model-based approach assumes the population units to be generated from some superpopulation, and the assumed superpopulation model governs any subsequent inference.

There has been a long-standing debate among survey statisticians regarding which one of the two is the preferred inferential approach. The advocates of design-based methods often criticize model-based inference regarding its failure to guard against any possible model misspecification. On the other hand, those advocating the use of models, question the ability of design-based methods to provide inference with sufficient accuracy in the face of small sample sizes, and often in the neeeded justification of large sample approximations for small or moderate samples. Fortunately, in these days, one notices occasional reconciliation of these two approaches (see e.g. Sarndal, 1984; Prasad and Rao, 1999, among others).

While the basic conceptual disagreement between the two approaches cannot be resolved, from an operational point of view, it is often possible

[1]University of Florida.

to find an agreeement between the two. The present article is a modest attempt to provide some general results showing either exact or large sample agreement. We will illustrate our procedures with several examples.

Section 2 of this paper gives some general results showing model-based interpretation of some of the classical design-based estimators including the celebrated Horvitz-Thompson and ratio estimators. In the process, we revisit the one parameter exponential family model in a slightly non-conventional framework. We show that the said estimators plus others can be deduced as special cases of a general expression for the posterior mean under a certain diffuse prior. In Section 3, we continue with the exponential family model as considered in Section 2, and establish design consistency of Bayes estimators of the finite population mean for a wide class of priors. Here, design consistency is defined in the sense of Lahiri and Mukherjee (2007), and will be made precise in Section 3. Lahiri and Mukherjee concluded that a "subjective Bayes estimator is, in general, not design consistent". They also provided an adjustment to the Bayes estimator to achieve design consistency. While not refuting the word "general" in the statement of these authors, we will show in Section 3 that it is sometimes possible to achieve design consistency exactly in the same sense as of Lahiri and Mukherjee (2007) for a general class of heavy-tailed priors, without seeking any adjustment. We will also point out in this section why the adjustment was needed in their Bayesian framework. Some final remarks are made in Section 4.

2. Some exact results

Consider a finite polpulation with units labelled $1, \ldots, N$. Associated with these units are the characteristics of interest denoted by $y_1, \ldots, y_N$. A sample s of fixed size n is drawn from the population. We will denote by $y(s)$ the set of y_i such that $i \in s$. Similarly, we denote by $\bar{s}$ the set of unsampled population units, and by $y(\bar{s})$ the set of y_j such that $j \in \bar{s}$. The objective is to estimate the finite population mean $m(y) = N^{-1} \sum_{i=1}^{N} y_i$.

Under simple random sampling without replacement, the standard design unbiased estimator of $m(y)$ is the sample mean $\bar{y}_s = \sum_{i \in s} y_i/n$. More generally, with unequal probability sampling, the most well-used estimator of $m(y)$ is

$$N^{-1} \sum_{i \in s} y_i/\pi_i,$$

the Horvitz-Thompson estimator, where π_i denotes the probability of selecting unit i, $i = 1, \ldots, N$. Clearly one must have $\sum_{i=1}^{N} \pi_i = E[\sum_{i=1}^{N} I_{[s \ni i]}] = n$, I denoting the usual indicator function.

With auxiliary information x_i available with the y_i, the well-known ratio estimator of

$$m(y) = [(\sum_{i\in s} y_i)/(\sum_{i\in s} x_i)] \sum_{i=1}^{N} x_i/N.$$

Many other alternative estimators of $m(y)$ have been proposed including an estimator of Hajek (1971), and the celebrated "generalized regression estimator" of Sarndal, Swensson and Wretman (1992), but they will not be considered here.

We provide in this section model-based interpretation of some of the well-known design-based estimators including the Horvitz-Thompson and ratio estimators. It is convenient to begin with a version of the one paramater exponential family model, and obtain the posterior mean of $m(y)$ under a diffuse prior. To this end, we prove the following theorem.

Theorem 1. Suppose $y_i|\theta$ are independently distributed with pdf's $f(y_i|\theta) = \exp[(\theta y_i - a_i\psi(\theta))/\sigma_i^2 + h(y_i)]$, $i = 1, \ldots, N$. Here, θ is an unknown parameter, but the a_i and σ_i^2 are known constants. Consider the prior $\pi(\theta) = c$. Then

$$E[m(y)|y(s)] = N^{-1}[\sum_{i\in s} y_i + (\sum_{i\in s} y_i\sigma_i^{-2}/\sum_{i\in s} a_i\sigma_i^{-2}) \sum_{j\in \bar{s}} a_j]. \quad (1)$$

Proof. First note that solving $E[(d\log f(y_i|\theta)/d\theta)|\theta] = 0$ (the first Bartlett identity), one gets $E(y_i|\theta_i) = a_i\psi'(\theta)$. The posterior

$$\pi(\theta|y(s)) \propto \exp[\theta \sum_{i\in s} y_i\sigma_i^{-2} - \psi(\theta) \sum_{i\in s} a_i\sigma_i^{-2}].$$

Now, by the Bayesian analog of the first Bartlett identity, namely,

$$E[(d\log\pi(\theta|y(s))/d\theta)|y(s)] = 0,$$

one gets

$$E[\psi'(\theta)|y(s)] = \sum_{i\in s} y_i\sigma_i^{-2}/\sum_{i\in s} a_i\sigma_i^{-2}. \quad (2)$$

Next, observe that

$$E[m(y)|y(s)] = N^{-1}[\sum_{i\in s} y_i + \sum_{j\in \bar{s}} E(y_j|y(s))]. \quad (3)$$

Now, noting that for a given $j \in \bar{s}$, $E[y_j|y(s)] = EE[\{y_j|\theta, y(s)\}|y(s)] = a_jE[\psi'(\theta)|y(s)]$, one gets (1) from (2) and (3).

As mentioned earlier, some of the well-known design-based estimators can be derived as special cases of the above result.

Example 1. Let $a_i = \pi_i$ and $\sigma_i^2 = \pi_i/(1-\pi_i)$, where we may recall that π_i is the selection probability of the ith unit and $\sum_{i=1}^N \pi_i = n$. In this case from Theorem 1, $E[m(y)|y(s)]$ simplifies to

$$E[m(y)|y(s)] = N^{-1}[\sum_{i\in s} y_i + \{\sum_{i\in s}((1-\pi_i)/\pi_i)y_i / \sum_{i\in s}(1-\pi_i)\} \sum_{j\in \bar{s}} \pi_j. \quad (4)$$

Since $\sum_{j\in\bar{s}} \pi_j = \sum_{i=1}^N \pi_i - \sum_{i\in s}\pi_i = n - \sum_{i\in s}\pi_i = \sum_{i\in s}(1-\pi_i)$, from (4), one gets $E[m(y)|y(s)] = N^{-1}\sum_{i\in s} y_i/\pi_i$, which is the celebrated Horvitz-Thompson estimator.

Remark 1. Little (2004) gave an asymptotic model-based interpretation of the Horvitz-Thompson estimator. Ghosh and Sinha (1989) provided an exact model-based justification, but restricted only to the normal model. The result is established now under broader generality, and the present result is believed to be new.

Example 2. Suppose now $a_i = x_i$ and $\sigma_i^2 = 1$. Then, from Theorem 1

$$E[m(y)|y(s)] = N^{-1}[\sum_{i\in s} y_i + (\sum_{i\in s} y_i / \sum_{i\in s} x_i) \sum_{j\in\bar{s}} x_j] =$$

$$= (\sum_{i\in s} y_i / \sum_{i\in s} x_i) \sum_{i=1}^N x_i/N,$$

the well-known ratio estimator.

Example 3. Suppose now $a_i = \sigma_i^2 = x_i$. Then one gets $E[m(y)|y(s)] = N^{-1}\sum_{i\in s} y_i + n^{-1}(\sum_{i\in s} y_i/x_i)\sum_{j\in\bar{s}} x_j$, an example originally considered in Royall (1970). Basu (1971) gave a very intuitive justification of this estimator.

Remark 2. With the available auxiliary information x_i, various choices $a_i = h(x_i)$ and $\sigma_i^2 = v(x_i)$ will produce different model-based estimators of the finite population mean which could potentially be useful also in design-based analysis.

Remark 3. Although phrased in a Bayesian framework, one can think of an alternate model-based interpretation of the result of Theorem 1. To see this, we may note that $E(y_i|\theta) = a_i\psi'(\theta)$ and $V(y_i|\theta) = a_i\sigma_i^{-2}\psi''(\theta)$. The latter follows from the second Bartlett identity $E[\{dlogf(y_i|\theta)/d\theta)\}^2|\theta] = E[(-d^2logf(y_i|\theta)/d\theta^2)|\theta]$.

Then, under the standard quasi-likelihood approach, one obtains the unbiased estimating equation $\sum_{i\in s}\{y_i - E(y_i|\theta_i)\}/V(y_i|\theta_i) = 0$, which is equivalent to $E[\sum_{i\in s}\{y_i - a_i\psi'(\theta)\}^2/(a_i\sigma_i^{-2}) = 0$. This leads to the same estimator of $\psi'(\theta)$ as given in Theorem 1.

Remark 4. The special case $a_i = 1$ for all i and $\sigma_i^2 = \sigma^2$ for all i is of interest. In this case $E[m(y)|y(s)]$ simplifies to $N^{-1}[\sum_{i\in s} y_i + n^{-1}(N - n)\sum_{i\in s} y_i] = n^{-1}\sum_{i\in s} y_i$, the standard design-based estimator of the finite population mean under simple random sampling without replacement. We will revisit this point again in Section 3.

3. Some asymptotic results

The exact results of the previous section require a flat prior for θ. However, design consistency of model-based estimators in an asymptotic sense can often be justified for a wide class of priors. We will show in this section how mathematical limits of certain Bayes estimators result in standard design-based estimators. We use the term "mathematical limit" in the sense of Lahiri and Mukherjee (2007), where the limiting operation is performed in the sense of ordinary calculus, keeping the observations fixed. The latter came to the conclusion that as the sample size goes to infinity, mathematical limits of subjective Bayes estimators of the finite population mean based on the one parameter exponential superpopulation family do not converge in general to the Horvitz-Thompson estimator. They also proposed an adjustment to their subjective Bayes estimators to achieve design consistency. What we show is that a slightly modified version of the one parameter exponential superpopulation family as considered in (1) can indeed lead to design consistent Bayes estimators for a wide class of priors without requiring any adjustment. We will also point out why Lahiri and Mukherjee (2007) needed an adjustment to their Bayes estimators to achieve design consistency.

To this end, we first prove the following theorem. We denote the expression given in the right hand side of (1) as r.

Theorem 2. Consider the one-parameter exponential family as given in Theorem 1. Consider priors $\pi(\theta)$ of θ which are differentiable in θ and satisfy

$$N^{-1}E[\pi'(\theta)/\pi(\theta)|y(s)]\sum_{j\in\bar{s}} a_j / \sum_{i\in s} a_i\sigma_i^{-2} \to 0 \text{ as } n \to \infty. \qquad (5)$$

Then, $E[m(y)|y(s)] - r \to 0$ as $n \to \infty$.

Proof. Once again we use the fact that $E[d\log\pi(\theta|y(s))/d\theta|y(s)] = 0$. In the present set up, this fact leads to the equation

$$\sum_{i\in s} y_i\sigma_i^{-2} - E[\psi'(\theta)|y(s)]\sum_{i\in s} a_i\sigma_i^{-2} + E[\pi'(\theta)/\pi(\theta)|y(s)] = 0,$$

solving which we get

$$E[\psi'(\theta)|y(s)] = \sum_{i\in s} y_i\sigma_i^{-2}/\sum_{i\in s} a_i\sigma_i^{-2} + E[\pi'(\theta)/\pi(\theta)|y(s)]/\sum_{i\in s} a_i\sigma_i^{-2}.$$

The result follows now from Theorem 1 and (5).

While (5) seems somewhat artificial and complicated, it does lead to some simple readily verifiable conditions in some special cases. We begin with the situation where $a_i = \pi_i$ and $\sigma_i^2 = \pi_i/(1-\pi_i)$, $i = 1, \ldots, N$ and $\sum_{i=1}^N \pi_i = n$. Then, r simplifies to the Horvitz-Thompson estimator. In this scenario, $\sum_{j\in\bar{s}} a_j/\sum_{i\in s} a_i\sigma_i^{-2} = 1$ so that (5) simplifies to the condition $N^{-1}E[\pi'(\theta)/\pi(\theta)|y(s)] \to 0$ as $n \to \infty$. For instance, for a prior $\pi(\theta)$ for which $|\pi'(\theta)/\pi(\theta)|$ is bounded uniformly in θ, this condition holds trivially since $n \to \infty$ implies $N \to \infty$.

The boundedness of $|\pi'(\theta)/\pi(\theta)|$ is not all that restrictive either. It holds for many heavy-tailed priors. For example, if $\theta|\sigma^2 \sim N(0, \sigma^2)$ and $\sigma^2 \sim$ inverse gamma$(\beta/2, \alpha/2)$, that is $\pi(\sigma^2) \propto (\sigma^2)^{-\beta/2-1}\exp(-\alpha/2)$, $\alpha > 0$ and $\beta > 0$, then θ has the marginal prior $\pi(\theta) \propto (\theta^2+\alpha)^{-(\beta+1)/2}$. This is immediately recognized as a t-density. In this case $|\pi'(\theta)/\pi(\theta)| = (\beta+1)|\theta|/(\theta^2+\alpha) \le (1/2)(\beta+1)/\alpha^{1/2}$ uniformly in θ. A second example is the logistic prior $\pi(\theta) = \exp(\theta)/[1+\exp(\theta)]^2$ which leads to $\pi'(\theta)/\pi(\theta) = [1-\exp(\theta)]/[1+\exp(\theta)]$. Then, $|\pi'(\theta)/\pi(\theta)| \le 1$ uniformly in θ.

A similar phenomenon occurs for ratio estimators. Here $a_i = x_i$ and $\sigma_i^2 = 1$ for all i. Then, $\sum_{j\in\bar{s}} a_j/\sum_{i\in s} a_i\sigma_i^{-2} = \sum_{j\in\bar{s}} x_j/\sum_{i\in s} x_i$. If now $C_1 \le x_i \le C_2$ for all $i = 1, \ldots, N$, then $\sum_{j\in\bar{s}} x_j/\sum_{i\in s} x_i \le (N-n)C_2/(nC_1)$. Then, (5) reduces to the simple condition $n^{-1}E[\pi'(\theta)/\pi(\theta)|y(s)] \to 0$ as $n \to \infty$. Again, for the t and logistic priors considered in the previous paragraph, (5) trivially holds.

A standard normal prior for θ leads to $\pi'(\theta)/\pi(\theta) = \theta$ and in this case (5) may not hold true because of the unboundedness of θ. Accordingly, the subjective Bayes estimator of Ericson (1969) may not achieve design robustness except under very special circumstances, for example, under simple random sampling with replacement. But even under the normal superpopulation

model, which is a special case of (1), a heavy-tailed prior such as the t or logistic can lead to design consistency of the Bayes estimator of the finite population mean.

It is important to point out why Lahiri and Mukherjee (2007) needed an adjustment to their subjective Bayes estimator. They introduced a new random variable T_n to this end. If one examines carefully the pdf of T_n given in their (2.4), then it is clear that it is equivalent to the posterior pdf given in our Section 2 with the flat prior $\pi(\theta) = c$, $a_i = 1$ and $\sigma_i^2 = \sigma^2$ (ϕ in their notation) for all i. As pointed out in our Remark 4, one then has $E[\psi'(T_n)] = \bar{y}_s$. This is an exact relation. Obviously, $\bar{y}_s$ is not necessarily equal to a weighted estimator, say, $\bar{y}_w$ unless the design is self weighted, which was noted also by Lahiri and Mukherjee. On the other hand, as evidenced in our Section 2, what one needs is a suitable choice of the a_i and σ_i^2 to agree with the Horvitz-Thompson or the ratio estimator.

4. Summary and conclusion

The paper derives Bayes estimators under the one-parameter exponential family superpopulation model, and demonstrates design consistency of these estimators under certain conditions. It is needless to say that not all model-based estimators can achieve design consistency. Many authors have shown design consistency of certain model-based estimators in a probabilistic sense. What we have shown here is that often it is possible to achieve design consistency of Bayes estimators in a pure mathematical way, holding the observations as fixed numbers. A possible extension of our work is to consider the multiparameter situation, and show design consistency of model-based estimators, for example, in the regression model, by considering mathematical rather than probabilistic limits.

Acknowledgement

This research was partially supported by an NSF Grant SES-1026165.

REFERENCES

(1) BASU, D. (1971). An essay on the logical foundations of survey sampling, part 1. In *Foundations of Statistical Inference*. Eds. V.P. Godambe and D.A. Sprott. Holt, Rinehart and Winston, Toronto, Canada, 203-242.

(2) SARNDAL, C-E, SWENSSON, B. and WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

(3) ERICSON, W.A. (1969). Subjective Bayesian models in sampling finite populations (with discussion). *J. Roy. Statist. Soc. B*, 31, 195-233.

(4) GHOSH, M. and SINHA, B.K. (1989). On the consistency between model and design based estimators in survey sampling. *Comm. Statist.*, 20, 689-702.

(5) HAJEK, J. (1971). Discussion of 'an essay on the logical foundations of survey sampling, part one' by D. Basu. In *Foundations of Statistical Inference*. Eds. V.P. Godambe and D.A. Sprott. Holt, Rinehart and Winston, Toronto, Canada, p 236.

(6) LAHIRI, P. and MUKHERJEE, K. (2007). On the design consistency property of hierarchical Bayes estimators in finite population sampling. *Ann. Statist.*, 35, 724-737.

(7) LITTLE, R.J.A. (2004). To model or not to model? Comparing modes of inference for finite population sampling. *J. Amer. statist. Assoc.*, 99, 546-556.

(8) PRASAD, N.G.N. and Rao, J.N.K. (1999). On robust small area estimation using a single random effects model. *Survey Methodology*, 25, 67-72.

(9) ROYALL, R.M. (1970). On finite population sampling theory under certain regression models. *Biometrika*, 57, 377-387.

(10) SARNDAL, C-E. (1984). Design-consistent versus model-dependent estimation in small domains. *J. Amer. statist. Assoc.*, 79, 624-631.

STATISTICS IN TRANSITION-new series, Summer 2012
Vol. 13, No. 2, pp. 243—260

APPLICATION OF ROTATION METHODS IN SAMPLE SURVEYS IN POLAND[1]

Jan Kordos[2]

ABSTRACT

The author reviews theory and application of rotation methods in sample surveys in Poland. He begins with reviewing designs of the surveys across time, depending on different objectives, focusing on partial rotation of sub-samples, and considers estimation problems and data quality issues generally. Next, he refers to some articles and books about surveys published over time, starting with Wilks (1940), Patterson (1950), Eckler (1955), Woodruff (1963), Rao and Graham (1964), Bailar (1975), Duncan and Kalton (1987) and Kalton and Citro (1993). He mentions also early Polish papers on rotation methods (Kordos (1966, 1967, 1971, 1982); Lednicki, 1982; Szarkowski and Witkowski, 1994), and concentrates on Polish household surveys, mainly Household Budget Survey (HBS), Labour Force Survey (LFS) and EU Statistics on Living Conditions and Income (EU-SILC). Special attention is devoted to last research on rotation sampling done by Polish sampling statisticians: Ciepiela et al. (2012), Kordos (2002), Kowalczyk (2002, 2003, 2004), Kowalski (2006, 2009), Kowalski and Wesołowski (2012) and Wesołowski (2010). Concluding remarks are given at the end.

Key Words: Rotation sampling; sampling on successive occasions; survey across time; panel survey; data quality; sample survey.

1. Introduction

At the beginning the author examines the interplay between sample survey theory and practice in Poland over approximately the past 50 years. He begins with the Neymans's (1934) classic landmark paper which laid theoretical foundations to the probability sampling (or design-based) approach to inference from survey samples. Main ideas of that paper were first published in Polish in

[1] This is an extended and updated version of the paper connected with the author's paper entitled "*Interplay Between Sample Survey Theory and Practice in Poland"* presented at the Congress of Polish Statistics held in Poznań, Poland, 18-20 April 2012. The author concentrates here on application of rotation methods in sample surveys in Poland.

[2] Warsaw School of Economics/Warsaw Management Academy.

1933 (Neyman, 1933) and had a significant impact on sampling practice in Poland before and after World War II. Sample surveys conducted in 1950s and 1960s were consulted with J. Neyman during his visits in Poland in 1950 and 1958 (Fisz, 1950; Zasepa, 1958).

Some practical problems encountered in the design and analysis of sample surveys were partly solved by the Mathematical Commission of the CSO which was established in 1949 as an advisory and opinion-making body to the CSO President in the field of sample surveys. The Commission concentrated specialists in the sampling methods both from the CSO and research centres in the country (Kordos, 1975, 2012). The Commission had a significant impact on sampling practice in Poland and was active until 1993.

Polish sampling statisticians had problems with sampling in time (household budget surveys, agricultural sample surveys, living conditions surveys) and some research works were devoted to rotation methods and panel surveys (Kordos, 1967, 1985). They had problems with size of sample determination, sample allocation and estimation methods (Bracha, 1996; Greń, 1964, 1966, 1970; Kordos, 1985, 1988a, 2002; Zasępa, 1962, 1970, 1993) and with data quality (Kordos, 1988bc). There were problems with calibration estimation that ensures consistency with user specific totals of auxiliary variables, unequal probability sampling without replacement, analysis of survey data (Kordos, Zięba-Pietrzak, 2010).

Very important and difficult problems for sampling practice are data for small areas. That is why Polish statisticians are interested in the issues related to estimation methods for small areas (Kalton et al., 1993; Domański and Pruska, 1996; Golata, 2012; Paradysz, 1998).

The Central Statistical Office of Poland (GUS) was deeply involved in rotation methods after consultation with Prof. Jerzy Neyman in 1958 (Zasepa, 1958), who criticized current application of sampling method in the panel household budget survey in Poland. Sampling statisticians in GUS, in cooperation with the Mathematical Commission of GUS (Kordos, 1975, 2012) have undertaken research in application of rotation methods in sampling surveys in Poland (GUS, 1972, 1979, 1987; Kordos, 1971, 1982, 1985, 1996; Lednicki, 1982).

Before presenting application of rotation methods in the sample surveys in Poland, I begin with reminding different objectives for surveys over time, appropriate survey designs for these surveys, estimation problems and data quality issues. General overview of issues in repeated surveys is also presented. Next, the involvement of the Central Statistical Office of Poland in theory and application of rotation methods in sampling surveys is presented.

2. Rotation sampling

The name „*rotation sampling*" (suggested by Wilks, 1940) refers to the process of eliminating some of the old elements from the sample and adding new elements to the sample each time a new sample is drawn. This method of

sampling is also called *sampling on successive occasions with partial replacement of units* (Patterson, 1950; Yates, 1949) and *sampling for a time series* (Hansen, Hurwith and Madow, 1953). *Double sampling* can be regarded mathematically as rotation sampling involving a present sample and one overlapping an earlier sample.

In general, we have a population of units (segments) which is to be sampled for T consecutive periods (months, quarters, years). In any proposed sampling design, the units to be sampled can change from period to period but not at time points within the period. In addition, there is a positive correlation between the responses from a unit in consecutive time periods which can be utilized to reduce the standard errors of the estimators of the end of period means. The problem is to determine a T period sampling scheme which is optimal in some sense. Rotation designs are a natural starting point in the search for a solution of this problem.

In many repeated surveys, samples are collected routinely (e.g., monthly or quarterly) from a finite population, and the characteristics of the sampling units change over time. Many of these surveys also provide information on relevant auxiliary variables. Thus, it is possible to improve on the current direct survey estimators of the finite population parameters by using data from earlier surveys, in conjunction with the auxiliary variables. For example, consider the Household Budget Survey (HBS), the Labour Force Survey (LFS) and the Statistics of Living Conditions Survey (EU-SILC).

A similar situation may be cited in small-area estimation problems, which have received considerable attention in recent years (Rao, 2003). Estimate of a small area (finite population) characteristic may be improved by utilizing data from related areas and the auxiliary variables.

3. Objectives for surveys over time

Changes in population characteristics and composition over time lead to variety of objectives for surveys across time. These objectives include the following (Kalton and Citro, 1993):

a. Estimating population parameters at distinct time points.
b. Estimating average values of population parameters.
c. Estimating net change.
d. Estimating gross and others components of individual change.
e. Averaging data for individuals over time.
f. Collecting data on events occurring in specified time period.
g. Cumulating samples over time.
h. Keeping contact with members of a rare population identified at a point of time.

To estimate these parameters for different objectives from sample surveys there are various problems connected with sample design, size of sample,

allocation of sampling units in space and time, collection of required data from these units, quality of obtained information, etc. Some of these problems may be partly solved by applying rotation sampling, i.e. changing sub-samples of units at different periods of time.

4. Survey designs over time

A number of survey designs have been developed to provide data to the address of the above mentioned objectives. These designs are (Duncan and Kalton, 1987; Kalton and Citro, 1993):

1) ***Repeated survey:*** a series of separate cross-sectional surveys conducted at different time points.
2) ***Panel survey***: collecting of the survey data for the same sample elements at different points of time.
3) ***Repeated panel survey***: is made up of a series of panel surveys, each of fixed duration:
   i. ***with no overlap*** (one panel may start only as the previous one ends),
   ii. ***with overlap*** (with two or more panels covering the same period of time).
4) ***Rotating panel survey:*** is equivalent to a repeated panel survey with overlap. Both limit the length of a panel and have two or more panels in the field at the same time. Rotating panel surveys are widely used to provide a series of cross-sectional estimates and estimates of net change, whereas repeated panel surveys with overlaps also have a major focus on longitudinal measurement. In consequence, repeated panel surveys tend to have longer duration and have fewer panels in operation at any given time than rotating panel surveys.
5) ***Split panel surveys***: is a combination of a panel survey and repeated survey or rotating panel survey.
6) ***Repeated survey with overlap***: like a repeated survey and overlapping survey, it is a series of cross-sectional surveys conducted at different time points, and is designed to provide overlaps.

The choice of design in a particular case depends on the desired objectives to be satisfied. Some designs are better than others for some objectives but less suitable for other objectives.

5. Reasons for sample rotation

Response burden among the households or other units can be reduced by periodic sample rotation. However, rotation of units increases the cost of the survey because of additional sample maintenance, possible additional training of interviewers, extra costs of initially collecting baseline information, and difficulties in training new units to provide data. Partial rotation of sampled units

at some fixed rate is undertaken as a compromise between total rotation, i.e. 100% of units, that is very expensive and gives poor estimates of change and no rotation at all (panel) that would result in an unacceptable distribution of response burden. The rotation schemes keep a unit in the sample for a given period after which the unit becomes ineligible for re-selection by the same survey for a minimum period.

One can think of a rotation design as a compromise between a complete sample overlap and taking independent samples. Each extreme has advantages and disadvantages. By using a rotation design, one hopes to realize some of the variance reduction of the complete sample overlap, while reducing its excess burden.

However, it should be recognized that retention of the same units for a long series of surveys presents practical and theoretical difficulties since the frame is changing; repeated surveys on the same units may also lead to atypical changes in these units. The continuing operation of a survey agency on successive surveys of the same type also usually leads to a gradual improvement in the quality of the information.

6. Estimation problems

An objective of a survey over time is usually to estimate more than one population parameters. Very often several objectives given in section 3 are to be achieved. Two of many important choices that must be made in a survey are (Kordos, 2002):

i. The choice of sampling design and a sample selection scheme to implement the design.
ii. The choice of a formula (an estimator) by which to calculate an estimate of a given parameter of interest.

The two choices are not independent of each other. For example, the choice of the estimator will usually depend on the choice of sampling design. It is necessary to use an appropriate strategy.

A strategy is the combination of a sampling design and an estimator. For a given parameter, the general aim is to find the best possible strategy, that is one that estimates the parameter as accurately as possible.

In the above sections we note five objectives and six designs. Most periodic studies have several purposes and thus we should face – not necessarily solve – difficult problems of multipurpose designs. Actually, objectives from (a) to (c) can be met with any of the six listed designs, but in some they increase in the variances or in costs. But estimates with gross and individual change need panels, and objective (e) need some changes. Often reasonable compromises become possible – to the degree that purposes can be defined. The chief variation in these designs concerns the amount (and kind) of overlaps between periods (Kish, 1987, pp. 159 – 169).

Table 1. Purposes and Designs for Periodic Samples

Purposes	Designs	Rotation Scheme
A. Current levels	A. Partial overlaps 0<P,1	abc–cde–efg
B. Cumulating	B. Non-overlaps P = 0	aaa–bbb–ccc
C. Net changes (means)	C. Complete overlaps P = 1	aaa–aaa–aaa
D. Gross changes (individuals)	D. Panels	Same elements
E. Multipurpose, time series	E. Combination F. Master frames	

Source: Kish, 1987, p. 160.

The rotation scheme of complete overlaps shows, with aaa–aaa, that the periods have all common parts; the non-overlaps with aaa–bbb shows none, and the partial overlaps abc–cde–efg shows c and e as one-third overlaps between succeeding periods only.

Here, we concentrate on the effects of varying proportions P in divers designs on different purposes; in complete overlaps P = 1, in non-overlaps P = 0, and in partial overlaps 0<P<1. The purposes are discussed in terms of variances for estimated means, because means (and percentages, rates, proportions) are both the most used and the simplest estimates to be treated. Effects on other estimates will not be entirely different but they are too numerous, diverse and difficult to be explored here.

Effects on the variances of means from different portions P can be treated clearly in this brief section. Other questions of biases, of feasibility, of costs are often more important, but also more difficult. They are treated in all sampling books. We assume here for simplicity that the period samples are of the same size or of the same sampling fraction; but changes in sizes, fractions and designs are possible. We start below with current levels and partial overlaps.

6.1. Current levels and partial overlaps

Variances of current estimates are the same for complete overlaps P = 1 and non-overlaps P = 0; they can be expressed briefly for means as $Deft^2=S^2/v$, where $Deft^2$ is the effect of the sample design on either the element variance S^2 or the sample size n (Kish, 1987, p.162-163).

That simple formula also holds for *simple* means from partial overlaps (0<P<1). But statistics based on them can utilize the overlap P for a reduction of the variance with a complex mean: with help of the correlation R^2 between surveys within the same overlap P, the portion (1-P) of the *preceding* sample is combined with the current mean to improve it. The variances are reduced by the factor:

$$\frac{1-(1-P)R^2}{1-(1-P)^2R^2} \qquad \text{(Cochran, 1977, secs.12.11-12.12)}$$

The actual gains unfortunately tend to be modest in most practical situations, the maximum reduction in variance, utilizing optimal proportions P and optimal weights, is the ratio

$[1+((1-R^2)^{0,5}0]/2$. Reductions increase to about 33% only for very high R^2 values, seldom seen in practice; for R = 0.9, for example, $[1+((1-R^2)^{0,5}0]/2=0.72$. This ratio is obtained either with the optimal P = 0.30 or with P = 1/3. For R = 0.6 that ratio becomes 0.9, only 10 percent reduction of the variance. It is worth stressing that in a long series the complex mean from the preceding sample can already benefit from reductions from its predecessors, and that using a longer series provides further slight reductions. Fortunately, for other purposes of repeated surveys statistical theory is more productive as well as simpler.

6.2. Net changes of Means and Overlaps

Net change refers to the difference $d=\bar{x}_1-\bar{x}_2$ of means between two periods, whereas *gross change* deals with the total changes of individuals, some of which remain hidden, because they cancel, in the net change of means. Measuring net changes are common and important aims of surveys, and they are also related to other issues of the data. Perhaps the most forms are differences in dichotomies, denoted by proportions $d = p_1 - p_2$, and in similar rates and ratios.

The variance of $(\bar{x}-\bar{y})$ can be greatly reduced when the pairs of variables have high positive correlation R in overlapping samples. We will discuss here several cases shortly.

1. The variance of mean differences are reduced by factors (1 – R) in complete overlaps; this is the extreme (with P = 1) of the factor (1 –PR) that may be obtained from the partial overlaps. Hence, for minimizing variance Var $(\bar{x}-\bar{y})$ complete overlaps would be the best. But partial overlaps are used in practice: for reasons of feasibility, to reduce burdens, fatigue, and biases of respondents; and reduce variances of other statistics in multipurpose designs.
2. We may obtain almost the full reductions of complete overlaps even from partial overlaps by using improved estimators of the differences. These estimators are useful when circumstances may present complex overlaps but still permit partial overlaps. In those estimators the overlaps portion P gets larger weights than the non-overlaps portion 1– P = Q, by the factor 1/(1–R), because elements in the overlap contribute much less to the variance. This improved estimator of the difference is (Kish, 1965, table 12.4. III):

$$\hat{D}(\bar{y}-\bar{x})=[P(\bar{y}-\bar{x})_p+Q(1-R)(\bar{y}-\bar{x})]/(1-QR)$$

Its variance may be expressed, for two *srs* samples of size n, as:

$$Var([\hat{D}(\bar{y}-\bar{x})] = \frac{(1-R)S^2}{(1-QR)n}$$

The factor (1- R)/(1-QR) approaches (1-R) for high values of R and for higher values of P, say p = 2/3. High value of R are common for stable characteristics that can be well measured.

7. Overview of Issues in Repeated Surveys

Patterson (1950), following the initial work by Jessen (1942), provided the theoretical foundations for design and estimation for repeated surveys, using generalized least squares procedures. For the Current Population Survey, Hansen et al. (1955) proposed a simpler estimator, called the K composite estimator. Gurney and Daly (1965) presented an improvement to the K composite estimator, called the AK composite estimator with two weighting factors A and K. Breau and Ernst (1983) compared alternative estimators to the K composite estimator for the CPS. Rao and Graham (1964) studied optimal replacement schemes for the K composite estimator. Eckler (1955) and Wolter (1979) studied two level rotation schemes such as the one used in the U.S. Retail Trade Survey. Yansaneh and Fuller (1998) studied optimal recursive estimation for repeated surveys. Fuller (1990) and Lent, Miller, Cantwell and Duff (1999) developed the method of composite weights for the CPS. The composite weights are obtained by raking the design weights to specified control totals that included population totals of auxiliary variables and K composite estimates for characteristics of interest. Using the composite weights, users can generate estimates from micro-data files for the current month without recourse to data from previous months.

Ciepiela et al. (2012) propose a dynamic version of the K-composite estimator (DK-composite estimator) without any restrictions on the rotation pattern. This estimator gives an alternative solution to quasi-optimal estimation under rotation sampling when it is allowed that units leave the sample for several occasions and then come back. Such situations happen frequently in real surveys and are not covered by the recursive optimal estimator introduced by Patterson (1955). However, the K-composite estimator suffers from certain disadvantages. It is designed for a stable situation in the sense that its Basic parameter is kept constant on all occasions. Additionally, it is restricted only to a certain family of rotation designs. The authors propose a dynamic version of the K-composite estimator (DK-composite estimator) without any restrictions on the rotation pattern. Mathematically, the algorithm they develop is much simpler than the one for the classical K-composite estimator with optimal weights. Moreover, it is precise in the sense that it does not use any approximate or asymptotic approach (opposed to the method used in Rao and Graham (1964) for computing optimal weights).

The above authors used the traditional design based approach, assuming the unknown totals on each occasion to be fixed parameters. Other authors (Scott, Smith and Jones 1977; Jones 1980; Binder and Dick 1989; Bell and Hillmer 1990; Tiller 1989 and Pfeffermann 1991) developed estimates for repeated surveys under the assumption that the underlying true values constitute a realization of a time series.

8. Data quality issues

Survey data quality issues, is a concept with many dimensions linked with each others. In theory, all dimensions of data quality are very important, but in practice, it is usually not possible to place high importance on all dimensions. Thus, with fixed financial resources, an emphasis on one dimension will result in a decrease in emphasis on another. More emphasis on accuracy can lead to less emphasis on timeliness and accessibility, or emphasis on timeliness may result in early/preliminary release data of significantly lower accuracy. Each dimension is important to an end user, but each user may differ in identifying the most important priorities for a data collection program.

An extensive literature exists and continues to grow on the topic of survey data quality. Special attention to data quality has been paid in last years by Eurostat (2007, 2009), and relating to household surveys (e.g. Bailar, 1975, 1979; Brackstone, 1999; CPS, 2002; GUS, 1972, 1979, 1987, 2009; Kordos, 1988ab; Lyberg et al. 1997).

For a sample survey over time, the following sources of errors are very important:

a. Non-response losses,
b. Time-in-sample, or conditioning effects,
c. Recall errors, including the seam effect, and other non-sampling errors.

In Polish household surveys, such as the HBS, the LFS, and the EU-SILC, only sampling errors and non-response rates are reported. Conditioning effects, recall errors, the seam effects and other non-sampling errors are neglected. Additional research in these fields is needed.

9. Application of rotation sampling in the sample surveys of the Central Statistical Office of Poland (GUS)

We started applying rotation sampling in statistical sample surveys in 1960s, after the consultation with Prof. Jerzy Neyman (Kordos, 2011; Zasepa, 1958). At the beginning we studied such articles as Wilks (1940), Jessen (1942), Patterson (1950), Eckler (1955), Woodruff (1963), and Rao and Graham (1964), and books as Yates (1949), Deming (1950) and Hansen et al. (1953).

In 1967 I published a review paper on rotation method in sampling surveys (Kordos, 1967). Later the following articles describing application of rotation sampling were published by GUS (1969, 1972, 1979. 1987), Kordos (1966, 1971, 1982, 1988ab, 2002), Szarkowski and Witkowski (1994) and Popiński (2006).

Theory for applying rotation method in the Polish sample surveys was rather simple. First, we applied cross-sectional surveys, and were criticized for missing seasonal effects, and data related to generalized for a given year.

The following surveys were carried out with application of rotation sampling:

1) Rotation method in morbidity survey (1967 – 1968).
2) Time use surveys (1967, 1976, 1984, 1996).
3) Experiments of HBS by rotation method (1968 – 1969, 1981).
4) Surveys of workers starting first job (1971).
5) Epilepsy survey in Warsaw ((1971).
6) Household Budget Survey (since 1982 +).
7) Labour Force Survey (since 1992 +).
8) EU Statistics on Income and Living Conditions Survey (UE-SILC) (since 2005 +).

10. Some theoretical research in rotation sampling in Poland in the last decade

Factors affecting the design of a sample over time include the key estimates to be produced, the type and level of analyses to be carried out, cost, data quality and reporting load. The interaction between sampling over time and features of the design, such as stratification and cluster sampling also needs to be decided. Duncan and Kalton (1987), Kalton and Citro (1993) and Steel (2004) give a general review of issues in the design and analysis of repeated surveys. Kasprzyk *et al*. (1989) cover many of the important issues associated with panel surveys.

Several Polish statisticians have undertaken some research in rotation sampling in last decade. I would like to mention in a synthetic way some of the results of Ciepiela et al. (2012), Kowalczyk (2002, 2003, 2004), Kowalski (2006, 2009), Kowalski and Wesołowski (2012), Popiński (2006) and Wesołowski (2010). Ciepiela et al. (2012) contributions have been presented above.

Kowalczyk (2003) gives the theory for the estimation of the population total on the current occasion under two-stage sampling design with a partial replacement of the second stage units. Under given rotation pattern she presents composite estimator of the population total, which exploits information from the previous occasion. She derives the variance of the estimator and compares it with the variance of the usual estimator. It has been shown that the composite estimator is better than the usual one applied both for a sample selected independently of the previous period and a sample selected according to the given rotation pattern.

The aim of Kowalski (2009) paper is to examine the setting of surveys repeated over time when the elements in the sample are rotated in a pre-designed

way. On each occasion the best linear unbiased estimator (BLUE) of the current population mean, built on all past responses is to be found. The most straightforward approach would be to compute the estimator as a solution of a least squares problem with linear restrictions. However, this method has certain drawbacks related to the fact that the size of the response data set increases over time. They follow a different approach based on finding linear recurrence relationships between optimal estimators obtained on successive occasions. Most of the original disadvantages are then corrected. In this context we present the solution to the BLUE estimation problem for some — sufficiently regular — classes of rotation patterns.

Kowalski and Wesołowski (2012) consider linear recurrence for BLUE estimators of the unknown population mean obtained on successive occasions. They work in the framework of sample rotation where evolution of the sample across time is pre-designed. It has been known since Patterson (1950) that the essential difficulties arise when it is allowed that units may return to the sample after being absent in the sample for several occasions. It means that contrary to the Patterson setting, holes in rotation patterns are allowed. These difficulties are to a large extent overcome in the present paper. They prove that under some assumptions, linear recurrence holds and its order is closely linked to the rotation scheme. Of special importance in this setting appear to be roots of certain polynomial conveniently expressed in terms of Chebyshev polynomials. An effective and easily implemental algorithm for calculation of the recurrence coefficients is given. It is illustrated through examples of rotation schemes which are used in concrete surveys.

Wesołowski (2010) paper is devoted to the Szarkowski's involvement in the Polish Labour Force Survey design, who observed that under the rotation pattern typical for the LFS the recursion for the optimal estimator of the mean on a given occasion has to use estimators and observations only from three last occasions. Since the fundamental work of Patterson (1950) it has been known that for rotation patterns with "holes" it is a difficult problem to determine the depth of such recursion formulas. Under special assumptions the problem has been settled only recently in Kowalski and Wesołowski (2012). In the present paper it is shown that these assumptions are always satisfied in the case of the Szarkowski rotation pattern 110011. Moreover, explicit formulas for the coefficients of recursion are derived. As the author has stressed "It took more than ten years to answer in affirmative Szarkowski's three steps conjecture. It is based on a general approach described in Kowalski and Wesołowski (2012 (KW in the sequel). Earlier the problem for rotation schemes with singleton holes was solved in Kowalski (2009) (particular cases of 1011 and 1101 rotation patterns were covered even earlier, in (2004)). Moreover, explicit formulas for the coefficients of the recursion are derived".

11. Concluding remarks

The aim of rotating panels is twofold: firstly, it allows increasing the precision of estimates of change between two different points in time and, secondly, producing flow estimates, thus allowing the calculation of important indicators for the analysis of dynamics characteristics. A further advantage is the possibility to make use of dependent interviewing to reduce non-response burden. However, rotating panels present the typical drawbacks of panels, although these problems are less critical in the light of the short panel duration. Possible drawbacks include panel attrition, panel conditioning and misreporting. Furthermore, because population evolves in time, the longer panels remain in the sample the more they diverge from the actual population's structure. Overlap between months or quarters may also cause some inefficiency in annual estimates. Besides, rotation can be the underlying cause of other problems, such as non-response and measurement inconsistencies between subsequent survey waves. Comparability of longitudinal data could also be of concern with different rotation schemes. Overall, however, the advantages of rotation patterns outweigh their disadvantages.

In this paper, I reminded objectives of surveys in time, sampling designs, estimation methods and data quality issues connected with such surveys. Next, general review of application of rotation methods in sample surveys in Poland, and some theoretical research of surveys over time were presented. However, more research is needed in the field of sampling design and data quality for surveys in time, taking into account sampling and non-sampling errors. In social surveys, such as HBS, LFS and EU-SILC collected data are biased for different reasons (e.g. non-response, measurement errors, response errors, etc.). In such cases, it is unreasonable to assess accuracy of results, using MSE or confidence intervals if additional efforts are not undertaken to improve accuracy of the collected data. For quality assessment, it seems reasonable to use only precision or relative standard error (CV). If a confidence interval must be used, then its interpretation should be changed, avoiding words “to cover the true value...”.

REFERENCES

BAILAR, B. (1975). The effects of rotation group bias on estimates from panel surveys. J. Amer. Statist. Assoc. 70, 23-30.

BELL, W. R., HILLMER, S. C., (1990). The time series approach to estimation for repeated surveys. *Survey Methodology*, 16, pp. 195–215.

BINDER, D. A., AND DICK, J. P. (1989). Modeling and estimation for repeated surveys. *Survey Methodology*, 15, pp. 29-45.

BRACHA, CZ. (1996). Theoretical Background of Sampling Methods (in Polish), PWN, Warszawa.

BRACKSTONE, G. (1999). Managing Data Quality in a Statistical Agency. *Survey Methodology*, 25, nr. 2, pp. 139-149.

BREAU, P., ERNST, L. (1983). Alternative estimators to the current composite estimator. *Proc. Sec. Survey Res. Meth., Amer. Statist. Assoc*., 397-402.

CIEPIELA, P., GNIADO, K. ,WESOŁOWSKI, J. and WOJTY, M (2012). Dynamic K-Composite Estimator for an arbitrary rotation scheme, *Statistics in Transition – new series*, Vol. 13, No. 1, s. 7-20.

COCHRAN, W. G. (1977). Sampling Techniques, third edition. John Wiley and Sons, New York.

Current Population Survey (2002). *Design and Methodology*, Technical Paper 63RV, Bureau of Labour Statistics, U.S. Census Bureau.

DEMING, W. E. (1950). *Some Theory of Sampling*, New York: Wiley.

DUNCAN, G. J., KALTON, G.(1987). Issues of Design and Analysis of Surveys Across Time, *International Statistical Review*, 55, pp. 97-117.

ECKLER, A. R. (1955). Rotation Sampling. *Annals of Mathematical Statistics,* vol. 26, pp. 664-685.

Eurostat (2007). *Handbook on Data Quality Assessment: Methods and Tools*.

Eurostat (2009*). Handbook on Quality Reports.*

FULLER, W. A. (1990). Analysis of repeated surveys, *Survey Methodology*, 16, 167-180.

FULLER, W. A. and RAO, J.N.K. (2001). A Regression Composite Estimator with Application to the Canadian Labour Force Survey, *Survey Methodology*, Vol. 27, No. 1, pp. 45-51.

GOŁATA E. (2004). *Problems of Estimate Unemployment for Small Domains in Poland*, "Statistics in Transition", Vol. 6, No. 5, s. 755-776.

GOŁATA, E. (2012). Data integration and small domain estimation in Poland – experiences and problems. *Statistics in Transition – new series*, vol. 13, No. 1, pp. 107-142.

GURNEY, M., DALY, J. F. (1965). A multivariate approach to estimation in periodic sample surveys. Proc. Amer. Statist. Assoc., Sect. Soc. Statist., 242-257.

GUS (1969). *Application of Mathematical Methods in Statistics (in Polish*), "*Biblioteka Wiadomości Statystycznych",* vol. 7, Warszawa.

GUS (1972). *Experimental Research of Household Budget Survey by Rotation Method* (in Polish), "*Biblioteka Wiadomości Statystycznych"* vol. 18, Warszawa.

GUS (1979). *Methodology of Sampling Surveys in GUS* (in Polish) – Works of Mathematical Commission, GUS, Warszawa.

GUS (1987)**.** *Application of Sampling Method in Statistical Investigations of GUS* (1981-1986) (in Polish). "Z prac Zakładu Badań Statystyczno-Ekonomicznych ", GUS, z. 166, Warszawa.

GUS (2009). *Incomes and Living Conditions of the Population in Poland* (report from the EU-SILC survey of 2007 and 2008), Statistical Information and Elaborations, Warsaw.

FISZ, M. (1950). Consultation with Prof. Neyman and Conclusions (in Polish), "*Studia i Prace Statystyczne*", nr 3-4.).

HANSEN, M. H., HURWITZ, W. N., MADOW, W. G. (1953). *Sample Survey Methods and Theory*, Vol. 1 i 2, New York.

HANSEN, M.H., W.H. HURWITZ, H. NISSELSON, and J. STEINBERG (1955). The redesign of the Census Current Population Survey. *J. Amer. Star. Assoc*. 50, 701-709.

JESSEN, R. J. (1942). Statistical investigation of a sample survey for obtaining farm facts. Iowa Ag. Exper. Station Research Bull. 304, 54-59.

JONES, R. G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society*, B, 42, 221-226.

KALTON, G. and CITRO, C. F. (1993). Panel surveys: adding the fourth dimension. *Survey Methodology*, vol. 19, pp. 205-215.

KALTON, G., KORDOS, J. and PLATEK, R. (1993). *Small Area Statistics and Survey Designs,* Vol. I: Invited Papers; Vol. II: Contributed Papers and Panel Discussion**.**GUS, Warsaw.

KASPRZYK, D., DUNCAN, G., KALTON, G., SINGH, M. P., (Eds) (1989). *Panel Surveys*, J. Wiley, New York.

KISH, L. (1965). *Survey Sampling*, New York 1965.

KISH, L. (1987). *Statistical Design for Research*, New York.

KISH, L. (1998). Space/time variations and rolling samples. *Journal of Official Statistics*, 14, 31-46.

KORDOS, J. (1966). Application of Rotation Method in Sample Survey of Morbidity and Incidence Rates in Poland (in Polish). *Przegląd Statystyczny*, No. 4, 1966, pp. 341-352.

KORDOS, J. (1967). Rotation Method in Sample Surveys (in Polish), *Przegląd Statystyczny*, No. 4, 1967, pp. 373-394.

KORDOS, J. (1971). Experimental Household Budget Survey by Rotation Method (in Polish), *Przegląd Statystyczny*, No.3-4, 1971, pp. 227-246.

KORDOS, J. (1975). 25 Years of Activity of the GUS Mathematical Commission (in Polish), *Przegląd Statystyczny*, No. 1, pp. 171-173.

KORDOS, J. (1982). Rotation Method in Household Budget Surveys in Poland (in Polish), "*Wiadomości Statystyczne*", No. 9, pp. 1-6.

KORDOS, J. (1985). Towards an Integrated System of Household Surveys in Poland. „*Bulletin of the International Statistical Institute",* (invited paper), vol. 51, Amsterdam 1985, pp. 1.3. 1-17.

KORDOS, J. (1988a). Methods of Controlling Quality in Household Budget Survey in Poland. *Seminar on Statistical Methodology. Conference of Europeans Statisticians*. Geneva, 1-4 February 1988., pp. 1-14.

KORDOS, J. (1988b). *Quality of Statistical Data* (in Polish), PWE, Warsaw, 244 pages.

KORDOS, J. (1996). Forty Years of the Household Budget Surveys in Poland, *Statistics in Transition*, Vol. 2, No. 7, pp. 1119-1138.

KORDOS, J. (2002). Estimation Problems and Data Quality in Rotation Samples*,* In*: Rotierende Stichproben - Datenkumulation und Datenqualität, Spektrum Bendesstatistik,* Band 21, Statistisches Bundesamt, 2002, pp. 57-73.

KORDOS, J. (2012a). Activities of the Mathematical Commission of GUS in 1950–1993, (in Polish), *Wiadomości Statystyczne,* No. 9.

KORDOS, J. (2012b). Interplay between sample survey theory and practice in Poland (in Polish), *Przegląd Statystyczny* (in preparing).

KORDOS, J. and ZAGÓRSKI, K. (1971). Rotation Method in Sample Survey of Persons Taking First Job (in Polish*), Wiadomości Statystyczne*, No. 9, 1971, pp. 11-13.

KORDOS, J., ZIĘBA-PIETRZAK, A. (2010). Development of Standard Error Estimation Methods in Complex Household Sample Surveys in Poland, , *Statistics in Transition – new series*, Vol. 11, No.2, pp. 231-265.

KOWALCZYK, B. (2002). Sampling in Time (in Polish). Ph.D. Thesis, Warsaw, School of Economics, Warsaw, Poland.

KOWALCZYK, B. (2003). Estimation of the population total on the current occasion under second stage unit rotation pattern, *Statistics in Transition,* December 2003, Vol. 6, No. 4, pp. 503-513.

KOWALCZYK, B. (2004). Number Index Estimation Using Auxiliary Information in Repeated Rotating Surveys,: *Sampling Designs for Environmental, Economic and Social Surveys: Theoretical and Practical Perspectives*, Universita degli Studi di Siena, 23-24 September 2004, Italy.

KOWALSKI, J. (2006). Rotation in sampling patterns, in review for *Journal of Statistical Planning and Inference,* 2006.

KOWALSKI, J. (2009). Optimal Estimation in Rotation Pattern, *J. Statist. Plann. Infer. 139 (4), pp. 2429-2436.*

KOWALSKI, J. and WESOŁOWSKI, J. (2012). Recurrence optimal estimators for rotation cascade patterns with holes (in preparing).

LEDNICKI, B. (1982). Sampling Design and Estimation Method in Household Budget Rotation Survey (in Polish). "*Wiadomości Statystyczne*", No. 9.

LENT, J., MILLER, S., CANTWELL, P. (1994). Composite weights for the Current Population Survey. *Proc. Sec. Survey Res. Meth., Amer. Statist. Assoc*., 867-872.

LYBERG, L., BIEMER, P, COLLINS, M., DELEEUW, E., DIPPO, C., SCHWARZ, N., TREWIN, D. (eds, 1997). *Survey Measurement and Process Quality*. New York.

NEYMAN, J. (1933). An Outline of the Theory and Practice of Representative Method Applied in Social Research (in Polish), *Instytut Spraw Społecznych*, Warszawa.

NEYMAN, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, vol. 97, s. 558-625.

PARADYSZ, J. (1998). Small Area Statistics in Poland - First Experiences and Application Possibilities, *Statistics in Transition,* Vol. 3, Number 5, pp. 1003—1015.

PATTERSON, H. D. (1950). Sampling on successive occasions with partial replacements of units. *Journal of the Royal Statistical Society*, ser. B, vol. 12, s.241-255.

PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, 9, 163-175.

POPIŃSKI, W., (2006). Development of the Polish Labour Force Survey, *Statistics in Transition*, Vol. 7, No. 5, pp. 1009-1030.

RAO, J. N. K., (2003). Small Area Estimation, Wiley, London.

RAO, J. N. K., GRAHAM, J. E. (1964). Rotation designs for sampling on repeated occasions. *Ann. Math. Statist*. 35, 492-509.

SCOTT, A. J., SMITH, T. M. F. and JONES, R. G. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review*, vol. 45, s. 13-28.

SZARKOWSKI, A., WITKOWSKI, J. (1994). The Polish Labour Force Survey, *Statistics in Transition*, Vol. 1, No. 4, pp. 467-483.

TILLER, R. (1989). A Kalman filter approach to labor force estimation using survey data. Proceedings of the Section on Survey Research Methods, *American Statistical Association*, 16-25.

WESOŁOWSKI, J. (2010). Recursive Optimal Estimation in Szarkowski Rotation Scheme, *Statistics in Transition – new series*, Vol. 11, No.2., pp. 267-285.

WILKS, S.S, (1940). Representative Sampling and Poll Reliability, *Pub. Opin. Quart.*, 4.

WOLTER, K. M. (1979). Composite estimation in finite populations. Journal of the American Statistical Association, 74, 604-613.

WOODRUFF, R. S. (1963). The Use of Rotating Samples in the Census Bureau's Monthly Surveys, *Jour. Am. Stat. Ass.*, vol. 58, pp, 454-467.

YANSANEH, I. S., FULLER, W. (1998). Optimal recursive estimation for repeated surveys. *Survey Meth. 24, 31-40.*

YATES, F. (1949). Sampling Methods for Censuses and Surveys. First Edition. Charles Griffin, London.

ZASĘPA, R. (1958). Problems of Sampling Surveys of GUS in the Light of Consultation with Prof. J. Neyman (in Polish). "*Wiadomości Statystyczne*", nr 6, pp. 7-12.

ZASĘPA, R. (1962). Statistical Investigations by Sampling Method, PWN, Warszawa.

ZASĘPA, R. (1972). *Sampling Methods*, PWE, Warszawa.

ZASĘPA, R. (1993). Precision of household budget surveys, "*Wiadomości Statystyczne"*, No. 3.

STATISTICS IN TRANSITION-new series, Summer 2012
Vol. 13, No. 2, pp. 261—278

ESTIMATION OF PARAMETERS FOR SMALL AREAS USING HIERARCHICAL BAYES METHOD IN THE CASE OF KNOWN MODEL HYPERPARAMETERS

Jan Kubacki[1]

ABSTRACT

In the paper the method of parameters estimation using hierarchical Bayes (HB) method in the case of known model hyperparameters for *a priori* conditionals was presented. This approach has some advantage in comparison with subjective model parameters selection because of more simulation stability and allows obtaining estimates that has more regular distribution. As an example the data about average per capita income from Polish Household Budget Survey for counties (NUTS4) and auxiliary variables from Polish Tax Register (POLTAX) were used. The computation was done using WinBUGS software and R-project environment with R2WinBUGS package, which control the simulations in WinBUGS, and coda package, which allows performing the analysis of simulation results. In the paper sample code in R-project that can be used as a pattern for further similar applications was also presented. The efficiency of hierarchical Bayes estimation with other small area methods was compared. Such comparison was done for HB and EBLUP techniques, for which some consistency related to the precision of estimates obtained using both techniques was achieved.

Key words: Small area estimation, hierarchical Bayes estimation, WinBUGS.

1. Introduction

Small area estimation methods are obviously used in the situations where there is a need to "borrow strength" to determine the estimation using sample survey, but the sample of considered subpopulation is not large enough, what causes too large estimation error. Here "small area" can be understood as smaller administrative units (for example counties – in Polish poviats) or specific groups extracted from the population (for example specific socio-economic groups). This problem can concern also mini-domains or rare features, which are observed with

[1] Centre for Mathematical Statistics, Statistical Office in Łódź, ul. Suwalska 29, 93-176 Łódź, Poland. E-mail: j.kubacki@stat.gov.pl.

smaller frequency, and because of this the estimates of such variables may cause difficulties even for larger administrative units (for example regions). The estimates for income from unemployment benefits for regions from Household Budget Survey may be a good example here. Relative estimation error here may be sometimes large and may exceed 20%. Application of the small area methods may be justified in such a case.

The small area estimation methodology has been systematically developed since 1980's. Here we can mention books from J.N.K Rao (2003) and N.T. Longford (2005) and Mukhopadhyay (1998). In Polish literature one can also find some examples of more comprehensive studies of this topic. Here we can point out works by Bracha, Lednicki and Wieczorkowski (2003, 2004), Domański and Pruska (2001), Gołata (2004), Dehnel (2003) and Żądło (2008). Small area issues were also the topic of many scientific conferences. Here we can recall one of the first small area estimation conference that was held in Warsaw in 1992 (see Kalton, G., Kordos, J., and Platek, R., 1993) and series of the conferences entitled "Small Area Estimation" that have been organized every two years since 2005. First was the conference organized in Jyväskylä, Finland (see http://www.stat.jyu.fi/sae2005/index.html), than conference that took place in 2007 in Pisa, Italy (see http://sae2007.dsm.unipi.it), next was the conference organized in 2009 in Elche, Spain (see http://icio.umh.es/congresos/sae2009) and the last conference took place in 2011 in Trier, Germany (see http://www.uni-trier.de/index.php?id=30789). Small area estimation topics were also presented at the conferences that were organized in Poland. Here we can mention the "Survey Sampling in Economic and Social Research" conference that is organized by the University of Economics in Katowice (see http://web2.ue.katowice.pl/metoda) and the conference "Multivariate Statistical Analysis" that is organized by University of Łódź (see http://www.msa.uni.lodz.pl). Thus, we can see that literature related to the small area estimation is relatively large and contains wide theoretical material, with application examples, what allows for implementation of small area methods in statistical practice.

Hierarchical Bayes estimation method is one of the most often applied small area estimation method. In the last years the growth of interest of this technique is observed. Here we can mention for example PhD thesis that was prepared by M. Vogt (2010) and B. Liu (2009). This method assumes that both *a priori* distributions $f(\lambda)$ of model parameters and conditional distributions $f(\mu,y|\lambda)$ of small area parameters μ (given the model parameter values) are known. Here also data from survey y should be included. Using Bayes theorem one can obtain *a posteriori* distribution $f(\mu|y)$. In simple cases such distribution can be obtained analytically, but more complex cases require special computational methods using MCMC (Markov Chain Monte Carlo) techniques, which are implemented numerically using Gibbs sampler methods.

2. Hierarchical Bayes (HB) method – application for small areas

Here the assumption for HB method will be presented more accurate. First, it is assumed, that we should obtain the following *a posteriori* distribution:

$$f(\mu \mid \mathbf{y}) = \int f(\mu, \lambda \mid \mathbf{y}) d\lambda \tag{2.1}$$

Using Bayes inference we can obtain the following dependence:

$$f(\mu,\lambda \mid \mathbf{y}) = \frac{f(\mathbf{y},\mu \mid \lambda) f(\lambda)}{f_1(\mathbf{y})} \tag{2.2}$$

where $f_1(\mathbf{y})$ is the marginal distribution and has the form:

$$f_1(\mathbf{y}) = \int f(\mathbf{y},\mu \mid \lambda) f(\lambda) d\mu d\lambda \tag{2.3}$$

As it was mentioned in the introduction, in particular cases to perform such calculations the knowledge about *a priori* distributions is needed. This knowledge can be used in construction of particular models for small areas. In the case considered here we take into account the type A model, and, speaking more precisely, basic area level model, which has the following form:

$$\hat{\theta}_i = \mathbf{z}_i^T \beta + b_i v_i + e_i \tag{2.4}$$

where $\hat{\theta}_i$ is small area estimator of particular variable for small area i, $\mathbf{z_i}$ is vector of explanatory variable, β is vector of regression coefficients, b_i is known positive constants, v_i represents the model error, and e_i represents the sample design error. It is often assumed, that the values of component v_i constitutes variables that are independent and identically distributed (iid) having the following properties:

$$E_m(v_i) = 0, V_m(v_i) = \sigma_v^2 \tag{2.5}$$

where E_m is the expected value for the component v for model, and V_m is the model variance. It is assumed for design error, that (for direct estimates)

$$E_p(e_i \mid \theta_i) = 0, V_p(e_i \mid \theta_i) = \psi_i \tag{2.6}$$

It is also assumed that estimation error for direct estimates ψ_i is also known. Taking into consideration the (2.4-2.6) and assuming that the distribution of model error σ_v^2 is also known and has the inverse Gamma distribution $G^{-1}(a,b)$ having parameters a and b (where a is the shape parameter and b is the scale parameter) the hierarchical model can be written in the following form:

(i) $\hat{\theta}_i \mid \theta_i, \beta, \sigma_v^2 \overset{ind}{\sim} N(\theta_i, \psi_i)$ $i=1,\ldots m$

(ii) $\theta_i \mid \beta, \sigma_v^2 \overset{ind}{\sim} N(\mathbf{z}_i^T \beta, b_i^2 \sigma_v^2)$ $i=1,\ldots m$

(iii) $f(\beta) \sim 1$

(iv) $\sigma_v^2 \mid \beta, \boldsymbol{\theta}, \hat{\boldsymbol{\theta}} \sim \mathrm{G}^{-1}(a,b)$ (2.7)

and here the case of known distribution of σ_v^2 and "flat" prior for β, given by $f(\beta) \sim 1$ is considered. It is also assumed that (in contrast to model (10.3.1) from Rao book), values of the parameters a and b in Gamma distribution for σ_v^2 are

known, what is a good approximation for the model from paragraph 10.3.3 in Rao. These values can be obtained from empirical distribution of model estimates that can be determined from linear regression models. Because models that have identical explanatory variables and similar variability of the estimates for both direct estimates and regression coefficients are considered, such approximation may lead to correct estimates of *a posteriori* for hierarchical model. According to Rao suggestion (p. 237) "when σ_v^2 is assumed to be known and $f(\beta) \sim 1$, the HB and BLUP approaches under normality lead to identical point estimates and measures of variability". However, it should be noted that model (10.3.1) in our opinion reflects the variability of σ_v^2 slightly less, what leads to consistency but with more simplified variance measure (see for example equation (7.1.6) in Rao)

$$MSE(\widetilde{\theta}_i^H) = E(\widetilde{\theta}_i^H - \theta_i)^2 = g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2) \tag{2.8}$$

Thus, taking into consideration such variability, obtained estimates are more consistent with EBLUP estimates (and incorporating full model variability). More details about this issue will be presented in experimental section.

3. Markov chain Monte Carlo (MCMC) methods

Assuming that $\boldsymbol{\eta} = (\boldsymbol{\mu}^T, \boldsymbol{\lambda}^T)^T$ is the vector of small area parameters $\boldsymbol{\mu}$ and model parameters $\boldsymbol{\lambda}$, it should be noted that for more complex models, which model (2.7) is a good example of, obtaining a sample from *a posteriori* distribution that has the form like (2.2) may be difficult because of complex nature of the denominator $f_1(y)$. Application of MCMC method in such a case may allow avoiding such difficulties. Here Markov chain $\{\boldsymbol{\eta}^{(k)}, k=0,1,2,\ldots\}$ is constructed, that the distribution of $\boldsymbol{\eta}^{(k)}$ is converged to unique stationary distribution given by $f(\boldsymbol{\eta}|\mathbf{y})$ denoted by as $\pi(\boldsymbol{\eta})$. Thus, neglecting the first d samples (drawing in the burn-in phase), we can obtain D dependent samples $\boldsymbol{\eta}^{(d)},\ldots,\boldsymbol{\eta}^{(d+D)}$, drawing from the target distribution $f(\boldsymbol{\eta}|\mathbf{y})$. Such sample is independent from starting point $\boldsymbol{\eta}^{(0)}$.

Such Markov chain construction requires that one-step transition probability $P(\boldsymbol{\eta}^{(k+1)}, \boldsymbol{\eta}^{(k)})$ be dependent only on the current state $\boldsymbol{\eta}^{(k)}$. As a consequence it leads to the conclusion, that conditional distribution of $\boldsymbol{\eta}^{(k+1)}$ given $\boldsymbol{\eta}^{(0)},\ldots, \boldsymbol{\eta}^{(k)}$ is independent on the chain history $\{\boldsymbol{\eta}^{(0)},\ldots,\boldsymbol{\eta}^{(k-1)}\}$. In such case the stationary condition for the transition kernel should be satisfied:

$$\int \pi(\boldsymbol{\eta}^{(k)}) P(\boldsymbol{\eta}^{(k+1)} \mid \boldsymbol{\eta}^{(k)}) d\boldsymbol{\eta}^{(k)} = \pi(\boldsymbol{\eta}^{(k+1)}) \tag{3.1}$$

The equation (3.1) shows, that if $\boldsymbol{\eta}^{(k)}$ can be obtained from $\pi(\cdot)$, then also $\boldsymbol{\eta}^{(k+1)}$ can be obtained from $\pi(\cdot)$. It is also necessary to ensure that the distribution of $\boldsymbol{\eta}^{(k)}$ given $\boldsymbol{\eta}^{(0)}$, denoted as $P^{(k)}(\boldsymbol{\eta}^{(k)}| \boldsymbol{\eta}^{(0)})$ converge to $\pi(\boldsymbol{\eta}^{(k)})$ regardless of that how the $\boldsymbol{\eta}^{(0)}$ is chosen. Thus, the chain considered here should be irreducible and aperiodic. Irreducible means that for all starting points η (0) the chain reach some

not empty set in the state space with positive likelihood. Aperiodicity means, that the chain should not oscillate between different set of states in a periodical manner.

4. Gibbs sampler

The computational implementation of MCMC can be performed using the method called Gibbs sampler. We briefly present this method here. The Gibbs sampler assumes that we obtain the series of the samples $\boldsymbol{\eta}^{(k)}$ with partitioning $\boldsymbol{\eta}$ vector into blocks $\boldsymbol{\eta}_1,\ldots,\boldsymbol{\eta}_r$. These blocks can contain one or more elements. For example, for basic area level model we have $\boldsymbol{\mu}=(\theta_1,\ldots,\theta_m)^T=\boldsymbol{\theta}$ and $\lambda=(\beta^T,\sigma_v^2)^T$. In such case $\boldsymbol{\eta}$ can be constituted with the following blocks $\eta_1=\beta$, $\eta_2=\theta_1,\ldots,\ \eta_{m+1}=\theta_m$, $\eta_{m+1}=\sigma_v^2$, assuming that $r=m+2$. It is also required that the following Gibbs conditional should be considered: $f(\boldsymbol{\eta}_1 \mid \boldsymbol{\eta}_2,\ldots,\boldsymbol{\eta}_r,\mathbf{y})$, $f(\boldsymbol{\eta}_2 \mid \boldsymbol{\eta}_1,\boldsymbol{\eta}_3,\ldots,\boldsymbol{\eta}_r,\mathbf{y}),\ldots,f(\boldsymbol{\eta}_r \mid \boldsymbol{\eta}_1,\ldots,\boldsymbol{\eta}_{r-1},\mathbf{y})$. The Gibbs sampler uses the conditionals mentioned above in construction of the transition kernel $P(\cdot|\cdot)$, for which stationary distribution of the Markov chain is equal to $\pi(\boldsymbol{\eta})=f(\boldsymbol{\eta}|\mathbf{y})$. This result is the consequence of the fact that $f(\boldsymbol{\eta}|\mathbf{y})$ is uniquely determined by the Gibbs conditionals.

Gibbs sampler algorithm can be described as follows:

Step 0. Choose the starting point $\boldsymbol{\eta}^{(0)}$ for components $\eta_1^{(0)},\ldots,\eta_r^{(0)}$, assuming, that k is equal 0. We can for example choose as the starting points the REML estimates for model parameters λ and EB estimates for μ parameters. But it can be an arbitrary set of points.

Step 1. Generate $\eta^{(k+1)}=(\eta_1^{(k+1)},\ldots,\eta_r^{(k+1)})$ in the following way. Draw $\eta_1^{(k+1)}$ using $f(\eta_1 \mid \eta_2^{(k)},\ldots,\eta_r^{(k)},y)$, than $\eta_2^{(k+1)}$ using $f(\eta_2 \mid \eta_1^{(k+1)},\eta_3^{(k)},\ldots,\eta_r^{(k)},y),\ldots,$ and finally draw $\eta_r^{(k+1)}$ from $f(\eta_r \mid \eta_1^{(k+1)},\ldots,\eta_r^{(k+1)},y)$

Step 2. Set the $k=k+1$ and go to step 1.

The steps 1-2 constitute one cycle for each k. The sequence $\{\boldsymbol{\eta}^{(k)}\}$ generated by Gibbs sampler is the Markov chain with stationary distribution $\pi(\boldsymbol{\eta})=f(\boldsymbol{\eta}|y)$.

5. Assumptions for hierarchical model and model hyperparameters

As it was shown earlier (see (2.7)), the hierarchical model should contain several assumptions connected with *a priori* distributions that include the sampling scheme, the model that explains the observations and the model variability. Because in the paper estimates for counties (poviats) are considered, some difficulties here that arise mainly from too small sample size should be overcome. Direct estimates and their standard error were determined using a specific technique that assumes using balanced repeated replication technique (BRR) in situations where application of BRR is possible and bootstrap method, where using the BRR is impossible. This method was analyzed earlier (see Kubacki, Jędrzejczak and Piasecki (2011) or Kubacki, Jędrzejczak (2011)) and

reveals effectiveness of such approach. The comparison of bootstrap precision estimates with other techniques, including Taylor linearization methods, indicates that both these techniques are nearly consistent. It should be noted that BRR method is applied now in Polish Household Budget Survey.

In the work considered here the following variables describing some income related categories were investigated:

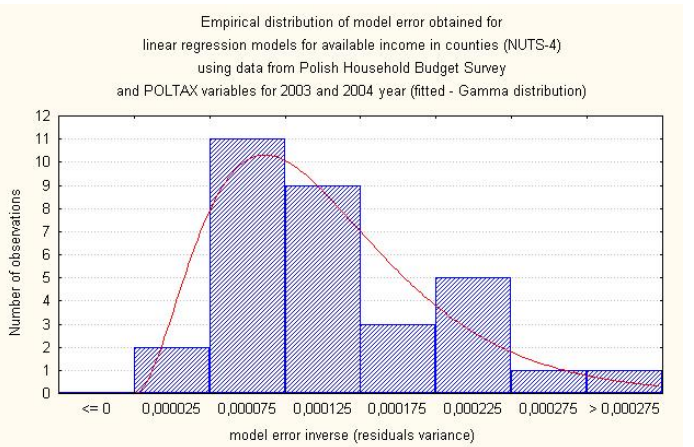
- available income
- income from hired work
- income from self-employment
- income from social security benefits
- retirement pays
- pensions resulting from inability to work
- family pensions
- income from other social benefits
- unemployment benefits.

The explanatory variables for the regression models come from POLTAX register and describe the following categories of income:

1. income from salary, related to employment
2. income from pension, rent (domestics)
3. income from economic activity carried out personally
4. income from property rights
5. income from tenancy or lease
6. income from other sources
7. income from special kind of agriculture production
8. discount from income (revenue) of universal insurance premium contribution
9. discount from tax (lump sum) of universal health insurance premium contribution,

and variables 5,6,7 were linked in one value (as a sum). These data was aggregated at the county - NUTS-4 - level (the anonymous POLTAX file contains the information about administrative unit down to NUTS-5 level) and then the indicator about average income from the mentioned above sources was determined by dividing the sums of this variable for NUTS-4 by the facto population (number of persons) for particular NUTS-4 unit. Such kind of explanatory variables was used for all target variables mainly because of time limit in the considered project. However, it seems that other sources of explanatory data could be used here. Here we can mention data from Polish Social Insurance Company (ZUS) and Labour Offices. This can be treated as an interesting investigation proposition due to the fact that the definitions of the described POLTAX variables only partially corresponds with Household Budget Survey income variables that can weaken the models for small areas.

Figure.1. Empirical distribution of model error obtained for linear regression for available income in counties (NUTS-4) using data from Polish Household Budget Survey and POLTAX variable for 2003 and 2004 year (fitted with Gamma distribution)



Source: Own calculations.

Parameters of distributions for model (2.7) were determined using shape parameters and scale parameters for Gamma distribution estimated from empirical distribution, achieved from the NUTS-4 level models (constructed separately for each region-voivodship NUTS-2). An example of such distribution is shown above.

6. Implementation of the hierarchical model in WinBUGS

In computation the WinBUGS and R-project software was used (also modules R2WinBUGS, coda and MASS). Special macro for R-project was prepared (its simplified example will be shown later), which was used as a connector with data input, performing necessary computations (including simulations in WinBUGS) and automatic visualization (here coda module was used).

In simulations the following computational schema was used. Similar schema was also used in earlier works that was done for hierarchical Bayes applications for small areas. Here we can mention two works: "Small Area Estimation with R Unit 5: Bayesian Small Area Estimation" (see Gomez-Rubio, V., 2008) and „Bayesian Spatial Modeling: Propriety and Applications to Small Area Estimation with Focus on the German Census 2011" (see Vogt, M., 2010). This scheme was as follows.

In the situation presented here Y[p] is related to the direct estimates, their estimation error tau[p], values from A[p] to G[p] are determined by values of explanatory variables for the model, parameters a0 and b0 come from empirical distribution of model error for linear regression and alphas are related to the linear regression coefficients.

```
model
{
 for(p in 1 : N) {
   Y[p] ~ dnorm(mu[p], tau[p])
   mu[p] <- alpha[1] + alpha[2] * A[p] + alpha[3] * B[p] + alpha[4] * C[p] + alpha[5] *
D[p] + alpha[6] * E[p] + alpha[7] * F[p] + alpha[8] * G[p] + u[p]
   u[p] ~ dnorm(0, precu)
 }
 precu ~ dgamma (a0,b0)
 alpha[1] ~ dflat()
 alpha[2] ~ dflat()
 alpha[3] ~ dflat()
 alpha[4] ~ dflat()
 alpha[5] ~ dflat()
 alpha[6] ~ dflat()
 alpha[7] ~ dflat()
 alpha[8] ~ dflat()
 sigmau<-1/precu
}
```

The macro in R-project environment has a (simplified) form like the code presented below. The code includes (for clarity of expression) only sections that present how the model parameters are determined and where simulations are done - with WinBUGS call. The rest of the code has more orderliness character and includes loading the necessary packages (here RODBC, R2WinBUGS and MASS is needed), setting the gamma parameters for σ_v^2 (here fitdistr function is called), reading the input data for particular region (here functions from RODBC package is used), and – after completing the simulations in WinBUGS – arranging the results and estimating the mean and variance (previously using read.coda function) as well as saving the results to the file (here standard cat and format function is used).

```
# determining the model parameters
model_HB<-paste("C:/Documents and Settings/PTS/Moje
dokumenty/model_kongres_demo.txt", sep = "")
infile  <- "coda1.txt"
indfile <- "codaindex.txt"
burn_in <- 3000
a0 <- dochg_shape
b0 <- dochg_rate
data <- list(N=N, Y=Y, tau=tau, A=A, B=B, C=C, D=D, E=E, F=F, G=G, a0=a0, b0=b0)
model <- lm( Y ~ 1 + A + B + C + D + E + F + G)
mod_smry <- summary(model)
alpha <- as.vector(mod_smry$coefficients[,1])
sigma_2 <- (mod_smry$sigma)*(mod_smry$sigma)
precu <- 1/sigma_2
u <- vector(mode = "numeric", length = N)
inits <- list(list(alpha=alpha, precu=precu, u=u))
parameters <- c("mu", "alpha", "precu", "u")

# simulations -  WinBUGS call
sim_HB  <-  bugs(data,  inits,  parameters,  model_HB,n.chains=1,  n.burnin  =  1,
n.iter=10000, n.thin = 1)
```

7. Results and discussion

As it was mentioned earlier estimates from model for HB method (including assumptions for model (2.7)) have similar values as for EBLUP estimator, both for point estimates and for estimation error. The method applied here allows also for obtaining relatively stable simulation history, and the distributions for linear model μ have normal distribution. Normality is achieved also for model error components, and the distribution of σ_v^2 reveals consistency with Gamma distribution. The simulation history also does not have autocorrelation and achieve stability already from the beginning of the simulation. Below, the results of computations for Wielkopolskie voivodship were presented.

Some specific attribute for the computations here is the presence of autocorrelation for model error component in the case of Oborniki county (u[13] denotation). It is connected with relatively low direct estimation error, compared with simulation history for other counties. Such behaviour in MCMC simulation is observed also for other explanatory variables. But existence of such autocorrelation does not change much the normality of their distribution.

The dependencies above for MCMC simulations are observed also for other variables, but fitting the data is sometimes weaker. Achieving normality in such situations may indicate that the assumptions about normality for distributions about estimates and model errors may be in such situation satisfied. However, it is difficult to say whether this fact can be confirmed empirically, because in real situations the change of socio-economic conditions often can be observed what may change the level of the phenomenon (for example because of prize changes and GDP changes), so observed regularities may be characteristic for hypothetical populations often know as superpopulations.

The computations performed for Wielkopolskie voivodship reveal differences between estimation error for EBLUP and HB method, but for majority of similar models the estimation error estimates obtained using these two methods are relatively close. The comparison of REE distribution is presented in Figure 6. However, some differences are observed, and are shown in Figure 7. It is evident from that distribution, that for most cases the HB method has higher REE reduction, then EBLUP estimator. However, REE reduction for EBLUP has more flat patterns that REE reduction for HB method.

Table 1. Values of available income estimate obtained from Polish Household Budget Survey and selected variables from POLTAX register for 2003 and Wielkopolskie voivodship with their precision estimate and relative estimation

error reduction obtained using direct estimation method and EBLUP method using REML technique

County (NUTS-4 unit)	Available income						
	Direct estimates			Estimates for EBLUP method (REML variant - SAE package)			REE reduction
	Parameter estimate	Estimation error	REE (%)	Parameter estimate	Estimation error	REE (%)	
Chodzieski	599.35	63.27	10.56	560.36	33.69	6.01	1.756
Czarnkowsko-Trzcianecki	503.02	80.88	16.08	565.86	28.15	4.97	3.233
Gnieźnieński	506.33	47.71	9.42	586.35	34.20	5.83	1.616
Gostyński	556.11	76.08	13.68	575.33	29.10	5.06	2.705
Grodziski	530.14	51.71	9.75	534.09	36.75	6.88	1.417
Jarociński	731.52	129.69	17.73	581.59	28.62	4.92	3.603
Kępiński	552.20	16.41	2.97	555.65	21.06	3.79	0.784
Kolski	634.46	54.89	8.65	545.68	33.38	6.12	1.414
Koniński	530.42	78.88	14.87	537.14	36.91	6.87	2.164
Kościański	547.35	43.21	7.89	563.69	31.80	5.64	1.399
Krotoszyński	580.99	52.75	9.08	560.27	30.59	5.46	1.663
Nowotomyski	759.51	196.83	25.92	561.16	42.17	7.51	3.449
Obornicki	667.71	4.06	0.61	667.25	4.36	0.65	0.932
Ostrowski	619.02	37.61	6.08	615.20	31.63	5.14	1.182
Ostrzeszowski	579.69	43.57	7.52	569.91	33.26	5.84	1.288
Pilski	728.53	94.61	12.99	625.12	38.75	6.20	2.095
Pleszewski	598.08	86.58	14.48	571.59	34.60	6.05	2.392
Poznański	683.95	85.94	12.57	754.01	43.44	5.76	2.181
Rawicki	694.54	63.63	9.16	571.44	42.61	7.46	1.229
Słupecki	526.62	52.33	9.94	555.81	33.18	5.97	1.665
Szamotulski	588.32	45.80	7.78	586.45	34.01	5.80	1.342
Średzki	594.31	54.73	9.21	610.19	30.31	4.97	1.854
Śremski	670.13	57.39	8.56	583.47	35.56	6.09	1.405
Turecki	457.04	48.58	10.63	513.10	42.97	8.37	1.269
Wągrowiecki	505.59	51.85	10.26	573.06	29.37	5.12	2.001
Wolsztyński	567.58	44.26	7.80	575.68	33.81	5.87	1.328
Wrzesiński	568.85	39.09	6.87	580.39	30.90	5.32	1.291
Złotowski	558.94	45.12	8.07	567.62	31.85	5.61	1.438
m. Kalisz	635.61	13.24	2.08	638.30	16.99	2.66	0.783
m. Konin	699.53	119.79	17.12	622.06	52.57	8.45	2.026
m. Leszno	664.60	74.80	11.26	690.10	53.55	7.76	1.450
m. Poznań	931.31	44.42	4.77	915.60	46.62	5.09	0.937

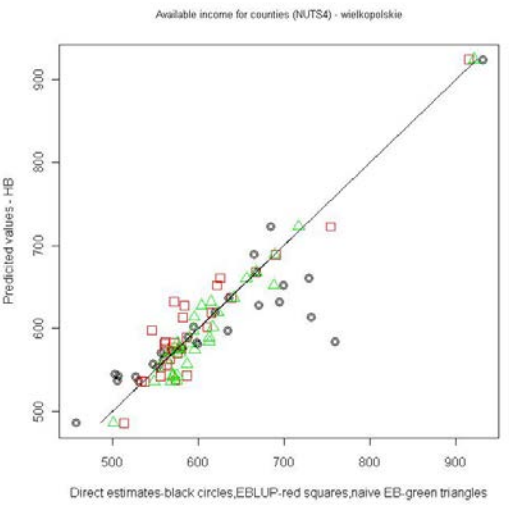
Source: Own calculations.

Table 2. Values of available income estimate obtained from Polish Household Budget Survey and selected variables from POLTAX register for 2003 and Wielkopolskie voivodship with their precision estimate and relative estimation error reduction obtained using direct estimation method and hierarchical Bayes estimation

County (NUTS-4 unit)	Available income						
	Direct estimates			Estimates using hierarchical Bayes method			REE reduction
	Parameter estimate	Estimation error	REE (%)	Parameter estimate	Estimation error	REE (%)	
Chodzieski	599.35	63.27	10.56	581.41	48.17	8.28	1.274
Czarnkowsko-Trzcianecki	503.02	80.88	16.08	544.23	52.31	9.61	1.673
Gnieźnieński	506.33	47.71	9.42	542.99	41.29	7.60	1.239
Gostyński	556.11	76.08	13.68	570.27	51.95	9.11	1.502
Grodziski	530.14	51.71	9.75	536.06	43.51	8.12	1.202
Jarociński	731.52	129.69	17.73	613.32	61.94	10.1	1.756
Kępiński	552.20	16.41	2.97	552.89	15.88	2.87	1.035
Kolski	634.46	54.89	8.65	597.46	45.14	7.56	1.145
Koniński	530.42	78.88	14.87	535.81	55.43	10.4	1.437
Kościański	547.35	43.21	7.89	556.84	36.49	6.55	1.205
Krotoszyński	580.99	52.75	9.08	575.07	41.91	7.29	1.246
Nowotomyski	759.51	196.83	25.92	583.59	78.11	13.4	1.936
Obornicki	667.71	4.06	0.61	667.47	4.09	0.61	0.993
Ostrowski	619.02	37.61	6.08	618.99	33.83	5.47	1.112
Ostrzeszowski	579.69	43.57	7.52	576.71	38.49	6.67	1.126
Pilski	728.53	94.61	12.99	660.38	63.46	9.61	1.351
Pleszewski	598.08	86.58	14.48	582.44	56.15	9.64	1.502
Poznański	683.95	85.94	12.57	722.69	64.37	8.91	1.411
Rawicki	694.54	63.63	9.16	631.92	53.71	8.50	1.078
Słupecki	526.62	52.33	9.94	541.89	42.12	7.77	1.278
Szamotulski	588.32	45.80	7.78	589.18	39.06	6.63	1.174
Średzki	594.31	54.73	9.21	601.61	43.88	7.29	1.263
Śremski	670.13	57.39	8.56	627.53	46.43	7.40	1.157
Turecki	457.04	48.58	10.63	485.88	45.10	9.28	1.145
Wągrowiecki	505.59	51.85	10.26	536.97	41.73	7.77	1.320
Wolsztyński	567.58	44.26	7.80	571.44	38.10	6.67	1.170
Wrzesiński	568.85	39.09	6.87	574.71	33.81	5.88	1.168
Złotowski	558.94	45.12	8.07	562.99	38.27	6.80	1.187
m. Kalisz	635.61	13.24	2.08	636.71	12.91	2.03	1.028
m. Konin	699.53	119.79	17.12	651.80	77.76	11.9	1.436
m. Leszno	664.60	74.80	11.26	689.04	63.24	9.18	1.226
m. Poznań	931.31	44.42	4.77	924.24	43.95	4.76	1.003

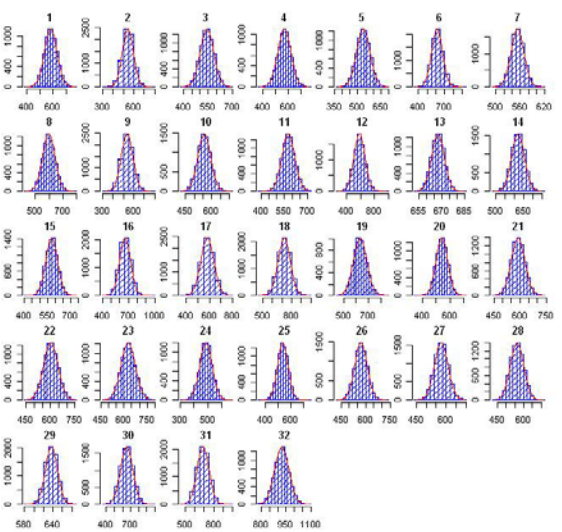
Source: Own calculations.

Figure 2. Observed vs. predicted plot for available income per capita estimates obtained from Polish Household Budget Survey and selected variables from POLTAX register for 2003 and counties in Wielkopolskie voivodship estimated by direct estimator (black circles), EBLUP estimator (red squares) naïve EB estimator (green triangles) and hierarchical Bayes estimator



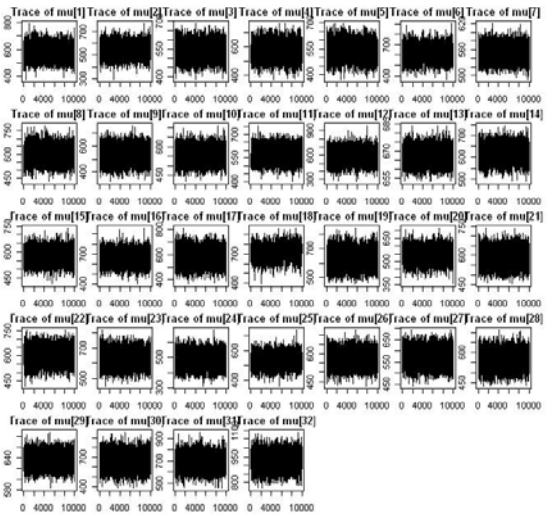
Source: Own calculations.

Figure 3. Plots of distributions of model estimates for available income per capita obtained from Polish Household Budget Survey and selected variables from POLTAX register for 2003 year and counties in Wielkopolskie voivodship obtained by MCMC simulation using Gibbs sampler



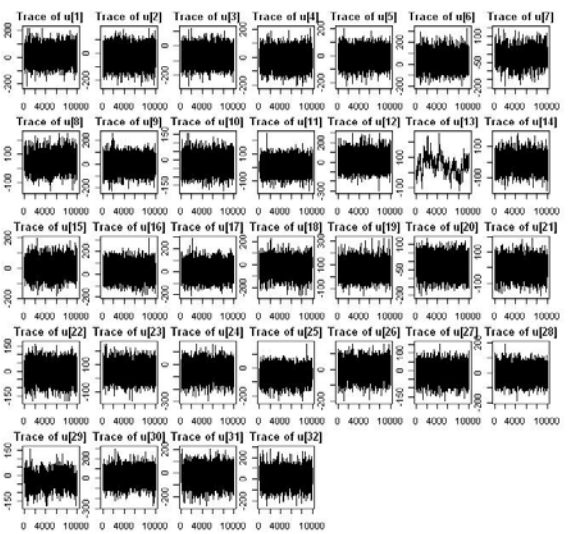
Source: Own calculations.

Figure 4. Plots of simulation history for model estimates of available income per capita obtained from Polish Household Budget Survey and selected variables from POLTAX register for 2003 year and counties in Wielkopolskie voivodship obtained using Gibbs sampler



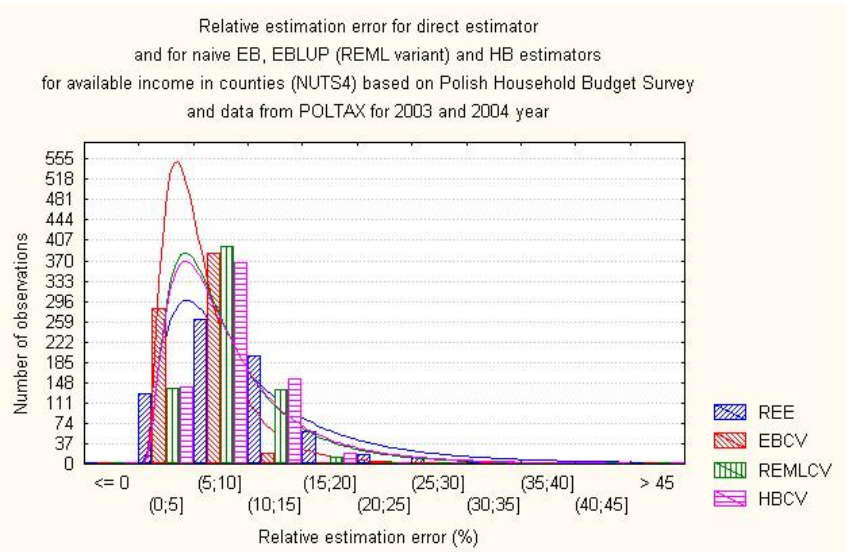
Source: Own calculations.

Figure 5. Plots of simulation history for model error of available income per capita obtained from Polish Household Budget Survey and selected variables from POLTAX register for 2003 and counties in Wielkopolskie voivodship obtained using Gibbs sampler



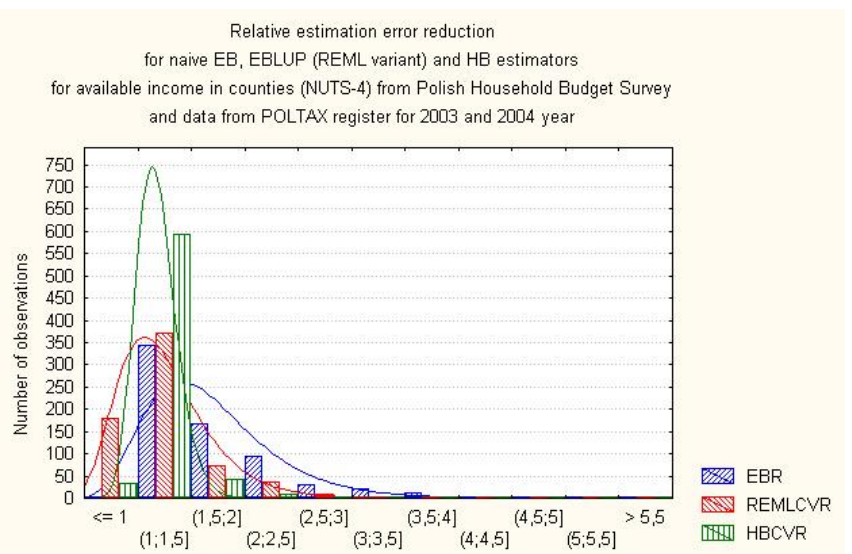
Source: Own calculations.

Figure 6. Distribution of relative estimation error for direct estimator and for naïve EB, EBLUP (REML variant) and HB estimators for available income in counties (NUTS4) based on Polish Household Budget Survey and data from POLTAX register for 2003 and 2004 year



Source: Own calculations.

Figure 7. Distribution of relative estimation error reduction for naïve EB, EBLUP (REML variant) and HB estimators for available income in counties (NUTS4) based on Polish Household Budget Survey and data from POLTAX register for 2003 and 2004 year



Source: Own calculations.

The differences observed for Wielkopolskie voivodship can be explained by weaker fit of the model. Such behaviour for Wielkopolskie region is visible also for ordinary regression models and that in fact can be a limitation on using the HB methods. However, it should be mentioned, that for more specific variables (for example for family pensions or unemployment benefits) the hierarchical models considered here have such an advantage that they rapidly achieve convergence, in contrast to loss of convergence, as it can be observed for some EBLUP models. It is, however, not the property of the hierarchical model itself, but the selection of the parameters of the model. As it was confirmed empirically, other parameters set for Gamma distribution using for σ_v^2 (as it was used for example in Vogt (2010) work, equal a=0.5, b=0.0005) do not behave properly for more specific variables. For that parameters of σ_v^2 the autocorrelation and sometimes the lack of stability (for example the oscillations for longer runs) are observed. Thus, application of such more general approach is not always efficient.

It should be noted here that such selection of parameters is only possible when more cases of similar models are available (as it was characteristic for counties models considered here). In more individual cases (for example when model for available income for voivodship is considered), the availability of more model cases is reduced. In such situation application of other strategy can be more suitable. One of such approaches (when σ_v^2 is not known) was shown in Rao book in part 10.3.3. The comparison of two of these methods may allow for more comprehensive assessment of methods used in this work.

8. Conclusions

In the paper the usefulness of estimates conducted by the hierarchical Bayes estimation in the case of known values of hyperparameters was demonstrated. Some consistency between hierarchical Bayes and other types of small area methods, for example EBLUP method, was shown. For this technique slightly better efficiency than EBLUP estimators was observed, but for less fitted model it could not be the rule. Because of good properties of computations shown in the paper (lack of autocorrelation and practically neglect of burn-in), it can be judged that such approach may be applied in practice. Unfortunately, in the situation described here some preliminary knowledge about the distribution of σ_v^2 is required, what may be sometimes difficult to obtain. In the case of counties it is, however, possible, and may be beneficial for practical reasons.

REFERENCES

BRACHA, CZ., LEDNICKI, B, WIECZORKOWSKI, R. (2003). Data Estimation for Polish Labour Force Survey for counties in 1995-2002 (in Polish - Estymacja danych z Badania Aktywności Ekonomicznej Ludności na poziomie powiatów dla lat 1995-2002), GUS, Warszawa.

BRACHA, CZ., LEDNICKI, B., WIECZORKOWSKI, R. (2004). Application of Complex Estimation Methods to the Disaggregation of data from Polish Labour Force Survey in 2003 (in Polish - Wykorzystanie złożonych metod estymacji do dezagregacji danych z Badania Aktywności Ekonomicznej Ludności w roku 2003), GUS, Warszawa, seria „Z prac Zakładu Badań Statystyczno-Ekonomicznych", z.299 .

CENTRAL STATISTICAL OFFICE (2000-2011). Household Budget Surveys (years 1999-2010) Statistical Information and Elaborations (in Polish Budżety gospodarstw domowych, (lata 1999-2010) Informacje i opracowania statystyczne), Warszawa, http://www.stat.gov.pl/gus/5840_3467_PLK_HTML.htm.

CENTRAL STATISTICAL OFFICE (2010). Polish Household Budget Survey Methodology (in Polish Metodologia Badania Budżetów Gospodarstw Domowych, Zeszyt metodologiczny zaopiniowany przez Komisję Metodologiczną GUS). Warszawa. http://www.stat.gov.pl/cps/rde/xbcr/gus/PUBL_WZ_meto_badania_bud__gospod__dom.pdf.

DEHNEL, G., (2003). Small Area Statistics as a Tool for Assessment of Regions Economic Development (In Polish: Statystyka małych obszarów, jako narzędzie oceny rozwoju ekonomicznego regionów), Wydawnictwo Akademii Ekonomicznej, Poznań.

DOMAŃSKI, CZ., PRUSKA, K. (2001). Methods of Small Area Statistics (in Polish - Metody statystyki małych obszarów), Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

GOŁATA, E. (2004). Indirect Estimation of Unemployment for the Local Labor Market (In Polish: Estymacja pośrednia bezrobocia na lokalnym rynku pracy), Wydawnictwo Akademii Ekonomicznej w Poznaniu, Poznań.

GOMEZ-RUBIO, V. (2008). "Small Area Estimation with R Unit 5: Bayesian Small Area Estimation", useR! 2008 11 August 2008, Dortmund (Germany), http://www.bias-project.org.uk/SAE_tutorial/useR08-tutorial.tgz.

KALTON, G., KORDOS, J., and PLATEK, R. (1993). Small Area Statistics and Survey Designs Vol. I: Invited Papers: Vol. 11: Contributed Papers and Panel Discussion, Warszawa, Główny Urząd Statystyczny.

KUBACKI, J. (2004). Application of the Hierarchical Bayes Estimation to the Polish Labour Force Survey, Statistics in Transition, Vol. 6, No. 5, 785-796. http://www.stat.gov.pl/cps/rde/xbcr/gus/PTS_sit_6_5.pdf.

KUBACKI, J. (2006). The Problems of Small Area Parameters Estimation in Polish Labor Force Survey (In Polish: Problematyka szacowania parametrów dla małych obszarów w badaniu aktywności ekonomicznej ludności), unpublished PhD thesis prepared in connection of PhD Studies in the College of Economic Analysis, Warsaw School of Economics.

KUBACKI, J., JĘDRZEJCZAK, A., PIASECKI, T. (2011). Application of Small Area Statistics Methods in Elaboration of Sample Surveys Results, Report from methodological study 3.065, Statistical Office in Łódź (in Polish Wykorzystanie metod statystyki małych obszarów do opracowania wyników badań statystycznych, Raport z pracy metodologicznej 3.065), Ośrodek Statystyki Matematycznej, Urząd Statystyczny w Łodzi.

KUBACKI, J., JĘDRZEJCZAK, A. (2011). The Comparison of Generalized Variance Function with Other Methods of Precision Estimation for Polish Household Budget Survey, Studia Ekonomiczne, Uniwersytet Ekonomiczny w Katowicach (in preparation).

LIU, B. (2009). Hierarchical Bayes Estimation and Empirical Best Prediction of Small Area Proportions, Dissertation submitted to the Faculty of the Graduate School of the University of Maryland, College Park, http://drum.lib.umd.edu/bitstream/1903/9149/1/Liu_umd_0117E_10245.pdf.

LONGFORD, N.T. (2005). Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician, Springer-Verlag, New York.

MUKHOPADHYAY, P. (1998). Small Area Estimation in Survey Sampling, Narosa Pub House.

PLUMMER, M., BEST, N., COWLES, K. and VINES, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC, R News, vol. 6, 7-11.

RAO, J.N.K. (2003). Small Area Estimation, Wiley Interscience, Hoboken, New Jersey.

SALVATI, N., GÓMEZ-RUBIO, V. (2006). SAE: Small Area Estimation with R. R package version 0.07, http://www.bias-project.org.uk/software/SAE_0.07.zip.

SPIEGELHALTER, D.J., THOMAS, A., BEST, N., and LUNN, D. (2003). WinBUGS User Manual, Version 1.4.

STURTZ, S., LIGGES, U., and GELMAN, A. (2005). R2WinBUGS: A Package for Running WinBUGS from R., Journal of Statistical Software, 12(3), 1-16.

VENABLES, W.N. RIPLEY, B.D. (2002). Modern Applied Statistics with S, Fourth Edition. Springer, New York.

VOGT, M. (2010). Bayesian Spatial Modeling: Propriety and Applications to Small Area Estimation with Focus on the German Census 2011, PhD Thesis, University of Trier, http://ubt.opus.hbz-nrw.de/volltexte/2010/578/pdf/Dissertation_Martin_Vogt.pdf.

ŻĄDŁO, T. (2008). Elements of small area statistics with R software (in Polish - Elementy statystyki małych obszarów z programem R), Akademia Ekonomiczna Katowice.

STATISTICS IN TRANSITION-new series, Summer 2012
Vol. 13, No. 2, pp. 279—286

APPLICATION OF ORDER STATISTICS OF AUXILIARY VARIABLE TO ESTIMATION OF THE POPULATION MEAN

Janusz L. Wywiał[1]

ABSTRACT

Estimation of the population average in a finite population by means of sampling strategies dependent on an auxiliary variable highly correlated with a variable under study is considered. The sample is drawn with replacement on the basis of the probability distribution of an order statistic of the auxiliary variable. Observations of the variable under study are the values of the concomitant of the order statistic. The mean of the concomitant values is the estimator of a population mean of the variable under study. The expected value and the variance of the estimator are derived. The limit distributions of the considered estimators were considered. Finally, on the basis of simulation analysis, the accuracy of the estimator is considered.

Key words*:* Order statistic, sample quantile, auxiliary variable, sampling scheme, sampling design, concomitant.

1. Introduction and basic definitions

A fixed and finite population of size N denoted $U = (1,...,N)$ will be considered. The observation of a variable under study and an auxiliary variable are identifiable and denoted by y_i and x_i, $i = 1,...,N$ respectively. The values of both variables are fixed. Our purpose is the estimation of the population mean $\bar{y} = \frac{1}{N}\sum_{k=1}^{N} y_k$. We assume that $x_i < x_{i+1}$, $i = 1,...,N-1$. Let U_x be the set of all distinct values of the auxiliary variable.

Let Q be the set of all possible samples selected from the population U. The sampling design $P(s)$ has to fulfil the assumptions: $P(s) \geq 0$ for all $s \in Q$ and $\sum_{s \in Q} P(s) = 1$.

[1] Katowice University of Economics, Department of Statistics, 14 Bogucicka, 40-226 Katowice, Poland. Email: janusz.wywial@ue.katowice.pl.

Let us consider simple random sampling design with replacement and with unequal probabilities of drawing population elements. The sample size is: $n > 2$. Let p_k be the probability of selecting the k-th population element in each draw. Let (X_i, Y_i) be a random variable whose values are the observations of an auxiliary variable and a variable under study, respectively, in the i-th draw. Thus, the probability function of the random variables (X_i, Y_i) is: $P(X_i = x_k, Y = y_k) = f(x_k) = p_k$ for $i = 1,...,n$ and $k = 1,...,N$. Thus, each random variable (X_i, Y_i) has the same probability function $f(x_k)$. Hence, the pairs of random variables $(X_i, Y_i), (X_2, X_2), ..., (X_i, Y_i), ..., (X_n, Y_n)$ are independent and they have the same distribution function defined by $P(X = x_k, Y = y_k) = f(x_k)$, $k = 1,...,N$. Thus, the sequence $((X_i, Y_i), i = 1,...,n)$ can be treated as data observed in the sample drawn with replacement from the population U with unequal probabilities $f(x_k)$, $k = 1,...,N$.

Let the sample $((X_i, Y_i), i = 1,...,n)$ ordered by the values of X_i be denoted by $((X_{r:n}, Y_{[r:n]}), r = 1,...,n)$ where $X_{r:n}$ is the r-th order statistic (so, $X_{r:n} \le X_{r+1:n}$, $r = 1,...,n-1$) and $X_{r:n}$ is concomitant of $X_{r:n}$, $r = 1,...,n$, see David and Nagaraja (2003), pp. 144. The distribution function of the r-th order statistic is as follows, see Arnold, Balakrishnan and Nagaraja (2008), pp. 42:

$$p_k(r:n) = P(X_{r:n} = x_k) = \sum_{i=0}^{r-1} \sum_{j=0}^{n-r} \frac{n! \left(F(x_{k-1})\right)^{r-1-i} \left(1 - F(x_k)\right)^{n-r-j} \left(f(x_k)\right)^{i+j+1}}{(r-1-i)!(n-r-j)!(j+j+1)!} \quad (1)$$

Let a k-th $k = 1,...,N$ population element be selected with replacement to the sample s of the size m in a single draw with the probability $p_k(r:n) = P(X_{r:n} = x_k)$. Hence, the sampling design of the defined sample is as follows:

$$P_{r:n}(s) = \prod_{k \in s} p_k(r:n)$$

Hence, the above sampling design is proportional to the product of the appropriate probabilities of the distribution function of r-th order statistic.

Moreover, let us note that another sampling scheme for selection of the sample is as follows. The simple sample of the size n is replicated m-times and in each of them the values x_k ($k = 1,...,N$) of the order statistic $X_{r:n}$ is observed.

The above defined sampling design $P_{r:n}(s)$ leads to selecting the sample s in which values $(x_{k_i}, y_{k_i})k \in s$ are observations of the independent random pair $(X_{r:n}^{(k)}, Y_{[r:n]}^{(k)})$ where $X_{r:n}^{(k)}$ has the same distribution as the order statistic $X_{r:n}$ and $Y_{[r:n]}^{(k)}$ has the same distribution as the concomitant variable $Y_{[r:n]}$, $i = 1,...,z$.

Let us note that the proposed sampling design is similar to the sampling design considered by Wywiał (2) but the former one is proportional to the singular value of the order statistic of a positively valued auxiliary variable and it is drawn without replacement.

2. Estimation strategies

We are going to consider the basic properties of the strategy $\left(\bar{y}_s, P_{r:n}(s)\right)$ where $\bar{y}_s$ is the ordinary sample mean:

$$\bar{y}_s = \frac{1}{m}\sum_{i\in s} y_i = \frac{1}{m}\sum_{k=}^{m} Y_{[r:n]}^{(k)} \tag{2}$$

.

The expected value and the variance of the strategy $\left(\bar{y}_s, P_{r:n}(s)\right)$ are as follows:

$$E(\bar{y}_s, P_{r:n}(s)) = E(Y_{[r:n]}) \tag{3}$$

where

$$E(Y_{[r:n]}) = \sum_{k=1}^{N} y_k p_k(r:n) \tag{4}$$

$$V(\bar{y}_s, P_{r:n}(s)) = \frac{1}{m}\sum_{k=1}^{N} (y_k - E(Y_{[r:n]}))^2 p_k(r:n) \tag{5}$$

The unbiased estimator of the variance:

$$V_s(\bar{y}_s, P_{r:n}(s)) = \frac{1}{m-1}\sum_{k\in s} (y_k - \bar{y}_s)^2 \tag{6}$$

David and Nagaraja (2003), pp. 145, show that

$$E(Y_{[r:n]}) = \bar{y} + \frac{v_{xy}}{v_x}\left(E(X_{r:n}) - \bar{x}\right) \tag{7}$$

where

$$v_{xy} = \frac{1}{N}\sum_{k=1}^{N}(y_k - \bar{y})(y_k - \bar{y}), \quad v_x = \frac{1}{N}\sum_{k=1}^{N}(x_k - \bar{x})^2.$$

This straightforwardly leads to the following conclusion: when $E(X_{r:n}) = \bar{x}$, the estimation strategy $\left(\bar{y}_s, P_{r:n}(s)\right)$ is unbiased for the population mean. Thus, the parameters r and n should be assigned in such a way that $E(X_{r:n}) = \bar{x}$.

In the next section, the accuracy of the proposed strategy will be compared with the following ones. The first of them is the well known simple sample mean $\bar{y}_s = \sum_{k\in s} y_k$. The sampling design of the sample of the size n drawn without replacement is: $P_0(s) = \binom{N}{n}^{-1}$. The strategy $\left(\bar{y}_s, P_0(s)\right)$ is unbiased and its variance is $V(\bar{y}_S, P_0(s)) = \frac{N-n}{Nn} v_{*y}$, $v_{*y} = \frac{1}{N-1}\sum_{k=1}^{N}(y_k - \bar{y})^2$.

Another well known strategy is called the mean from the sample of the fixed size n drawn with replacement. Each population element is drawn to the sample from a population with probability proportional to a value of the auxiliary variable. Let $p_k (k = 1,...,N, \sum_{k=1}^{N} p_k = 1)$ be the probability of selection of a k-th population element in a single draw. Using the simplified notation the multinomial sampling design explains the following expression, see e.g. Tillé (2006):

$$P_1(s) = \prod_{k\in s} p_k .$$

Usually the probabilities p_k, $k = 1,...,N$ are determined by the expression:

$$p_k = \frac{x_k}{\sum_{i=1}^{N} x_i}, \text{ for } x_k > 0, \; k = 1,...,N.$$

The unbiased estimation strategy is denoted by $\left(\hat{y}_s, P_1(s)\right)$ where the estimator is of the Hansen-Hurvitz (1943) type:

$$\hat{y}_s = \frac{1}{m}\sum_{k\in s}\frac{y_k}{p_k} \tag{8}$$

The variance of the strategy is:

$$V\left(\hat{y}_s, P_1(s)\right) = \frac{1}{m}\sum_{k=1}^{N}\left(\frac{y_k}{Np_k} - \bar{y}\right)^2 p_k \tag{9}$$

The last estimator using the sample drawn according to the sampling design $P_{r:n}(s)$ is as follows:

$$\tilde{y}_s = \frac{1}{m}\sum_{k\in s}\frac{y_k}{p_k(r:n)}. \tag{10}$$

$E(\tilde{y}_{m,z,s}) = \bar{y}$ and

$$V\left(\tilde{y}_s, P_{r:n}(s)\right) = \frac{1}{m}\sum_{k=1}^{N}\left(\frac{y_k}{Np_k(r:n)} - \bar{y}\right)^2 p_k(r:n) \tag{11}$$

.

The construction of the strategy $\left(\bar{y}_s, P_{r:n}(s)\right)$ leads to the conclusion that its distribution converges to the normal distribution with parameters given by the expression (4) and (5) if $m \to \infty$. This conclusion results straightforwardly from the well known Lindeberg theorem, see e.g. Billingsley (1979).

On the basis of the same theorem it is easy to proof that under sufficiently large sample size m the strategies $\left(\hat{y}_s, P_1(s)\right) \sim N(\bar{y}, V(\hat{y}_s, P_1(s))$ and $\left(\tilde{y}_s, P_{r:n}(s)\right) \sim N(\bar{y}, V(\tilde{y}_s, P_{r:n}(s))$ where $V\left(\hat{y}_s, P_1(s)\right)$ and $V\left(\tilde{y}_s, P_{r:n}(s)\right)$ are given by the expressions (9) and (11), respectively.

3. Comparison of the estimators' accuracy

The accuracy of the considered strategies is based on the following relative accuracy coefficient:

$$e\left(t_s, P(s)\right) = \frac{V\left(t_s, P(s)\right)}{V\left(\bar{y}_s, P_0(s)\right)}$$

Thus, it is the ratio of the variance $V\left(t_s, P(s)\right)$ and the variance of the simple sample (drawn without replacement) mean. We use the following notation:

$$e_1 = e(\bar{y}_s, P_{r:n}(s)), \quad e_2 = e(\hat{y}_s, P_2(s)).$$

The simulation experiments were based on the procedure prepared by means of the program R. Firstly, according to a theoretical probability distribution,

pseudo-values of the random variable (X,Y) have been generated. Two two-dimensional distribution functions have been considered. The first of them has been two dimensional normal distribution denoted by $N(100,0,1,1,\rho)$ where $E(X)=100$, $E(Y)=0$, $V(X)=V(Y)=1$ and the correlation coefficient $\rho=\rho(X,Y)$. The other distribution of the random variable (X,Y) was a two-dimensional exponential one where $X=Z+U$, $Y=Z+V$ and U, V, Z are independent, $Z \sim EXP(\alpha^{-1})$, $U \sim EXP(\beta^{-1})$, $V \sim EXP(\beta^{-1})$, $\alpha^2+\beta^2=1$, $\rho=\beta^2$. The considered sets of generated pseudo-values were of size $N=500$. The program has calculated the inclusion probabilities and finally the mean square error for different population for the considered strategies.

We assume that $m=n$. Thus, in Tables 1 and 2, the parameter r is the rank of the auxiliary variable order statistic for which $|E(X_{r:n})-\bar{x}|=minimum$.

The analysis of Table 1 lets us say that in the case of the normal distribution of the variables (X,Y), the strategy $(\bar{y}_s, P_{r:n}(s))$ is evidently better than $(\hat{y}_s, P_1(s))$. Moreover, the strategy is less accurate than the simple sample mean.

Table 1. The relative efficiency coefficients (%) of the strategies $(X,Y) \sim N(100,0,1,1,\rho)$.

	ρ:	0.5		0.8		0.95	
n	r	e_1	e_2	e_1	e_2	e_1	e_2
10	6	62	102	35	102	20	102
14	8	59	103	31	103	16	103
20	11	57	104	29	104	12	104
24	13	56	105	28	105	11	104
30	16	56	106	27	106	9	106
40	21	56	109	26	108	8	108
50	26	56	111	25	111	7	111

Source: The author's own calculations.

Table 2. The relative efficiency coefficients (%) of the strategies. The exponential distribution of (X,Y).

	ρ :	0.5		0.8		0.95	
n	r	e_1	e_2	e_1	e_2	e_1	e_2
10	2	62	138	38	54	20	15
14	3	67	140	40	55	21	15
20	4	58	141	38	55	16	15
24	5	62	142	39	56	17	16
30	5	58	144	39	57	15	16
40	7	62	147	40	58	16	16
50	7	63	151	41	59	15	17

Source: The author's own calculations.

The accuracy of the estimation in the case of a highly asymmetric two-dimensional exponential distribution of the variable under study and an auxiliary variable is considered in Table 2. When the correlation coefficient is high, $\rho = 0.95$, the accuracy of the strategies $(\hat{y}_s, P_1(s))$ and $(\bar{y}_s, P_{r:n}(s))$ is comparable. We infer that for a rather small correlation coefficient $\rho = 0.5$, the strategy $(\bar{y}_s, P_{r:n}(s))$ is evidently better than the strategy $(\hat{y}_s, P_1(s))$. In this case, the strategy $(\hat{y}_s, P_2(s))$ is less accurate than the simple sample mean. Finally, in the medium case, when $\rho = 0.8$, the strategy $(\bar{y}_s, P_1(s))$ is a little better than $(\hat{y}_s, P_1(s))$.

4. Conclusions

The proposed $(\bar{y}_s, P_{r:n}(s))$ strategy for the population mean (total) is based on sampling design dependent on order statistics of an auxiliary variable. The basic parameters of the strategy are derived. It was shown when it is unbiased. It is quite

easy to show that the considered estimators has an approximately normal distribution when the sample size is sufficiently large. Thus, it is possible to construct confidence intervals for a population mean (total) of a variable under study as well as to test statistical hypotheses about those parameters.

The simulation analysis let us conclude that in the case of two-dimensional distribution of variable under study and the auxiliary one, the proposed estimation strategy $\left(\overline{y}_s, P_{r:n}(s)\right)$ is not worse than the strategy $\left(\hat{y}_s, P_1(s)\right)$.

Acknowledgements

The research was supported by the grant number N N111 434137 from the Ministry of Science and Higher Education. I am grateful to professor Malay Ghosh for his valuable comments.

REFERENCES

ARNOLD, B.C., BALAKRISHNAN N., NAGARAJA H.N. (2008). *A First Course in Order Statistics*. Society for Industrial and Applied Mathematics. Philadelphia.

BILLINGSLEY, P. (1979). *Probability and Measure*. John Wiley & Sons, New York-Chichester-Brisbane- Toronto.

DAVID, H.A., NAGARAJA, H.N. (2003). *Order statistics*. John Wiley & Sons.

HANSEN, M.H., HURVITZ, W.N. (1943). *On the theory of sampling from finite population. Annals of Mathematical Statistic*, 14, 333-362.

TILLE, Y. (2006). *Sampling algorithms*. Springer.

WYWIAŁ, J.L. (2008). Sampling design proportional to order statistic of auxiliary variable. *Statistical Papers* vol. 49, Nr. 2/April, pp. 277-289.

STATISTICS IN TRANSITION-new series, Summer 2012
Vol. 13, No. 2, pp. 287—300

ESTIMATION FOR SHORT TERM STATISTICS

G. Dehnel[1]

ABSTRACT

Economic processes taking places in the EU markets require systematic changes that will facilitate the international exchange of socio-economic information. All this creates a need for modification of the business statistics information system. System modification is just one element of transformations undertaken by the Central Statistical Office aimed at reforming most of its surveys. These changes focus among other things on exploiting administrative sources for purposes of public statistics. Participation of Poland in the MEETS program offered a unique chance to take steps towards a reform of the present state of business statistics. The results obtained in the MEETS program study was the basis for this article. The objective of the paper was to present the possibility of using administrative registers to short term statistics in the light of DG1 survey.

Key words: Domain estimation, Short term statistics, Small business statistics.

1. Introduction

Changes in the information system of statistics are geared towards a greater reliance on administrative data. Information stored in administrative register is expected to improve the effectiveness of statistics. The changes are designed to reduce the response burden for companies owing to statistical reporting, reduce the cost of obtaining data by making direct use of administrative data in the case of non-response, and change the composition of the statistical office staff by increasing the number of employees responsible for data analysis at the expense of those involved in data collection.

All the benefits of using administrative data listed above are especially significant when it comes to short term statistics, which requires adequate estimation methods as well as systematic and timely provision of data by its administrators. This explains the need for continued efforts to find methodological solutions that will improve the effectiveness of the short term statistics research system.

[1] Poznan University of Economics, Department of Statistics. E-mail: g.dehnel@ue.poznan.pl.

A new system should enable quick access to basic measures about, for example, business activity.

Changes should take into account the STS Recommendations issued by Eurostat (cf. Methodology term business statistics, 2006).

They concern the following areas:

1. working-day adjustment[1],
2. seasonal adjustment,
3. data transmission[2],
4. common information policy on STS data revisions,
5. publishing of STS data,
6. treatment of data that should not be published by Eurostat.

The need to change the system of public statistics to enable a greater reliance on administrative registers is a good opportunity to modify short term business statistics methodology. Such modification should involve the use of modern estimation methods, which are based on "borrowing strength" from outside the sample by relying on auxiliary variables. Including administrative data will increase the information scope of databases. Such an approach is aimed at improving estimates and extending the range of levels at which results can be presented across different variables.

Work in the area of estimation can be conducted in at least two directions:

a) studies aimed at making use of information from administrative sources for calibration estimation in an effort to deal with non-response; and

b) studies focused on making use of information from administrative sources for small area estimation methodology.

A full definition of calibration approach was formulated by C.E. Särndal (2007). According to Särndal, the calibration approach to estimation for finite populations consists in:

(a) the computation of weights that incorporate specified auxiliary information and are restrained by calibration equation(s),

(b) the use of these weights to compute linearly weighted estimates of totals and other finite population parameters: weight times variable value, summed over a set of observed units,

(c) satisfying the objective of obtaining nearly design unbiased estimates given that non-response and other non-sampling errors are absent.

[1] The term 'working-day adjustment' concerns both calendar and working/trading day effect adjustments. The calendar effect is related to the fact that the economic activity varies around the special periods and dates in the year while the working/trading day effect originates from the varying number of days of the week in each month.

[2] A reliable and timely transmission of STS data from the Member States to Eurostat is of critical importance for the quality of the data and it's processing without delays. Although many improvements can be stated – among which the implementation of GESMES/TS as a common data transmission protocol is the most important one – there are still issues to be solved.

The second area of research focuses on applications of small area estimation (SAE). Using administrative data to produce business statistics does not fully solve all the problems connected with economic statistics. Distributions of units by variables of interest are strongly right skewed as well as highly differentiated and heavily concentrated. This calls for the application of new, non-traditional methods of small area estimation. They are based on a more specific approach to produce estimates at different levels across variables. SAE comprises such methods as: robust estimation, modification of GREG estimation, Kernel Regression.

2. The aim of the study

Earlier studies on the subject indicate that the short term system of business statistics can rely on data from personal income tax and corporate tax registers and the social insurance register, as well as data from the VAT register. Information about VAT and income tax should be collected directly from businesses. This information should be obtained from accounts of monthly payments (for banks) and tax returns (for the Ministry of Finance). The response burden owing to statistical reporting should be limited to a minimum.

Considering the fact that sampling survey is conducted on a monthly basis and depends on data timeliness, the available registers cannot be used as sources of up-to-date information. All enterprises have an account system, independent of statistical requirements. Thus, the ideal solution would be to collect statistical information from enterprises automatically, in the background, and transfer it to the statistical office without any extra effort on the part of the enterprise. However, this approach requires that a number of conditions are met, it entails certain technical difficulties and requires changes in legal regulations, which exceed the competence of public statistics (cf. MEETS[1] report).

One way to cope with this problem is to apply estimation methods relying on delayed register data. This approach should improve the completeness and quality of statistical data, which in turn will produce better estimates and help to correct information collected incrementally.

One of the objectives of the article is to highlight possibilities of using administrative register resources to improve estimation precision of variables describing short term business statistics. Another task is to extend the scope of estimates by taking into account estimation on a monthly basis at the low level of aggregation.

The necessity to produce estimates of business indicators for short term statistics, especially at a low level of aggregation, is a pressing need for the business sector, government and local authorities. Statistical institutes collect

[1] *Modernization of European Enterprise and Trade Statistics*, a project co-financed by the European Union (30121.2009.004-2009.807) was conducted by the Statistical Office in Poznan in cooperation with Poznan University of Economics.

information which can be used to generate business estimates. Even when register data is available, surveys often provide the only source of data in short term statistics. The question of how to estimate for short term statistics leads to the consideration of estimation methods. For present purposes, there are two of them:

- design-based methods – make use of survey weights and resulting inferences are based on the probability distribution introduced by the sampling design [cf. Rao, 2003 p.1]. The prime example of this approach is the HT estimator. In the design-based theory, although a model is used to assist estimation, no assumption is made that the population is actually defined by a model (Särndal et al., 1992, p.238-239). The approach is therefore also known as model-assisted estimation. The example of this kind of approach is the GREG estimator and synthetic estimator.
- model-based methods, where the actual finite population is regarded as a finite instance of an infinite superpopulation defined on a model incorporating distributional assumptions.

Under both approaches it is possible to strengthen estimates by using, in addition to survey data, other known data, which correlate with the variable of interest. These are known as auxiliary variables or covariates. Both design-based and model-based methods involve constructing a statistical model connecting the variable of interest to covariates. However, in both cases, models are used in different ways.

3. Assumptions of the research

The study made use of four types of estimators. Three of them represent the design-based approach: direct, GREG and regression synthetic. The fourth one – EBLUP - represents model-based methods. The EBLUP estimator is commonly used in for comparing estimator performance.

Direct estimator

The direct estimator **is** commonly used in small area estimation studies as a benchmark for comparing estimator performance.

$$\hat{\bar{Y}}_d^{DIRECT} = \frac{1}{\hat{N}_d} \sum_{i \in u_d} w_{id}\, y_{id} \tag{1}$$

where $\hat{N}_d = \sum_{i \in u_d} w_{id}$ $w_{id} = 1/\pi_{id}$

assuming that $\pi_{id,jd'} = 0$, for each $d \neq d'$ or $i \neq j$. Standard estimation error is calculated using the following formula:

$$M\hat{S}E(\hat{\bar{Y}}_d^{DIRECT}) = \left(\frac{1}{\hat{N}_d}\right)^2 \sum_{i \in u_d} w_{id}\,(w_{id}-1)(y_{id} - \hat{\bar{Y}}_d^{DIRECT})^2 \tag{2}$$

It is characterised by high variability for most small areas; besides, its application does not guarantee estimates of the target variable for all domains – particularly with respect to cases of non-inclusion in the sample for a given domain. For this reason it is not very useful for estimation (cf. also Särndal et al. 1992, Ghosh, Rao, 1994, Lehtonen, Veijanen, 1998, Eurarea Documents: Standard Estimators, 2001).

Generalised REGression estimator – GREG

The Greg estimator is treated as a specific case of direct estimator. The direct estimator for a given small area is adjusted for differences between the sample and population area means of covariates. Auxiliary variables are transformed and adapted to the value of the target valuable. For this purpose, various models are used, which describe the relationship between the target variable Y and the auxiliary variable X. The standard approach is to use the ordinary regression model:

$$\hat{Y}_d^{GREG} = \frac{1}{\hat{N}_d}\sum_{i\in s_d}\frac{y_i}{\pi_i} + \left(\overline{\mathbf{X}}_d^T - \frac{1}{\hat{N}_d}\sum_{i\in s_d}\frac{\mathbf{x}_i}{\pi_i}\right)^T \hat{\boldsymbol{\beta}} \tag{3}$$

where $\hat{N}_d = \sum_{i\in s_d}\frac{1}{\pi_i}$ and $\hat{\boldsymbol{\beta}}$ are estimated using the least square method.

When a domain contains no data, the GREG estimator reduces to a synthetic estimator, $\overline{\mathbf{X}}_d^T\hat{\boldsymbol{\beta}}$. The formula for the MSE estimator is:

$$\hat{MSE}(\hat{\overline{Y}}_d^{GREG}) = \sum_{i\in u_d}\sum_{j\in u_d}\frac{\pi_{ijd} - \pi_{id}\pi_{jd}}{\pi_{ijd}\pi_{id}\pi_{jd}} g_{id} r_{id} g_{jd} r_{jd} \tag{4}$$

The use of the auxiliary variable X can be justified by its strong correlation with the target variable Y. In this case, the variance of the GREG estimator is lower than the variance of the direct estimator. A small sample size in a domain is conducive to an increase in variance, but with increasing correlation between variables Y and X, variance is considerably reduced. One advantage of GREG estimator is its lack of bias. Assuming that multiple samples are drawn, the expected value of the GREG estimator for a domain is close to the real value of the variable for this domain in the population.

Synthetic estimator

In synthetic estimation for a population divided into homogenous categories it is assumed that the means computed for units belonging to each category are identical. Estimation for domains is the weighted mean of estimated means determined on the basis of sampled units. The weight depends on the share of a small area within a category. The synthetic estimator is unbiased provided the assumption is met. In reality, however, this happens extremely rarely. The regression synthetic estimator is constructed on the basis of a two-level model for

unit data of the variable Y, accounting for the correlation with the values of covariates X at the level of individual units and territorial units:

$$y_{id} = x_{id}^T \beta + u_d + e_{id} \tag{5}$$

where: $u_d \sim iid\ N(0, \sigma_u^2), \quad e_{id} \sim iid\ N(0, \sigma_e^2)$ described by the formula:

$$\hat{\bar{Y}}_d^{SYNTH} = \overline{X}_{.d}^T \hat{\beta} \tag{6}$$

The estimator does not account for sampling weights, and MSE can be estimated using the formula:

$$\hat{MSE}(\hat{\bar{Y}}_d^{SYNTH}) = \hat{\sigma}_u^2 + \overline{\mathbf{X}}_{.d} \hat{\mathbf{V}} \overline{\mathbf{X}}_{.d}^T \tag{7}$$

where $\hat{V}$ is the covariance matrix of auxiliary variables.

The EBLUP estimator

Empirical Best Linear Unbiased Predictors (EBLUP) can be explained in the following manner. They are predictors for small areas, and are the best in the sense of having the least model variance; they are linear in the sense of having a linear function of the sample values y; they are unbiased in the sense of lacking model-based bias. EBLUP is a composite estimator, combining direct linear estimators and regression synthetic estimators with weights depending on the value of MSE estimators. In the case of unit-level model, EBLUP can be defined as a weighted mean of the synthetic and GREG estimators. In the area-level model, EBLUP is a weighted mean of the direct and regression synthetic estimators. The EBLUP estimator is constructed by replacing the unknown value of variance with its estimate. The general formula of the EBLUP estimator takes the following form:

$$\hat{\bar{Y}}_d^{EBLUP} = w_d^{EBLUP} \hat{\bar{Y}}_d^{GREG} + (1 - w_d^{EBLUP}) \hat{\bar{Y}}_d^{SYNTH} \tag{8}$$

In a more developed form, the models can be described as:

$$\hat{\bar{Y}}_d = \gamma_d (\bar{y}_{.d} - \overline{x}_{.d}^T \hat{\beta}\ ) + \overline{X}_{.d}^T \hat{\beta} \tag{9}$$

where: $$w_d^{EBLUP} = \gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2 / n_d} \tag{10}$$

$\bar{y}_{.d}$ and $\bar{\mathbf{x}}_{.\mathbf{d}}^{\mathbf{T}}$ are mean sample values y and covariates for area *d* respectively, and $\hat{\beta}, \hat{\sigma}_e^2, \hat{\sigma}_u^2$ are parameters estimated on the basis of the standard linear two-level model. The MSE estimator can then be estimated using the formula:

$$\hat{MSE}(\hat{\bar{Y}}_d) = \frac{\gamma_d \hat{\sigma}_e^2}{n_d} + (1-\gamma_d)^2 \overline{X}_{.d}^T \hat{V} \overline{X}_{.d} \tag{11}$$

where $\hat{V}$ is the covariance matrix of auxiliary variables.

A description of estimators used in the study can also be found in the book by R. Chambers and A. Saei (2003). The process of estimating statistics relied on the findings of the EURAREA project[1] [cf. Eurarea_Project_Reference_Volume, 2004] and Report of the MEETS program[2].

The study focused on monthly-based statistics. Information for the study came from the survey conducted by the Statistical Office in Poznan. The survey is conducted in the form of monthly reports submitted by all large and medium-sized enterprises and a 10% sample of small enterprises. Auxiliary data came from administrative registers made available by the Central Statistical Office, the Ministry of Finance and the Social Insurance Institution.

Registers provided by the Ministry of Finance included: National Register of Taxpayers, National Register of VAT Payers, a database of Personal Income Tax (PIT) payers and Corporate Income Tax (CIT) payers. In addition, there were two other databases from the Social Insurance Institution (ZUS): the register of natural persons and the register of legal persons.

All variables from databases were used as auxiliary variables:
- value added tax from the VAT database,
- number of employees from the ZUS register,
- revenue from the PIT or CIT register,
- cost from the PIT or CIT register [3],
- profit from the PIT or CIT register.

Average revenue was the target variable. Since administrative register data were only available as an annual release, they could only be used in the model as

[1] The European project entitled EURAREA IST-2000-26290 *Enhancing Small Area Estimation Techniques to meet European needs* was part of the Fifth framework programme of the European Community for research, technological development and demonstration activities. The project was coordinated by ONS – Office for National Statistics, UK) with the participation of six countries: The United Kingdom, Finland, Sweden, Italy, Spain and Poland. Poland was represented by a team of statisticians from the Chair of Statistics at the Academy of Economics in Poznan, (http://www.statistics.gov.uk/eurarea).

[2] The MEETS program research was conducted using the EBLUPGREG program (Veijanen A., Djerf K., Sőstra K., Lehtonen R., Nissinen K., 2004, EBLUPGREG.sas, a program for small area estimation borrowing Strength Over Time and Space using Unit level model, Statistics Finland, University of Jyväskylä).

[3] *Cost* was not included in the version of the CIT database made available for purposes of the project. Therefore, to complete information coming from this database, the value of cost had to be computed using the following algorithm: Cost = Revenue – Profit + Loss.

delayed variables. Short term statistics requires a certain timeliness of data, which in effect leads to the necessity of performing estimation based on variables from periods prior to the study period.

The level of aggregation adopted for the study was economic activity classification (NACE Rev.2). Estimation was conducted for two consecutive months: *t* and *t+1* (using administrative register data with time delay).

According to the second study objective, estimation was also conducted at a lower level of aggregation. The level of aggregation adopted for the study was a combination of economic activity classification (NACE Rev.2) and the territorial division by province.

The population consisted of business units which participated in the survey sampling and for which it was possible to match data from any of the administrative registers. Estimation was conducted as a simulation study to evaluate the performances of the estimators. 1,000 samples were drawn, which were then used to estimate basic characteristics of enterprises by economic activity sections (NACE Rev.2) and province.

The analysis refers to data from the years 2008 and 2009 for which databases were made available by the Polish Central Statistical Office (CSO).

Estimation was conducted for two consecutive months: *t* and *t+1*.

4. Precision assessment methods

As a consequence of adopting a composite estimation method, the results are biased, owing to sampling error and non-sampling error. Both components are affected by non-sampling error. As for the sampling error, it can only be determined with respect to the estimation component.

For each of the estimators used in the study, expected values were computed based on results obtained in 1,000 samples to determine estimator variance, relative estimation error and relative bias. Measures of estimation precision were determined for each domain and for all domains combined. Thus, it was possible to make both a synthetic assessment of estimator properties and one that accounted for domain size and their unique characteristics. Mean values of estimator properties were estimated following 1,000 samples during the simulation study. In addition, distribution characteristics of the estimators were presented where possible.

The mean value of estimates after 1,000 samples can be calculated from:

$$\hat{\bar{Y}}_d = \frac{1}{1000} \sum_{p=1}^{1000} \hat{Y}_{dp} \tag{12}$$

Where:

d - domain

p=1, ... , 1000 – denotes the sample number;

The approximate value of the variance estimator was thus expressed as [cf. Bracha, 1994 p.33]:

$$\hat{V}(\hat{Y}_d) = \frac{1}{999}\sum_{p=1}^{1000}(\hat{Y}_{dp} - \hat{\bar{Y}}_d)^2 \tag{13}$$

The approximate value of the MSE was computed using the following formula [cf. Choudhry, Rao, 1993 p. 276]:

$$\hat{MSE}(\hat{Y}_d) = \frac{1}{999}\sum_{p=1}^{1000}(\hat{Y}_{dp} - Y_d)^2 \tag{14}$$

where Y_d denotes the known domain characteristic. Root MSE is a measure which combines variance and squared bias. Its estimate is defined on the basis of MSE:

$$\hat{S}(\hat{Y}_d)_{MSE} = \sqrt{\hat{MSE}(\hat{Y}_d)} \tag{15}$$

Relative Error of the Estimate (REE) was calculated on the basis of the value of MSE:

$$\hat{REE}(\hat{Y}_d)_{MSE} = \frac{\sqrt{\hat{MSE}(\hat{Y}_d)}}{Y_d} \tag{16}$$

Absolute bias of the estimator was defined as the difference between the expected and real value.

$$\left|\hat{BIAS}(\hat{Y}_d)\right| = \left|\hat{\bar{Y}}_d - Y_d\right| = \left|\frac{1}{1000}\sum_{p=1}^{1000}\hat{Y}_d - Y_d\right| \tag{17}$$

On the basis of the above characteristics computed for each domain it was possible to assess estimation precision for a domain, accounting for its specific nature, especially, the number of units.

5. Estimation results and assessment of their precision

Estimating *revenue* by NACE sections

The results of estimating *revenue* at the level of selected economic activity sections (NACE Rev.2) for period *t* are presented in Tables 1-6. Table 1 contains expected values obtained in the simulation study after 1,000 samples. The last column contains mean revenue within each section. It is used as the benchmark to

assess the convergence of estimates. The actual assessment of estimation precision and bias is possible using information presented in tables 2–6.

Table 1. The expected value of estimators for revenue, period *t*

SECTION	Estimator				POPULATION MEAN
	DIRECT	GREG	SYNTHETIC	EBLUP	
Manufacturing	54585.85	54625.55	54768.17	54661.80	54576.28
Construction	34855.68	34836.24	34559.73	34703.67	34898.88
Trade	80320.49	80244.88	79884.69	80201.53	80280.19
Transport	63016.47	63255.07	63625.85	63386.54	63028.05

Source: Own tabulation based on the MEETS real dataset.

Table 2. Variance of estimators for revenue, period *t*

SECTION	Estimator			
	DIRECT	GREG	SYNTHETIC	EBLUP
Manufacturing	89293.41	37769.69	35070.17	21969.24
Construction	744019.25	70269.94	42931.32	47628.88
Trade	3042933.29	231401.36	1290654.38	271080.04
Transport	646764.21	1136134.85	56900.57	686058.41

Source: Own tabulation based on the MEETS real dataset.

Table 3. MSE of estimators for revenue, period *t*

SECTION	Estimator			
	DIRECT	GREG	SYNTHETIC	EBLUP
Manufacturing	89384.99	40198.74	71926.17	29289.69
Construction	745887.78	74198.57	158070.54	85775.56
Trade	3044559.05	232648.72	1447231.29	277272.30
Transport	646898.32	1187722.65	414625.08	814702.77

Source: Own tabulation based on the MEETS real dataset.

Table 4. Root MSE of estimators for revenue, period *t*

SECTION	Root MSE			
	DIRECT	GREG	SYNTHETIC	EBLUP
Manufacturing	298.97	200.50	268.19	171.14
Construction	863.65	272.39	397.58	292.87
Trade	1744.87	482.34	1203.01	526.57
Transport	804.30	1089.83	643.91	902.61

Source: Own tabulation based on the MEETS real dataset.

Table 5. REE of estimators for revenue, period *t*

SECTION	REE (%)			
	DIRECT	**GREG**	**SYNTHETIC**	**EBLUP**
Manufacturing	0.55	0.37	0.49	0.31
Construction	2.47	0.78	1.14	0.84
Trade	2.17	0.60	1.50	0.66
Transport	1.28	1.73	1.02	1.43

Source: Own tabulation based on the MEETS real dataset.

Table 6. Absolute bias of estimators for revenue, period *t*

SECTION	Absolute bias of estimators			
	DIRECT	**GREG**	**SYNTHETIC**	**EBLUP**
Manufacturing	9.57	49.26	191.88	85.52
Construction	43.20	62.65	339.15	195.21
Trade	40.30	35.30	395.50	78.65
Transport	11.58	227.02	597.80	358.49

Source: Own tabulation based on the MEETS real dataset.

To assess the estimation one can use REE (cf. Table 5). This measure is based on estimates of MSE, which can be compared with its known domain characteristic, thus accounting for estimation precision and bias. The GREG estimator and *EBLUP estimator* yielded similar estimates for each of NACE sections. A significant improvement in estimation precision could be observed. For *manufacturing*, where the best results were obtained, REE is at 0.3% of the known domain characteristic. The bias of the GREG estimator is considerably lower than that of the EBLUP estimator, which often yields better general results owing to its lower variance. In the case of the *transport* section, however, none of the estimators used produced better results than those obtained by means of direct estimation.

Tables 7–8 present selected characteristics of the estimation of revenue for period *t+1*. The currently available data from administrative databases, updated annually or quarterly, can be used to inform a model based on delayed auxiliary data.

Table 7. The expected value of estimators for revenue, period $t + 1$

SECTION	Expected values of estimators				POPULA-TION MEAN
	DIRECT	**GREG**	**SYNTHETIC**	**EBLUP**	
Manufacturing	4140.10	4130.69	4146.73	4131.16	4138.36
Construction	4325.28	4310.89	4111.81	4300.93	4319.91
Trade	704.06	724.35	993.19	733.68	704.23
Transport	6630.63	6636.30	6597.27	6632.67	6632.82

Source: Own tabulation based on the MEETS real dataset.

Table 8. REE of estimators for revenue, period $t + 1$

SECTION	REE (%)			
	DIRECT	GREG	SYNTHETIC	EBLUP
Manufacturing	0.79	0.47	0.51	0.45
Construction	2.51	1.71	4.88	1.70
Trade	1.64	6.32	42.57	7.14
Transport	1.03	0.90	0.66	0.82

Source: Own tabulation based on the MEETS real dataset.

Estimating *revenue* by NACE section and province

In order to analyse estimates at the level of NACE sections and province one needs to choose a method of presentation. Owing to limited space, the results are confined to the expected value of revenue for two NACE sections (for all provinces).

Table. 9. REE of estimators for revenue, in the *manufacturing* section by province

Province	REE (%)			
	DIRECT	GREG	SYNTHETIC	EBLUP
Dolnośląskie	30.19	13.23	37.09	17.00
Kujawsko-Pomorskie	39.33	25.08	32.09	17.68
Lubelskie	54.80	27.54	32.00	17.34
Lubuskie	150.60	11.81	14.12	8.29
Łódzkie	49.21	12.85	24.74	11.05
Małopolskie	32.36	16.27	25.61	12.18
Mazowieckie	36.54	53.83	47.79	45.07
Opolskie	70.01	17.84	20.21	10.58
Podkarpackie	37.93	24.66	28.29	14.25
Podlaskie	41.01	35.82	36.14	22.95
Pomorskie	39.52	34.41	24.26	16.72
Śląskie	23.77	19.77	22.46	11.76
Świętokrzyskie	64.00	23.87	26.32	16.35
Warmińsko-Mazurskie	112.50	35.42	17.89	14.72
Wielkopolskie	36.02	11.37	17.77	9.27
Zachodniopomorskie	34.42	17.83	30.30	14.31

Source: Own tabulation based on the MEETS real dataset.

Table. 10. REE of estimators for revenue, in the *construction* section by province

Province	REE (%)			
	DIRECT	**GREG**	**SYNTHETIC**	**EBLUP**
Dolnośląskie	32.09	19.79	17.02	9.25
Kujawsko-Pomorskie	40.01	15.49	23.71	14.08
Lubelskie	42.32	18.34	20.47	13.85
Lubuskie	70.40	21.34	21.93	11.31
Łódzkie	42.68	18.56	28.84	14.56
Małopolskie	53.21	14.27	22.15	12.68
Mazowieckie	54.81	20.02	13.77	9.01
Opolskie	56.66	22.50	30.17	17.60
Podkarpackie	39.10	18.79	39.15	23.01
Podlaskie	58.30	73.16	22.77	19.41
Pomorskie	91.56	19.28	24.54	18.47
Śląskie	29.52	17.92	24.65	11.71
Świętokrzyskie	136.00	34.22	29.27	25.34
Warmińsko-Mazurskie	43.70	12.70	25.19	14.78
Wielkopolskie	106.50	27.77	24.94	24.76
Zachodniopomorskie	54.24	19.28	21.37	13.22

Source: Own tabulation based on the MEETS real dataset.

Measures of precision and bias presented in Tables 9 and 10 show an evident improvement in efficiency due to the use of estimation using auxiliary data from administrative databases.

6. Conclusion

The use of data from the tax system and the system of social insurance in short term statistics of small, medium and big enterprises in the proposed range undoubtedly shall have a positive influence on the completeness and accuracy of statistical data as far as full surveys are concerned, and in the case of representative surveys it shall improve the quality of estimates. In consequence, it shall contribute to the decrease of responsibilities of enterprises resulting from statistical reporting.

A basic problem in direct use of administrative data in short term statistics of enterprises is too long period of their elaboration and distant time limits of data transfer by administrators, what makes impossible to carry out responsibilities in the scope of time-limits of providing result data to users. Cooperation with administrators aiming at minimising the period of elaboration of administrative data and shortening time limits of their transfer to statistics may be the solution to this question.

The use of information from various data sources about business activity of enterprises can also help to improve the quality of short term statistics of small, medium-sized and large enterprises. This is because these databases are a rich source of potential auxiliary variables, which can be used to increase estimation precision by reducing the negative effect of missing data in statistical reporting.

REFERENCES

BRACHA, C., 1994. Metodologiczne aspekty badania małych obszarów, „Studia i Materiały. Z Prac Zakładu Badań Statystyczno-Ekonomicznych" No. 43, CSO, Warsaw.

CHAMBERS, R., SAEI, A., 2003. Linear Mixed Model with Spatial Correlated Area Effect in Small Area Estimation.

Eurarea_Project_Reference_Volume, 2004.

GHOSH, M., RAO J.N.K., 1994. *Small Area Estimation, An Appraisal,* „Statistical Science", Vol. 9, No.1.

LEHTONEN, R., VEIJANEN, A., 1998. *On multinomial logistic generalized regression estimators*, Department of statistics, University of Jyväskylä, No. 22, Jyväskylä.

Report of the MEETS program, 2011. CSO, Warsaw.

Methodology term business statistics, Associated documents, Luxembourg: Office for Official Publications of the European Communities, 2006.

RAO, J.N.K., 2003. Small Area Estimation, John Wiley & Sons, Inc., Hoboken, New Jersey.

SÄRNDAL, C.E., 2007. The Calibration Approach in Survey Theory and Practice, Survey Methodology. Vol. 33. No. 2. 99–119.

SÄRNDAL, C.E., SWENSSON B., WRETMAN J., 1992. *Model Assisted Survey Sampling*, Springer Verlag, New York Inc.

STATISTICS IN TRANSITION-new series, Summer 2012
Vol. 13, No. 2, pp. 301—310

METHODS OF REPRESENTATION FOR KERNEL CANONICAL CORRELATION ANALYSIS

Mirosław Krzyśko[1], **Łukasz Waszak**[2]

ABSTRACT

Classical canonical correlation analysis seeks the associations between two data sets, i.e. it searches for linear combinations of the original variables having maximal correlation. Our task is to maximize this correlation. This problem is equivalent to solving the generalized eigenvalue problem. The maximal correlation coefficient (being a solution of this problem) is the first canonical correlation coefficient. In this paper we construct nonlinear canonical correlation analysis in reproducing kernel Hilbert spaces. The new kernel generalized eigenvalue problem always has the solution equal to one, and this is a typical case of over-fitting. We present methods to solve this problem and compare the results obtained by classical and kernel canonical correlation analysis.

Key words: Canonical correlation analysis, generalized eigenvalue problem, reproducing kernel Hilbert space.

1. Introduction

The classical tool for studying the association between a dependent variable Y and a set of p explanatory variables $\boldsymbol{X} = (X_1, \ldots, X_p)'$ is multiple regression. Often we are interested in a more complicated interaction, i.e. an interaction between a set of q dependent variables $\boldsymbol{Y} = (Y_1, \ldots, Y_q)'$ and a set of p explanatory variables $\boldsymbol{X} = (X_1, \ldots, X_p)'$. This method was proposed by Hotelling (1936), and is referred to in the literature as canonical correlation analysis.

Our task is to find the strength of association between two vectors $\boldsymbol{Y}$ and $\boldsymbol{X}$. For this purpose we construct new variables V and U, being a linear combination of the original vectors $\boldsymbol{Y}$ and $\boldsymbol{X}$, i.e. $U = \boldsymbol{a}'\boldsymbol{X}$ and $V = \boldsymbol{b}'\boldsymbol{Y}$, where $\boldsymbol{a} \in \mathbf{R}^p, \boldsymbol{b} \in \mathbf{R}^q$. The variables U and V obtained in this way are real one-

1 Faculty of Mathematics and Computer Science at Adam Mickiewicz University in Poznań, mkrzysko@amu.edu.pl.
2 Faculty of Mathematics and Computer Science at Adam Mickiewicz University in Poznań, lwaszak@amu.edu.pl.

dimensional variables ($U,V \in \mathbf{R}$). The dependence between $\boldsymbol{Y}$ and $\boldsymbol{X}$ can be expressed as the classical correlation coefficient $\rho(U,V)$ between V and U.

2. Classical Canonical Correlation Analysis (CCA)

Suppose $\widetilde{\boldsymbol{X}} = (\boldsymbol{X_1},\ldots,\boldsymbol{X_N}) \in \mathrm{R}^{p\times N}$ and $\widetilde{\boldsymbol{Y}} = (\boldsymbol{Y_1},\ldots,\boldsymbol{Y_N}) \in \mathrm{R}^{p\times N}$ are two centred (standardized) N-element data sets, i.e. $E\widetilde{\boldsymbol{X}} = \boldsymbol{0_N}$, $Cov\widetilde{\boldsymbol{X}} = \boldsymbol{1_N}$ and $\boldsymbol{X_i} = (X_{1i},\ldots,X_{pi})' \in \mathrm{R}^p$, $\boldsymbol{Y_i} = (Y_{1i},\ldots,Y_{qi})' \in \mathrm{R}^q$, for $i = 1,\ldots,N$.

The classical task of canonical correlation analysis (CCA) is to find the linear combinations of original vectors $\boldsymbol{Y} \in \widetilde{\boldsymbol{Y}}$ and $\boldsymbol{X} \in \widetilde{\boldsymbol{X}}$ maximizing $\rho(U,V)$, i.e.:

$$\max \rho(U,V),$$

where $U = \boldsymbol{a}'\boldsymbol{X}$ and $V = \boldsymbol{b}'\boldsymbol{Y}$, where $\boldsymbol{a} \in \mathrm{R}^p, \boldsymbol{b} \in \mathrm{R}^q$. Obviously:

$$\rho(U,V) = \frac{cov(U,V)}{\sqrt{Var(U)Var(V)}},$$

$cov(U,V) = E(UV) - E(U)E(V)$,
$Var(U) = E(U^2) - E^2(U)$,
$Var(V) = E(V^2) - E^2(V)$.

Because $\widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{Y}}$ are centred, $E(U) = 0 = E(V)$. Moreover, it can be assumed without loss of generality that $E(U^2) = 1 = E(V^2)$. Consequently we obtain $\rho(U,V) = E(UV)$. Now we want to maximize $E(UV)$, i.e.:

$$\max \rho(U,V) = \max E(UV)$$

To maximize $E(UV)$ we construct a Lagrangian on $\widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{Y}}$:

$$F(\boldsymbol{A},\boldsymbol{B}) = \boldsymbol{A}'\boldsymbol{S_{XY}}\boldsymbol{B} - \frac{\lambda}{2}(\boldsymbol{A}'\boldsymbol{S_{XX}}\boldsymbol{A} - N) - \frac{\mu}{2}(\boldsymbol{B}'\boldsymbol{S_{YY}}\boldsymbol{B} - N) \tag{1}$$

where $\boldsymbol{A} = (\boldsymbol{a_i}),\ \boldsymbol{a_i} \in \mathbb{R}^p, \boldsymbol{B} = (\boldsymbol{b_i}),\ \boldsymbol{b_i} \in \mathbb{R}^q, i = 1,\ldots,N$ and $\boldsymbol{S_{XX}} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}'$, $\boldsymbol{S_{YY}} = \widetilde{\boldsymbol{Y}}\widetilde{\boldsymbol{Y}}'$, $\boldsymbol{S_{XY}} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{Y}}' = \boldsymbol{S_{YX}}'$ and $\frac{\lambda}{2}, \frac{\mu}{2}$ are Lagrange multipliers.

Taking the derivates of the components of vectors $\boldsymbol{a_i}$ and $\boldsymbol{b_i}$ and equating to zero we obtain:

$$\frac{\partial F}{\partial \boldsymbol{A}} = \boldsymbol{S_{XY}}\boldsymbol{B} - \lambda\boldsymbol{S_{XX}}\boldsymbol{A} = \boldsymbol{0} \tag{2}$$

$$\frac{\partial F}{\partial \boldsymbol{B}} = \boldsymbol{S_{XY}}'\boldsymbol{A} - \mu\boldsymbol{S_{YY}}\boldsymbol{B} = \boldsymbol{0} \tag{3}$$

where $\frac{\partial F}{\partial \boldsymbol{A}} = \left(\frac{\partial F}{\partial \boldsymbol{a_i}}\right), \frac{\partial F}{\partial \boldsymbol{B}} = \left(\frac{\partial F}{\partial \boldsymbol{b_i}}\right)$.

Because $Var(U) = Var(V) = 1$, we arrive at $\lambda = \mu = \boldsymbol{A}'\boldsymbol{S}_{XY}\boldsymbol{B} = \rho$. Therefore (2) and (3) are equivalent respectively to (4) and (5):

$$\boldsymbol{S}_{XY}\boldsymbol{B} = \rho\boldsymbol{S}_{XX}\boldsymbol{A} \tag{4}$$

$$\boldsymbol{S}_{XY}'\boldsymbol{A} = \rho\boldsymbol{S}_{YY}\boldsymbol{B} \tag{5}$$

Using a matrix representation, equations (4) and (5) are equivalent to:

$$\boldsymbol{S}_1\boldsymbol{C} = \rho\boldsymbol{S}_2\boldsymbol{C} \tag{6}$$

where

$$\boldsymbol{S}_1 = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{S}_{XY} \\ \boldsymbol{S}_{XY}' & \boldsymbol{0} \end{bmatrix},\ \boldsymbol{S}_2 = \begin{bmatrix} \boldsymbol{S}_{XX} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S}_{YY} \end{bmatrix},\ \boldsymbol{C} = (\boldsymbol{c}_i),\ \boldsymbol{c}_i = \begin{bmatrix} \boldsymbol{a}_i \\ \boldsymbol{b}_i \end{bmatrix}.$$

Equation (6) is a classical case of the generalized eigenvalue problem. The matrix $\boldsymbol{S}_2$ is non-singular and can be presented as the classical eigenvalue problem:

$$(\boldsymbol{S} - \rho\boldsymbol{I})\boldsymbol{C} = \boldsymbol{0} \tag{7}$$

where $\boldsymbol{S} = \boldsymbol{S}_2^{-1}\boldsymbol{S}_1$.

The solutions $\rho_i(\boldsymbol{S})$ of equation (7) $\det(\boldsymbol{S} - \rho\boldsymbol{I}) = 0$ determine the *i*-th canonical correlation coefficient, and the corresponding vector $\boldsymbol{c}_i(\boldsymbol{S})$ the *i*-th pair of canonical variables $(U_i, V_i) = (\boldsymbol{a}_i{}'\boldsymbol{X}, \boldsymbol{b}_i{}'\boldsymbol{Y})$.

3. Introduction to reproducing kernel Hilbert spaces (RKHS)

We introduce some facts about reproducing kernel Hilbert spaces (RKHS) which will be used in our analysis (Preda (2006)). Let $H \subseteq \mathbb{R}^d$ be a set and $\mathcal{H}$ be a Hilbert space of functions on H. Denote by $\langle\cdot,\cdot\rangle$ the inner product of $\mathcal{H}$. A bivariate real-valued function $\boldsymbol{K}$ on H is said to be a reproducing kernel for $\mathcal{H}$ if:
(RK1) $\forall x \in H\colon\ \boldsymbol{K}(\cdot, x) \in H$
(RK2) (reproducing property) $\forall x \in H\ \forall F \in \mathcal{H}\colon F(x) = \langle F, \boldsymbol{K}(\cdot, x)\rangle$.

If $\mathcal{H}$ admits an reproducing kernel $\boldsymbol{K}$, then $\boldsymbol{K}$ has the following properties (Aronszajn, 1950):
(K1) $\boldsymbol{K}$ is the unique reproducing kernel for $\mathcal{H}$.
(K2) $\boldsymbol{K}$ is symmetric and non-negative definite.
(K3) Elements of the form
$\sum_{i=1}^{n} a_i\boldsymbol{K}(t_i, \cdot), n \in \mathbb{N}, \{a_i \in \mathbb{R},\ i = 1, \ldots, n\}, \{t_i \in H,\ i = 1, \ldots, n\}$ are dense in $\mathcal{H}$.

In view of *(K3)*, if $\boldsymbol{K}$ is a symmetric and non-negative definite function, one can construct a Hilbert space $\mathcal{H}_{\boldsymbol{K}}$ which is the completion of all functions on H of the form $\sum_{i=1}^{n} a_i \boldsymbol{K}(t_i,\cdot)$ under the inner product:

$$\langle \sum_{i=1}^{n} a_i \boldsymbol{K}(t_i,\cdot), \sum_{i=1}^{n} b_j \boldsymbol{K}(s_j,\cdot) \rangle = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \boldsymbol{K}(t_i,s_j)$$

Thus, $\mathcal{H}_{\boldsymbol{K}}$ is an RKHS with reproducing kernel $\boldsymbol{K}$ and we have the well known result (Aronszajn, 1950):

Moore–Aronszajn Theorem. To every non-negative definite function $\boldsymbol{K}$ on $H \times H$ there is a corresponding unique RKHS $\mathcal{H}_{\boldsymbol{K}}$ of real-valued functions on H and vice versa.

Examples of reproducing kernels:

Gaussian kernel: $\boldsymbol{k_G(x,y)} = \exp(-\frac{\|x-y\|^2}{2\sigma^2}) = \exp(-c \parallel \boldsymbol{x}-\boldsymbol{y} \parallel^2), c = \frac{1}{2\sigma^2} > 0.$

Polynomial kernel: $\boldsymbol{k_P(x,y)} = (b\langle \boldsymbol{x},\boldsymbol{y}\rangle + c)^d = (b\boldsymbol{x}'\boldsymbol{y}+c)^d,\ b,c \in \mathbb{R}, d \in \mathbb{N}.$

Theorem. Let $F \in \mathcal{H}$ and $\widetilde{\boldsymbol{K}}$ be the kernel matrix for non-centred data. Then the kernel matrix $\boldsymbol{K}$ for centred data is expressed by the formula:

$$\boldsymbol{K} = \boldsymbol{P}\widetilde{\boldsymbol{K}}\boldsymbol{P},$$

where $\boldsymbol{P} = \boldsymbol{I}_N - \frac{1}{n}\mathbf{1}_N\mathbf{1}_N'$ and $\mathbf{1}_N$ is N-dimensional vector consisting of one.

4. Kernel Canonical Correlation Analysis (KCCA)

Let space $\mathbb{R}^p$ be mapped to a reproducing kernel Hilbert space: $\boldsymbol{\varphi}: \mathbb{R}^p \longrightarrow H(k_x),\ \boldsymbol{X}_i \mapsto \boldsymbol{\varphi}(\boldsymbol{X}_i), i = 1,\dots,N,$ and space $\mathbb{R}^q$ to a reproducing kernel Hilbert space $H(k_y)$: $\boldsymbol{\psi}: \mathbb{R}^q \longrightarrow H(k_y),\ \boldsymbol{Y}_i \mapsto \boldsymbol{\psi}(\boldsymbol{Y}_i), i = 1,\dots,N.$ We formulate a new CCA task on the sets $\boldsymbol{\varphi}(\widetilde{\boldsymbol{X}})$ and $\boldsymbol{\psi}(\widetilde{\boldsymbol{Y}})$, i.e. on elements $\boldsymbol{\varphi}(\boldsymbol{X}_i) \in \boldsymbol{\varphi}(\widetilde{\boldsymbol{X}})$ and $\boldsymbol{\psi}(\boldsymbol{Y}_i) \in \boldsymbol{\psi}(\widetilde{\boldsymbol{Y}})$, for $i = 1,\dots,N$, assuming $\boldsymbol{\varphi}(\widetilde{\boldsymbol{X}})$ and $\boldsymbol{\psi}(\widetilde{\boldsymbol{Y}})$ are centred. Then each of the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is in the subspaces stretched respectively on $\boldsymbol{\varphi}(\boldsymbol{X}_i)$ and on $\boldsymbol{\psi}(\boldsymbol{Y}_i)$:

$$\boldsymbol{a} = \sum_{i=1}^{N} \alpha_i \boldsymbol{\varphi}(\boldsymbol{X}_i)$$

$$\boldsymbol{b} = \sum_{i=1}^{N} \beta_i \boldsymbol{\psi}(\boldsymbol{Y}_i)$$

The kernelized problem of CCA has the same form (KCCA):

$$\max \rho(U,V) = \max E(UV)$$

where

$$U = \boldsymbol{\alpha}'\boldsymbol{\varphi}(X) = \sum_{i=1}^{N} \alpha_i \boldsymbol{\varphi}'(X_i)\boldsymbol{\varphi}(X)$$

$$V = \boldsymbol{b}'\boldsymbol{\psi}(Y) = \sum_{i,j=1}^{N} \beta_i \boldsymbol{\psi}'(Y_i)\boldsymbol{\psi}(Y)$$

To maximize $E(UV)$ we construct analogously to CCA a new Lagrangian on $\boldsymbol{\varphi}(\tilde{X})$ and $\boldsymbol{\psi}(\tilde{Y})$:

$$F_H(\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_{i=1}^{N}\left(\sum_{j=1}^{N}\alpha_j\boldsymbol{\varphi}'(X_j)\right)\boldsymbol{\varphi}(X_i)\boldsymbol{\psi}'(Y_i)\left(\sum_{k=1}^{N}\beta_k\boldsymbol{\psi}(Y_k)\right)$$

$$-\frac{\lambda}{2}\left[\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\alpha_j\boldsymbol{\varphi}'(X_j)\right)\boldsymbol{\varphi}(X_i)\boldsymbol{\varphi}'(X_i)\left(\sum_{j=1}^{N}\alpha_j\boldsymbol{\varphi}(X_i)\right) - N\right]$$

$$-\frac{\mu}{2}\left[\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\beta_j\boldsymbol{\psi}'(Y_j)\right)\boldsymbol{\psi}(Y_i)\boldsymbol{\psi}'(Y_i)\left(\sum_{k=1}^{N}\beta_k\boldsymbol{\psi}(Y_k)\right) - N\right] \qquad (8)$$

Using the Moore-Aronszajn theorem and kernel trick, i.e. $k(X_i,X_j) = \langle\boldsymbol{\varphi}(X_i),\boldsymbol{\varphi}(X_j)\rangle$ and $k(Y_i,Y_j) = \langle\boldsymbol{\psi}(Y_i),\boldsymbol{\psi}(Y_j)\rangle$, (8) can be presented as:

$$F_H(\boldsymbol{\alpha},\boldsymbol{\beta}) = \boldsymbol{\alpha}'\boldsymbol{K_X}\boldsymbol{K_Y}\boldsymbol{\beta} - \frac{\lambda}{2}\left(\boldsymbol{\alpha}'\boldsymbol{K_X^2}\boldsymbol{\alpha} - N\right) - \frac{\mu}{2}\left(\boldsymbol{\beta}'\boldsymbol{K_Y^2}\boldsymbol{\beta} - N\right) \qquad (9)$$

where $\boldsymbol{K_X} = \left[k(X_i,X_j)\right]$ and $\boldsymbol{K_Y} = \left[k(Y_i,Y_j)\right]$ are kernel matrices and where $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_N)'$ and $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_N)'$.

Taking the derivates of the components of vectors $\boldsymbol{\alpha_i}$ and $\boldsymbol{\beta_i}$ and equating to zero we obtain:

$$\frac{\partial F_H}{\partial\boldsymbol{\alpha}} = \boldsymbol{K_X}\boldsymbol{K_Y}\boldsymbol{\beta} - \lambda\boldsymbol{K_X^2}\boldsymbol{\alpha} = \boldsymbol{0} \qquad (10)$$

$$\frac{\partial F_H}{\partial\boldsymbol{\beta}} = \boldsymbol{K_Y}\boldsymbol{K_X}\boldsymbol{\alpha} - \mu\boldsymbol{K_Y^2}\boldsymbol{\beta} = \boldsymbol{0} \qquad (11)$$

where $\frac{\partial F}{\partial\boldsymbol{\alpha}} = \left(\frac{\partial F}{\partial\boldsymbol{\alpha_i}}\right), \frac{\partial F}{\partial\boldsymbol{\beta}} = \left(\frac{\partial F}{\partial\boldsymbol{\beta_i}}\right)$.

Analogously, we arrive at $\lambda = \mu = \boldsymbol{\alpha}'\boldsymbol{K_X}\boldsymbol{K_Y}\boldsymbol{\beta} = \rho$. Therefore, (10) and (11) are equivalent to (12) and (13) respectively:

$$\boldsymbol{K_X}\boldsymbol{K_Y}\boldsymbol{\beta} = \rho\boldsymbol{K_X^2}\boldsymbol{\alpha} \qquad (12)$$

$$\boldsymbol{K_Y}\boldsymbol{K_X}\boldsymbol{\alpha} = \rho\boldsymbol{K_Y^2}\boldsymbol{\beta} \qquad (13)$$

Using a matrix representation, equations (12) and (13) are equivalent to:

$$\boldsymbol{K_1 C(k) = \rho(k) K_2 C(k)} \tag{14}$$

where

$$\boldsymbol{K_1 = \begin{bmatrix} 0 & K_X K_Y \\ K_Y K_X & 0 \end{bmatrix},\ K_2 = \begin{bmatrix} K_X^2 & 0 \\ 0 & K_Y^2 \end{bmatrix},\ C(k) = (c_i(k)), c_i(k) = \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix}}.$$

Equation (14) is the generalized eigenvalue problem and could again be presented as the classical eigenvalue problem:

$$\boldsymbol{(K - \rho(k) I) C(k) = 0} \tag{15}$$

where $\boldsymbol{K = K_2^{-1} K_1}$. The matrix $\boldsymbol{K_2}$ is non-negative determined and can include singularity values. The generalized Moore–Penrose pseudoinverse of the matrix is ambiguous.

In order to solve this problem we first apply the idea used in **ridge regression**, namely regularization of the matrix $\boldsymbol{K_2}$, i.e.

$$\boldsymbol{K_2 \mapsto K_2 + \varepsilon I} \tag{16}$$

where $\varepsilon > 0$; it is enough to take $\varepsilon = 10^{-5}$.

Consequently, the matrix $\boldsymbol{K}$ has the form:

$$\boldsymbol{K = (K_2 + \varepsilon I)^{-1} K_1} \tag{17}$$

A second concept for solving this problem is the idea of using **SVD decomposition** of the matrix $\boldsymbol{K_2}$, i.e.

$$\boldsymbol{K_2 = QGQ'} \tag{18}$$

where $\boldsymbol{Q}$ is the orthogonal ($\boldsymbol{QQ' = I}$) matrix and $\boldsymbol{G}$ is the diagonal matrix of degree t, where t is the number of eigen-values of matrix $\boldsymbol{K_2}$ greater then 10^{-6}.

As a result, we obtain the matrix $\boldsymbol{K}$:

$$\boldsymbol{K = K_2^{-1} K_1 = (QGQ')^{-1} K_1 = QG^{-1} Q' K_1} \tag{19}$$

The solutions $\rho_i(K)$ of equation (15) $\det(K - \rho I) = 0$ determine the *i*-th kernel canonical correlation coefficient, and the corresponding vector $c_i(K)$ the *i*-th pair of kernel canonical variables $(U_i, V_i) = (a_i K_X, b_i K_Y)$.

5. Quasi Kernel Canonical Correlation Analysis (Q-KCCA)

In this case let only the space $\mathbb{R}^p$ be mapped to a reproducing kernel Hilbert space: $\varphi: \mathbb{R}^p \longrightarrow H(k_x)$ and $Y = (Y_1, \ldots, Y_q) \in \mathbb{R}^q$ (Zheng et al. (2006)). We present a concatenation of the methods proposed in section 3 and 4.

$$U = \alpha' \varphi(X) = \sum_{i=1}^{N} \alpha_i \varphi'(X_i)\varphi(X)$$

$$V_k = b'Y = \sum_{i=1}^{N} b_i Y_i$$

Using the Moore-Aronszajn theorem and kernel trick again, i.e. $k(X_i, X_j) = (\varphi(X_i), \varphi(X_j))$, we obtain:

$$F_H(\alpha, B) = \alpha' K_X \tilde{Y} B - \frac{\lambda}{2}\left(\alpha' K_X^2 \alpha - N\right) - \frac{\mu}{2}\left(B'\tilde{Y}\tilde{Y}'B - N\right) \tag{20}$$

Taking the derivates of the components of vectors α_i and b_i and equating to zero we obtain:

$$\frac{\partial F_H}{\partial \alpha} = K_X \tilde{Y} B - \lambda K_X^2 \alpha = 0 \tag{21}$$

$$\frac{\partial F_H}{\partial B} = \tilde{Y}' K_X \alpha - \mu \tilde{Y}\tilde{Y}' B = 0 \tag{22}$$

where $\frac{\partial F}{\partial \alpha} = \left(\frac{\partial F}{\partial \alpha_i}\right), \frac{\partial F}{\partial B} = \left(\frac{\partial F}{\partial b_i}\right)$.

Therefore, (21) and (22) are equivalent to (23) and (24) respectively:

$$K_X \tilde{Y} b = \lambda K_X^2 \alpha \tag{23}$$

$$\tilde{Y}' K_X \alpha = \mu \tilde{Y}\tilde{Y}' b \tag{24}$$

Using a matrix representation, equations (23) and (24) are equivalent to:

$$A_1 A(k) = \rho(k) A_2 A(k) \tag{25}$$

where

$$A_1 = \begin{bmatrix} 0 & K_X \tilde{Y} \\ \tilde{Y}' K_X & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} K_X^2 & 0 \\ 0 & \tilde{Y}\tilde{Y}' \end{bmatrix}, \quad A(k) = (a_i(k)), a_i(k) = \begin{bmatrix} \alpha_i \\ b_i \end{bmatrix}.$$

Again, we can transform problem (25) again into the classical eigenvalue problem:

$$(A - \rho(k) I) A(k) = 0 \tag{26}$$

Using regularization of the matrix K_X^2 we obtain:

$$A = \begin{bmatrix} K_X^2 + \varepsilon I & 0 \\ 0 & \tilde{Y}' \end{bmatrix}^{-1} \begin{bmatrix} 0 & K_X \tilde{Y}' \\ \tilde{Y} K_X & 0 \end{bmatrix} \tag{27}$$

Analogously, using SVD decomposition of the matrix K_X^2 we obtain

$$A = \begin{bmatrix} QG^{-1}Q' & 0 \\ 0 & \left(\tilde{Y}\tilde{Y}'\right)^{-1} \end{bmatrix} \begin{bmatrix} 0 & K_X \tilde{Y} \\ \tilde{Y}' K_X & 0 \end{bmatrix} \tag{28}$$

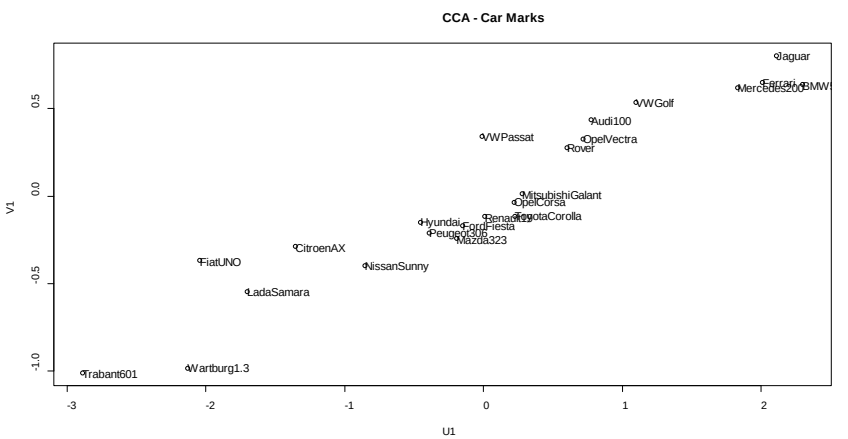
6. Example - Car Marks Data Set

These data come from the book *Applied Multivariate Statistical Analysis* by Wolfgang Härdle and Léopold Simar (2007). They are averaged marks for 24 car types from a sample of 40 persons. The marks range from 1 (very good) to 6 (very bad), on the pattern of German school marks. The variables are (X_1) Economy, (X_2) Service, (X_3) Non-depreciation of value, (X_4) Price (mark 1 for very cheap cars), (X_5) Design, (X_6) Sporty car, (X_7) Safety, (X_8) Easy handling. The exact description of the data set is found in the book in section B.7 Car Marks (p. 434-435). In particular, we would like to investigate the relation between the two variables representing non-depreciation of value and price of the car, and all other variables, i.e. $\boldsymbol{Y} = (X_3, X_4)'$ and $\boldsymbol{X} = (X_1, X_2, X_5, X_6, X_7, X_8)'$. All variables are standardized.

In the classical case of CCA we obtain two non-zero correlation coefficients $\rho_1 = 0.9792$ and $\rho_2 = 0.8851$. For the largest correlation coefficient ρ_1 we obtain vectors $\boldsymbol{a}_1' = [-0.3632, 0.1504] \in \mathbf{R}^2$ and $\boldsymbol{b}_1' = [0.0039, 0.4574, 0.2429, 0.2064, 0.6216, -0.3850] \in \mathbf{R}^6$ which correspond to the canonical variables (U_1, V_1), where:

$$(U_1^i)_{i=1,\ldots,N} = (\boldsymbol{a}_1' \boldsymbol{X}_i)_{i=1,\ldots,N}$$
$$(V_1^i)_{i=1,\ldots,N} = (\boldsymbol{b}_1' \boldsymbol{Y}_i)_{i=1,\ldots,N}$$

A projection (CCA) into the coordinate system of the canonical variables corresponding to the canonical coefficients ρ_1 is shown below:



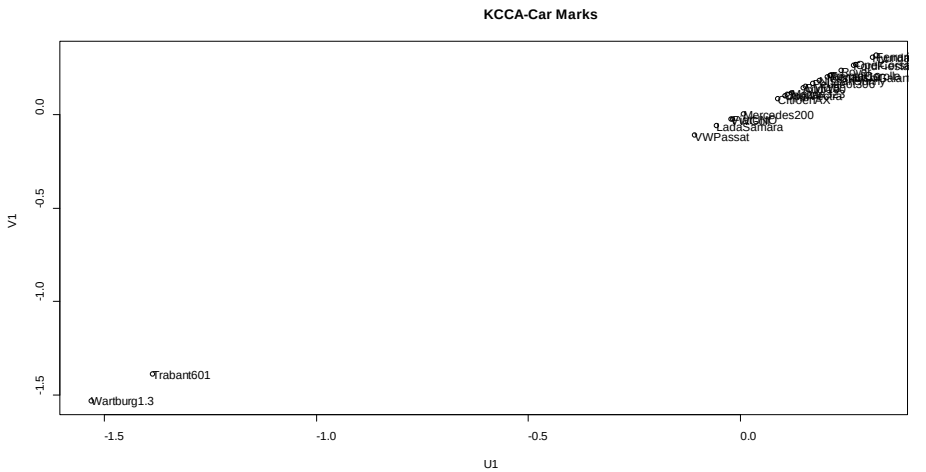
In the kernel case (KCCA) we can use the polynomial kernel $\boldsymbol{k_P}(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}'\boldsymbol{y} + 1)^2$. We obtain 26 non-zero correlation coefficients. We present only the six largest:

$$\rho_1 = 1.000000, \rho_2 = 0.999998, \rho_3 = 0.999980, \rho_4 = 0.999385, \rho_5 = 0.998763, \rho_6 = 0.000002.$$

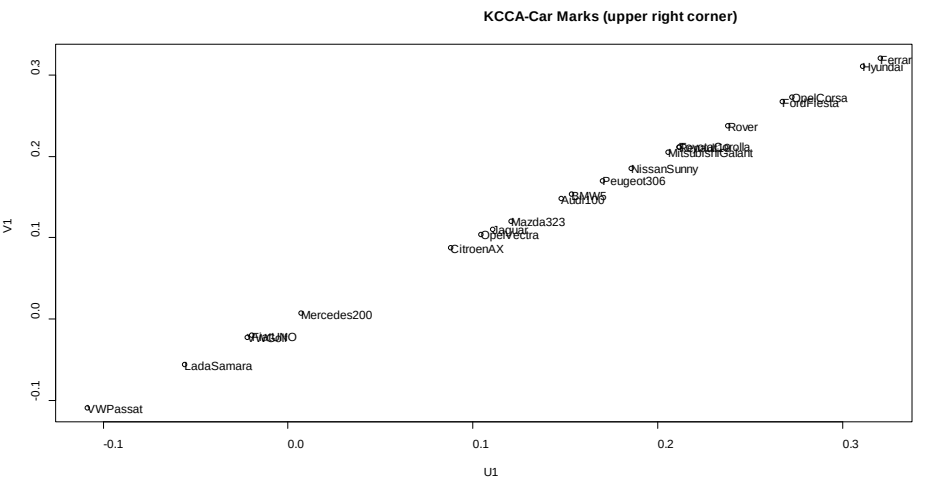
For the largest correlation coefficient ρ_1 we obtain vectors $\boldsymbol{a}_1' = [0.0664, \ldots, 0.0536] \in \mathbf{R}^{28}$ and $\boldsymbol{b}_1' = [-0.0234, \ldots, 0.2440] \in \mathbf{R}^{28}$ which correspond to the canonical variables (U_1, V_1), where:

$$(U_1^i)_{i=1,\ldots,N} = \boldsymbol{a}_1' \boldsymbol{K}_X$$
$$(V_1^i)_{i=1,\ldots,N} = \boldsymbol{b}_1' \boldsymbol{K}_Y$$

A projection (KCCA) into the coordinate system of the canonical variables corresponding to the canonical coefficients ρ_1 is shown below:



Because in this image the elements in the upper right corner are not clearly visible, we present below a magnification of the upper part:

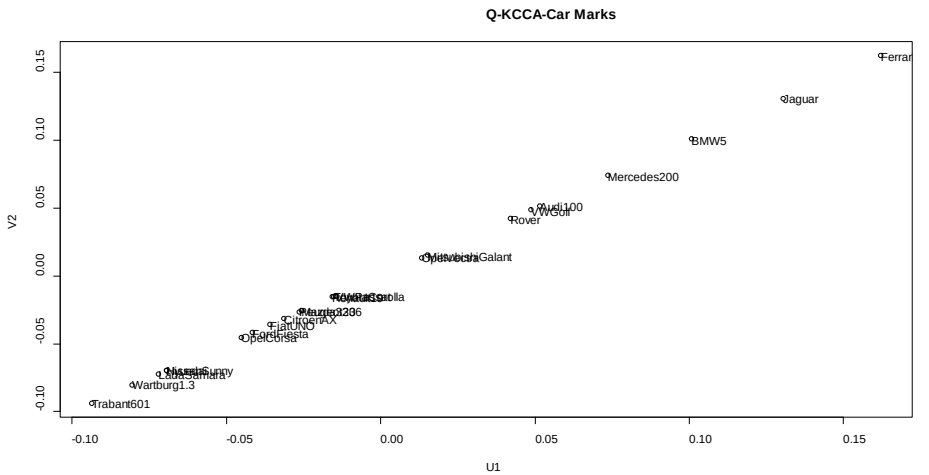


Finally, in the quasi kernel case (Q-KCCA) we obtain two non-zero correlation coefficients $\rho_1 = 0.99995$ and $\rho_2 = 0.99935$. For the largest correlation coefficient ρ_1 we obtain vectors $\boldsymbol{a}_1' = [-0.0916, \ldots, 0.0153] \in \mathbb{R}^{23}$ and $\boldsymbol{b}_1' = [-0.0101, -0.0774] \in \mathbb{R}^2$ which correspond to the canonical variables (U_1, V_1), where:

$$(U_1^i)_{i=1,\ldots,N} = \boldsymbol{a}_1' \boldsymbol{K}_X$$

$$(V_1^i)_{i=1,\ldots,N} = (\boldsymbol{b}_1' \boldsymbol{Y}_i)_{i=1,\ldots,N}$$

A projection (Q-KCCA) into the coordinate system of the canonical variables corresponding to the canonical coefficients ρ_1 is shown below:



7. Conclusions

Kernel methods compared to the classical methods give a correlation coefficient close to 1, i.e. two data sets can be presented as 100% correlated data sets. In the discussed example it can be seen that the largest correlation coefficients for Q-KCCA and KCCA are similar (respectively 0.99995 and 1.00000). However, intuitively the projection into the coordinate system of the canonical variables (corresponding to the largest canonical coefficients) in the case of Q-KCCA is more similar to CCA then KCCA. Therefore, Q-KCCA can be recommended as one of the best methods in nonlinear canonical correlation analysis.

REFERENCES

ARONSZAJN, N. (1950): Theory of reproducing kernels, Transactions of the American Mathematical Society 68, 337–404.

HARDLE, W., SIMAR, L. (2007): Applied Multivariate Statistical Analysis, Springer, 321-330 and 434-435.

HOTELLING, H. (1936): Relation between two sets of variates, Biometrika 28, 321-377.

PREDA, C. (2006): Regression models for functional data by reproducing kernel Hilbert spaces methods, Journal of Statistical Planning and Inference 137, 831.

ZHENG, W., ZHOU, X., ZOU, C., ZHAO L. (2006): Facial Expression Recognition Using Kernel Canonical Correlation Analysis, IEEE Transaction on Neural Networks 17(1), 233.

STATISTICS IN TRANSITION-new series, Summer 2012
Vol. 13, No. 2, pp. 311—320

INDEPENDENCE ANALYSIS OF NOMINAL DATA WITH THE USE OF LOG-LINEAR MODELS IN R

Justyna Brzezińska[1]

ABSTRACT

Log-linear models are used to analyze the relationship between two or more categorical (e.g. nominal or ordinal) variables. The term log-linear derives from the fact that one can, through logarithmic transformations, restate the problem of analyzing multi-way frequency tables in terms that are very similar to ANOVA. Specifically, one may think of the multi-way frequency table to reflect various main effects and interaction effects that add together in a linear fashion to bring about the observed table of frequencies. There are several types of models between dependence and independence: homogenous association, partial association, conditional association and null model. Expected cell frequencies are obtained with the use of iterative proportional fitting algorithm (IPF) [Deming, Stephen 1940]. The next step is to derive model coefficients for single variables as well as for interaction parameter and the most useful tool for interpreting model parameter is odds and odds ratio. Log-linear models are available in **R** software with the use of `loglm` function in `MASS` library and `glm` function in `stats` library. In this paper log-linear analysis will be presented with the use of available packages on empirical datasets in economic area.

Key words: Log-linear models, cross-tabulation, qualitative data, independence analysis of nominal data.

1. Introduction

Frequency counts of categorical variables are probably the most frequently encountered variables of research. Categorical data analysis has a long history. First papers were published in 1900 [Pearson, Yule] and were focused on independence analysis for categorical variables in two-way tables. Later on it was developed into multi-way contingency tables [Haberman 1974, Bishop, Fienberg Holland 1975]. In the middle of the twentieth century log-linear analysis for nominal [Bishop, Fienberg, Holland 1975, Knoke, Burke 1980, Christensen 1997] and ordinal data [Ishii-Kunts 1994] developed successfully. Nowadays, with the

[1] University of Economics in Katowice.

use of speed computers and professional software, advanced techniques are used for visualizing data structure [Friendly 2000, Zeileis, Mayer, Hornik 2006].

Cross-classification tables are used to answer questions such as whether the relation exists between categorical variables or whether the relation between variable is different for different groups of subject. The usual method for analyzing cross-classified tables, no matter how many variables are considered, is to test relations between variables taken one pair at a time. Most often this is accomplished using the chi-square, Yule`s, Cramer`s and Pearson`s coefficient, however, for higher-way tables more general and wider method should be used. A more general strategy for the analysis of cross-classified categorical data involves testing several models, including not only the model of independence but also models that represent various types of association and interactions included. This strategy is called log-linear analysis and it is one of the most interesting tool for categorical data analysis. We do not differentiate between dependent and independent variable and all response variables are treated as factors. They can be used as an explorative model-building fashion to find the most parsimonious model that describes the data best, as well as for hypothesis testing with simultaneous testing of all possible factors combination.

This paper is concerned with the independence type in log-linear analysis and its application in economic research. It explains the appropriate use of the analysis, describes briefly the calculations involved, and finally illustrates applications of the method based on empirical set of data with the use of **R**.

2. Independence analysis and log-linear models for nominal data

The analysis of cross-classified categorical data has occupied a prominent place in statistics, but most of techniques were associated with the analysis of two-dimensional contingency tables and the calculation of chi-square or related statistics. During last years, the availability of high-speed computers has led to major development in the analysis of multidimensional tables. However, the excellent guides are already available [Bishop, Fienberg, Holland 1975, Cox 1970, Haberman 1974, Lindsey 1973, Plackett 1974], the analysis of higher dimensional tables has fully developed much later.

Categorical data consists of variable whose values comprise a set of discrete categories. Such data requires different statistical methods from those commonly used for quantitative data. The aim of this paper is to provide theoretical and practical methods designed to reveal patterns of relationship among nominal variable with the use of log-linear analysis. The term log-linear derives from the fact that one can, through logarithmic transformations, restate the problem of analyzing multi-way frequency tables in terms that are very similar to ANOVA. Specifically, one may think of the multi-way frequency table to reflect various main effects and interaction effects that add together in a linear fashion to bring about the observed table of frequencies. Log-linear models are used for modelling

cell counts in multi-way contingency tables and it allows one to distinguish several types of association: conditional independence model, joint independence model, homogenous association model, partial independence model, saturated or null model.

In a three-way table with nominal variable X, Y and Z, several types of potential independence can be presented, for example homogenous association, joint independence, conditional independence, partial independence, etc. Saturated model includes all the possible effects in multiplicative form for three variables given as:

$$m_{hjk} = \eta \tau_h^X \tau_j^Y \tau_k^Z \tau_{hj}^{XY} \tau_{hk}^{XZ} \tau_{jk}^{YZ} \tau_{hjk}^{XYZ} . \quad (1)$$

By taking the natural logarithms we have additive equation given as:

$$\log\left(m_{hjk}\right) = \lambda + \lambda_h^X + \lambda_j^Y + \lambda_k^Z + \lambda_{hj}^{XY} + \lambda_{hk}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{hjk}^{XYZ} . \quad (2)$$

Saturated model reproduces perfectly the observed cell frequencies through the theoretical frequencies but such model is meaningless since the aim is to find a more parsimonious model with less parameters. In order to find the best model from a set of possible models, some additional measures can also be considered. Fitting a log-linear model is a process of deciding which association terms are significantly different from zero. These terms are included in the model that is used to explain the observed frequencies. Terms that are excludes from the model go into the residual or error term, which reflects the overall badness-of-fit of the model [Friendly 2000]. The goal of the analysis is to find a small model (with fewer association terms) that nonetheless achieves a reasonable fit.

A rule of thumb to determine the degrees of freedom is df = number of cells – number of free parameters [Agresti 2002]. With the use of backward elimination method the starting point is saturated model. Thus, the aim of a researcher is to find a reduced model. A reduced model is a more parsimonious model with fewer parameters and thus fewer dependencies and effects. The hierarchy principle reveals that a parameter of lower order cannot be removed when there is still a parameter of higher order that concerns at least one of the same variables.

The most useful statistic used to test the goodness of fit for log-linear model is the likelihood ratio statistic [Christensen 1997]:

$$G^2 = 2\sum_{h=1}^{H}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{hjk} \ln\left(\frac{n_{hjk}}{m_{hjk}}\right). \quad (3)$$

Therefore, larger G^2 values indicate that the model does not fit the data well and thus the model should be rejected.

Akaike information criterion [Akaike 1973] refers to the information contained in a statistical model according to equation:

$$AIC = G^2 - 2df . \quad (4)$$

Another information criterion is Bayesian information criterion [Raftery 1986]:

$$BIC = G^2 - df \cdot \ln n . \tag{5}$$

The model that minimizes AIC and BIC will be chosen.

The likelihood ratio can also be used to compare an overall model within a smaller, nested model. The equation is as follows:

$$\Delta G^2 = G_1^2 - G_2^2 , \tag{6}$$

with: $\Delta df = df_2 - df_2$ degrees of freedom. If the ΔG^2 comparison statistic is not significant, then the nested model (1) is not significantly worse than the saturated model (2). There are several models which appear to provide an adequate fit, therefore, the more parsimonious (nested) model will be chosen. Models are getting reduced through the hierarchy principle. The hierarchy principle reveals that a parameter of lower order cannot be removed when there is still a parameter of higher order that concerns at least one of the higher orders [Knoke, Burke 1980]. We obtain MLEs for elementary cells under any hierarchical model by iterative proportional fitting of the sufficient configuration [Bishop, Fienberg, Holland 1975]. An acceptable model is the one whose expected cell frequencies do not significantly differ from the observed data. Although the model has been described for three-dimensional tables, the extension to higher dimensional tables is straightforward.

3. Application in R

Empirical example presented in this paper is based on housing market dataset on Copenhagen housing conditions (library(MASS), data(housing)) with sample size of 1681. Multi-way table contains four categorical data [Cox, Snell 1984, Madsen 1976] (Table 1).

Table 1. Variables in the analysis

Variable	Factor levels
Sat (S)	Satisfaction of householders with their present housing circumstances (High, Medium. Low)
Infl (I)	Perceived degree of influence householders have on the management of the property (High, Medium, Low)
Type (T)	Type of rental accommodation (Tower, Atrium, Apartment, Terrace)
Cont (C)	Contact residents are afforded with other residents (Low, High)

*Source: **R** software, library(MASS), data(housing).*

We will use the loglm() function in the library(MASS) to fit log-linear models. From a different perspective equivalent models can also be fit as a generalized linear models with the use of glm(…,family=poisson) function in stats package.

Suppose for a four-dimensional table $H \times J \times K \times L$ with the total sample size N and that the cell count for $hjkl-$cell is n_{hjkl}. Given the large number of possible hierarchical models that can be fit to multidimensional table, it is reasonable to ask whether a systematic approach to model selection is possible. Many different approaches has been proposed, but none of them entirely satisfactory [Fienberg 1980]. In this paper stepwise procedures will be presented with a significance level 0.15. First, following the hierarchy principle and starting from saturated model $[SITC]$, the goodness of fit statistic G^2 with corresponding $p-$value are computed for one-, two- and three-ordered level of interaction (Table 2).

Table 2. Goodness of fit statistics

Model	G^2	df	$p-$value
S.I.T.C	295.352	63	0,000
SI.ST.SC.IT.IC.TC	43.9518	40	0.308
SIT.SIC.STC.ITC	5.94433	12	0.919

Source: own calculations in ***R.***

Concerning $\alpha-$level of 0.15 only models $[SI][ST][SC][IT][IC][TC]$ and $[SIT][SIC][STC][ITC]$ (and the saturated model) fit the data. However, the likelihood ratio can be used for comparing two nested models against each other. Comparison of nested models can be done using ANOVA method.

```
Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1   43.951777 40
Model 2    5.944334 12  38.007443        28        0.09826
Saturated  0.000000  0   5.944334        12        0.91886
```

According to the output, the only model that fits data well is model `Model 2` called model of conditional independence $[SIT][SIC][STC][ITC]$. Its deviance is close enough to the deviance for the saturated model to give a non-significant ($p > 0.15$) $p-$value. However, we would prefer a simpler model than all three-way interactions. It is possible to fit all-pairwise models with the addition of one three-way interaction $[SIT]$, $[SIC]$, $[STC]$ and finally $[ITC]$.

Table 3. Goodness of fit statistics

Model	G^2	df	$p-$value
SI.ST.SC.IT.IC.TC.SIT	22.132	82	0.775
SI.ST.SC.IT.IC.TC.SIC	42.286	36	0.218
SI.ST.SC.IT.IC.TC.STC	31.783	34	0.577
SI.ST.SC.IT.IC.TC.ITC	38.662	34	0.267

Source: own calculations in ***R.***

For all possible models with two-way interaction plus one three way interaction $p-$value is non-significant, and those models are simpler than three-way interactions model, so it is very difficult to choose the best fitting model. The comparison of goodness of fit statistics for all good models will be done again with the use of ANOVA method.

```
           Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1   43.951777 40
Model 2   42.285862 36   1.665915         4        0.79690
Model 3   31.783060 34  10.502802         2        0.00524
Model 4   38.662216 34  -6.879155         0        1.00000
Model 5   22.131810 28  16.530406         6        0.01117
Model 6    5.944334 12  16.187476        16        0.43995
Saturated  0.000000  0   5.944334        12        0.91886
```

Concerning $\alpha-$level of 0.15 and following the hierarchy principle, the difference in fit of model 1 and model 2 is non-significant (P(> Delta(Dev)= 0.79690); moreover Delta(df)=4 is close enough to Delta(df)= 1.6659 so model 1 $[SI][ST][SC][IT][IC][TC]$ as mode parsimonious, nested model should be the appropriate model fitting best. It means that this model includes the effect of single variables, as well as all possible to-way interactions on cell counts. It also means that none of three-way interaction $[SIT]$, $[SIC]$, $[STC]$ or $[ITC]$ have influence on cell counts. Satisfaction of householders with their present housing circumstances (S) depends on perceived degree of influence householders have on the management of the property (I), type of rental accommodation (T) and contact residents are afforded with other residents (C) individually, what is proved by interactions $[SI][ST][SC]$. Furthermore, perceived degree of influence householders have on the management of the property (I) depends on type of rental accommodation (T) and contact residents are afforded with other residents (C) individually: $[IT][IC]$. Finally, the type of rental accommodation (T) depends on contact residents are afforded with other residents (C): $[TC]$. The model equation for relationship $[SI][ST][SC][IT][IC][TC]$ can be described as:

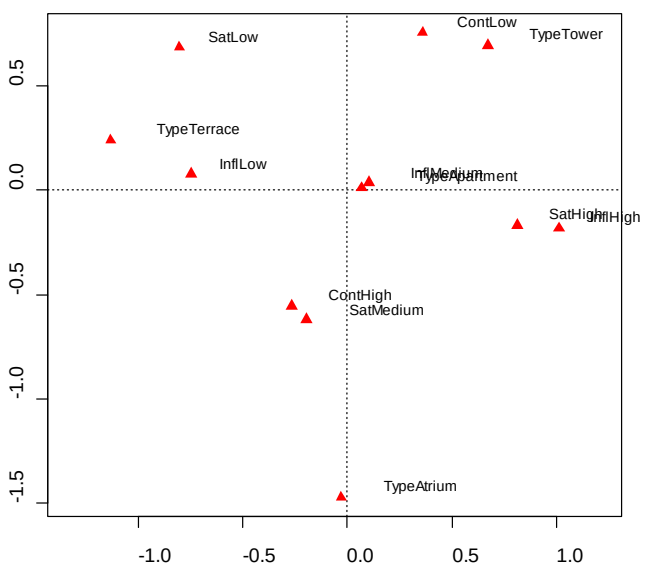
$$\log\left(m_{hjkl}\right)=\lambda+\lambda_h^S+\lambda_j^I+\lambda_k^T+\lambda_l^C+\lambda_{hj}^{SI}+\lambda_{hk}^{ST}+\lambda_{hl}^{SC}+\lambda_{jk}^{IT}+\lambda_{jl}^{IC}+\lambda_{kl}^{TC}. \qquad (7)$$

Sometimes it is convenient to interpret only the model equation and possible interactions, as with many variables included it might be difficult to interpret every single parameter for every cell. Parameters give information on each cell, if modelled cell is greater or smaller than empirical cell.

Another method widely used for categorical data analysis is correspondence analysis (CA) for two categorical variables and multiple correspondence analysis (MCA) for several variables. The method is particularly helpful in analyzing cross-tabular data in the form of numerical frequencies, and results in an elegant but simple graphical display which permits more rapid interpretation and understanding of data [Greenacre 1993]. Correspondence analysis is a statistical method for picturing the associations between the levels of a two- or multi-way contingency table. The observed association is done by cell frequencies, and typical inferential aspect is the study of whether certain levels of one characteristic are associated with some levels of another. Correspondence analysis is a geometric technique for displaying the rows and columns of a contingency table as a points in low-dimensional space. The main goal of the method is a global view of the data that is useful for interpretation with the use of perception map.

Symmetric map of the data housing, using the multiple correspondence analysis (MCA) in `ca` package is presented below.

Figure 1. Perception map: multiple correspondence analysis (Inertia=0.024)



*Source: own calculations in **R**.*

Total inertia is 0.024, so there is no relationship between variables. It is also seen that it would be quite difficult to group the points into well separated groups or clusters. Eigenvalue for first and second axis are: 59% and 8.2%, and it means that over 67.2% of the association between satisfaction of householders with their present housing circumstances (S), influence householders have on the management of the property (I), type of rental accommodation (T) and contact residents are afforded with other residents (C) can be represented well in two dimensions. This perception map might be helpful, when it is difficult to see any type of relationship, especially for multi-way tables. However, in this case log-linear analysis provided much more detailed information on data structure and pattern of the relationship.

With the use of vcd package it is also possible to present the data structure graphically with the use of mosaic and sieve plot [Friendly 2000].

Figure 2. Mosaic plot for housing data.

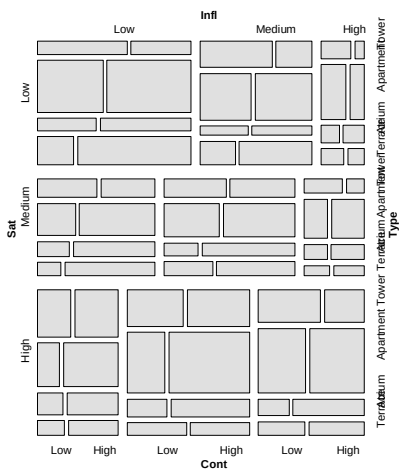
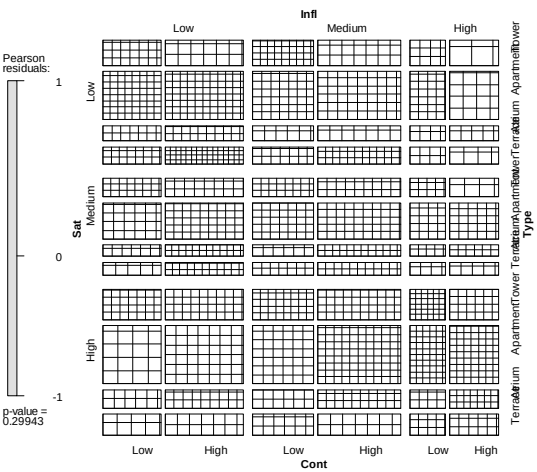


Figure 3. Sieve plot for housing data.



*Source: own calculations in **R**.*

The mosaic plot (Figure 2) and sieve plot (Figure 3) indicate that the model $[SI][ST][SC][IT][IC][TC]$ fits well (residuals in mosaic plot, as well as differences between observed and theoretical counts in sieve plot, are small).

The advantage of log-linear models is that they can be used for any number of categorical variables and there are no limits with the table dimension. Different types of association can be distinguished which gives more detail information about relation between variables and visualization methods can be applied (ca, vcd, vcdExtra).

4. Conclusions

Log-linear analysis is a method of statistical analysis that is used when all the variables of interest are categorical. The method has had wide use in the economic, psychological and social sciences and has several advantages. In log-linear analysis there is no dependent variable that can be predicted, instead cell frequencies are modelled. Log-linear models can easily incorporate more than two categorical variables and are useful for patterns of association analysis. We can also use them to identify different types of independence.

The purpose of this paper was to provide an outline of log-linear models and its application in economic research. The `loglm` package in **R** was used as well as multiple correspondence analysis (`ca`) and visualizing tools for multi-way tables (`vcd`, `vcdExtra`) to facilitate interpretation of results. In the paper the log-linear analysis was presented using dataset Copenhagen housing conditions. The advantages to be gained from the model-fitting techniques are that they provide a systematic approach to the analysis of multidimensional tables and also estimates of the magnitude of effects interest. The method allows for distinguishing different types of association: saturated model, complete (mutual) independence, joint independence, conditional independence or homogeneous association. However, Agresti [2002] says: "*there is no guarantee that either strategy will lead to a meaningful model*". We should also look for a model that is simple to interpret and smoothes rather than overfits the data.

REFERENCES

AGRESTI, A., 2002. Categorical Data Analysis, Wiley & Sons, Hoboken, New Jersey.

AKAIKE, H., 1973. Information theory and an extension of the maximum likelihood principle, in: Proceedings of the 2nd International Symposium on Information, Petrow B. N., Czaki F., Budapest: Akademiai Kiado.

BISHOP, Y. M. M., FIENBERG E. F., HOLLAND P. W., 1975. Discrete Multivariate Analysis, MIT Press, Cambridge, Massachusetts.

CHRISTENSEN, R., 1997. Log-Linear Models and Logistic Regression, Springer–Verlag, New York.

COX, D. R., 1970. Analysis of Binary Data, London, Mathuen.

COX, D. R. and SNELL, E. J., 1984. Applied Statistics, Principles and Examples. Chapman & Hall.

FIENBERG, S., 1980. The analysis of cross-classified categorical data, MIT Press, Cambridge.

FRIENDLY, M., 2000. Visualizing categorical data, SAS Institute Inc.

GREENACRE, M. J., 1993. Correspondence analysis in practice, London Academic Press.

HABERMAN, S. J., 1974. The Analysis of Frequency Data, Chicago, University of Chicago Press.

HORNIK, K., MAYER D., ZEILEIS A., 2006. The strucplot framework: visualizing multi-way contingency tables with vcd, Journal of Statistical Software, 17 (3), 1-48.

ISHII-KUNTS, M., 1994. Ordinal log-linear models, Sage University Papers.

KNOKE, D., BURKE P. J., 1980. Log–linear Models, Quantitative Applications in the Social Science" 20, Sage University Papers, Sage Publications, Newbury Park, London, New Delhi.

LINDSEY, J. K., 1973. Inferences from Sociological Survey Data: A Unified Approach, New York, Elsevier.

MADSEN, M., 1976. Statistical analysis of multiple contingency tables. Two examples. Scand. J. Statist. 3, 97–106.

PEARSON, K., 1900. On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, Philos. Mag. Ser. 5, 50, 157-175.

PLACKETT, R. L., 1974. The Analysis of Categorical Data, London, Griffin.

RAFTERY, A. E., 1986. A note on Bayesan Factors for log-linear contingency table models with vague prior information, Journal of the Royal Statistical Society, Ser. B, 48, 249-250.

YULE, G. U., 1900. On the association of attributes in statistics, Phil. Trans. Ser. Q 194, 257-319.

STATISTICS IN TRANSITION-new series, Summer 2012
Vol. 13, No. 2, pp. 321—334

INCOME SATISFACTION AND RELATIVE DEPRIVATION

Hanna Dudek[1], **Joanna Landmesser**[2]

ABSTRACT

The main objective of the study is to identify determinants of income satisfaction in Poland. For this purpose, income situation of households in relative terms is analyzed. The effects of the relative deprivation on income satisfaction in various socio-demographic groups of households are also examined. The method of partially generalized ordered logit models is used in the paper. The empirical investigation is based on data from Household Budget Survey carried out by the Polish Central Statistical Office in 2009.

Key words: Income Satisfaction, Relative Deprivation, Partially Generalized Ordered Logit Model.

1. Introduction

Economic research on income satisfaction is sparse but growing rapidly. Many studies of household wealth in developed countries are not limited to the analysis of the consumption and the objective income condition. They take into account the issue of deprivation covering many areas of life including subjective deprivation. Understanding the determinants of income satisfaction may help in the creation of the social policy aimed at mitigating the effects of subjective poverty.

The main objective of the study is to identify determinants of income satisfaction in Poland. For this purpose, the income situation of households, both in absolute and relative terms, is analyzed. Some authors suggest that the natural relationship is greater between subjective well-being and relative deprivation rather than between subjective well-being and income itself (Easterlin, 1995). Relative deprivation is a concept assuming that people compare themselves with other individuals or groups when evaluating their own situation. Therefore, for

[1] Warsaw University of Life Sciences, Faculty of Applied Informatics and Mathematics, Nowoursynowska 159, 02-776 Warsaw, Poland. E-mail: hanna_dudek@sggw.pl.
[2] Warsaw University of Life Sciences, Faculty of Applied Informatics and Mathematics, Nowoursynowska 159, 02-776 Warsaw, Poland. E-mail: joanna_landmesser@sggw.pl.

each household the value of the relative deprivation function is computed in the form proposed by D'Ambrosio and Frick (D'Ambrosio, Frick, 2007). Regressors in the econometric model include not only relative deprivation index, but also the other socio-demographic characteristics such as gender, age, marital status, and dummies for employment status. The effects of relative deprivation on income satisfaction in various socio-demographic groups of households are examined.

The method of partially generalized ordered logit models is used. The empirical investigation is based on data from household budget surveys carried out by the Central Statistical Office (CSO) in 2009. The methodology proposed makes a modification of the approach presented in (Vera-Toscano et al., 2006) and (Ferrer-i-Carbonell, 2005), where to estimate the model explaining income satisfaction the authors used ordered logit and probit models without verifying strong assumptions imposed on these models.

2. Income satisfaction researches – review of empirical research

Since the end of the last century more and more work on the explanation of the satisfaction with income have begun to appear in the scientific literature in economics and sociology. Individual satisfaction is the satisfaction derived from income, with evidence showing that age, education and individual income appear to have significantly positive impact on individual's income satisfaction. The set of factors considered as potential explanatory variables in income satisfaction research may be divided into two groups: attributes of the household head and attributes of the whole household. The first group encompasses characteristics such as: age of the family head, gender of the household head, the level of education of the household head. The second group comprises, for instance, the following attributes: household disposable income, the number of household members, the place of residence (rural, small towns, large cities).

Many studies have found a negative correlation between age and subjective well-being, but only up until to a certain age (Ferrer-i-Carbonell, Van Praag, 2003; D'Ambrosio, Frick, 2007; Vera-Toscano et al., 2006; Van Praag et al., 2010). The relationship is U-shaped and has its turning point around a certain age and, after this point, subjective welfare is likely to increase with age. In the works cited (in addition to various socio-demographic characteristics) the level of current income in absolute or relative terms is used to explain the satisfaction with income. The paper (D'Ambrosio, Frick, 2007) explains the perception of the income situation of households by their own relative deprivation.

A personal determinant such as gender is also significant for subjective well-being. Some research studies prove males to be less satisfied than females (for example, (Van Praag et al., 2000, 2003) for Germany). Ulman has observed a positive impact of education on financial satisfaction - the higher the level of education, the less subjective poverty (Ulman, 2006).

The point of view approved by economists is that household income and size exert statistically significant influence on measures of subjective economic well-being. Having considered the type of residence, it is stated that there is no firm opinion.

A few papers on the analysis of Polish households' subjective assessment of their own income situation are (Liberda et al., 2011), (Dudek, 2009) and (Dudek, 2012). In the first one it is presented that the determinants of income satisfaction include age, gender and educational level of household heads, the source of income, the place of residence and affiliation to the income group. The main objective of the (Dudek, 2009) and (Dudek, 2012) studies was to estimate the equivalence scale so as to include explanatory variables: the logarithms of income, the number of persons in the household as well as age, sex, educational level, the fact of being in formal or informal relation to the household head, and the place of residence.

The literature lacks modelling of the impact of relative deprivation on subjective assessment of personal income of Polish households. This paper aims to fill in this gap. One of the purposes of this study is to examine the impact of various potential determinants of income satisfaction in Polish households.

3. The data

Data employed in this study come from the Household Budget Survey (HBS) carried out by the Central Statistical Office in 2009. The observation unit is a household. One-person household is defined as a person not sharing his/her income with any other person, whether living alone or not. Multi-person household is defined as a group of people living together and sharing their incomes and expenditures. The Household Budget Survey does not contain any information referring to households from collective homes, such as students' hostels, social welfare homes (the so-called collective households) as well as households of the diplomatic corps of foreign countries. The households of foreign citizens speaking Polish with permanent or long-lasting residence in Poland are included in the survey. The number of households participating in the survey in each year was about 30000. The monthly rotation of households implemented assumes that every month of the year a different group of households participated in the survey (*Household Budget Surveys in 2009*).

The study focuses on households of employees whose exclusive or prevailing source of livelihood is income from employment in either the public or private sector. Subjective measures are based on households' answers to the question: "Considering your monthly disposable income, is your household able to make ends meet: (1) with great difficulty, (2) with difficulty, (3) with some difficulty, (4) without difficulty, (5) with ease, (6) with great ease?"

Table 1 presents the structure of Polish employees' households according to the assessment of subjective income situation.

Table 1. Structure of households by categories with respect to income perception in 2009.

Level of income satisfying household needs	Category	The percentage of employees' households
Very poor	*j*=1	6.72%
Poor	*j*=2	15.45%
Insufficient	*j*=3	42.30%
Scarcely enough	*j*=4	26.89%
Good	*j*=5	6.89%
Very good	*j*=6	1.76%

Source: Own calculations based on 2009 Household Budget Survey data.

Due to the small number of households declaring their situation as very good, they are joined with those assessing their income position as good. Therefore, in the econometric analysis five levels (categories) of income assessment are taken into account.

4. Econometric Framework

In order to estimate the impact of relative deprivation on income satisfaction the so-called equivalent income was considered at the beginning, taking into account the modified OECD scale. The disposable income of each household was divided by the corresponding value of equivalence scale, yielding $y_1, y_2, \ldots, y_n$, where y_i – the equivalent income of *i*-th household, $i = 1, 2, \ldots, n$, $n = 18240$ – number of households of employees in the sample. The equivalence scale values were calculated as follows: 1+0,5(*a*-1)+0,3*c*, where *a* – the number of adults, *c* – the number of children under 14 years of age in the household. This formula is applied to the data from the 2009 in CSO publications. Equivalence scales allow one to compare the situation of households of varying size and demographic structure. They reflect the influence of household demographic structure on its living costs.

The equivalent incomes were arranged in non-decreasing order obtaining $y_{(1)} \le y_{(2)} \le \ldots \le y_{(n)}$. For each household with income $y_{(i)}$ the value of the relative deprivation function is computed as follows:

$$d(y_{(i)}) = \frac{1}{n} \sum_{j=i+1}^{n} (y_{(j)} - y_{(i)}), \quad d(y_{(n)}) = 0 \text{ (D'Ambrosio, Frick, 2007).} \qquad (1)$$

The above formula shows that the higher the equivalent income, the lower the value of the relative deprivation of income. Setting the households from the smallest to the largest value of relative deprivation the same rank ordering is obtained as in terms of decreasing equivalent income (from largest to smallest income value). One of the first papers presenting the idea of relative deprivation in a descriptive way is the monograph (Runciman, 1966). The quantification of this concept was presented in the work (Yitzhaki, 1979).

Preliminary data analysis revealed subjective assessment of own income from relative deprivation. Basic information on this subject is provided in Table 2.

Table 2. Subjective assessment of own income and relative deprivation.

Level of income satisfying household needs	Category	Value of relative deprivation		
		Median	Mean	Standard deviation
Very poor	*j*=1	974.10	964.92	335.05
Poor	*j*=2	811.38	810.81	331.53
Insufficient	*j*=3	602.56	627.452	315.18
Scarcely enough	*j*=4	351.54	402.58	271.77
Good or very good	*j*=5	157.86	217.31	208.55

Source: Own calculations based on 2009 Household Budget Survey data.

To compare the strength of the relationship between subjective assessment of income, relative deprivation and equivalent income in the study of D'Ambrosio and Frick (D'Ambrosio, Frick, 2007), the Pearson correlation coefficients were used. Just as in the present analysis a stronger relationship was found between the subjective assessment of income and relative deprivation than between the subjective assessment of income and equivalent income.

To analyze the formation of income satisfaction in the paper (D'Ambrosio, Frick, 2007) the linear models for panel data were used. In works (Ferrer-i-Carbonell, Van Praag, 2003) (Schwarze, 2003) (Stanovnik, Verbič, 2006), (Vera-Toscano et al., 2006) more appropriate models were used to explain the variable

expressed at an ordinal scale. Subjective perception of income can be treated as a self-reported measure of utility. In order to explain it, the ordered logit model is applied. The starting point in such a case is usually a model with latent variable y^*:

$$y_i^* = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i,\ i = 1,2,\ldots,n, \tag{2}$$

where: y^* – an unobserved latent variable which represents the response, if it could be measured accurately on the continuous scale,

$\mathbf{x}_i$ – a row vector of explanatory variables representing the characteristics of individual *i*,

$\boldsymbol{\beta}$ – a column vector of parameters $\beta_1, \beta_2, \ldots, \beta_k$ to be estimated,

ε_i – a random component for the *i*-th observation,

n – a number of individuals,

the subscript *i* refers to the observation number.

Let us assume a set of cut-points $\delta_0, \delta_1, \ldots, \delta_m$, such that $-\infty = \delta_0 < \delta_1 < \ldots < \delta_m = \infty$, that divide $(-\infty, \infty)$ into *m* intervals. The relationship between the latent variable and the realized outcome is: $y_i = j$ if and only if

$$\delta_{j-1} < y_i^* \le \delta_j,\ i = 1,2,\ldots,h,\ j = 1,2,\ldots,m. \tag{3}$$

The $\delta_0, \delta_1, \ldots, \delta_m$ are unknown parameters to be estimated with $\beta_1, \beta_2, \ldots, \beta_k$. The substitution of (2) into (3) yields:

$$\delta_{j-1} - \mathbf{x}_i\boldsymbol{\beta} < \varepsilon_i \le \delta_j - \mathbf{x}_i\boldsymbol{\beta}. \tag{4}$$

It leads to the following probabilities of each outcome:

$$P(y_i = j \,|\, \mathbf{x}_i) = F(\delta_j - \mathbf{x}_i\boldsymbol{\beta}) - F(\delta_{j-1} - \mathbf{x}_i\boldsymbol{\beta}), \tag{5}$$

where F – cdf of iid error terms ε_i. In practical applications the following models are usually used:

- ordered logit model with $F(z) = \Lambda(z) = \dfrac{1}{1+\exp(-z)}$, (6)

- ordered probit model with $F(z) = \Phi(z) = \displaystyle\int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$. (7)

From an empirical point of view, it usually does not matter which model is used. Logit and probit models typically provide very similar results. This is all because the distribution functions for the logit and probit are comparable, differing slightly only in the tails of their respective distributions. In this research paper logit model is employed.

The slope parameters $\beta_1, \beta_2, \ldots, \beta_k$ have no intuitive interpretation. For the probabilities, the marginal effects of the regressors are:

$$\frac{\partial P(y_i = j \mid \mathbf{x}_i)}{\partial x_l} = -\beta_l \left\{ \frac{d\Lambda(\delta_j - \mathbf{x}_i\boldsymbol{\beta})}{dz} - \frac{d\Lambda(\delta_{j-1} - \mathbf{x}_i\boldsymbol{\beta})}{dz} \Big|_{z=\mathbf{x}_i\boldsymbol{\beta}} \right\} \tag{8}$$

The term in braces can be positive or negative, so one must be very careful in interpreting the slope parameters $\beta_1, \beta_2, \ldots, \beta_k$ in the ordered logit model. Only the signs of the changes in $P(y_i = 1 \mid \mathbf{x}_i)$ and $P(y_i = m \mid \mathbf{x}_i)$ are unequivocal. The marginal effects of the regressor x_l on the probabilities $P(y_i = 1 \mid \mathbf{x}_i)$ are:

$$\frac{\partial P(y_i = 1 \mid \mathbf{x}_i)}{\partial x_l} = -\beta_l \left\{ \Lambda(\delta_1 - \mathbf{x}_i\boldsymbol{\beta}) \left(1 - \Lambda(\delta_1 - \mathbf{x}_i\boldsymbol{\beta})\right) \right\}, \; l = 1, 2, \ldots, k. \tag{9}$$

As $\Lambda(1 - \Lambda) \geq 0$, the derivative of $P(y_i = 1 \mid \mathbf{x}_i)$ has the opposite sign from β_l. Similarly, as

$$\frac{\partial P(y_i = m \mid \mathbf{x}_i)}{\partial x_l} = \beta_l \left\{ \Lambda(\delta_{m-1} - \mathbf{x}_i\boldsymbol{\beta}) \left(1 - \Lambda(\delta_{m-1} - \mathbf{x}_i\boldsymbol{\beta})\right) \right\}, \tag{10}$$

the change in $P(y_i = m \mid \mathbf{x}_i)$ must have the same sign as β_l.

The parameters of ordered response model can be estimated by maximum likelihood method. *LR* test may be conducted for the selection between nested models. Akaike (AIC) and Bayesian (BIC) information criteria are used to compare alternative non-nested models. The model with smaller values of information criteria is preferred. Information criteria penalize models with additional parameters. Therefore, the AIC and BIC model order selection criteria are based on parsimony.

More information about the properties of ordered category models and their estimation and verification can be found in the works: (Dudek 2007), (Greene, Hensher 2010), (Książek 2010).

The models of ordered category are quite restrictive, because they assume that:

$$P(u_i \leq j \mid \mathbf{x}_i) = F(\delta_j - \mathbf{x}_i\boldsymbol{\beta}), \tag{11}$$

i.e. the parameters of the explanatory variables do not depend on the category j, $j = 1,2,\ldots,m$. Ordered logit model approach embodies the restriction that the parameters $\beta_1, \beta_2, \ldots, \beta_k$ are to be the same for all categories. This assumption is called the parallel regression assumption. In order to verify it we used the test proposed in the publication (Brant, 1990). The idea of this test is based on the consideration of the *m*-1 binary regression for variables:

$$u_j^{**} = \begin{cases} 1, & \text{if} \quad u > j \\ 0, & \text{if} \quad u \leq j \end{cases}, \tag{12}$$

where $j = 1,2,\ldots,m-1$ and the **β**'s are allowed to differ across regression. The parallel regression assumption implies that parameters slope vector **β** should be the same in every equation. The LR test is an omnibus test for all variables. It does not allow to determine whether the coefficients for some variables are identical across the binary equations while other coefficients are different. The Wald test developed by Brant allows to test the parallel regression assumption for each variable individually. This could be helpful in identifying individual variables that were problematic.

The following approach is presented in (Brant, 1990), (Long, 1997) and (Książek, 2010). After estimating the total variance and covariance matrix of parameters in all *m*-1 binomial models, Wald tests are carried out - total and individual tests for individual variables. The first test is used to verify the null hypothesis of equality of the relevant parameters for all m-1 binomial models jointly for the whole set of explanatory variables. The test statistic in the Brant test has the asymptotic chi-square distribution with *p* (*m*-2) degrees of freedom, where *p* - number of parameters by the explanatory variables, *m* - number of categories of an ordinal variable corresponding to the unobservable response variable. The rejection of the null hypothesis means that at least for one explanatory variable parameters significantly differ in at least two binomial models, following which the parallel regression assumption is not satisfied. Conducting individual Wald tests can identify the variables "responsible" for the violation of this assumption. The null hypothesis then says about the equality of parameters for a particular explanatory variable in all binomial models. As a consequence of rejecting the null hypothesis, the model for ordered categories should not be used, and then other methods to estimate parameters of the model (2) have to be applied. One of the approaches that can be used in this case is the estimation of a generalized model of an ordered category (Greene, Hensler, 2010):

$$P(u_i \leq j \,|\, \mathbf{x}_i) = F(\delta_j - \mathbf{x}_i \boldsymbol{\beta}_j), \text{ for } j = 1,2,\ldots,m-1. \tag{13}$$

In the model above, it is allowed that the parameters by the explanatory variables are dependent on the category of an ordinal variable corresponding to the unobservable response variable. In the literature, such models are referred to

as partial proportional odds models for ordinal response variables (Peterson i Harrell, 1990; Williams, 2006). If the individual Wald tests do not require the rejection of the hypothesis of some parameters equality, it is possible to use partially generalized ordered models. In models of this type, the parameters for some of explanatory variables do not depend on the category of ordinal variable. This approach is on the one hand less restrictive than the use of generalized ordered model but, on the other hand, it is more "parsimonious" and allows easier interpretation of structural parameters due to the inclusion of a smaller number of parameters in the model.

In this paper we review the parallel regression assumption and propose some solutions in this field. The following model was considered:

$$u^* = \alpha_0 + \alpha_1 y + \sum_{k=1}^{K} \gamma_k s_k \tag{14}$$

where: u^* – the income satisfaction (unobservable variable; only the level measured at an ordinal scale can be observed),
y – the value of the relative income deprivation,
s_k – k-th control variable, $k = 1, 2, \ldots, K$; e.g. such as gender, education level, age, marital status, place of residence,
$\alpha_0, \alpha_1, \gamma_1, \gamma_2, \ldots, \gamma_K$ – parameters.

We undertook to verify the hypothesis that the impact of relative deprivation on subjective assessment of income is the same in different groups defined by the above-mentioned features. For this purpose, models with different interactions between the explanatory variables were considered. For example, if y is the value of the relative deprivation and z refers to the dummy variable, such as gender, the expression $ay+bzy = y(a+bz)$ for $z = 0$ means an impact equal to ay and for $z = 1$ the impact will be $y(a+b)$. The final model included those interactions of deprivation index with socio-demographic variables that were statistically significant.

5. Results and discussion

We estimated a number of models explaining the formation of income satisfaction. The selection of variables was influenced by the substantive and statistical considerations. To compare models with different set of explanatory variables we used Akaike and Bayesian information criteria.

In the first step ordered logit models were considered. After performing the total Brant test we found out that the parallel regression assumption should be rejected. The results of individual Brant test showed that the "responsibility" for the violation of this assumption bear some variables.

Finally, we obtained model, for which verification results are presented in table 3. As quantitative explanatory variables were included: deprivation index

(label: *depryw*), the number of people (label: *LOS*), the number of children, age of reference person, and some interaction of these characteristics. Qualitative variables (listed in the fourth column of the table) are binary variables receiving a value of 1 for the indicated variant and 0 otherwise. We considered all possible interactions of deprivation index with socio-demographic variables. In the model we included only those that are statistically significant at the 0.05 level.

Table 3. Results of the Brant test for ordered logit model.

Variable	Value of the test statistic	p-value	Variable	Value of the test statistic	p-value
depryw	6.21	0.102	*higher education*	5.04	0.169
*depryw*LOS*	1.25	0.742	*secondary education*	4.99	0.172
*depryw*LOS*^{2}	1.55	0.671	*vocational education*	2.46	0.483
*depryw*country side*	2.01	0.571	*countryside*	3.33	0.343
number of children	4.83	0.183	*workers position*	5.44	0.142
age	0.12	0.990	*women*	5.81	0.121
age^2	0.20	0.991	*partnership*	12.07	0.007

Source: Own calculations made in the Stata v. 10.

The results presented in Table 3 were obtained from data on households, which consist of a maximum of 7 persons, for whom the values of equivalent income and deprivation index values were from the interval $[Q_1\text{-}2*IQR; Q_3+2*IQR]$, where Q_1 and Q_3 are the first and third quartile, IQR - interquartile range. Other observations, representing approximately 4.5% of the sample, were considered outliers and were excluded from the analysis.

For the whole set of variables presented in Table 3, the value of the total Brant test statistic was 152.92, which indicates the rejection of the parallel regression assumption (the critical value, read from the tables for the chi-square distribution for 42 degrees of freedom and significance level 0.05 is equal to 58.12). Information presented in Table 3 show that only for the variable relating to the person staying in the relationship, the hypothesis that the parameters do not depend on the category was rejected. Therefore, the partially generalized ordered model was used. The estimation results of the models parameters whose values do

not depend on the category of ordinal variable, which refers to the subjective assessment of income, are presented in Table 4.

Table 4. The results of parameters estimation in the partially generalized ordered logit model.

Variable	Coefficient	Standard error	Variable	Coefficient	Standard error
depryw	-0.0046	0.0002	*higher education*	0.7185	0.0695
*depryw*LOS*	0.0004	0.0001	*secondary education*	0.3931	0.0568
*depryw*LOS*^{2}	-0.00003	0.0000	*vocational education*	0.1955	0.0531
*depryw*countryside*	0.0007	0.0001	*countryside*	-0.2031	0.0639
number of children	-0.0937	0.0214	*workers position*	-0.1833	0.0374
age	-0.0563	0.0079	*women*	-0.2832	0.0341
age^2	0.0006	0.0001	-	-	-

Source: Own calculations made in the Stata v. 10.

Estimates of other parameters of the partially generalized ordered logit model, which differ in each category, are given in Table 5.

Table 5. The results of parameters estimation in the partially generalized ordered logit model, cont.

Variable	For u>1	For u>2	For u>3	For u>4
partnership	0.4339 (0.0774)	0.3884 (0.0555)	0.1959 (0.0510)	0.3918 (0.0867)

Source: Own calculations made in the Stata v. 10, standard errors in the parentheses.

Some important results can be derived under *ceteris paribus* assumption:

- the higher the education level of the reference person (head of household), the greater the likelihood that the disposable income of the household allows to make ends meet with ease or with great ease;
- the probability of higher levels of income satisfaction of household whose reference person is a blue-collar worker was lower than in the case of white-collar workers;
- the probability of higher levels of income satisfaction of household whose reference person is a women was lower than in the case of a man;

- the probability of higher levels of income satisfaction of household in the countryside was lower than in the case of household in a town;
- if the reference person remained in a partnership (formal or not), then the probability of good or very good self-assessment of the income situation was greater than that of a person not being in such a relationship;
- the likelihood of higher levels of income satisfaction initially decreased with the age of household head and increased with the age over 50.

The influence of relative deprivation on the income satisfaction depended on the place of residence and the number of persons living in the household. The results on this issue are presented in table 6.

Table 6. The calculation of parameters for the influence of relative deprivation on the income satisfaction.

Place of residence	Number of persons living in the household						
	1	2	3	4	5	6	7
rural area	-0.0036	-0.0032	-0.0030	-0.0028	-0.0027	-0.0026	-0.0026
town	-0.0042	-0.0039	-0.0036	-0.0035	-0.0033	-0.0033	-0.0033

Source: Own calculations made in the Stata v. 10.

For the age, education level, number of children *et cetera* assumed we have: $depryw*(-0.0046+0.0004*LOS-0.00003\ LOS^2+0.007*countryside)$.

6. Concluding remarks

This study improves the understanding of the mechanism underlying the expression of income satisfaction. The analysis allows us to conclude that household employees' perception of their own income situation in 2009 depended on many factors, in particular on their relative deprivation. Moreover, the following factors should be considered as determinants of subjective assessment of income: the place of residence, the number of children in the household as well as age, sex, education, type of employment and the fact of being in a partnership for the reference person (household head). Subjective economic assessment of being "very poor" was more frequent in households with female heads than in those with male ones. The income satisfaction was U-shaped with age. It was found that the impact of relative deprivation on the subjective assessment of income is different in households of different sizes and places of residence.

In the paper some problems of model specification were emphasized. The ordered response model makes the assumption that the explanatory variables of the model will have the same impact across each of the categories of the

dependent variable, which is known as the "parallel regression assumption". This assumption was tested with a Brant test. It was found that the "ordinary" ordered logit model should not be used to analyze the situation in a considered sample. Therefore, using the method of partially generalized ordered logit models allows the dependence of structural parameters on the category of ordinal variable (which determines the degree of income satisfaction).

REFERENCES

BRANT, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression, Biometrics, 46(4), 1171-1178.

D'AMBROSIO, C. and FRICK J. (2007). Income satisfaction and relative deprivation: An empirical link, Social Indicators Research, 81(3), 497–519.

DUDEK, H. (2007). An identification of farmers' households in danger of poverty on the ground of ordered logit model, [in:] K. Jajuga i M. Walesiak (eds.): Klasyfikacja i analiza danych – teoria i zastosowania, Taksonomia 14, Wydawnictwo AE we Wrocławiu, 367-375.

DUDEK, H. (2009). Subjective aspects of economic poverty – ordered response model approach, [in:] W. Ostasiewicz (ed.), Quality of Life Improvement through Social Cohesion, Research Papers of Wrocław University of Economics, 73, 9-24.

DUDEK, H. (2011). Skale ekwiwalentności – estymacja na podstawie kompletnych modeli popytu, Rozprawy i Monografie, nr 377, Wydawnictwo SGGW, Warszawa.

DUDEK, H. (2012). Subiektywne skale ekwiwalentności – analiza na podstawie danych o satysfakcji z osiąganych dochodów, [in:] K. Jajuga i M. Walesiak (eds.): Klasyfikacja i analiza danych – teoria i zastosowania, Taksonomia 14, Wydawnictwo AE we Wrocławiu, accepted for print.

EASTERLIN, R. A. (1995). Will raising the incomes of all increase the happiness of all?, Journal of Economic Behavior & Organization, 27, 35-47.

FERRER-I-CARBONELL, A. (2005). Income and well-being: An empirical analysis of the comparison income effect, Journal of Public Economics, 89, 997-1019.

FERRER-I-CARBONELL, A. and VAN PRAAG B. (2003). Income satisfaction inequality and its causes, The Journal of Economic Inequality, 1(2), 107-127.

GREENE, W. H. and HENSHER D. A. (2010). Modeling ordered choices: a primer, Cambridge University Press, Cambridge.

KSIĄŻEK, M. (2010). Modele zmiennych wielomianowych uporządkowanych, [in:] M. Gruszczyński (ed.): Mikroekonometria, Oficyna Wolters Kluwer Polska, Warszawa, 103-152.

LIBERDA, B., PĘCZKOWSKI M. and GUCWA-LEŚNY E. (2011). How do we value our income from which we save?, University of Warsaw, Faculty of Economic Sciences Working Papers, 3(43), 1-19.

LONG, J. S. (1997). Regression models for categorical and limited dependent variables, Sage Publications, Thousand Oaks.

PETERSON, B. and HARRELL, F. E. JR. (1990). Partial proportional odds models for ordinal response variables, Journal of the Royal Statistical Society, Series C, 39(2), 205-217.

RUNCIMAN, W. G. (1966). Relative deprivation and social justice: a study of attitudes to social inequality in twentieth-century England, University of California Press.

SCHWARZE, J. (2003). Using panel data on income satisfaction to estimate equivalence scale elasticity, Review of Income and Wealth, 49, 359-372.

STANOVNIK, T. and VERBIČ M. (2006). Analysis of subjective economic well-being in Slovenia, Eastern European Economics, 44(3), 60-70.

ULMAN, P. (2006). Subjective Assessment of Economic Poverty in Poland, 25th SCORUS Conference on Regional and Urban Statistics and Research "Globalization Impact on Regional and Urban Statistics", Wrocław, Poland, 30.08 – 1.09. 2006.

VAN PRAAG, B. M. S., FRIJTERS, P., FERRER-I-CARBONELL, A. (2000). A Structural Model of Well-Being: With an Application to German Data, Tinbergen Institute Discussion Paper, 00-053/3, Amsterdam.

VAN PRAAG, B. M. S, FRIJTERS, P., FERRER-I-CARBONELL, A. (2003). The Anatomy of Subjective Well-being, Journal of Economic Behavior and Organization, 51, 29-49.

VAN PRAAG, B. M. S., ROMANOV D., and FERRER-I-CARBONELL, A. (2010). Happiness and financial satisfaction in Israel: Effects of religiosity, ethnicity, and war, Journal of Economic Psychology, 31, 1008–1020.

VERA-TOSCANO, E., ATECA-AMESTOY, V. and SERRANO-DEL-ROSAL, R. (2006). Building financial satisfaction, Social Indicators Research, 77(2), 211-243.

WILLIAMS, R. (2006). Generalized ordered logit/partial proportional odds models for ordinal dependent variables, The Stata Journal, 6(1), 58-82.

YITZHAKI, S. (1979). Relative deprivation and the Gini coefficient, Quarterly Journal of Economics, 93, 321–324.

STATISTICS IN TRANSITION-new series, Summer 2012
Vol. 13, No. 2, pp. 335—342

ENSEMBLE APPROACH FOR CLUSTERING OF INTERVAL-VALUED SYMBOLIC DATA

Marcin Pełka[1]

ABSTRACT

Ensemble approach has been applied with a success to regression and discrimination tasks [see for example Gatnar 2008]. Nevertheless, the idea of ensemble approach, that is combining (aggregating) the results of many base models, can be applied to cluster analysis of symbolic data.
The aim of the article is to present suitable ensemble clustering based on symbolic data. The empirical part of the paper presents results simulation studies (based on artificial data sets with known cluster structure) of ensemble clustering based on co-occurrence matrix for symbolic interval-valued data, compared with single clustering method. The results are compared according to corrected Rand index.

Key words: Ensemble clustering; interval-valued symbolic data.

1. Introduction

Ensemble techniques based on aggregating information (results) from different models have been applied with a success in context of supervised learning (discrimination and regression). The ensemble techniques are applied in order to improve the accuracy and stability of classification algorithms (Breinman 1996).

Ensemble clustering means combining (aggregating) N base clustering results (models) $P_1,\ldots,P_N$ into one model P^* with k^* clusters (see: Fred and Jain 2005).

Recently several studies on combination method have established a new area in classical taxonomy. Nevertheless, the idea of ensemble approach, that is combining (aggregating) the results of many base models, can be applied to cluster analysis of symbolic data.

There are several proposals of applying the idea of ensemble approach in the context of clustering – aggregation of results of different clustering algorithms,

[1] Department of Econometrics and Computer Science, Wroclaw University of Economics. E-mail: marcin.pelka@ue.wroc.pl.

receiving different partitions by resampling the data, applying different subsets of variables, applying a given algorithm many times with different values of parameters or different initializations.

2. Symbolic data

Symbolic objects, unlike classical objects, can be described by many different symbolic variable types. Bock and Diday have defined five different symbolic variable types (Bock and Diday 2000, p. 2) – see table 1 for examples of symbolic variables:

1) single quantitative value,
2) categorical value,
3) quantitative value of interval type,
4) set of values or categories (multivalued variable),
5) set of values or categories with weights (multivalued variable with weights),
6) modal interval-valued variable proposed in Billard and Diday (Billard and Diday 2006).

Regardless of their type symbolic variables also can be the following (Bock and Diday 2000, p. 2):

1) taxonomic – which present prior known structure,
2) hierarchically dependent – rules which decide if a variable is applicable or not have been defined,
3) logically dependent – logical rules which affect variable's values have been defined.

Table 1. Examples of symbolic variables

Symbolic variable	Realizations	Variable type
preferred price of a new car (in PLN)	<25000; 36000>, <28000; 37000>, <30000; 50000>, <33000; 58000>, <65000; 80000>, <66000; 90000>	interval-valued (non-disjoint)
engine capacity	<1000; 1200>, (1200; 1400>, (1400; 1600>, (1600; 1800>, (1800; 2000>, (2000; 2200>	interval-valued (disjoint)
colour	{green, black, yellow, red, purple, blue}	multivalued
preferred brand of a car	{60% Honda, 35% Toyota, 5% Audi} {40% Honda, 20% Skoda, 20% Toyota, 20% Audi} {80% Audi, 15% Opel, 5% Toyota}	multivalued with weights

Source: Own research.

There are two main symbolic objects types:

1. First order objects (simple objects, individuals) – single respondent, product, company, etc., described by symbolic variable types. This objects are individuals that are symbolic by their nature.
2. Second order objects (aggregate objects, super individuals) – more or less homogeneous classes, groups of individuals described by symbolic variables.

3. Ensemble clustering methods

There are two main approaches that can be applied in ensemble learning for symbolic interval-valued data (see: Gathemi *et al.* 2009; De Carvalho *et al.* 2012; Hornik 2005):

1. Clustering algorithm for multiple relational matrices – proposed by De Carvalho *et al.* 2012. This approach is based on different distance matrices. Those distance matrices can be obtained by applying different distance measures, or subsets of variables or subsets of objects. Distance matrices are used to calculate relevance weight vectors. Relevance weight vectors and distance matrices are then applied to cluster a set of objects into k clusters.
2. Clustering ensemble that apply consensus functions in clustering ensembles. There are five main consensus functions that are applied in clustering ensemble.

Hypergraph partitioning which assumes that clusters can be represented as hyperedges on a graph. Their vertices correspond to the objects to be clusters. Each hyperedge describes a set of objects belonging to the same cluster. The problem of consensus clustering is reduced to finding the minimum-cut of a hypergraph (Gathemi *et. al.* 2009, p. 638; Strehl and Gosh 2002). Different adaptations of hypergraph partitioning have been proposed by Strehl and Gosh (2002), Fern and Brodley (2004), Ng *et al.* (2002).

The main idea of the **voting approach** is to permute cluster labels in such a way that best agreement between the labels of two partitions is obtained. All the partitions from the cluster ensemble must be relabelled according to a fixed reference partition. This reference partition can be taken from the ensemble or from a new clustering of the data set. Fisher and Buhman, and Dudoit and Fridlyand have presented a combination of partitions by relabeling and voting (Gathemi *et al.* 2009, p. 639).

Mutual information approach assumes that the objective function of a clustering ensemble can be formulated as the mutual information between the empirical probability distribution of labels in the consensus partition and the labels in the ensemble. In this approach usually a generalized definition of mutual information is applied – for example in Topchy *et al.* (2003). Luo *et al.* (2006) have introduced consensus scheme via genetic algorithm based on information theory. Azimi *et al.* (2007) have proposed clustering ensemble method which generates a new feature space from initial clustering outputs (Gathemi *et al.* 2009, p. 640).

In the **finite mixture model** approach the main assumption is that the output labels are modelled as random variables drawn from probability distribution described as a mixture of multinomial component densities. The objective of consensus clustering is formulated as a maximum likelihood estimation. Usually the expectation maximization algorithm (EM) is used to solve the maximum likelihood problem. Such approach is presented by Topchy *et al.* (2004), Analoui and Sadighian (2006) (Gathemi *et al.* 2009, p. 641).

The **co-association based functions** operate on the co-association (co-occurrence) matrix. Numerous clustering methods can be applied to co-association matrix to obtain the final partition. By applying different clustering methods, resampling the data, different subsets of variables, or the same clustering with different values of parameters or initializations we obtain N partitions (each can have different number of clusters) of set E (set of objects to be classified):

$$\begin{aligned} P^1 &= \left\{C_1^1, C_2^1, \ldots, C_{k_1}^1\right\}, \\ &\vdots \\ P^N &= \left\{C_1^N, C_2^N, \ldots, C_{k_N}^N\right\} \end{aligned} \tag{1}$$

The algorithm of ensemble clustering that uses co-association matrix can be described as follows (Fred and Jain 2005, p. 848):

a) obtain different base partitions,

b) build the co-association matrix (co-occurrence matrix). The main idea of this matrix is that objects belonging to the same clusters ("natural clusters") are likely to be co-located in the same clusters in different partitions. The elements of the co-association matrix are defined as follows:

$$C(i,j) = \frac{n_{ij}}{N}, \tag{2}$$

where: i, j – pattern (objects) numbers, n_{ij} – number of times pattern (i,j) is assigned to the same cluster among N partitions, N – total number of partitions,

c) apply the co-association matrix as the data matrix for some classical clustering method – like single-link, average, *k*-means or pam,

d) choose the best partition. Fred and Jain (2002) propose to apply "lifetime" criterion in the case of hierarchical clustering methods. They define lifetime as the value of threshold values on the dendrogram that leads to the identification of k clusters – their suggestion is to look for the highest value of this threshold.

Also other methods that will lead to identification of the final number of clusters can be applied – for example Baker & Hubert, Hubert & Levine, Russeeuw's silhouette cluster quality indices (see for example Gatnar and Walesiak 2004, p. 342-343 for details).

4. Results of simulation studies

In order to compare the results of single clustering method (single model) with results of ensemble clustering the adjusted Rand index was applied in the case of single clustering method. In the case of ensemble clustering average ensemble accuracy (that is based on adjusted Rand index) is applied. Average ensemble accuracy can be defined as follows:

$$A_{agr} = \frac{1}{K}\sum_{k=1}^{K} AR\left(P_k^{agr}, P'\right), \tag{3}$$

where: K – number of ensembles, AR – adjusted Rand index, P_k^{agr} – classification on the base of k-th ensemble, P' – known class labels.

The individual accuracy is defined as follows:

$$A_i = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{J}\sum_{j=1}^{J} AR\left(P_k^j, P'\right), \tag{4}$$

where: J – number of ensemble members, P_j^k – classification on the base of j-th member of k-th ensemble.

To compare the results of single clustering methods with results of ensemble clustering four different artificial data sets where generated (models are obtained by applying culsterSim and mlbench packages of R software):

1. **Data set I** – 120 symbolic objects in three elongated clusters described by two interval-valued variables. The observations are independently drawn from bivariate normal distribution with means $(0, 0), (1.5, 7), (3, 14)$ and covariance matrix Σ $\left(\sigma_{jj} = 1, \sigma_{jl} = -0.9\right)$.

2. **Data set II** – 120 symbolic objects divided into five clusters in three dimensions that are not well separated. The observations are independently drawn from multivariate normal distribution with means equal to: $(5, 5, 5), (-3, 3, -3), (3, -3, 3), (0, 0, 0), (-5, -5, -5)$, and covariate matrix Σ, where $\sigma_{jj} = 1\,(1 \le j \le 3)$, and $\sigma_{jl} = 0.9\ (1 \le j \ne l \le 3)$.

To obtain symbolic interval data for data sets I and II the data were generated for each model twice into sets *A* and *B* and minimal (maximal) value of $\left\{x_{ij}^A, x_{ij}^B\right\}$ is treated as the beginning (the end of interval).

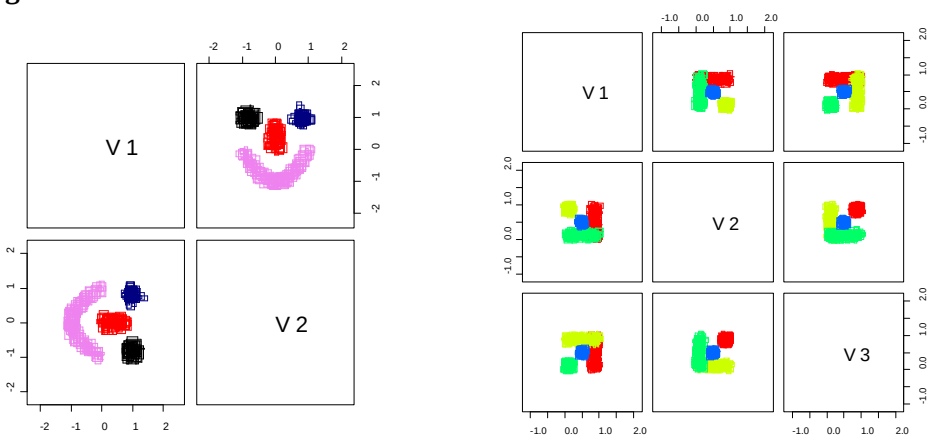
3. **Data set III** – is an adaptation of well-known cuboids data set (from mlbench package). Four clusters in three dimensions.

4. **Data set IV** – is an adaptation of well-known smiley data set (from mlbench package). Four clusters in two dimensions.

In order to build interval-valued variables from mlbench cuboids and smiley data sets the data obtained from mlbench package is treated as the "seed" of a rectangle. Each rectangle is therefore a vector of two intervals defined by:

$\left(\left[z_j - \gamma_j / 2, z_j + \gamma_j / 2\right]\right)$, where z_j – is the value of variable for j-th variable, γ_j – is the width and the height of the rectangle for j-th variable. The value γ_j is drawn randomly from the interval $[0, 1]$ for each variable. The figure 1 presents data sets III and IV.

Figure 1. Data sets III and IV



a) b)

a) – data set IV (smiley); b) – data set III (cuboids)

Source: own computations in R software.

To determine the final number of clusters Rousseeuw's Silhouette, Baker & Hubert, Hubert & Levine cluster quality indices were used (Ichino-Yaguchi distance measure was applied). The most common result was taken into consideration. Results of clustering with application of single model and ensemble clustering results for each data set (with application of adjusted Rand index) are presented in table 2.

Table 2. Results of clustering for four data sets

Clustering approach	**Data set I**		**Data set II**		**Data set III**		**Data set IV**	
	Number of clusters	Rand index	Number of clusters	Rand index	Number of clusters	Rand index	Number of clusters	Rand index
Single method:								
- single link	2	1	2	0.1744	11	0.3314	10	0.2212
- average link	2	1	2	0.3961	3	0.2943	2	0.1627
- pam	2	1	2	0.3786	3	0.3171	2	0.2302

Table 2. Results of clustering for four data sets (cont.)

Clustering approach	Data set I		Data set II		Data set III		Data set IV	
	Number of clusters	Rand index	Number of clusters	Rand index	Number of clusters	Rand index	Number of clusters	Rand index
Ensemble approach: - different clustering methods applied; number of clusters chosen at random from the interval [2; 15]	2	1	5	1	4	0.8266	4	0.8457

Source: Own research with application of R software.

5. Final remarks

Ensemble clustering methods that were developed to deal with classical data situation can be quite easily adapted to symbolic data situation. Ensemble clustering methods based on the co-association (co-occurrence) matrix can be applied to cluster symbolic interval-valued data.

Symbolic interval-valued data often tends to form not well-separated clusters of many different shapes. Single clustering methods (hierarchical, divisive or iterative) not always can detect correct number of clusters. Ensemble approach in clustering can be a solution to these problems.

For the purposes of simulation studies a R script was written by author. It allows co-occurrence matrix to be built and applied as the data matrix for any suitable clustering method.

Simulation studies have shown that ensemble clustering based on co-association matrix achieves better results (in terms of adjusted Rand index) than single clustering methods – especially when dealing not typical cluster structures, or not-well separated clusters.

The most important aims for future work are: comparing ensemble clustering based on co-association matrix with other ensemble clustering approaches, do more simulation studies on ensemble learning for symbolic data.

REFERENCES

BILLARD, L., DIDAY E. (2006). Symbolic Data Analysis: Conceptual Statistics and Data Mining, Wiley, Chichester.

BOCK, H.-H., DIDAY, E. (red.) (2000). Analysis of symbolic data. Explanatory methods for extracting statistical information from complex data, Springer Verlag, Berlin-Heidelberg.

BREIMAN, L. (1996). Bagging predictors, Machine Learning, 24(2), p. 123-140.

DE CARVALHO, F.A.T., LECHEVALLIER, Y., DE MELO, F.M. (2012). Partitioning hard clustering algorithms based on multiple dissimilarity matrices, Pattern Recognition, 45(1), p. 447-464.

FERN, X.Z., BRODLEY, C.E. (2004). Solving cluster ensemble problems by bipartite graph partitioning, Proceedings of the 21st International Conference on Machine Learning, Canada.

FRED, A.L.N., JAIN, A.K. (2005). Combining multiple clustering using evidence accumulation, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 27, p. 835-850.

GAHEMI, R., SULAIMAN, N., IBRAHIM, H., MUSTAPHA, N. (2009). A survey: Clustering ensemble techniques [in:] Proceedings of World Academy of Science, Engineering and Technology, Vol. 38, p. 636-645.

GATNAR, E. (2008). Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji, Wydawnictwo Naukowe PWN, Warszawa.

GATNAR, E., WALESIAK, M. (red.) (2004). Metody statystycznej analizy wielowymiarowej w badaniach marketingowych, Wydawnictwo AE, Wrocław.

HORNIK, K. (2005). A clue for cluster ensembles, Journal of Statistical Software, 14, 65-72.

NG, A., JORDAN, M., WEISS, Y. (2002). On spectral clustering: analysis and an algorithm, [In:] T. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14, MIT Press, 849-856.

STREHL, A., GHOSH, J. (2002). Cluster ensembles – A knowledge reuse framework for combining multiple partitions, Journal of Machine Learning Research, 3, p. 583-618.

STATISTICS IN TRANSITION-new series, Summer 2012
Vol. 13, No. 2, pp. 343—364

TAXONOMIC ANALYSIS OF THE POLISH PUBLIC HEALTH IN COMPARISON WITH SELECTED EUROPEAN COUNTRIES

Anna Wierzbicka[1]

ABSTRACT

As an international organization the European Union pursues a range of purposes including improvement of public health, prevention of human illness and diseases as well as elimination of the sources of danger to physical and mental health. To enable effective health policies supporting actions of the member states, the data and materials which are fundamental to public health assessments are collected at the Community level. They underpin a variety of analyses necessary to evaluate changes in medical systems and determine the degree of similarity between the EU member states.

This article analyses public health in Poland in relation to selected European countries. The study is based on the medical, economic and social indicators available from the EUROSTAT database. The taxonomic methods used in the study allowed ranking the sampled countries and identifying those with the highest level of public health. A more detailed assessment was based on the general presentation of health care systems in each country. In addition to Poland, the other post-socialist countries in the sample are Bulgaria, Estonia, Lithuania, Romania, Slovenia, Hungary and the Czech Republic. Importantly, the study covers the years 2004-2009, after most of the countries joined the European Union. The reason why other former Eastern bloc countries were omitted from the study was the unavailability of appropriate data in the EUROSTAT database. The paper discusses a group of developed European countries too. However, because the complicated historical past of the post-socialist countries and the socio-economic difficulties that result from it make their analysis more interesting from a comparative point of view, they are the primary group explored in this article.

This article is divided into theoretical and empirical section. It begins with a short introduction. Part two presents the concept of health and public health. The third part describes key factors differentiating health systems. It briefly presents the main healthcare models, that is: the Bismarck's, Beveridge's, Siemaszko's and the residual model. The fourth section refers to a grouping method based on Hellwig's taxonomic measure of development. Part five with its scope includes a

[1] Uniwersytet Łódzki, Zakład Demografii i Gerontologii Społecznej.
E-mail: wierzbicka.anna@venti.com.pl.

comparative analysis of public health status of selected European countries. The sixth part presents major determinants of cluster formation. The paper ends with a conclusion that summarizes the obtained results.

Key words: Public health, taxonomic measure of development, comparative analysis.

1. Introduction

Today's Europe is populated by several hundred ethnic groups and nations totalling over 730 million people[1]. In the year 1000 the population living on the old continent was less than 40 million, but within the next 900 years the number rose to 408 million, to be more than 725 million in the year 2000[2]. The recent generations were growing up in the period of "demographic explosion", when the number of population was expanding at a very fast rate. The growth was certain to continue and the only questions it brought about were: Will it be fast, very fast or perhaps lightning fast? Will our planet be able to provide space and food for so many inhabitants? These concerns turned out to be ungrounded. Today's forecasts show that by the year 2020 the European population will decline to 715 million and then to ca. 650 million in 2050[3]. The old continent, particularly the European Union, is faced by a low or negative natural increase and the aging of its population. These circumstances increase the interest in demographic transformations and, particularly, in their socio-economic impacts on social security systems whose main purpose is to protect people against so-called social risk. One element of this protection is healthcare[4] (others are social welfare, old age and disability pensions, protection against unemployment, etc.).

2. The notion of health and public health

„Health" may be interpreted from different angles and its meaning greatly varies depending on the accumulated knowledge, scientific achievements, attitudes as well as cultural norms prevailing in societies.The five-thousand year old Chinese medicine defines health as a balance between two types of energy: ying and yang[5]. If the balance is disturbed, then diseases appear. This approach

[1] United Nations Statistics Division (2011). The above number depends on how the geographic area of Europe is defined. In 2011 the EU population stood at over 500 million people. In the same year the number of the non-EU population in Europe was estimated at more than 230 million.

[2] United Nations Population Division and http://www.fordham.edu/halsall/source/pop-in-eur.asp (26 March 2012).

[3] United Nations, Eurostat, Berlin Institute for Population and Development.

[4] According to the ILO definition of 1955.

[5] Two opposing energies that joined to create the universe. One may not function without the other. In discussions on health ying represents systemic fluids and human tissue, while yang is energy, warmth, activity and productivity.

seems to explain why preventive measures are so popular in the Far East. Prevention comes before treatment (Lisowski, 2000). The Chinese tend to treat a human being rather than a disease, so analysis of the patient's environment is necessary. Considering the longevity of the Chinese, their young appearance and the low rate of civilisation diseases in China, this approach seems to work very well and, although not fully comprehensible to the Europeans, it has had a direct bearing on the practice of western medicine.

Hippocrates, one of the predecessors of modern European medicine known as "the father of medicine", was of the same opinion. He too associated health, that is feeling well, and illness, i.e. feeling bad, with a balance between the patient's environment and lifestyle (Korczak, 1977).

Medicine has made huge progress since Hippocrates times, mainly due to the scientific revolution in the 19th c. The meaning of most notions has been changed, or rather has been extended. One of the basic and most popular definitions of health that are in use today was formulated by the world Health Organization (WHO) in 1948: „Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity"(WHO, 1948). In the 21st c. the meaning of health is not limited to mental and physical well-being, i.e. to the correct functioning of the human organism. People live more intensive lives today, so social, economic and cultural, sometimes also spiritual determinants occurring in their environments must be taken into account. All of them affect health which is a multifaceted phenomenon.

The health of the general public is called public health. These two terms are only different in that health is an attribute of individuals, while public health is used with respect to populations. Summing up, public health equals population health. It is a multidisciplinary field of science incorporating the elements of many other disciplines, such as epidemiology, sociology, psychology, pedagogy, economics, and law. The notion of public health has lost its purely medical dimensions.

The origin of the term dates back to the 18th c.[1]. Traditionally, public health has been interpreted only in terms of hygiene and contagious diseases. Today it is understood as the science and the art of preventing diseases, extending life expectancy and promoting health through collective efforts of the society (Acheson, 1998). According to the WHO, public health in its broad sense, i.e. as „new public health", consists of:

issues concerning population health,
population health status,
general health services,
healthcare administration (WHO, 1973).

The WHO Health Promotion Glossary of 1998 explains that the reason for drawing a line between „public health" and „new public health" was the need to

[1] Johann Peter Frank (1745-1821) – a German physician who provided the fundamentals of modern hygiene, epidemiology, forensic medicine, state-managed healthcare. Frank is credited to develop the concept and the notion of „public health".

accentuate completely different approaches to describing and analysing health determinants and methods for solving population health problems that are used within these two areas (WHO, 1998). New public health is characterised by full knowledge of how lifestyles and living conditions determine human health status. It also requires that the needs relative to the use of resources are identified and that appropriate investments in solutions, services and institutions that create, maintain and protect health are made. These characteristics of public health have caused that the WHO views it as a socio-political concept serving the improvement of health and quality of life of whole populations through various kinds of interventions.

The definitional scope of both health and public health is very extensive. There are many definitions, concepts and interpretations that have been developed for these two terms, which are not only of crucial importance, but also, more importantly, complementary. Population health is an aggregate of the health of individuals.

The state must therefore understand the fundamental importance of both individuals and the entire population being able to enjoy the same degree of care. Otherwise statehood becomes a meaningless notion, because states may not exist without individual citizens.

3. The models of healthcare systems

Health represents the highest social priority and value. But it is also a product delivered by the sector of medical care. It is not possible for individual patients to influence the functioning of their organisms, which are exposed to various external stimuli.

According to Marc Lalonde's health field concept[1], the status of human health is determined by four factors: biology (genetic determinants), the environment, the lifestyle and healthcare. In his model, the greatest influence on human health is attributed to lifestyle (over 50%), while healthcare contributing only 10% is the least important. It is not likely, though, that a human with "good genes" and living healthy life will be safe from diseases and accidents. Therefore, nations need institutions and establishments providing medical assistance and consultations and taking care of the sick. This role is performed by healthcare understood as „a separate whole made of many diverse elements interconnected by various ties (representing different interactions) that pursues health-related objectives" (Włodarczyk, Poździoch, 2001). Accordingly, healthcare includes all kinds of activities whose main purpose is to promote, restore and sustain health (WHO 1984).

[1] Marc Lalonde – former Minister of National Health and Welfare in Canada. In 1974 he published the report "A New Perspective on the Health of Canadians" where he presented the "health fields" concept; http://pl.wikipedia.org/wiki/Marc_Lalonde (26 Februry 2012).

The healthcare sector is one of the main branches in the economy of every country. The health systems of particular countries pursue the same goals, but differ from each other in the circumstances that have led to their establishment. Countries can be roughly divided into developed, developing and the Third World countries in the poorer and less industrialised areas. Different countries are characterised by different standards of living and levels of development, as well as having distinctive historical past and unique socio-economic problems that arise from it. This means that all analyses, programmes, policies and solutions addressed to the different parts of the world may not miss the huge differences between nations, their rates of economic growth and health status.

The key factor differentiating health systems is the way they are funded and organized. The main healthcare models are the following:

- the Bismarck's model (based on insurance),
- the Beveridge's model (based on taxes, national budget, and national health service),
- the Siemaszko's model (in centrally-planned systems),
- the residual model (a market-based solution).

Table 1. Models of healthcare systems

Bismarck's model	**Beveridge's model**	**Siemaszko's model** (a bygone model – socialist health service)
Funded from predominantly mandatory premiums paid by employees and employers	Funded from general taxation or other state revenues	Funded from general taxation or other state revenues
All premium payers are entitled to use services, mainly employees and their family members	All (almost) citizens in the country are entitled to benefits	All (almost) citizens in the country are entitled to benefits
Funds are distributed through sickness funds	Funds are distributed through central (national) or decentralised (national or subnational) institutions of public administration	Funds are distributed through central (national) and regional institution of public administration
The „basket of services" is created by excluding some types of services (e.g. all or some dental services, plastic surgery, and physiotherapy)	The "basket of services" is very general and broad, practically depending on public investments	The "basket of services" is very general and broad, practically determined by public investments
Mainly private service providers acting for profit	Mainly public service providers	Only public service providers

Table 1. Models of healthcare systems (cont.)

Bismarck's model	**Beveridge's model**	**Siemaszko's model** (a bygone model – socialist health service)
Contracts between payers and service providers	The central level allocates funds to the intermediate level (including local governments) and to service providers using centrally determined rules	The central level allocates funds to the intermediate level (of governments administration) and to service providers according to centrally determined rules
Rates, frequently the same across the country, are determined in contracts via an administrative/negotiatory mechanism	Fund allocation rules are defined at the central or regional level in relation to the available infrastructure and population characteristics	Fund allocation rules are defined at the central level in relation to the infrastructure
Financing per service	Cavitation financing and master budgets (mainly for hospitals)	Financing rules based on conversion factors related to infrastructure
Co-payment for most services	Minimal co-payment	Co-payment is not used (expanding „grey zone")
Service providers of choice.	Regulated access to successive levels of healthcare	Assignment to regional/ local structures of service providers, so-called catchment areas

Source: Siwińska V., Brożyniak J., Iłżecka J., Jarosz M. J., Orzeł Z., Modele systemów opieki zdrowotnej w Polsce i wybranych państwach europejskich, Praca poglądowa, Zdrowie Publiczne 2008.

The fourth healthcare model which has been omitted from the table is the residual model using market solutions, where health is treated as a commodity. Healthcare services are considered goods to be exchanged via commercial transactions. They can be sold and purchased as any other product. The key factor is the purchasing power of the buyer. The model lays stress on personal responsibility. It is for an individual to make all decisions and to accept all health risk. The public sector protects poor people, the elderly, and mothers and children in difficult life situation. The government exercises very limited control of the expenses and the allocation of resources. All inputs invested in the functioning of healthcare (labour, capital) are expected to bring returns to those who made them, according to the free market principles (Włodarczyk, 1996).

Each of the healthcare models is different and all of them have their advantages and disadvantages. A single model that could solve the problem of diseases and medical dilemmas in the country does not exist. Which health care system is adopted is simply a consequence of the strategy that the state follows. The choice is determined by its health policy, available resources, the

opportunities for their redistribution, as well as the level of funding and its sources. In practice, countries mostly use the modified versions of the above models, which borrow solutions from each other.

Europe has more than a dozen of healthcare solutions that have evolved, to a different degree and in different proportions, from the main models. Most countries have built their systems on the oldest of the classical solutions, which was created by Chancellor of the German Reich, Otto von Bismarck. These are Germany, Austria, France, the Netherlands and Switzerland, as well as Japan and Israel outside Europe. The national healthcare system devised by Lord Beveridge after World War II has been adopted with many modifications by the Nordic countries, Spain, Portugal, Ireland, the United Kingdom, Italy, etc. Canada is one of the non-European users of this system. The least common model which represents a mixture of the two models is used in Belgium and Greece. The USA uses a system built on strictly market principles.

The system that Poland had in the inter-war period was structured on the Bismarck model. The historical events after World War II caused that the Eastern bloc countries, including Poland, adopted the Siemaszko's model, which strongly resembles that developed by Lord Beveridge. The system was never fully implemented in our country, but most of its solutions remained in force until 1998, when the health care system started to be remodelled towards the Bismarck's concept.

Table 2. Countries grouped by their healthcare systems

A Bismarckian-type model	**A Beveridge-type model**	**A mixed model with elements of the Bismarck's and Beveridge's models**	**A system originally based on the Siemaszko's model, evolving into the Bismarck model**
Germany, Austria, France, Netherlands, Switzerland, Luxemburg	Denmark, Finland, Norway, Iceland, Sweden, UK, Ireland, Spain, Portugal, Italy	Belgium, Greece	Albania, Bosnia-Herzegovina, Bulgaria, Croatia, Estonia, Hungary, Lithuania, Latvia, Poland, Romania, Czech Republic, Slovakia, Slovenia, Macedonia
outside Europe: Japan, Israel	outside Europe: Canada		And the former USSR countries, e.g. Armenia, Belarus, Georgia, Kazakhstan, Russia, Ukraine

Source: Developed by the author.

After the World War II, many countries started to restructure their healthcare systems, aiming to create modern, better organized and "welfare" systems. The specific solutions they chose were determined by their policy and potential. The different rates of social development and wartime losses have caused that any unification of the systems has not been possible to date. This seems to explain why particular countries use diverse methods, instruments and solutions, although the objectives and goals are similar.

The former Eastern-Bloc countries were faced with the gravest problems, because they had to switch from central-command economies to free market systems only several tens of years after the end of the greatest armed conflict in world history. Economic rigidities, shortage of goods, unrealistic economic plans, excessive employment ("hidden" unemployment), the lack of effective motivation systems are only few of the problems that haunted countries subordinated to communist ideology. They had a direct bearing on the situation of the thoroughly centralised healthcare systems, contributing to the low health status of population, lengthening queues of patients seeking consultations from specialist physicians and to technological backwardness. It was not until the collapse of the regime that the anachronistic nature of the system in the former socialist bloc became fully exposed. Hence, the systemic transformation complete, reforms of the health care systems had to be undertaken, whose main principles were borrowed from models used elsewhere.

4. A grouping method based on hellwig's taxonomic measure of development

Taxonomy is used for classifying and ordering objects characterised by many parameters and properties and its roots go back to biological sciences. It originally served the purpose of grouping live organisms based on their anatomic and physiological characteristics. A breakthrough in its use was a study by the Polish scientist Jan Czekanowski1, who developed his own taxonomic method in 1913. His method basically consists in ordering and sorting objects by grouping them into possibly homogenous classes2. Czekanowski's research gave impulse for using taxonomy in other fields of science, such as medicine, sociology, and economics.

The classification methods allow creating homogenous groups of objects and reducing a large amount of data to several major categories, thus enabling more general conclusions to be reached. Another advantage is that they help cut the

[1] Jan Czekanowski (1882-1965) – Professor at the Lvov University (1913-1941) and the Catholic University of Lublin, head of the Chair of Anthropology and Ethnography, of the Chair of Anthropology at the Adam Mickiewicz University in Poznan, full member of the Polish Academy of Sciences, a member of the Polish Statistical Association and the Polish Biometric Society.

[2] http://www.stat.gov.pl/cps/rde/xbcr/gus/POZ_Zasluzeni_statystycy_dla_nauki.pdf (5 March 2012).

time and costs of research, because the number of necessary analyses can be limited to the most important issues.

One of the major taxonomic methods is the Hellwig's synthetic measure of development. It is a linear ordering method, where the points in a multidimensional space are projected onto a straight line1. The analysed objects are ordered with respect to the accepted model development which is a synthetic measure integrating the characteristics and information from a series of variables into one aggregate indicator. These properties of the Hellwig's method make it possible for the analysed objects to be ordered by the level of phenomena that cannot be quantified with a single measure, e.g. the quality of life, population health status or technological progress. That multifaceted phenomena can be described with a single numerical value and makes the identification of their level much easier, as well as the facilitation of creation and analysis of comparative rankings.

The first step in constructing the Hellwig's taxonomic measure of development is to determine the range of diagnostic variables to be used in the study (x1, x2,…, xk) and to create a matrix of information on particular objects. The matrix is written as:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \dots \\ X_m \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mk} \end{bmatrix} \tag{1}$$

where: x_{ij} – the value of the *j*-th diagnostic variable for the *i*-th object; i = (1,2…,m),

j = (1,2…,k); m – the number of objects, k – the number of diagnostic variables.

In the next step, the matrix is standardised with the following formula:

$$Z_{ij} = \frac{x_{ij} - \overline{x}_j}{S_{xj}}, \quad \text{i} = (1,2\dots,\text{m}),\ \text{j} = (1,2\dots,\text{k}) \tag{2}$$

where: x_{ij} – the empirical value of the *j*-th diagnostic variable for the *i*-th object,

$\overline{x}_j$ – the arithmetic mean in the distribution of the diagnostic variable x_j,

S_{xj} – the standard deviation in the distribution of the diagnostic variable x_j.

[1] In the method of non-linear ordering points in a multidimensional space are projected onto a plane.

The synthetic model variable, i.e. the model development**,** is obtained from the standardised variables:

$$P_{0j} = [z_{01}, z_{02},\ldots,z_{0k}], \tag{3}$$

where: $z_{0j} = \max(z_{ij})$ for stimulant variables,

$z_{0j} = \min(z_{ij})$ for destimulant variables.

A stimulant and a destimulant are diagnostic variables whose higher values indicate better or worse situation of the object, respectively.

In the next step, the Euclidean distance between each object and the model development P_{0j} is determined:

$$d_{i0} = \sqrt{\sum_{j=1}^{k}(z_{ij} - z_{0j})^2}, \quad i = (1,2\ldots,m),\ j = (1,2\ldots,k), \tag{4}$$

where: z_{ij} – the normalized values of the *j*-th variable for the *i*-th object,

z_{0j} – the normalized value of the model development for the *j*-th variable.

In order to normalize coefficient d_{i0}, a relative taxonomic measure of development is computed for particular objects:

$$z_i = 1 - \frac{d_{i0}}{d_0} \tag{5}$$

where:

$$d_0 = \bar{d}_0 + 2S_0 \tag{6}$$

$$\bar{d}_0 = \frac{1}{m}\sum_{i=1}^{m} d_{i0} \tag{7}$$

$$S_0 = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(d_{i0} - \bar{d}_0)^2} \tag{8}$$

where: $\bar{d}_0$ – the arithmetic mean of the distances to the model development,

S_0 – the standard deviation of the distances to the model development.

The measure takes values within the range [0:1]. The higher its value, the closer the examined object to the model and the better its situation. A result close to zero points to a very disadvantageous situation of the object. Based on the values of the taxonomic measure of development, the objects can be ordered by the level of the analysed phenomenon.

With the arithmetic mean and the standard deviation calculated for the relative measure of development four categories (clusters) were distinguished, which were then assessed with a four point-scale ranging from very good to unsatisfactory. It

must be noted at this point, though, that the scale levels are relative, because they only rank the sampled countries and use the available variables.

I group – very good $\bar{z} + S_z < z_i$ (9)

II group – good $\bar{z} < z_i \leq \bar{z} + S_z$ (10)

III group – satisfactory $\bar{z} - S_z < z_i \leq \bar{z}$ (11)

IV group – unsatisfactory $z_i \leq \bar{z} - S_z$ (12)

where: $\bar{z}$ – the arithmetic mean of the relative measure of development,

S_z – the standard deviation of the relative measure of development.

5. A comparative analysis of public health status

Public health status is a multifaceted phenomenon whose complexity prevents analysis based on its one property only. A multivariate approach including, for instance, taxonomic methods is necessary.

Because of the limited availability of the data, this analysis has covered eight post-socialist countries: Bulgaria, the Czech Republic, Hungary, Estonia, Romania, Lithuania, Slovenia and Poland. To find out how their achievements compare with the situation in the non-Eastern Bloc countries, the analysis has been extended to three Nordic countries – Finland, Norway and Sweden – and six Western-European countries represented by Germany, France, the Netherlands, Austria, Portugal and Switzerland. The research period spanned the years 2004-2009, which is important to note because all the selected post-socialist countries are new members of the European Union that joined it just in 2004, except for Bulgaria and Romania that became EU members 3 years later. Among the sample countries, Germany, France and the Netherlands are the oldest members of the Community. Their accession to the EU, then called the European Economic Community[1], took place in 1957. Norway and Switzerland are non-EU countries, but one belongs to the European Economic Area[2] and the Schengen Area[1] (Norway) and the other to the Schengen Area (Switzerland).

[1] The European Economic Community (EEC) was an international organization formed as a result of integration processes initiated after WWII. It started to function on 1 January 1958 based on the Treaty of Rome whose signatories were Belgium, France, the Netherlands, Luxemburg, FRG (the United Germany since 1990) and Italy on 25 March1957. The EEC turned into the European Union. On 1 December 2009 the EU acquired legal personality and replaced the EEC (called the European Community since 1993), taking over all its powers.

[2] The European Economic Area (EEA) is a free-trade area and a single market for the EU and EFTA (European Free Trade Association) members (except for Switzerland). The EEA is based on four pillars, i.e. the free **movement of people, capital, goods and services**.
http://pl.wikipedia.org/wiki/Europejski_Obszar_Gospodarczy (6 March 2012).

The analysis started with the selection of appropriate diagnostic variables for describing public health status in the selected countries. The status was identified by means of medical and economic, as well as social indicators. All data were complete, available and relevant to the description of the analysed phenomenon.

The initial set of variables contained:

- ✓ the number of practicing primary care physicians per 100,000 population (a stimulant),
- ✓ the number of students graduating from medical studies (physicians) per 100,000 population (a stimulant),
- ✓ the number of hospital beds per 100,000 population (a stimulant),
- ✓ the number of patients discharged from hospitals per 100,000 population (a destimulant),
- ✓ average hospitalization time [days] (a destimulant),
- ✓ standardized mortality rate (a destimulant),
- ✓ female life expectancy (a stimulant),
- ✓ male life expectancy (a stimulant),
- ✓ healthcare spending [a percentage of GDP] (a stimulant),
- ✓ healthcare spending – long-term care [a percentage of GDP] (a stimulant).

It is worth noting that well-selected diagnostic variables should be weakly correlated (to avoid duplication of information), and should be characterized by a high degree of variability. For this purpose, on the basis of carried out calculations, too strongly correlated variables were removed from the set of initial variables, as well as those variables for which the coefficient of variation was not greater than 10%. These variables were in fact reconsidered as quasi-permanent, what means that they do not provide relevant information about the analyzed issue. Variables that were eliminated are female and male life expectancy, both stimulants. Therefore, the final analysis included eight variables, five stimulants and three destimulants.

The clarity of the paper requires to remark that variables were divided into stimulants and destimulants and their values were used to obtain the synthetic model variable according to the below mentioned rule:

$z_{0j} = \max(z_{ij})$ for stimulant variables, $z_{0j} = \min(z_{ij})$ for destimulant variables,

where: z_{ij} – the normalized values of the *j*-th variable for the *i*-th object.

The above chosen diagnostic variables were used for making calculations and ordering countries by public health status based on the Hellwig's taxonomic measure of development (Table 3).

[1] The Schengen Agreement was signed on 14 June 1985 in Luxembourg. It has removed all personal border control between the member states, while providing for closer cooperation between their security services.

Table 3. The selected European countries ordered by public health status according to the taxonomic measure of development (z_i), years 2004-2006

No.	2004		2005		2006	
	Country	z_i	**Country**	z_i	**Country**	z_i
1	Austria	0,405	Austria	0,398	Austria	0,395
2	France	0,388	France	0,392	Netherlands	0,381
3	Netherlands	0,365	Netherlands	0,366	France	0,380
4	Norway	0,335	Norway	0,319	Norway	0,311
5	Germany	0,308	Germany	0,308	Germany	0,298
6	Switzerland	0,307	Switzerland	0,286	Switzerland	0,276
7	Portugal	0,239	Portugal	0,251	Portugal	0,261
8	Sweden	0,224	Slovenia	0,231	Sweden	0,244
9	Slovenia	0,214	Sweden	0,226	Slovenia	0,183
10	Finland	0,156	Romania	0,196	Finland	0,171
11	Romania	0,155	Finland	0,156	Romania	0,169
12	Czech Republic	0,143	Czech Republic	0,147	Czech Republic	0,157
13	Hungary	0,129	Hungary	0,131	Hungary	0,137
14	Poland	0,108	Estonia	0,115	Estonia	0,128
15	Estonia	0,101	Poland	0,092	Poland	0,081
16	Bulgaria	0,097	Bulgaria	0,087	Bulgaria	0,058
17	Lithuania	0,054	Lithuania	0,029	Lithuania	0,055

Source: developed by the author based on the Eurostat data.

Table 4. The selected European countries ordered by public health status according to the taxonomic measure of development (z_i), years 2007-2009

No.	2007		2008		2009	
	Country	z_i	**Country**	z_i	**Country**	z_i
1	France	0,389	Austria	0,362	Austria	0,386
2	Austria	0,368	France	0,354	Netherlands	0,368
3	Netherlands	0,367	Netherlands	0,347	France	0,363
4	Norway	0,308	Norway	0,312	Norway	0,303
5	Switzerland	0,279	Portugal	0,266	Germany	0,280
6	Portugal	0,278	Switzerland	0,258	Portugal	0,270
7	Germany	0,273	Germany	0,251	Sweden	0,212
8	Sweden	0,227	Sweden	0,206	Switzerland	0,177
9	Romania	0,199	Slovenia	0,187	Czech Republic	0,176
10	Slovenia	0,167	Finland	0,186	Finland	0,174
11	Czech Republic	0,149	Romania	0,154	Slovenia	0,174
12	Finland	0,143	Czech Republic	0,136	Lithuania	0,158
13	Hungary	0,138	Estonia	0,101	Estonia	0,137
14	Poland	0,105	Hungary	0,095	Romania	0,118
15	Estonia	0,094	Poland	0,088	Hungary	0,112
16	Bulgaria	0,059	Lithuania	0,080	Poland	0,075
17	Lithuania	0,058	Bulgaria	0,046	Bulgaria	0,023

Source: developed by the author based on the Eurostat data.

The table reveals significant movements in the rankings of particular countries. In all analysed years but 2007 Austria ranked first. Between 2004 and 2009 France and Netherlands continued to rank alternatively second or third position except for 2007 when France took the first place. This means that the health status of the Austrian, Dutch and French populations was the closest to the model development. Between 2004 and 2009 Norway maintained its 4th place in the ranking. Except for Austria and Norway seven more countries maintained their places in the raking during the first three years of the analysis: Germany (5), Switzerland (6), Portugal (7), the Czech Republic (12), Hungary (13), Bulgaria (16) and Lithuania (17). In the years 2007-2009 position of these countries was not so stable. Other countries ranked differently in particular years.

In the analysed period, the following countries moved up in the ranking:

- the Czech Republic – 11th and 12th in the years 2004-2008, became 9th in 2009,
- Lithuania – last and penultimate places between 2004 and 2008, but 12th in 2009,
- Portugal – ranked 7th in the years 2004-2006 and 6th in 2007 and 2009 became 5th in 2008.

On the other hand, the countries below moved down the scale:

- Switzerland – ranked 5th and 6th in the years 2004-2008, but 8th in 2009,
- Poland – ranked 14th and 15th between 2004 and 2008, but 16th in 2009,
- Hungary – ranked 13th and 14th in the years 2004-2008, but in 2009 it was 15th,
- Romania – ranked 9th, 10th, and 11th in the years 2004-2008, but 14th in 2009,
- Germany – ranked 5th in the years 2004-2006 and in 2009, but 7th in the years 2007-2008,
- Slovenia – ranked 8th, 9th, and 10th in the years 2004-2008, but 11th in 2009,
- Finland – ranked 10th and 11th in the years 2004-2006 and in 2008-2009, but 12th in 2007.

The lowest level of public health in the analysed years was found in Bulgaria, which ranked last twice (2008-2009), being also penultimate four times (2004-2007). Lithuania also ranked last (2004-2007), but in 2008 it moved up by a notch to the penultimate place and then to the 12th place in 2009, which shows that public health status in this country was changing for the better.

The analysis leaves no doubts that the last five places in the ranking were typical of the former Eastern bloc countries. All changes in their rankings were limited to this group. The post-socialist country that ranked higher was Slovenia, which outdistanced Finland in the period 2004-2008. However, in 2009 the two countries swapped their places and Finland ranked higher again. Two more former Eastern bloc countries that ranked higher than Finland are Romania in 2005 and 2007 and also the Czech Republic in 2007 and 2009. None of other former Eastern bloc countries ranks higher than the Nordic or Western European countries in the sample. It is worth noting that across the analysed years public health status in the last countries in the ranking was quite far from the model development, clearly pointing to a large distance dividing the post-socialist countries and the ranking leaders.

Among the Eastern bloc countries being relatively new EU member states public health status was best in Slovenia, which ranked more or less in the middle of the sample. Interestingly, it was outdistanced only twice by the other post-socialist countries, by Romania in 2007 and by the Czech Republic in 2009. Positive trend was found in the Czech Republic, whose more or less regular place throughout the years 2004-2008 and its promotion to the 9th place in 2009 testifies to quite stable health policy pursued in the country. An interesting solution that has been implemented in the Czech Republic is that patients are required to cover some costs of medical services which generates additional revenues for healthcare; their 2008 volume was estimated at around €200 million.

The fact that Austria, France and the Netherlands invariably ranked in the first three places between 2004 and 2009 is not surprising. These countries are the oldest members and the founders of the European Union. All other countries in this analysis that usually ranked among the leaders joined the Community much later, except for Germany.

The basic descriptive statistics of the synthetic variables were used to put the countries into groups (to cluster them) according to their public health status in the years 2004-2009.

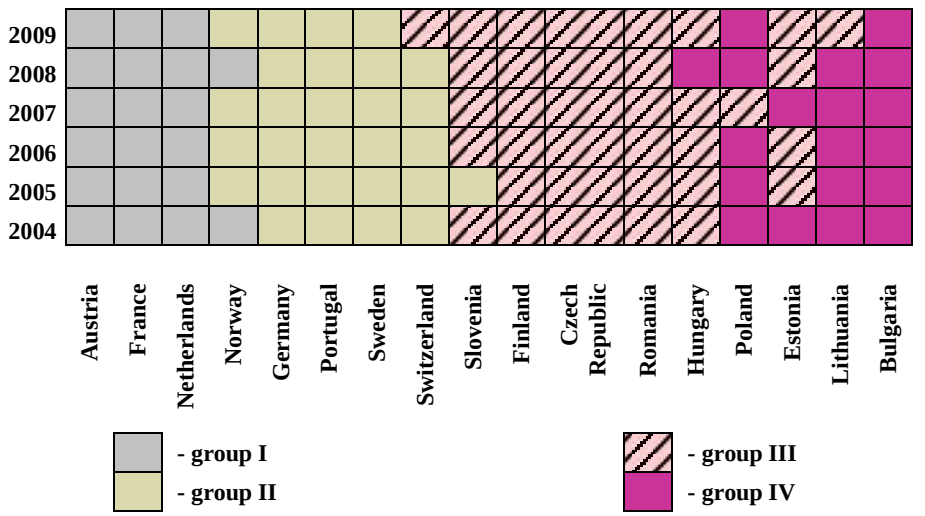
Table 5. The groups of the sampled European countries by their public health status, years 2004-2009

Group	2004	2005	2006	2007	2008	2009
I	Austria France Netherlands Norway	Austria France Netherlands	Austria Netherlands France	France Austria Netherlands	Austria France Netherlands Norway	Austria Netherlands France
II	Germany Switzerland Portugal Sweden	Norway Germany Switzerland Portugal Slovenia Sweden	Norway Germany Switzerland Portugal Sweden	Norway Switzerland Portugal Germany Sweden	Portugal Switzerland Germany Sweden	Norway Germany Portugal Sweden
III	Slovenia Finland Romania Czech Republic Hungary	Romania Finland Czech Republic Hungary Estonia	Slovenia Finland Romania Czech Republic Hungary Estonia	Romania Slovenia Czech Republic Finland Hungary Poland	Slovenia Finland Romania Czech Republic Estonia	Switzerland Czech Republic Finland Slovenia Lithuania Estonia Romania Hungary
IV	Poland Estonia Bulgaria Lithuania	Poland Bulgaria Lithuania	Poland Bulgaria Lithuania	Estonia Bulgaria Lithuania	Hungary Poland Lithuania Bulgaria	Poland Bulgaria

Source: Developed by the author.

The groups are represented graphically in Chart 1.

Chart 1. Groups distinguished in the set of 17 European countries, years 2004-2009



Source: Developed by the author.

Austria, France and the Netherlands are unquestionable leaders in the ranking. In the analysed period they always belonged to the top group. Norway joined it in 2004 and 2008. Group II consists of countries where public health status is good. Its steady members were Germany, Portugal and Sweden. Norway belonged to this group almost throughout the sample period, excluding years 2004 and 2008, when it was promoted to Group I. Besides these five countries, Switzerland was also a five-time member of Group II, in the years 2004-2008. The only post-socialist country that ranked that high is Slovenia which joined Group II in 2005. Group III was the largest in the year 2009, when its members were Switzerland, the Czech Republic, Finland, Slovenia, Lithuania, Estonia, Romania and Hungary. Group III is made of countries where public health status was found to be satisfactory. Finland, the Czech Republic and Romania were permanent members of this group; Lithuania joined it in 2009 and Estonia in the years 2005-2006 and 2008-2009. Group III was the smallest in the years 2004-2005 and 2008. In 2008 its only members were Slovenia, Finland, Romania, the Czech Republic and Estonia, while other countries felt to Group IV (Hungary, Poland). In 2008 Poland felt from Group III to Group IV, joining Hungary, Lithuania and Bulgaria. On the other hand, in 2009 countries in Group III were joined by Hungary and Lithuania; the latter managed to get out from the lowest group only once in the entire period. In the other years Lithuania was in the worst performing Group IV encompassing countries with unsatisfactory level of public health. Bulgaria did not manage to leave this group even once. It is worth stating that Poland entered into Group IV as many as five times, joining Lithuania and

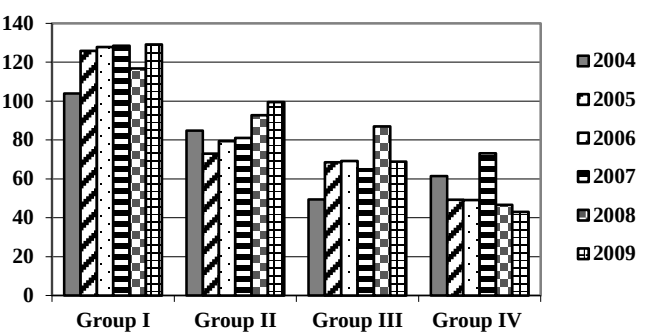
Bulgaria. Our country moved up to Group III only once, in 2007. Other members of Group IV were Estonia in 2004 and 2007, and Hungary in 2008. A quite surprising finding was to see Finland in Group III in all the analyzed period. It may have been expected that the principles of health and welfare would rank it higher in relation to other countries, closer to Norway.

Across the analysed years Austria, France and the Netherlands (Group I), Germany, Portugal and Sweden (Group II), Finland, the Czech Republic and Romania (Group III) and Bulgaria (Group IV) were the steady members of their groups. Bulgaria turned out to be the worst in the ranking, never leaving the group of countries characterised by the unsatisfactory status of public health (IV). Lithuania made the greatest progress, moving from the worst group in the years 2004-2008 to Group III in 2009. Quite interesting is the situation of Estonia, which felt to the worst group only twice. It is likely that Estonia owes this positive change to the increasingly close cultural relations with Finland and Sweden. Estonia presents herself as a Nordic country more and more often, which may have a positive influence on both her economy and the healthcare system. Slovenia was the most successful among the post-socialist countries – it outdistanced Sweden once, Finland five times and managed to retain is approximately middle place in the ranking throughout the analysed period. In other former Eastern bloc countries health status was not satisfactory. Regrettably, its level was very remote from the model development.

6. Major determinants of cluster formation

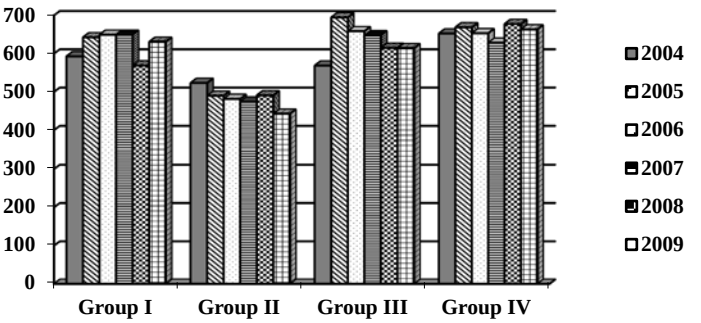
This section presents the evolution of the four main characteristics contributing to the emergence of a clear-cut gap between the Western European countries and the post-socialist countries. The characteristics (variables) are presented for particular groups in the years 2004-2009.

Chart 2. The average number of practicing primary care physicians per 100,000 population in each group, years 2004-2009



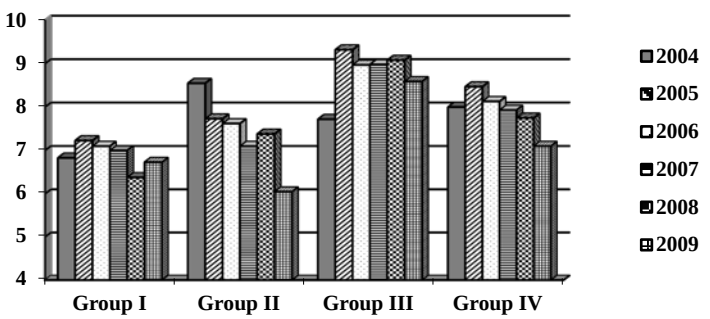
Source: Developed by the author.

Chart 3. The average number of hospital beds per 100,000 population in each group, years 2004-2009



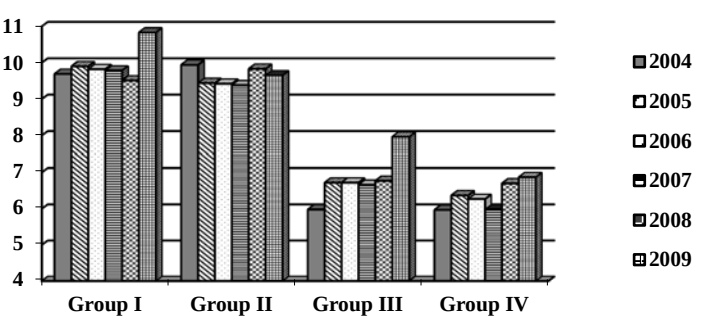
Source: Developed by the author.

Chart 4. The average patient hospitalization time in each group, years 2004-2009



Source: Developed by the author.

Chart 5. The average healthcare spending [% of GDP] in each group, years 2004-2009



Source: Developed by the author.

Charts 2-5 show that particular diagnostic variables contribute to the formation of clearly dissimilar clusters. In all analysed years, Groups I and II have higher average numbers of practicing primary care physicians (particularly cluster I), higher shares of healthcare expenditures (as % of GDP) and shorter average hospitalization time per patient compared with the other two groups. The latter finding also denotes a lower average number of hospital beds, because more efficient patient turnover allows better use of this resource.

7. Conclusion

The taxonomic analysis of public health presented in this article covered 17 European countries. Three of them were Nordic countries, six were Western European countries, and the other countries in the sample were former Eastern bloc countries, which makes their socio-economic situations and the resultant development opportunities very similar. The research has shown that in the 2004-2009 ranking the leaders were countries that based their healthcare systems on the Bismarck's model. Among the first four countries only one (Norway) has a healthcare system drawing on the Beveridge's concept. The post-socialist countries where historical circumstances determined the long-term use of the Siemaszko's model ranked the lowest. Most members of the former Eastern bloc started to reform their health systems only around 15 years ago, using the solutions adopted by the "old" EU members. The process can be expected to continue, bringing forth a sort of functional unification across the Community systems. A case in point is Slovenia where some functional characteristics of its healthcare system are not different from those in the leading countries. It is worrying, though, that Poland was in the worst group almost for the entire analyzed period. The country managed to change its position and move upper to Group III only once, in 2007. That year Poland outdistanced three post-socialist countries – Estonia, Lithuania and Bulgaria.

Poland's low ranking in the years 2004-2009 reveals the difficult situation of the Polish healthcare system. There are several issues that contribute to its poor performance. The first of them seems to be huge problems of public finances that necessitate austerity measures affecting also the healthcare sector. The second issue is inadequate health policy and poor adjustment between healthcare and changing demographic circumstances, particularly regarding fast aging of population. Aging increases the need for medical services, thus deepening the imbalance between their supply and demand (also spatially; for instance, the demand tends to concentrate in large urban centres). Moreover, Poland has not implemented solutions that have been proven effective in countries with healthcare systems based on the Bismarck's model, such as co-payment. Adjusted to the prevalent financial circumstances, co-payments are not burdensome for individuals, but their total volume may make a large difference in the amount of funding available to health service.

As shown by the analysis, between 2004 and 2009 mainly only two post-socialist countries in the sample (Lithuania and Bulgaria) had worse public health status than Poland. That shows that Poland is still very distant from the model development as well as from the Nordic and Western European countries. Besides, the status has not recently improved. For the sake of illustration, the Ranking Patient Empowerment Index developed in 2009 ranked the Polish healthcare system among the worst systems in Europe. Out of an attainable maximum of 1000 points Poland scored only 528, ranking 25th among 31 analysed countries. Although the scope of the Index and the variables it used somewhat differ from this research, the PEI report is an important point of reference that provides guidelines for future studies in this field.

REFERENCES

ACHESON, D., 1998. Inequalities in health: report of an independent inquiry, London: HMSO.

BERBEKA, J., 2006. Poziom życia ludności a wzrost gospodarczy w krajach Unii Europejskiej [Standard of living and economic growth in the European Union], Akademia Ekonomiczna [Academy of Economics], Kraków (in Polish).

CAMPBELL, A., 1976. Subjective measures of well- being, American Psychologist.

CAMPBELL, A., CONVERSE, P., RODGERS, W., 1976. The quality of American life, Russel Sage Fundation, New York.

Constitution of the World Health Organization, 1948.

GETZEN, T. E., 2000. Ekonomika zdrowia [Health Economics], PWN [Polish Scientific Publishers PWN] , Warszawa (in Polish).

Glossary of Terms used in Health, 1984, WHO, Geneva.

GRABIŃSKI, T., WYDYMUS S., ZELIAŚ A., 1989. Metody taksonomii numerycznej w modelowaniu zjawisk społeczno-gospodarczych [Methods of numerical taxonomy in modeling of socio-economic phenomena], PWN [Polish Scientific Publishers PWN], Warszawa (in Polish).

Health Promotion Glossary, 1998. World Health Organization, Geneva.

HELLWIG, Z., 1968. Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom rozwoju oraz zasoby i strukturę wykwalifikowanych kadr [Use of taxonomic method for the typological division of countries based on their level of development and the resources and structure of qualified personnel], Przegląd Statystyczny [Statistical Review] (in Polish).

HOSSMAN, I., KARSCH, M., KLINGHOLZ, R., KOHNCKE, Y., KROHNERT, S., PIETSCHMANN, C., SUTTERLIN, S., 2008. Europe's demographic future; growing imbalances [summary], Berlin Institute for Population and Development, Hannover, Germany.

http://pl.wikipedia.org/wiki/Marc_Lalonde (26 February 2012).

http://www.stat.gov.pl/cps/rde/xbcr/gus/POZ_Zasluzeni_statystycy_dla_nauki.pdf (5 March 2012).

http://pl.wikipedia.org/wiki/Europejski_Obszar_Gospodarczy (6 March 2012).

http://www.fordham.edu/halsall/source/pop-in-eur.asp (26 March 2012).

KORCZAK, C., J. LEOWSKI, J., 1977. Problemy higieny i ochrony zdrowia [Problems of hygiene and health], WSiP, Warszawa (in Polish).

LALONDE, M., 1974. A new perspective on the health of Canadians. A working document., Ottawa: Government of Canada.

LISOWSKI, K., 2000. Żyjmy dłużej 6, Medycyna chińska [Let us live longer 6, Chinese Medicine] (in Polish).

MALINA, A., ZELIAŚ, A. (1998). On Building Taxonometric Measures on Living Conditionsm, Statistics in Transition, vol. 3, No. 3.

MASLOW, A., 2006. Motywacja i osobowość [Motivation and personality], PWN [Polish Scientific Publishers PWN], September (in Polish).

MŁYNARSKA-WICHTOWSKA, A., 2004. Finansowanie ochrony zdrowia w krajach UE [Financing health care in EU countries], Kancelaria Sejmu, Biuro Studiów i Ekspertyz, Wydział Studiów Budżetowych [Chancellery of the Seym, the Office of Research, Division of the Budget Study] (in Polish).

SIWIŃSKA, V., BROŻYNIAK, J., IŁŻECKA, J., JAROSZ, M. J., ORZEŁ, Z., 2008. Modele systemów opieki zdrowotnej w Polsce i wybranych państwach europejskich [Models of health care systems in Poland and selected European countries], Praca poglądowa, Zdrowie Publiczne [Public Health] (in Polish).

The WHOQOL Group, Measuring quality of life, World Health Organization, 1997; www.who.int/mental_health/media/68.pdf

WŁODARCZYK, W. C., 1996. Polityka Zdrowotna w społeczeństwie demokratycznym [Health policy in a democratic society], Uniwersyteckie Wydawnictwo Medyczne „Vesalius" [University Medical Publisher "Vesalius"], Łódź-Kraków-Warszawa (in Polish).

WŁODARCZYK, C., POŹDZIOCH, S., 2001. Systemy zdrowotne. Zarys problematyki [Health systems. Draft of issue] , Wydawnictwo Uniwersytetu Jagiellońskiego [Jagiellonian University Publisher], Kraków (in Polish).

ZELIAŚ, A., MALINA, A., 1997. O budowie taksonomicznej miary jakości życia. Syntetyczna miara rozwoju jest narzędziem statystycznej analizy porównawczej [About the construction of taxonomic measure of quality of life. A synthetic measure of development is a tool for statistical comparative analysis], Taksonomia z. 4 (in Polish).

ZELIAŚ, A. (2002). Some Notes on the Selection of Normalization of Diagnostic Variables, Statistics in Transition, vo. 5, No. 5.

STATISTICS IN TRANSITION-new series, Summer 2012
Vol. 13, No. 2, pp. 365—386

STATISTICS AND SOCIOLOGY: THE MUTUALLY-SUPPORTIVE DEVELOPMENT FROM THE PERSPECTIVE OF INTERDISCIPLINARIZATION OF SOCIAL RESEARCH[1]

Włodzimierz Okrasa[2]

ABSTRACT

Statistics emerged as a scientific discipline and has been developed as such, especially extensively over the past century, not only due to an extraordinary service it provides to other disciplines but also thanks to ideas, questions and approaches originally formulated in different fields of empirical research, including sociology, which also contributed to statistics. The confluence of developments in these two disciplines (statistics and sociology) seems to be one of the most successful and beneficial for both of them. Yet, it has become a focus of systematic reflection only recently. The aim of this paper is to make a concise overview of the logical scheme of this interaction and to stress the importance of the counterfactual causal modeling being currently under constant refinement. A more explicit formula of interdisciplinarization that underlies such an interaction anyway would add to overcoming the methodological challenges it poses to either discipline. While contributing to the advancement of 'cause-and-effect' oriented quantitative sociology this would enhance methodology of social science research in general.

I. The key aspects of the interplay between statistics and sociology

In most of historical essays on the development of statistics that account for other disciplines' contributions to it – for instance, such as of S. Stigler's (1986) narrations on how statistics arose from the interplay of mathematical concepts with astronomy, geodesy, experimental psychology, genetics, and sociology – the last one, sociology, is typically being mentioned among the leading contributors. On the one hand, studying impact of statistics on other scientific fields, might allow us to look at the history of science through the window of its own progress

[1] This article is based on a paper presented at the Congress of Polish Statistics: 100th Anniversary of the Polish Statistical Association, Poznan, April 18-20, 2012.

[2] Central Statistical Office (GUS) and the University of Cardinal Stefan Wyszynski in Warsaw (UKSW).

(as suggested by Fienberg, 1992). And such a perspective seems to be an obvious way to trace the role statistics plays in the growing process of interdisciplinarization of research. On the other hand, adopting a more disciplinary-oriented view – as in this paper, with its primary interest in the question of whether and how has sociology played a part in the development of modern statistics – the issue goes beyond historical aspects. This paper focuses on the elements of confluence or 'commonalities' in the methodological advancement of both disciplines, statistics and sociology, and how this interplay contributes to the interdisciplinarization of social research.

Interdisciplinarization of empirical research seems to be an intrinsic effect of employing statistics. As J. Neyman used to say, "[S]tatistics is the servant to all sciences" (cf. Chiang), though in rather different way for each discipline. The nature of this service is, to a large extent, shaped by methodological framework and an overall research paradigm in particular field of science. Such a view can be justified on the historical ground – in science *en globe*, and also within the picture narrowed to a social science field of research. It allows for identifying a discipline-specific subject matter, on the one side, and statistical issues on the other, just providing a needed base for interaction between them, especially between sociology and statistics. Several statistical methods (developed within particular discipline) have been fruitfully combined on the ground of sociological research. With some time lag, however – for instance, *path analysis* has been imported from biometricians, *factor analysis* from psychometricians, and *structural modeling* from econometricians.

The process of mutual influence between statistics and sociology is taking place across different planes, with diverse intensity and character over the time. A typology of respective approaches can be based on specification of the main object of the focus - being it either (problem-oriented) statistical research method or type of data-oriented approach, or an approach focused on the type of analysis, or their cross-roads between disciplines. The development in quantitative methods in sociology will be discussed together with improvements in statistical data production during the post-war evolution of both disciplines. Looking back, and employing the four phase perspective of historical development of statistics since the mid of XVII to the mid of XX centuries, as it is typically proposed (cf., Fienberg, op cit.), it should be noted that sociology has entered into research interaction with statistics only during the third phase that spanned from 1820 to 1900. Owing it mainly to A. Quetelet, who is considered the father of quantitative social science (e.g., Fienberg, op. cit.) and this phase is called 'the socialization of statistics' due to his concept of "the average man" and work he did on fitting of the normal distribution (that was originally used by Gauss and Laplace in astronomy) to several social data sets. For instance, to the distribution of body heights measured in population of conscripts, and other data produced by censuses and within 'social physics' (Fienberg, op cit., p. 216), as the new field of research was called before the term 'sociology' was coined by A. Comte.

According to its founder, sociology was grounded in statistical data, what has soon been demonstrated in a spectacular way by Durkheim's studies of suicides.

In his more disciplinary-oriented narration on quantification in sociology, Paul Lazarsfeld (1961) talked about two-way path ('the two roots' in his words – op. cit., p. 283) that subsequently led the newly emerged discipline to its so much data-loaded version: English 'political arithmeticians' (W. Petty in 17th century) and German universities engaged in description of state's features. Originally it was made in non-quantitative way, as it did G. Achenwall, who created the term 'statistics' in 18th century; however, according to Lazarsfeld, it was Herman Conring who already in 17th century elaborated a system of tables (rows and columns), laying ground for 'university statistics' (*ibidem*, p. 286-291). Within each of these two paths efforts have been made to generate statistical knowledge that was indispensable for what Foucault called 'correct government'. Nb. as a university discipline, quantitative sociology was established at the end of 19th century, with the appointment of Franklin H. Giddings as professor of this field at Columbia University (1894)[1]

- **Two-way interaction**: ***statistics in sociology*** and ***sociology of statistics.*** Statistics, especially official statistics – meant as figures such as consumer prices, the unemployment rate, economic indicators, or data on crime, divorce, and poverty, being produced by government – affect both public and individual decisions in many ways. In doing so, official statistics do not only holds a mirror to reality but - as it was noted by Alonso and Starr (editors of the remarkable collection of papers entitled "The Politics of Numbers" (1988) - they also reflect presuppositions and theories about the nature of society. Such a reciprocal influence might be described with Winston Churchill's metaphor that *we shape our buildings and then they shape us* (Alonso and Starr – op cit., p. 3). And it does involve the sociological question of how quantification and statistical algorithms contribute to shaping the social world, as asked by Desrosierès (2011), who proposed a framework for analyzing this issue based on the idea of two-dimensional construction. The first concentrates on (i) how society and the economy are being conceptualized; (ii) modes of public action, and (iii) different forms of statistics and of their treatment. The second encompasses the five models of state which differ in terms of specific to each of them statistical instruments, such as (Desrosierès, op cit. p 42-45): (a) the French-style *engineering state* and centrally-planned economies – statistics specific to it focuses on production in physical quantity, input-output table, material balance and demography; (b) the *classical liberal state,* concerned about equal access to information for all stakeholders – statistics promoting market transparency,

[1] In addition to F. H. Giddings as the first who institutionally brought statistical methods to sociology, the authors identify such pioneers in other disciplines as well: McKeen Cattell in psychology, F. Boas in anthropology, and H. L. Moore in economics (all faculty members at Columbia University, which "provided a conducive setting for the interdisciplinary process of the incorporation of statistical methods into the social sciences" (Camic, Xie, 1994, p. 773.).

and market shares; (c) the *welfare state* concerned with employment and family protection schemes – interest in labor statistics, consumer price indexes, and sampling surveys of households; (d) crisis-related model of the *Keynesian state*, with policies dealing with aggregates (models like of Lawrence Klein's/LINK) – main interest in national accounting, in the figures on consumption and the price index; (e) *the neoliberal state* (that emerged in 1990's) focused on microeconomic dynamics – interest in incentive systems (according to the rational expectations theory), and in statistical indicators to evaluate performance.

The relationship between statistics and sociology that is of interest here is two-fold – methodological and institutional. From the methodological angle, statistics is meant as the means of data generating and analysis; its particularly extensive use observed in such disciplines as psychometrics, econometrics (the term 'sociography' was also proposed by F. Tönnis around the turn-of-the-century, though unsuccessfully). From the institutional standpoint, statistics is a domain of public activity ('public statistics'). One of the means to institutionalize the statistics was census of population, which is traced to ancient societies[1]. Institutionalization is being manifested in the best way by government statistics or official statistics, which itself becomes a research object within *sociology of official statistics*. It aims to shed light on "why and what is being counted – or sometimes even more interestingly – not counted" (as phrased by G. Sternlieb (1973) to whom creation of sociology of statistics is being attributed – see Starr (1987)). A classic framework of sociology of statistics encompasses, according to Starr (op cit.) the five main areas of interest: *system origins* and development – s*ocial organization* of statistical systems, including relationships between the agents involved in the distribution and use of the statistical information – c*ognitive organizations* of statistical system meant as the intellectual construction of presuppositions, rules, categories of classification and methods of measurement used to produce the information by the statistical institutions – *system uses* (how production and distribution of statistical information shape policy-making and society) – and contemporary *system changes*.

- **Approaches 'to historicize' the interplay of statistics and sociology over time.** The above mentioned three (post-war) generations of quantitative methods in sociology which have emerged in connection with the kinds of data they address, can be conceptualized after Raftery as follows: (i) the late 1940s to early 1960s – focus on cross-tabulations and on measures of association and *log-linear models* (the area of statistics to which sociology has contributed the most, following Pearson's and Yule's earlier ideas of cross-product and odds ratio); (ii) the second generation, which began in the

[1] Census was meant in ancient Rome as "a register of adult male citizens and their property for purposes of taxation, the distribution of military obligation, and the determination of political status." – see Wilcox W.F., 1930. "Census", *Encyclopedia of the Social Sciences*, New York: Macmillan.

1960s and continued to early 1980s, dealt with unit-level survey data (facilitated by the availability of high-speed computers), with special focus on *path analysis* and LISREL-type causal models, and *event-history analysis;* (iii) the third generation started in the late 1980s, with methods dealing with 'newer data forms' that comprise data that do not fall into none of the above categories or forms, such as *spatial* or social *network* data, and *texts* or *narratives*. Somewhat more detailed specification of phases in post-war era of statistics-and-sociology methodological interaction, albeit for only until 1980s, was proposed by Bernert (1983): post-war to 1950, after 1950 to late 1960s, and 1980s. [Omitted here is the pre-war era (categorized by Bernert into three periods: before WW1, 1920-30, and 1930s) during which so-called '-tricians' have contributed to statistical methods -biome*tricians* (path analysis), psychome*tricians* (factor analysis), and econome*tricians* (modeling theory) – which next attracted quantitative sociologists; especially those engaged in causal analysis, during the whole post-war period (until 1980s).]

Out of all the data-and-method types of advancements in methodology of social research during the post-war era it perhaps was *sampling survey* that has been of particular interest to sociologists – as 'a telescope on society' (House et al., 2004) – and influenced the interaction between statistics and sociology in a special way. Modern representative sampling method was, according to Kish (1995)[1] and others (e.g., Heeringa et al., 2010) possible only thanks to earlier works of J. Neyman (especially, his path breaking article (1933) that laid the ground for theory of inference along with works of R. Fisher (1935) and his collaborators involved in survey sampling (Cochran, Mahalanobis, Yates, Snedecor). In particular, sociologists were keen to the 'design-based' theory of inference employed to data generated under probability sample (following Neyman's original framework), as alternative to model-based method of inference from data collected under model-dependent sample. [A typology proposed by Hansen et al., (1983) that results from cross-classification of 'sampling plan' (probability and model-based sample) and 'method of inference' (design-based and model-based inference) does not involve non-probabilistic methods of sample selection – quota sampling, convenience sampling, snowball sampling and 'peer-nomination' – which, however, have also attracted the survey practicioners, including sociologists interested in variety of purposive sampling.]

Some of the major issues in quantitative methodology which have marked the post-war era are briefly discussed below in line with the above quoted periodization (i.e., following Raftery – op. cit., taking also into account some suggestions from Bernert, 1983 and Kish, 1995). In addition to the already stated question of mutually-supportive methodological developments in statistics and

[1] According to Kish, an official birth date for survey sampling marks the publication of paper by A. N. Kiaer (1897) on 'representative method' in Bulletin of International *Statistical Institute* (Kish. 1995).

sociology, some remarks will also be added on one of the hottest topic discussed in the literature, namely, the epistemological status of statistical associations (correlation) and causality, with special emphasis on *counterfactual framework* of causation in social research.

II. Three generations of interactional development

1. Association model methodology – *models for cross-tabulations*

Two types of analytical tools, *loglinear models* and *latent class models,* have emerged during the two post-war decades and contributed significantly to better use of categorical data in social research, and subsequently in other disciplines as well. The first one started with data on social mobility; the other with multivariate discrete data (such as indicators of directly unobserved constructs).

- *Loglinear models*: Thanks to them the social mobility tables have been given an appropriate specification in terms of associations – with separation of structural mobility from circular or exchange mobility – in the model originally defined by Birch (1963) as *loglinear model* for the observed counts $\{x_{ij}\}$:

$$\log(E[x_{ij}]) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$$

where *i, j* are index rows and columns, respectively; *u1(i)* and *u2(j)* are the *main effects* for the rows and columns, *u12(ij)* is the *interaction* term.

Since the model's applicability was limited given a large number of parameters required, Duncan (1979) and Goodman (1979) provided a more general approach to modeling the interaction terms using association model. Thanks to describing associations in terms of local odds ratios, Raftery and his colleagues (see Raftery, op. cit., p. 7)) managed to analyze the tables comprising 7,000 cells and five dimensions: (1) father's occupation, (2) offspring's occupation, (3) gender, (4) race, and (5) period. One important result of employing such models (by different authors) was obtaining a better picture of social mobility in westerns societies (e.g., that the mobility has been increasing in USA; that social mobility increased by 1 percent a year in industrialized countries, through the 2nd half of the 20th century). It also allowed clarifying some of social policy relevant issues like relationship between household vulnerability and recurrent poverty (Okrasa, 1999). In order to determine how vulnerability affects the household welfare status over time, the association between the two variables was estimated using a version of the above model for poverty pattern over time (welfare-path or trajectory) and household vulnerability (Okrasa, *ibidem*):

$$\ln(m_{ij}) = \mu + \lambda_i^{\text{welfare-path}} + \lambda_j^{\text{vulnerability}} + BU_iV_j$$

where μ represent the constant and lambdas are the main effects parameters, and *B* is a regression coefficient multiplied by the scores U and V assigned to the cell

at i row and j column. As expected, chronically vulnerable households were facing substantially higher risk of becoming long-term poor than non-vulnerable households. The odds of being in poverty for at least one year, rather than remaining outside the poverty during the whole four year period under study were nearly twice as high (1.8) for vulnerable households than for non-vulnerable households. The same odds ratio holds between any two adjacent categories of the poverty pattern over time.

It is worth mentioning that association models – as shown by Goodman (1985) – are related to *canonical correlations* and to *correspondence analysis*. According to Raftery (op.cit.), the use of association models has diffused from sociology to other disciplines, such as epidemiology.

- *Latent Class Models (LCM):* The latent class models were introduced to account for observed associations in multivariate discrete data (Lazarsfeld (1950), Goodman (op. cit.)). To some extent, it was done in analogy to the idea of *factor analysis* for multivariate continuous data, or to *cluster analysis* for the case of binary data (but knowing a priori how many clusters there might be, and that that allocation of an object to a cluster is probabilistic – Bartholomew et al., 2002). In brief (for details see for instance Bartholomew et al., *ibidem*, pp. 235-245) it is assumed that every object (out of n objects consisting a random sample from some population) belongs to just one of the J latent class.

Let $\pi_{ij} = Pr(x_i = 1|j)$ be the probability that an object from class *j* will answer positively to item *i* (i = 1, … ,p; j = 1, … , J); and η_j be the proportion of the population in latent class j (or equivalently the probability that a random selected object from the population belongs to latent class j, for (j = 1, … , J). The probability of giving a positive response to a particular item is the same for all objects in the same class; given the latent class to which an object belongs, its response to different items is conditionally independent. And η_j be the proportion of the population in latent class j, or equivalently the probability (the *prior* probability) that a random selected object from the population belongs to latent class j, for (j = 1, … , J). The model is fitted iteratively to obtain maximum likelihood estimates, $\hat{\pi}_{ij}$ of π_{ij} , and $\hat{\eta}_j$ of η_j. Allocation to classes takes place according to (the *posterior*) probability that an object with a particular response pattern falls into a particular class, is: $Pr(\text{object is in class } _j| x_1, \ldots, x_p)$, (j = 1, … ,J).

It is difficult sometimes to distinguish empirically between a latent class model and a latent *trait* model. For instance, Bartholomew and others indicate on the number of determination parameters – if it is large, a latent class model seems to be more appropriate (*ibidem,* p. 247). Several different versions of the extension of the basic latent class model have been proposed in the literature since its origination until recently – including introduction of the Markov Chain Monte Carlo (MCMC) methods for parameter estimation (one of the features facilitated by substantial growth in power of computers – cf. Bartholomew et al., 2011).

2. Unit-level survey data

The second generation of interaction between statistics and sociology is (according to Raftery, op. cit.) by researchers' efforts to shift from *functional* to *causal* type of analysis.

- *Path analysis & structural equation models.* Macro-sociologists preoccupied with analysis of stratification process in terms of inter-generational changes in occupational status have moved from linear regression toward a causal interpretation based on Blau's and Duncan's (1967) *path model* as the archetype model of occupational attainment (following an earlier idea of Wright (1921)). It started with the Duncan Socioeconomic Index (SIE) (according to which the prestige score of 45 occupations presented in the 1960 Census, based on the proportion of occupational incumbents who have completed high school and those who have earned more than $10 000, was predicted with high $R^2 = 0.91$ - see Raftery, p. 11). It was extended to a more general methodology of *structural equation models.* According to the Balu and Duncan famous model, which was replicated in various modifications and in different contexts in the literature for decades, two pairs of sequential variables were responsible for variability in occupational status. Namely, respondent's occupation showed to be of an effect of father's education and father's occupation (affecting mutually each other) which subsequently influenced both respondent's education and respondent's first job.

  Decomposition of a total effect into direct and indirect effects in structural model, along with causal interpretation given earlier to them by Blalock (1961) and, on the other hand, with presentation of the measurement model for unobservable latent variables (such as ability or attitude or motivation, etc.) while treating it as an integral part of the 'causal relation' model, led to linear structural relationship methodology (including LISREL-software by Jöreskog (1973)). Its different versions for various types of data – including categorical data, longitudinal data, and multilevel data, sometimes in combination with other models (e.g. graphical models, for instance by computer scientists) – continue to develop as one of the powerful tool for modeling causal type associations in sociological generalizations or theories. However, the causal interpretation of structural models' parameters has being vigorously criticized during the past few decades both by statisticians (Freedman, 1987, Sobel, 1995) and social science methodologists (Morgan and Winship, 2007). But this issue, and the state of the current discussion on modeling causation based on non-experimental design or observed data, represent a separate kind of problem that deserves discussion, and can only be briefly mentioned below.

In order to complete an overview of the methodological innovations which are worth mentioning as exemplifying another aspect of cooperation between statisticians and sociologists (during the second generations of their developments in Raftery's periodization) two other data-suited methods deserve to be brought up: *event history* analysis and *limited dependent variable*s.

- *Event history* methodology for analyzing data on such occurrences as marriages, divorces, births, job changes, going on or off welfare, along with factors influencing such events – e.g., as dropping out school, or falling in (or moving out of) poverty (Okrasa, 1999) – was introduced by Cox (1972). The Cox proportional hazard model has brought together essential elements of two approaches being previously in use for working with event-history data: life table analysis to count for time (from demography) and regression analysis to also count for time-related influencing factors (but troubled with the problem of censoring). Of particular interest to sociologists showed to be a class of discrete-time event-history models (Allison, 1982), extended to such complex areas of study as innovations and social influence due to, among others, using accelerated failure-time models rather than proportional hazards models (Yamaguchi (1994) – cf. Raftery (op. cit., p. 15).
- *Limited dependent variables*, as a method for analyzing categorical dependent variables – a special case of which is binary dependent variable, though nominal and ordinal variables are included too – is another example of a procedure in the development and implementation of which sociologists have been effective since its commencing (following its earlier application in health research) making significant contribution (cf. Long, 1997). *Logistic regression* showed not only to suit better many sociological data – not necessarily with binary dependent variables (e.g., Agresti, 1990) – than does linear regression model. It also is preferred by sociologists over another alternative to linear regression, such as *probit analysis*, due to convenient interpretation of the logistic regression coefficients in terms of odds ratios. From among different extensions of the limited dependent variables, two types of models are perhaps most significant and useful. One was developed by econometricians in the context of evaluation research (such as sample selection model with two-stage estimator – see Heckman, 1979). Another type of model was elaborated in the context of compositional data such as analysis of household expenditures represented in form of vector of shares of particular consumer items (summing up to one).

3. Causality and spatial data

According to the adopted here perspective (after Raftery, op. cit.) on the confluence of statistical and sociological methodologies based on types of data emerging as specific to a given period of time, in the last (third) phase of the post-war era the following new kind of data showed to be of special interest to sociologists: social networks and spatial data – textual and qualitative data –

narrative and sequence analysis. For each of them new statistical methods have been proposed, but given the limitation of space it is not possible to characterize them here. Therefore, leaving aside all other types of data some remarks confine to spatial data and spatial analysis.

Spatial methodologies. Spatial data and analysis are, surprisingly enough, considered to be less advanced in sociology than in other disciplines (e.g. Urry). This contrasts with the fact that not only social data are inherently spatial – since all social phenomena take place in space – but also because they seem to provide a natural basis for integration of social science (Goodchild et al., 2000) and for interdisciplinarization of social research (Okrasa, 2012). According to Fischer and Getis (2010, p.2) there are disciplinary distinctions in the interest in particular types of spatial data: sociologists share strong interest with anthropologists and geographers in point and area (chloropleth) data; economists and political scientists are especially fond of time series data; environmentalists use remotely sensed spatial data; planners, especially in domain of transport, prefer to work with network data. These distinctions are, to some extent, reflected in different 'schools of thought' on spatial analysis methodology. There are four such schools identified (*ibidem,* pp. 3-6), which however overlap with each other in terms of certain concepts and statistics (such as spatial autocorrelation statistics), as follows: (i) *exploratory spatial data analysis (ESDA*) which is an extension of so-called Tukey-type data exploration (or Tukey's exploratory data analysis /EDA) to geo-referenced data and is traditionally used prior to the model building phase, employing such summary measures like histograms or box plots, or three dimensional scatter diagram; (ii) *spatial statistics*, with focus on testing map patterns of phenomena (like crime of diseases), the tendency for them to cluster or disperse, and how spatially defined objects interact with one another, either statically or over time; (iii) *spatial econometrics*, which flourished after 1988 – the date of publication Anselin's influential book (under such a title) in which the well-established econometrics is related to spatial models in the form of the fundamental regression tools of the spatial econometrics, with recent extension in geographically weighted regression (GWR) to a spatial econometric system that allows regression parameters to vary over space (Fischer and Gits, *ibidem,* p. 5); (iv) *geostatistics*, defined as a way to describe and explain physical phenomena in a continuous spatial data environment (with such major topics as study of *variograms* or using *kriging,* predictive techniques of simulation applied in mostly natural resource exploration and earth science). So far, the latter are rarely used in sociological analysis, but the previous ones are being employed in analyzing spatial aspects of phenomena (ideally, using GIS data), such as inequality, power, politics, interaction and social network, community, social movements, poverty, deviance, crime, study of residential segregation, etc.

Association and causality *– counterfactual account in statistics and sociology.* Within the traditional (Hempelian) explanatory framework that has prevailed over the post-war methodology of empirical research for a while to establish causality

seemed to be like finding the Holy Grail of theorization (especially within the neo-positivist perspective). However, according to Bunge (1979) quoted by Bernert - op cit.,) not all thinkers shared such a view – for instance causality has been declared a 'fetish' (Karl Pearson), or the 'relic of a bygone age' (Bertrand Russell), or a 'superstition' (Ludwig Wittgenstein) and a 'myth' (Stephen Toulmin), (*ibidem,* p. 231). Moreover, sociological interpretations of underlying mechanism being called into question to provide explanation in terms of causality has not generally been compatible with interpretation of association by statisticians, who typically avoid such a language. Some of them criticized causal type interpretation of earlier versions of path analysis coefficients and structural equation models often employed by sociologists (Freedman, 1987, Sobel, 1998). In brief, non-experimental, observational data which social researchers typically work with put several challenges to classic causality-based explanatory framework. On the other hand, this has inspired social science methodologists to deal with some dilemma and controversies – starting with revising some of unnecessary assumptions, such as (that) explanation is synonymous with causal explanation or that to explain also means to predict. For instance, the latter may be met by spurious causation or time series, while sociologists called for preventing against taking *Granger case* for causation (e.g., Goldthorpe, 2001) stressing that causality needs more than *robust dependency.* [So-called 'Granger causation' was originally discussed in the context of econometric time-series analysis (Granger, 1969) demonstrating that predictive power cannot be a criterion for causality which is meant on this ground as follows: "A variable X 'Granger causes' Y if, after taking into account all information apart from values of X. these values still add to one's ability to predict future values of Y" (*ibidem,* p. 2).] And that, in addition to formal conditions – considered the "holy trinity" of causality (association, asymmetry and non-spuriousness or 'third variable problem', e.g. Mutz, 2011, p. 9) – should involve a type of a 'generative process' within a 'substantive' model' that meet a 'pragmatic utilization' criteria while being empirically testable (*ibidem,* p. 15). Such a concern was raised in line of earlier sociologists' caution – Bernert indicates to Znaniecki's (1934) warning against the danger of withdrawing 'into the realm of pure mathematical concepts' (*ibidem,* p. 233) while speaking only of association in lieu of cause (*ibidem* p. 245).

Somewhat alternative to the Goldthorp's approach to causality though congruent with it in ignoring statistical models – including recent advances in counterfactual causal analysis (discussed below) – is a mechanism-based approach to causality being under intensive development during the past decade within *analytical sociology* (AS). It is a new approach in sociology, focused on providing explanation (in the vein of Hempel's covering law models of explanation) "by specifying mechanisms that show how phenomena are brought about" (Hedström, 2005, p. 24). It seems worthwhile to mention about AS for its proneness to greater interdisciplinarization of social research as being open to

various types of the mechanism and for its distinguishing feature which is its aspiration to provide "a syntax of explanation…(that) comprises a set of constraints on how an explanation should be constructed and empirically tested" (cf. Manzo, 2010). It is enough to confine these remarks to express the view that this approach could benefit from extending by directly embracing the counterfactual model of causality (e.g., Morgan and Winship, op cit., p. 233).

The issue of causality was more or less explicitly present in practically all the important classes of methods which were developed – some with active involvement of sociologists – during the period 1950-90, such as (e.g., Heeringa et al., 2010, op cit.,): log-linear models and related methods for contingency tables – generalized linear models, e.g. logistic regression –survival analysis model – general linear mixed models (hierarchical linear models) – structural equation models – and latent variables models. Following Lazarsfeld's 'elaboration' approach to explore relationships among the measured variables (i.e., to explain the correlation), sociologists sought to accomplish what multiple regression methods could do in experimental situations through employing path-modeling with latent variables in non-experimental analysis (i.e., using survey data in *analytical* as opposite to *descriptive* mode in Skinner's terminology – see Heeringa et al., op cit. p. 4).

During the 1970's and 1980's sociologists were refining *structural equation models* – following two paradigmatic traditions: a 'Simon-Blalock tradition' and 'Duncan-Blau' (path analysis) tradition. Further technical advancement in structural equation modeling (SEM) was due to the graphical presentation of the problem by computer scientists (Pearl, 2000). But none of them was recognized by statisticians as a satisfactory way of dealing with cause-and-effect question which motivate much of the empirical (non-experimental) work in social science. The new opportunity for quantitatively oriented social sciences emerged with moving beyond the structural equation and path-model techniques that have been prevailing during 'the age of regression' – when some of their leading proponents, notably Blalock (1964), overstressed "focus on causal laws as represented by regression equations and their coefficients' (*ibidem,* p. 177).

The move was toward experimental type of reasoning based on 'potential outcomes framework' (POM) proposed by Neyman (1923) formalized as the *counterfactual model of causal inference* (CMCI) for observational data by statisticians (Rubin, 1974 and series of works by him and his colleagues, Holland, 1986, Rosenbaum, 1989), and by econometricians (especially, works of Heckman, 1989 and Manski, 1995). The counterfactual model that brings experimental language into observational data analysis ('experimental metaphor') – in line with Stouffer's advice for sociologists to "always keep in mind the model of a controlled experiment, even if in practice we may have to deviate from an ideal model"– is now being used with increasing frequency in sociology, psychology and political science, see Morgan and Winship (op. cit., p. 4-7)". Most of the paradigmatic applications of this model which they discuss are provided by studies of cause-and-effect type of problem in the evaluation context

– for instance: (a) the causal effect of school vouchers on learning (Chubbs and Moe, 1990); (b) the causal effect of Catholic schooling on learning (Coleman and Hoffer, 1987); (c) the causal effect of manpower on earnings (LaLonda, 1995) [for instance, 'treatment' is having or not having college degree and earning (as an effect) – four theoretically possible cases/answers to 'what-if earnings' for each person, including two what-is potential outcomes which are counterfactual: adults who have completed high school, *what-if* earning under the state „having a college degree", and adults who have completed college degree, *what-if* earning under „have only a high-school degree", etc.]; (d) the causal effect of alternative voting technology on valid voting ((Wand et al., 2001).

Operationally, the essence of the counterfactual model of causality (or CMCI), a counterfactual account ("what would have been") that is built-in potential outcomes framework, refers to the relationship between potential and observed variables given the causal states, 'treatment' and 'control', respectively. Interestingly enough, the statistical models of counterfactual causation have been developed independently on counterfactual theories of causation proposed in the philosophy of science following David Lewis's (1973) conceptualization of 'world semantics' for counterfactuals ("If A had not occurred, C would not have occurred"). The central notion of this semantics is a relation of comparative similarity between possible worlds (for a pair of possible worlds, the world that resembles the actual world more than the other is said to be closer to actuality) – see Menzies (2008, p. 3). It was only after innovative re-formulation of the counterfactual causation in graph-theoretic version of the structural equation framework by Pearle (2000/2009, op cit.) when the two strands of the literature met each other (Hitchcock (2001), Woodward (2003). Some authors (e.g., Kluve, 2001) distinguish between three approaches to modeling causation for observational data: structural equation models (SEM), potential outcome model (POM), and Directed Acyclic Graphs (DAG) dominating respectively in econometrics, statistical, and formal (or computer-science) domains of applications. However, others consider these approaches as different stages in the development of structural equations for modeling causality rather than separate methodologies – see Grace et al., (2012). According to Grace et al., the first generations of SEM consists of the early works (following Wright's (1921) ideas) of path analysis that spreads to econometrics in 1940s, and to sociology (Blalock 1964), being limited however to the analysis of correlation matrices. The second generation (continuing to the present) consists of the already mentioned works of Jöreskog's (in 1970s.) synthesizing factor and path relations under the LISREL model involving both latent and observed variables. The third has just begun with contribution from computer science (Pearl 2000/2009).

The statistical counterfactual causal modeling gained recently new impetus from a graph-theoretic implementation of structural equation modeling (following ideas and methods proposed by Pearl (2000/2009)) that subsumes the historical matrix approach and is incorporated into general SEM practice (Shipley, 2009). Most characteristic of this new approach – reckoned by Grace et al., (op cit.) as

third generation of structural equations, which are considered by Pearl (op. cit.) the natural language for representing and studying causal relations – is the generalization of the structural equation model as a causal graph along with extending it to the nonparametric level.

For the purpose of presenting the idea of the statistical approach to the problem – let Y^t and Y^c denote potential outcome random variables defined over all individuals in the population under study. Accordingly, Y_i^t is the potential outcome in the treatment state for individual i, and Y_i^c is the potential outcome in the control state for individual i. And the individual-level causal effect of the treatment is defined as:

$$\delta_i = Y_i^t - Y_i^c .$$

For two causal states (a binary case), a causal exposure variable D is equal 1 for members of the population who are exposed to the treatment state, and equal 0 to others (non-exposed to the treatment). Formally, the observed outcome variable Y is defined as

$$Y = Y^t \text{ if } D = 1$$
$$Y = Y^c \text{ if } D = 0$$

Or, equivalently

$$Y = DY^t + (1\text{-}D)\, Y^c$$

Consequently, in terms of the Neyman potential outcomes' conceptualization the problem –called in the literature the fundamental problem of causal inference (after Holland, op cit.) – can be presented as below:

a) Treatment Group $(D = 1)$ & $Y = Y^t \rightarrow$ Observable as Y
b) Treatment Group $(D = 1)$ & $Y = Y^c \rightarrow$ Counterfactual
c) Control Group $(D = 0)$ & $Y = Y^t \rightarrow$ Counterfactual
d) Control Group $(D = 0)$ & $Y = Y^c \rightarrow$ Observable as Y

Since, in reality, the potential outcome under the treatment state can never be observed for those observed in the control state, and *vice-versa* (rows b and c), the outcome variables (such as labor market earning, or test score, or voting results in the above cited examples) contain only a portion of the information needed to calculate individual-level causal effects. [This is why, following Rubin (1978), counterfactual causal analysis at its core is considered a missing data problem (Winship and Sobel, 2001, p. 14).] As a consequence of impossibility to calculate individual-level causal effects, the average treatment effect (ATE) in the population is estimated as the expectation, E(.) of a difference between potential outcomes:

$$E[\delta] = E[Y^t - Y^c]$$
$$= E[Y^t] - E[Y^c]$$

For instance, the average causal effect in the case of study by Coleman et al. (op cit.) is then estimated as the mean value among all students in the population of these what-if differences in test score (for a randomly selected students from the population). Often, the average treatment effect is defined separately – as conditional average treatment effect – for those who typically take the treatment, ATT (assuming the linearity of the expectation operator)

$$E[\delta \,|D=1] = E\,[Y^t \;-\; Y^c \,|\, D=1]$$

$$= E[\quad Y^t \,|\, D=1] - E[Y^c \,|\, D=1]$$

and for those who typically do not take the treatment, ATC

$$E[\delta \,|D=0] = E\,[Y^t \;-\; Y^c \,|\, D=0]$$

$$= E[\quad Y^t \,|D=0] - E[Y^c \,|D=0]$$

Framing the experimental type of reasoning (causal inference) for observational data through employing counterfactual models in social science has resulted in elaborating analytical instruments capable to compensate for inherently deficient data (POM as 'missing data problem'). The matching estimators that are the classic technique for estimating causal effects are typically interpreted either (a) as a method to form quasi-experimental contrast by sampling comparable units from a population under study, or (b) as a nonparametric method of adjustment for treatment in the situation when parametric regression is not persuaded. The significant use of this technique in sociology was as early as mid-1980s. (e.g. by Berk and Newton, 1985). Different procedures of matching estimators are under continuing refinement, such as: matching as conditioning via stratification – matching as weighting (using propensity score as the estimated probability of taking the treatment as a function of variables that predict treatment assignment) – matching as a data analysis algorithm – matching when treatment assignment is non-ignorable (Morgan and Harding, 2006). However, none of these methods is considered perfect, being either limited in practical applications (e.g., only recently included are multivalued treatments to procedures being essentially appropriate for binary treatments or exposures). Also, predicting treatment status from the observed variables with a logit model may not solve the causal inference problem (e.g., if no variable yielding a 'perfect stratification' is known).

In their excellent exposition of methodological issues involved in counterfactual causal modeling, Morgan and Winship (2007/2009) discuss the prospects and limitations of this approach in empirical research in social sciences. At first, they stress its utility *vis-à-vis* traditional structural equation modeling pointing first to data related advantages (such as sufficiency of the requirement that data be balanced with respect to the determinants of treatment assignment,

while easing the problem of omitted variables). On the other hand, they specify some objections to employing counterfactual modeling, such as (1) incapability to non-manipulable causes, (2) inappropriate for discovering the causes of effects, and (3) causal inference should not depend on 'metaphysical quantities' (*ibidem,* p.278-282).

They offer four modes of causal inquiry in observational social science which constitute a sequence of cumulative proprieties: (a) *associational analysis* – association as a precondition of causation; (b) *conditional associational analysis* – to eliminate sources of spuriousness; (c) *mechanism-based analysis* – efforts to provide mechanistic explanation of the process that generates the causal effect; and (d) *all-cause structural analysis* – complete set of variables between the putative causal variable and the outcome variable (the structural equation models in economics provide an exemplary successful usage of such a mode of causal analysis). Of course, the last approach that could be seen as an ultimate goal of causal inquiry – attempting to provide answer to all of the "who, when, where, and how" questions – calls for a theory and data, as any type of causal approaches does. In this point, however, counterfactual models of causation are useful for identifying in a more accurate way which stage (or mode) of causal analysis is actually feasible, given a theory and data availability.

As regards the problem of data that strongly interferes with making a choice among feasible strategies of causal analysis, a final observation relates to an alternative approach to causality – which could be considered as data-based pragmatic approach due to actually abandoning the problem of causality and interpretation of the nature of cause-and-effect relationship. Namely, it is a newly proposed data collection technique which under the name *population-based survey experiments* uses "survey sampling methods to produce a collection of experimental subjects that is representative of the target population of interest for a particular theory" and is easily implemented through utilizing computer-assisted telephone interviewing (CATI) and internet-based interview (Mutz, 2011, p. 2-10). The underlying idea is to enable the researchers to control the random assignment of participants to variations of the independent variable in order to observe their effect on a dependent variable. Their proponents, who tested efficiency of such quasi-experimental reasoning based on sample survey in Time-sharing Experiments for the Social Science (TESS) while stressing internal validity of the approach as its strength are, however, aware of the problem of external validity and generalizability as its fragile part, that needs further elaboration. (see Mutz op cit., p. 22). Anyway, it illustrates another attempt of sociologists and other social scientists to contribute to the statistical methods in the difficult context of establishing empirically unbiased causal inferences. [It seems promising in the theory-based evaluation research context, albeit not discussed in the literature yet.]

III. Conclusions

The mutually supportive interplay between statistics and sociology over the post-war period that contributed to the developments in each of these disciplines substantially has become an object of systematic reflection (starting as early as with Lazarsfeld's historical notes on quantitative sociology in late 1950s and with Bernert's paper at early 1980s of the 20th century – not to mention several important papers written by statisticians, such as by Kish on sampling or Stigler and Fienberg and others on different issues). The confluence could be observed in many areas, especially of those being motivated by going beyond descriptive use of statistics and searching for an explanation – from Lazarsfeld's 'elaboration' program and Goodman's log-linear models through logistic regression to path analysis and structural equation models of causation (Blalock, Duncan, others) to recent counterfactual causal modeling. A lot of new hopes have been created with introduction of the counterfactual model of causal inference for observational data due to, in a summary: (i) the possibility that the size of a treatment effect may vary across individuals (*effect heterogeneity*); (ii) the researcher need to fully specify the *implicit manipulation* or "experiment" associated with the estimation of a causal effect ('no causation without manipulation' as emphasized by Holland and Rubin); (iii) selection of variables as controls given that they substantially change the estimate of the treatment effect; (iv) using *matching* method due to its ability to provide a powerful nonparametric alternative to regression (for the estimation of a causal effect); (v) limiting *instrumental variable* estimators only to estimating the effect of the treatment for those individuals whose treatment status is changed; and (vi) longitudinal data do not suffice for causal inference without prior assumptions about the values of the outcome (under the counterfactual condition).

One of the important lesson learnt from such studies is the need for a more interdisciplinary orientation than it was during 20th century in quantitative methods, marked by isolation from each other (and often from statistics as a whole, as pointed by the critics of structural equation models with latent variables in 1960-70s). Interdisciplinarization is predicted (and advocated) by leading social methodologists and is stressed as chief direction of developments in quantitative sociology. Importance of such an approach has been recognized in contemporary institutional efforts towards *interdisciplinarization* of social science research through establishing several joint programs of research and education, both by departments of statistics and by several social science departments in Europe and the USA. In the context of the Congress of the Polish Statistical Association, it seems worthwhile mentioning that such a kind of institutional interdisciplinary cooperation – sociology and statistics – was created in the late 1960s., along with establishing of a Statistical-Sociological Research Unit at the Central Statistical Office of Poland[1].

[1] The Unit was created under leadership of Aleksander Wallis and resumed (after his depart) by Krzysztof Zagórski. (The author was a part of the 'first generation' Unit's members).

REFERENCES

AGRESTI A., 1990. Categorical Data Analysis. Wiley. N. Y.

ALLISON P., 1982. Discrete-Time Methods for the Analysis of Event Histories. *Sociological Methodology,* 13, pp. 61-98.

ALONSO, W., STARR, P., 1987. The Politics of Numbers, Russell Sage Foundation, NY.

BARTHOLOMEW, D. J., STEELE F., MOUSTAKI I., and GALBRAITH J. I., 2002. The Analysis and Interpretation of Multivariate Data for Social Scientists, Chapman and Hall/CRC, Boca Raton, Fl.

BARTHOLOMEW, D. J., KNOTT M., MOUSTAKI I., 2011 Latent Variable Models and Factor Analysis: A Unified Approach. 3rd edition; *Wiley Series in Probability and Statistics*. John Wiley and Sons, Ltd.

BERNERT, CH., 1983. The Career of Causal Analysis in American Sociology, *The British Journal of Sociology,* Volume 34 Number 2 (June).

BLALOCK, H. M., 1961. Causal Inferences in Nonexperimental Research. New York: W.W. Norton.

BLAU, P. M., DUNCAN O. D., 1967. American Occupational Structure. Free Press, N.Y.

BUNGE, M. A., 1979. Causality and Modern Science. (3rd ed.), New York: Dover.

Camargo Alexandre de Paiva Rio, Sociology of Statistics: possibilities of a new field of investigation, Hist. cienc. saude-Manguinhos, vol.16 no.4 Rio de Janeiro Oct./Dec. 2009.
(http://dx.doi.org/10.1590/S0104-59702009000400004)

CAMIC, CH., XIE, Y., 1994. The Statistical Turn in American Social Science: Columbia University, 1890 to 1915. *American Sociological Review,* Volume: 59, Issue: 5.

COX, D. R., 1971. Regression Models and Life Tables (with discussion). *Journal of the Royal Statistical Society*, Ser. B, 34, pp. 187-220.

DESROSIERÈS, A., 2011. Words and Numbers. For a Sociology of the Statistical Argument, chapt. 2 [in] Saetnan A. R., Lomell H. M., Hammer S. (ed.): The Mutual Construction of Statistics and Society, Routlege, New-York, pp. 41-63.

DUNCAN, O. D., 1979. How Destination Depends on Origin in the Occupational Mobility Table. *American Journal of Sociology* 84:793-803**.**

FIENBERG, S. E., A Brief History of Statistics in Three and One-Half Chapters: A Review Essay, *Statistical Science,* Vol. 7, No. 2 (1992), pp. 208-225.

FISCHER, M. M., GETIS, A., 2010, Handbook of Applied Spatial Analysis. Software Tools, Methods and Applications, Springer, Berlin Heidelberg.

FISHER, R. A., 1935, The Design of Experiments. Oliver and Boyd, Edinburgh.

FREEDMAN, D. A., 1987. As Others See Us: A Case Study in Path Analysis. *Journal of Educational Statistics,* 12: 101-223.

GOLDTHORPE, J., 2001. Causation, Statistics, and Sociology, *European Sociological Review*, Vol. 17 No. 1, 1-20.

GOODCHILD, M. F., ANSELIN, L., APPELBAUM, R. P., HARTHORN, B. H., 2000. Toward Spatially Integrated Social Science, *International Regional Science Review* 23, pp. 139-159, (April).

GOODMAN, L. A., 1979. Simple Models for the Analysis of Association in Cross-Classifications Having Ordered Categories, *Journal of the American Statistical Association* 74:537-52.

GOODMAN, L. A., 1985. The Analysis of Cross-Classified Data Having Ordered and/or Unordered Categories**.** *Annals of Statistics* 13:10-69.

GRACE, J. B., SCHOOLMASTER jr., D. R., GUNTENSPERGEN, G. R., LITTLE A. M., Mitchell B. R., Miller K. M., Schweiger E. W, 2012. Guidelines for a graph-theoretic implementation of structural equation modeling, *Ecosphere* 3(8):73.

HANSEN, M. H., MADOW, W. G., TEPPING, B. J.(ed.), 1983. An evaluation of model-dependent and probability-sampling inference in sample surveys, *Journal of the American Statistical Association,* 78: 776-793.

HECKMAN, J. J., 1979. Sample Selection Bias as a Specification Error, *Econometrica,* 47:153-61.

HEERINGA, S. G., WEST B. T., BERGLUND P. A., 2010. Applied Survey Data Analysis, Chapman and Hall/CRC, Boca Raton, FL.

HITCHCOCK, C., 2001. The Intransitivity of Causation Revealed in Equations and Graphs. *Journal of Philosophy,* 98: 237-299.

HOUSE, J. S., JUSTER, F. T., KAHN, R. L., SCHUMAN, H., and SINGER, E. (ed.), 2004. A Telescope on Society: Survey Research and Social Science at the University of Michigan and Beyond, University of Michigan Press.

JORESKÖG, K. G., 1973. A General Method for Estimating a Linear Structural Equa-tion System. Pp. 85-112 [in] A. S. Goldberger and O. D. Duncan (ed.) *Structural Equation Models in the Social Sciences*, Seminar. N. Y.

KISH, L., 1995. The Hundred Years' Wars of Survey Sampling, chap. 3 [in] Graham Kalton and Steven Heeringa (ed.)**,** *Leslie Kish: Selected Papers,* 2003. John Wiley & Sons, Hoboken, NJ.

KLUVE, J., 2001. On the Role of Counterfactuals in Inferring Causal Effects of Treatments. Alfred Weber Institute, University of Heidelberg and IZA Bonn Discussion Paper No. 354.

LAZARSFELD, P. F., 1961. Notes on the History of Quantification in Sociology--Trends, Sources and Problems. *Isis*, Vol. 52, No. 2. (Jun., 1961), pp. 277-333.

LAZARSFELD, P. F., 1950."The Logical and Mathematical Foundation of Latent Structure Analysis. p. 362-412 [in] *Studies in Social Psychology in World War II. Vol. 4, Measurement and Prediction*, edited by E. A. Schulman, P. F. Lazarsfeld, S. A. Starr, and J. A. Clausen. Princeton, NJ: Princeton University Press.

MANZO, G., 2010. Analytical Sociology and Its Critics, *Archives of European Sociology*. LI, 1 pp. 129–170.

MENZIES, P., 2008. Counterfactual Theories of Causation. *The Stanford Encyclopedia of Philosophy*, Vol: 2011, Issue: Dec. 16, Stanford University.

MORGAN, S. L., HARDING, D. J., 2006. Matching Estimators of Causal Effects: Prospects and Pitfalls in Theory and Practice. *Sociological Methods and Research* 35: 3-60.

MORGAN, S. L., WINSHIP, C., 2007. Counterfactuals and Causal Inference. Methods and Principles for Social Research. Cambridge University Press, N.Y.

MUTZ, D. C., 2011. Population-Based Survey Experiments, Princeton University Press, Princeton, NJ.

NEYMAN, J., 1933. An outline of the theory and practice of the representative methods: The method of stratified sampling and the method of purposive selection (in Polish – English version published in 1934. On the two different aspects of the representative method. *Journal of the Royal Statistical Society*, 97: 558-625).

OKRASA, W., 1999. The Dynamics of Poverty and the Effectiveness of Poland's Safety Net. *Policy Research Working* Paper No 2221. The World Bank, Washington D. C.

OKRASA, W., 1999. Who Avoids and Who Escapes from Poverty during the Transition. Evidence from Polish Panel 1993-96, *Policy Research Working Paper* No 2218. The World Bank, Washington D. C.

OKRASA, W., 2012. Spatially Integrated Social Research and Official Statistics: Methodological remarks and empirical results on local development, Paper presented at the *International Conference of Spatial Econometrics and Regional Economic Analysis*, Łódź, 4-5 June (2012).

PEARL, J., 2009. Causality: Models, Reasoning and Inference. (2nd edition), Cambridge University Press, N. Y.

PEARL J., 2012. The causal foundations of structural equation modeling. Pages 68–91 in R. H. Hoyle (ed.). *Handbook of structural equation modeling*. Guilford Press, New York, New York, USA.

RAFTERY, A. E., 2001. Statistics in Sociology, 1950-2000: A Selective Review. *Sociological Methodology*, Vol. 31, pp. 1-45 American Sociological Association.

SHIPLEY, B., 2009. Confirmatory path analysis in a generalized multilevel context, *Ecology* 90:363–368.

STIGLER, S. M. (1986), The History of Statistics: The Measurement of Uncertainty before 1900, Harvard Univ. Press, p. 410 [Reissued in paperback edition (1990)].

SOBEL, M. E., 1995. Causal inference in the social and behavioral sciences, [in] G. Arminger, C. C. Clogg, and M. E. Sobel ed. *Handbook of Statistical Modeling for the Social and Behavioral Sciences*,Plenum Press, N. Y.

SOBEL, M. E., 1998. Causal Inference in Statistical Models of the Process of Socio-economic Achievement: A Case Study. *Sociological Methods and Research* 27: 318-48.

STARR, P., 1987. The Sociology of Official Statistics [in] Alonso, W., Starr, P., (ed.). 1987. *The Politics of Numbers, Russell Sage Foundation*, N. Y., pp.7–58.

Statisticians in History.
http://www.amstat.org/about/statisticiansinhistory/index.cfm?fuseaction=biosinfo&BioID=11.

URRY, J., 2004, The Sociology of Space and Place, [in] Judith R. Blau (ed.) *The Blackwell Companion to Sociology*. Blackwell Publishing, Malden, MA.

WILCOX, W. F., 1930. Census, *Encyclopedia of the Social Sciences*, New York: Macmillan.

WOODWARD, J., 2003. Making Things Happen. A Theory of Causal Explanation. Oxford University Press, Oxford.

ZNANIECKI, F., The Method of Sociology, Farrar & Rinehart, New York 1934.

STATISTICS IN TRANSITION-new series, Summer 2012
Vol. 13, No. 2, pp. 387—404

FORECASTING OF MIGRATION MATRICES IN BUSINESS DEMOGRAPHY

Piotr Gurgul[1], **Paweł Zając**[2]

ABSTRACT

This paper demonstrates that the forecast of migration matrices can be conducted by means of updating procedures, well-known in the I-O theory. The authors use some of the most popular I-O updating procedures (RAS and some non-biproportional approaches) and calculate measures of the ex-post error of predictions. While taking into account the measures of distance between two matrices, a ranking of forecasting methods of migration matrices (forecast horizon one) is established. Finally, the advantages and drawbacks of particular forecasting methods with respect to one-step ex-post forecasts of migration matrices are discussed.

JEL Classification: C02, L25.

Keywords: Births and deaths of enterprises, migration in branches of industry, prediction, updating methods.

1. Introduction

Firm bankruptcies always bring significant economic losses to stockholders, employees, customers, management, and others. This is accompanied by a substantial social and economic cost to the country. The recent collapse of many global corporations, e.g. Enron, WorldCom, Global Crossing, Adelphia Communications, Tyco, Vivendi, Royal Ahold, HealthSouth, and, in Australia, HIH, One.Tel, Pasminco and Ansett, has aroused interest in various aspects of corporate bankruptcy modelling (also corporate disclosure and auditing regulation). A model forecasting corporate bankruptcy would serve to reduce losses by providing an early warning to stockholders. Early signs of probable distress will enable both management and investors to adopt counter measures. The goal is to shorten the length of time of losses.

[1] Piotr Gurgul, AGH University of Science and Technology, Department of Computer Science, Cracow, Poland. E-mail: pgurgul@o2.pl.
[2] Paweł Zając, AGH University of Science and Technology, Department of Applications of Mathematics in Economics, Cracow, Poland. E-mail: pzajac@zarz.agh.edu.pl.

The present study is an attempt to forecast the birth and death of enterprises within input-output methodology.

The next section contains a short overview of existing models in the literature and the methods used for forecasting insolvencies. In section 3, the methodology of enterprise migration matrices is presented, in section 4 methods for forecasting migration matrices are shown, and in section 5 the results of computations are presented and discussed in detail. Finally, section 6 presents the conclusions of the paper.

2. Literature on start-ups and insolvencies

According to the Austrian economist Joseph Schumpeter (1982) the failure of company results from coexisting processes of destruction and creation. The same point is shared by Foster and Kaplan (2001). The authors are convinced that processes of creative destruction are often started by the financial markets. This takes place because financial markets support companies by supplying financial capital as well as withdrawing it with decreasing competitiveness of the company. A developing company can get financial resources. However, after signs of a reduction in competitiveness this support is refused. In case of bankruptcy the company not only needs to deal with its current activity, but also with the mutually connected processes of destruction and creation.

In the 1960's quantitative methods were commonly used to predict the risk of bankruptcy. The pioneer of such research was Altman (1968). The author applied discrimination methods. Discrimination methods allowed a classification of firms into two clusters, those vulnerable to bankruptcy and those not. The clustering was conducted on the basis of financial indicators defined for a sample of companies. The main indicators included the working capital to total assets ratio, the retained earnings to total assets ratio (retained earnings - profits which have not been paid out in dividends and which can be re-invested in the business), the EBIT (Earnings Before Interest and Taxes) to total assets ratio, the market value of equity to the book value of total liabilities and sales to total assets. By means of these indicators Altman classified correctly up to 95 percent of companies the year before their bankruptcy and 83 percent two years before.

Beaver (1966) started empirical research into the prediction of business failure. He used a univariate model. In spite of some shortcomings of the univariate approach, especially a lack of integration of the various ratios, Beaver's model achieved a very good level of predictive accuracy. However, later researchers applied a multivariate discriminant analysis (for short MDA) developed by Altman (1968). Deakin (1972) considerably improved the accuracy of the forecasts of the Altman model. He used all the fourteen ratios which Beaver had identified as good predictors of bankruptcy. Moreover, he used a probabilistic classification rule rather than a critical cut-off point. The model by Blum (1974) included variables expressed in terms of change over time. Wilcox (1971, 1973)

developed the variables applied in the discriminant function by deriving a random walk with a drift formulation of the transition to the failure state. Diamond (1976) developed the multivariate model by using realistic prior probabilities. He also conducted an error cost estimation.

The research started by Altman and continued by others led to the foundation of the American Bankruptcy Institute (ABI) in the USA in 1982. This institute supplies the Congress and public opinion with reports describing cases of company bankruptcy. In these reports its causes and results are analyzed. This Institute employs attorneys, accountants, auctioneers, assignees, bankers, lenders and academic professors and conducts not only research activities concerning bankruptcy in the USA, but is also involved in many educational projects on bankruptcy. Moreover, the institute owns an extensive empirical dataset with insolvency data. It is also responsible for permanent cooperation with the media. The institute tries to forecast the future of companies. This is very important because of changes in the corporate environment and very complex intercorporate connections that may pose difficulties in formulating a long-term growth strategy. The Institute discusses past events, but the most important part of the ABI's activity is the investigation and forecast of the future performance of American companies.

Greiner and Schein (1988) stressed that the flexibility of a company is a reflection of the creativity of the owner. The improvements resulted from the problems which a company encounters in existence should lead to improvements. The management should recognize problems early on. Otherwise, the company will tend to an end. General uncertainty, a rapidly changing situation, unexpected problems and new goals have an impact on the immunity to change business life cycles, which as a result become shorter. Uncompetitive companies are eliminated from the market.

There are similar ideas in Handy (1995), who formulated what is known as the 'S' shape. According to this entrepreneurs should already be preparing the company for the crisis period even in the time of its prosperity. They should provoke artificial symptoms of crisis and undertake proper countermeasures. These actions can weaken the organization in the short term. However, the changes developed by the fake crisis can strengthen and immunize the company.

Frederick et al. (1988) stressed that company activities are always connected with a company's social function as its resources come from society. If the organization terminates the proper exercise of these functions by not using those resources in a way determined as adequate by society, it starts to tend towards its end. This situation is implied by the decline of the financial and human resources which are necessary for the proper functioning and existence of the company. The insolvency of a company is the result of the social necessity to extend the efficiency of economic processes.

The removal of unprofitable firms, the protection of debtees, employees and the whole country against dishonest debtors as well as the elimination from the market of companies which have lost their cost-effectiveness are main

advantages. As a result of this process a bad company is either reorganized or disposed of which is supposed to protect it and make possible a better usage of social resources.

In practice bankruptcy is mainly a tool of control for the protection of the market. Davydenko and Franks (2008) compared British, French, and German insolvency codes. They checked whether outside creditors can adjust low protection of creditors through changing their rules for granting loans. Kahl (2002) created a model explaining different options for creditors in a dynamic context. This problem became particularly important during the last banking and economic crisis which resulted in a credit crunch. It led to the bankruptcy of many small and middle sized companies.

That is why, in order to answer the question whether and when a company will default, a structured approach was needed. Over the years a wide range of credit risk and distress prediction models have been developed. This created a new economic subfield of corporate bankruptcy prediction. The most recent studies give a good overview of developments in the field of insolvency research. An introduction to the methods of corporate failure prediction are presented by Thiele and Lohmann (1995) or Baetge and Ströher (2005).

In Matschke (1979) previous research contributions on the problems of company classification by MDA in international studies are reviewed. In the seventies and eighties the number of insolvencies and bankruptcies rose. The use of early warning signals of insolvency and distress in the forecasts became very important. Numerous studies, e.g. Eisenbeis (1977), Moyer (1977), Altman et al. (1981), Zmijewski (1984), Platt and Platt (1991) concerned the methodological and statistical problems arising from the application of multiple discriminant models. Thus, new methods, e.g. logistic analysis were developed and became popular.

The last method uses a firm's financial ratios, weighs each of the ratios by its respective weight and aggregate products to a probability measure of insolvency for a company. Corporate failure predictions were constructed with the use of logistic analysis, e.g. Ohlson (1980), Zavgren (1985) and Peel and Peel (1988).

Ohlson (1980) used a logit of the maximum likelihood method to build and analyze a model which sampled 105 failed companies and 2058 non-failed companies from 1970 to 1976. He set up 3 models from 9 explanatory variables to predict corporate failure.

Keasey and Watson (1987) employed a logit to build a prediction model. They sampled 73 failed companies and 73 non-failed companies from 1970 to 1983, using 28 financial variables and 18 non-financial variables in their study.

Logit techniques are free of the restrictive assumptions typical of the MDA. Moreover, they allow an interpretation of individual coefficients. Other methods without restrictive statistical assumptions are non-parametric techniques. The best known of these is the neural networks method. This tries to improve computerized pattern recognition. It develops models based on the functioning of human brain.

The neural network attempts to implement learning behaviour into computing systems.

In spite of the application of new methods to bankruptcy forecasts, MDA is still the most frequently used and developed method in different insolvency problems. In a more recent study Hesselmann (1995) reported that a discriminant approach based on "z-score" model of Altman (1968) achieved a predictive accuracy of 70 percent. Thus, the MDA method became an integral part of the monitoring system in the banking sector. The drawback of the MDA model is, according to the author, the lack of all qualitative factors. Baetge and Ströher (2005), stressed that MDA is not a proper tool for forecasting an acute crisis of a company. They suggested the usage of neural networks combined with neuro-fuzzy systems for an evaluation of a firm's survival chances. However, MDA can help to detect the symptoms of problems which can raise the awareness of a coming crisis. MDA can also help to find the relevant measures against coming distress. The future development of a company can be frequently derived from historical data because certain patterns of the symptoms which can lead to insolvency are visible several years before the death of a company.

Some contributors criticize MDA because of a lack of an economic theory to support this method. Muche (2007) tried to relate his stochastic model of insolvency prediction with the extraction and loading of common financial ratios from financial statements. Agarwal and Taffler (2011) were concerned with accounting-ratio-based models. In their opinion the traditional models were no worse than market-based models for credit risk assessment. Moreover, they bettered them in terms of potential bank profitability, especially in the case of error misclassification costs. Diacogiannis (1996) expressed his doubts concerning taking into account security prices for corporate insolvency forecasts. He argued that macro-economic variables such as inflation could improve the predictive power of the known models.

Jacobson and Lindé (2000) checked whether macro-economic variables can improve the accuracy of prediction models. They worked on credit ratings produced by the two biggest Swedish credit rating agencies monitoring Swedish financial markets. The focus of the research was on the assessment of the stability of a financial system. The authors advised taking into account macroeconomic developments in order to determine risk in the banking system. In their opinion macroeconomic variables have large explanatory power in the explanation of the actual proportion of company bankruptcies. This is important because companies are responsible for a large part of a bank's credit losses. In their recent studies Jones and Hensher (2008) reviewed different modern methods used in the field of credit risk and corporate bankruptcy prediction. They included probit models, advanced logistic regression models, survival analysis models, non-parametric techniques, structural models and a model in reduced form. The contributors advised using several new methods, like the mathematical and theoretical systems known as "belief functions" and insolvency modelling for public sector entities.

The financial statements of companies are often the basis for statistical forecast models like univariate or MDA. The assessment of the likelihood of a firm's insolvency is frequently conducted by market-based econometric models which are based on security prices. This spectrum of models includes logistic regression models and non-parametric techniques, e.g. neural networks and recursive partitioning models. The results of insolvency research can significantly differ from one another. There are many approaches to predicting bankruptcy though the results depend to large extent on the method used.

Recently Gurgul and Zając (2011) introduced the bankruptcy prediction model which was used for empirical purposes. It did not refer to financial indicators but described the processes of destruction and creation of companies in a national economy.

In the next section we will apply different techniques aimed at forecasting the births and deaths of companies in different sectors of national economy.

3. Enterprise migration matrices

Define matrix **X** as follows:

$$\mathbf{X}^{(t)} = \left[x_{ij}^{(t)}\right] = \begin{bmatrix} \mathbf{M}^{(t)} & \mathbf{d}^{(t)} \\ \mathbf{b}^{(t)} & 0 \end{bmatrix}$$

Matrix **X** is formed using the following matrices:

$\mathbf{M}^{(t)} = \left[m_{ij}^{(t)}\right]$ - migration matrix representing the number of enterprises crossing between sectors *i* and *j* during a period from *t*-1 to *t*.

$\mathbf{d}^{(t)} = \left[d_i^{(t)}\right]$ - death vector presenting the number of dead enterprises in each sector; by a dead enterprise we understand an enterprise that existed in year *t-1* and simultaneously did not exist in *t*.

$\mathbf{b}^{(t)} = \left[b_j^{(t)}\right]$ - a birth vector containing the number of new enterprises in each sector; by a new enterprise we understand an enterprise that did not exist in year *t*-1 and existed in *t*.

A zero in matrix **X** indicates that moving from 'birth' state to 'death' in the same year is not possible by this model.

In addition, by **r** and **c** we understand vectors holding the sums of rows and columns of matrix **X**:

$$\mathbf{r}^{(t)} = \left[r_i^{(t)}\right] = \left[\sum_j x_{ij}^{(t)}\right]$$

$$\mathbf{c}^{(t)} = \left[c_j^{(t)}\right] = \left[\sum_i x_{ij}^{(t)}\right]$$

The main properties of matrix $\mathbf{X}^{(t)}$ are:

- the sum of each row ($r_i^{(t)}$) describes the number of enterprises in corresponding sectors in year t-1,
- the sum of each column ($c_j^{(t)}$) describes the number of enterprises in sectors in year t,
- the diagonal values of matrix $\mathbf{X}^{(t)}$ represent the number of enterprises that survived year t and remained in the same sector,
- for each sector the sum of a column in year t becomes the sum of a row in year t+1 ($c_i^{(t-1)} = r_i^{(t)}$).

An example of the matrices defined above is presented in table 1. Table 1 holds data that describes migration, births and deaths in Belgian enterprises between the years 2000 and 2001. Belgian companies are divided into sectors such as agriculture, industry, construction trade and transport, financial activities, other services and activities. Those enterprises that do not belong to these sectors are gathered in the group named unknown.

Table 1. Matrix $\mathbf{X}^{(2001)}$ for Belgian companies in the year 2001

	Agriculture	Industry	Construction	Trade and Transport	Financial activities	Other services and activities	Unknown	Death	Total
Agriculture	26165	6	13	42	13	12	104	1527	27882
Industry	2	42731	48	148	50	29	183	2713	45904
Construction	18	60	65718	55	53	7	310	5035	71256
Trade and Transport	171	151	77	229481	280	104	1434	20839	252537
Financial activities	23	63	55	321	137713	100	536	10479	149290
Other services and activities	4	6	4	59	96	70519	226	4107	75021
Unknown	46	70	166	497	2310	314	12628	5299	21330
Birth	1689	2317	5288	16936	15298	6062	2469	0	50059
Total	28118	45404	71369	247539	155813	77147	17890	49999	

Source: Crossroads Bank for Enterprises.

In the first row one can see that 27882 enterprises were active in "Agriculture" branch in 2000 (Total - sum of the first row). Between the year 2000 and 2001: 26165 of them remained in the "Agriculture" branch in 2001, 6 of them migrated to the "Industry" branch, 2 enterprises migrated to "Agriculture" from "Industry", 1527 enterprises from "Agriculture" died (Death – first row), which means they were no longer economically active in 2001, 1689 new enterprises in "Agriculture" were created (Birth – the first column). Finally, in 2001 the "Agriculture" branch had 28118 active enterprises.

In our next step we will define a matrix $\mathbf{P}^{(t)}$ that holds the probability of company migration between sectors, "birth" and "death" states in year *t*. For each year a matrix will be defined as a fraction of enterprises migrating between possible states. Every element in matrix **X** would be divided by the sum of the corresponding row. In this way we obtain a probability matrix which is analogous to input matrix in the context of IO analysis:

$$\mathbf{P}^{(t)} = \left[p_{ij}^{(t)}\right] = \left[\frac{x_{ij}^{(t)}}{\sum_i x_{ij}^{(t)}}\right]$$

Matrix $\mathbf{P}^{(2001)}$ (in %) calculated from matrix $\mathbf{X}^{(2001)}$ is presented in table 2.

Table 2. Matrix $\mathbf{P}^{(2001)}$ derived from matrix $\mathbf{X}^{(2001)}$

	Agriculture	Industry	Construction	Trade and Transport	Financial activities	Other services and activities	Unknown	Death	Total
Agriculture	93.8419	0.0215	0.0466	0.1506	0.0466	0.0430	0.3730	5.4767	27882
Industry	0.0044	93.0877	0.1046	0.3224	0.1089	0.0632	0.3987	5.9102	45904
Construction	0.0253	0.0842	92.2280	0.0772	0.0744	0.0098	0.4351	7.0661	71256
Trade and Transport	0.0677	0.0598	0.0305	90.8702	0.1109	0.0412	0.5678	8.2519	252537
Financial activities	0.0154	0.0422	0.0368	0.2150	92.2453	0.0670	0.3590	7.0192	149290
Other services and activities	0.0053	0.0080	0.0053	0.0786	0.1280	93.9990	0.3012	5.4745	75021
Unknown	0.2157	0.3282	0.7782	2.3301	10.8298	1.4721	59.2030	24.8429	21330
Birth	3.3740	4.6285	10.5635	33.8321	30.5599	12.1097	4.9322	0	50059
Total	28118	45404	71369	247539	155813	77147	17890	49999	

Source: Crossroads Bank for Enterprises, own calculations.

Coppens and Verduyn (2009) assumed that matrix $\mathbf{P}^{(t)}$ is constant over time. Based on this assumption they used a matrix $\overline{\mathbf{P}} = \frac{1}{6}\sum_{i=2001}^{2006}\mathbf{P}^{(i)}$ for all their calculations. But in fact this assumption is a simplification. We will show methods for the prediction of further values of matrix **X** based on input-output matrix updating procedures. In our studies we use the same data as Coppens and Verduyn (2009). Our research is based on data from the Crossroads Bank for Enterprises published by Belgostat about Belgian enterprise migration in industry sectors in the period between 2000-2006 (see: http://www.belgostat.be - select option: "Business demography").

4. Forecasts of migration matrices

The problem of forecasting migration matrices can be conducted by means of updating procedures well-known in I-O theory. In this paper, we use some of the most popular I-O updating algorithms (RAS and some non-biproportional approaches) and measure the ex-post error of prediction. Non-biproportional algorithms generally define a measure of distance between the elements of two matrices and minimize it. Potentially a measure of the distance between two matrices can be identified in many ways, but those based on absolute differences and/or squared differences are most popular among researchers (Jackson and Murray (2004)).

For simplicity of notation, let $p_{ij} \in \mathbf{P}$ be the last known probability coefficient matrix, and let $q_{ij} \in \mathbf{Q}$ be the prediction of the probability matrix for the next year. Similarly as in the previous section, **c** and **r** represent known vectors containing the sum of columns and the sum of rows in matrix **Q**.

1. *Constant Value Matrix*

The method is based on research done by Coppens and Verduyn (2009).

$$\mathbf{P} = \overline{\mathbf{P}} = \frac{1}{6}\sum_{i=2001}^{2006}\mathbf{P}^{(i)}$$

2. *Absolute Differences*

The main goal of this method is to minimize the sum of absolute differences between the elements of the last known matrix and the prediction of the matrix for the following year. Unfortunately, some zeros may appear in the algorithmic solution.

$$\sum_i \sum_j \left| p_{ij} - q_{ij} \right| \rightarrow \min$$

$$s.t. \sum_j q_{ij} = 1 \text{ for all } j$$

$$\sum_i q_{ij} r_i = c_j \text{ for all } i$$

$$q_{ij} > 0 \text{ for all } i, j$$

3. *Weighted Absolute Differences*

This method assigns weight to absolute differences by the size of coefficients. A large (small) coefficient implies a large (small) assigned weight. In the case of our data, the diagonal values of matrix **P** are more "important" than the rest of the matrix. The percentage of enterprises remaining in the same sector is more stable than migrating portion.

$$\sum_i \sum_j p_{ij} \left| p_{ij} - q_{ij} \right| \to \min$$

$$s.t. \sum_j q_{ij} = 1 \text{ for all } j$$

$$\sum_i q_{ij} r_i = c_j \text{ for all } i$$

$$q_{ij} > 0 \text{ for all } i, j$$

4. *Normalized Absolute Differences (Matuszewski 1964)*

The fourth method also assigns weights to absolute differences, but in contrast to method 3 small (large) coefficients are assigned to large (small) weights. In this approach the diagonal elements of matrix **P** which are relatively high in level are significantly less "important" than the other coefficients of this matrix. The percentage of migrating enterprises is more stable than the ratio of enterprises remaining in the same sector.

$$\sum_i \sum_j \frac{\left| p_{ij} - q_{ij} \right|}{p_{ij}} \to \min$$

$$s.t. \sum_j q_{ij} = 1 \text{ for all } j$$

$$\sum_i q_{ij} r_i = c_j \text{ for all } i$$

$$q_{ij} > 0 \text{ for all } i, j$$

5. *Squared Differences (Almon 1968)*

The fifth method was formulated by Almon (1968). It is worth noticing that in this method we minimize the Euclidean distance between matrices **P** and **Q**. This fifth method in fact minimizes the sum of squares, and because of that has the properties of ordinary least squares (Durieux and Payen, 1976).

$$\sum_i \sum_j \left(p_{ij} - q_{ij}\right)^2 \rightarrow \min$$

$$s.t. \sum_j q_{ij} = 1 \text{ for all } j$$

$$\sum_i q_{ij} r_i = c_j \text{ for all } i$$

$$q_{ij} > 0 \text{ for all } i, j$$

6. *Weighted Squared Differences*

The sixth method assigns weights to squared differences. The weights are analogous to those applied in the third method (small coefficient implies small weight).

$$\sum_i \sum_j p_{ij} \left(p_{ij} - q_{ij}\right)^2 \rightarrow \min$$

$$s.t. \sum_j q_{ij} = 1 \text{ for all } j$$

$$\sum_i q_{ij} r_i = c_j \text{ for all } i$$

$$q_{ij} > 0 \text{ for all } i, j$$

7. *Normalized Squared Differences (Friedlander 1961)*

The seventh method was formulated by Friedlander. This method is a normalized version of Almon's method and has weights similar to those of the fourth method.

$$\sum_i \sum_j \frac{\left(p_{ij} - q_{ij}\right)^2}{p_{ij}} \rightarrow \min$$

$$s.t. \sum_j q_{ij} = 1 \text{ for all } j$$

$$\sum_i q_{ij} r_i = c_j \text{ for all } i$$

$$q_{ij} > 0 \text{ for all } i, j$$

The main difficulty connected to last three methods is resolving the non-linear optimization problem. Solutions can often be local rather than global. In our simulations we used the Simplex method and a Generalized Reduced Gradient (GRG2) algorithm. As a starting point actual numbers from target matrices were taken.

8. *RAS method*

Last but not least is the biproportional method RAS (Bacharach,1970). The iterative proportional fitting procedure IPFP, known as biproportional fitting in statistics, the RAS algorithm in economics and matrix scaling in computer science, is an iterative algorithm for estimating cell values of a contingency table such that the marginal totals remain fixed and the estimated table decomposes into an outer product.

The RAS method was developed by Stone (1961). It is based on a biproportional technique which is an iterative procedure that allows the computation (on condition that the marginal totals remain fixed) of non-negative matrix elements. At the beginning the estimated matrix is equal to the prior matrix. In further stages the rows and columns of the estimated matrix are scaled alternately in order to achieve the desired properties. The scaling is conducted by the multiplication of each element in a row (column) by the proportion of the desired and actual sum of that row (column). In the case of our data about 200 iterations were performed for single migration matrix.

5. Empirical results

In order to decide which method gives better predictions we used four matrix comparison methods. We used Theil's U statistic (Theil 1971), the weighted absolute difference (WAD, Lahr, 2004), index C (Roy et al., 1982) and finally standardized total percentage error (STPE) for multiplier assessment (Miller and Blair, 1985). In fact these comparison methods are measures of the ex-post error of prediction. In order to present the formula of the ex-post statistics, we used elements $a_{ij} \in \mathbf{A}$, where $\mathbf{A}$ denotes the actual matrix from a certain year.

Numbers $q_{ij} \in \mathbf{Q}$ are the forecasts of the elements of an actual matrix **A**. We applied the following measures of ex-post errors:

$$U = \sqrt{\frac{\sum\sum\left(a_{ij} - q_{ij}\right)^2}{\sum\sum a_{ij}^2}}$$

$$WAD = \frac{\sum\sum a_{ij}\left|a_{ij} - q_{ij}\right|}{\sum\sum\left(a_{ij} + q_{ij}\right)}$$

$$C = \frac{\left|\sum\sum q_{ij}\log q_{ij} - \sum\sum a_{ij}\log a_{ij}\right|}{\sum\sum a_{ij}\log a_{ij}}$$

$$STPE = 100\frac{\sum\sum\left|a_{ij} - q_{ij}\right|}{\sum\sum a_{ij}}$$

We calculated these statistics for all the methods used. Migration matrices from years 2000-2006 were used. In table 3 we present the best methods of predicting the following year's migration according to ex-post error statistics for the matrices.

Table 3. Best methods for predicting the following year's migration according to ex-post error statistics

	2001/2002	2002/2003	2003/2004	2004/2005	2005/2006
STPE	M8	M2	M8	M1	M8
Theil's U	M4	M2	M8	M1	M8
WAD	M8	M2	M7	M1	M8
C	M8	M2	M4	M1	M1

Source: Own calculations.

For each method of comparison the mean value of statistics based on the matrices were calculated. In order to complete the ranking of the methods applied, we assessed them according to all four ex-post error statistics. The results are presented in table 4.

Table 4. Results of matrix updating. Average index values for years 2000-2006

	STPE	rank	Theil's U	rank	WAD	rank	C	rank	Average Rank	Combined Rank
M1	1.7876	4	0.0427	2	0.0052	3	0.0227	1	2.5	2
M2	1.7822	3	0.041	1	0.0049	1	0.0271	2	1.75	1
M3	2.1837	7	0.0431	3	0.0052	2	0.0289	3	3.75	3
M4	1.8037	5	0.0544	8	0.0063	8	0.0353	7	7	8
M5	2.0286	6	0.0542	7	0.006	6	0.0358	8	6.75	7
M6	2.2417	8	0.054	6	0.006	7	0.0344	5	6.5	6
M7	1.7684	2	0.0531	5	0.0058	5	0.0335	4	4	5
M8	1.6440	1	0.0518	4	0.0056	4	0.0347	6	3.75	3

Source: Own calculations.

Based on these results, we see that all predictions have comparable statistical values. That is why on the basis of these rankings we recommend the use of minimizing absolute differences and the RAS method.

6. Conclusions

As a first conclusion we must draw attention to the size of prediction errors made by using these methods. The standard prediction error (STPE) for all the years and methods is between 0.39% and 3.8%, which means that effectiveness of this methodology is very high.

According to the ranking based on the mean error statistics the second method is the best one for predicting migration matrices. The second method is a non-biproportional algorithm minimizing absolute differences. The unexpectedly good ex-post forecasts were achieved by the average matrix (i.e. mean value of matrices from the period under study). One of the reasons for that may be that the target matrix is also included in the average matrix calculation (as one of six). It is worth noting that although the RAS method was ranked as third it was most often the best method, and exhibited the lowest values of STPE statistics. The latter observation means that the probability of wrong classification of an enterprise by this method is lower in comparison to the others.

The comparison of methods that favour large diagonal values of a matrix and methods which attach large weights to relatively small matrix elements is also interesting. Our investigation shows that generally the second group of algorithms gives better results. This means that the migration process which applies to a small percentage of enterprises is stable in time. Irregularity connected to changes in the numbers of enterprises involves mostly companies that stick to a certain branch of industry.

REFERENCES

AGARWAL V., TAFFLER R. (2011). Comparing the performance of market-based and accounting based bankruptcy prediction models, Journal of Banking and Finance, 32, pp. 1541-1551.

ALMON C. (1968). Recent methodological advances in input-output in the United States and Canada, Fourth International Conference on Input-Output Techniques, Geneva.

ALTMAN E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, Journal of Finance, 23, pp. 589-609.

ALTMAN E., AVERY R., EISENBEIS R., SINKEY J. (1981). Application of classification techniques in business, banking, and finance, Greenwich, Conn, JAI Press.

BAETGE J., STROEHER T. (2005). Empirische Insolvenzforschung zur Beurteilung der Bestandsfestigkeit von Unternehmen, in: Burmann, Chr. (Eds.), Management von Ad-hoc-Krisen: Grundlagen, Strategien und Erfolgsfaktoren, Gabler, Wiesbaden, pp. 151-167.

BEAVER W.H. (1966). Financial ratios as predictors of failure, in: Empirical research in accounting: Selected studies, Supplement for Journal of Accounting Research, 4, pp. 71-111.

BACHARACH M. (1970). Biproportional Matrices and Input-Output Change, Cambridge, U.K., Cambridge University Press.

BLUM M. (1974). Failing Company Discriminant Analysis, Journal of Accounting Research, Spring, pp. 1 -25.

COPPENS F., VERDUYN F. (2009). Analysis of business demography using Markov chains: an application to Belgian data, Working Paper Research 170, National Bank of Belgium.

DAVYDENKO S., FRANKS J. (2008). Do bankruptcy codes matter? A study of defaults in France, Germany, and the U.K., Journal of Finance, 63, pp. 565-608.

DEAKIN E.B. (1972). A Discriminant Analysis of Predictors of Business Failure, Journal of Accounting Research, Spring, pp. 167- 179.

DIACOGIANNIS G.P. (1996). The usefulness of share prices and inflation for corporate failure prediction, University of Piraeus Journal of Economics Business, Statistics and Operations Research, 46, pp. 135-156.

DIAMOND U.S. (1976). Pattern Recognition and the Detection of Corporate Failure, Ph.D. dissertation, New York University.

EISENBEIS R.A. (1977). Pitfalls in the application of discriminant analysis in business, finance, and economics, Journal of Finance, 32, pp. 875-900.

FREDERICK W.C., DAVIS K., POST J.E. (1988). Corporate Social Responsibility and Business Ethics, McGraw-Hill Publishing Company, New York, pp. 28-29.

FOSTER R., KAPLAN S. (2001). Creative Destruction, Financial Times.

GURGUL H., ZAJAC P. (2011). The dynamic model of birth and death of enterprises, Statistics in Transition, 12/ 2, pp. 381–400.

GREINER L.E., SCHEIN V.E. (1988). Power and Organization Development, Addison-Wesley.

HANDY CH. (1995). The Age of Paradox, Harvard Business School Press.

HESSELMANN S. (1995). Insolvenzprognose mit Hilfequalitativer Faktoren, Shaker Verlag, Aachen.

JACKSON R.W., MURRAY A.T. (2004). Alternative input–output matrix updating formulations, Economic Systems Research, 16, pp. 135–14.

JACOBSON T., LINDE J. (2000). Credit rating and the business cycle: can bankruptcies be forecast?, Sveriges Riksbank economic review, 4, pp. 11-33.

JONES S., HENSHER D.A. (2008). Advances in credit risk modeling and corporate bankruptcy Prediction, University Press, Cambridge, UK.

KAHL M. (2002). Economic distress, financial distress and dynamic liquidation, Journal of Finance, 57, pp. 135-168.

KEASEY K., WATSON R. (1987). Non-financial symptoms and the prediction of small company failure: a test or Argenti's hypotheses, Journal of Business Finance and Accounting, 14 (3), 335-354.

LAHR M., de MESNARD L. (2004). Biproportional Techniques in Input-Output Analysis: Table Updating and Structural Analysis, Economic Systems Research, 16 (2), 115-34.

MATSCHKE M.J. (1979). Insolvenzprognose aus vergangenheitsorientierten Jahresabschlüssen als Basis von Kreditentscheidungen, Betriebswirtschaftlich eForschung und Praxis, 31, pp. 485-504.

MILLER R.E., BLAIR P.D. (1985). Input-output analysis: Foundations and extensions, Prentice-Hall, Englewood Cliffs, N.J.

MOYER R.C. (1977). Forecasting financial failure: a re-examination, Financial Management, 6, pp.11-17.

MUCHE T. (2007). Ein stochastisches Modell zur Insolvenzprognose auf der Basis von Jahresabschlußdaten, Betriebswirtschaftliche Forschung und Praxis, 59, pp. 376-399.

OHLSON J.A. (1980). Financial Ratios and Probabilistic Prediction of Bankruptcy, Journal of Accounting Research, Spring, pp. 109-131.

PEEL M.J., PEEL D.A. (1988). A Multilogit Approach to Predicting Corporate Failure: Some Evidence for the UK Corporate Sector, Omega, pp. 309-318.

PLATT H., PLATT M. (1991). A note on the use of industry-relative ratios in bankruptcy prediction, Journal of Banking and Finance, 15, pp. 1183-1194.

ROY J., BATTEN D., LESSE P. (1982). Minimizing information loss in simple aggregation, Environment and Planning, 14, 973-980.

SHUMPETER J. (1982). The Theory of Economic Development: An inquiry into profits, capital, credit, interest and the business cycle, Transaction Publisher.

STONE R.A. (1961). Input-Output Accounts and National Accounts. Paris, Organization for European Economic Cooperation.

THEIL H. (1971). Applied Economic Forecasting. Amsterdam, North-Holland.

THIELE O., LOHMANN K. (1995). Möglichkeiten und Grenzen der Insolvenzprognose auf der Grundlage von Jahresabschluss informationen, Freiberger Arbeitspapiere, 95/17.

WILCOX J.W. (1971), A Simple Theory of Financial Ratios as Predictors of Failure, Journal of Accounting Research, Autumn, pp. 389-395.

ZAVGREN C.V. (1985). Assessing the Vulnerability to the Failure of American Industrial Firms: A Logistic Analysis, Journal of Business Finance and Accounting, pp. 19-45.

ZMIJEWSKI M. (1984). Methodological issues related to the estimation of financial distress prediction models, Journal of Accounting Research (Supplement), pp. 59-82.

STATISTICS IN TRANSITION-new series, Summer 2012
Vol. 13, No. 2, pp. 405—418

TIME SERIES MODEL FOR PREDICTING THE MEAN DEATH RATE OF A DISEASE

D. K. SHANGODOYIN[1], **J. F. OJO**[2], **J. O. OLAOMI**[1], **A. O. ADEBILE**[3]

ABSTRACT

This study develops a time series model to estimate the mean death rate of either an emerging disease or re-emerging disease with a bilinear induced model. The estimated death rate converges rapidly to the true parameter value for a given mean death at time t. The derived model could be used in predicting the m-step future death rate value of a given disease. We illustrated the new concept with real life data.

Keywords: Mean death, bilinear model, Death cases, emerging and re-emerging diseases.

1. Introduction

A measure of death rate gives an indicator to quality of health services available within a locality as well as the standard of preparedness of the health provider to curb the menace of the disease; the most important to health care provider is on how best to estimate the death rate of a given contagious disease. Globally, World Health Organization (WHO) is saddled with the responsibility of how best to estimate the death rate arising from a global pandemic. The World Health Organization (WHO) convectional formula for computation of death rate is simply the ratio of the number of known deaths to the total number of confirmed cases (Mathers and Loncar (2006), WHO (2005) Chronic disease report), however, this formula is likely to underestimate the true death rate because the outcomes of many cases might be unknown or uncertain at the time these figures are compiled. Some other methods for estimating death rate have varying degrees of challenges as mentioned in Shangodoyin (2010), in particular, the generalized mixed effect model of estimating death rate discussed by Tong

[1] Department of Statistics, University of Botswana, Botswana. Corresponding author: shangodoyink@mopipi.ub.bw.
[2] Department of Statistics, University of Ibadan, Ibadan, Nigeria.
[3] Department of Statistics, Federal Polytechnic, Ede, Nigeria.

and Chang (2006) although converges quickly to the death rate computed from the complete data, the linear model specified leads to a singular precision matrix for the unknown parameters, also the choice of the singular value decomposition presented may restrain this approach for practical use and in their paper, it was concluded that further research on how to carry out the estimation of the conditional mean death rate model with constraint in which the estimated death rate should be greater than or equal to zero.

Linear models are widely used because of their unrivalled simplicity, but they cannot be applied for data that have a turning-or rate-change-point, even if the data show good linearity sufficiently far from this point. To describe such bilinear-type data so as to make possible smooth and fully parameterizable transitions between two linear segments while still maintaining a clear connection with the linear models we require a bilinear type of model. Buchwald, P (2007) utilized a completely generalized version of a linearized biexponential model (LinBiExp) to various time profiles of biological and medical interest including growth profiles, such as those of human stature, agricultural crops and fruits, multicellular tumor spheroids, single fission yeast cells, or even labour productivity, and decline profiles, such as age-effects on cognition in patients who develop dementia and lactation yields in dairy cattle.

In this study, an attempt is made to develop a bilinear time series model to estimate the overall death rate of an emerging or re-emerging disease with the inclusion of bilinear parameters on the conditional mean death model.

2. Basic theory of bilnear models

There are situations when it is felt that linear time series models may not be adequate in explaining the underlying random mechanisms as in for instance sunspot data, Canadian lynx data which cannot be adequately described by linear time series models and the test proposed by Rao and Gabr (1981) does confirm that the above series cannot be described by linear Gaussian models. However, one may ask if there exist other models, which can provide a better fit. One is led then to consider non-linear models. Jones (1978), Granger and Andersen (1978), Haggan and Ozaki (1980), Priestly (1980), Tong and Lim (1980) and Rao(1981) have considered particular types of nonlinear time series models. The nonlinear models considered by Granger and Andersen (1978) and Rao (1981) are known as bilinear time series models. This class of time series has been found to provide a better fit as well useful in many areas for example biological sciences, ecology and engineering (see Mohler (1973), Bruni, Dupillo and Koch (1974)). From what we have above, one could say that the time series models for which it may be possible to obtain optimal forecasts for several steps ahead and which can perform better than a linear model are the bilinear time series models.

The most general form of the bilinear model, as defined in Granger and Andersen (1978a) which is denoted by BL (p, q, r, s) is given as:

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + e_t - \sum_{j=1}^{q} \theta_j e_{t-j} + \sum_{i=1}^{r} \sum_{j=1}^{s} \beta_{ij} X_{t-i} e_{t-j}$$

where the noise e_t consists of independent, identically distributed random variables with zero means and variance σ^2. We assume also that the model is invertible. Also e_t is assumed to be independent of X_s. The generality of the above model makes it too complicated to be analyzed successfully (Tuan and Lanh,1981); as a result these authors considered the first order bilinear model BL (1, 0, 1, 1) given as $X_t = \phi_1 X_{t-1} + e_t + b_{11} X_{t-1} e_{t-1}$ also, Rao (1980) considered the bilinear model BL (p, 0, r, s) given as

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{i=1}^{r} \sum_{j=1}^{s} \beta_{ij} X_{t-i} e_{t-j} + e_t$$

2.1. The vector form

It is well known that the linear autoregressive moving average models can be written in the form of a first-order vector difference equation. Anderson, 1971; Priestley, 1978, 1980 and this vector form is known as the state space form. It is convenient to study the properties of the process when the model is in the state space form because of the Markovian nature of the model (Akaike, 1974).

Consider the model BL (p, d, 0, r, 1) given as

$$X_t = \psi_1 X_{t-1} + \ldots\ldots\ldots + \psi_{p+d} X_{t-p-d} + \left(\sum_{k=1}^{r} b_{k1} X_{t-k} \right) e_{t-1} + e_t \qquad (3.2.1)$$

Let us define the matrices

$$\Psi = \begin{bmatrix} -\psi_1 & -\psi_2 & \psi_3 & & -\psi_{p+d-1} & \psi_{p+d} \\ 1 & 0 & 0 & \ldots\ldots & 0 & 0 \\ 0 & 1 & 0 & \ldots\ldots & 0 & 0 \\ 0 & 0 & 0 & \ldots\ldots & 1 & 0 \end{bmatrix},$$

$$B = \begin{pmatrix} b_{11} & b_{21} & b_{31} & \ldots\ldots & b_{r1} \\ 0 & 0 & 0 & \ldots\ldots\ldots & 0 \\ 0 & 0 & 0 & \ldots\ldots\ldots & 0 \end{pmatrix}$$

and vectors C^T=(1, 0, 0,…..0), H^T=(1, 0, 0,…..0), and let $\mathbf{X}_t^T = (X_t, X_{t-1}, ..., X_{t-p+d})$,
t = …..-1, 0, 1,…. With this notation, we can write the model in sub section 2.1 in the vector form as:

$$\mathbf{X}_t = \Psi\mathbf{X}_{t-1} + \mathbf{B}\mathbf{X}_{t-1}e_{t-1} + \mathbf{C}e_t \qquad \textbf{and} \qquad \mathbf{X}_t = \mathbf{H}^T\mathbf{X}_t$$

It is pertinent to ask ourselves that under what conditions on the given matrices C, ψ, B and on the given sequence $\{e_t$, t = …-1, 0, 1, …} of independent identically distributed random variables with common mean zero and common variance $\sigma^2 < \infty$, does there exist a vector-valued strictly stationary process {X_t, t= …-1, 0, 1, …} satisfying $\mathbf{X}_t = \Psi\mathbf{X}_{t-1} + \mathbf{B}\mathbf{X}_{t-1}e_{t-1} + \mathbf{C}e_t$ for every t = …-1, 0, 1,… The answer to the above is considered in sub section 2.2

2.2. Stationarity and Convergence of BL (p, d, 0, r, 1)

In this section, we give a sufficient condition for the existence of strictly stationary process and convergence conforming to the bilinear model defined in section 2.1, in the case when s =1. This we do through the following theorem.

Theorem: Let $\{e_t, t \in Z\}$ be a sequence of independent identically distributed random variables defined on a probability space (Ω, IR, P) such that E e_t = 0 and $Ee_t^2 = \sigma^2 < \infty$. Let Ψ and B be two matrices of order p x p such that $\rho(\Psi \otimes \Psi + \sigma^2 B \otimes B) = \lambda < 1$. Let C be any column vector with components $c_1, c_2, \ldots, c_p$. Then the series of random vectors

$$\sum_{r\geq 1}\prod_{i=1}^{r}(\Psi + Be_{t-j})Ce_{t-r}$$

converges absolutely almost surely as well as in the mean for every fixed t in Z. Further, if

$$X_t = Ce_t + \sum_{r\geq 1}\prod_{i=1}^{r}(\Psi + Be_{t-j})Ce_{t-r}, \quad t \in Z,$$

then $\{X_t, t \in Z\}$ is a strictly stationary process conforming to the bilinear model $\mathbf{X}_t = \Psi\mathbf{X}_{t-1} + \mathbf{B}\mathbf{X}_{t-1}e_{t-1} + \mathbf{C}e_t$, for every t in Z. Conversely, if $\{X_t, t \in Z\}$ is a strictly stationary process satisfying $X_t = Ce_t + \Psi X_{t-1} + BX_{t-1}e_{t-1}$, for every t in Z for some sequence $\{e_t, t \in Z\}$ of independent identically distributed random variables with $Ee_t = 0$ and $Ee_t^2 = \sigma^2 < \infty$ and some matrices Ψ , B, C of orders

p x p, r x r, p x 1 respectively with $\rho\left(\Psi \otimes \Psi + \sigma^2 B \otimes B\right) = \lambda < 1$, then $X_t = Ce_t + \sum_{r\geq 1} \prod_{i=1}^{r} \left(\Psi + Be_{t-j}\right) Ce_{t-r}$, for every t in Z.

The proof of this theorem was considered in Shangodoyin, Ojo and Marcin (2010).

3. Mean Death rate model specification

Assume that p_j is the probability that a confirmed case ends up in death on the j-th day after the confirmation of the disease; also q_j is the corresponding probability of being recovered and discharged from the hospital on the j-th day. Then $P = \sum_{j=0}^{\infty} p_j$ is the death rate of the disease and $Q = 1 - P = \sum_{j=0}^{\infty} q_j$ is the recovery rate; all cases are assumed to be independent of each other and have identical death rate and probability distribution of time to death or discharge. Let C_t be *the* number of confirmed cases at time t, and D_t be the *number* of deaths at time t. The death on the t-th time must be from the previously known cases at time t-j, then the probability that it is from time t-j equals, p_j, say, and there are C_{t-j} at time t-j (Tong and Chan (2006)). The conditional mean of D_t given C_{t-j}; for j=0,1..... as

$$E\left(D_t / C_{t-j}\right) = \mu_t = \sum_{-\infty}^{\infty} p_j C_{t-j} \tag{1}$$

For a finite general linear process, the expression in (1) is written as

$$\mu_t = \sum_{a_1}^{a_2} p_j C_{t-j} \tag{2}$$

Where $a_1, a_2 \geq 0$ are lower and upper bounds of the time to death.

The expected value of death at time t is a linear function of the number of confirmed cases in time t-j, with assumption that all cases are independent of each other and have identical death and probability distribution of time to death may tempt analyst to adopt linear estimation techniques to determine p_j; this could pose some problems because if $\langle C_t \rangle = 0 \ \forall$ t<0, then theoretically D_t are serially dependent since they come from the earlier confirmed cases. Serial

correlation among the D_t may be accounted for by including a latent process on the right side of (2).

Also for finite uncensored aggregate data on re-emerging and/or emerging disease the analysis of the expected deaths taking into consideration the distribution of its past realizations and coupled with the large number of lags of C_{t-j} and $\mu_{\text{t-j}}$ can easily introduce multicollinearity in the model. This problem may not be adequately captured by linear model presented in equation (2), a problem identified by Tong and Chan (2006); and it could be corrected by incorporating smoothness assumption in the estimation of the unknown parameters using some parsimonious parametric class of models. A class of parsimonious models that have been found to provide a better fit some biological, ecological and medical data is the bilinear time series models (Mohler (1973) and Bruni etal (1974)). Martins (1999) and Gonclaves (2000) have shown that this class of models provide an optimal forecast for several steps ahead and can perform better than a linear model. In order to account for serial correlation in μ_t and to ensure that ε_j and p_j vary smoothly with lag j, we will adopt a modified bilinear model (MBL) (1, 0, 1, 1) to the left hand side of equation (2); for details see Rao (1981), Chen (1992), Shangodoyin, Ojo and Kozak (2010).
Assuming that u is obtained using the order selection criterion AIC with starting lower bound j=1, then the derived model is:

$$\mu_t - \alpha\mu_{t-1} - \beta\mu_{t-1}e_{t-1} = \sum_{u_1}^{u_2} p_j C_{t-j} + e_t \tag{3}$$

We then write the bilinear model for estimating the mean of the deaths at time t as:

$$\mu_t = \alpha\mu_{t-1} + \beta\mu_{t-1}e_{t-1} + \sum_{1}^{u} p_j C_{t-j} + e_t \tag{4}$$

4. Model estimation

Suppose that μ_t are generated by equation (4), the sequence of random deviates $\{e_t\}$ could be determined from the relation

$$e_t = \mu_t - \alpha\mu_{t-1} - \beta\mu_{t-1}e_{t-1} - \sum_{1}^{u} p_j C_{t-j} \tag{5}$$

To estimate the unknown parameters in equation (5), we make the following assumptions:

(i) The errors $\{e_t\}$ are independent and identically distributed with mean zero and variance σ^2 with finite kurtosis.

(ii) The root of the $p(B)=\sum p_j B^j$ polynomial lies outside the unit circle, this ensures stationary of $\{C_{t-j}\}$.

(iii) The values of $|\alpha|<1$ and $|\beta|<1$, ensure that invertibility condition required of the bilinear process is satisfied.

Following Walker (1964), we shall refer to the parameters $\boldsymbol{\lambda}^{'}=[\alpha,\beta,\mathbf{p}]$ as the structural parameters, the complete set of r=1+1+u+1 parameters will be denoted by $\boldsymbol{\Omega}=[\boldsymbol{\lambda},\sigma^2]$. The task here is to get the estimator of $\boldsymbol{\lambda}$ using information on all realizations of the entire history of the system generating μ_t. The effect of transients can be minimized if the difference equation (5) is started off from a value of time t for which all the previous values of both C_t and μ_t are known. Let us specify the starting values of μ_t as $\dot{\mu}_t=\frac{\sum_1^t D_t}{t}$; $\forall$t=1,2.....

Thus, with the errors in equation (5) and appropriate initial values, say, $e_{t,0}$

The least squares estimator of $\boldsymbol{\lambda}^{'}=[\alpha,\beta,\mathbf{p}]$ are obtained by minimizing the function $Q(\lambda)=\sum_1^n e_t^2$ with respect to initial values $\boldsymbol{\lambda}_0$. These estimators are derived by non-linear least squares estimation method discussed as follows:

Let $\boldsymbol{\lambda}_0^{'}=[\alpha,\beta,p_1,p_2,.....p_u]$ for explicitness, we shall write $\lambda_{1,0}=\alpha,\ \lambda_{2,0}=\beta,\ \lambda_{3,0}=p_1,\ \lambda_{4,0}=p_2,.....,\lambda_{r-1,0}=p_{u-1}$ and $\lambda_{r,0}=p_u$ for the r-structural parameters. The first and second order partial derivatives of $Q(\boldsymbol{\lambda})$ with respect to the parameters $\boldsymbol{\lambda}_0^{'}$ are respectively

$$G_i=\frac{\partial\left(Q(\boldsymbol{\lambda})\right)}{\partial\dot{\lambda}_i}=2\sum_1^n e_t\left(\frac{\partial e_t}{\partial\dot{\lambda}_i}\right),\forall i=1,2,.....,r$$

and $$h_{ij}=\frac{\partial^2\left(Q(\boldsymbol{\lambda})\right)}{\partial\dot{\lambda}_i\partial\dot{\lambda}_j}=2\sum_1^n\left(\frac{\partial e_t}{\partial\dot{\lambda}_i}\right)\left(\frac{\partial e_t}{\partial\dot{\lambda}_j}\right)+2\sum_1^n e_t\frac{\partial^2 e_t}{\partial\dot{\lambda}_i\partial\dot{\lambda}_j}$$

The elements in G_i are derived from:

$$\frac{\partial e_t}{\partial \alpha} = \dot{\mu}_{t-1}; \frac{\partial e_t}{\partial \beta} = \dot{\mu}_{t-1} e_{t-1}; \frac{\partial e_t}{\partial \varepsilon_j} = C_{t-j}, \forall j = 1, 2, ..., u$$

To estimate the structural parameters, let $\dfrac{\partial^2 e_t}{\partial \dot{\lambda}_i \partial \dot{\lambda}_j} = 0$

Such that for any given values of C_{t-j} and $\dot{\mu}_t$, we evaluate the recursive equations $\sum_1^n e_t \left(\dfrac{\partial e_t}{\partial \dot{\lambda}_i} \right) = 0$ and $\sum_1^n \left(\dfrac{\partial e_t}{\partial \dot{\lambda}_i} \right) \left(\dfrac{\partial e_t}{\partial \dot{\lambda}_j} \right) = 0$

let $\mathbf{V}'(\boldsymbol{\lambda}) = \left[\dfrac{\partial Q(\boldsymbol{\lambda})}{\partial \dot{\lambda}_1}, \dfrac{\partial Q(\boldsymbol{\lambda})}{\partial \dot{\lambda}_2}, ..., \dfrac{\partial Q(\boldsymbol{\lambda})}{\partial \dot{\lambda}_r} \right]$ an $\mathbf{H}(\boldsymbol{\lambda}) = \left[\dfrac{\partial Q(\boldsymbol{\lambda})}{\partial \dot{\lambda}_i \partial \dot{\lambda}_j} \right]$

be the matrices of the first and second partial derivatives. Expanding $V(\boldsymbol{\lambda})$ near $\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}$ in Taylor series, we obtain:

$$\left[\mathbf{V}(\hat{\boldsymbol{\lambda}}) \right]_{\lambda = \hat{\lambda}} = 0 = \mathbf{V}(\boldsymbol{\lambda}) + \mathbf{H}(\boldsymbol{\lambda})(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda})$$

This gives the estimators of the structural parameters as $\hat{\boldsymbol{\lambda}} = \dot{\boldsymbol{\lambda}} - \mathbf{H}^{-1}(\dot{\boldsymbol{\lambda}}) \mathbf{V}(\dot{\boldsymbol{\lambda}})$

Now, suppose that we restrict to u=1, in this case the lag to death is a unit of the measuring period, say, one (year or month of week or day). The simultaneous equations resulting from substituting all the necessary elements of G_i and $\mathrm{h_{ij}}$ are :

$$\sum_1^n e_t \left(\frac{\partial e_t}{\partial \dot{\alpha}} \right) = \sum_1^n \dot{\mu}_t \dot{\mu}_{t-1} - \dot{\alpha} \sum_1^n \dot{\mu}_{t-1}^2 - \dot{\beta} \sum_1^n \dot{\mu}_{t-1}^2 e_{t-1} - \dot{p}_1 \sum_1^n \dot{\mu}_{t-1} C_{t-1} = 0 \tag{6}$$

$$\sum_1^n e_t \left(\frac{\partial e_t}{\partial \dot{\beta}} \right) = \sum_1^n \dot{\mu}_t \dot{\mu}_{t-1} e_{t-1} - \dot{\alpha} \sum_1^n \dot{\mu}_{t-1}^2 e_{t-1} - \dot{\beta} \sum_1^n \dot{\mu}_{t-1}^2 e^2{}_{t-1} - \dot{\varepsilon}_1 \sum_1^n \dot{\mu}_{t-1} C_{t-1} e_{t-1} = 0 \tag{7}$$

$$\sum_1^n e_t \left(\frac{\partial e_t}{\partial \dot{p}_1} \right) = \sum_1^n \dot{\mu}_t C_{t-1} - \dot{\alpha} \sum_1^n \dot{\mu}_{t-1} C_{t-1} - \dot{\beta} \sum_1^n \dot{\mu}_{t-1} C_{t-1} e_{t-1} - \dot{p}_1 \sum_1^n C_{t-1}^2 = 0 \tag{8}$$

The equations derived from the second partial derivatives are:

$$\sum_1^n \left(\frac{\partial e_t}{\partial \dot{\alpha}} \right) \left(\frac{\partial e_t}{\partial \dot{\beta}} \right) = \sum_1^n \dot{\mu}_{t-1}^2 e_{t-1} = 0 \tag{9}$$

$$\sum_1^n \left(\frac{\partial e_t}{\partial \dot{\alpha}} \right) \left(\frac{\partial e_t}{\partial \dot{p}_1} \right) = \sum_1^n \dot{\mu}_{t-1} C_{t-1} = 0 \tag{10}$$

$$\sum_1^n \left(\frac{\partial e_t}{\partial \dot{\beta}} \right) \left(\frac{\partial e_t}{\partial \dot{p}_1} \right) = \sum_1^n \dot{\mu}_{t-1} C_{t-1} e_{t-1} = 0 \tag{11}$$

Substituting equations (9) and (10) in (6) gives:

$$\dot{\alpha} = \frac{\sum_{1}^{n} \dot{\mu}_t \dot{\mu}_{t-1}}{\sum_{1}^{n} \dot{\mu}_{t-1}^2} \tag{12}$$

Also, using equations (9) and (11) in (7) gives:

$$\dot{\beta} = \frac{\sum_{1}^{n} \dot{\mu}_t \dot{\mu}_{t-1} e_{t-1}}{\sum_{1}^{n} (\dot{\mu}_{t-1} e_{t-1})^2} \tag{13}$$

The estimator for $\dot{p}_1$ is obtained using equations (10) and (11) in (8) gives:

$$\dot{p}_1 = \frac{\sum_{1}^{n} \dot{\mu}_t C_{t-1}}{\sum_{1}^{n} C_{t-1}^2} \tag{14}$$

To determine the mean and variance of the estimators defined in equations (12), (13) and (14) we start by re-writing equation (12) as

$$\hat{\alpha} = \frac{\sum_{1}^{n} \hat{\mu}_t \hat{\mu}_{t-1}}{\sum_{1}^{n} \hat{\mu}_{t-1}^2} = f(\hat{\mu}_o, ..., \hat{\mu}_n)$$

Assuming Taylor expansion is valid, we can approximate

$$\hat{\alpha} = f(\hat{\mu}_o, ..., \hat{\mu}_n) \cong f(\mu_o, ..., \mu_n) + \sum_{i=0}^{n} (\hat{\mu}_i - \mu_i) \frac{\partial f}{\partial \hat{\mu}_i} |_{\hat{\mu}_i = \mu_i} + ... \tag{15}$$

The equation (15) yields

$$E(\hat{\alpha}) \cong f(\mu_o, ..., \mu_n) \tag{16}$$

and

$$V(\hat{\alpha}) \cong \sum_{i=0}^{n} \{(\frac{\partial f}{\partial \hat{\mu}_i})^2 |_{\hat{\mu}_i = \mu_i}\} V(\hat{\mu}_i) + \sum_{i \neq}^{n} \sum_{j=1}^{n} \{(\frac{\partial f}{\partial \hat{\mu}_i} \frac{\partial f}{\partial \hat{\mu}_j}) |_{\hat{\mu}_i = \mu_i, \hat{\mu}_j = \mu_j}\} Cov(\hat{\mu}_i, \hat{\mu}_j) \tag{17}$$

The expression of variance (17) is extremely complex and hence no analytical form of $V(\hat{\alpha})$ is available. However, one can compute the $V(\hat{\alpha})$ numerically by using any of the variance estimation methods such as Jackknife method, Bootstrap method and Random group method.

5. Data Analysis and Discussion

This new methodology shall be considered on two real life data; data was collected on Tuberculosis incidence recorded monthly at Julia Molefe Clinic, Gaborone, Botswana and Malaria incidence data was recorded on annual basis at the University College hospital, Ibadan, Nigeria. We utilized equation (14) to compute the death rate; this formula relies both on the moving death average and the number of cases observed for each respective month or year over time.

In table 1, we presented the computed statistics for the death rate of Tuberculosis data using equation (14) and have 0.07358 about 7.3% with estimator variance 0.0016; using the WHO formula we obtained 0.076 about 7.6% the variance of estimator is 0.0684.

In table 2, we presented some computed statistics based on estimator derived in equation (14) and WHO convectional estimator for death rate using Malaria data; the estimator in equation (14) is 0.004 about 0.4% with estimator variance 0.0000005, whereas the WHO estimator for this data is 0.006 about 0.6% with estimator variance 0.000385. Since the ratio of estimators $\frac{Var(estimator \text{ given in equation 14})}{\text{Var(WHO estimator)}} < 1$ in both real life data considered, we affirmed that the estimator presented in equation (14) is more efficient than the WHO estimator; also, WHO formula is not likely to give the true estimate of death rate because the outcomes of some cases might not be known at the time of compilation besides equation (14) may track the monthly (or periodic) number of deaths relatively well.

To predict the mean death rate for the Tuberculosis data we used equations (12), (13) and (14) to compute the estimates of $\dot{\alpha}$, $\dot{\beta}$ and $\dot{p}_1$ as 0.857, -1.21 and 0.081 respectively, thus the prediction model for this data is $\mu_t = \underset{(0.002)}{0.857}\,\mu_{t-1} - \underset{(0.078)}{1.21}\,\mu_{t-1}e_{t-1} + \underset{(0.0006)}{0.081}C_{t-1}$ assuming a one-step time to death for this disease. Similarly for the Malaria data, the computed estimates of

$\dot{\alpha}$, $\dot{\beta}$ and $\dot{p}_1$ are 1.10, -0.1161 and 0.004 respectively; also the prediction model for this data is $\mu_t = \underset{(0.056)}{1.10}\,\mu_{t-1} - \underset{(0.145)}{0.12}\,\mu_{t-1}e_{t-1} + \underset{(0.000001)}{0.004}\,C_{t-1}$ assuming a one-step time to death for this disease.

6. Conclusion

The assumption that the time to death is at most one (day or month or year depending on the nature of the disease); then the overall death rate can be computed using $\dot{p}_1$ provided $C_{t-j} > 0$, then the quantity in (14) will be less than unity, this is another view to the further research advocated by Chan and Tong (2006) approach. Also we can use equation (4) to predict the death rate of a given disease over a given period. It is interesting to note that the estimator in equation (14) performs better than the WHO conventional estimator in the two real-life cases considered.

Appendix

Table 1. TB DEATH STATISTICS (January, 2010- July, 2010)

TIME (t) in Month	C_t	D_t	μ_t	C_{t-1}	e_t
1	16	1	1	14	-0.134
2	9	1	1	16	-1.173
3	11	0	1	9	-0.606
4	12	2	1	11	-0.768
5	5	2	1	12	-0.849
6	13	0	1	5	-0.282
7	12	0	1	13	-0.930
Total	78	6			

Appendix

Table 2. ANNUAL MALARIA DEATH STATISTICS

TIME (t) in Month	C_t	D_t	μ_t	C_{t-1}	e_t
2000	3664	10	10	1670	-5.15
2001	2483	5	7.5	3664	-18.17
2002	2507	13	9.3	2483	-8.88
2003	2113	9	9.3	2507	-9.38
2004	2004	11	9.6	2113	-9.08
2005	2712	37	14.2	2004	-4.38
2006	3072	25	15.7	2712	-10.77
Total	18555	110			

REFERENCES

AKAIKE, H. (1973). Maximum Likelihood Identification of Gaussian Auto-regresssive Moving Average Models. Biometrika 60, 255-265.

ALTMAN, K.L. (2003). The SARS epidemic. The front line research. New York Times. May 07 Edition.

ANDERSON, T.W. (1971). The Statistical Analysis of Time Series, New York and London Wiley.

BUCHWALD, P. (2007). A general bilinear model to describe growth or decline time profiles Math Biosci. 2007 205(1):108-36.

BRUNI, C., DUPILLO, G. and KOCH, G. (1974). Bilinear Systems: An Appealing Class of Nearly Linear System in Theory and Application. IEEE Trans. Auto Control Ac-19, 334-338.

CHEN, C.W.S. (1992). Bayesian Inferences and forecasting in bilinear time series models. Communications in Statistics – Theory and Methods 21(6), 1725-1743.

DONNELLY, C.A., et. al (2003). Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. The Lancet, 361,pg 1761-1766.

GRANGER, C.W.J. and ANDERSON, A.P. (1978). Introduction to Bilinear Time Series Models. Vandenhoeck and Puprecht.

GONCLAVES, E., JACOB P. and MENDES-LOPES N. (2000). A decision procedure for bilinear time series based on the Asymptotic Separation. Statistics, 333-348.

HAGGAN, V. and OZAKI, T. (1980). Amplitude Dependent Exponential AR Model Fitting for Non-Linear Random Vibrations. Proc. International Time Series Meeting, Nottinghamm, ed. O. D. Anderson. North Holland.

JONES, D.A. (1978). Non-linear Autoregressive Processes. Proc. Royal Soc. London (A) 360, 71-95.

KINDIG, D.A., SEPLAKI, C.L. and LIBBY, D.L. (2002). Death variation in US subpopulations. Bulletin of WHO, 80(1), pg 9-15.

LI, W.K. and HUI, Y.V. (1983). Estimation of random coefficient autoregressive process: An empirical Bayes approach . Journal of Time Series Analysis, Vol. 4, No. 2. pg 89-94.

LOHMANN, M.S. (2005). Data quality control and observations error estimation. www.ucar.edu.

MARTINS, C.M. (1997). A Note on the Third-Order Moment Structure of a bilinear model with Non-Independent Shocks. Potrugaliae Mathematica vol 56, 58-89.

MATHERS, C.D. and LONCAR, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. PLos Medicine. Vol. 3. Issue 11, pp 2011-2030.

MOHLER, R.R. (1973). Bilinear Control Processes. New York: Academic Press.

PRIESTELY, M.B. (1978). Non-Linear Models in Time Series Analysis. The Statistician 27, 159-176.

SHANGODOYIN, D.K., OJO, J.F. and KOZAK, M. (2010). Subsetting and identification of optimal models in one-dimensional bilinear time series modelling. International Journal of management science and Engineering management England, UK, 5(4), pp 252-260.

SHANGODOYIN, D.K. (2010). Using time series models to determine the death rate of a given disease. To appear in International Encyclopaedia of Statistical Science (Springer, USA).

SUBBA RAO, T. (1981). On the theory of Bilinear time Series Models ,Jour. R. Statist. Soc.B. 43, 244-255.

TONG, H. AND CHAN, K. (2006). Estimating the death rate of an emerging disease by Time Series Analysis. Technical report, Department of Statistics & Actuarial Science, University of Iowa, Iowa, USA.

TUAN DINH PHAM and LANH TAT TRAN (1981). On the First Order Bilinear Time Series Model. Jour. Appli. Prob. 18, 617-627.

WALKER, A.M. (1962). Large sample estimation of parameters of autoregressive processes with moving average residuals. Biometrika, 49, 117-131.

STATISTICS IN TRANSITION-new series, Summer 2012
Vol. 13, No. 2, pp. 419—438

RESEARCH ON CONSTRUCTING COMPOSITE INDEX OF OBJECTIVE WELL-BEING FROM CHINA MAINLAND*

Zhanjun Xing[1], **Lei Chu**[2]

ABSTRACT

This paper presents some results of the study employing the Quality of Life (QOL) index that was initiated in China during 1980s. Currently, some progress has been made in the research field. For one thing, many researchers defined QOL as a composite structure which included subjective and objective elements. Considerable summary of well-being indices has been constructed from different analysis units (e.g., cities, province, and nations). For another, statistical methods played an important part in determining the weights of these indices, either objective weighting or subjective weighting. This paper defines the concept of QOL as the quality of people's being, which is used in analyzing the well-being of people' living condition and also reflected their subjective well-being to the living condition. Based on this concept, an analytical system of Chinese people's well-being was elaborated. 115 indices from five aspects reflecting the characteristics of the Chinese people's objective well-being have been initially proposed. Economic well-being, health and basic survival well-being, social well-being, cultural well-being, political well-being and environmental well-being have been taken into consideration. Through the analysis of content validity of experts and correlation analysis, 58 indices were selected. The weights of the subjective and objective indices were determined through analytic hierarchy process (AHP) and principal component method respectively. By combining the weight coefficients obtained from the two approaches, the composite indicator of Chinese people's objective well-being was performed and the objective well-being level of 30 Chinese provinces was evaluated.

Key words: Quality of life, Objective well-bell index, Analytic hierarchy process, China Mainland.

[1] School of Political Science and Public Administration, Shandong University, Jinan City, P.R. China. E-mail: xingzhanjun@163.com.
[2] School of Philosophy and Social Development, Jinan City, P.R. China. E-mail: alexchu0220@hotmail.com.
* Acknowledgement: Authors appreciate support from the Research Institute of Statistical Science of NBSC.

1. Introduction

People's Republic of China, which has a total population of about 1.3 billion, economically frail before 1978, has again become one of the world's major economic powers with the greatest potential. In the last 30 years following the reform and opening-up in 1979 in particular, China's economy developed at an unprecedented rate, and that momentum has been held steady into the 21st century. But as the biggest developing country, China's rapid development faces a series of new problems including environmental pollution, income inequality, food safety, etc.; the effort to safeguard and improve people's quality of life (QOL) has been adversely affected. Now, the public policy aims at improving people's well-being and raise the level of public services included in the Twelfth Five-Year Plan (2011-2015) (China adopts the "five-year-plan" strategy for economic). This is an appropriate time to compile a composite index to monitor the well-being in China mainland.

The study of QOL indices had been initiated since the 1980s last century in China. Currently, some progress has been made in this research field and it can be summarized into three stages.

The Earlier research in China mainland can be traced back to the draft of social statistical indicators prepared by National Bureau of Statistics of China in 1983. In the mid-1980s, Chinese-American sociologist Lin cooperated with Tianjin Academy of Social Sciences and Shanghai Academy of Social Sciences to study the urban residents' quality of life (Lin, et al. 1989). This research defined quality of life based on the satisfaction preferences, obviously by the subjective well-being tradition. For example, through factor analysis and structural model analysis, Lin and his co-workers set forth a series of structures and indicators for urban residents' quality of life. Specifically, this study selected 22 aspects of people's lives, and classified into five categories by factor analysis: the social characteristics of work, the economic characteristics of work, the family relationships, the relationships outside the family and leisure time. In contrast with scholars' definitions of quality of life, the government departments and policy research institutions put emphasis on objective well-being that can reflect social development, for the data availability and effectiveness of policies.

Started from 1990s, Chinese Academy of Social Science, National Bureau of Statistics of China, Ministry of Agriculture of China and other organizations have attempted to define the concept of well-off and use a statistical monitoring indicator system to describe the process of Chinese building a well-off society in an all-around way (National Bureau of Statistics of China, 1992-2010). As an indispensable part of this indicator system, quality of life, especially the objective content, has been concerned by policy makers and society. Meanwhile, in order to narrow the gap with foreign researchers, Chinese researchers remain active in cooperation and exchanges with Western scholars (Zhou. et al., 2004). For this reason, much more advanced technologies and approaches have been used in

QOL researches, particularly in the constructions of objective well-being indicators.

Recently, with the strategic decision to build a harmonious society in China, subjective well-being（as the core indicators of QOL）has been brought into public focus. Meanwhile, a group of young scholars are becoming the main force in this field (Zhou. et al., 2004; Xing et al., 2007). In addition, in order to make the achievement can serve the public policy-making, Chinese scholars pay more attention on public policy-oriented of well-being and quality of life researches. It reflects in the concept of well-being, the selection of indicators and the presentation of the QOL reports.

From what has been discussed above, the goal of this research is to develop a composite indicator of objective well-being from China mainland, which could provide both theoretical and practical implications for public policy.

2. Concept of Well-being

Reviewed from literature on QOL or well-being, there is an absolute consensus that a comprehensive definition of QOL must include two linked dimensions, namely an objective one and a subjective one, and an increasingly number of comprehensive concepts of well-being which reflect both individual well-being and social well-being have been put forward. Examples can easily be found to show this trend.

Erik Allardt (Allardt, 1973) started from distinguishing between three basic needs of human beings, namely Having, Loving, and Being, to propose his concept of well-being, which includes objective conditions and subjective need satisfactions. Obviously, there is a great deal of difference between such a broader concept of quality of life and the traditional Scandinavian level of living approach (Erikson, 1993; Uusitalo, 1994). In addition, based on the constellation of objective living conditions and subjective well-being across different life domains, German notion of QOL also combines with objective and subjective dimensions together (Zapf, 1984). In this concept, well-being means good living conditions and positive subjective well-being. Another comprehensive concept of well-being was explored by the UK policy makers and stakeholders who identify the well-being as an implicit goal of Government policy. Such a common understanding of well-being was developed on the UK 2005 sustainable development strategy, *Securing the future*, written by Defra (Department for Environment, Food and Rural Affairs) and other government department of UK (Defra, 2005):

"Well-being is a positive physical, social and mental state; it is not just the absence of pain, discomfort and incapacity. It requires that basic needs are met, that individuals have a sense of purpose, that they feel able to achieve important personal goals and participate in society."

Moreover, this statement considers that well-being is enhanced by some essential conditions, "Supportive personal relationships, strong and inclusive communities, good health, financial and personal security, rewarding employment, and a healthy and attractive environment." Among others things, such understanding of well-being links with public policy by emphasizing the role of government about people's well-being, which is to enable people have a fair access to achieve the social, economic and environment resources now and in the future.

Therefore, taking into account this trend on concept of well-being that has been discussed above, this paper defines the concept QOL or Well-being as the quality of people's being, which is used in analyzing the well-being of people's living condition and reflecting their subjective well-being to the living condition (**Fig. 1**). In fact, this definition can be explained from four aspects as follows:

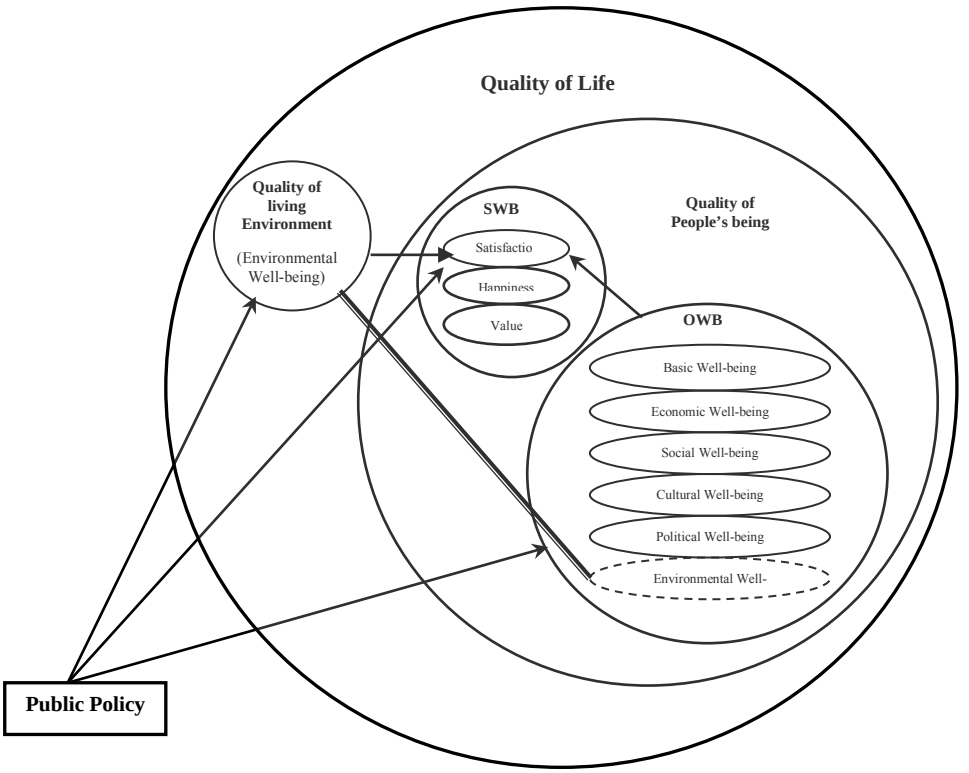
1. Combining the quality of people's being as well as the quality of the natural environment (Environmental well-being); Quality of life could be regarded as a complex system which adheres to the concept of sustainable development.

2. The quality of people's being means the degree of people's need are being met, it is reflected in citizens' satisfaction with objective living conditions and the level of subjective of well-being.

3. Subjective well-being here is defined as people's own subjective experience to their survival and development conditions. It mainly consists of satisfaction, happiness and value components. Specifically, satisfaction refers to the residents' cognition and evaluation of different domains; Happiness is the positive emotion that people can experience; Sense of self-worth means a person's target location and spiritual support of life.

4. Objective well-being state refers to the conditions that citizens achieve from social system, including health and survival well-being, economic well-being, political well-being, social well-being and cultural well-being. Furthermore, as a component of objective conditions, quality of natural environment also could be included, and could be operationalized as environmental well-being.

Figure 1. Theoretical framework of quality of life



3. Methods and Results

Regarding different characters of objective indicators and subjective indicators, we adopt distinct approaches of obtaining weight for each type of them. And in this research we put focus on obtaining weights of objective well-being and the final composite of objective well-being index.

3.1. Design of Life domains and sub-classifications

According to the concept of QOL above, this research defines six life domains of well-being, namely health and basic survival well-being, economic well-being, environmental well-being, cultural well-being, social well-being, and political well-being. Definitions of these six life domains are presented as follows:

Health and basic survival well-being, a major domain of quality of life index, is defined as the conditions provided by society to meet citizens' need of health

and basic survival as well as citizens' satisfaction of those conditions. To meet the needs of health and basic necessities of life is the most necessary conditions to ensure that people can engage in normal economic, political, social, and cultural activities. According to this definition, we will make evaluation both from health and basic survival.

Economic well-being is defined as the situation of residents' material wealth and its fairness, which are reflected by the sub-classifications of income level, consumption level, as well as labor and employment. While emphasizing residents' real income level, economic well-being pays more attention to fairness and rationality of the results of economic development. Specially, income level focuses on measuring of residents' income and its fairness, while consumption level focuses on measuring of residents' consumption. In addition, Labor and employment is regarded as the measurement of residents' employment status and the rationality of overall economic structure.

Environmental well-being is that some appropriate conditions provided by natural environment to meet the survival, reproduction and development of human being during a period of time. As a result of the protection and the governance of the natural environment by people in order to maintain these conditions favorable, people can maintain positive interaction between human activities and the environment. According to the definition of environment well-being which is mainly based on the theory of sustainable development, we set up three sub-classifications: resources and environmental quality, environmental pollution and governance, and the subjective evaluation. These three parts represent the evaluation of the quality of existing environment and resources, the situation of environmental pollution and governance, and citizens' satisfaction of the natural environment.

Cultural well-being could be defined as the security condition of a variety of spiritual activities and products as well as the degree of satisfaction of the needs of cultural life, which is a reflection of the quality of spiritual life of the residents and one of the crucial criteria of the comprehensive and sustainable development of human beings. According to the above definition and a synthesis of current situation of the research at home and abroad, we will construct a primary index framework of culture well-being in terms of the education level and leisure resources (including multifarious arts, sports and recreational activities). The education level refers to the potency dimension of the input and achievements of the output while the leisure resources refer to the security condition and degree of satisfaction of the leisure products and services of the inhabitants.

Social well-being is the state which universally provides all citizens with a social system in accordance with law to guarantee a certain standard of living and improve the fund and services for quality of life as much as possible. It aims to solve all aspects of social welfare issues for the majority of social members, including protection and non-protection of the welfare as a whole. In the sub-classification level: (1) relief protection is a way of guarantee that the nation and region provided the basic living security relief to poverty residents or ones with temporary difficulties. It covered protection situation of minimum living groups, social vulnerable groups and temporary disadvantaged groups. (2) social

interaction is the government encouragement and support, social organizations and social members voluntarily organize and participate in the activities to help the weak; their funding mainly come from donations and the voluntary payment, and the government often provides taxation support. (3) welfare service is the state and region input state in terms of infrastructure and public services to protect and improve quality of life of the residents. (4) public safety is the life, health of most people and safety of public property, which is the basic element to protect the residents enjoying a normal quality of life.

Political well-being refers to citizens in certain society on democracy, freedom and human rights, rule of law and the results of psychological satisfaction. Political demand is developed on the basis of a higher level of economic demand, this demand mainly for the desire for political freedom, the acquisition and use of political power and the participation in political process, etc. Political well-being, namely to the extent that political rights are fully realized, it is an important component of individuals' whole quality of life. Furthermore, political well-being will be analyzed at the national level only and not included in the next research procedures.

3.2 The Selection of indicators

Based on a great number of western studies on well-being indicators construction, and previous Chinese scholars' researches of QOL, this research chose 115 indicators mainly from China Statistical Yearbook for the selection by Content Validity Ratio analysis and Correlation analysis. The indicators of national level were not analyzed in next step.

However, because of differences among social indicators, the non-dimensional indicators of treatment must be done before data analysis. This treatment was computed from the following formulas:

$$Z_i = \frac{X_i - X^i_{\min}}{X^i_{\max} - X^i_{\min}} \quad \text{(For the positive indicators, such as Life expectancy)}$$

$$Z_i = \frac{X^i_{\max} - X_i}{X^i_{\max} - X^i_{\min}} \quad \text{(For the negative indicators, such as Perinatal mortality)}$$

Where X_i is the one index data of one province in China mainland for a year, $X^i_{\min}$ is the minimum data of the same index in the same year among 31 provinces, $X^i_{\max}$ is the maximum data of the same index in the same year among 31 provinces, Z_i is the new data as a result of the non-dimensional treatment, and will be analyzed in next steps.

3.2.1 Content Validity Ratio analysis

Through literature searching, we chose a number of scholars in the QOL or other related research fields and got contact with them by mail, E-mail or telephone to help us on selecting the appropriate indicators from these alternative ones. Moreover, most of those experts also took part in another survey of our study for the AHP analysis.

The content validity ratio (CVR) is a psychometric concept representing the rating of item relevance to select items for the test. By completing the Likert scale we designed, every expert was asked to evaluate the relevance of each indicator that we chose to people's quality of life. Then, the content validity of ratio of each indicator was computed from the following formula (Lawshe, 1975):

$$\mathrm{CVR}=\frac{n_i-\frac{N}{2}}{\frac{N}{2}} \tag{1}$$

3.2.2 Correlation analysis

The correlation between social indicators is another factor that should be taken into account during the process of selection of indicators. It is unaccepted neither high nor low correlation. First of all, we collected data of each social indicator from 2000 to 2009 to form a database and did the non-dimensional treatment. Secondly, the correlation analysis was operated between every two indicators which were in the same sub-classification, and between the sum of all indicators from the same sub-classification and one of those indicators. Thirdly, we set the criterion in this step for indicators selection: if an indicator presents a highly positive correlation with the sum of all indicators in the same sub-classification as well as presents a low positive correlation with other indicators in the same sub-classification, it should be selected into our well-being index.

Then, by content validity ratio analysis and correlation analysis, a total of 56 social indicators were selected to constitute the objective part of this well-being index (**Table 1**).

Table 1. Description of composite indicator of objective well-being from China mainland

Well-being index	Sub-classification	Indicators
Health and Survival well-being	Health	A1 Number of doctors per thousand population
		A2 Mortality rate of children under-five
		A3 Perinatal mortality
		A4 Life expectancy
	Basic survival	A5 Coverage rate of rural population with access to running water
		A6 Consumer spending accounts for clothing in the ratio of per capita annual consumption expenditure
		A7 Per capita floor space of residential building
		A8 Coverage rate of urban population with access to gas
		A9 Coverage rate of rural population with access to sanitary latrines

		A10 Number of civilian passenger cars per thousand population
		A11 Coverage rate of road accessibility in rural and towns
		A12 Mortality rate per ten thousand cars
Economic well-being	Income level	B1 Residents' Per Capita Income
		B2 Gini Coefficient
	Consumption level	B3 Annual Per Capita Living Expenditure of Urban and rural residents Households
		B4 Consumer Price Index
		B5 Urban and rural residents Household's Engel's Coefficient
	Labor and Employment	B6 Registered Urban Unemployment Rate
		B7 The Ratio of Value-added of the Tertiary Industry to GDP
		B8 The Ratio of Total Wages Bill of Staff and Workers to GDP
Environment well-being	Resources and Environment quality	C1 Energy Consumption per Unit of GDP
		C2 Urban air quality compliance rate
		C3 Per Capita Green Covered Area in the urban
	Environment pollution and governance	C4 Total Volume of Industrial Waste Gas Emission
		C5 Industrial Wastewater Discharge compliance rate
		C6 Proportion of Comprehensive utilization of industrial solid waste
		C7 Investment in Anti-pollution Projects as Percentage of GDP
		C8 Proportion of Harmless Treated Garbage in the urban
Culture well-being	Education level	D1 Teacher-student ratio of elementary, middle and high education
		D2 Adult Literacy Rate
		D3 Average years of education of six years old and above
		D4 Ratio of Cultural and recreational consumption in aggregate consumption expenditure
	Leisure resources	D5 Per capita expenditure for operating of cultural undertakings
		D6 Users accessing the Internet per million
		D7 Popularity rate of the color television receiver
		D8 Numbers of published books, newspapers and magazines
Social well-being	Security and relief	E1 Basic social insurance coverage ratio
		E2 Capita unemployment insurance benefits paid
		E3 Urban low level of average expenditure
		E4 Rural minimum insurance average level of expenditure
		E5 Per expenditure on civil affair expenses
	Social solidarity	E6 Ten-thousand capita the number of social organizations
		E7 Number of social donations capita
	Welfare services	E8 Thousand capita the number of hospital beds
		E9 Ten-thousand capita the number of urban benefit the public networks
		E10 Ten-thousand capita the number of urban public toilets
Political well-being (National level)		

3.3. A Subjective and Objective Integrated Approach to Determine Attribute Weights

The combined method of analytic hierarchy process (AHP) and principal component analysis (PCA) was applied to the comprehensive evaluation of objective well-being from China mainland. In the first stage, the weights of each indicators, sub-classification and life domain were obtained by AHP, than an

objective analysis was made, namely PCA, to evaluate the objective well-being in this study.

3.3.1. Analytical Hierarchy Process

The Multi Criteria Decision Making (MCDM) is a set of techniques which is able to weight and score a range of criteria and then the scores are ranked by the expertise and other related interested groups. Analytical Hierarchy Process (AHP) is one of the MCDM methods which are widely applied to human field such as project design, policy evaluation, and resources allocation (Cheng et al., 2005). In addition, the previous researches show that this subjective approach of obtaining weights is very suitable for solving complicate issues (Yuksel et al. 2007), and it was applied in obtaining subjective weights for composite well-being index (Sedigheh, et al., 2009; Lee, et al., 2011).

In this research, the AHP procedure involves three basic steps (Satty, 1985):

(1) The hierarchy construction: in this stage, we broke down our well-being concept into its component parts of which every possible attributes are arranged into multiple hierarchical levels (**Fig. 2**). The criteria and sub-criteria are not each equally important to the decision at each level of the hierarchy and each alternative rates differently on each criteria. Then, we combined and consolidated the evaluations of the alternatives and criteria by group involved in the decision-making task by the analytical process provided by AHP. For the characteristics of respondents (experts) for AHP, you can see from the part of content validity ratio analysis in this paper.

(2) Defining and executing data collection to obtain pair-wise comparison data on elements of the hierarchical structure. Given a pair-wise comparison, the analysis involves two tasks: ① Developing a comparison matrix at each level of the hierarchy starting from the second level and working down. Those comparisons were carried out through expert survey. **Table 2** shows the 9-point scale used in typical analytic hierarchy studies and also applied in our study ranging from 1 (indifference or equal importance) to 9 (extreme preference or absolute importance). This pair-wise comparison enabled the decision maker (experts) to evaluate the contribution of each factor to the objective independently, thereby simplifying the decision making process. ② Computing the relative weights for each element of the hierarchy and estimating the consistency ratio (CR) to check the consistency of the judgment (Chen, C.F., 2006). After the survey, each expert's choice in AHP judgment was collected by computer software and the relative weights of the objectives and corresponding criteria and the consistency ratios of the matrices could be calculated.

Figure 2. Structured hierarchy of objective well-being from China mainland

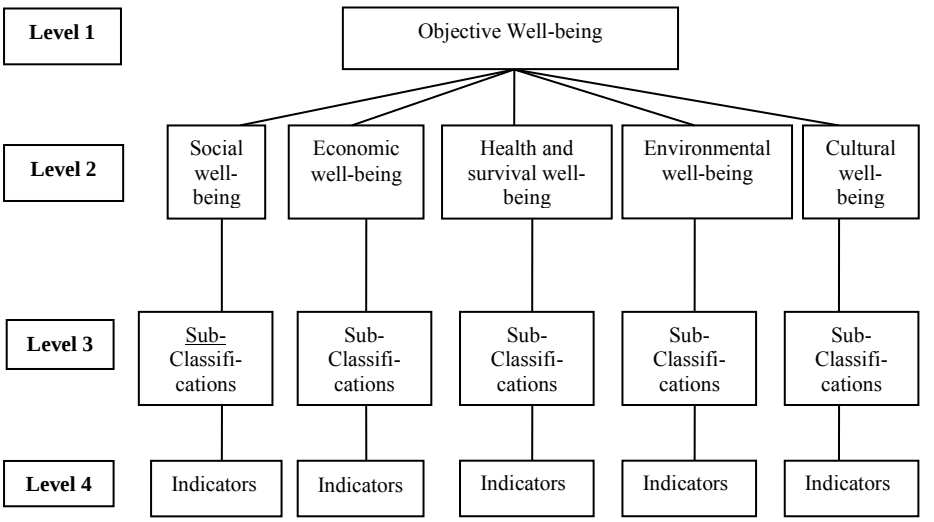


Table 2. 9-point intensity of relative importance scale (Satty and Kearns, 1985)

Intensity of Importance	Definition	Explanation
1	Equal importance	Two activities contribute equally to the objective
3	Weak importance of one over another	Experience and judgment slightly favor one activity over another
5	Essential or strong importance	Experience and judgment strongly favor one activity over another
7	Demonstrated importance	An activity is strongly favored and its dominance
9	Absolute importance	The evidence favoring one activity over another is of the highest possible order of affirmation
2,4,6,8	Intermediate values Between the two adjacent judgments	When compromise is needed
Reciprocals of above non-zero	If activity *i* has one of the above non-zero numbers assigned to it when compared with activity *j*, then *j* has the reciprocal value when compared with *i*	

Elements in each level are compared in pairs with respect to their importance to an element in the next higher level. Starting at the top of the hierarchy and working down, the pair-wise comparisons at a given level can be reduced to a number of square matrices A = $[\alpha_{ij}]_{n\times n}$ as in the following:

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ a_{1n} & a_{2n} & \cdots & a_{nn} \end{pmatrix}$$

The matrix has reciprocal properties, which are:

$$a_{ji} = 1/a_{ij}$$

After all pairwise comparison matrices are formed, the vector of weights, w = $[w_1, w_2, \ldots, w_n]$, is computed on the basis of Satty's eigenvector procedure. The computation of the weights involves two steps. First, the pairwise comparison matrix A = $[\alpha_{ij}]_{n\times n}$ is normalized by Eq. (2) and then the weights are computed by Eq. (3):

$$a_{ij}^{*} = a_{ij} / \sum_{i=1}^{n} a_{ij} \tag{2}$$

for all j = 1, 2, …,n

$$w_i = \sum_{j=1}^{n} a_{ij}^{*} /\mathrm{n} \tag{3}$$

for all j = 1, 2, …,n

Satty showed that there is a relationship between the vector weights, w and the pairwise comparison matrix, A, as shown in Eq. (4):

$$\mathrm{A}\, w = \lambda_{\max} w \tag{4}$$

The $\lambda_{\max}$ value is an important validating parameter in AHP and is used as a reference index to screen information by calculating the Consistency Ratio (CR) of the estimated vector. To calculate the CR, the Consistency Index (CI) for each matrix of order n can be obtained from Eq. (5):

$$CI = \lambda_{\max} \text{-n/n-1} \tag{5}$$

Then, CR can be calculated using Eq. (6):

$$CR = CI / RI \tag{6}$$

where, RI is the random consistency index obtained from a randomly generated pair-wise comparison matrix. **Fig. 3-8** shows the value of the CR. Suggested by Satty (Satty, 1980), if $CR < 0.1$, then the comparisons are

acceptable. If, however, $CR \geqslant 0.1$, the values of the ratio are indicative of inconsistent judgments and this expert's choice on this matrix will be deleted.

(3) Constructing an overall priority rating. The final weight was computed by computer software at group level (**Figure 3-8**).

Figure 3. Causal model of health & basic survival well-being (n=22)

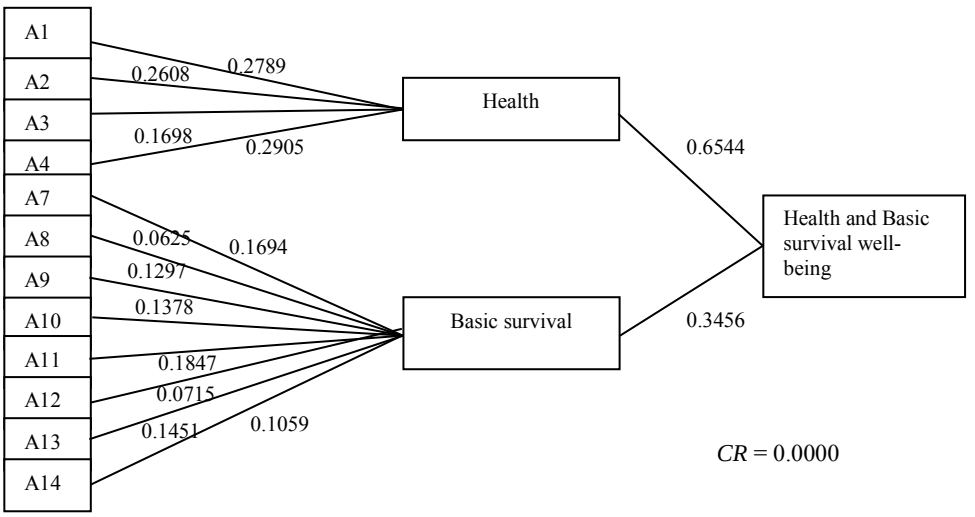


Figure 4. Causal model of economic well-being (n=22)

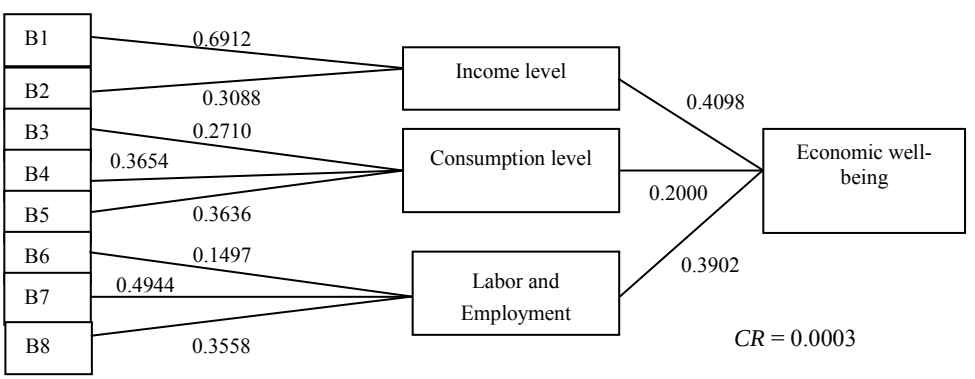


Figure 5. Causal model of environmental well-being (n=23)

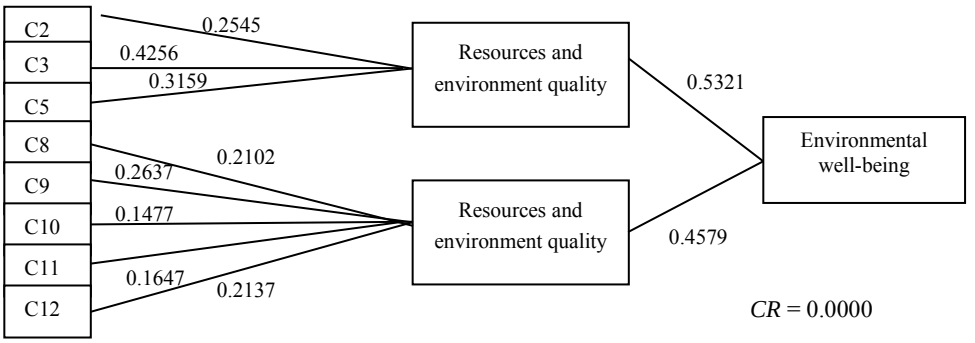


Figure 6. Causal model of cultural well-being (n=22)

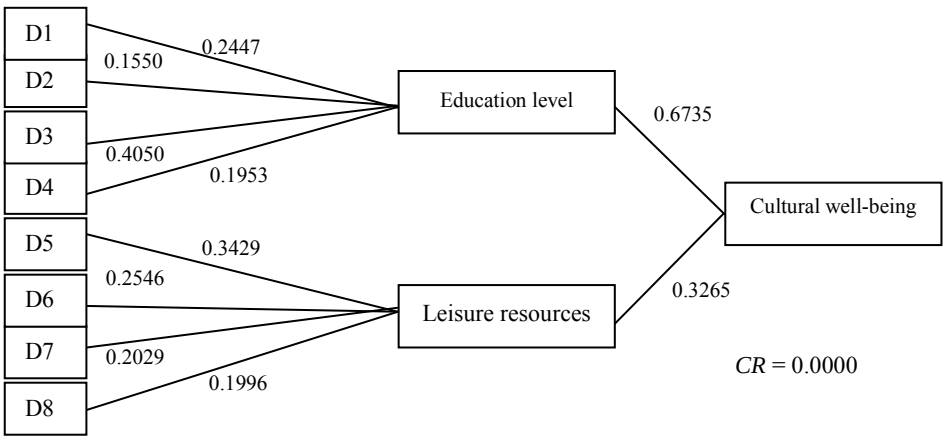


Figure 7. Causal model of social well-being (n=20)

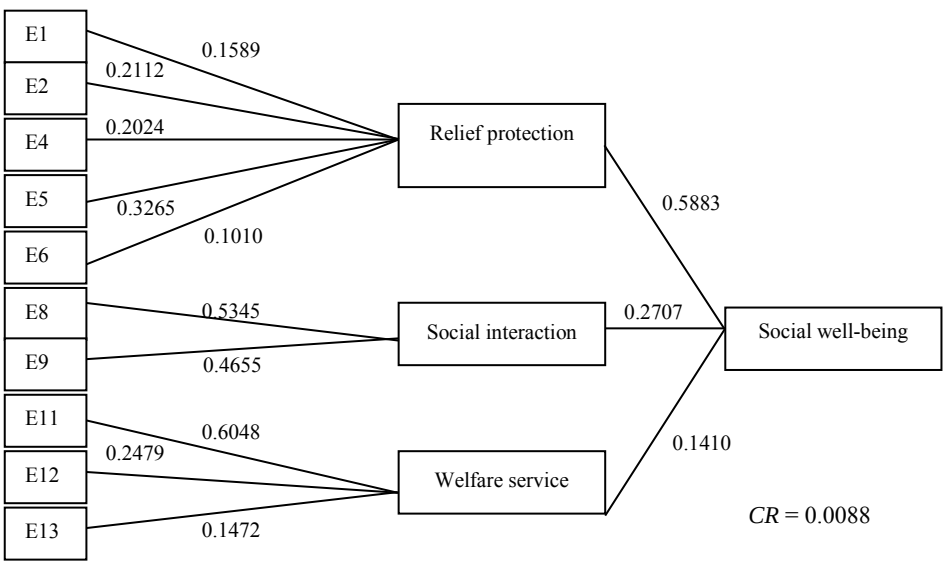
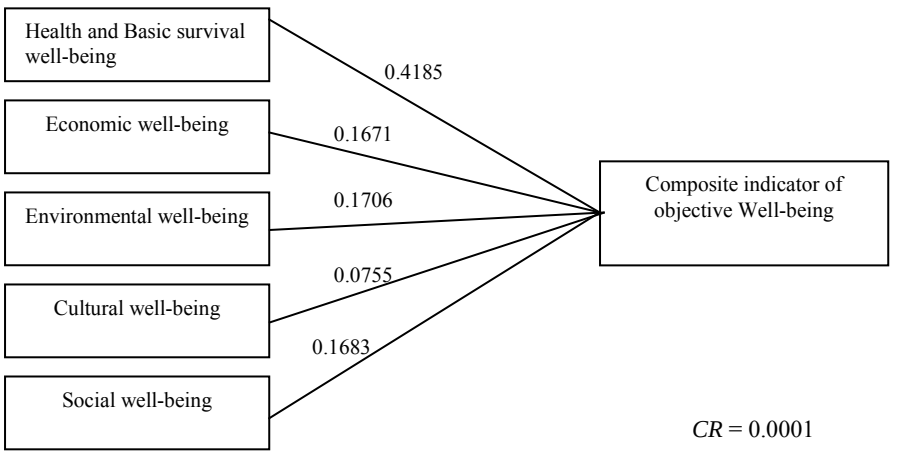


Figure 8. Causal model of objective well-being (n=22)



3.3.2. Principal Component Analysis

There is no denying that the AHP is a good approach to make the qualitative indicators can be measured. However, this approach has difficulty in reflecting the relationship between indicators, so in this step we introduced the Principal Component Analysis (PCA) to compensate for this limitation of AHP. The PCA is a multivariate statistical method for selecting the major factor from the objective data based on the variance to the target of an evaluation (Jolliffe, 2002). Regarding this principle of the PCA, first of all, we made transformation of 46 dimensionless indicators (the indicators of national level were not included) through weighted treatment by the following formulas:

$$Z_i^* = W_i Z_i \text{ (i=1, 2, 3…)} \tag{7}$$

Where Z_i^* is the new data for each indicator to be transformed by the weighted treatment, W_i is the weight for each indicator obtained from the AHP in last step, Z_i is the dimensionless indicator data without weighting. Then the PCA analysis based on covariance and involving three basic steps was made to this new database (there is a new corresponding data from 31 provinces in China mainland during 2004-2008 for each indicator), and these steps were shown by the example of the 'Health', a sub-classification from health and basic survival well-being:

(1) Principal component extraction. According to the principle of multivariate statistical analysis, if the contribution rate of the first *k* principal components reaches 85%, it shows that these *k* components contains the basic information of

all indicators. The contribution rate of variance here means the relative importance of each major component (Jolliffe, 2002). In order to make each principal component extracted reflects one single quality to the greatest extent, we did the extraction for each sub-classifications of life domains. From the extraction by SPSS 15.0 on computer, we got two major components for health, and they contain 87.06% of information of this part.

(2) Computing score for every major component. In this stage, initial principal component analysis, eigenvalue and contribution rate were taken into account to get an expression for each major component. For the start, the eigenvectors were computed by the following formula:

$$a_{jl} = \sqrt{\frac{\mu_{jl}}{\lambda_l}} \quad (j=1, 2 \ldots p; l=1, 2 \ldots k) \tag{8}$$

Where μ_{jl} is the initial loading value of the j-th indicator corresponding to the l-th factor (**Table 3** shows the initial loading values of 'Health'), while λ_l is the eigenvalue corresponding to the first factor. Second, we got the expressions of two major components or facts for 'Health' as follows:

$$F_1 = 0.392 W_{a1} Z_{a1} + 0.533 W_{a2} Z_{a2} + 0.503 W_{a3} Z_{a3} + 0.555 W_{a4} Z_{a4} \tag{9}$$

$$F_2 = 0.839 W_{a1} Z_{a1} - 0.217 W_{a2} Z_{a2} - 0.493 W_{a3} Z_{a3} + 0.065 W_{a4} Z_{a4} \tag{10}$$

Table 3. Component Matrix (Health)

	Component	
	1	2
1(A1)	.643	.745
2(A2)	.876	-.193
3(A3)	.826	-.438
4(A4)	.912	.058

Third, the evaluation score of 'Health' can be computed by this formula as follows:

$$Y_{HOWB} = 0.6737 F_1 + 0.1970 F_2 \tag{11}$$

Where '0.6737' and '0.1970' are the contribution rate of variance corresponding to these two components respectively.

Then, we could get the evaluation score of another sub-classification of Health and basic survival well-being by the same steps. Meanwhile, these two score together is the score of Health and basic survival well-being.

Through the approach and procedures above-mentioned we have got the score of each life domain, then the total scores of well-being of every province in China mainland was computed through the following formulas indirectly:

$$Y_{HOWB\&SUOWB}=0.429W_{a1}Z_{a1}+0.316W_{a2}Z_{a2}+0.241W_{a3}Z_{a3}+0.387W_{a4}Z_{a4}+0.175W_{a5}Z_{a5}+0.183W_{a6}Z_{a6}+0.300W_{a7}Z_{a7}+0.226W_{a8}Z_{a8}+0.232W_{a9}Z_{a9}+0.121W_{a10}Z_{a10}+0.207W_{a11}Z_{a11}+0.154W_{a12}Z_{a12} \quad (12)$$

$$Y_{ECOWB}=0.707W_{b1}Z_{b1}+0.472W_{b2}Z_{b2}+0.388W_{b3}Z_{b3}+0.418W_{b4}Z_{b4}+0.265W_{b5}Z_{b5}+0.510W_{b6}Z_{b6}+0.315W_{b7}Z_{b7}+0.313W_{b8}Z_{b8} \quad (13)$$

$$Y_{ENOWB}=0.021W_{c1}Z_{c1}+0.236W_{c2}Z_{c2}+0.535W_{c3}Z_{c3}+0.112W_{c4}Z_{c4}+0.276W_{c5}Z_{c5}+0.328W_{c6}Z_{c6}+0.063W_{c7}Z_{c7}+0.263W_{c8}Z_{c8} \quad (14)$$

$$Y_{COWB}=0.435W_{d1}Z_{d1}+0.331W_{d2}Z_{d2}+0.402W_{d3}Z_{d3}+0.263W_{d4}Z_{d4}+0.254W_{d5}Z_{d5}+0.430W_{d6}Z_{d6}+0.237W_{d7}Z_{d7}+0.406W_{d8}Z_{d8} \quad (15)$$

$$Y_{SOOWB}=0.292W_{e1}Z_{e1}+0.290W_{e2}Z_{e2}+0.304W_{e3}Z_{e3}+0.242W_{e4}Z_{e4}+0.416W_{e5}Z_{e5}+0.214W_{e6}Z_{e6}+0.707W_{e7}Z_{e7}+0.566W_{e8}Z_{e8}+0.150W_{e9}Z_{e9}+0.117W_{e10}Z_{e10} \quad (16)$$

$$Y_{OWB}=0.418Y_{HOWB\&SUOWB}+0.167Y_{ECOWB}+0.170Y_{ENOWB}+0.075Y_{COWB}+0.168Y_{SOOWB} \quad (17)$$

3.4. Results

By the process above, namely ACP-PCA approach, we got the score of objective well-being for 30 provinces of China mainland during 2004-2008 (owing to the lack of some data, the score of wellbeing for Xizhang Aut.Reg was not computed and did not participate in the final ranking). See in **Table 4** for the final results.

From the results in our research, we could find that the inequalities of objective well-being between different provinces and between different regions in China mainland are prominent issues due to the economic inequalities that are long-standing here. Moreover, this kind of regional disparity especially exists between the Western and Eastern area in China.

4. Conclusion and Discussion

Through constructing the comprehensive concept of well-being, this study is an attempt to compose objective well-being and subjective well-being together. In addition, it introduces the public policy-oriented perspective to the well-being study, which is reflected in the concept of well-being, the framework of indicators system, and the criteria in selecting indicators. Furthermore, according to the different disadvantages of subjective weighting approach and objective weighting approach, namely AHP and PCA respectively, this research adopts a synthesized method into the evaluation of objective well-being. So through this study, we gain experience in constructing the composite well-being index for a developing country. Obviously, since the concept of quality of life is multi-dimensions and involves many indicators, this paper only accomplished the task in evaluation of the objective part of well-being, the subjective well-being should be complemented next as well as the international comparison that is not included in this research.

Table 4. The score of objective well-being for every provinces of China mainland (2004-2008)

Objective well-being from China mainland (2004-2008)										
Provinces in China mainland	2004	Rank	2005	Rank	2006	Rank	2007	Rank	2008	Rank
Beijing	0.3899	1	0.3824	1	0.3965	1	0.3719	1	0.3782	1
Tianjin	0.2634	3	0.2689	3	0.2775	3	0.2786	3	0.2803	3
Heibei	0.1604	16	0.1685	13	0.1708	15	0.1635	17	0.1683	17
Shanxi	0.1723	11	0.1768	11	0.1828	11	0.1755	12	0.1833	11
Inner Mongolia Aut.Reg	0.1577	18	0.1630	17	0.1723	13	0.1682	14	0.1740	14
Liaoning	0.1902	8	0.2029	7	0.2071	8	0.1991	8	0.2080	7
Jilin	0.1886	9	0.1918	9	0.1949	10	0.1915	9	0.1946	10
Heilongjiang	0.1671	12	0.1744	12	0.1775	12	0.1789	11	0.1773	12
Shanghai	0.3274	2	0.3358	2	0.3428	2	0.3497	2	0.3500	2
Jiangsu	0.2102	6	0.2223	6	0.2323	6	0.2282	6	0.2336	6
Zhejiang	0.2398	4	0.2574	4	0.2622	4	0.2582	4	0.2602	4
Anhui	0.1347	26	0.1396	26	0.1503	24	0.1405	25	0.1494	26
Fujian	0.1846	10	0.1915	10	0.1982	9	0.1911	10	0.1992	9
Jiangxi	0.1446	23	0.1543	21	0.1570	23	0.1555	21	0.1648	19
Shandong	0.1908	7	0.1990	8	0.2110	7	0.2054	7	0.2054	8
Henan	0.1450	21	0.1511	23	0.1583	22	0.1545	22	0.1577	21
Hubei	0.1621	13	0.1661	14	0.1716	14	0.1642	16	0.1669	18

Table 4. The score of objective well-being for every provinces of China mainland (2004-2008) (cont.)

Objective well-being from China mainland (2004-2008)										
Provinces in China mainland	2004	Rank	2005	Rank	2006	Rank	2007	Rank	2008	Rank
Hunan	0.1456	20	0.1548	20	0.1625	20	0.1565	20	0.1641	20
Guangdong	0.2244	5	0.2308	5	0.2351	5	0.2308	5	0.2366	5
Guangxi Zhuang Aut.Reg	0.1449	22	0.1536	22	0.1593	21	0.1479	23	0.1518	25
Hainan	0.1602	17	0.1617	19	0.1634	19	0.1712	13	0.1754	13
Chongqing	0.1472	19	0.1621	18	0.1658	18	0.1648	15	0.1722	15
Sichuan	0.1307	27	0.1341	27	0.1398	27	0.1393	26	0.1522	22
Guizhou	0.1024	30	0.1115	30	0.1121	30	0.1037	30	0.0996	30
Yunnan	0.1096	29	0.1170	29	0.1198	29	0.1215	29	0.1296	28
Shaanxi	0.1426	24	0.1440	24	0.1465	25	0.1444	24	0.1525	23
Gansu	0.1294	28	0.1329	28	0.1319	28	0.1263	28	0.1306	29
Qinghai	0.1361	25	0.1397	25	0.1458	26	0.1358	27	0.1386	27
Ningxia Hui Aut.Reg	0.1618	15	0.1649	15	0.1697	16	0.1625	18	0.1717	16
Xinjiang Uygur Aut.Reg	0.1621	14	0.1645	16	0.1696	17	0.1567	19	0.1530	22
Xizhang Aut.Reg	—	—	—	—	—	—	—	—	—	—

REFERENCES

ALLARDT, E. (1993). Having, loving, being: An alternative to the Swedish model of welfare research, in M. Nussbaum and A. Sen (eds.), *The Quality of Life* (Oxford University Press, Oxford, pp. 88–94).

BANAI, R. (2005). Anthropocentric problem solving in planning and design, with analytic hierarchy process. *J. Arch. Plann. Res*, 22: 107-120. http://direct.bl.uk/bld/PlaceOrder.do?UIN=169190 608&ETOC=RN&from=searchengine.

CHEN, C., 2006. Applying the Analytical Hierarchy Process (AHP) approach to convention site selection. J. Travel Res., 45: 167-174. http://jtr.sagepub.com/cgi/reprint/45/2/167.

CHENG, E., LI, H. and YU, L. (2005). The Analytic Network Process (ANP) approach to location selection, A shopping mall illustration. Construct. Innovat., 5: 83-97. DOI: 10.1108/14714170510815195.

DIENER, E., & SUH, E. (1997). Measuring quality of life: Economic, social, and subjective indicators, *Social Indicators Research*, 40, 189–216.

Department of Environment, Food and Rural Affairs in UK (DEFRA) (2005), The wellbeing measurement in the national sustainable development indicators, http://www.defra.gov.uk/sustainable/government/progress/national/68.htm.

ERIKSON, R. (1993). Descriptions of inequality: The Swedish approach to welfare research, in M. Nussbaum and A. Sen. (eds.), *The Quality of Life* (Oxford University Press, Oxford, pp. 67-83).

FENG, X., YI S. (2000). The quality of family life in urban areas: indicators and the structures, *Sociological Studies (Chinese), 4*, 107-125.

JOLLIFFE, I.T.(2002). Principal component analysis, New York, Springer series in statistics, ISBN:0-387-95442-2, pp: 10-33.

LAWSHE, C.H. (1975). A quantitative approach to content validity. *Personnel Psychology,28,* 563-575.

LEE JAEHYUK, HAESEONG, J.E., JEONGSOO BYUN. (2011). Well-being index of super tall residential buildings in Korea, *Building and Environment, 46*, 1184-1194.

LIN, N., WANG L., PAN, Y., YUAN, G. (1989). The structures and the indicators of quality of life, *Sociological Studies (Chinese), 6,* 73-89.

LU, S. (1992). An analysis between marriage and quality of life of family in Chinese cities, *Sociological Studies (Chinese), 4,* 84-91.

SATTY, T.L. (1980). *Analytic Hierarchy Process*. New York, McGraw-Hill, ISBN: 096203178X, pp: 320.

SATTY, T.L. and K.P. KEARNS. (1985). *Analytical Planning: The Organization of Systems. Pergamon*, Oxford, ISBN: 0080325998, pp: 212.

SEDIGHEH LOTFI , KARIM SOLAIMANI. (2009). An assessment of Urban Quality of Life by Using Analytic Hierarchy Process Approach, *Journal of Social Sciences, 5,* 123-133.

SHEN, Q., LO K. and WANG, Q. (1998). Priority setting in maintenance: A modified multi-attribute approach using analytical hierarchy process. Construct. Manage. Econ., 16: 693-702.

http://ideas.repec.org/a/taf/conmgt/v16y1998i6p693-702.html.

YUKSEL, I. and DAGDEVIREN, M. (2007). Using the Analytic Network Process (ANP) in a SWOT analysis - A case study for a textile firm. J. Inform. Sci., 177: 3364-3382. DOI: 10.1016/j.ins.2007.01.001.

UUSITALO, H. (1994). Social statistics and social reporting in the Nordic countries, in P. Flora, F. Kraus, H.-H. Noll and F. Rothenbacher (eds.), *Social Statistics and Social Reporting in and for Europe* (Informationszentrum Sozialwissenschaften, Bonn, pp. 99–120).

XING, Z., HUANG, L. (2007). An analysis of subjective well-being among major social groups: survey from the coastal province in China mainland, *Social Sciences in Nanjing (Chinese), 1,* 83-97.

ZAPF, W. (1984). Individuaelle Wohlfahrt: Lebensbedingungen und wahrgenommene Lebensqualität, in W. Glatzer and W. Zapf (eds.), *Lebensqualität in der Bundesrepublik* (Campus, Frankfurt a.M./New York, pp. 13–26).

ZHOU, C., CAI, J. (2004). Quality of life indicators: from subjective perspective, *Wuhan University Journal (Philosophy and Social Sciences edition)(Chinese),5,* 582-587.

STATISTICS IN TRANSITION-new series, Summer 2012
Vol. 13, No. 2, pp. 439—444

REPORT

The XXX Conference on Multivariate Statistical Analysis MSA 2011, 7-9 November 2011, Łódź, Poland

The 30th Anniversary International Conference on **Multivariate Statistical Analysis** took place on November 7th-9th 2011 in **Łódź, Poland**. The organization of the conference was entrusted to **Professor Czesław Domański**, the Chair of **Department of Statistical Methods** in University of Łódź and the President of **Polish Statistical Association**. Multivariate Statistical Analysis Conference 2011 was dedicated to the Professor of mathematics Antoni Łomnicki, Professor of statistics and administrative law Józef Kleczyński and Professor of mathematics and history of science, a methodologist and a champion of science, an eminent teacher Samuel Dickstein.

The conference presented the latest theoretical achievements in the field of the multivariate statistical analysis and its applications. This was a continuation of the issues undertaken on the conferences of the past years. The scientific programme of MSA 2011 covered a wide range of various statistical problems, such as multivariate distributions, statistical tests, non-parametric inference, discrimination analysis, Monte Carlo analysis, Bayesian inference, application of statistical methods for finance, insurance, capital markets and risk management.

The MSA 2011 was opened by the Chairman of the Organizing Committee, **Professor Czesław Domański**. The opening speech also included the University of Łódź Rector, **Professor Włodzimierz Nykiel** and the Dean of the Faculty of Economics and Sociology of the University of Łódź, **Professor Jan Gajda**.

This year Multivariate Statistical Analysis Conference began with the historical session, which was leaded by **Professor Eugeniusz Gatnar** (University of Economics in Katowice). Firstly, **Professor Mirosław Krzyśko** (Adam Mickiewicz University in Poznań) presented Antoni Łomnicki's biography, character and scientific achievements of this great mathematician, and also discussed the main results which Łomnicki gained in the probability theory, mathematical statistics, cartography and teaching of mathematics (in the memory of Antoni Łomnicki on the seventieth anniversary of his death).

As the session was also dedicated to the honoured Józef Kleczyński and Samuel Dickstein, **Professor Czesław Domański** (University of Łódź) presented their profiles and achievements in his speeches.

Also, for this session **Professor Jerzy Kowaleski** (University of Łódź) prepared a speech about life and work of Zbigniew Michałkiewicz. Moreover, a tribute was given to the following scientists, who died in 2010: Wiesław Sadowski (by Władysław Welfe, University of Łódź), Zofia Zarzycka (by Wacława Starzyńska, University of Łódź), Zbigniew Czerwiński (by Grażyna Dehnel, Poznań University of Economics), Wiesław Wagner (by Bronisław Ceranka, Poznań University of Life Sciences) and Piotr Chrzan (by Janusz Wywiał, University of Economics in Katowice)).

Altogether there were nearly 100 participants from various academic and research centres in Poland and abroad. Concerning the papers, 46 papers were presented in 12 sessions.

In addition, on the 8th November a meeting of **The Main Board of Polish Statistical Association** and The Organizing Committee of **The Congress of Polish Statistics** took place.

For the second plenary session (**Professor Janusz Wywiał,** University of Economics in Katowice, was the Chair of this session) the following papers were presented:

- Samaradasa Weerahandi (Pfizer Inc., New York), *Generalized Point Estimation with Application to Mixed Models*
- Kian-Guan Lim (Singapore Management University), *Statistical Tests of Conditional Shortfalls and Change-Point Estimation in Risk Management*
- Muhammad Aslam (Universite de Bourgogne, Dijon), *Robust Inference for Linear Regression Model in the Presence of Heteroscedasticity and Multicollinearity*
- Sajjad Haider Bhatti (Universite de Bourgogne, Dijon), *A Panel Data Analysis of the Relationship between GDP per Capita and its Shares of Consumption, Government Expenditures and Investment*

During the third plenary session (**Professor Walenty Ostasiewicz,** University of Economics in Katowice, was the Chair of this session) the following papers were presented:

- Eugeniusz Gatnar (University of Economics in Katowice), *Central Bank in the system of official statistics*
- Bronisław Ceranka, Małgorzata Graczyk (Poznań University of Life Sciences), *Construction of regular E-Optimal weighing design*
- Małgorzata Graczyk (Poznań University of Life Sciences), *Regular A-Optimal spring balance weighing design with correlated errors*
- Tomasz Żądło (University of Economics in Katowice), *On parameter estimation of some longitudinal model*

The second day of the MSA 2011 Conference was opened by the invited paper of **Professor Stanisława Bartosiewicz** (Wroclaw School of Banking) titled *Some "unkempt" thoughts on the subject of treatise on dynamics, condition and prospects of development of Polish agriculture.* Professor Bartosiewicz proposed the formation of a Treatise on Polish Agriculture, Forestry, Hunting and Fishery in order to commemorate the Centenary of the Polish Statistical Association, which falls in the year 2012. The following problems were pointed out:

- How Polish society benefits from agriculture, hunting, forestry and fishing?
- What is the economic efficiency of Polish agriculture, hunting, forestry and fishery?
- What is the dynamics and the state of Polish agriculture in comparison to other countries?
- Is subjectivity in creating scenarios of development useful in social terms?

Professor Czesław Domański also invited **Professor Mirosław Krzyśko** (Adam Mickiewicz University in Poznań) to present a lecture about *Kernel version of the functional principal components* (prepared by Tomasz Górecki and Mirosław Krzyśko). In this presentation a new construction of functional principal components were proposed, based on principal components for vector data.

Titles of the papers of the next sessions of the MSA Conference, with the authors' names are presented below:

8th November 2011

Session IV. (**Professor Marek Walesiak,** Wroclaw University of Economics, was the Chair of the session)

- Already mentioned two invited papers by Professor Stanisława Bartosiewicz and Professor Mirosław Krzyśko
- Wojciech Zieliński (Warsaw University of Life Sciences – SGGW), *Comparison of confidence intervals for fraction in finite populations*
- Joanna Dębicka (Wroclaw University of Economics), *Matrix approach to analysis of a portfolio of multistate insurance contracts*

Session V. (**Professor Wojciech Zieliński,** Warsaw University of Life Sciences – SGGW, was the Chair of the session)

- Justyna Brzezińska (University of Economics in Katowice), *Model selection methods in log-linear analysis*

- Joanna Kisielińska (Warsaw University of Life Sciences – SGGW), *The exact bootstrap method shown on the example of the mean and variance estimation*
- Wojciech Gamrot (University of Economics in Katowice), *On Fattorini's Estimator*
- Magdalena Chmielińska (University of Economics in Katowice), *The influence on choice of the method of inspection at Total Quality Control cost*

Session VI A. (**Professor Agnieszka Rossa,** University of Łódź, was the Chair of the session)

- Grażyna Dehnel, Elżbieta Gołata (Poznań University of Economics), *The social dimension of demographic processes*
- Dorota Raczkiewicz (Warsaw School of Economics), *Educational, professional and family careers of women in Poland*
- Beata Bieszk-Stolorz, Iwona Markowicz (University of Szczecin), *Men's and women's economic activity in Poland. Interpretation of the logit model with one and many explanatory variables*
- Andrzej Dudek (Wroclaw University of Economics), *Classification of large data sets. Problems, methods, algorithms*

Session VI B (**Professor Grażyna Trzpiot,** University of Economics in Katowice, was the Chair of the session)

- Jerzy Korzeniewski (University of Łódź), *Assessment of Talavera method of variable selection in clustering on real world data sets*
- Meral Yay (Mimar Sinan University of Fine Arts), *A Content Analysis of the Representation of the European Uprising in Turkish News Media*
- Marta Małecka (University of Łódź), *GARCH process application in risk valuation for WIG20 index*

Session VII A. (**Professor Grzegorz Kończak,** University of Economics in Katowice, was the Chair of the session)

- Janusz Wywiał (University of Economics in Katowice), *On estimation of domain mean*
- Mariusz Kubus (The Opole University of Technology), *On model selection in some regularized linear regression methods*
- Paweł Kobus (Warsaw University of Life Sciences – SGGW), *Application of copulas to modelling joint distribution of crop plants yield and price*

Session VII B. (**Professor Elżbieta Gołata,** Poznań University of Economics, was the Chair of the session)

- Artur Zaborski, Marcin Pełka (Wroclaw University of Economics), *Geometrical presentation of preferences by using PROFIT analysis and R program*
- Marcin Pełka (Wroclaw University of Economics), *Clustering of symbolic data with application of ensemble approach*
- Małgorzata Szerszunowicz (University of Economics in Katowice), *Some construction of factorial designs*

9th November 2011

Session VIII A. (**Professor Bronisław Ceranka,** Poznań University of Life Sciences, was the Chair of the session)

- Aleksandra Baszczyńska (University of Łódź), *Some Remarks on the Symmetry Kernel Test*
- Dorota Pekasiewicz (University of Łódź), *Bayesian statistical tests for fraction for independent and dependent sampling*
- Jacek Stelmach (University of Economics in Katowice), *About the testing of the differences between population with eigenvectors*
- Magdalena Makuch (University of Economics in Katowice), *On cluster sampling strategy*

Session VIII B. (**Professor Janusz Wywiał,** University of Economics in Katowice, was the Chair of the session)

- Małgorzata Misztal (University of Łódź), *Some remarks on the data imputation using „missForest" method*
- Jacek Białek (University of Łódź), *Proposition of the price index*
- Robert Pietrzykowski (Warsaw University of Life Sciences – SGGW), *The application of modified weight matrix defining the spatial interactions*
- Krystyna Pruska (University of Łódź), *Probit and logit models in small area estimation of proportion*

Session IX. (**Professor Krystyna Pruska,** University of Łódź, was the Chair of the session)

- Dorota Rozmus (University of Economics in Katowice), *Comparison of accuracy of spectral clustering and cluster ensembles based on co-occurrence matrix*

- Grzegorz Kończak (University of Economics in Katowice), *On testing directional hypothesis in the data homogeneity analysis*
- Wioletta Grzenda (Warsaw School of Economics), *The significance of prior information in Bayesian parametric survival models*

The next Multivariate Statistical Analysis Conference aka MSA 2012 is planned on **November 12th-14th, 2012** and will take place in **Lodz**. The Chairman of the Organizing Committee, **Professor Czesław Domański** informs this will be the 31st edition of the Conference and kindly invites all interested scientists, researchers and students to take part in it.

Prepared by:
Aleksandra Kupis - Fijałkowska
Anna Witaszczyk