STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

Volume 11, Number 1, July 2010

ISSN 1234-7655

Address for correspondence

GUS, al. Niepodległości 208, 00-925 Warsaw, POLAND, Tele/fax: 00 48 22 – 825 03 95

STATISTICS IN TRANSITION-new series, July 2010
Vol. 11, No. 1, pp. 1—4

FROM THE EDITOR

This issue of the *Statistics in Transition new series* (the first number of its eleventh volume) is composed of two sets of articles prepared by 25 authors. First group of papers is devoted to sampling and estimation methods, another to different and mutually unrelated topics. They are complemented by reports from two conferences, including one devoted to comparative statistics on income in the European Union countries that took place in Warsaw on early springtime.

The issue starts with **Amjad D. Al-Nasser**'s paper *An information-theoretic approach to the measurement error model* in which the idea of generalized maximum entropy (GME) estimation approach is employed to fit the general linear measurement error model and to improve the parameter estimation while signifying the additional assumptions that are needed in the traditional MLE. A Monte Carlo comparison shows that the GME is outperformed the MLE estimators in terms of mean squared error. Also, the real data analysis gives similar results that confirm the robustness of GME over the MLE estimation method, just suggesting that the GME can be considered a good alternative in estimating the ultrastructural models.

In the next paper *Clustering financial data using the Copula–GARCH model in an application for main market stock returns* **Anna Czapkiewicz** and **Beata Bisiura** show how to obtain the dependency parameter between time series with the Copula–GARCH model when returns are non-normal and it is impossible to specify the multivariate distribution relating two or more return series. The dissimilarity measure based on the maximum likelihood parameter obtained from the Normal or t-Student copula is proposed and applied to classify forty two indices from American, European, and Asian stock markets.

A group of authors, **Gyan K. Agarwal, Manish Trivedi, Usha Jha, Rajendra K. Jha** and **Ripunjai K. Shukla** who conducted an intensive survey among the persons living in the newly-formed Jharkhand State of Eastern India - to investigate the tobacco use in various age groups in adults - reports on the survey results in the paper *Tobacco Use And Its Impact On Respiratory Health: A Statistical Approach.* Using the multiple regression, Chi-square test and analysis of variance (ANOVA), they found that about 66.5 percent of the adult population exposed to tobacco is defined by the 18-25 years while 21 percent as defined by the 26-33 years. They conclude also that the Eastern India, especially the Jharkhand state, has a huge problem of widespread tobacco use among adults, particularly among the 18-25 year age group.

In the paper *Bhattacharya and Holla's Distribution and some of its Interesting Properties and Applications,* **Anwar Hassan Mehraj** and **Ahmed**

Bhat discuss important structural properties of this distribution they derive due to employing trigonometric transformation functions to this type of compound distributions; they also show how to obtain its higher order moments with an alternative expression. An attempt is also being made to obtain estimate of its parameters – a model is fitted to the set of observed data (with relative precision and a comparison of the estimates).

Krishan K. Pandey and **G.C. Tikkiwal** in the paper *Generalized Class of Synthetic Estimators for small areas under systematic sampling scheme* address the need to develop alternative estimators to provide small area statistics using auxiliary information under – specifically, the data already collected through large-scale surveys. The generalized class of synthetic estimators – that, among others, includes the simple, ratio and product synthetic estimators – are used for estimating crop acreage for small domain. Also, their relative performance is compared empirically with direct estimators through a simulation study.

The next paper, *Extreme Value Modelling of the Maximum Temperature: A Case Study of Humid Subtropical Monsoon Region in India* by **Ripunjai K. Shukla, Manish Trivedi** and **Manoj Kumar** is devoted to a special type of problem of fitting a distribution model to the long-run climate data, focusing on extreme events. In particular, they attempt to identify and fit the generalized extreme value distribution for extreme maximum temperature data of Ranchi District (located in Humid Sub-Tropical Monsoon Region) by using the method of maximum likelihood. The estimates of 10, 50, 100 and 200 years return level for yearly extreme maximum temperature are described in that how they vary in future.

G. N. Singh and **Kumari Priyanka,** in their paper *Estimation of Population Mean at Current occasion in Presence of Several Varying Auxiliary Variates in Two-Occasion Successive Sampling*, attempt to improve the precision of estimates at current occasion in two-occasion successive sampling using several varying auxiliary variates. They propose two different efficient estimators and examine their theoretical properties. Relative comparison of efficiencies of the proposed estimators with the sample mean estimator when there is no matching from previous occasion, and the optimum successive sampling estimator when no auxiliary information is used have been incorporated. Empirical studies are significantly justifying the composition of proposed estimators.

The paper *Robust estimation of finite population total* by **Srivastava M. K**. and **Srivastava N**., deals with the robust prediction of finite population total under the superpopulation model while involving reweighed iterative algorithm for Robust Prediction and making comparison among all resistant estimators. The discussion also involves the calculation of asymptotic bias and variance in terms of the influence function computed for these predictors, which are sensitive to the presence of outliers. Two populations have been considered for simulation study to judge the performance of proposed predictors with conventional and model based existing alternatives.

The 'other' articles section that contains five papers begins with **Jacek Białek**'s paper *The Generalized Formula For Aggregative Price Indexes*. The proposed formula for aggregative price indices intends to satisfy most of the postulates coming from axiomatic price index theory while embracing the ideal Fisher index and the Lexis index as its special cases. It is also shown how the generalized formula can be used to define new indices, which also satisfies defined postulates. The geometric mean of the indexes coming from the considered class generates the new indexes, which also belong to the same class.

Czesław Domański in *Polish Statistics Day* provides us with a brief description of the National Day of Polish Statistics organized by the Main Council of the Polish Statistical Society in cooperation with the Committee for Statistics and Econometrics of the Polish Academy of Sciences and the President of the Central Statistical Office. They all agreed upon the particular date for celebrating it, 9th of March, to commemorate the anniversary of the first General Census that was ordered by the Polish parliament (so called Four-Year Sejm) on March 9, 1789.

In the paper *The economic aspects of contradiction between generations in Poland* **Izabella Kudrycka** and **Małgorzata Radziukiewicz** discuss the issue of inter-generation conflicts focusing on the economic aspects of contradictions between age cohorts, especially on young-old groups comparisons in the contexts of selected dimensions of welfare (Social Security system, credit and debt on macroeconomic level, social welfare, income, consumption and taxation, revenue and expenditure of the state budget).

Piotr Tarka in *Statistical Choice between Rating or Ranking Method of Scaling Consumer Values* addresses some measurement problems that are caused by methodological inadequacies and assumptions concerning different types of scales. Focusing on rating and ranking methods applied to consumers' values evaluation, the author recommends a cautionary approach, emphasizing a unique conditions under each particular measurement process takes place.

Acknowledging the rising importance of spatial units in public statistics, **Wiesław Wagner** and **Andrzej Mantaj** in *Contiguity Matrix of Spatial Units and its Properties on Example of Land Districts of Podkarpackie Province* discuss the issue of their description and their properties while using contiguity matrix approach to determine mutual positions of spatial units. Such an approach allows for broadening the interpretation of researched social-economic phenomena since its elements can be used to determine connected domains, constituting some subsets of spatial units, which is met in taxonomical methods.

Reports on two conferences conclude this issue. The *2010 International Conference on Comparative EU Statistics on Income and Living Conditions* organised by the Statistical Office of the European Communities (Eurostat) and the Network for the analysis of EU-SILC (Net-SILC) was hosted by the Central Statistical Office of Poland in Warsaw (from 25th to 26th March 2010). And 15th annual conference of SCORUS (the Standing Committee on Regional and Urban

Statistics) devoted to *Relations between Generations and Challenges of an Ageing Society* was held in Berlin on 29-31 March 2010.

Wlodzimierz OKRASA
Editor-in-Chief

STATISTICS IN TRANSITION-new series, July 2010
Vol. 11, No. 1, pp. 5—6

Message from The Editor

COMPARATIVE SURVEYS: A NEW SECTION IN THE JOURNAL

The need for and the availability of data for comparative, multi-population research has greatly increased over recent years. Diverse factors have contributed to the rise of internationally comparable sources of statistics, and there are many consequences of this development. "Though international comparisons in statistics ... have been made for along time, the deliberate design of valid and efficient multinational surveys is new and is increasing."[1] The need for comparative information is manifest at both the national and the international level. Countries need to assess their place in relation to other countries, especially their geo-political neighbours. International agencies require similar data on different countries. The rise of super-national organisations and of processes occurring trans-nationally is a major factor, of special interest to us in the context of the EU. Researchers made already impressive progress in generating and analyzing data from deliberately designed comparative surveys in multinational, multiregional and multicultural contexts.[2] And are eager to employ the newly elaborated methodological strategies, and to share the results for supporting research-informed policies in analogous, supranational contexts.

All this is helped by the development of an intellectual atmosphere in which researchers are increasingly looking for internationally comparable datasets.[3]

We would like to announce the establishment of a new section named "Comparative Surveys" as a regular part of *Statistics in Transition*. Comparability of statistical data, i.e. their usefulness in drawing comparisons and contrast among different populations, is a complex concept; nevertheless, it is a fundamental requirement for any data to be used in multi-population comparisons and contrasts. Many additional conceptual, technical and operational issues need to be taken into account in generating comparable results – a set of comparable surveys is more than the sum of the individual surveys comprising it. The broad objectives

[1] Kish, L. (1994). Multi-population survey designs: five types with seven shared aspects. *International Statistical Review*, 62(2), pp. 167–186.

[2] Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L., Mohler, P. Ph., Pennell B-L., Smith, T.W. (2010) Survey Methods in Multinational, Multiregional and Multicultural Contexts. John Wiley & Sons, Inc., Hoboken, New Jersey.

[3] Verma, V. (2002). Comparability in multi-country survey programmes. *Journal of Statistical Planning and Inference,* vol. 102(1), pp. 189–210.

of the papers included in this section of the Journal will be the discussion of issues relating to the meaning, generation and use of comparable data, whether across countries, regions, sub- or super-populations or time.

We look forward to receiving interesting and informative contributions from the scientific community for the new Comparative Surveys section, the launching of which will take place in the next (Autumn) issue of the journal.

I have invited one of our Associate Editors, Professor Vijay Verma of Siena University, to assist in the development of this new section of the Journal. As many of our readers are well aware, Professor Verma has been involved in the development of international statistics throughout his working life and contributed significantly to several areas of comparative survey methodology. I have requested him to join me in soliciting, selecting and reviewing articles for the new section, and hopefully, also to continue to make regular contributions himself. I am grateful to Professor Verma for his generous acceptance of my request to play such an exceptional role in (co)editing of the Journal's new section.

Wlodzimierz OKRASA
Editor-in-Chief

STATISTICS IN TRANSITION-new series, July 2010
Vol. 11, No. 1, pp. 7

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, July 2010
Vol. 11, No. 1, pp. 9—24

AN INFORMATION – THEORETIC APPROACH TO THE MEASUREMENT ERROR MODEL

Amjad D. Al-Nasser

ABSTRACT

In this paper, the idea of generalized maximum entropy estimation approach (Golan et al. 1996) is used to fit the general linear measurement error model. A Monte Carlo comparison is made with the classical maximum likelihood estimation (MLE) method. The results showed that, the GME is outperformed the MLE estimators in terms of mean squared error. A real data analysis is also presented.

Key words: Measurement Error Model, Generalized Maximum Entropy, Maximum Likelihood

1. Introduction

The traditional maximum entropy formulation is based on the entropy-information measure of Shannon (1948). It is developed and described in Jaynes (1957a, 1957b). Shannon's entropy measure reflects the uncertainty about the occurrence of a collection of events. Suppose we have a set of events $\{x_1, x_2, \ldots, x_k\}$ whose probabilities of occurrence are $p_1, p_2, \ldots, p_k$. then using an axiomatic method to define a unique function to measure the uncertainty of a collection of events, Shannon (1948) defines the entropy of the distribution (discrete events) as the average of self-information $H(P) = -\sum_{i=1}^{k} p_i \ln(p_i)$, where 0ln(0) = 0.

Since the 1990's many attempts have been made to apply the method of maximum entropy in the area of linear models. In 1996, Golan et al. proposed an estimator based on the maximum entropy formalism of Jaynes that they called the generalized maximum entropy (GME) estimator. The logic of using GME estimation method is that GME tends to dominate traditional estimation methods with small samples, and it does not rely on any distributional assumption, also it is robust in fitting nonlinear models (Golan, 2008; Peeter, 2004). The idea underling the GME approach is to view each unknown parameter and error terms as an expected value of some proper probability distribution; then by maximizing

the joint entropies subject to the data, represented by each unobserved value, and the requirement for proper probability distributions; better estimates can achieved with less assumptions (Csiszar (1991), Donho et al. (1992), Golan, et al. (1997), Al-Nasser (2003), Al-Nasser (2004) and Al-Nasser et al (2006)).

The remainder of this paper is divided into six sections. Section 2 presents the Ultrastructural model. Section 3 presents the generalized maximum entropy estimation approach idea in general linear model. Section 4 introduces the GME to ultrastructural model, Section 5 gives a real data analysis to study the relationships between number of crimes and number of unemployed citizens in Jordan , Section 6 presents Monte Carlo evidence on the numerical performance of GME and maximum likelihood estimation (MLE), and Section 6 presents concluding comments.

amjadn@yu.edu.jo
Department of Statistics, Science faculty
Yarmouk University, 21163 Irbid
Jordan

2. The Model

Consider the simple linear relationship between two mathematical variables ξ and η

$$\eta = \alpha + \beta\xi$$

where α is the intercept and β is the slope. The classical theory of regression analysis assumes that these variables are measured without error; particularly in the social sciences and natural science this assumption is often violated. Hence, this linear relationship is reformulated such that both variables are contaminated with measurement errors. Then, the observed values can be defined by:

$$\begin{aligned} x_i &= \xi_i + \delta_i , \\ y_i &= \eta_i + \varepsilon_i , \quad i = 1,2,...,n \end{aligned} \tag{1}$$

where δ and ε are the measurement errors associated with ξ and η, respectively. It is assumed that $\delta_1, \delta_2, ..., \delta_n$ are identically and independently distributed (*i.i.d*) with mean 0 and variance σ_δ^2. Similarly, $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ are *i.i.d* with mean 0 and variance σ_ε^2. Further, suppose that the true values of the variable (ξ) have possibly different means; say $\mu_1, \mu_2, ..., \mu_n$, so that we can write

$$\xi_i = \mu_i + \omega_i , \quad i = 1,2,...,n \tag{2}$$

where $\omega_1, \omega_2, ..., \omega_n$ are *i.i.d* with mean 0 and variance σ_ω^2. Finally, assume that the distribution of $(\delta_1, \delta_2, ..., \delta_n)$, $(\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)$ and $(\omega_1, \omega_2, ..., \omega_n)$ are mutually independent of each other. This provides the specification of ultrastructural model.

This model can be reduced to the functional measurement error model if we assumed that ξ_i are fixed; i.e. $\omega_i = 0$, $i = 1,2,...n$. Also it can be reduced to the structural measurement model if we assumed $\mu_i = \mu_j, \forall i, j = 1,2,..., n$. On the other hand, if $\delta_i = 0$, $i = 1,2,...n$ then the ultrastructural model reduces to the classical regression model with no measurement error; for more details see, Dolby (1976) and Cheng and Van Ness (1999).

Assuming Normal measurement errors, then the MLE for the parameters of the ultrastructural model are unidentifiable, Shalabh (1997). Hence, to solve the ultrastructural model (1-2), the MLE method required additional assumptions; Dolby (1976) derived the MLE estimates when the ratios $\lambda = \dfrac{\sigma_\varepsilon^2}{\sigma_\delta^2}$, and $\nu = \dfrac{\sigma_\omega^2}{\sigma_\delta^2}$ are known. Then, under the customary assumptions, the MLE for the ultrastructural model are the results in a quintic equation for $\hat{\beta}$;

$$h(\hat{\beta}) = \nu S_{xx}\hat{\beta}^5 + \left(3\nu\lambda S_{xx} - \nu S_{yy}\right)\hat{\beta}^3 - 2\lambda(\nu - 1)S_{xy}\hat{\beta}^2$$
$$+ \left\{2\lambda^2(\nu + 1)S_{xx} - \lambda(\nu + 2)S_{xx}\right\}\hat{\beta} - 2\lambda^2(\nu + 1)S_{xy} = 0$$

This generally should be solved by iterative numerical methods. However, Gleser (1985) showed that the likelihood is maximized on the $\sigma_\omega^2 = 0$; boundary of the parameter set. Thus, the ML estimates for the ultrastructural model are just the ML estimates of the functional model; which leads to

$$\hat{\beta} = \frac{(S_{yy} - \lambda S_{xx}) + ((S_{yy} - \lambda S_{xx})^2 + 4\lambda S_{xy}^2)^{1/2}}{2S_{xy}}$$

The ML solution of the other parameters will be:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \quad \hat{\mu}_i = \frac{\lambda x_i + \hat{\beta}(y_i - \hat{\alpha})}{\lambda + \hat{\beta}^2}$$

where

$S_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2 \Big/ n$, $S_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2 \Big/ n$, $S_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \Big/ n$, $\bar{x}=\sum_{i=1}^{n} x_i \Big/ n$ and $\bar{y}=\sum_{i=1}^{n} y_i \Big/ n$

For more details see Cheng and Van Ness (1999), Carroll et al. (1995).

3. Generalized Maximum Entropy

Consider the general linear model; $y_i = f(x_i, \beta) + \varepsilon_i$; i = 1,2,…n. where $\beta = (\beta_1, \beta_2, ..., \beta_K)'$ is the vector of parameters to be estimated, the regressor variable x_i, $i = 1,2,...n$ are *K*-dimensional vectors whose values are assumed known and ε_i, $i = 1,2,...,n$ is the random error.

In GME, the model is fitted after some reformulation of the unknown parameters β, and the unknown error term ε_i, i = 1,2,3,…, n; if they are not in probability format. This can be done by reparameterize their possible outcome values probabilistically as a convex combination of random variables. This combination is presented as expected value of some proper probability distribution. For each unknown, assume that there exists a discrete probability distribution that defined over the parameter space [0, 1]; by a set of equally distanced discrete points; then the formulation of model parameter will be of the form $\beta = ZP$; where Z is a ($K \times KR$) matrix and P is a KR-vector of weights such that $p_k > 0$ and $P_k' 1_R = 1$ for each k. Simply, each β_k, $k = 1,2,...,K$ can be defined by a set of equally distanced discrete points $Z_k' = [z_{k1}, z_{k2}, z_{k3}, ..., z_{kR}]$, where $R \geq 2$ with corresponding probabilities $P_k' = [p_{k1}, p_{k2}, p_{k3}, ..., p_{kR}] \in [0,1]$. That is,

$$\beta_k = \sum_{r=1}^{R} z_{kr} p_{kr} \ , \sum_{r=1}^{R} p_{kr} = 1, \ 0 \leq p_{kr} \leq 1, \text{ k = 1,2,…,K}$$

The support points Z for β are known. Golan et al. (1996) suggested that these values can be specified uniformly symmetric around *0* with large value of the lower and the upper bounds. For example, one can select the support point $z = (-c, 0, c)$, given that c is a large value (i.e. c = 100 or 1000 etc.). Moreover, assuming one specify Z to span the true values of β then the GME is a consistent estimator, which is an advantage of this method.

The disturbance ε_i can be treated in similar fashion. For a set $T \geq 2$ support points ε_i is assumed to be bound between two finite values v_t , v_T, which are symmetric around zero with corresponding unknown probability weights w_{1t}, w_{nT} .

That is, each error term may be modeled as; $\boldsymbol{\varepsilon} = VW$, where V is a $(n \times nT)$ matrix and W is a nT-dimensional vector of weights; that is to say:

$$\varepsilon_i = \sum_{j=1}^{T} v_{ij} w_{ij} \quad , \quad \sum_{j=1}^{T} w_{ij} = 1 \, , 0 \le w_{ij} \le 1 \, , \text{ i= 1,2,…,n}$$

Placing bounds for v_j is difficult in practice. Alternatively, Chebychev's inequality Pukelsheim (1994) may be used as a conservative means of specifying sets of error bounds. The empirical GME literature indicates that, in general, the number of support values R for unknown parameters is 5 and the number of support values of T of the error term is 3; see for example Paris (2001), Al-Nasser (2003), Al-Nasser (2005) and Golan et al. (1996).

Now, using the reparameterized unknowns' $\beta = ZP$ and $\varepsilon = VW$, we rewrite the general linear model as follows:

$$y = f(x_i, ZP) + VW$$

Then, maximum entropy principle may be stated in scalar summations with two nonnegative probability components, and the GME estimators can be achieved by solving the following non-linear programming problem:

$$\left.\begin{array}{l} Max \;\; H(P,W) = -P' \ln(P) - W' \ln(W) \\ \text{subject to the following constraints;} \\ \text{(i)} \;\; y = f(x, ZP) + VW \\ \text{(ii)} \quad 1_K = (I_K \otimes 1'_R)P \\ \text{(iii)} \quad 1_n = (I_n \otimes 1'_J)W \end{array}\right\} \quad (3)$$

Note that $\otimes$ is the Kronecker product, 1_K is a K-dimensional vector of ones, and $\ln(P) = (\ln(p_{11}), \ln(p_{12}), ..., \ln(p_{KR}))$. The GME system in (3) is a non-linear programming system and can be solved by applying the Lagrangian method, in which after finding the Lagrangian function, the first order conditions can be solved . For more details see Golan et al. (1996).

3.1.GME Procedure to ME Model

Without losing of generality, assume that ξ_i are fixed; i.e. $\omega_i = 0, \;\; i = 1,2,...n$. The compound model of ultrastructural model given in (1-2) can be rewritten as follows:

$$y_i = \alpha + \beta(x_i - \delta_i) + \varepsilon_i \quad , i = 1,2,...,n \qquad (4)$$

Then, by using the GME we can solve the problem after some reformulation of the unknown parameters α and β, and the unknown error terms δ_i, and ε_i, i = 1,2,3,…, n. In the compound model (4), there are two unknown parameters reparametrized as

$$\alpha = AQ \text{ ; where } 1_R^{'} Q = 1$$

where A is a row vector of size R and Q is a R-dimensional column vector of weights. The slope will be in the form

$$\beta = ZP \text{ ; where } 1_K^{'} P = 1$$

where Z is a row vector of size K and P is a K-dimensional column vector of weights. Also, there are two error terms that are reparametrized in following terms

$$\delta = V^* W^* \text{ where } \left(I_n \otimes 1_T^{'}\right) W^* = 1_n$$

$$\varepsilon = VW \text{ ; where } \left(I_n \otimes 1_J^{'}\right) W = 1_n$$

noting that V^* and V are ($n \times nT$) and ($n \times nJ$) matrices respectively, and W^* and W are nT-dimensional and nJ-dimensional vectors of weights respectively. Based on these reparametrization, the new ultrastructural model can be rewritten as

$$Y = AQ + X\left(ZP + V^* W^*\right) + VW$$

3.2. GME Solution

Generalized Maximum Entropy (GME) for the model given in (4) may be formulated as nonlinear programming (NP) system. The NP selects q, p, w^* and w ≥ 0 to maximize the joint Shannon entropy:

$$H(Q,P,W,W^*) = -Q^{'} \ln(Q) - P^{'} \ln(P) - W^{'} \ln(W) - W^{*'} \ln(W^*)$$

Subject to

$$Y = AQ + X\left(ZP + V^* W^*\right) + VW$$

$$1_R^{'} Q = 1$$

$$1_K^{'} P = 1$$

$$\left(I_n \otimes 1_T^{'}\right) W^* = 1_n$$

$$\left(I_n \otimes 1_J^{'}\right) W = 1_n$$

Here, we have $3n+2$ constraints and $R+K+n(T+J)$ unknowns. The solution of this system can be found by deriving the first order conditions of the Lagrangian function:

$$L = H(Q,P,W,W^*) + \gamma'[Y - AQ - X(ZP + V^*W^*) \cdot$$

$$- VW] + \theta_1'[1 - 1_R' Q] + \theta_2'[1 - 1_K' P]$$

$$+ \psi'[(I_n \otimes 1_T')W^* + \zeta'[(I_n \otimes 1_J')W$$

where, $\gamma' \in \Re^n, \theta_1 \in \Re^R, \theta_2 \in \Re^K, \psi \in \Re^n, and\ \zeta \in \Re^n$ are the associated vectors of Lagrangian multipliers. Taking the gradient of L to derive the first order condition and solving these conditions, the GME solution selects the most uniform distribution consistent with the information provided in the data and the add up constraints:

$$\hat{Q} = \exp(-A1_n'\hat{\gamma}) \odot \{1_R' \exp(-A1_n'\hat{\gamma})\}^{-1}$$

$$\hat{P} = \exp(-Z1_n'\hat{\gamma}(X - V^*\hat{W}^*)) \odot 1_K' \{\exp(-Z1_n'\hat{\gamma}(X - V^*\hat{W}^*))\}^{-1}$$

$$\hat{W} = \exp(-V'\hat{\gamma}) \odot \{(I_n \otimes 1_J 1_J')\exp(-V'\hat{\gamma})\}^{-1}$$

$$\hat{W}^* = \exp(-V^{*'}\hat{\gamma}1_K' Z\hat{P}) \odot \{(I_n \otimes 1_T 1_T')\exp(-V^{*'}\hat{\gamma}1_K' Z\hat{P})\}^{-1}$$

where $\odot$ is the Hadamard (element wise) product. To simplify matter the individual probabilities take the forms:

$$\hat{q}_r = \frac{\exp\left(-a_r \sum_{i=1}^{n} \hat{\gamma}_i\right)}{\sum_{r=1}^{R} \exp\left(-a_r \sum_{i=1}^{n} \hat{\gamma}_i\right)}, r = 1,2,...,R;$$

$$\hat{p}_k = \frac{\exp\left(-z_k \sum_{i=1}^{n} \hat{\gamma}_i \left(x_i - \sum_{t=1}^{T} v_t^* \hat{w}_{it}^*\right)\right)}{\sum_{k=1}^{K} \exp\left(-z_k \sum_{i=1}^{n} \hat{\gamma}_i \left(x_i - \sum_{t=1}^{T} v_t^* \hat{w}_{it}^*\right)\right)}, k = 1,2,3,...,K$$

$$\hat{w}_{ij} = \frac{\exp(-\hat{\gamma}_i v_j)}{\sum_{j=1}^{J} \exp(-\hat{\gamma}_i v_j)} \quad , i = 1,2,3,\ldots,n \quad , j = 1,2,3,\ldots,J;$$

$$\hat{w}_{it}^{*} = \frac{\exp\left(-\hat{\gamma}_i v_t^{*} \sum_{k=1}^{K} z_k \hat{p}_k\right)}{\sum_{t=1}^{T} \exp\left(-\hat{\gamma}_i v_t^{*} \sum_{k=1}^{K} z_k \hat{p}_k\right)} \quad , i = 1,2,3,\ldots,n \quad , t = 1,2,3,\ldots,T$$

Then the intercept and the slope of model (4) can be estimated as

$$\left.\begin{aligned} \hat{\alpha} &= A\hat{Q} \\ \hat{\beta} &= Z\hat{P} \end{aligned}\right\} \quad (5)$$

Lemma.1 The GME estimators given in (5) are unbiased estimators.

Proof: Simply, by taking the expected value of $\hat{\alpha}$, we have;

$$E(\hat{\alpha}) = E(\sum_r a_r \hat{p}_r) = E(a) \sum_r \hat{p}_r$$

since $\sum_r \hat{p}_r = 1$, then

$$E(a) = \sum_r a_r p_r = \alpha$$

Therefore, $\hat{\alpha}$ is an unbiased estimator of the intercept. Consequently, the estimated variance of $\hat{\alpha}$ is:

$$Var(\hat{\alpha}) = Var\left(\sum_r a_r \hat{p}_r\right) = \sum_{r=1}^{R} \hat{p}_r^2 \operatorname{var}(a_r)$$

$$= \sum \hat{p}_r^2 \left[\left(\sum a_r^2 \hat{p}_r\right) - \left(\sum a_r \hat{p}_r\right)^2\right]$$

$$= \sum \hat{p}_r^2 \left[\left(\sum a_r^2 \frac{\exp\left(-\hat{\gamma}_r a_r\right)}{\sum_{r=1}^{R} \exp\left(-\hat{\gamma}_r a_r\right)} \right) - \left(\sum a_r \frac{\exp\left(-\hat{\gamma}_r a_r\right)}{\sum_{r=1}^{R} \exp\left(-\hat{\gamma}_r a_r\right)} \right)^2 \right]$$

In similar ways we can prove that $\hat{\beta}$ is an unbiased estimator of the slope, and consequently we can derive the associated variance.

4. Monte Carlo Experiments

Monte Carlo experiments were performed to comment on the choice of the support points and number of support points of the unknown parameters and error terms in GME formulations. For fixed sample size, $n = 20$, a simulation study was carried out by generating 1000 samples according to the ultrastructural relationship $y_i = 1 + x_i + \varepsilon_i$ and $x_i = \frac{2+i}{2} + \delta_i$, $i = 1,2,...,n$. The errors terms generated independently from standard normal, i.e., $\varepsilon \sim N(0,1)$ and $\delta \sim N(0,1)$. Then, a comparison between the GME and MLE estimation methods was made in terms of bias and means squared error (*MSE*)

$$MSE(\hat{\beta}) = \frac{1}{1000} \sum_{i=1}^{1000} \left(\hat{\beta}_i - \beta\right)^2 \text{ and } Bias(\hat{\beta}) = \frac{1}{1000} \sum_{i=1}^{1000} \left(\hat{\beta}_i - \beta\right)$$

Three experiments were conducted:

Experiment 1.

The support parameter space of the error terms, V^* and V were fixed to be three in the intervals [-3S_x, 0, 3S_x] and [-3S_y, 0, 3S_y] for δ and ε; respectively. Then, this experiment was conducted for selecting the support values (a, z) and number of support values (R, K) for the unknown parameters $\alpha = AQ$ and $\beta = ZP$; i.e.,

$$\{a_i, i = 1,2,...,R\};\ \{z_j, j = 1,2,...,K\}$$

In the first part of this experiment, we fixed the number of these support values to be 3 in the intervals [-1, 0, 1] to [-500, 0, 500]. The results of this simulation study in Table.1 indicate that the best support values for both parameters should be in the interval [-100, 0, 100]. Hereafter, in the second part of this experiment, we start to increase number of support values within this interval to have 4, 5… or 7 support points and allocate them in an equidistant fashion. The results in Table.1 indicate that the greatest improvement in precision comes when

R and K were equal to 7. Moreover, it could be noted that for all choices of the parameter spaces, the GME results were more accurate and more efficient than the MLE results.

Table 1. Selecting the parameters support

Method		$\hat{\alpha}$		$\hat{\beta}$	
		Bias	*MSE*	*Bias*	*MSE*
MLE		-0.1465	0.6190	-0.0903	0.0549
GME	[-1,0,1]	-0.0906	0.0824	-0.0546	0.0323
	[-10,0,10]	-0.0765	0.0599	-0.0445	0.0296
	[-100,0,100]	-0.0579	0.0386	-0.0471	0.0311
	[-500,0,500]	-0.0647	0.0465	-0.0450	0.0294
	Increasing number of the support points				
	[-100,-50,50,100]	-0.0608	0.0405	-0.0464	0.0308
	[-100,-50,0,50,100]	-0.0744	0.0582	-0.0443	0.0291
	[-100,-50,-25,25,50,100]	-0.0588	0.0393	-0.0466	0.0307
	[-100,-50,-25,0,25,50,100]	-0.0544	0.0354	-0.0436	0.0278

Experiment 2.

To check the impact of error terms support space, and based on the experimental design outlines above, the Monte Carlo trial were repeated under the results of Experiment.1. The number of support values for each parameters A and Z were fixed to be 7 support values within the interval [-100, 0, 100]. Also, this experiment started by fixing the number of support values (J, T) to be 3, then the experiment repeated by shifting the support values V^* and V in the interval [$-hS$, 0, hS], where h = 1, 2, …, 7. Under the simulation assumptions, the results in Table.2 indicate that the greatest improvement in the precision comes from using h = 3. In similar ways of Experiment.1, the simulation is repeated by start increasing the number of support points in the interval [$-3S$, 0, $3S$] for each error term. Taking in our consideration both parameters, the greatest improvement comes when number of support points equals to 3.

Table 2. Selecting the error terms support

Method		$\hat{\alpha}$		$\hat{\beta}$	
		Bias	*MSE*	*Bias*	*MSE*
MLE		-0.1465	0.6190	-0.0903	0.0549
GME	[-S, 0, S]	-0.0540	0.0361	-0.0445	0.0282
	[-2S, 0,2 S]	-0.0534	0.0337	-0.0437	0.0279
	[-3S, 0, 3S]	-0.0544	0.0354	-0.0436	0.0278
	[-4S, 0, 4S]	-0.0557	0.0384	-0.0470	0.0311
	[-5S, 0, 5S]	-0.0714	0.0541	-0.0449	0.0296
	Increasing number of the support points				
	[-3S, -1.5 S,1.5S, 3S]	-0.0819	0.0672	-0.0397	0.0254
	[-3S, -1.5 S,0,1.5S, 3S]	-0.0771	0.0633	-0.0405	0.0260
	[-3S, -1.5 S, -0.75S, 0.75S, 1.5S, 3S]	-0.0603	0.0432	-0.0426	0.0266
	[-3S, -1.5 S, -0.75S,0, 0.75S, 1.5S, 3S]	-0.0644	0.0467	-0.0366	0.0219

Experiment 3.

Based on the previous experiments results, which related to the choice of parameter support, this experiment starts to increase the sample size; n = 20, 30, …, 100. The simulation results in Table.3 showed that the GME estimators have a lower *MSE* and lower bias for all sample sizes.

Table 3. Comparisons between GME and MLE

n	Method	$\hat{\alpha}$		$\hat{\beta}$	
		Bias	*MSE*	*Bias*	*MSE*
20	GME	-0.0544	0.0354	-0.0436	0.0278
	MLE	-0.1465	0.6190	-0.0903	0.0549
30	GME	-0.0488	0.0313	-0.0449	0.0284
	MLE	-0.1303	0.6721	-0.0896	0.0501
40	GME	-0.0396	0.0218	-0.0398	0.0240
	MLE	-0.0769	0.4213	-0.0798	0.0481
50	GME	-0.0411	0.0211	-0.0382	0.0239
	MLE	-0.0818	0.4186	-0.0764	0.0397
100	GME	-0.0375	0.0191	-0.0394	0.0234
	MLE	-0.0734	0.3381	-0.0789	0.0268

5. Empirical Example

To illustrate the computation of estimates by using the GME and MLE methods, a small data set adapted from department of statistics (2007) is used. The data given in Table.4 are number of crimes and number of unemployed collected at twelve main directorates in Jordan. The estimates of both variables contain measurement error arising from the observational and sampling error associated with the use of the register cases only (crimes or unemployed) to represent the population.

Table 4. Number of Unemployed and Number of General Crimes in Jordan by Location of Occurrence, 2007

Police Directorate	Population Size	Number of Crimes	Number of unemployed
Amman	2216000	20166	243760
Balqa	383400	2261	52526
Zarqa	852700	5649	113409
Madaba	143100	790	26617
Irbid	1018700	6296	197628
Mafraq	269000	949	43578
Jarash	171700	755	30563
Ajlun	131600	713	27504
Karak	223200	1199	29016
Tafiela	80100	405	13857
Ma'an	108800	805	21216
Aqaba	124700	1745	18082

Source: Statistical Yearbook, Amman, Jordan, 2007.

We believe that the number of crime is a linear function of the number of unemployed, and based on Figure.1 the suggested model for this data is

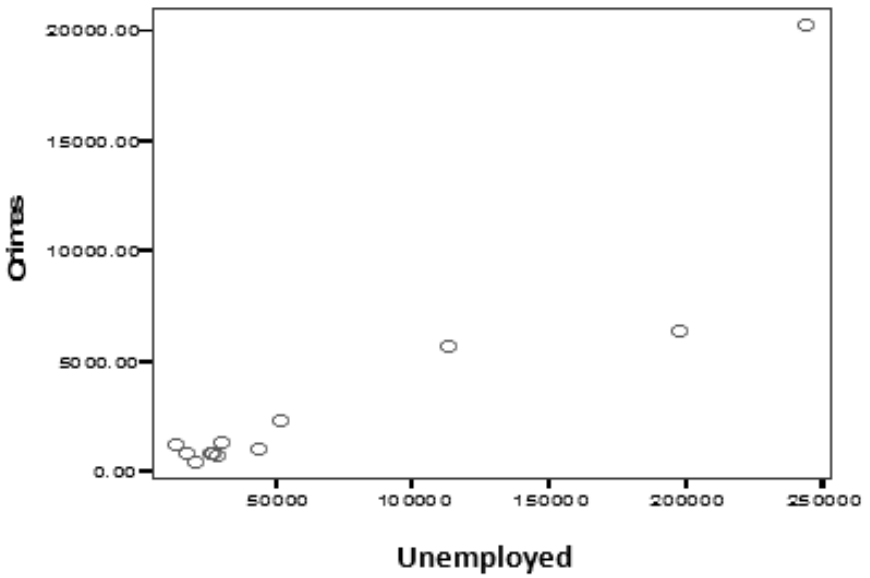
$$\eta_i = \alpha + \beta\xi_i \quad , i = 1,2,...,12$$

such that

$$x_i = \xi_i + \delta_i \,,$$
$$y_i = \eta_i + \varepsilon_i \,, \quad i = 1,2,...,12$$

where η is the number of crimes and ξ is the number of unemployed.

Figure 1. Scatter plot of number of Crimes vs Number of Unemployed



The statistics associated with the data of Table.4 are $(\bar{x}, \bar{y})$ = (68146.26, 3437.5833); (S_{xx}, S_{yy}, S_{xy}) = (5.373197E+09, 2.884094E+07, 3.557909E+08).

This model is solved by using both approaches (GME and MLE). While using the GME method we carried out all numerical calculations parallel to the theoretical development. The number of support values for each parameter were fixed to be 7 support values within the interval [-100, 0, 100]; however, the support parameter space of the error terms were fixed to be three in the intervals [-3S_x, 0, 3S_x] and [-3S_y, 0, 3S_y] for δ and ε, respectively. The results are given in Table.10

Table 5. MLE and GME estimates of the number of crimes data

Method	$\hat{\alpha}$	$\hat{\beta}$	Total Residual sum of squares (TRSS)
MLE	-1039.039	0.0663	-15.75
GME	-2656.31	0.0900	10.68

Figure 2. Comparison between the observed residuals of both methods

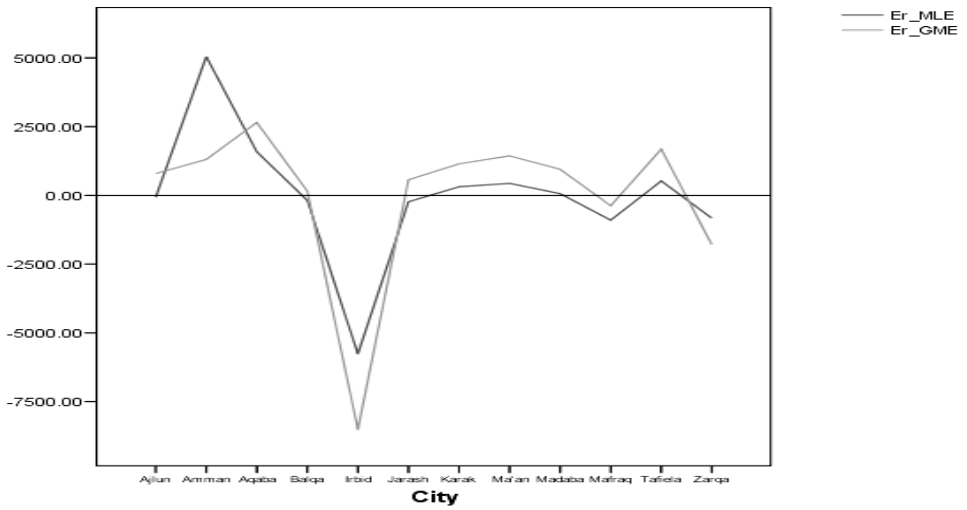


Figure.2 depicts a distinction between the observed errors by using both methods. The TRSS is -15.75 and 10.68 for the MLE and GME estimation methods, respectively; this appears to support that the GME is a good alternative estimation method to the MLE in fitting the measurement error models.

6. Concluding Remarks

This study gives the researcher a more precise method for estimating the parameters of the ultrastructural model by applying the GME estimation approach. The main idea of using GME is to improve the parameter estimation in the generalized measurement error models and to abstract the additional assumptions that are needed in the traditional MLE. In fact, all that the GME needs to be applicable can be obtained from the sample or can be specified by the researcher experiences. The Monte Carlo simulations provide a good evidence for the superiority of the GME on the MLE in terms of *MSE*. Also, the real data analysis gives similar results that confirm the robustness of GME over the MLE estimation method. Hence, the GME considered a good alternative in estimating the ultrastructural models.

REFERENCES

AL-NASSER, A. 2005. Entropy Type Estimator to Simple Linear Measurement Error Models. *Austrian Journal of Statistics*. 34(3). 283–294.

AL-NASSER, A. 2004. Estimation of Multiple Linear Functional Relationships. *Journal of Modern Applied Statistical Methods*.3 (1), 181–186.

AL-NASSER, A. 2003. Customer Satisfaction Measurement Models: Generalized Maximum Entropy Approach. *Pakistan Journal of Statistics*. 19(2), 213–226.

CARROLL, R. J., RUPPERT, D. and STEFANSKI, L. A. 1995. *Measurement Error in Nonlinear Models*. Chapman and Hall, London.

CHI-LUN CHENG and JOHN W. VAN NESS. 1999. Statistical Regression with Measurement Error. Arlond: N.Y: USA.

CSISZAR, I. 1991. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *The Annals of Statistics*, 19, 2032–2066.

Department of Statistics. 2007. Statistical Yearbook, 58th Issue, Amman, Jordan.

DOLBY, G. R., 1976, The ultra-structural model: A synthesis of the functional and structural relations, *Biometrika* 63, 39–50.

DONHO, D. L, JOHNSTONE, I.M, HOCH, J.C, and STERN A.S 1992. Maximum entropy and nearly black object. *J. Royal, Statistical Society, Ser B*, 54, 41–81.

GOLAN, A. (2008). Information and Entropy Econometrics – A Review and Synthesis. Foundations and Trends in Econometrics. 2, 1–2, 1–145.

GOLAN, A. JUDGE, G. MILLER, D.1996. *A maximum Entropy Econometrics: Robust Estimation with limited data*, Wiley, New York.

GOLAN, A. JUDGE, G. PERLOFF, J. 1997. Estimation and Inference with Censored and Ordered Multinomial Response Data. *J. Econometrics*. 79, 23–51.

GLESER, L. J .1985. A note on G. R. Dolby's unreplicated ultrastructural model. *Biometrika*, 72, 117–124.

JAYNES, E. T. 1957(a,b). Information and Statistical Mechanics (I, II). *Physics Review* (106,108), (620–630, 171–190).

QUIRINO PARIS. 2001. Multicollinearity and Maximum Entropy Estimators. *Economics Bulletin*. Vol.3, No.11, 1–9.

PEETERS, L. (2004). Estimating a random-coefficients sample-selection model using generalized maximum entropy. Economics Letters. 84: 87–92

PUKELSHEIM, F. 1994. The Three Sigma Rule. *The American Statistician*, Vol.48, no.2, 88–91.

SRIVASTAVA, A. K, SHALABH. 1997. Consistent estimation for the non-normal ultrastructural model, *Statist. Probab. Lett.* 34 .67–73.

SHANNON C, E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal.* 27, 379–423.

STATISTICS IN TRANSITION-new series, July 2010
Vol. 11, No. 1, pp. 25—45

CLUSTERING FINANCIAL DATA USING COPULA-GARCH MODEL IN AN APPLICATION FOR MAIN MARKET STOCK RETURNS

Anna Czapkiewicz[1], Beata Basiura[2]

ABSTRACT

There are many statistical techniques that allow us to find similarities among variables. Cluster analysis discovers structure within sets of data. The choice of a relevant metric is a fundamental problem in the case of clustering financial data. In this paper, the Copula–GARCH model is used to obtain the dependency parameter between time series. The dissimilarity measure based on the maximum likelihood parameter obtained from the Normal or t-Student copula is proposed and applied to classify forty two indices from American, European, and Asian stock markets.

Key words: Clustering stock indices, dependence parameter, Copula–GARCH model, Copula function, Skewed distribution

1. Introduction

The identification of similarities or dissimilarities in financial time series has become an important research area in finance and empirical economics. This problem has been widely studied in the discrimination and clustering literature. It is important to classify time series into groups with similar dependency structures. The choice of a relevant metric is a fundamental problem in clustering time series. Some studies use approaches based on the Euclidean distance. This metric has important limitations of being invariant to transformations that modify the order of observations over time and it does not allow for the correlation structure of time series. Financial time series are often characterized by volatility structures that might be represented by GARCH processes. Piccolo (1990) introduced a metric for ARIMA models based on the autoregressive representation. In this approach, time series are close when the models characterizing them are similar.

[1] Faculty of Management, AGH University of Science and Technology, Krakow, Poland, e-mail: gzrembie@cyf-kr.edu.pl.
[2] Faculty of Management, AGH University of Science and Technology, Krakow, Poland, e-mail: bbasiura@zarz.agh.edu.pl.

The distances between GARCH processed were discussed by Otranto (2004). A closeness between such processes is affirmed when the underlying conditional variance equations are similar in terms of their structure. However, this measure does not take into account the dependence structure between time series.

In this paper, we consider the problem of clustering financial time series. We propose a similarity measure based on the dependency structure between time series. To specify the behaviour of financial data, the GARCH(1,1) model is recommended (Bollerslev, 1986) and widely used. It is well known that for this class of models, the obtained residuals are generally non-normal. The evidence of heavy tails and skewness is also provided. That is why the Pearson correlation coefficient, which describes the dependency only for multivariate normal distribution, is not a good measure of dependence between financial time series. Without the multivariate normality assumption, two pairs of markets can have equal linear correlation coefficient while they can still differ in terms of dependence structure.

Embrechts et al. (2001, 2003) proposed an application of copula function to model the multivariate distribution. Using a copula approach, data modelling can be split into two steps: modelling marginal distributions, and then modelling a dependence structure between marginal distributions. Thus, in the first step we can model the financial data using a GARCH process. Next, the dependency parameter can be obtained from the copula taken as a function connecting the univariate marginal distributions and the multivariate distribution.

A variety of copula functions, already described in the literature (Nelsen, 1999), provides us with a wide range of models of dependence structure, but only some of them are appropriate for financial markets. The most popular are Normal copula and *t*-Student copula. Both Normal and *t*-Student copula functions have different characteristics in terms of tail dependence. The *t*-Student copula is recommended by several authors such as Mashal, Zevi (2002) and Breymann et al. (2003).

There are two important issues arising in describing empirical data using a copula function. Firstly, copula parameters must be accurately estimated. Secondly, the chosen copula specification should be tested. In this context, different statistical tests could be applied. Some, for example Junker and May (2005), used a chi-square test. Breymann et al. (2003) suggested the test based on the probability transform and on projection of the multidimensional problem into its univariate representation, where Anderson-Darling test statistic could be applied. Some other tests have also been discussed in the Chen et al. (2004), Genest et al. (2008)

The aim of this paper is to classify forty two indices coming from different emerging as well as developed, European, American, and Asian stock markets. The empirical data covers period from January 2002 to December 2008. In order to obtain correctly specified marginal distribution, the GARCH models with different conditional distribution of innovations are tested. Next, the dependency

structure between the GARCH processes is estimated. For relatively short time series, we assume that such a dependency is constant over the time.

The elements of the correlation matrix obtained from Normal or *t*-Student copula is proposed as the similarity measure between data. Estimating multidimensional copula would be a challenging task considering the number of data. Therefore, the bivariate copula parameters for all pairs of markets are estimated using the Maximum Likelihood (ML) method. We verify if the two-dimensional Normal or *t*-Student copula is able to capture the dependency structure between two markets.

For a clustering purposes we adapted the Ward (1963) clustering algorithm (see Mirkin (2005)) with the distance based on the estimator of the correlation parameter obtained from the Normal or *t*-Student copula.

The paper is organized as follows: In Section 2, the discussion of the univariate model is presented. Section 3 contains the copula function characteristics, description of the copulas used in the empirical studies, and the tests for models of dependence. The clustering method and the empirical results are presented in Section 4. The main results are summarized in the last section.

2. The return distribution

The advantage of using copulas, as mentioned in the introduction, stems from the fact that marginal distributions can be separated from the underlying dependency structure. Many univariate models have been proposed to describe the dynamics of returns. In this paper we investigate the univariate generalized autoregressive conditional heteroscedastic GARCH(1,1) model proposed by Bollerlav (1986). The GARCH(1,1) model is defined as follows:

$$y_t = \mu + \varepsilon_t, \qquad \varepsilon_t = \sqrt{h_t}\eta_t, \quad \eta_t \sim iid(0,1), \quad h_t = a_0 + a_1\varepsilon_{t-1}^2 + a_2 h_{t-1} \qquad (1)$$

This is the simplest and so far the most popular GARCH parameterization. With the restrictions $a_0 > 0$, $a_1, a_2 > 0$, and $a_1 + a_2 < 1$, the conditional variance process is stationary with finite second moments. In the classical GARCH model, the obtained residuals are assumed to be normally distributed. However, it is well known that in practice such residuals are far from normality. Scrutiny of daily returns led to the introduction of fat-tailed distributions for this residuals. For instance, Nelson (1991) considered the GED (generalized error distribution), while Bollerslev and Wooldridge (1992) focused on Student's *t* distributed innovations. Fat-tails are not the only problem in the context of conditional distribution, the skewness can also be noticed. With the skewness parameter being introduced, we can consider skewed *t*-Student or skewed GED distribution of the η_t (Arellano-Valle et. al. 2004).

Generally, the considered skewed distribution has the form:

$$f_{SKEW}(x;\psi,\nu)=\frac{2}{a(\psi)+b(\psi)}\left[g\left(\frac{x}{a(\psi)}\right)I_{(x<0)}+g\left(\frac{x}{b(\psi)}\right)I_{(x>0)}\right],$$

where ψ and ν denote, respectively, the asymmetry and the degrees-of-freedom parameters. The $a(\psi)$, $b(\psi)$ are some normalizing functions. The most known versions are $a(\xi)=\xi$ and $b(\xi)=\xi^{-1}$ (Fernandez, Steel, 1998) or $a(\lambda)=1-\lambda$ and $b(\lambda)=1+\lambda$ (Hansen, 1994). In our empirical research, we use the first weight. For several selected indices we used both weights, but the results were similar. The $g(\cdot)$ denotes symmetrical function distribution, it is *t*-Student or GED here. The density of Student's *t* distribution has the form:

$$f_S(x,\nu)=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{(\nu-2)\pi}}\left(1+\frac{x^2}{(\nu-2)}\right)^{\frac{\nu+1}{2}},$$

whereas the density of the GED distribution is:

$$f_G(x,\nu)=\frac{\nu\exp\left\{-\frac{1}{2}\left|\frac{x}{\lambda}\right|^{\eta}\right\}}{\lambda\Gamma\left(\frac{1}{\nu}\right)}2^{\frac{\nu+1}{\nu}},\text{ where }\lambda=\left[\frac{\Gamma\left(\frac{1}{\nu}\right)}{\Gamma\left(\frac{3}{\nu}\right)}2^{-\frac{2}{\nu}}\right]^{1/2}.$$

The maximum likelihood method is a general approach to estimate the parameters of GARCH model. Alternatively, a nonparametric approach can be followed by computing the empirical *cdf* of the data. We focus on the parametrical methods due to using chi-square test to verify the adapted model. Several restriction of the model specification are tested in the empirical section. In particular, we test within the class of GARCH models with different conditional distributions: Normal, *t*-Student, GED, skewed *t*-Student and skewed GED.

3. The Copula functions

The copula function is a multivariate distribution defined on the unit cube $[0,1]^d$, with uniformly distributed margins. The function $C:[0,1]^d\to[0,1]$ is a *d*-dimensional copula if it the satisfies the following properties:

1. For all $u_i\in[0,1]$, $C(1,..,1,u_i,1,..,1)=u_i$.
2. For all $u\in[0,1]^d, C(u_1,...,u_d)=0$, if at least one coordinate $u_i=0$.
3. *C* is *d*-increasing.

The importance of the copula function stems from the fact that it captures the dependence structure of the multivariate distribution. Let $X = (X_1,...,X_d) \in R^d$ be a random vector with a multivariate distribution $F(x_1,...,x_d) = P(X_1 \le x_1,...,X_d \le x_d)$, and continuous margins $F_n(x_n) = P(X_n \le x_n)$, $x_n \in R$, $n = 1,...,d$.

According to Sklar's theorem (Sklar, 1959), there exists a unique copula $C:[0,1]^d \to [0,1]$ of F such that for all $x \in R^d$,

$$F(x_1,...,x_d) = P(X_1 \le x_1,...,X_d \le x_d) = C\big(F_1(x_1),...,F_d(x_d)\big).$$

A fundamental conclusion of this theorem states that in multivariate continuous distribution functions, the dependence structure and the margins can be separated, and the dependence structure can be represented by the copula. This property allows us to fit the marginal distribution of each series of data separately, and to model the dependence structure between different data. Furthermore, the copulas are invariant under strictly increasing transformations of the variables. This property implies that the dependence structure between the variables is captured by the copula, regardless of the margins.

The most important copula in finance is the Normal copula:

$$C(u_1,...,u_d) = \Phi_\Sigma(\Phi^{-1}(u_1),...,\Phi^{-1}(u_d)),$$

where Φ is a cumulative normal distribution function and, Φ_Σ is a d-dimensional cumulative normal distribution function with correlation matrix Σ. The bivariate Normal copula is defined by:

$$C(u_1,u_2;\rho) = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{s^2 - 2\rho st + t^2}{2(1-\rho^2)}\right) dsdt, \quad (2)$$

and ρ denotes the correlation coefficient.

Another important copula in finance is the t-Student copula:

$$C\ (u_1,...,u_d) = t_{\Sigma,\eta}(t_\eta^{-1}(u_1),...,t_\eta^{-1}(u_d)),$$

where t_η is the Student's t cumulative distribution with η degrees of freedom and $t_{\Sigma,\eta}$ is the Student's t cumulative distribution with η degrees of freedom and the correlation matrix Σ. The bivariate case of t-Student is given by:

$$C(u_1,u_2;\rho) = \int_{-\infty}^{t_\eta^{-1}(u_1)} \int_{-\infty}^{t_\eta^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho^2}} \left(1 + \frac{s^2 - 2\rho st + t^2}{\eta(1-\rho^2)}\right)^{-\frac{\eta+2}{2}} dsdt, \quad (3)$$

where ρ is the correlation ratio.

In our empirical work, we focus on the copulas defined above. Both copulas have different characteristics in terms of tail dependence. The Normal copula has no tail dependence. The *t*-Student has symmetric tail dependence: large joint positive realizations have the same probability of occurrence as large negative realizations. Such a difference has important consequences to the modelling of the dependency parameter. In the next research we take up this subject.

3.1. Estimation of copula model

The theory of copulas shows that any joint distribution can be decomposed into its univariate marginal distributions and a copula function, which describes the dependency between the variables. Let $X = (X_1, X_2, ..., X_d)$ be random variables of interest, where X_i has parametric cumulative distribution function $F_i(x_i;\alpha_i)$ and corresponding densities denoted by $f_i(x_i;\alpha_i)$. The parameters of the marginal distribution are $\theta_1 = \left(\alpha_1^T, ..., \alpha_d^T\right)^T$. Let $F(x_1,...,x_d;\theta) = C\left(F_1(x_1,\alpha_1),...,F_d(x_d,\alpha_d);\theta_2\right)$, where C is a copula distribution function with parameters θ_2, and let $\theta = \left(\theta_1^T, \theta_2^T\right)^T$. Given a marginal distributions, different multivariate distributions can be formed simply by applying different copula functions. The density is given by

$$f\left(x_1,...,x_d;\theta\right) = c(F_1(x_1,\alpha_1),...,F_d(x_d,\alpha_d);\theta_2) f_1(x_1,\alpha_1)...f_d(x_d,\alpha_d).$$

The log-likelihood function of a sample $x_t = (x_{t1}, x_{t2}, ..., x_{td})$ is

$$l(\theta) = \sum_{t=1}^{T} \ln c\left(F_1(x_{t1};\alpha_1),...,F_d(x_{t2};\alpha_d);\theta_2\right) + \sum_{t=1}^{T}\sum_{i=1}^{d} \ln f_i(x_{ti};\alpha_i).$$

The log-likelihood function can be decomposed as: $l(\theta) = l_c(\theta_1,\theta_2) + l_m(\theta_1)$ – this decomposition could not hold if some parameters are common across the margins and the copula (Patton, 2006). The ML estimation involves maximizing the log-likelihood function with respect to all the parameters θ simultaneously. Because of complicated form of the likelihood function, it is difficult to accomplish this goal. That is why some simplifications were introduced. In the empirical work we consider IFM strategy (Shih, Louis, 1995; Joe, Xu, 1996). IFM proceeds in two steps. Firstly, $\hat{\theta}_1$ parameter estimators of margins distribution were estimated. Secondly, the estimator of copula dependency parameter was obtained as $\hat{\theta}_2 = \text{argmax } l_c\left(\hat{\theta}_1,\theta_2\right)$. Under some regularity conditions, Patton (2006) shows that the IFM procedure yields consistent and asymptotically normal

estimates. However, there is some loss of efficiency in estimation because Step 1 ignores the dependence between margins. The other algorithms were also proposed by Lui, Luger (2009).

3.2. Testing procedure

The maximum likelihood value or other measures derived from the likelihood method, such as the Akaike (AIC) or Schwarz, or Bayesian information criterion are currently used in model selection. However, those criteria are not sufficient to accept a copula as being enough representative. To show that the considered copulas are representative enough, the testing of goodness of fit is required. Generally, the goodness-of-fit testing is equivalent to test the hypothesis

$$H_0 : C(u_1,...,u_d) = C(u_1,...,u_d;\theta) \text{ for } \theta \in \Theta ,$$

i.e., the unknown copula $C(u_1,...,u_d)$ is a member of the parametric family, against the alternative hypothesis

$$H_1 : C(u_1,...,u_d) \neq C(u_1,...,u_d;\theta) \text{ for } \theta \in \Theta$$

i.e., the unknown copula is not a member of the parametric family.

3.2.1. A test based on Rosenblatt's transform

This test is based on the probability integral transformation due to Rosenblatt(1952). Let $C_j(u_1,...u_j,\theta)$ denote the joint distribution function of $U_1,...,,U_j$ under H_0. It is given by $C_j(u_1,...u_j,\theta) = C(u_1,...,u_j,1,...,1;\theta)$. In addition, let $C_j(u_j;\theta \,|U_1,...,U_{j-1})$ denote the conditional distribution function of U_j given $(U_{j-1},...,U_1)$ under H_0. It can be derived via:

$$C_j(u_j;\theta \,|U_1,...,U_{j-1}) = \frac{\partial^{j-1} C_j(u_1,...u_j,\theta)}{\partial u_1 ... \partial u_{j-1}} / \frac{\partial^{j-1} C_{j-1}(u_1,...,u_{j-1};\theta)}{\partial u_1 ... \partial u_{j-1}}.$$

Define $Z_1 = U_1$ and $Z_j = C_j(U_j;\theta \,|U_1,...,U_{j-1})$ for $j = 2,...,d$.

For the two-dimensional case,

$$Z_1 = U \text{ , } Z_2 = C_2(v;\theta|U),$$

where

$$C_2(v;\theta|U) = \frac{\partial C\ (u,v;\theta)}{\partial u}.$$

Since the copula function is a multivariate distribution function, the H_0 holds if and only if the probability integral transformed random variables $Z_1,...,Z_d$ are independent and identically distributed as a Uniform $[0, 1]$ (Rosenblatt, 1952).

Exploiting that fact, the random variable $W = \sum_{j=1}^{d} \left[\Phi^{-1}(Z_j)\right]^2$ should follow a χ_d^2 distribution. To test this distribution, Breymann et al. (2003) applied the Anderson-Darling statistic. Since W is not observable, the pseudo observations $\hat{W}_t$ are taken into consideration:

$$\hat{W}_t = \sum_{j=1}^{d} \left[\Phi^{-1}(Z_{ij})\right]^2,\ i = 1,...,n.$$

For more details on this testing procedure see Genest and Remillard (2008), Genest et.al. (2009), Kojdanovic and Yan (2009).

3.2.2. Chi-square test

The chi-square test can be applied only in the case when known margins are estimated using parametric methods. Each marginal distribution model must be accepted which requires some previous estimation and validation procedures. Furthermore, the observations are assumed to be independent and identically distributed. Otherwise, the chi-square test does not hold. The procedure for testing bivariate copulas is as follows. Let A_{ij} be bins of $[0, 1]^2$ (the *y*-axis is divided into *r* parts and the *x*-axis is divided into *s* parts), O_{ij} be the observed frequency for bin A_{ij}, and let $E_{ij}(\hat{\theta}) = n \cdot p_{ij}(\hat{\theta})$ be the expected frequency for A_{ij}, where $p_{ij}(\hat{\theta}) = \iint_{A_{ij}} dC(u,v;\hat{\theta})$.

The chi-square statistic is calculated by formula:

$$\chi^2(\hat{\theta}) = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(O_{ij} - E_{ij}(\hat{\theta})\right)^2}{E_{ij}(\hat{\theta})},$$

where $\hat{\theta}$ is the vector of estimated parameters. The test statistic follows, approximately, a chi-square distribution with $(rs-1-d)$ degrees of freedom. This analysis is sensitive to the selection of the number of bins (Roch, Alegree (2006)). In the empirical study, we used two partitions. Firstly, the unit square is divided into bins of equal area, being $r=s=10$. Secondly, the unit square was separated into eight squares, as it was suggested by Patton (2003). This second partition is useful for expressing the tail dependences.

4. Empirical results

4.1. The data

We investigate the relationships between forty two selected stock indices. Table 1 presents the variables under consideration.

Table 1. The indices from forty two countries

	Country	Name	Code		Country	Name	Code
1	Argentine	MERVAL	ME	22	Japan	NIKKEI	NI
2	Australia	AS30	AS	23	Latvia	RGSE	RG
3	Austria	ATX	ATX	24	Malaysia	KLSE	KL
4	Belgium	BEL20	BE	25	Mexico	IPC	IPC
5	Brazil	BOVESPA	BO	26	New Zealand	NZSE50	NZ
6	Bulgarian	SOFIX	SO	27	Norway	OSE	OS
7	Canada	TSE300	TS	28	the Philippines	PSE	PS
8	Chile	IPSA	IP	29	Poland	WIG	WI
9	China	SHANGHAI	SH	30	Romania	BET	BET
10	Czech Republic	PX50	PX	31	Russia	RTS	RT
11	Estonia	TLSE	TL	32	Singapore	STI	ST
12	Finland	HEX	HE	33	Slovakia	SAX	SA
13	France	CAC40	CA	34	South Korea	KOSPI	KO
14	Germany	DAX	DA	35	Spain	IBEX	IB
15	Greece	ATGI	AT	36	Switzerland	SMI Swiss	SM
16	Nederland	AEX	AE	37	Taiwan	TWSE	TW
17	Hong Kong	HSI	HS	38	Thailand	SET	SE
18	Hungary	BUX	BU	39	Turkey	ISE100	IS
19	India	BSE30	BS	40	the UK	FTSE250	FT
20	Indonesia	JKSE	JK	41	the USA	DJIA	DJ
21	Italia	MIB30	MI	42	the USA	NASDAQ	NA

The sample consists of daily frequency data and covers the period from January 2002 to December 2008. Missing data were filled by linear interpolation from the preceding to the following missing quotations. The return of indices is defined as $r_t = \ln P_t / P_{t-1}$ where P_t (t =1, …, T) is an adjusted index value at period t. In our empirical work we used *R-project* program.

4.2. Estimation of marginal model

In a preliminary step of our empirical work, we investigate the structure of the univariate marginal returns. We consider several restrictions as a possible candidates for adjusting the empirical return distribution. The model GARCH(1, 1) with symmetrical and skewed conditional distributions is specified. The unknown parameters of this specification are estimated using the ML method. We have selected one of discussed conditional distributions that fits best to the data in terms of the values of the log-likelihood. The analysis of the log-likelihood or the Akaike criterion implies that skewed distributions fit better to the data than symmetrical ones. Thus, the procedure of testing the goodness-of-fit is carried out only for skewed Student's t distribution and skewed GED distribution. For the testing purposes, we follow the procedure described in Diebold et al. (1998). If a marginal distribution is correctly specified, the margin u_{it} denoting the transformed GARCH(1,1) standardized residuals should be *iid* Uniform$[0, 1]$. The test is performed in two steps. Firstly, we evaluate whether u_{it} is serially independent. Doing so, we separately examine the serial correlation of $(u_{it} - \overline{u}_i)^k$, for $k = 1,...,4$, on 20 own lags. The k-th test statistic is defined as $R(k) = (T-20)R^2$, where R^2 is the coefficient of determination of the regression, and is distributed as a χ^2_{20} under the null hypothesis. Secondly, we test the null hypothesis that u_{it} is Uniform$[0, 1]$. For the testing purposes, we divide the empirical and theoretical distributions into N bins and then we test whether the two distributions significantly differ at each bin. We consider the case with N=10 and we adapt the chi-square test. Table 2 reports goodness-of-fit statistics. The first four columns contain p-values, labeled $p(k)$, of $R(k)$ under the null hypothesis of no serial correlation of the k-th centered moments of the u_{it}. Next column reports p-values for the chi-square test statistics under the null hypothesis that *cdf* of residuals is Uniform [0, 1]. Last column presents the maximum likelihood values for selected conditional distributions.

Table 2. The p-values for test specification and maximum likelihood values

State	*p*(1)	*p*(2)	*p*(3)	*p*(4)	*p*	Max likelihood
Argentina	0.741	0.984	0.509	0.981	0.528	4534.17
Australia	0.880	0.195	0.780	0.379	0.020	5938.64
Austria	0.568	0.949	0.360	0.978	0.056	5343.64
Belgium	0.975	0.155	0.569	0.351	0.001	5448.18
Canada	0.709	0.399	0.349	0.417	0.702	5720.46
China	0.000	0.183	0.000	0.171	0.060	4698.27
Czech R	0.258	0.993	0.057	0.972	0.002	5228.22
Estonia	0.000	0.001	0.000	0.001	0.071	5805.34
Finland	0.332	0.371	0.435	0.273	0.253	4945.83
France	0.029	0.011	0.306	0.022	0.125	5187.11
Germany	0.419	0.101	0.507	0.308	0.932	5070.78
Greece	0.000	0.135	0.000	0.257	0.769	5242.24
Netherlands	0.134	0.045	0.579	0.063	0.411	5187.72
Hong Kong *	0.094	0.090	0.072	0.250	0.503	5170.27
Hungary	0.470	0.935	0.037	0.817	0.031	4922.01
Italy *	0.695	0.409	0.660	0.640	0.382	5471.54
Japan *	0.097	0.589	0.062	0.411	0.186	5018.13
Mexico	0.087	0.880	0.113	0.656	0.044	5154.21
Norway	0.864	0.579	0.420	0.554	0.052	5110.74
Poland*	0.135	0.005	0.025	0.007	0.976	5167.47
Russia	0.001	0.582	0.000	0.889	0.608	4555.85
Slovakia	0.000	0.000	0.000	0.000	0.722	5546.35
South Korea *	0.144	0.344	0.060	0.542	0.813	4850.86
Spain	0.667	0.573	0.744	0.547	0.149	5332.16
Switzerland	0.188	0.035	0.289	0.144	0.071	5405.75
Taiwan	0.008	0.075	0.003	0.018	0.105	5013.62
Turkey	0.165	0.513	0.056	0.306	0.001	4286.35
UK	0.002	0.011	0.023	0.250	0.069	5682.96
USA NASDAQ *	0.710	0.081	0.929	0.127	0.433	5092.67
USA DJIA *	0.265	0.049	0.057	0.972	0.077	5601.52

** GARCH with skewed GED conditional distribution*
Note: p(k) - p-values for the null hypothesis of no serial correlation of the k-th centered moments; p are p-values for the chi-square test statistics for the null hypothesis that cdf of residuals is Uniform [0, 1];

The skewed *t*-Student is the most frequently chosen conditional distribution. Only in the case of a few indices, the skewed GED fits better. We present all indices with at last one searching conditional distribution not being fulfilled at a

1% significance level. The results for data satisfying both conditions, *iid* and uniformity, are written in italics. We can notice that specifications of the margins are quite proper for the well developed countries. For other indices, it is difficult to fix a proper specification. The model is not correctly specified for Romania, Brazil, India, Chile, Indonesia, Malaysia, Latvia, Thailand, Bulgaria, Singapore, the Philippines. For these indices we failed to obtain the independence as well as to select the proper conditional distribution. For Belgium, Turkey and Czech Republic the specification of GARCH(1, 1) is correct, but the null hypothesis of uniformity were rejected. For Poland, Grace, the UK, Taiwan, Russia, Slovakia, China, Estonia, the GARCH(1,1) is not sufficient enough, whereas the conditional distributions were selected properly.

4.3. Estimation of the multivariate model

In the empirical work, we consider the case of constant parameter dependency with respect to time. This assumption is justified by relatively short sample size. It is impossible to estimate multidimensional copula function due to the fact that the number of data to be investigated would be large. Therefore, the bivariate copula parameters for all market pairs are estimated. The combination of the data vectors yields 861 pairs. For almost all market pairs, the estimates of the dependency parameter for the Normal and *t*-Student copula is found to be positive and strongly significant. Strong dependency between European pairs is observed. Table 3 reports some of the parameters estimated.

Table 3. Parameter estimates of the Normal copula and *t*-Student copula

	Canada	Brazil	France	Germany	Nederland	Italy	Japan	South Korea	Switzerland	USA (DJIA)
Canada		0.43	0.48	0.63	0.49	0.49	0.22	0.24	0.40	0.63
Brazil	0.44		0.45	0.43	0.46	0.45	0.32	0.36	0.41	0.32
France	0.49	0.46		0.92	0.93	0.90	0.37	0.33	0.85	0.51
Germany	0.63	0.44	0.92		0.89	0.87	0.33	0.32	0.82	0.55
Nederland	0.49	0.47	0.93	0.89		0.86	0.37	0.33	0.83	0.51
Italy	0.50	0.45	0.90	0.87	0.86		0.29	0.25	0.79	0.50
Japan	0.23	0.33	0.38	0.34	0.38	0.31		0.63	0.37	0.14
South Korea	0.25	0.37	0.33	0.32	0.34	0.30	0.64		0.34	0.14
Switzerland	0.42	0.42	0.85	0.82	0.83	0.80	0.38	0.35		0.43
USA (DJIA)	0.63	0.32	0.53	0.56	0.52	0.51	0.14	0.14	0.45	

Note: This table supply estimated parameters of dependency; for that parameters all p – values are smaller than 0.0001 Estimates for the Normal copula are reported below the main diagonal, while for the t-Student copula above it.

To investigate whether a given copula function is able to capture the dependence structure present in the data, we perform two goodness-of-fit tests. It is known, that chi-square test requires proper specifications of the margins. That's the reason tests were made only for pairs formed from correctly specified margins. The tests were carried out for Normal copula and *t*-Student copula. Table 4 presents the results. To use the chi-square, the unit square $[0, 1]^2$ is divided into 100 bins of equal area.

Table 4. The percentage of cases where we do not to reject the null at a 1% significance level

Copula function	Test based on the probability transform	Chi-square test
Normal	70	74
t-Student	75	80

Note: Only for pairs formed from correctly specified margins.

In both tests, the *t*-Student copula better describes the dependence structure between correctly specified margins. The results indicates that the null hypothesis is not rejected only in approximately eighty percent of cases, which is not a satisfactory result. Hence, we cannot directly assess the adequacy of the models. Thus, we now turn to the hit test suggested by Patton (2003). This test is based on the chi-square statistic associated with a partition of the unit square into eight zones. Instead of the total chi-square value, we examine the contribution of each particular region to the global test statistic to find whether this copulas are unveiled. Such evidence were presented in the Canela, Collazo (2006). In Table 5, we document the average of the contributions of each region to the total value.

Table 5. The average of the contribution to the chi-square value (%)

	Regions							
Copula	1	2	3	4	5	6	7	8
Normal	12	16	13	6	3	21	23	5
t-Student	7	40	10	7	8	8	6	13

Note: Zones: the first zone is a square at lower left top (0, 0) and upper right top (0.1, 0.1); the second square has lower left top (0.9, 0.9) and upper right top (1, 1); the third square has tops (0.1, 0.1); (0.25, 0.25); the fourth - (0.75, 0.75); (0.9, 0.9); the most important is the fifth region - a big square with tops (0.25, 0.25); (0.75, 0.75); the sixth has lower left (0, 0.75) and upper right (0.25, 1); the seventh is a square with lower left top (0, 0.75) and upper right top (1, 0.25); the eight region is the rest area.

Only pairs formed from correctly specified margins; the region 5 determine only 8% of the whole chi-square value (*t*-Student copula), in contrast to region 2 which determine 40% of the whole chi-square value (the same copula).

The region 2 is the worst one for the *t*-Student copula. The other regions are clarified well enough. On the basis of the above results, we come to the conclusion that rejection of null hypothesis is caused by predictably bad performance in the upper-right quadrant. The rejected copulas do not therefore describe upper tail dependence.

5. The cluster analysis

In our study, we use Ward's clustering algorithm. Given a dissimilarity matrix $D = (d_{ij})$, the Ward algorithm can be formulated as follows. In an initial setting, all the entities are considered as singleton clusters. Assuming that there is N elements to classify, begin with N clusters consisting exactly of one entity, search the similarity matrix, reduce the number of clusters by one through merging the most similar pair of clusters. Perform those steps until all clusters are merged. The Ward is different from all other methods because it uses an analysis of variance approach to evaluate the distances between clusters. In short, this method attempts to minimize the Sum of Squares of any two hypothetical clusters that can be formed at each step.

For the choice of relevant metric, some authors used the Pearson correlation coefficient as a similarity measure of a pair of time series. The square of this distance is a popular and widely used dissimilarity index between standardized variables (Krzanowski, 1988; Kaufman, Rousseeuw, 1990) but requires the assumption of the normal margins. Although this metric can be useful to ascertain the structure of stock returns movements, it could not be used in the case of non-elliptical distributions.

As a dissimilarity measure between y_i and y_j time series we propose to take, instead of Pearson correlation, the dependency parameter ρ_{ij} from Normal (2) or *t*-Student (3) copula. (In the case of the multivariate normal distribution, the proposed ρ_{ij} distance is the same as the distance based on the Pearson correlation). For y_i and y_j specified by equation (1), the $\hat{\rho}_{ij}$ parameter is computed via MLM for $u_1 = F_i\left(\varepsilon_{ti} / \sqrt{h_{ti}}; \alpha_i\right)$ and $u_2 = F_j\left(\varepsilon_{tj} / \sqrt{h_{tj}}; \alpha_j\right)$. For the clustering purposes we considered the dissimilarity index between two time series as $1 - \hat{\rho}_{ij}$.

Fig. 1 presents the dendogram obtained with the distance discussed above based on the parameter obtained from Normal copula, whereas Fig. 2 shows the results for the distance based on the parameter obtained from the *t*-Student copula. We can notice that the cluster groups slightly differ.

Figure 1. The dendrogram for Normal copula

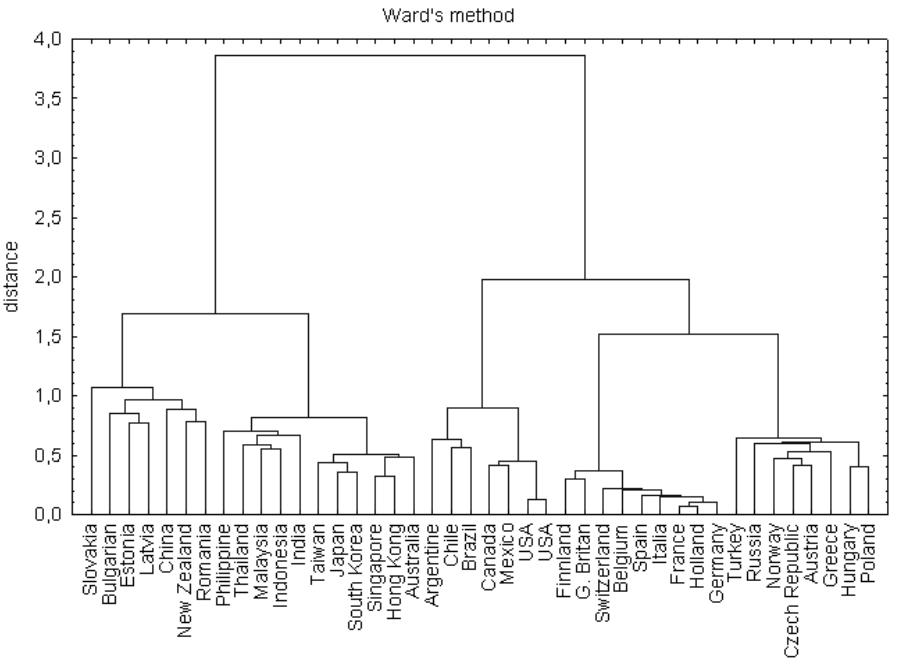
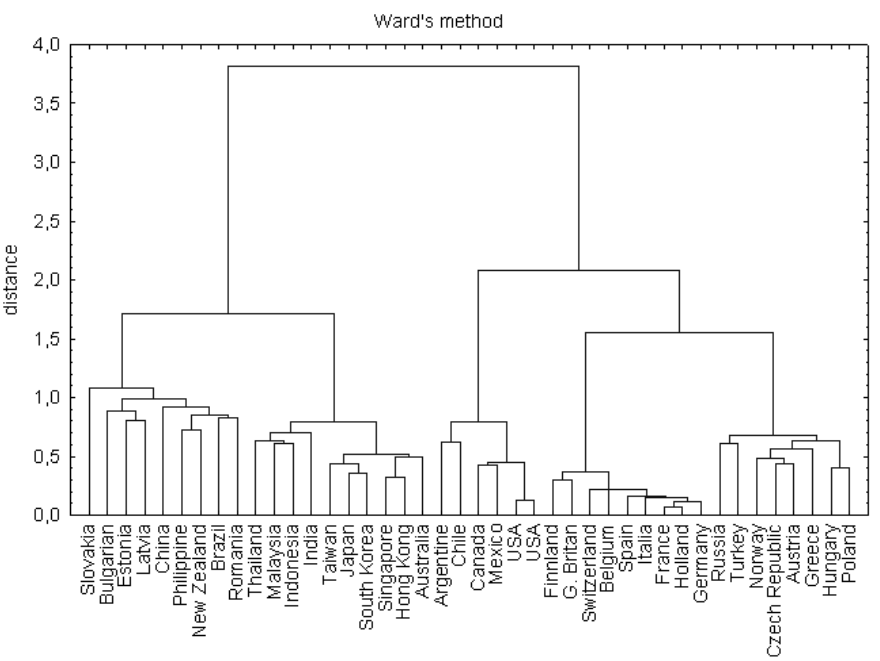


Figure 2. The dendrogram for *t*-Student copula

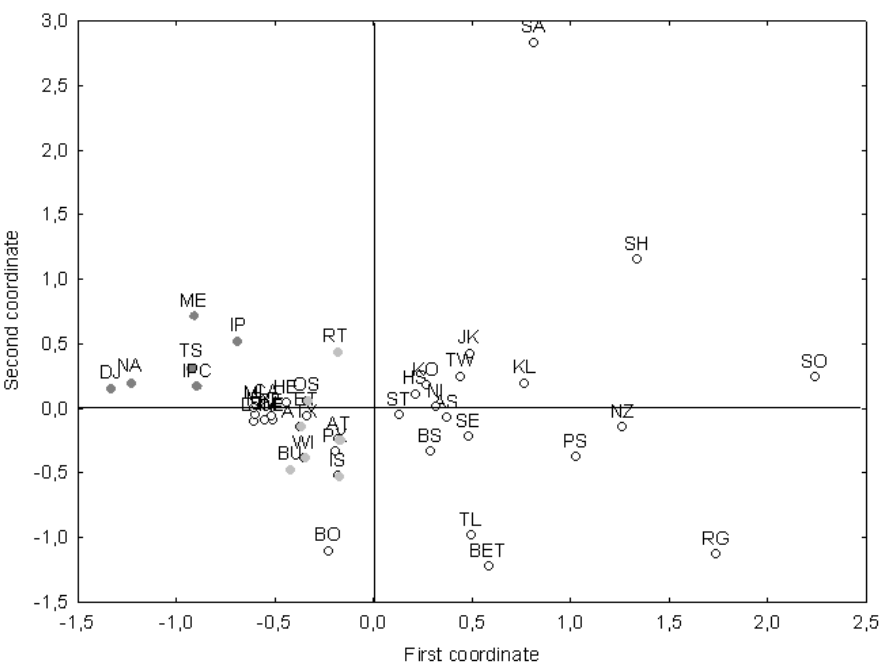


We obtained five groups. First one includes indices from West Europe: Switzerland, Belgium, Spain, Italy, France, the Netherlands, Germany. Finland and the UK also belong to this group. The second group consists of three small subgroups. It is formed by Poland, Hungary, Norway and Czech Republic, Austria and Russia, Turkey. These group includes some countries from East Europe. The third group involves American countries. Next group merges three subgroups of indices from Asia: Thailand, Malaysia, Indonesia, India and Taiwan, Japan, South Korea and the last: Singapore, Hong Kong, Australia. Finally, we have a group of outlier countries with weak connection with another ones. It includes: China, Slovakia, Bulgaria, Estonia, Latvia, the Philippines, New Zealand, Brazil and Romania.

The results of goodness-of-fit testing indicate that for properly specified margins the conclusions should be drawn from Fig. 2. However, as can be seen from the figures, for properly specified margins both copulas gave identical clustering results.

In order to summarize and better interpret the results, the multidimensional scaling map is presented. Fig. 3. presents the map for all discussed indices.

Figure 3. The multidimensional scaling map for all indices



The cluster groups are confirmed by the scaling map results. We can notice the indices concentrations according to the geographical regions, especially very dense concentration for West Europe and Asia indices. The outliers are noticeable

too. The most irrelevant indices are Slovakian (SA), Latvian (RG), Chinese (SH) and Bulgarian (SO).

Figure 4. The multidimensional scaling map for GARCH indices. First distance is based on the dependency structure, second is based on the fitted autoregressive expansions.

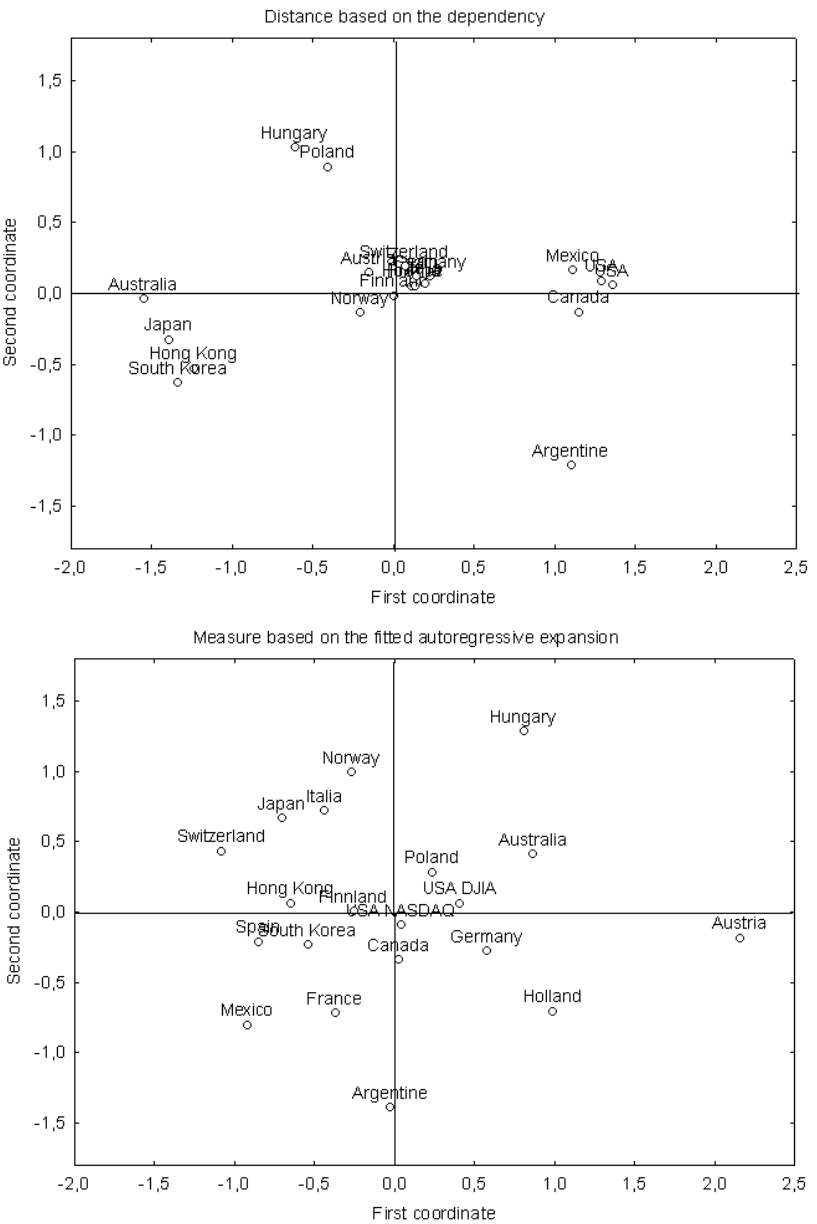


Fig. 4. presents the scaling map for the well specified model GARCH(1,1). The measure discussed in this paper and based on the fitted autoregressive expansions suggested by Otranto (2004) is taken as a dissimilarity measure. There are presented clustered groups in the case of the method discussed in this paper in contrary to dispersed map of indices where the distance is based on the fitted autoregressive expansions.

6. Conclusions

The aim of the work was to separate, by means of cluster analysis, forty two indices returns coming from European, American, and Asian stock markets, into similar groups. Modelling the dependency between stock market returns is a difficult task when returns follow a complicated dynamics. When returns are non-normal, it is often simply impossible to specify the multivariate distribution relating two or more return series. To cope with this problem, the Copula-Garch model is adapted. Firstly, the specification of GARCH(1,1) model for indices returns was verified. We have found this specification to be proper for developed markets, whereas it failed for emerging markets. Secondly, the hypothesis that the Normal and *t*-Student copulas are able to capture the dependency structure of the markets were tested. In the case of *t*-Student copula, this hypothesis was rejected approximately in 30 percent of all pairs, and approximately in 20 percent in the case of properly specified margins. These rejections were due to bad performance in the model upper tail dependence.

For the clustering purposes, we have adapted Ward's algorithm. The similarity measure as well as the dissimilarity index were based on the parameter from Normal or *t*-Student copula obtained by the maximum likelihood method. We have received five clustered groups tied to geographical regions: West, East Europe, American regions and Asia, which confirms our expectations. We have also received the group of countries weakly connected between them and the other markets.

Researches were prepared without taking into consideration area of time. The time dependences were reflected in groups. The lags behind for time series will be included in the next study. We also intend to take into consideration the tail dependence and dynamic copulas for long time series in the future. For the specifications of the margins, the GARCH model in which the first four moments are conditional and time varying, will also be discussed.

REFERENCES

ARELLANO-VALLE, R., and GÓMEZ, H., and QUINTANA, F., 2004, A New Class of Skew-Normal Distributions, *Communications in Statistic*s, Series A, 33(7), 1465–1480.

BOLLERSLEV T., 1986, Generalized Autoregressive Conditional Heteroskedasticity, *Journal of Econometrics*, 31, 307–327.

BOLLERSLEV T., WOOLDRIDGE J.M., 1992, Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models with Time-Varying Covariance's, *Econometric Reviews*, 11, 143–172.

BREYMANN W., DIAS A., EMBRECHTS P., 2003, Dependence Structures for Multivariate High-Frequency Data in Finance, *Quantitative Finance*, 3, 1–14.

CAIADO J., CRANTO N., PENA D., 2006, A periodogram-based metric for time series classification, *Computation Statistic & Data Analysis* v50 i10, 2668–2684.

CANELA M.A., COLLAZO E. P., 2005, *Modeling Dependence in Latin American Markets Using Copula Function*, Working Paper.

CHEN X., FAN Y., PATTON, A.J., 2004, *Simple Tests for Models of Dependence Between Multiple Financial Time Series, with Applications to U.S. Equity Returns and Exchange Rates*, FMG Discussion Papers, Financial Markets Group.

DIEBOLD F.X., GUNTHER T.A., TAY A.S., 1998, Evaluating Density Forecasts with Applications to Financial Risk Management, *International Economic Review*, 39(4), 863–883.

EMBRECHTS P., LINDSKOG F., MCNEIL A., 2001a, *Modeling Dependence with Copulas and Applications to Risk Management*, ETHZ, Working Paper.

EMBREECHT P., MCNEIL A.J., STRAUMANN D., 2001b, *Correlation and dependency in risk management: properties and pitfalls*. [In:] M. Dempster, H. Moffant, Risk Management, Cambridge University Press , New York, 176–223.

EMBRECHTS P., LINDSKOG F., MCNEIL A., 2003, *Modeling Dependence with Copulas and Applications to Risk Management*, In: Rachev, S.T. (Ed.), Handbook of Heavy Tailed Distributions in Finance, Elsevier/Noth-Holland, Amsterdam.

FERNANDEZ C., STEEL M., 1998, On Beyesian Modeling of Fat Tails and Skewness, *Journal of the American Statistical Association*, 93, 359–371.

EMBRECHTS P., LINDSKOG F., MCNEIL A., 2001a, Modeling *Dependence with Copulas and Applications to Risk Management*, ETHZ, Working Paper.

EMBREECHT P., MCNEIL A.J., STRAUMANN D., 2001b, *Correlation and dependency in risk management: properties and pitfalls*. [In:] M. Dempster, H. Moffant, Risk Management, Cambridge University Press , New York, 176–223.

EMBRECHTS P., LINDSKOG F., MCNEIL A., 2003, Modeling Dependence with Copulas and Applications to Risk Management, [In:] Rachev, S.T. (Ed.), *Handbook of Heavy Tailed Distributions in Finance*, Elsevier/Noth-Holland, Amsterdam.

FERNANDEZ C., STEEL M., 1998, On Beyesian Modeling of Fat Tails and Skewness, *Journal of the American Statistical Association*, 93, 359–371.

GENEST R., REMILLARD B., 2008, Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models, *Annales de l'Institut Henri Poincare: Probabilites et Statistiques*, 44, 1096–1127.

GENEST R., REMILLARD B., BEAUDOIN D., 2009, Goodness-of-fit tests for copulas: A review and a power study, *Insurance: Mathematics and Economics*, 44, 199–214.

HANSEN B., 1994, Autoregressive Conditional Density Estimation, *International Economic Review*, v.35, no. 3,705–730.

JOE H., XU J.J., 1996, *The estimation method of inference function for margins for multivariate models*, Technical Report, Departments of Statistics, University of British Columbia.

JUNKER M., MAY A., 2005, Measurement of aggregate risk with copulas, *Econometrics Journal, Royal Economic Society*, 8(3): 428–454, December.

KAUFMAN L., ROUSSEEUW P., 1990, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York.

KOJADINOVIC I., YAN J., 2009a, *Fast large-sample goodness-of-fit tests for copulas*, Submitted.

KOJADINOVIC I., YAN J., 2009b, A goodness-of-fit test for multivariate multiparameter copulas based on multiplier central limit theorems, *Statistics and Computing*,. In press.

KRZANOWSKI W., LAI Y., 1988, A Criterion For Determining the Number of Groups In Data Set Using Sum of Squares Clustering, *Biometrics* 44(1), 23–34.

LUI Y., LUGER R., 2009, Efficient estimation of copula-GARCH models 1, *Computational Statistics & Data Analysis*, 53, 2284–2297

MASHAL R., ZEEVI A., 2002, *Beyond Correlation: Extreme co-movements Between Financial Assets*, Mimeo, Columbia Graduate School of Business.

MIRKIN B., 2005, *Clustering for Data Mining: A Data Recovery Approach*, Boca Raton Fl., Chapman and Hall/CRC.

NELSEN B. R., 1999, *An Introduction to Copulas*, Springer Verlag, New York.

NELSON D.B., 1991, Conditional Heteroskedasticity in Asset Returns: A New Approach, *Econometrica*, 59(2), 397–370.

OTRANTO E., 2004, Classifying the Markets Volatility with ARMA Distance Measures, *Quaderni di Statistica*, 6,1–19.

PATTON A.J., 2003, Modeling Asymmetric Exchange Rate Dependence, Working paper, University of California, San Diego.

PATTON A.J., 2004. On the Out-of-Sample Importance of Skewness and Asymmetric Dependence for Asset Allocation, *Journal of Financial Econometrics*, Oxford University Press, 2(1), 130–168.

PATTON A.J., 2006, Estimation of multivariate models for time series of possibly different lengths, *Journal of Applied Econometrics*, John Wiley & Sons, Ltd., 21(2), 147–173.

Piccolo, 1990, A distance measure for classifying ARIMA models, *Journal of Time Series Analysis* v11, 153–164.

R Development Core Team, 2004, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN is 3-900051-07-0 URL <http://www.Rproject.org>.

ROCH O., ALEGRE A., 2006, Testing the bivariate distribution of daily equity returns using copulas. An application to the Spanish stock market, *Computational Statistics& Data Analysis*, 51, 1312–1329.

ROSENBLATT M., 1952, Remarks on a Multivariate Transformation, *The Annals of Mathematical statistics*, 23, 470–472.

SHIH J., LOUIS T.A., 1995, Inference on the Association Parameter in Copula Models for Bivariate Survival Data, *Biometric*s, 51: 1384–1399.

SKLAR A., 1959, Fonction de Repartition a n Dimension et Leur Marges, *Publications de L'Institut de Statistiques de L'Universite de Paris*, 8, 229–231.

WARD J. H., 1963, Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, 58, 236–244.

STATISTICS IN TRANSITION-new series, July 2010
Vol. 11, No. 1, pp. 47—62

TOBACCO USE AND ITS IMPACT ON RESPIRATORY HEALTH: A STATISTICAL APPROACH

Gyan K. Agarwal[1], Manish Trivedi[2*], Usha Jha[1], Rajendra K. Jha[3] and Ripunjai K. Shukla[2]

ABSTRACT

Background: Tobacco use, which is the cause of several respiratory diseases, generally starts in the teens. Therefore, we have examined the risk of respiratory symptoms with major focus upon active and passive smoke exposure in adults among the persons living in the newly-formed Jharkhand State of Eastern India. An intensive survey is an initiative to investigate the tobacco use in various age groups especially in adults.
Methods: The questionnaire data of the population-based sample (n = 600) was analyzed. Multiple regression models were carried out to study the current respiratory health as dependent variables, whereas Tobacco Intake, Continuous Tobacco use, Age, Education level, Job Category, Income, General Health and Past Health of the tobacco users in the study area as independent variables after adjusting for age, gender, smoking and socioeconomic status. Data analysis was performed using SPSS 16.0 software and the results accounted are based upon the analysis obtained from the Multiple Regression, Chi-square test and Analysis of Variance (ANOVA) Test.
Results: Out of the total 700 distributed questionnaires, 600 of them were accurately filled and therefore considered for the present investigation. As per our analysis, about 66.5% of the adult population exposed to tobacco is defined by the 18-25 years while 21% as defined by the 26-33 years. Of this, 38% of the reported individuals were "students" who were exposed either to the tobacco smoking or chewing.
Conclusion: The Eastern India, especially the Jharkhand state, has a huge problem of widespread tobacco use in both forms among adults, particularly among the 18-25 year age group. The prevalence of smoking is higher among rural population, while the school and college going youth are affected in the ratio of 38% from tobacco exposure in the Ranchi province.

[1] Department of Applied Chemistry, Birla Institute of Technology Mesra, Ranchi(JH).
[2] Dr. Manish Trivedi, Reader In Statistics, School of Sciences, Indira Gandhi, National Open University, New Delhi-110086.
[3] Department of Medicine, Rajendra Institute of Medical Sciences, Ranchi(JH).

Key words: Tobacco Use, Respiratory Health, Environmental Tobacco smoke, Tobacco control, Intensive Survey.

1. Introduction

Globally, tobacco is responsible for the death of 1 in 10 adults (about 5 million deaths each year) with 2.41 (1.80.3.15) million deaths in developing countries and 2.43 million in developed countries [3] and [6]. Among these, 3.84 million deaths were reported for men. The leading causes of death from smoking were found to be cardiovascular diseases (1.69 million deaths), chronic obstructive pulmonary disease (0.97 million deaths) and lung cancer (0.85 million deaths) [6]. Fifty per cent of unnecessary deaths due to tobacco occurs in middle age (35.69 years), robbing around 22 years of normal life expectancy [3]. In developed countries, smoking is estimated to cause over 90% of the lung cancer in men and about 70% of it among women. In these countries, 56%-80% of deaths are due to chronic respiratory diseases, and 22% of cardiovascular deaths are attributable to tobacco. The attributed mortality is greater in males (13.3%) than in females (3.8%). Globally, the attributable fractions for mortality due to tobacco smoking are about 12% for vascular disease, 66% for the cancer of the trachea, and bronchus and lung cancer combined, and 38% for chronic respiratory diseases, [7]. Globally, the disease consequences of tobacco use (smoking) have been more extensively and better documented than perhaps for any comparable risk factor. This is partly due to the fact that for decades, until recently, the tobacco industry kept on challenging the validity of the findings and refused to accept results that were long accepted by all health scientists. In part, it is also due to the fact that the spectrum of the diseases caused by tobacco use is very large.

Tobacco is the most important preventable cause of disease burden and death all over the world [10] and [12]. In spite of the known association of major diseases with tobacco, its continued use is very bothersome for both the health professionals and the policy-planners alike [2] and [4]. There is an urgent need to face this challenge and curb its use. This is especially important among the youth as they are more likely to start the habit in their formative years but are also more likely to quit the habit in time before any harm occurs. There is enough evidence to show that majority of smokers starts the tobacco use before 18 years of age [8]. Such information is however lacking in India.

Cigarette smoking causes 5 million deaths per year worldwide, and it is estimated that the annual death toll from smoking will climb to 10 million deaths by 2030, with 7 million deaths in developing countries [10] and [12]. Cigarette smoke damages the lower respiratory tract [17], increases oxidative stress, and increases the risk of bronchitis, chronic obstructive lung disease, cancer, and death [10], [12] and [16]. Tobacco companies have gradually shifted their market from high-income to low-income countries, where many people are poorly informed about the health risks of tobacco use and antismoking policy is relatively weak [18]. Although much research has been focused on the

relationship between smoking and adverse outcomes such as cancer, respiratory illnesses, and cardiovascular disease, the problem of smoking and its relation to malnutrition, child survival, and poverty have not been well characterized. Tobacco-related disease kills an estimated half million people a year in India, [4]. Most adult addicts to tobacco start young. Data on tobacco use by rural children or youth in India [5], [8] and [14] are few and only recently available [15]. This pilot survey assessed the degree, nature and pattern of tobacco use by children in rural areas and the need for a large study.

2. Material and methods

2.1.Study design and sampling

The above is a questionnaire-based cross-sectional study of different regions in and around the Ranchi district of the Jharkhand state of India. The data collection for this study was from January to June 2005; January to April 2006; August to November 2008; and January to April 2009.

2.2.Study Population

The entire population type was clearly differentiated as urban, semi-urban and rural type depending upon the socio-economic background of the people of the Ranchi city and its province. The people in the said area belong to either middle or lower income groups and people in higher income group is less commonly seen. Majority of the study population comprised of the age between 18-25 years which is believed to be from the student-type from the host institution, local schools and colleges as well as the University. This was followed by workers working as employees/staff of the governmental and private sectors, people within home domain included housewives, retired employees, other elderly people, and businessmen, retailers, etc., which were classed as “adults” from the urban or semi-urban population; while “adults” from rural set-up included the local tribes, working class included guards, peons, carpenters, sweepers, maidservants, etc. The observed age spectrum was between 10 and 81 years, with a few exceptions. “Adults” were appropriately graded between (2)–(9) for different age groups. Grade 1 for 10–17, Grade 2 for 18–25, Grade 3 for 26–33, Grade 4 for 34–41, Grade 5 for 42–49, Grade 6 for 50–57, Grade 7 for 58–65, Grade 8 for 66-73 and Grade 9 for 74–81.

2.3.Questionnaire

The questionnaire was adapted and standardized from its original one, as prepared for a study by the State University of New York at Buffalo [11]. Special

attention was given to include the Indian conditions for both population types. The objective of the questionnaire was to observe a relation between tobacco and health. The questionnaire was framed as Passive Smoking and Respiratory Health, to observe the impact of the environmental tobacco smoke along with respirable suspended particulate matter on an individuals' respiratory health over a period of time, as short and long term exposure [1]. This was further re-framed as Tobacco Use and Health, to comparatively assess the impact of tobacco smoke on respiratory health and tobacco chewing on oral health. Both versions of the questionnaire were chiefly divided into nine balanced sections as: Smoke Exposure History; Breathing Trouble; Hospital Utilization and Access to Care; Recent Disability; Health-related Quality of Life; the Epworth Sleepiness Scale; Respiratory Specific Quality of Life; Identification and Demographic Information.

3. Survey administration

The present study is part of the main project titled "Impacts of Environmental Tobacco Smoke on Respiratory Health of Adult" initiated at the State University of New York at Buffalo.

3.1.Surveillance

Out of the total of 700 questionnaires distributed, only 600 were returned completely filled and with other appropriate information needed, and at the same time their peak flow rates are also measured using a handy *AsthmaMentor* Peak Flow Meter (© 2002 Respironics HealthScan, Inc.; US Patent No. 5,320,170; 391,506) . The remaining 100 are not included in the present study. 20 of them were not returned to us, 50 were incomplete in different respects while 30 were soiled beyond recognition. Thus, we have considered them as no-response.

3.2.Exploration

This paper focuses upon the smoker types and their respiratory health conditions, and therefore suitable ways to monitor both short- and long-term exposure to tobacco on respiratory health of adult are explored statistically.

4. Data interpretation

We have analyzed the data using SPSS version 16.0 (SPSS South Asia Pvt. Limited). Firstly, we have calculated a tobacco exposure score for male and female exposure percentages (see Table 1). We have found that 87% male and 13% female are habitually smoking or chewing tobacco. Among the male

exposures around 70% are exposed with the effect of the tobacco smoking or chewing. Secondly, we have also calculated a similar score for various types of exposure, e.g. active smoking, passive smoking, for Ex-smoker and Non-Smoker (see Table 1). We have observed about 45% of the population is of active smoking type while 47% is of passive smoking type. Only 6% out of the 100% is not exposed to any form of tobacco (see Table 1).

Table 1. Percentage Ratio of the Tobacco Users and Their Health Status

Sex Ratio	% Ratio	Type of smokers	% Ratio	Breathing Status	% Ratio	Age Group	% ratio	Health Status (Current / Past)	% Ratio Difference
Female	12.9	Active	44.8	Acute	2.2	10-17	4.38	Excellent	2.5
						18-25	66.42	Fair	1.8
		Ex-Smoker	0.4	Fine	51.2	26-33	20.99	Good	6.0
						34-41	10.22	Poor	1.8
		Non-Smoker	1.5	Rarely	36.6	42-49	7.66	Very Good	0.1
Male	87.1					50-57	5.84	Very Poor	0.0
		Not Exposed	6.0	Repeated	8.6	58-65	2.01		
						66-73	0.55		
		Passive	47.4	Worst	1.5	74-81	0.18		
TOTAL	100	Total	100	Total	100	Total		Total	100

We have found that 1.5% had the worst breathing and coughing, while around 11% of the reported cases were facing the breathing problem either repeatedly or acutely, and 37% of the ETS exposed cases were facing the breathing problem rarely however, the intensity of the problem was fairly high. To evaluate exposure stability over time, we have compared the self-reported duration of ETS exposure during the previous 4 weeks at the various levels. We have found that the percentage difference of persons having "excellent" and "fair health" conditions are 4.3% (2.5% + 1.8%) getting their health deteriorated from 24.8% to 29.1% within a month (see Table 1). In the list of original data the proportion of the persons having excellent and fair health condition is 29.1% at the time of collection of information which was 24.8% before one month. It was also observed that where individuals reported of having poor to very poor health conditions, the ratio increased from 3.6% to 5.4% after one month.

Upon evaluating the ETS exposure effect with respect to the age factor, it was seen that about 66.5% of the population of tobacco consumers is limited to the age grade (2) and 21% of the same is restricted to the age grade (3) (see Table 1).

The implication of the above observation is alarming and it indicates that it is the youth mass (age grade 2) of the population type which is exposed to the tobacco use very prominently. While considering the youth category, 38% are students while 7% is the unemployed class that is exposed to the tobacco smoking and chewing. While studying the form of the tobacco exposure we have found that around 41% are using the smoking forms of the tobacco and 7% are regional *Khaini* (chewing tobacco) users where, as around 30% are using both tobacco forms.

5. Statistical approach

Multiple regression model is carried out for current respiratory health as dependent variable, and Tobacco Intake, Continuous Tobacco use, Age, Education level, Job Category, Income, General Health and Past Health of the tobacco users in the study area as independent variables after adjusting for age, gender, smoking and socioeconomic status. Data analysis was performed using SPSS 16.0 software. We have carried out the ANOVA test to check the variability in the social characters with reference to the tobacco exposure. We have taken the five social characteristics under consideration. The characteristics are Gender, Education Level, Job Category, Religion and Income of the tobacco users in the study area. By carrying out the ANOVA test for the Respiratory Health with respect to the different respiratory symptoms, we have found that some of the symptoms are unanimously present in each and every group of the tobacco users. We have used the chi-square goodness of fit test to study the discrepancy in between the different natures of the tobacco intake. After performing the chi-square goodness of fit test we have found that in the study region the different forms of tobacco used by the tobacco users are not homogeneous.

5.1. ANOVA test for tobacco use in relation with social characters

In this paper we have carried out the ANOVA test to check the variability in the social characters with reference to the tobacco exposure. We have taken the five social characteristics under consideration. Each characteristic is described into some sub characteristics. The characteristics are Gender, Education Level, Job Category, Religion and Income of the tobacco users in the study area. By carrying out the ANOVA test on the Gender, it is found that the average variation between the groups of the tobacco users is more than the average variation within the group. The reason behind this is that the 87% male and 13% female are habitual of smoking or chewing tobacco. This large difference in the proportions of the male and female tobacco users indicate that in the study region tobacco exposure is more prevalent in the males as compared to the females.

While considering the other form of social characteristic which is the Income under study, we have found that the habit of tobacco use is equally distributed

among the various income groups whereas in terms of the job category there are no observed variations in the tobacco exposure because in each job category the type of exposure is only changed. In this study, the proportion of the tobacco exposure is on an average same in every religion and hence, the habit of tobacco use becomes irrelevant, whereas in terms of the Education Level, it is observed that at different stages of education there are no variations in the tobacco users.

Table 2. Result of the ANOVA Test for different Social Characteristics

		Sum of Squares	df	Mean Square	F	Sig.
Gender	Between Groups	24.327	4	6.082	40.143	.000
	Within Groups	90.146	595	.152		
	Total	114.473	599			
Education level	Between Groups	3504.490	4	876.123	5.129	.000
	Within Groups	101644.428	595	170.831		
	Total	105148.918	599			
Job category	Between Groups	1696.115	4	424.029	8.075	.000
	Within Groups	31245.078	595	52.513		
	Total	32941.193	599			
Religion	Between Groups	26.917	4	6.729	.411	.801
	Within Groups	9739.401	595	16.369		
	Total	9766.318	599			
Salary	Between Groups	7.433	4	1.858	1.718	.144
	Within Groups	643.525	595	1.082		
	Total	650.958	599			

5.2. ANOVA test for respiratory health in relation with respiratory symptoms

By carrying out the ANOVA test for the Respiratory Health with respect to the different respiratory symptoms like Wheezing, Tightness of chest, Shortness of breath, Coughing, Asthma, Frequency of attack, and Exacerbations, we have found that some of the symptoms are unanimously present in each and every groups of the tobacco users. We have also observed that the symptoms like Wheezing, Asthma, Frequency of attack, Frequency at night time, Exacerbations are also present in each and every group and some of the symptoms like Tightness of chest, Shortness of breath or Coughing are found in only age grade 2 and 3.

This may be because the respiratory health of these user groups is very low ranked. It is understood to have come from the way they have rated their respiratory health scale.

Table 3. ANOVA Test for Respiratory Symptoms

		Sum of Squares	df	Mean Square	F	Sig.
Wheezing	Between Groups	1.297	4	.324	1.964	.099
	Within Groups	98.243	595	.165		
	Total	99.540	599			
Tightness of chest	Between Groups	4.463	4	1.116	6.857	.000
	Within Groups	96.802	595	.163		
	Total	101.265	599			
Shortness of breath	Between Groups	7.598	4	1.900	8.529	.000
	Within Groups	132.520	595	.223		
	Total	140.118	599			
Coughing	Between Groups	5.361	4	1.340	5.572	.000
	Within Groups	143.139	595	.241		
	Total	148.500	599			
Asthma	Between Groups	.241	4	.060	1.510	.198
	Within Groups	23.718	595	.040		
	Total	23.958	599			
Frequency of attack	Between Groups	635.208	4	158.802	1.232	.296
	Within Groups	76710.652	595	128.925		
	Total	77345.860	599			
Frequency at night time	Between Groups	934.267	4	233.567	2.078	.082
	Within Groups	66866.692	595	112.381		
	Total	67800.958	599			
Exacerbations	Between Groups	.168	4	.042	.170	.954
	Within Groups	147.030	595	.247		
	Total	147.198	599			

5.3. Chi-Square test for tobacco users of different nature

In this paper, we have made use of the chi-square goodness of fit test to study the discrepancy in between the different natures of the tobacco intake. We have taken the four different types of tobacco intake such as Smokers, Chewers, Mixed of both types and others. In different groups of tobacco users more than 50% are the smokers, 25% are the chewers and 15% tobacco users take in both forms of the tobacco.

Table 4. Chi-Square Goodness of fit test for the Tobacco intake

	Observed N	Expected N	Residual
No tobacco	253	240.0	13.0
Smokers	183	180.0	3.0
Chewers	88	90.0	-2.0
Mixed	62	60.0	2.0
Others	14	30.0	-16.0
Total	600		

After performing the chi-square goodness of fit test, we have found that in the present study region the different forms of tobacco used by the tobacco users are not homogeneous. The proportion of the tobacco intake in different forms is not the same as the smokers are more than 50%. Having considered the expected frequencies of the different tobacco intake forms to be 0.40, 0.30, 0.20 and 0.10, the critical value of the chi-square test statistics falls within the acceptance region (Table 5).

Table5. Result of the Chi-Square Goodness of Fit Test Statistics

	Tobacco intake
Chi-Square	9.399[a]
df	4
Asymp. Sig.	.052

However, if the chi-square test was to be carried out with the assumptions of the homogeneous expected frequencies, then the assumptions are not fitted with the observed value of the frequencies of the tobacco users of different categories.

5.4. Regression Model for the Respiratory Health

In this paper, we have studied the average relationship of 8 different variables with respect to the Respiratory Health. These variables are understood to affect the Respiratory Health of the tobacco users. Such variables are Tobacco Intake, Continue Tobacco use, General Health, Past Health, Age, Education level, Job Category and Income (Table 6).

Table 6. Regression model for the Respiratory Health

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
(Constant)	124.969	1.076		116.179	.000	122.856	127.082
Tobacco intake	-.022	.016	-.017	-1.359	.175	-.055	.010
Continue tobacco use	.050	.057	.010	.883	.378	-.061	.161
Age	-.051	.022	-.032	-2.280	.023	-.095	-.007
Education level	.033	.018	.023	1.860	.063	-.002	.069
Job category	.030	.034	.012	.904	.366	-.036	.096
Salary	.154	.225	.008	.687	.492	-.287	.596
Your general health	-11.693	.382	-.614	-30.640	.000	-12.442	-10.943
Your past health	-7.737	.389	-.380	-19.882	.000	-8.501	-6.972

a Dependent Variable: Respiratory Health on a scale of 0 to 100.

Our regression analysis shows that some of the parameters may have a negative effect on the Respiratory Health and rest of the same may have a positive effect. Tobacco intake and the Age factor very badly affect the respiratory health of the tobacco user. The general health and the past health of the tobacco user also have a negative effect on the respiratory health whereas the Education level, Job category and the Income structure show the positive effect on the respiratory health. The possible explanation for this trend could be observed from the identification of an individual on the basis of education and living style.

In the present study, we have also conducted the multiple regression analysis by making use of the expression of the multiple regression model for the respiratory health. This expression may suitably be defined by the dependent variable (Y) which represents the Respiratory health while independent variables (X1, X2, ….., X8) represent Tobacco Intake, Continuously Tobacco use, Age, Education level, Job Category, Income, General Health and Past Health of the tobacco users in the study area. The multiple regression model can be interpreted

as Y =124.969 - 0.022 X1 + 0.050 X2 - 0.051 X3 + 0 .033 X4 + 0.030 X5 + 0.154 X6 -11.693 X7 -7.737 X8

From the histogram created for the Respiratory Health (Fig. 1) on a scale of 0 to 100 where 0 represents "dead health" and 100 represents "full health", we have observed the frequency distribution of the regression standardized residual to be following the standard normal distribution with the mean 6.06 E -16, and the standard deviation is found to be 0.993.

Figure 1. Histogram

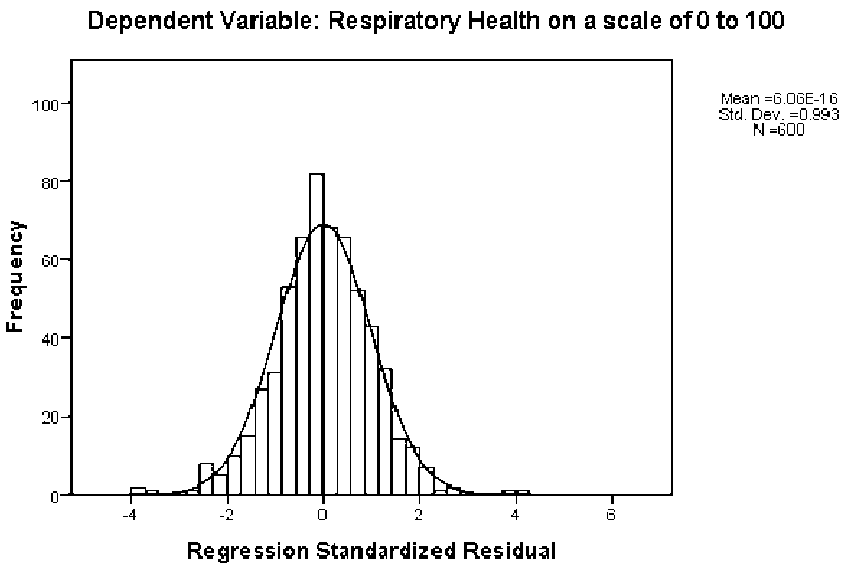
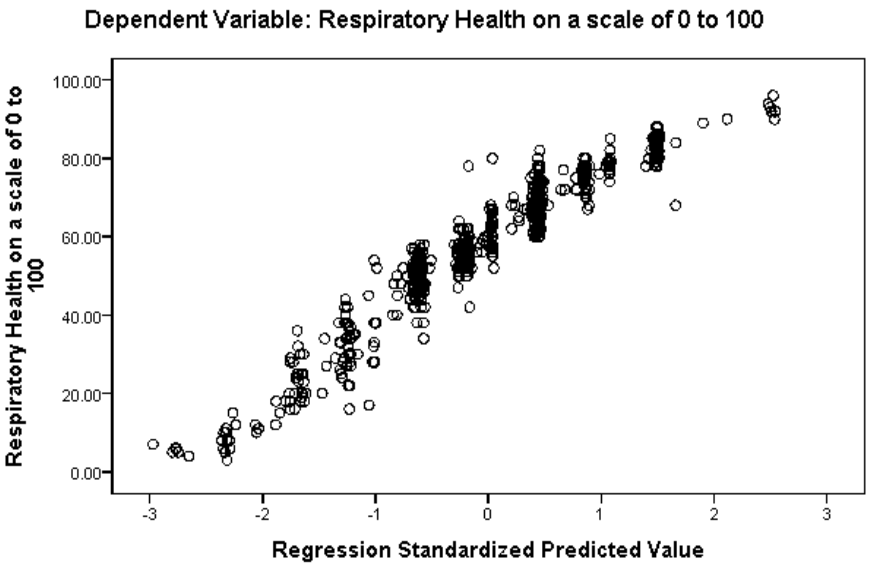


Figure 2. Scatter plot



In (Fig. 2), the scatter plot of the regression standardized predicted values for respiratory health on scale of 0 to 100 is shown. It is found that most of the tobacco users have rated their respiratory health from 50 to 80. As a result, much of the predicted values are scattered around 50 to 80. This is evident as the values are very densely marked in the scatter plot.

6. Discussion & conclusion

There have been studies in the past to investigate the trends of tobacco use in the previously combined state of Bihar [13] and also in Darabhanga district (Bihar) in 1967 [9]. However, this is possibly the first time that such a study reports exclusively for the newly formed state of Jharkhand.

Our study establishes statistically that of the 600 subjects from random populations, there were 253 (42.2%) non-consumers and 347 (57.8%) consumer of tobacco. There were 42.2% of active smokers who reported for repeated breathing trouble observed from their peak flow rates marked as PFM1, PFM2 and PFM3. Regarding the passive smokers, there were 47.4% who developed respiratory symptoms where a majority were those exposed for long term within their homes and as a result showed a decreased peak flow rate. Remaining tobacco users that belong both to smokeless tobacco and mixed type showed varying troubles with their respiratory symptoms: 14% wheezing, 12% Tightness of chest, 15.5% shortness of breath, (34%) coughing. For smokeless tobacco

users, they are also impairing their oral health but this is outside the scope of the present study.

It is understood from our present observations that the demographic parameters such as gender, age, caste, religion, education, work etc. under the set of environmental situations arising from rural or urban setup form the governing basis for getting hooked to tobacco habits. The age of an individual is regarded highly susceptible to developing habits like those of tobacco and alcohol. And as has been reported in the past, this study points to impaired respiratory health due to tobacco use. This is also evident from our results obtained from regression analysis. Ranchi and its province have shown vividness in its tobacco habits and consumption types which may mainly be attributed to its topology and culture. Further, it is also seen that the youth of this region is highly susceptible to developing the respiratory symptoms particularly if they are from low economic background and using tobacco.

Funding

The present study was supported by the Birla Institute of Technology, Mesra, Ranchi, India.

Declaration of interests

The authors have no competing interests to declare. The views expressed in this publication are sole responsibility of the authors and there is no direct or indirect influence of any outside source. None of the authors of the present paper is affiliated to any tobacco-related agency.

REFERENCES

[1] AGARWAL, G. K., JHA, U. & JHA, R. K., (2005), Study of the impacts of Environmental Tobacco Smoke and Respiratory Suspended Particulate Matter on Respiratory Health of Adult. Proceedings of the First International Conference on Environmental Exposure and Health, Atlanta, Georgia.

[2] BLANC P. & TOREN K. (1999), How much asthma can be attributed to occupational factors? American Journal of Medicine,107:580–7.

[3] CHAN-YEUNG M, MALO J. L. & TARLO SM, (2003), Proceedings of the first Jack Pepys Occupational Asthma Symposium, American Journal of Respiratory and Critical Care Medicine;167:450–71.

[4] GUPTA P. C., HAMNER J. & MURTI P., (1992), Control of Tobacco-Related Cancers and Other Disease. Proceedings of an International Symposium. Mumbai, Oxford University Press, pp 353–355.

[5] JAYANT K., (1991), Identifying specific tobacco problems in Children and Tobacco: The Wider View, Eds. Charlton A, Moyer C. International Union against Cancer, Geneva, p 27.

[6] KOGEVINAS M., ANTO J. M. & SUNYER J., (1999), Occupational asthma in Europe and other industrialised areas: a population-based study, European Community Respiratory Health Survey Study Group. Lancet, 353:1750–4.

[7] KOGEVINAS M., ZOCK J. & JARVIS D., (2007), Exposure to substances in the workplace and new-onset asthma: an international prospective population based study (ECRHS-II). Lancet 370:336–41.

[8] LAHIRI V., (1991), Identifying specific tobacco problems In Children and Tobacco: The Wider View Eds. Charlton A, Moyer C, International Union against Cancer (UICC), Geneva, pp 25–26.

[9] MEHTA F. S., PINDBORG J. J, HAMNER J. E. (1971); Oral cancer and precancerous condition in India, Tata Institute of Fundamental Research, Mumbai, Munksgaard, Copenhagen, 113–126.

[10] MEYER J. D., HOLT D. L. & CHERRY N. M., (1999), SWORD '98: Surveillance of work-related and occupational respiratory disease in the UK, Occupation Medicine, 49:485–9.

[11] MISHRA, A., (2003), Questionnaire on the study: Impacts of Environmental Tobacco Smoke (ETS) on respiratory health of adults, State University of New York at Buffalo, USA.

[12] PETO R, LOPEZ A. D., BOREHAM J., THUN M. & HEATH C. J., (1996), Doll R. Mortality from smoking worldwide, Britain Medicinal Bulletin, 52: 12–21.

[13] SINHA D. N., GUPTA P. C. & PEDNEKAR M. S., (2003), Tobacco use in a rural area of Bihar, India, Indian Journal of Community Medicine, 28(4): 167–170.

[14] Survey Results, (1990), Children and Tobacco: Newsletter No. 1, International Union against Cancer (UICC), Geneva.

[15] U S Department of Health and Human Services, (1994), Preventing tobacco use among young people: A report of the Surgeon General, Atlanta, Georgia: Public.

[16] VAIDYA S.G., VAIDYA N. S. & NAIK U. D., (1992), Epidemiology of tobacco habits in Goa, India. *In:* Control of Tobacco-Related Cancers and

Other Diseases. Proceedings of an International Symposium, Eds. Gupta PC, Hamner J. and Murti P., Mumbai, Oxford University Press, pp 315–320.

[17] World Health Report, (2002), Quantifying selected major risks to health, World Health Organization 2002, Available from http://www.who.int.whr/2002/chapter4.pdf.

[18] YACH D. & BETTCHER D., (2000), Globalization of tobacco industry influence and new global responses, Tobacco Control, 9: 206-16, Health Service, Centre for Disease Control and prevention, Office on Smoking and Health, 1994.

STATISTICS IN TRANSITION-new series, July 2010
Vol. 11, No. 1, pp. 63—74

BHATTACHARYA AND HOLLA DISTRIBUTION AND SOME OF ITS INTERESTING PROPERTIES AND APPLICATIONS

Anwar Hassan, Mehraj Ahmed Bhat[1]

ABSTRACT

In this paper we consider Bhattacharya and Holla's (1965) distribution and derive its important structural properties which have not been studied so far. The distribution has very interesting properties when transformed in trigonometric functions which are also useful and alternative expression for obtaining its higher order moments. We also make an attempt to obtain estimate of its parameters. The model is fitted to the set of observed data and relative precision and a comparison has also been made.

Key words: Generating function, trigonometric function, compound distribution.

1. Introduction

Gurland (1957) studied interrelations among compound and generalized distributions. Gurland (1958) also defined a general class of contagious distributions. Patil (1964) studied on certain compound Poisson and compound binomial distribution. Blischke (1963) defined mixture of discrete distributions. Cohen (1963) obtained estimation in mixtures of discrete distributions. Johnson and Kotz (1969) and Johnson, Kotz and Kemp (1992) also discussed some mixture, compound and contagious distributions.

One of the interesting and useful compound distributions is defined by Bhattacharya and Holla (1965). This distribution showed its importance in the theory of accident proneness. This distribution is obtained by mixing Poisson distribution with rectangular distribution denoted as Poisson $\hat{\theta}$ rectangular (a, b) as

$$P(X=k) = \frac{1}{b-a}\int_a^b \frac{e^{-\theta}\theta^k}{k!}d\theta; \quad k = 0, 1,2,\ldots \tag{1.1}$$

[1] P. G. Department of Statistics, University of Kashmir, Srinagar, India, anwar.hassan2007@gmail.com, anwar_hassan2007@hotmail.com, mehraj_stat@yahoo.co.in

$$P(X=k)=\frac{1}{b-a}\left[e^{-a}\left(1+\frac{a}{1!}+\frac{a^2}{2!}+\ldots+\frac{a^k}{k!}\right)-e^{-b}\left(1+\frac{b}{1!}+\frac{b^2}{2!}+\ldots+\frac{b^k}{k!}\right)\right] \quad (1.2)$$

$$k = 0, 1, 2, \ldots$$

This is known as Bhattacharya and Holla (1965) distribution.

As regards the applications, the use of (1.1) in the field of the Poisson queuing system, accident statistics and acceptance sampling plans of manufactured articles based on the counting of defects, etc. are well known, the variations in service facilities from one server to another in a parallel service channels may result in fluctuations in individuals in the first case, in variations in the accident liabilities from individual to individual in the second case and, in the later case, in an inevitable and continuous changing in manufacturing conditions leading to fluctuations in the Poisson distribution parameter involved.

In this study we derive its important structural properties which have not been done earlier so far. The important feature of this paper is that to obtain moments and moment generating function in terms of trigonometric functions. This transformed form gives very interesting properties of the distribution and higher order moments can also be obtained using this transformed form. This expression is an alternative for obtaining moments of the proposed model. We also provide estimation of the parameters of the distribution. The Poisson model has been fitted by many statisticians. The same data has been fitted to this model. This model gives better result as compared to Poisson distribution.

2. Properties of Compound Distribution

2.1. Moments about origin

The first four moments of (1.1) about origin are derived as

$$\mu_1' = \frac{b+a}{2} \quad (2.1)$$

$$\mu_2' = \frac{a^2+b^2+ab}{3}+\frac{b+a}{2} \quad (2.2)$$

$$\mu_3' = \frac{b^3+ba^2+ab^2+a^3}{4}+(a^2+b^2+ab)+\frac{b+a}{2} \quad (2.3)$$

$$\mu_4' = \frac{a^4 + b^4 + a^3b + ab^3 + a^2b^2}{5} + \frac{3(a^3 + b^3 + a^2b + ab^2)}{2} + \frac{7(a^2 + b^2 + ab)}{3} + \frac{a+b}{2} \tag{2.4}$$

The central moments of (1.1) are obtained as

$$\mu_2 = \frac{(b-a)^2}{12} + \frac{b+a}{2} \tag{2.5}$$

$$\mu_3 = \frac{b+a}{2} + \frac{(b-a)^2}{4} \tag{2.6}$$

$$\mu_4 = \frac{(b-a)^4}{80} + \frac{(a+b)(b-a)^2}{4} + \frac{4a^2 + 4b^2 + ab}{3} + \frac{a+b}{2} \tag{2.7}$$

2.2. Moment Generating Function

We also find the moment generating function of (1.1)

$$M_x(t) = \frac{e^{-a(1-e^t)} - e^{-b(1-e^t)}}{(b-a)(1-e^t)}$$

$$M_x(t) = \frac{1}{b-a}\left[\begin{array}{l}(b-a) - \frac{(b^2-a^2)}{2!}(1-e^t) \\ + \frac{(b^3-a^3)}{3!}(1-e^t)^2 - \frac{(b^4-a^4)}{4!}(1-e^t)^3 + \ldots\end{array}\right] \tag{2.8}$$

2.3. Characteristic Function

$$\phi_X(t) = \frac{e^{-a(1-e^{it})} - e^{-b(1-e^{it})}}{(b-a)(1-e^{it})}$$

$$\phi_x(t) = \frac{1}{b-a}\left[\begin{array}{l}(b-a) - \frac{(b^2-a^2)}{2!}(1-e^{it}) + \frac{(b^3-a^3)}{3!}(1-e^{it})^2 \\ - \frac{(b^4-a^4)}{4!}(1-e^{it})^3 + \ldots\end{array}\right] \tag{2.9}$$

2.4.Cumulant Generating Function

It provides a technique for obtaining moments of the distribution in a very simple and convenient way. The provided Log M_x (t) is a convergent function of t.

$$K_X(t)=\left\{\left[\frac{b^2-a^2}{2!(b-a)}\left(t+\frac{t^2}{2!}+\dots\right)+\frac{b^3-a^3}{3!(b-a)}\left(t+\frac{t^2}{2!}+\dots\right)^2\right.\right.$$

$$\left.+\frac{b^4-a^4}{4!(b-a)}\left(t+\frac{t^2}{2!}+\dots\right)^3+\cdots\right]$$

$$-\frac{1}{2}\left[\frac{b^2-a^2}{2!(b-a)}\left(t+\frac{t^2}{2!}+\dots\right)+\frac{b^3-a^3}{3!(b-a)}\left(t+\frac{t^2}{2!}+\dots\right)^2\right.$$

$$\left.+\frac{b^4-a^4}{4!(b-a)}\left(t+\frac{t^2}{2!}+\dots\right)^3+\cdots\right]^2$$

$$+\frac{1}{3}\left[\frac{b^2-a^2}{2!(b-a)}\left(t+\frac{t^2}{2!}+\dots\right)+\frac{b^3-a^3}{3!(b-a)}\left(t+\frac{t^2}{2!}+\dots\right)^2\right.$$

$$\left.\left.+\frac{b^4-a^4}{4!(b-a)}\left(t+\frac{t^2}{2!}+\dots\right)^3+\cdots\right]^3-\cdots\right\} \tag{2.10}$$

$$K_X(t)=Y-\frac{1}{2}Y^2+\frac{1}{3}Y^3-\frac{1}{4}Y^4+\dots$$

where

$$Y=\frac{b^2-a^2}{2!(b-a)}\left(t+\frac{t^2}{2!}+\dots\right)+\frac{b^3-a^3}{3!(b-a)}\left(t+\frac{t^2}{2!}+\dots\right)^2+\frac{b^4-a^4}{4!(b-a)}\left(t+\frac{t^2}{2!}+\dots\right)^3+\cdots$$

Now equating the coefficients of like powers of t on both sides of equation (2.10), we get the cumulants as

$$K_1=\frac{a+b}{2} \tag{2.11}$$

$$K_2 = \frac{(b-a)^2}{12} + \frac{b+a}{2} \tag{2.12}$$

Similarly

$$K_3 = \frac{b+a}{2} + \frac{(b-a)^2}{4} \tag{2.13}$$

$$\mathrm{K}_4 = -\frac{(\mathrm{b}-\mathrm{a})^4}{120} + \frac{7(\mathrm{b}-\mathrm{a})^2}{12} - \frac{\mathrm{a}+\mathrm{b}}{2} \tag{2.14}$$

$$K_4 + 3K_2^2 = \frac{(b-a)^4}{80} + \frac{(a+b)(b-a)^2}{4} + \frac{4a^2+4b^2+ab}{3} + \frac{a+b}{2} \tag{2.15}$$

The (2.12), (2.13) and (2.15) are also central moments equal to (2.5) to (2.7)

2.5.Skewness

The Skewness of the (1.1) can be obtained by using the expressions given in (2.5) and (2.6) as

$$\gamma_1 = \sqrt{108}\,\frac{\left[2(\mathrm{b}+\mathrm{a})+(\mathrm{b}-\mathrm{a})^2\right]}{\left[6(\mathrm{b}+\mathrm{a})+(\mathrm{b}-\mathrm{a})^2\right]^{3/2}} \tag{2.16}$$

It is clear from above for given a and b the Skewness is positive for $\frac{(\mathrm{b}-\mathrm{a})^2}{\mathrm{b}+\mathrm{a}} > -2$, and the Skewness is negative for $-6 < \frac{(\mathrm{b}-\mathrm{a})^2}{\mathrm{b}+\mathrm{a}} < -2$ provided $\mathrm{b}+\mathrm{a} \neq 0$. In case $\mathrm{b}+\mathrm{a}=0$

$$\gamma_1 = \frac{\sqrt{108}}{\mathrm{b}-\mathrm{a}}$$

This means Skewness is inversely proportional to (b – a)

2.6.Kurtosis

$$\beta_2 = 3 + \frac{6}{5}\,\frac{\left[60(\mathrm{a}+\mathrm{b})+70(\mathrm{b}-\mathrm{a})^2-(\mathrm{b}-\mathrm{a})^4\right]}{\left[36(\mathrm{a}+\mathrm{b})^2+12(\mathrm{b}+\mathrm{a})(\mathrm{b}-\mathrm{a})^2+(\mathrm{b}-\mathrm{a})^4\right]} \tag{2.17}$$

For given values of a and b the Kurtosis is greater than 3 if the condition $\frac{60}{(b-a)^2-70} > \frac{(b-a)^2}{a+b}$ provided a + b ≠ 0 is satisfied, this reveals the models (1.1) and (1.2) are leptokurtic. However, if the condition $\frac{60}{(b-a)^2-70} < \frac{(b-a)^2}{a+b}$ provided a + b ≠ 0 is satisfied. The models are platykurtic. In case a + b = 0, $\beta_2 = 3 + \frac{84}{(b-a)^2} - 6/5$

This means coefficient of Kurtosis is a decreasing function of (b – a). The models take normal form at b – a = $\pm\sqrt{70}$.

3. Recurrence Relation for Moment Generating Function and Moments in terms of Trigonometric function

We have MGF (2.8) which can also be written as

$$M_X(t) = \frac{1}{b-a}\sum_{r=0}^{\infty}\frac{(b^{r+1}-a^{r+1})}{(r+1)!}(-1)^r(1-e^t)^r \tag{3.1}$$

Taking the n-th term (3.1) and let

$$D = \frac{b^{n+1}-a^{n+1}}{(b-a)(n+1)!}(-1)^n(1-e^t)^n \tag{3.2}$$

$$= A_n(1-e^t)^n \tag{3.3}$$

where $A_n = \frac{(b^{n+1}-a^{n+1})}{(b-a)(n+1)}(-1)^n$

$$\text{Let } D = A_n(1-e^t)^n \tag{3.4}$$

Put $e^t = \text{Sin}^2\theta$

Since we find moments at t = 0, when t = 0, θ = π/2 so in the transformed case we will find moments of the distribution for θ = π/2

$$D = A_n \text{ Cos}^{2n}\theta \tag{3.5}$$

where $e^t = \text{Sin}^2\theta$

$$\frac{dD}{dt} = D' = A_n\left[\frac{d}{d\theta}\text{Cos}^{2n}\theta\right]\frac{d\theta}{dt}$$

$$D^1 = -A_n \; n \, Cos^{2n-2}\theta \; Sin^2 \theta \tag{3.6}$$

$$D^2 = D' + n(n-1) \, A_n Cos^{2n-4}\theta \, Sin^4\theta \tag{3.7}$$

$$D^3 = D^2 + n(n-1)A_n \Big[4 \, Cos^{2n-4}\theta \, Sin^3\theta \, Cos \, \theta$$

$$- (2n-4)Cos^{2n-5} Sin \, \theta \, Sin^4\theta\Big]\frac{Sin\,\theta}{2 \, Cos \, \theta}$$

$$D^3 = -2 \, D' + 3 \, D^2 - n(n-1)(n-2)A_n \; Cos^{2n-6}\theta \, Sin^6\theta \tag{3.8}$$

$$D^4 = 6D^1 - 11D^2 + 6D^3 + n(n-1)(n-2)(n-3)A_n \; Cos^{2n-8}\theta \, Sin^8\theta \tag{3.9}$$

$$D^5 = -24D^1 + 50D^2 - 35D^3 + 10D^4$$

$$- n(n-1)(n-2)(n-3)(n-4)A_n \; Cos^{2n-10}\theta \, Sin^{10}\theta \tag{3.10}$$

$$\begin{aligned} D^6 &= 120D^1 - 274D^2 + 225D^3 - 85D^4 + 15D^5 \\ &+ n(n-1)\ldots(n-5)A_n \; Cos^{2n-12}\theta \, Sin^{12}\theta \end{aligned} \tag{3.11}$$

$$D^7 = -720D^1 + 1764D^4 - 1624D^3 + 735D^4 - 175D^5 + 21D^6$$

$$- n(n-1)\ldots(n-6)A_n \; Cos^{2n-14}\theta \, Sin^{14}\theta \tag{3.12}$$

Where $D^1, D^2, D^3, D^4 \ldots$ are $1^{st}, 2^{nd}, 3^{rd}, 4^{th} \ldots$ derivatives of D (3.4)
Now writing the general term as

$$D^n = (-1)^n \left\{ \left[a_1 D^1 - a_2 D^2 + a_3 D^3 - a_4 D^4 + \ldots - (n-1)\text{terms} \right] + n! \, A_n Sin^{2n}\theta \right\} \tag{3.15}$$

$$D^n = (-1)^n \left\{ f_{n-1}(D) + n! \, A_n Sin^{2n}\theta \right\} \tag{3.16}$$

where n = 2, 3, 4, …
where $f_{n-1}(D) = a_1 D^1 - a_2 D^2 + a_3 D^3 - a_4 D^4 + \ldots$, where n = 2, 3, 4, …
where the coefficient a_1, a_2, a_3, … a_{n-1} are determined by using the following iterative table which is discussed below:

$f_n(D)$	n	a_1	a_2	a_3	a_4	a_5	a_6
$f_1(D)$	1	1					
$f_2(D)$	2	2	3				
$f_3(D)$	3	6	11	6			
$f_4(D)$	4	24	50	35	10		
$f_5(D)$	5	120	274	225	85	15	
$f_6(D)$	6	720	1764	1624	735	175	21

The coefficient matrix is formed by a straightforward method. First column, i.e. coefficient column a_1 is obtained by writing down n! in each row for each n = 1, 2, …. Diagonal of the matrix, i.e. coefficient of a_1, a_2, a_3, … along the diagonal is obtained by summing up the integers up to n = 1, 2, 3, … in each row. In other words, we use $n(n+1)/2$ for writing down diagonal elements.

If c_{ij} denotes the element of the given matrix being in the i-th row and j-th column then c_{ij} $c_{ij} = i\, c_{(i-1)j} + c_{(i-1)(j-1)}$ for every i=2,…,n , j=2,3,…, i-1

Moreover, the above recurrence relation can also be written in matrix form as:

$$\begin{bmatrix} D^2 \\ -D^3 \\ D^4 \\ -D^5 \\ D^6 \\ -D^7 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & -3 & 0 & 0 & 0 & 0 \\ 6 & -11 & 6 & 0 & 0 & 0 \\ 24 & -50 & 35 & -10 & 0 & 0 \\ 120 & -274 & 225 & -85 & 15 & 0 \\ 720 & -1764 & 1624 & -735 & 175 & -21 \end{bmatrix} \begin{bmatrix} D^1 \\ D^2 \\ D^3 \\ D^4 \\ D^5 \\ D^6 \end{bmatrix} + \begin{bmatrix} 2\,A_2 \mathrm{Sin}^4\theta \\ 6\,A_3 \mathrm{Sin}^6\theta \\ 24\,A_4 \mathrm{Sin}^8\theta \\ 120\,A_5 \mathrm{Sin}^{10}\theta \\ 720\,A_6 \mathrm{Sin}^{12}\theta \\ 5040\,A_7 \mathrm{Sin}^{14}\theta \end{bmatrix}$$

Solving this matrix equation we can obtain moments of any order. (3.17)

4. Estimation

We estimate a and b by moments method, $m_r = \mu'_r$ where μ'_r and m_r are population moments and sample moments about origin respectively. The estimates of a and b are

$$\hat{b} = \bar{x} + \sqrt{3(s^2 - \bar{x})} \tag{4.1}$$

$$\hat{a} = \bar{x} - \sqrt{3(s^2 - \bar{x})} \tag{4.2}$$

where $\bar{x}$ and s^2 are sample mean and variance.

5. Goodness of Fit

The proposed model seems to explain observed data with great accuracy. The model is fitted with the observed data in comparison with Poisson distribution. It seems that proposed model is relatively more appropriate for natural behaviour.

We present the fitting to a data to which Poisson has been be fitted by Consul and Jain (1973). The same set of observed data is fitted to the proposed model (1.1) which has not been fitted earlier. The relative precision is also presented graphically in the following tables given below.

Table 5.1. Accidents of 647 women working on H.E. Shells during five weeks

No. of Accidents	Observed Frequency	Expected frequencies	
		Poisson Distribution PD	Bhattacharya and Holla Distribution BHD
0	447	406	454
1	132	189	113
2	42	44	59
3	21	7	16
4	3	1	4
5	2	0	1
TOTAL	647	647	647
Mean	0.465224		
Variance	0.690826		
Estimates		λ =0.465224	a= -0.356
			b= 1.287
χ^2		49.42	11.008
d.f		2	1

Table 5.2.

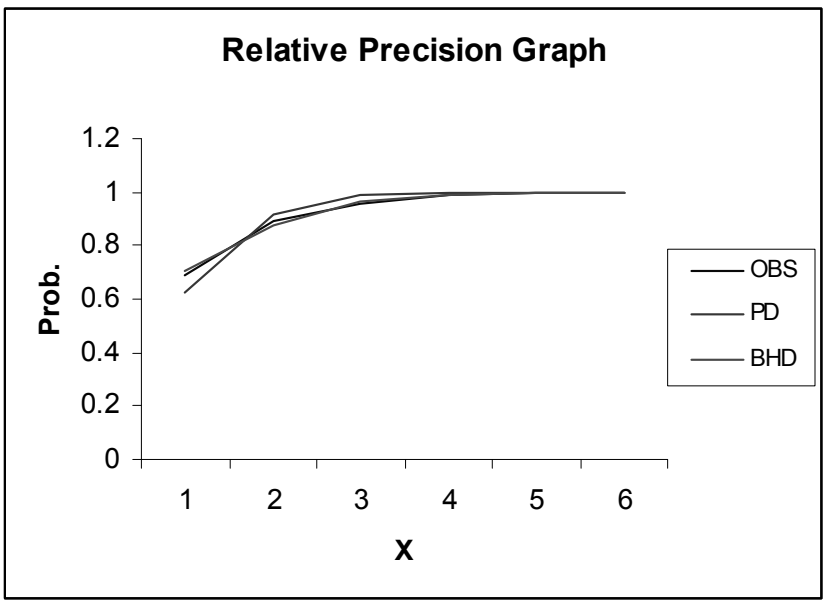
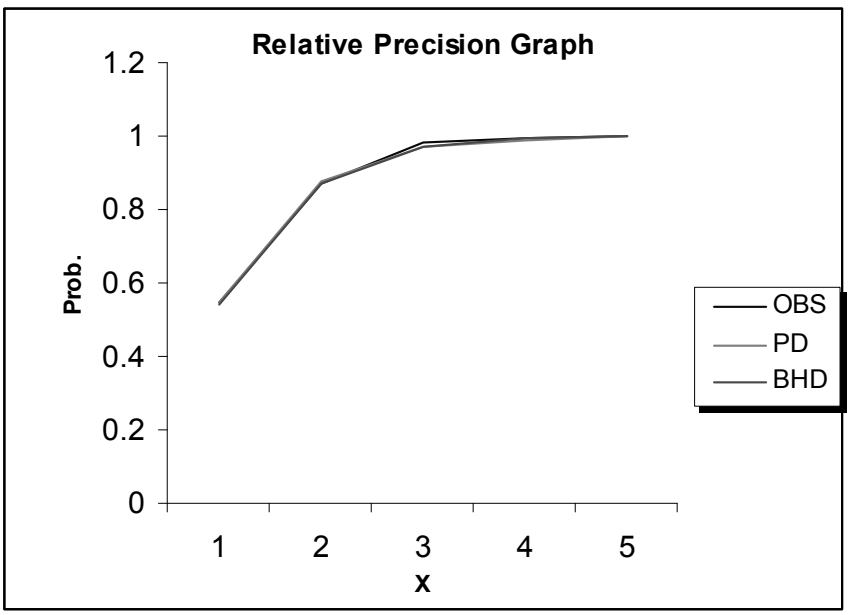


Table 5.3. Death due to horse kicks in the Prussian army

Number of deaths	Observed Frequency	Expected frequencies	
		Poisson Distribution PD	Bhattacharya and Holla Distribution BHD
0	109	109	108
1	65	66	66
2	22	19	20
3	3	4	5
4	1	2	1
TOTAL	200	200	200
Mean	0.61000		
Variance	0.610955		
Estimates		λ= 0.61000	a=0.5565
			b= 0.6635
χ^2		1.24	1.006
d.f		2	1

Table 5.4.



It is evident from the tables 5.1 , 5.2, 5.3 and 5.4 that the values of chi-square, and graph of relative precision in all the cases, the model gives the remarkable best fit as compared to Poisson distributions.

Conclusion

In this paper, we considered the Bhattacharya and Holla (1965) distribution as possible alternative to accident statistics and acceptance sampling plans of manufactured articles based on the counting of defects and for analyzing data from different behavior. We studied various properties of the model interims of generating function (MGF). The model has interesting properties when moment generating function is transformed in trigonometric functions.

REFERENCES

BHATTACHARYA, S. K. and HOLLA, M. S. (1965): On a discrete distribution with special references to the theory of accident proneness; Journal of the American Statistical Association, **60**:1060–1066.

BLISCHKE, W. R. (1963): Mixture of discrete distribution, proceedings of the International Symposium on Discrete Distributions, Montreal, 385–397.

COHEN, A. C. (1963): Estimation in mixtures of discrete distributions, proceedings of International Symposium on Discrete Distributions, Montreal, 373–378.

GURLAND, J. (1957): Some interrelations among compound and generalized distributions, Biometrika, **44**:265–268.

GURLAND, J. (1958): A general class of contagious distributions, Biometrics, **14**: 229–249.

PATIL, G. P. (1964): On certain compound Poisson and compound binomial distribution, Synkhya, Series A, **26**:293–294.

JOHNSON and KOTZ, (1969) Univariate discrete distribution, John Wiley & Sons Johnson, Kotz and Kemp, (1992): Univariate discrete distribution, John Wiley & Sons.

CONSUL, P.C and JAIN (1973): A generalization of Poisson distribution: Technometrics, 15(4)791–799.

STATISTICS IN TRANSITION-new series, July 2010
Vol. 11, No. 1, pp. 75—89

GENERALIZED CLASS OF SYNTHETIC ESTIMATORS FOR SMALL AREAS UNDER SYSTEMATIC SAMPLING SCHEME

Krishan K. Pandey[1], G. C.Tikkiwal[2]

ABSTRACT

This paper defines and discusses a generalized class of synthetic estimators for small domain, using auxiliary information, under systematic sampling scheme. The generalized class of synthetic estimators, among others, includes the simple, ratio and product synthetic estimators. Further, it demonstrates the use of the generalized synthetic and ratio synthetic estimators for estimating crop acreage for small domain and also compares their relative performance with direct estimators, empirically, through a simulation study.

Key words: Synthetic Estimation; Small Domain; Inspector Land Revenue Circles (ILRCs); Timely Reporting Scheme (TRS); Absolute Relative Bias (ARB); Simulated relative standard error (Srse).

1. Introduction

The common feature of small area estimation problem is that when large-scale sample surveys are designed to produce reliable estimates at the national or state level, generally they do not provide estimates of adequate precision at lower levels like District, Tehsil / County, and Inspector land Revenue Circle. This is because the sample sizes at the lower level are generally insufficient to provide reliable estimates using traditional estimators. Therefore, the need was felt to develop alternative estimators to provide small area statistics using the data already collected through large-scale surveys. The traditional design based and alternative estimators are also termed, in the literature of small area estimation, respectively as direct and indirect estimators.

The indirect estimators are based on methods which increase the effective sample size either by (i) simulating enough data through appropriate analysis of available data under appropriate modelling or (ii) by using data from other

[1] Assistant Professor, CMES, University of Petroleum & Energy Studies, Dehradun- 248007, India.
[2] Department of Mathematics & Statistics, J.N.V. University, Jodhpur, Rajasthan -342011, India.

domains and /or time periods through models that assume similarities across domain and /or time periods. The only known method so far belonging to category (i) is SICURE- modelling [TIKKIWAL.(1993)].The other methods of estimation like Synthetic, Composite, and Generalized Regression belong to category (ii). Among these the synthetic estimators are used for small area estimation, mainly because of its simplicity, applicability to general sampling design and potential to increase accuracy in estimation. However, if the implicit model assumption of similarities across domain and /or time period fails, the synthetic estimator may be badly design biased. GONZALEZ (1973), GONZALEZ and WAKESBERG (1973), GHANGURDE AND SINGH (1977, 78), Tikkiwal & Pandey (2007) among others study the synthetic estimator based on auxiliary variables, viz. the ratio synthetic estimator. These studies show that synthetic estimators provide reliable estimates to some extent.

Tikkiwal and Ghiya (2000) define and discuss a generalized class of synthetic estimators for small domains, using auxiliary information, under simple random sampling and stratified random sampling schemes. The generalized class, among others, includes simple, ratio and product synthetic estimators. The two authors compare empirically the relative performance of various direct and synthetic estimators for estimating crop acreage for small domains.

This paper discusses the generalized class of synthetic estimators using auxiliary information under systematic sampling scheme. The systematic sampling scheme, being operationally more convenient in practice, is often used in large-scale field surveys under multistage design. In such survey, like crop acreage surveys in India, ultimate stage of sampling units like villages / households / agricultural fields etc. are selected by systematic sampling scheme. Systematic sampling scheme, apart from operationally more convenient, provides more efficient estimators under certain conditions [Cf. Cochran (1977), Sukhatme et al. (1984), Madow (1946) & Osborne, J.G. (1942)].

2. Formulation of the problem & Notations

Let us suppose that we have a finite population $U = (1,...,i,...,N)$ which is divided into 'A' non-overlapping small areas U_a of size N_a $(a = 1,.....,A)$ for which estimates are required. Let the characteristic under study be denoted by 'y' and also assume that the auxiliary information is available, which is denoted by 'x'. Suppose the population units in small area 'a' are numbered 1 to N_a, i.e. $U_a = (1,.....,N_a)$ and n_a units are to be selected by systematic sampling scheme. A systematic sample of size n_a is selected from each small area 'a', $(a = 1,.....,A)$ either (i) by linear systematic sampling scheme, (when

$N_a = n_a k_a$, k_a being an integer) or (ii) by circular systematic sampling scheme, (when $N_a \neq n_a k_a$). Consequently,

$$\sum_{a=1}^{A} N_a = N \quad \text{and} \quad \sum_{a=1}^{A} n_a = n \ ,$$

The various population and sample means for characteristic X &Y can be denoted by:

$\bar{X}$ & $\bar{Y}$ = Means of the population based on N observations.

$\bar{X}_a$ & $\bar{Y}_a$ = Population means of domain 'a' based on N_a observations.

$\bar{x}_{ai}$ & $\bar{y}_{ai}$ = Sample means of domain 'a' based on n_a observations.

Case (i): For the case $N_a = n_a k_a$, i.e. for linear systematic sampling scheme, arrange the population units into $n_a k_a$ arrays and select a random number, say i between 1 and k_a then every k_a^{th} unit thereafter. So the sample consists of n_a units from $N_a (= n_a k_a)$ units, and the sample is $\{i, i+k_a, ..., i+(n_a-1)k_a\}$. The number i is called random start and k_a is the sampling interval. Further, let x_{aij} & y_{aij} denote the values of the auxiliary variate and characteristic under study respectively for the jth unit of the ith sample bearing serial number i +(j-1) k_a , i =1,....., k_a ; j =1,....., n_a . Therefore,

$$\bar{X}_a = \frac{1}{N_a}\sum_i \sum_j x_{aij} \ , \ \bar{x}_{ai.} = \frac{1}{n_a}\sum_{j=1}^{n_a} x_{aij} \ , \ \bar{Y}_a = \frac{1}{N_a}\sum_i \sum_j y_{aij} \ \text{and}$$

$$\bar{y}_{ai.} = \frac{1}{n_a}\sum_{j=1}^{n_a} y_{aij}$$

Various mean squares and coefficient of variations of subpopulation 'U_a' for auxiliary variate x & characteristic under study, y is denoted by

$$S_{x_a}^2 = \frac{1}{k_a - 1}\sum_{i=1}^{k_a} (\bar{x}_{ai.} - \bar{X}_a)^2 \ , \ C_{x_a} = \frac{S_{x_a}}{\bar{X}_a} \ \text{and} \ S_{y_a}^2 = \frac{1}{k_a - 1}\sum_{i=1}^{k_a} (\bar{y}_{ai.} - \bar{Y}_a)^2 \ ,$$

$$C_{y_a} = \frac{S_{y_a}}{\bar{Y}_a}$$

The coefficient of covariance between X and Y is denoted by

$$C_{x_a y_a} = \frac{S_{x_a y_a}}{\bar{X}_a \bar{Y}_a} \text{ , where } S_{x_a y_a} = \frac{1}{k_a - 1} \sum_{i=1}^{k_a} (\bar{y}_{ai.} - \bar{Y}_a)(\bar{x}_{ai.} - \bar{X}_a)$$

Case (ii): For the case $N_a \neq n_a k_a$, i.e. for those small areas where N_a / n_a is not an integer but k_a is the integer nearest to N_a / n_a, Lahiri (1954) suggested to use circular systematic sampling design. Here, in this case a random number is chosen from 1 to N_a and the units corresponding to this random number are chosen as the random start. There, after every k_a^{th} unit is chosen in a cyclic manner until a sample of n_a units is selected. Thus, if i is a number selected at random from 1 to N_a, the sample consists of units corresponding to these numbers are

$$\{i + (j-1)k_a\} \quad \text{if} \quad i + (j-1)k_a \leq N_a$$

$$\{i + (j-1)k_a - N_a\} \quad \text{if} \quad i + (j-1)k_a > N_a, \text{ j =1,2,……..,} \ n_a$$

In this case the x_{aij} & y_{aij} denote the values of the auxiliary variate and characteristic under study respectively for the jth unit of the ith sample bearing the number $\{i + (j-1)k_a\}$ or $\{i + (j-1)k_a - N_a\}$ as the case may be for $j = 1, 2, ..., n_a$. The various mean squares and coefficient of variations of sub population 'U_a' for auxiliary variate x & characteristic under study, y in this case will be as follows:

$$S_{1x_a}^2 = \frac{1}{N_a - 1} \sum_{i=1}^{N_a} (\bar{x}_{ai.} - \bar{X}_a)^2 , \quad C_{1x_a}^2 = \frac{S_{1x_a}^2}{\bar{X}_a^2} \quad \text{and} \quad S_{1y_a}^2 = \frac{1}{N_a - 1} \sum_{i=1}^{N_a} (\bar{y}_{ai.} - \bar{Y}_a)^2 ,$$

$$C_{1y_a}^2 = \frac{S_{1y_a}^2}{\bar{Y}_a^2}$$

The coefficient of covariance between X and Y is denoted by

$$C_{1x_a y_a} = \frac{S_{1x_a y_a}}{\bar{X}_a \bar{Y}_a}, \text{ where } S_{1x_a y_a} = \frac{1}{N_a - 1} \sum_{i=1}^{N_a} (\bar{y}_{ai.} - \bar{Y}_a)^2 (\bar{x}_{ai.} - \bar{X}_a)^2$$

3. Generalized Class of Synthetic Estimators

Following Srivastava (1967), we define in this section a generalized class of synthetic estimators of population mean $\bar{Y}_a$ based on the auxiliary variable 'x' under Systematic Sampling Scheme as follows.

$$\bar{y}_{s,a} = \bar{y}_w \left(\frac{\bar{x}_w}{\bar{X}_a} \right)^{\beta} \tag{3.1}$$

Where β is a suitably chosen constant, and

$$\left.\begin{aligned} \bar{y}_w &= \sum{}' \, p_a \bar{y}_{ai.} + \sum{}'' \, p_a \bar{y}_{ai.} \\ \bar{x}_w &= \sum{}' \, p_a \bar{x}_{ai.} + \sum{}'' \, p_a \bar{x}_{ai.} \end{aligned}\right\} \tag{3.2}$$

Where $\sum{}'$ denotes the summation over those small areas where $N_a = n_a k_a$ and $\sum{}''$ denotes summation over those small areas where $N_a \neq n_a k_a$ and $p_a = \frac{N_a}{N}$. Here, clearly

$$E(\bar{y}_w) = \bar{Y} \text{ and } E(\bar{x}_w) = \bar{X} \tag{3.3}$$

The above estimator $\bar{y}_{s,a}$ perform well under the following condition

$$\bar{Y}_a \left(\bar{X}_a \right)^{\beta} \cong \bar{Y} (\bar{X})^{\beta} \tag{3.4}$$

It is noted that the synthetic estimator $\bar{y}_{s,a}$ is consistent if the condition given in (3.4) is satisfied.

Remark 3.1.

If β = 0, -1, 1, the estimator $\bar{y}_{s,a}$ in (3.1) reduces to $\bar{y}_{s,s,a} = \bar{y}_w$, $\bar{y}_{s,r,a} = \left(\frac{\bar{y}_w}{\bar{x}_w} \right) \bar{X}_a$, and $\bar{y}_{s,p,a} = \bar{y}_w \left(\frac{\bar{x}_w}{\bar{X}_a} \right)$ respectively with synthetic condition $\bar{Y}_a \cong \bar{Y}$, $\frac{\bar{Y}_a}{\bar{X}_a} \cong \frac{\bar{Y}}{\bar{X}}$, and $\bar{Y}_a (\bar{X}_a) \cong \bar{Y} \bar{X}$.

4. Design Bias and Mean Square Error of Generalized Synthetic Estimator

Design Bias and Mean Square Error of generalized synthetic estimator, under the synthetic condition given in (3.4), is as follows

$$B(\bar{y}_{s,a}) = \bar{Y}_a \left[\beta \left\{ \sum{}'_a p_a^2 \frac{(k_a - 1)}{k_a} \frac{S_{y_a x_a}}{\bar{X}\,\bar{Y}} + \sum{}''_a p_a^2 \frac{(N_a - 1)}{N_a} \frac{S_{1x_a y_a}}{\bar{X}\,\bar{Y}} \right\} \right.$$

$$+\frac{\beta(\beta-1)}{2}\left\{\sum_a{}' p_a^2 \frac{(k_a-1)}{k_a}\frac{S_{x_a}^2}{\bar{X}^2}+\sum_a{}'' p_a^2 \frac{(N_a-1)}{N_a}\frac{S_{1x_a}^2}{\bar{X}^2}\right\}\Bigg] \tag{4.1}$$

for a = 1, …., A.

$$MSE(\bar{y}_{s,a}) = E(\bar{y}_{s,a}-\bar{Y}_a)^2 = \bar{Y}_a^2\Bigg[\left\{\sum_a{}' p_a^2 \frac{(k_a-1)}{k_a}\frac{S_{y_a}^2}{\bar{Y}^2}+\sum_a{}'' p_a^2 \frac{(N_a-1)}{N_a}\frac{S_{1y_a}^2}{\bar{Y}^2}\right\}$$

$$+(2\beta-1)\beta\left\{\sum_a{}' p_a^2 \frac{(k_a-1)}{k_a}\frac{S_{x_a}^2}{\bar{X}^2}+\sum_a{}'' p_a^2 \frac{(N_a-1)}{N_a}\frac{S_{1x_a}^2}{\bar{X}^2}\right\}$$

$$+4\beta\left\{\sum_a{}' p_a^2 \frac{(k_a-1)}{k_a}\frac{S_{y_a x_a}}{\bar{X}\bar{Y}}+\sum_a{}'' p_a^2 \frac{(N_a-1)}{N_a}\frac{S_{1x_a y_a}}{\bar{X}\bar{Y}}\right\}\Bigg]$$

$$-2\bar{Y}_a^2\Bigg[\beta\left\{\sum_a{}' p_a^2 \frac{(k_a-1)}{k_a}\frac{S_{y_a x_a}}{\bar{X}\bar{Y}}+\sum_a{}'' p_a^2 \frac{(N_a-1)}{N_a}\frac{S_{1y_a x_a}}{\bar{X}\bar{Y}}\right\}$$

$$+\frac{\beta(\beta-1)}{2}\left\{\sum_a{}' p_a^2 \frac{(k_a-1)}{k_a}\frac{S_{x_a}^2}{\bar{X}^2}+\sum_a{}'' p_a^2 \frac{(N_a-1)}{N_a}\frac{S_{1x_a}^2}{\bar{X}^2}\right\}\Bigg] \tag{4.2}$$

The suitable value of β is the one for which $MSE(\bar{y}_{s,a})$ is minimum. So minimizing the $MSE(\bar{y}_{s,a})$ with respect to β under synthetic condition, gives simplified expression for β, if $\bar{X} \cong \bar{X}_a$ as follows

$$\beta = \frac{-\left\{\sum_a{}' p_a^2 \frac{(k_a-1)}{k_a}\frac{S_{y_a x_a}}{\bar{X}\bar{Y}}+\sum_a{}'' p_a^2 \frac{(N_a-1)}{N_a}\frac{S_{1x_a y_a}}{\bar{X}\bar{Y}}\right\}}{\left\{\sum_a{}' p_a^2 \frac{(k_a-1)}{k_a}\frac{S_{x_a}^2}{\bar{X}^2}+\sum_a{}'' p_a^2 \frac{(N_a-1)}{N_a}\frac{S_{1x_a}^2}{\bar{X}^2}\right\}} \tag{4.3}$$

It is noted that the expression of MSE for direct estimator under linear & circular systematic sampling design, is minimum if $\alpha = -\frac{C_{x_a y_a}}{C_{x_a}^2}$ [Cf. Srivastava (1967)].

5. Estimation of Mean square errors

Since a systematic sample can be regarded as a random selection of one cluster, it is not possible to give an unbiased or even consistent estimator of the design variances of $\bar{y}_{ai.}$ or $\bar{x}_{ai.}$. A common practice in applied survey work is to regard the sample as random and, for lack of knowing what else to do, estimate the variance using simple random sample formulae. Unfortunately, if followed indiscriminately this practice can lead to badly biased estimators and incorrect inferences concerning the population parameters of interest.

Wolter (1984, 1985) investigate several biased estimators of variances with a goal of providing some guidance about when a given estimator may be more appropriate than other estimators. The criterion to judge the various estimators on the basis of their bias, their mean square error, and proportion of confidence interval formed using the variance estimators which contain the true population parameter of interest. This study suggests the use of biased but simple estimator v_{2y} for $V(\bar{y}_{ai.})$, when sample size is very small for both the situations, viz. when $N_a = n_a k_a$ and $N_a \neq n_a k_a$. The expression of v_{2y} is given as follows:

$$v_{2y} = (1-f)\left(\frac{1}{n_a}\right)\sum_{j=2}^{n_a}\frac{a_{ij}^2}{2(n_a-1)} \tag{5.1}$$

$$\left.\begin{aligned} &where \quad a_{ij} = \Delta y_{ij} = y_{ij} - y_{i,\,j-1} \\ &and \quad f = \frac{n_a}{N_a} \end{aligned}\right\} \tag{5.2}$$

Similarly, estimate of $V(\bar{x}_{ai.})$ is given by v_{2x}, where

$$v_{2x} = (1-f)\left(\frac{1}{n_a}\right)\sum_{j=2}^{n_a}\frac{b_{ij}^2}{2(n_a-1)} \tag{5.3}$$

$$\left.\begin{aligned} &where \quad b_{ij} = \Delta x_{ij} = x_{ij} - x_{i,\,j-1} \\ &and \quad f = \frac{n_a}{N_a} \end{aligned}\right\} \tag{5.4}$$

We note that above estimators v_{2y} and v_{2x} are based on overlapping differences of Δy_{ij} & Δx_{ij} respectively. Further, the estimate of covariance term between $\bar{y}_{ai.}$ and $\bar{x}_{ai.}$, given by Swain (1964), is

$$\hat{Cov}(\bar{y}_{ai.}, \bar{x}_{ai.}) = r\sqrt{v_{2y} v_{2x}} \tag{5.5}$$

where r is correlation coefficient between x and y observations based on the sample of size n_a.

5.1. Estimation of mean square error of direct estimator

Following Srivastava (1967), the generalized class of direct estimators of $\bar{Y}_a$ under systematic sampling scheme is $\bar{y}_{d,a}^G = \bar{y}_{ai.}\left(\frac{\bar{x}_{ai.}}{\bar{X}_a}\right)^{\alpha}$.

Its mean square under case (i) is

$$MSE(\bar{y}_{d,a}^G) = \bar{Y}_a^2\left[\frac{V(\bar{y}_{ai.})}{\bar{Y}_a^2} + \frac{V(\bar{x}_{ai.})}{\bar{X}_a^2} + \frac{2\alpha Cov(\bar{y}_{ai.}, \bar{x}_{ai.})}{\bar{X}_a \bar{Y}_a}\right]$$

$$\text{or } MSE(\bar{y}_{d,a}^G) = V(\bar{y}_{ai.}) + \alpha^2 R_a^2 V(\bar{x}_{ai.}) + 2\alpha R_a Cov(\bar{y}_{ai.}, \bar{x}_{ai.}) \tag{5.6}$$

where $R_a = \frac{\bar{Y}_a}{\bar{X}_a}$, thus a consistent estimator of $MSE(\bar{y}_{d,a}^G)$ is given by

$$mse(\bar{y}_{d,a}^g) = v_{2y} + \alpha^2 r_a^2 v_{2x} + 2\alpha r_a r\sqrt{v_{2y} v_{2x}} \tag{5.7}$$

Where $r_a = \frac{\bar{y}_a}{\bar{x}_a}$ is the ratio of sample means. It is also observed that the mean square error for direct estimator in case of circular systematic sampling is given by

$$MSE\left(\bar{y}_{d,a}^G\right)_c = \bar{Y}_a^2\left[\frac{V(\bar{y}_{ai.})_c}{\bar{Y}_a^2} + \alpha^2\frac{V(\bar{x}_{ai.})_c}{\bar{X}_a^2} + 2\alpha\frac{Cov(\bar{y}_{ai.}, \bar{x}_{ai.})_c}{\bar{X}_a \bar{Y}_a}\right]$$

$$MSE\left(\bar{y}_{d,a}^G\right)_c = V(\bar{y}_{ai.})_c + \alpha^2 R_a^2 V(\bar{x}_{ai.})_c + 2\alpha R_a Cov(\bar{y}_{ai.}, \bar{x}_{ai.})_c \tag{5.8}$$

Thus, consistent estimator of $MSE\left(\bar{y}_{d,a}^G\right)_c$ is given by

$$mse(\bar{y}_{d,a}^g)_c = v_{2y}^{'} + \alpha^2 r_a^2 v_{2x}^{'} + 2\alpha r_a r\sqrt{v_{2y}^{'} v_{2x}^{'}} \tag{5.9}$$

where $v_{2y}^{'}$ and $v_{2x}^{'}$ are the estimates of variances of $V(\bar{y}_{ai.})_c$ and $V(\bar{x}_{ai.})_c$ respectively in case of circular systematic sampling design. To be calculated similarly as of v_{2x} and v_{2y}.

5.2. Estimation of mean square error of synthetic estimator

The expression for the Mean Square Error given in (4.2), can be approximated under the synthetic condition given in (3.4) as follows:

$$MSE(\bar{y}_{s,a}) = \sum_a^{'} p_a^2 V(\bar{y}_{ai.}) + \sum_a^{''} p_a^2 V(\bar{y}_{ai.})_c$$

$$+ \beta^2 R_a^2 \left\{ \sum_a^{'} p_a^2 V(\bar{x}_{ai.}) + \sum_a^{''} p_a^2 V(\bar{x}_{ai.})_c \right\}$$

$$+ 2\beta R_a \left\{ \sum_a^{'} p_a^2 Cov(\bar{y}_{ai.}, \bar{x}_{ai.}) + \sum_a^{''} p_a^2 Cov(\bar{y}_{ai.}, \bar{x}_{ai.})_c \right\} \qquad (5.10)$$

Thus, a consistent estimator of $MSE(\bar{y}_{s,a})$ is given by

$$mse(\bar{y}_{s,a}) = \left\{ \sum_a^{'} p_a^2 v_{2y} + \sum_a^{''} p_a^2 v_{2y}^{'} \right\} + \beta^2 r_a^2 \left\{ \sum_a^{'} p_a^2 v_{2x} + \sum_a^{''} p_a^2 v_{2x}^{'} \right\}$$

$$+ 2\beta r_a \left\{ \sum_a^{'} p_a^2 r \sqrt{v_{2y} v_{2x}} + \sum_a^{''} p_a^2 r \sqrt{v_{2y}^{'} v_{2x}^{'}} \right\} \qquad (5.11)$$

Where $r_a = \dfrac{\bar{y}_a}{\bar{x}_a}$ is the ratio of sample means.

6. Crop Acreage Estimation for Small Domain – A Simulation Study

This section demonstrates the use of the generalized synthetic and ratio synthetic estimators to obtain crop acreage estimates for small domain and also compare their relative performance with the corresponding direct estimators empirically, through a simulation study. This is done by taking up the state of Rajasthan, one of the states in India, for case study [Cf. Tikkiwal & Ghiya (2000)].

6.1. Existing methodology for estimation

In order to improve timelines and quality of crop acreage statistics, Timely Reporting Scheme (TRS) is used by most of the States of India. The TRS has the

objective of providing quick and reliable estimates of crop acreage statistics and thereby production of the principle crops (i.e. Jowar, Bajra, Maize etc.) during each agricultural season. Under the scheme, the Patwari (Village Accountant) is required to collect acreage statistics on a priority basis in a 20 percent sample of villages, selected by stratified linear systematic sampling design, taking Tehsil (a sub-division of the District) as a stratum. These statistics are further used to provide state level estimates using direct estimators, viz. unbiased (based on sample mean) and ratio estimators.

6.2. Details of the simulation study

For collection of revenue and administrative purposes, the State of Rajasthan, like most of the other states of India, is divided into a number of districts. Further, each district is divided into a number of Tehsils and each Tehsil is also divided into a number of Inspector Land Revenue Circles (ILRCs). Each ILRC consists of a number of villages. For the present study, we take ILRCs as small domains.

In the simulation study, we undertake the problem of crop acreage estimation for all Inspector Land Revenue Circles (ILRCs) of Jodhpur Tehsil of Rajasthan. They are seven in number and these ILRCs contain respectively 29, 44, 32, 30, 33, 40 and 44 villages. These ILRCs are small domains from the TRS point of view. The crop under consideration is Bajra (Indian corn or millet) for the agriculture season 1993-94. The Bajra crop acreage for agriculture season 1992-93 is taken as the auxiliary characteristic x. The various information regarding the ILRCs of Jodhpur Tehsil is provided in the Table 6.2.1.

Table 6.2.1. Total Area (Irrigated and Unirrigated) under Bajara Crop in ILRCs of Jodhpur Tehsil for Agricultural seasons 1992–93 and 1993–94

S.No	ILRCs of Jodhpur Tehsil	No. of villages in ILRC	Total area(Irr.+U.Irr.) under the crop Bajra in 1992-93	Total area(Irr.+U.Irr.) under the crop Bajra in 1993-94
1	Jodhpur (1)	29	7799.5899	5696.5000
2	Keru (2)	44	21209.5880	15699.6656
3	Dhundhada (3)	32	19019.0288	16476.4863
4	Bisalpur (4)	30	15153.9248	14269.0000
5	Luni (5)	33	19570.1323	16821.4508
6	Dhava (6)	40	25940.0979	25075.5000
7	Jajawal Kalan (7)	44	18007.4120	15875.0000
	Total	252	126699.7737	109913.6027

Below is the list of all those estimators, whose relative performance is to be assessed for estimating population total T_a of small domain for 'a' = 1, 2 …7.

Direct estimators

Direct ratio estimator $\hat{T}_{1,a} = N_a\, \bar{y}_{d,r,a} = N_a \left(\dfrac{\bar{y}_{ai.}}{\bar{x}_{ai.}} \right) \bar{X}_a$

Direct general estimator $\hat{T}_{2,a} = N_a\, \bar{y}^{G}_{d,a} = N_a \bar{y}_{ai.} \left(\dfrac{\bar{x}_{ai.}}{\bar{X}_a} \right)^{\alpha}$

Where $\bar{y}_{ai.} = \dfrac{1}{n_a} \sum_{j=1}^{n_a} y_{aij}$; and $\bar{x}_{ai.} = \dfrac{1}{n_a} \sum_{j=1}^{n_a} x_{aij}$

Indirect estimators

Ratio synthetic estimator $\hat{T}_{3,a} = N_a\, \bar{y}_{s,r,a} = N_a \left(\dfrac{\bar{y}_w}{\bar{x}_w} \right) \bar{X}_a$

Generalized synthetic estimator $\hat{T}_{4,a} = N_a\, \bar{y}_{s,a} = N_a \bar{y}_w \left(\dfrac{\bar{x}_w}{\bar{X}_a} \right)^{\beta}$

where $\bar{y}_w = \sum{}' p_a \bar{y}_{ai.} + \sum{}'' p_a \bar{y}_{ai.}$; and $\bar{x}_w = \sum{}' p_a \bar{x}_{ai.} + \sum{}'' p_a \bar{x}_{ai.}$

Before simulation, we examine the condition of generalized synthetic and synthetic ratio estimators as given in Eq. (3.4) and in remark (3.1). These results are presented in following tables 6.2.2 & 6.2.3 respectively. We note that both the above conditions met for ILRCs (3), (5), (7) deviate moderately for ILRCs (4) & (6) and deviate considerably for ILRC (7).

Table 6.2.2. Absolute Differences (Relative) under Synthetic Assumption of Synthetic Ratio Estimator for Various ILRCs

ILRC	$\bar{Y}_a / \bar{X}_a$	$\bar{Y} / \bar{X}$	$\left[\left\|(\bar{Y}_a / \bar{X}_a) - (\bar{Y} / \bar{X})\right\| \div (\bar{Y}_a / \bar{X}_a)\right] \times 100$
(1)	0.73036	0.86751	18.17
(2)	0.7402	0.86751	17.19
(3)	0.8663	0.86751	0.13
(4)	0.9416	0.86751	7.86
(5)	0.8595	0.86751	0.91
(6)	0.9666	0.86751	10.25
(7)	0.8815	0.86751	1.58

Table 6.2.3. Absolute Differences under Synthetic Assumption of Generalized Synthetic Estimator for Various ILRCs

ILRC	$\bar{Y}_a(\bar{X}_a)^\beta$	$\bar{Y}(\bar{X})^\beta$	$\left[\left\|\{\bar{Y}_a(\bar{X}_a)^\beta - \bar{Y}(\bar{X})^\beta\}\right\| \div \bar{Y}_a(\bar{X}_a)^\beta\right]\times 100$
(1)	3.31157	4.6578	40.65232
(2)	2.11349	2.4947	18.03699
(3)	0.77584	0.7791	0.42019
(4)	1.23143	1.1343	7.887578
(5)	0.8136	0.82231	1.070551
(6)	0.14251	0.13789	3.241878
(7)	2.40008	2.44412	1.834939

Now, for simulation study, taking villages as sampling units, 500 independent systematic samples each of size 25, 50, 63, 76 and 88 are selected by the procedure described in section 2 from the population of 252 villages of Jodhpur Tehsil. The simulation length was estimated with the help of the concept discussed by Whitt, W. (1989) & Murphy, K.E. Carter, C.M. & Wolfe, L. H. (2001) based on the steady state condition, that is selecting approximately 10 percent, 20 percent, 25 percent, 30 percent and 35 percent villages independently form each ILRC. For each small area estimator under consideration and for each sample size we compute Absolute Relative Bias (ARB) and Average Square Error (ASE), as defined below.

$$ARB(\hat{T}_{k,a}) = \frac{\left|\frac{1}{500}\sum_{s=1}^{500}\hat{T}_{k,a}^{s} - T_a\right|}{T_a}\times 100 \tag{6.1}$$

and $$Srse(\hat{T}_{k,a}) = \frac{\sqrt{ASE(\hat{T}_{k,a})}}{E(\hat{T}_{k,a})}\times 100 \tag{6.2}$$

Where $ASE(\hat{T}_{k,a}) = \frac{1}{500}\sum_{s=1}^{500}\left(\hat{T}_{k,a}^{s} - T_a\right)^2$ and $E(\hat{T}_{k,a}) = \frac{1}{500}\sum_{s=1}^{500}\hat{T}_{k,a}^{s}$

For k = 1,……,4 and a =1,……., 7.

6.3. Results

We present the results of ARB and Srse in Table (6.3.1) only for $n = 50$, (a sample of 20 present villages, as presently adopted in TRS) as the findings from other tables are similar.

For assessing the relative performance of the various estimators, we have to adopt some rule of thumb. Here, we adopt the rule that at the ILRCs level, an estimator should not have Srse more than 10 % and bias more than 5%.

Table 6.3.1. Simulated relative standard error (in %) and Absolute Relative Bias (in %) for various ILRCs under SRSWOR scheme, for n = 50

Estimator	ILRCs						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$\hat{T}_{1,a}$	37.83 (20.00)	24.91 (20.83)	8.63 (0.81)	16.63 (9.87)	13.01 (0.193)	17.87 (12.00)	15.41 (1.181)
$\hat{T}_{2,a}$	19.67 (19.12)	21.31 (19.60)	8.21 (0.75)	14.44 (9.66)	9.03 (0.085)	17.56 (11.53)	10.47 (1.071)
$\hat{T}_{3,a}$	18.46 (9.80)	17.62 (10.18)	6.18 (0.98)	12.02 (7.32)	8.13 (0.523)	11.86 (6.61)	6.51 (1.68)
$\hat{T}_{4,a}$	17.02 (9.00)	13.99 (10.09)	4.82 (0.8)	11.12 (7.10)	7.06 (0.47)	8.99 (5.20)	5.53 (1.50)

Note: *The figures shown in parentheses are the Absolute Relative Biases in percentage.*

We note from the table that none of the estimators satisfy the rule in ILRCs 1 and 2. This may be because, in these circles, there is a considerable deviation from the synthetic condition, as observed earlier. In ILRCs 4 and 6, where the condition deviate moderately, $\hat{T}_{4,a}$ alone satisfies the rule to some extent. In ILRCs 3, 5 and 7, where the synthetic condition closely meet, both $\hat{T}_{3,a}$ and $\hat{T}_{4,a}$ satisfy the rule but $\hat{T}_{4,a}$'s performance is slightly better than $\hat{T}_{3,a}$.

From the above analysis it is clear that if the synthetic estimators do not deviate considerably from their corresponding synthetic condition then, performance of the synthetic estimators $\hat{T}_{3,a}$ and $\hat{T}_{4,a}$, based on a sample of 20 present villages (as presently being taken under TRS), is satisfactory at the level of ILRCs. Therefore, these estimators are also likely to perform better both at Tehsil and district levels. When the synthetic estimators deviate considerably from their corresponding synthetic condition, then we should look for other types of estimators such as those obtained through the ***SICURE MODEL*** [TIKKIWAL, B.D. (1993)] and assess their relative performance through studies of the kind, in series, over some years for crop acreage estimation.

Acknowledgement

The author is very thankful to the referee for his valuable suggestions.

REFERENCES

GHANGURDE, P.D. and SINGH, M.P. (1978).“*Evaluation of efficiency of synthetic estimates*”, Proceeding of the Social Statistical Section of the American Statistical Association, 53–61.

COCHRAN, W.G. (1977).“*Sampling Techniques*”, Wiley & Sons.

GONZALEZ, M.E. (1973). “*Use and evaluation of synthetic estimates”,* Proceedings of the social statistical section of American Statistical Association, 33–36.

GONZALEZ, M. E. and WAKSBERG, J. (1973).“*Estimation of the error of synthetic estimates”,* paper presented at first meeting of the International association of survey statisticians, Vienna, Austria, 18–25, august 1973.

LAHIRI, D.B. (1954).“*On the question of bias of systematic sampling*”, Proceedings of the world Population Conference, 6, 349–62.

MURPHY, K.E. CARTER, C.M. & WOLFE, L.H. (2001) “*How long should I simulate, and for how many trials? A practical guide to reliability simulations”*, Reliability and Maintainability Symposium, 2001, Proceedings, Philadelphia, p.p. 207–212.

MADOW, L.H. (1946).“*Systematic sampling and its relation to other sampling designs”,* Journal of American Statistical Association 41:204–217.

Osborne, J.G. (1942).“*Sampling errors of systematic & random survey of covertype areas*”, Journal of American Statistical Association 37:256–264.

SRIVASTAVA, S.K. (1967). “*An estimator using auxiliary information in sample surveys”,* Calcutta Statist. Assoc. Bull. , 121–132.

SWAIN, A.K.P.C. (1964).“*The use of systematic sampling in ratio estimates*”, JISA, 2, 160–164.

SUKHATME, P. V. SUKHATME, B. V., SUKHATME, S. and ASOK, C. (1984). “*Sampling Theory of Surveys with Applications (3rd Edition)”*, Indian Society of Agricultural Statistics, New Delhi.

TIKKIWAL, G.C. and GHIYA, A. (2000). “*A generalized class of synthetic estimators with application of crop acreage estimation for small domains*”, Biom, J, 42, 7, 865–876.

TIKKIWAL, B.D. (1993). "*Modelling through survey data for small domains",* Proceedings of International Scientific conference on small area statistics and survey Design (held in September, 1992 at Warsaw, Poland).

TIKKIWAL, G.C. and PANDEY, K.K. (2007)."*On Synthetic and Composite Estimators for Small Area Estimation under Lahiri – Midzuno Sampling Scheme*", Statistics in Transition New Series, vol. 8. No. 1, 111–123.

WOLTER, K.M. (1984, 1985)."*An investigation of some estimators of variances for systematic sampling*", J. Amer. Stat. Assoc., 79, 781–90.

WHITT, W. (1989)."*Simulation run length planning"*, in the proceedings of the 1989 winter simulation conference, AT & T Bell laboratories U.S.A., pp-106–112.

STATISTICS IN TRANSITION-new series, July 2010
Vol. 11, No. 1, pp. 91—103

EXTREME VALUE MODELING OF THE MAXIMUM TEMPERATURE: A CASE STUDY OF HUMID SUBTROPICAL MONSOON REGION IN INDIA

Ripunjai K. Shukla, Manish Trivedi*, Manoj Kumar[1]

ABSTRACT

This experiment is sought to identify and fit the generalized extreme value distribution for extreme maximum temperature data of Ranchi by using the method of maximum likelihood. Ranchi District is located in Humid Sub-Tropical Monsoon Region in Jharkhand state in India. To examine the uncertainty of estimated parameters Q-Q plot and goodness of fit (K-S test) criteria were applied. The study revealed that three parameter Generalized Extreme Value Distribution fitted very well to the data. The estimates of 10, 50, 100 and 200 years return level for yearly extreme maximum temperature are described in that how they vary in future. Further, Exponential Smoothing technique is also applied to capture the trend of extreme maximum temperature in which the residuals fulfilled their assumptions, i.e. randomness and normality.

Key words: Extreme maximum temperature, Maximum likelihood, Generalized Extreme Value etc.

Introduction

Modelling the behaviour of extreme events is important in many fields. These include hydrology, climatology, engineering as well as insurance and finance. In these fields, one is often concerned about the maximum or minimum level of some values rather than the expected or average case.

Extreme events manifest themselves in a variety of ways: large insurance claims, flood levels of rivers, extreme temperature, value-at-risk analysis of a portfolio, wind-speeds, and wave heights during a storm, to name a few. With the ever-growing securitization of risk, there is a commensurately growing need to properly model events that could shock the financial system [Parisi, (2000)].

1 Department of Applied Mathematics, Birla Institute of Technology Mesra, Ranchi (JH)-835215, Email: manish_trivedi1976@yahoo.com (Dr. Manish Trivedi - Corresponding Author), msinha_09@rediffmail.com (Dr. Manoj Kumar), ripu121@yahoo.co.in (Mr. Ripunjai K. Shukla).

Among several popular extreme value models, the Generalized Extreme Value (GEV) model play leading role for extreme data. Emil Julius Gumbel, a German mathematician, developed new distributions in the 1950s for extreme events, known as Gumbel distribution. The GEV distribution model combines three types of extreme value model into one expression, with Type I for Gumbel model, Type II for Frechet model and Type III for Weibull model [Huang et al, (2008)]. The GEV Distribution model has been used by [Onofrio et al., (1999)] and [Phein and Fang, (1989)] in Hydrology, Meteorology and Flood analysis. [Michele and Salvadori, (2005)] used GEV distribution as well as Pareto distribution also to estimate extreme of the parameters. [Morrison and Smith, (2002)] developed the model for flood peaks using the generalized extreme value distribution. The Gumbel-distribution, GEV distribution and the Generalized Pareto Distribution is new and quickly growing branch of statistics [Demoulin and Roehrel (2004)].

There is a growing dissatisfaction with the use of standard statistical tools for the prediction of extremes and rare events. Examples abound in several scientists disciplines of gross under estimation, based on historical data, of the probabilities of extreme events that subsequently occur and cause catastrophic damage. This is not restricted to hydrological events: examples appear in ecology (estimating the probability of species extinction), [Ludwig (1996)]; fish management (estimating the exhaustion of fisheries), [Hilborn and Mangel (1997)], [Malakov, (1999)]; insurance (calculating the probabilities of enormous claims, [Smith and Goodman, (2000)]; and geophysical sciences [Smith, (1989)]. Recently [Pinter et al. (2001)] have also emphasized the need for a reassessment of food hazards, highlighting the potential impact of the dynamical structure of rivers which have been modified due to different land usage patterns. Standard methodology for modelling extremes consists of adopting an asymptotic model to describe stochastic variation at extreme levels of a process, inference, and forecasting on the basis of the inferred model. Asymptotically motivated models remain the centrepiece of our modelling strategy, since without such an asymptotic basis, models have no rationale for extrapolation beyond the level of observed data. However, in this article we argue that there are two principal reasons why naive adoption of this paradigm leads to systematic underestimation of the probability of disastrous events. First, model and prediction uncertainty are often overlooked or ignored. Since such uncertainties can be substantial, designs made without allowance for these effects can be disastrously anti-conservative. Moreover, it is commonplace to reduce the dimensionality of extreme value models and to proceed on the basis of a simplified model. This procedure may lead to similar estimated models, but is likely to lead to overly confident measures of precision as the principal source of model uncertainty is artificially discarded. The second aspect is that of model homogeneity. We argue that a false assumption of a stationary model may also lead to considerable underestimation of the probability of a disastrously extreme event. While recent techniques, such as the local likelihood method of [Ramesh and Davison (2002)], identify local temporal

variation by parameter estimation at each individual time point via appropriate data weightings, we take a different approach. With no empirical evidence of temporal trends, we model non-stationary effects through the assumption of within-season homogeneity for model parameters within a number of seasons, whilst allowing for variations across seasons whose starting point and duration are treated as unknown. Our main departure from standard methodology is a preference for Bayesian inference. Though we also include classical likelihood analyses for comparison, we argue that the Bayesian approach provides a more coherent framework for keeping track of, and incorporating, all of the uncertainties involved in the prediction process. Computations for such models are intractable using conventional techniques, but are now almost routine using stochastic algorithms such as Markov chain Monte Carlo (MCMC).

Extremes are unusual or rare events and in classical data analysis often labelled as outliers and even ignored. For the events which do not happen very often for that one should apply extreme value theory. Earthquakes, hurricanes, and stock market crashes are surprising phenomena which do not follow any rule, but can be modelled by the distributions of extreme value models. For example, very high temperature can be fatal and is therefore more important than average temperature when assessing the consequences of climate changes. Policy makers are concerned with changes in the probability of extremes of climate parameters, i.e. extremely hot summer in the next coming decades. Therefore, the present investigation was carried out to assess and predict the changes in extreme maximum temperature of Ranchi.

Methodology

Distributional Study

The annual maximum of temperature was modelled using GEV distribution. Denoting daily observations by X_1, $X_{2,\ldots\ldots}$, the classical model for extremes is obtained by studying the behaviour of $M_n = \max [X_1, \ldots\ldots, X_n]$ for large values of n, where M_n corresponds naturally to the annual maximum. Asymptotic considerations suggest that the distribution of M_n should be approximately that of a member of the generalized extreme value distribution family [NERC,(1975)]; [Leadbetter et al., (1983)], for example, having distribution function

$$f(x;\mu,\sigma,k)=\frac{1}{\sigma}\left[1+k\left(\frac{x-\mu}{\sigma}\right)\right]^{-1-1/k}\exp\left\{-\left[1+k\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/k}\right\},\ k\neq 0$$

$$\text{and}\qquad f(x;\mu,\sigma,k)=\frac{1}{\sigma}\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\cdot\exp\left\{-\left[\left(\frac{x-\mu}{\sigma}\right)\right]\right\},\ k=0 \tag{1}$$

Where $\sigma>0$, $-\infty<\mu$, $k<\infty$ and the range of x is such that $[1+k(x-\mu)/\sigma]>0$.

The location, scale and shape of the distribution are controlled by the parameters μ, σ and k respectively.

For a random sample of annual maxima $x_1, \ldots, x_n$, the log likelihood is written as

$$l(\mu,\sigma,k)=-n.\log\sigma\left[-\left(\frac{1}{k}+1\right)\right]\sum_{i=1}^{n}\log\left[1+k\left(\frac{x_i-\mu}{\sigma}\right)\right]-\sum_{i=1}^{n}\left[1+k\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-1/k} \quad k\neq 0$$

$$\text{and,} \quad l(\mu,\sigma,0)=-n.\log\sigma-\sum_{i=1}^{n}\log\left[\left(\frac{x_i-\mu}{\sigma}\right)\right]-\sum_{i=1}^{n}\left[\left(\frac{x_i-\mu}{\sigma}\right)\right], \quad k=0 \tag{2}$$

This expression allows us to calculate the maximum likelihood estimates of the parameters. In the most applications the quartiles of the fitted distribution are to interest as they allows us to make predictions about the level exceeded once every 1/(1-T) years on the average. This is called the return level and is given by [Nadarajah and Choi, (2007)].

$$X_T=\mu-\frac{\sigma}{k}\left[1-\left\{-\log(1-\frac{1}{T})\right\}^{-k}\right], \quad k\neq 0$$

$$X_T=\mu-\sigma.\log\left\{-\log\left(1-\frac{1}{T}\right)\right\}, \quad k=0 \tag{3}$$

If L_1 is the maximum likelihood of three parameter distribution and L_2 is the maximum likelihood of the two parameter distribution, then the preferable model from the both fitted models can be judged by test statistic $\lambda = -2.\log(L_2/L_1)$, would be assumed to follow the χ^2 distribution with 1 degree of freedom (Wald, 1943). Since the number of parameters differs by 1 but it would be asymptotically true as the number of observations tend to infinity.

Model Incorporating Trend

The time series forecasting assumes that a time series is a combination of a pattern and some random error. The goal is to separate the pattern from the error by understanding the pattern trend, its long term increase and decrease and its seasonality. Several methods of time series forecasting are available such as the moving averages methods, linear regression with time, exponential smoothing, etc. This study concentrates on Exponential Smoothing technique as applied to extreme temperature time series data.

Single Exponential Smoothing

This is also known as simple exponential smoothing. Simple smoothing is used for short-range forecasting, usually just one step ahead into the future. The model assumes that the data fluctuates around a reasonably stable mean (no trend

or consistent pattern of growth). The following specific formula for simple exponential smoothing is given in [Makridakis et al, (2003)]:

$$F_t = \alpha * X_t + (1-\alpha) * F_{t-1} \tag{4}$$

Where X_t is the observation at time t, F_t is the forecasted value at time t and α is the damping factor having the range $0 < \alpha < 1$.

When applied recursively to each successive observation in the series, each new smoothed value (forecast) is computed as the weighted average of the current observation and the previous smoothed observation; the previous smoothed observation was computed in turn from the previous observed value and the smoothed value before the previous observation, etc.

Initial Value

The initial value of F_t, plays an important role in computing all the subsequent values. Setting it to x_1 is one method of initialization. Another possibility would be to average the first four or five observations. The smaller the value of α, the more important is the selection of the initial value of F_t.

Testing the homogeneity of Mean Square of Errors

Bartlett test carried out to check the homogeneity of error Mean Square of Error that whether they are homogeneous or not at various values of damping factor α. The test statistic of Bartlett test is given by the following formula [Gupta and Kapoor, (1970)]

$$\chi^2 = \sum_{i=1}^{d} \left(v_i \log \frac{S^2}{S_i} \right) \Big/ \left[1 + \frac{1}{3(d-1)} \left\{ \sum_{i=1}^{d} \left(\frac{1}{v_i} \right) - \frac{1}{v} \right\} \right] \tag{5}$$

Where $S^2 = \dfrac{\sum_{i=1}^{d} v_i S_i^2}{\sum_{i=1}^{d} v_i} = \dfrac{\sum_{i=1}^{d} v_i S_i^2}{v}, \quad \sum_{i=1}^{d} v_i = v$

In the above expression S_i^2 is the ith Mean Square of Error and v_i is the corresponding degree of freedom. The test statistic follows χ^2 distribution with (d-1) degree of freedom.

Data source

The data of daily basis maximum temperature for the years 1960 to 2006 were collected form Birsa Agricultural University, Ranchi (Jharkhand) and daily wise extreme maximum temperature has been extracted from the whole data set then

the analysis of that extreme maximum temperature has been carried out to further study.

Fitting of generalized extreme value distribution

The data are presented graphically in Fig.1, which does not give much insight to any long term change. The summary statistics given in Table-1 revealed that data showed slight longer tail on left side with more flatness, i.e. negatively skewed and platykurtic with slow variability. The standard deviation (S.D.) and standard error (S.E.) are 1.182 and 0.174 respectively.

Figure 1. Yearly extreme maximum temperature

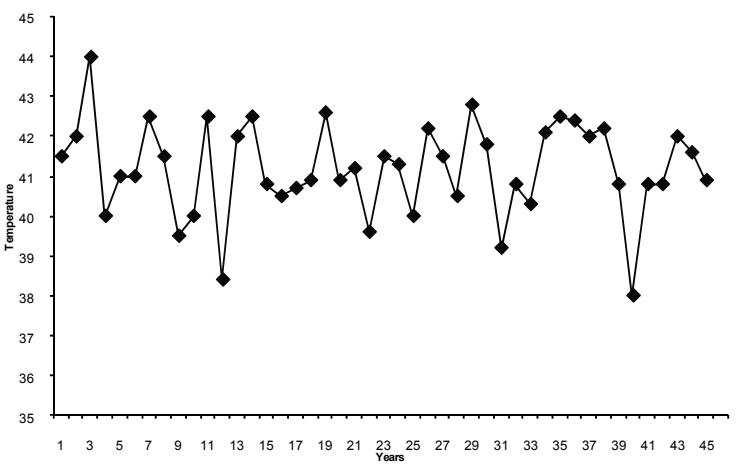


Table 1. Summary Statistics of extreme maximum temperature data

Statistics	Value
Sample Size	46
Range	6
Mean	41.213
Standard Deviation	1.182
Coefficient of Variation	2.870
Standard Error	0.174
Skew ness	-0.505
Kurtosis	0.661

Maximum likelihood estimates of the various parameters of GEV distribution with and without shape parameter are given in Table-2. The probability density function (PDF) and cumulative density function (CDF) of GEV distribution (only for $k \neq 0$) indicated in Fig.-2 and Fig.-3 respectively. From the both types of fitted model the GEV, when $k \neq 0$, is better due to likelihood ratio test failed, i.e. λ (Wald's statistics)= $0.6721 < \chi^2$ (3.84) at 1 degree of freedom. The bracketed values in table-2 are the S.E. of the respective parameters.

Table 2. The maximum likelihood estimates of GEV distribution

Estimated parameters		
Location ($\hat{\mu}$)	Scale($\hat{\sigma}$)	Shape ($\hat{k}$)
40.907 (0.2108)	1.2578 (0.1987)	-0.4707 (0.1404)
40.681 0.1497	0.9215 0.0570	–

Figure 2. PDF of the GEV distribution for Extreme Maximum Temperature

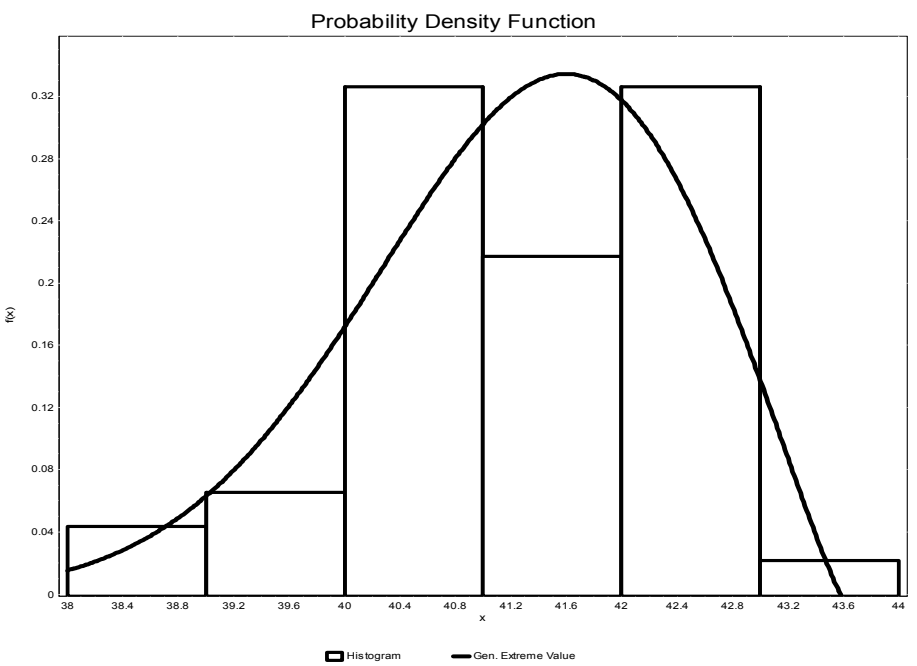


Figure 3. CDF of the GEV distribution for Extreme Maximum Temperature

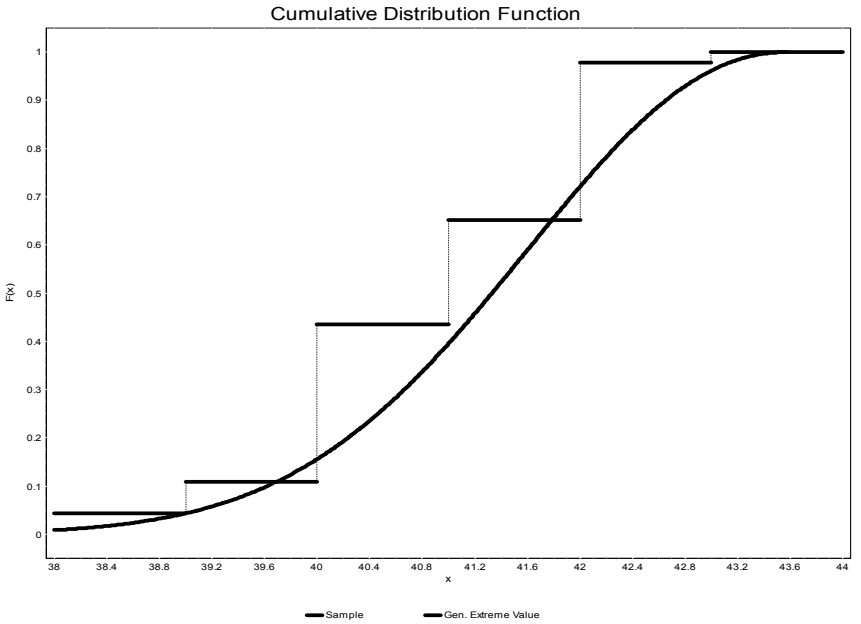
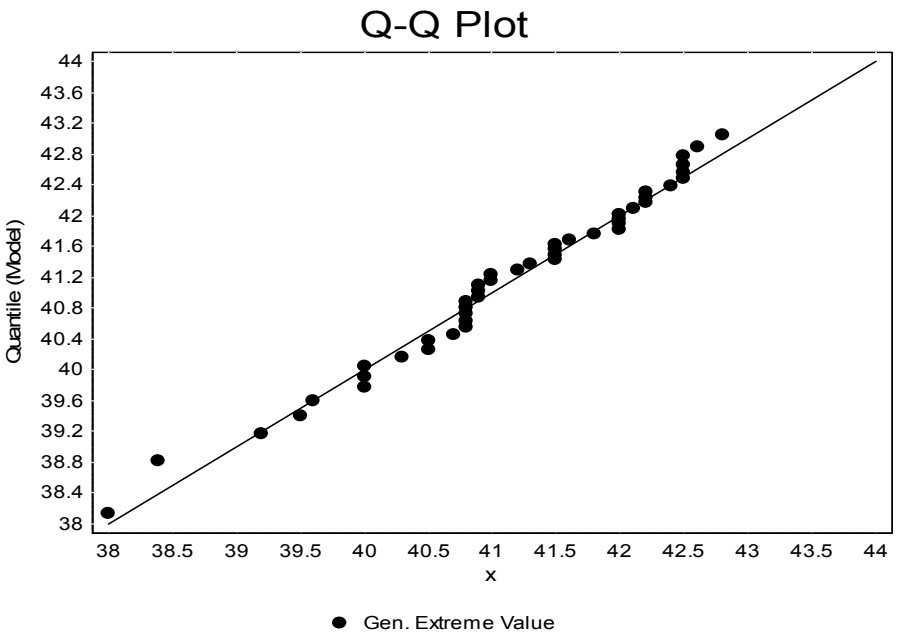


Table 3. Goodness of fit of the GEV Distribution model

Kolmogorov-Smironov Test				
Sample Size Statistic P-Value	46 0.08285 0.88448			
p	0.2	0.1	0.05	0.01
Critical Value	0.15457	0.17665	0.19625	0.23544
Null Hypothesis	Accept	Accept	Accept	Accept

The diagnosis of the model by the Q-Q plot indicated that quantiles are much closer to the diagonal reference line (Fig.4). A Q-Q plot is where the observed quantile is plotted against the quantile predicted by the fitted model. For example, to check the goodness of fit of model 1, we would plot the sorted values (in the ascending order) of the observed annual maximum daily rainfall versus expected quantiles Y_i determined by $F(Y_i) = (i - 0.375)/(n + 0.25)$ (Nadarajah and Choi, 2007), where F is the CDF of GEV distribution and n is the number of observations in the data series. Confirmation of good fitting is more emphasized by non-significant Kolmogorov-Smironov (K-S) test (Table-3), in which the null hypothesis of goodness of fit is accepted.

Figure 4. Q-Q plot of GEV Distribution model



Once the best model for the data is determined the next interest to derive the return levels of the Extreme Maximum Temperature in the summer seasons. The p year return level, say x_p, is the level exceeded on average only once in T years already indicated in equation (3). The return level for 10, 50, 100 and 200 years are consistently increasing towards long run in future indicated in the Table -4**.**

Table 4. Return level of maximum extreme temperature

	Return Level			
Year	T= 10	T= 50	T=100	T= 200
Temperature (°C)	42.90	43.30	43.37	43.43

Modeling by exponential smoothing of extreme maximum temperature

The asymptotic arguments which lead to the generalized Pareto model as a distribution of threshold excesses by an independent sequence can be generalized to consider sequences of dependent, though stationary, series. Under a weak condition that limits the effect of any long-range dependence, it can be shown that the generalized Pareto family has the same status as a limiting family for threshold excesses in this more general setting (see [Leadbetter et al, (1987)] for

example). This provides some justification for use of the generalized Pareto distribution as a model for extreme daily rainfalls, even if there is some temporal dependence in the series. The exponential smoothing was carried out at various value of damping factor (α). The Mean Square of Errors (MSE) and errors assumption (i.e. randomness and normality) and Bartlett test to check the homogeneity of Mean Square of Errors are illustrated in Table-5.

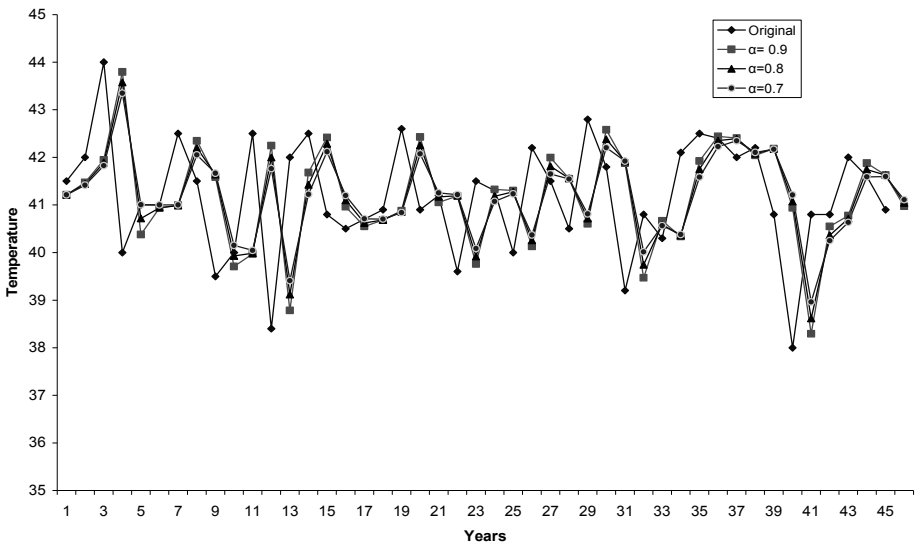
Table 5. Validation of the model at various values of α

α	MSE	Run Test (p = 5%)	K-S test (p = 5%)	Bartlett Test (χ^2) (p= 5%)
0.9	2.639	PASS	PASS	PASS
0.8	2.442	PASS	PASS	
0.7	2.272	PASS	PASS	

From Table-5, it is seen that hypothesis is successfully passed for all the values of α varying from 0.7 to 0.9 with an increment of 0.1 for both Run Test and K-S test. Bartlett test is also passed for Mean Square of Errors, i.e. there is no significant difference among the Mean Square of Errors and the predicted values by this technique are much closer to observed values when α = 0.9. Moreover, it is observed that the Mean Square of Errors is slightly higher for larger values of α but these are statistically same as indicated by Bartlett test (see Table-5).

The observed and predicted extreme of maximum temperature for different values of α is given in Fig.5. Form Fig. 5, it is clearly visible that when α = 0.9, the predicted values are much closer with observed values. Generally speaking, the more the value of α is the more the fitness of the predicated value with the observed value.

Figure 5. Observed and predicted values by exponential smoothing at various values of α



Conclusion

From the present study it is concluded that (i) the model is best fitted when the damping factor is closer to 1. (ii) The extreme maximum temperature gradually and slightly increases from time to time for the next 200 years. An overall look of the study shows that by using these techniques one can be risk-free of extreme temperature arising in the environmental disaster in the sectors like agriculture, forest fire, drought, etc. For further study, researcher can make a forecast at one step ahead whereas in distributional study one can predict the return level of extreme temperature for long run of time in the future.

Acknowledgement

The authors are thankful to the Birsa Agricultural University, Ranchi (Jharkhand) for providing necessary data by which the present investigation has been possible.

REFERENCES

D' ONOFRIO, E. E., FIORE, M. M. and ROMERO, S. I. (1999); Return periods of extreme water levels estimated for some vulnerable areas of Buenos Aires. *Continental Shelf Research* 19, 1681–1693.

DAVISON, A. C. and SMITH, R. L. (1990) Model for exceedances over high thresholds. Journal the Royal Statistical Society, Series B, 52, No. 3, 393–442.

DE MICHELE, C. and SALVADORI, G. (2005); Some hydrological applications of small sample estimators of Pareto and Extreme Value distributions, *Journal of Hydrology* 301, 37–53.

DEMOULIN, V. C. and ROEHREL, A. (2004); Extreme Value Theory can save your neck. *White Paper Jan-08, 2004* (Supported by Swiss National Science Foundation).

GUPTA, S. C., KAPOOR, V. K. (1970); Fundamental of Mathematical Statistics. *Chapter- 13*, pp-13.68.

HILBORN, R. and MANGEL, M. (1997) The ecological detective: Confronting models with data. Monographs on Population Biology. Princeton University Press.

HUANG, W., SUDONG XU and NNAJI, S. (2008); Evaluation of GEV model for frequency analysis of annual maximum water levels in the coast of United States, *Ocean Engineering*, doi: 10.1016/j.oceaneng.2008.04.010.

LEADBETTER, M.R., LINDGREN, G. and ROOTZEN, H. (1983); Extremes and related properties of random sequences and processes. Springer-Verlag, New York.

LUDWIG, D. (1996) Uncertainty and the assessment of extinction probabilities. Ecological Applications, 6 (4), pp. 1067–1076.

MALAKOV, D. (1999) Bayes offers a New Way to make sense of numbers. Science, Vol. 286, pp. 1460–1464.

MAKRIDAKIS, S., WHEELWRIGHT, S. C. and HYNDMAN, R. J. (2003); FORECASTING: Methods and Applications (Third Edition), Chapter-4, pp 136–180.

MORRISON, J. E. and SMITH, J. A. (2002); Stochastic modeling of flood peaks using the generalized extreme value distribution. *Water Resource Research* 38, 411–412.

NADARAJAH, S. and CHOI., D. (2007); Maximum daily rainfall in South Korea. *J. Earth Syst. Sci.* 116, No. 4, pp. 311–320.

PARISI, F. (2000); Special Report: Extreme Value Theory and Standard and Poor's Rating. *ABC research,* New York :(1) 212–438–2570.

PHEIN, H. N., FANG, T. S., (1989); Maximum likelihood estimation of the parameters and quartiles of the general extreme-value distribution for censored samples. *Journal of Hydrology* 105, 139–155.

PINTER N., THOMAS R. and WLOSINSKI H. (2001) Assessing ood hazards on dynamic rivers. EOS, Transactions, American Geophysical Union, Vol. 82, Number 31, 333 and 339–340.

RAMESH, N. I. AND DAVISON, A. C. (2002) Local models for exploratory analysis of hydrological extremes, Journal of Hydrology, 256, 106–119.

SMITH, R. L. (1989) Extreme value analysis of environmental times series: an example based on ozone data (with discussion), Statistical Science, 4, 367–393.

SMITH, R. L. and D. J. GOODMAN (2000) Bayesian risk analysis. Technical Report, Dept. Statistics, UNC-Chapel Hill.

WALD, A. (1943). Test of statistical hypothesis concerning several parameters, when the number of observation is large. Transactions of American Mathematical Society, 54:426–483.

STATISTICS IN TRANSITION-new series, July 2010
Vol. 11, No. 1, pp. 105—126

ESTIMATION OF POPULATION MEAN AT CURRENT OCCASION IN PRESENCE OF SEVERAL VARYING AUXILIARY VARIATES IN TWO-OCCASION SUCCESSIVE SAMPLING

G. N. Singh, Kumari Priyanka[1]

ABSTRACT

The present work intended to emphasize the role of several varying auxiliary variates at both the occasions to improve the precision of estimates at current occasion in two-occasion successive sampling. Two different efficient estimators are proposed and their theoretical properties are examined. Relative comparison of efficiencies of the proposed estimators with the sample mean estimator when there is no matching from previous occasion, and the optimum successive sampling estimator when no auxiliary information is used have been incorporated. Empirical studies are significantly justifying the composition of proposed estimators.

Key words: Successive sampling, varying auxiliary variates, optimum replacement policy, bias, mean square error.

1. Introduction

There are many practical instances, where the survey often needs to be repeated many times. The aim of a repeated survey is to allow one or more items to be monitored over time. For survey design purposes this aim has often been simplified to two objectives: good estimates of the item for each occasion, and good estimates of change from occasion to occasion. The follow-up of these objectives is achieved by mean of sampling on successive occasions according to a specified rule, with partial replacement of units, called successive sampling. A key issue is the extent to which elements sampled at a previous occasion should be retained in the sample selected at the current occasion; this is termed as optimum replacement policy.

[1] Department of Applied Mathematics, Indian School of Mines, Dhanbad-826 004, India, E-mail: gnsingh_ism@yahoo.com.

The problem of sampling on two successive occasions with a partial replacement of units was first considered by Jessen (1942). Estimation on more than two occasions is due to Patterson (1950). Patterson's theory was examined and extended by Eckler (1955), Rao and Graham (1964), Cochran (1977), Gupta (1979), Das (1982), Chaturvedi and Tripathi (1983), among others. Sen (1971) developed estimators for the population mean on the current occasion using information on two auxiliary variates available on previous occasion. Sen (1972, 73) extended his work for more than two auxiliary variates. In addition to the information from previous occasion, Singh et al. (1991) and Singh and Singh (2001) used an additional auxiliary information on current occasion for estimating the current population mean in two-occasion successive sampling. Singh (2003) extended this methodology for h-occasions successive sampling.

In many situations, information on an auxiliary variate may be readily available on the first as well as on the second occasion, for example tonnage (or seat capacity) of each vehicle or ship is known in survey sampling of transportation, number of polluting industries is known in environmental survey. Many other situations in biological (life) sciences could be explored to show the benefits of the present study. Utilizing the auxiliary information available for both the occasions Feng and Zou (1997), Biradar and Singh (2001), Singh (2005), Singh and Priyanka (2006, 2007, 2008 a, b) have proposed several chain-type ratio, difference and regression estimators for estimating the population mean at current (second) occasion in two-occasion successive sampling.

Following the works of Singh and Priyanka (2008 a, b), the objective of the present work is to propose estimators for estimating the population mean at current occasion using several varying auxiliary variates available at both the occasions. Chain-type difference and regression estimators are proposed, utilizing p-dynamic auxiliary variates. Practicability of the proposed estimator has been discussed. A relative comparison of efficiencies of the proposed estimators with the sample mean estimator when there is no matching from previous occasion, and the optimum successive sampling estimator when no auxiliary information has been used is discussed. Empirical studies show highly significant gains for the proposed estimators.

2. Formulation of Estimator

2.1. Notations

Let the character under study on the first (second) occasion be denoted by x (y). It is assumed that information on p (non negative integer constant) auxiliary variates z_{1j} (z_{2j}), j = 1, 2, - - -, *p* whose population means are known, are available on the first (second) occasion respectively. The auxiliary variates z_{1j} (z_{2j}), are correlated to x and y on the first and second occasions respectively. For convenience, it is assumed that the population under consideration is considerably

large enough. A simple random sample (without replacement) of n units is taken on the first occasion. A random sub-sample of m = n λ units is retained (matched) for its use on the second occasion, while a fresh simple random sample (without replacement) of u = (n - m) = nμ units is drawn on the second occasion from the non-sampled units of the population so that the sample size on the second occasion is also n. λ and μ are the non-negative fractions of matched and fresh samples respectively at the current (second) occasion such that λ+ μ =1. Hence onwards the following notations are employed:

$\bar{X}$, $\bar{Y}$: Population means of the study variate x and y respectively.

$\bar{Z}_{1j}$, $\bar{Z}_{2j}$: Population means of the jth (j = 1, 2, - - -, *p)* auxiliary variates at first and second occasion respectively.

$\bar{x}_n$, $\bar{x}_m$, $\bar{y}_u$, $\bar{y}_m$, $\bar{z}_{1nj}$, $\bar{z}_{1mj}$, $\bar{z}_{2mj}$, $\bar{z}_{2uj}$: Sample means of the respective variates of the sample sizes shown in suffices, where j = 1, 2, - - -, *p*.

ρ_{yx}, ρ_{xz1j}, ρ_{yz1j}, ρ_{z1jz2j}, ρ_{xz2j}, ρ_{yz2j}: Correlation coefficients between the variates shown in suffices, where j = 1, 2, - - -, *p*.

S_x^2, S_y^2, S_{z1j}^2, S_{z2j}^2: Population mean squares of x, y, z_{1j} and z_{2j} (j = 1, 2, - - -, *p*) respectively.

fpc: Finite population correction

2.2. Formulation of the Estimator T

To estimate the population mean $\bar{Y}$ on the second occasion, utilizing information on p varying auxiliary variates, two different estimators are suggested. One is a difference estimator based on sample of size u = (n μ) drawn afresh on the second occasion and given by

$$T_1 = \bar{y}_u + \sum_{j=1}^{p} \beta_{yz2j} \left(\bar{Z}_{2j} - \bar{z}_{2uj} \right) \tag{1}$$

Second estimator is a chain-type difference to difference estimator based on the sample of size m (= n λ) common with both the occasions and defined as

$$T_2 = \bar{y}_m^* + \beta_{yx} \left(\bar{x}_n^* - \bar{x}_m^* \right) \tag{2}$$

where $\bar{y}_m^* = \bar{y}_m + \sum_{j=1}^{p} \beta_{yz2j} \left(\bar{Z}_{2j} - \bar{z}_{2mj} \right)$

$$\bar{x}_n^* = \bar{x}_n + \sum_{j=1}^{p} \beta_{xz1j} \left(\bar{Z}_{1j} - \bar{z}_{1nj} \right)$$

$$\bar{x}_m^* = \bar{x}_m + \sum_{j=1}^{p} \beta_{xz1j}\left(\bar{Z}_{1j} - \bar{z}_{1mj}\right)$$

and β_{yx}, β_{xz1j} and β_{yz2j} (j = 1, 2, - - -, *p*) are known population regression coefficients between the variates shown in suffices. We assume that these population regression coefficients are known from past surveys or may be known through a pilot survey.

Combining the estimators T_1 and T_2, we have the final estimator of $\bar{Y}$ as

$$T = \varphi T_1 + \left(1 - \varphi\right) T_2 \tag{3}$$

where φ is an unknown constant to be determined so as to minimize the variance of the estimator T.

Remark 2.1: For estimating the mean on each occasion the estimator T_1 is suitable, which implies that more belief on T_1 could be shown by choosing φ as 1 (or close to 1), while for estimating the change from one occasion to the next, the estimator T_2 could be more useful so φ might be chosen as 0 (or close to 0). For asserting both the problems simultaneously, the suitable (optimum) choice of φ is required.

2.3. Estimator T in Practice

The main difficulty in using the estimator T, is the non-availability of population regression coefficients β_{yz2j}, β_{yx} and β_{xz1j}. Under such situations, the following working version of estimator T may be considered.

$$\Delta = \psi\, \Delta_1 + \left(1 - \psi\right) \Delta_2 \tag{4}$$

where $\Delta_1 = \bar{y}_u + \sum_{j=1}^{p} b_{yz2j}(u)\left(\bar{Z}_{2j} - \bar{z}_{2uj}\right)$ and $\Delta_2 = \bar{y}_m^{**} + b_{yx}(m)\left(\bar{x}_n^{**} - \bar{x}_m^{**}\right)$

where $\bar{y}_m^{**} = \bar{y}_m + \sum_{j=1}^{p} b_{yz2j}(m)\left(\bar{Z}_{2j} - \bar{z}_{2mj}\right)$

$$\bar{x}_n^{**} = \bar{x}_n + \sum_{j=1}^{p} b_{xz1j}(n)\left(\bar{Z}_{1j} - \bar{z}_{1nj}\right)$$

$$\bar{x}_m^{**} = \bar{x}_m + \sum_{j=1}^{p} b_{xz1j}(m)\left(\bar{Z}_{1j} - \bar{z}_{1mj}\right)$$

ψ is an unknown constant to be determined so that it minimizes the mean square error of the estimator Δ, $b_{yz2j}(u)$, $b_{yx}(m)$, $b_{yz2j}(m)$, $b_{xz1j}(n)$ and $b_{xz1j}(m)$

are the sample regression coefficients between the variates shown in suffices and based on the sample sizes indicated in the braces, where j = 1, 2, - - - , *p*.

2.4. Particular Cases

Case 1: If p = 1, estimators T and Δ defined in equations (3) and (4) respectively reduces to the estimators proposed by Priyanka (2008).

Case 2: If the auxiliary variates considered over the two-occasion are not changing much rapidly over time, i.e., $z_{1j} \approx z_{2j} \approx z_j$ (say), then the estimator T defined in equation (3) reduces to Singh and Priyanka (2008 b) estimator.

3. Properties of the Estimator T

Theorem 3.1: T is an unbiased estimator of $\bar{Y}$.

Proof: Since T_1 and T_2 are difference type estimators, they are unbiased for $\bar{Y}$. The final estimator T is a convex linear combination of T_1 and T_2, therefore, T is also an unbiased estimator of $\bar{Y}$.

Theorem 3.2: The variance of the estimator T (ignoring fpc) is given by

$$V(T) = \varphi^2 V(T_1) + (1 - \varphi)^2 V(T_2) \tag{5}$$

$$V(T_1) = \frac{1}{u}\left(1 - \sum_{j=1}^{p} \rho_{yz2j}^2 + \sum_{j \neq k=1}^{p} \rho_{yz2j}\, \rho_{yz2k}\, \rho_{z2jz2k}\right) S_y^2 \tag{6}$$

$$V(T_2) = \left[\frac{1}{m}\left(1 - \sum_{j=1}^{p} \rho_{yz2j}^2 + \sum_{j \neq k=1}^{p} \rho_{yz2j}\, \rho_{yz2k}\, \rho_{z2jz2k}\right)\right.$$

$$+ \left(\frac{1}{m} - \frac{1}{n}\right)\left(- \rho_{yx}^2 \left(1 + \sum_{j=1}^{p} \rho_{xz1j}^2\right) + 2\, \rho_{yx} \sum_{j=1}^{p} \left\{\rho_{xz1j}\rho_{yz1j} + \rho_{yz2j}\rho_{xz2j} - \rho_{yz2j}\rho_{xz1j}\rho_{z2jz1j}\right\}\right.$$

$$\left.\left. + \sum_{j \neq k=1}^{p} \left\{\rho_{yx}^2 \rho_{xz1j}\, \rho_{xz1k}\, \rho_{z1jz1k} - 2\rho_{yx}\rho_{yz2j}\, \rho_{xz1k}\, \rho_{z2jz1k}\right\}\right)\right] S_y^2 \tag{7}$$

Proof: It is clear that the variance of the estimator T is given by

$$V(T) = \varphi^2 V(T_1) + (1 - \varphi)^2 V(T_2) + 2\,\varphi(1 - \varphi)\mathrm{Cov}(T_1, T_2) \tag{8}$$

Variance and covariance terms of equation (8) can be derived as

$$V(T_1) = E\left[T_1 - \bar{Y}\right]^2 = \left(\frac{1}{u} - \frac{1}{N}\right)\left(1 - \sum_{j=1}^{p}\rho_{yz2j}^2 + \sum_{j \neq k=1}^{p} \rho_{yz2j}\ \rho_{yz2k}\ \rho_{z2jz2k}\right) S_y^2 \quad (9)$$

$$V(T_2) = E\left[T_2 - \bar{Y}\right]^2 = E\left[\left(\bar{y}_m^* - \bar{Y}\right) + \beta_{yx}\left\{\left(\bar{x}_n^* - \bar{X}\right) - \left(\bar{x}_m^* - \bar{X}\right)\right\}\right]^2$$

Taking expectation we get the variance of T_2 as

$$V(T_2) = \left[\left(\frac{1}{m} - \frac{1}{N}\right)\left(1 - \sum_{j=1}^{p}\rho_{yz2j}^2 + \sum_{j \neq k=1}^{p} \rho_{yz2j}\ \rho_{yz2k}\ \rho_{z2jz2k}\right)\right.$$

$$+ \left(\frac{1}{m} - \frac{1}{n}\right)\left(-\ \rho_{yx}^2\left(1 + \sum_{j=1}^{p}\rho_{xz1j}^2\right) + 2\ \rho_{yx}\sum_{j=1}^{p}\left\{\rho_{xz1j}\rho_{yz1j} + \rho_{yz2j}\rho_{xz2j} - \rho_{yz2j}\rho_{xz1j}\rho_{z2jz1j}\right\}\right.$$

$$\left.\left. + \sum_{j \neq k=1}^{p}\left\{\rho_{yx}^2\rho_{xz1j}\ \rho_{xz1k}\ \rho_{z1jz1k} - 2\rho_{yx}\rho_{yz2j}\ \rho_{xz1k}\ \rho_{z2jz1k}\right\}\right)\right] S_y^2 \quad (10)$$

Since the estimators T_1 and T_2 are unbiased estimators based on two independent samples of sizes u and m respectively,

$$\text{Cov}\ (T_1, T_2) = 0 \quad (11)$$

Further, we consider population size is sufficiently large, therefore, finite population corrections (fpc) are ignored.

Applying the above assumption in the equations (9) and (10), we get the expressions for the variances of the estimators T_1 and T_2 as given in equations (6) and (7) respectively.

Now, substituting the values of V (T_1), V (T_2) and Cov (T_1, T_2) in equation (8) we get the V (T) as in equation (5).

Since variance of the estimator T in equation (5) is a function of the unknown constant φ, it is minimized with respect to φ and subsequently the optimum value of φ is obtained as

$$\varphi_{opt.} = \frac{V(T_2)}{V(T_1) + V(T_2)} \quad (12)$$

Substituting this optimum value φ in equation (5) we obtain the minimum variance of T with respect to φ as

$$V(T)_{opt.} = \frac{V(T_1)V(T_2)}{V(T_1) + V(T_2)} \quad (13)$$

Further, substituting the values from equations (6) and (7) in equation (13), the simplified value of $V(T)_{opt.}$ is obtained as shown below in theorem 3.3.

Theorem 3.3: The $V(T)_{opt.}$ is derived as

$$V(T)_{opt.} = \frac{A[A + \mu B]}{n[A + \mu^2 B]} S_y^2 \tag{14}$$

Where $A = \left(1 - \sum_{j=1}^{p} \rho_{yz2j}^2 + \sum_{j \neq k=1}^{p} \rho_{yz2j}\ \rho_{yz2k}\ \rho_{z2jz2k}\right)$,

$$B = \left[- \rho_{yx}^2 \left(1 + \sum_{j=1}^{p} \rho_{xz1j}^2\right) + 2\ \rho_{yx} \sum_{j=1}^{p} \left\{\rho_{xz1j}\rho_{yz1j} + \rho_{yz2j}\rho_{xz2j} - \rho_{yz2j}\rho_{xz1j}\rho_{z2jz1j}\right\}\right.$$

$$\left. + \sum_{j \neq k=1}^{p} \left\{\rho_{yx}^2 \rho_{xz1j}\ \rho_{xz1k}\ \rho_{z1jz1k} - 2\rho_{yx}\rho_{yz2j}\ \rho_{xz1k}\ \rho_{z2jz1k}\right\}\right)\right], \text{ and } \mu\left(= \frac{u}{n}\right).$$

Corollary 3.1: If there is complete matching, i.e., $\mu = 0$ then

$$V(T)_{opt.} = \frac{A}{n} S_y^2 \tag{15}$$

Corollary 3.2: If there is no matching, i.e., $\mu = 1$ then

$$V(T)_{opt} = \frac{A}{n} S_y^2 \tag{16}$$

In both the cases $V(T)_{opt.}$ has the same value, this gives an implication that there must be an optimum choice of other than extreme values so that $V(T)_{opt.}$ will be smaller than the quantity given in equations (15) or (16). Thus, for making current estimate (neither the case of "complete matching" nor the case of "no matching") better, it is always preferable to replace the sample partially. For such situation a sensible strategy is to look for an optimum choice of μ, which is an important factor in reducing the cost of the survey.

Hence, $V(T)_{opt.}$ defined in equation (14) is minimized with respect to μ, resulting in a quadratic equation in μ, which shown as

$$B\,\mu^2 + 2\,A\,\mu - A = 0 \tag{17}$$

Solving equation (17) for μ, the solutions are given as

$$\hat{\mu} = \frac{-A \pm \sqrt{A(A+B)}}{B} \tag{18}$$

The real values of $\hat{\mu}$ exists if $A(A+B) \geq 0$. For any combination of the correlations that satisfies the above condition two real values of $\hat{\mu}$ are possible, however only the value of $\hat{\mu}$ that falls in the unit interval, $0 \leq \hat{\mu} \leq 1$ is admissible. Substituting this value of $\hat{\mu}$ (say μ_0) from equation (18) in equation (14), we have

$$V(T)_{opt.^*} = \frac{A[A+\mu_0 B]}{n[A+\mu_0^2 B]} S_y^2 \tag{19}$$

4. Properties of the Estimator Δ

Since Δ_1 and Δ_2 are the simple linear regression and chain type regression estimators respectively, they are biased for $\bar{Y}$. Therefore, the resulting estimator Δ defined in equation (4) is also a biased estimator of $\bar{Y}$. The bias B (.) and mean square error MSE (.) up to the first order of approximations and for large population size (ignoring fpc) are derived under large sample approximations, considering the following transformations:

$$\bar{y}_u = \bar{Y}(1+e_1),\ \bar{y}_m = \bar{Y}(1+e_2),\ \bar{x}_n = \bar{X}(1+e_3),\ \bar{x}_m = \bar{X}(1+e_4),$$

$$s_{yx}(m) = S_{yx}(1+e_5),\ s_x^2(m) = S_x^2(1+e_6),\ s_{yz2j}(u) = S_{yz2j}(1+e_{7j}),$$

$$s_{yz2j}(m) = S_{yz2j}(1+e_{7j}^*),\ s_{z2j}^2(u) = S_{z2j}^2(1+e_{8j}),\ s_{z2j}^2(m) = S_{z2j}^2(1+e_{8j}^*),$$

$$s_{xz1j}(n) = S_{xz1j}(1+e_{9j}),\ s_{xz1j}(m) = S_{xz1j}(1+e_{9j}^*),\ s_{z1j}^2(m) = S_{z1j}^2(1+e_{10j}),$$

$$s_{z1j}^2(n) = S_{z1j}^2(1+e_{10j}^*),\ \bar{z}_{2uj} = \bar{Z}_{2j}(1+e_{11j}),\ \bar{z}_{2mj} = \bar{Z}_{2j}(1+e_{12j}),$$

$\bar{z}_{1mj} = \bar{Z}_{1j}(1+e_{13j})$, $\bar{z}_{1nj} = \bar{Z}_{1j}(1+e_{14j})$ such that $|e_i| < 1$, $|e_{lj}| < 1$ and $|e_{kj}^*| < 1$ $\forall$ i = 1, 2, - - -, 6; l = 7, 8, 9, 10, 11, 12, 13, 14; k = 7, 8, 9, 10 and j= 1, 2, - - -, *p*.

Under the above transformations Δ_1 and Δ_2 take the following forms:

$$\Delta_1 = \left[\bar{Y}(1+e_1) - \sum_{j=1}^{p} \beta_{yz2j}\, \bar{Z}_{2j}\, e_{11j}(1+e_{7j})(1+e_{8j})^{-1}\right] \tag{20}$$

$$\Delta_2 = \left[\bar{Y}(1+e_2) + \beta_{yx}\bar{X}(e_3 - e_4 + e_3e_5 - e_4e_5 - e_3e_6 + e_4e_6)\right.$$

$$+ \sum_{j=1}^{p} \Big\{ - \beta_{yz2j} \bar{Z}_{2j} \left(e_{12j} - e_{12j} e^*_{8j} + e^*_{7j} e_{12j} \right)$$

$$+ \beta_{xz1j} \; \bar{Z}_{1j} \; \beta_{yx} \left(e_{13j} - e_{10j} e_{13j} + e_{9j} e_{13j} - e_{14j} - e^*_{9j} e_{14j} + e_{14j} e^*_{10j} \right.$$

$$\left. + e_5 e_{13j} - e_5 e_{14j} - e_6 e_{13j} + e_6 e_{14j} \right) \Big\} \Big] \tag{21}$$

Thus, we have the following theorems.

Theorem 4.1: $B(\Delta) = \psi \, B(\Delta_1) + (1 - \psi) B(\Delta_2)$ (22)

where $B(\Delta_1) = - \left(\frac{1}{u} \right) \sum_{j=1}^{p} \beta_{yz2j} \left[\frac{C_{0102}}{S_{yz2j}} - \frac{C_{0003}}{S^2_{z2j}} \right]$ (23)

and $B(\Delta_2) = \frac{1}{m} \sum_{j=1}^{p} \beta_{yz2j} \left[\frac{C_{00003}}{S^2_{z2j}} - \frac{C_{0102}}{S_{yz2j}} \right] + \left(\frac{1}{m} - \frac{1}{n} \right) \left[\beta_{yx} \left\{ \frac{C_{3000}}{S^2_x} - \frac{C_{2100}}{S_{yx}} \right\} + \right.$

$$\left. \beta_{yx} \sum_{j=1}^{p} \beta_{xz1j} \left\{ \frac{C_{1110}}{S_{yx}} - \frac{C_{2010}}{S^2_x} - \frac{C_{0030}}{S^2_{z1j}} + \frac{C_{1020}}{S_{xz1j}} \right\} \right] \tag{24}$$

where $C_{abcd} = E\left[\left(x_i - \bar{X} \right)^a \left(y_i - \bar{Y} \right)^b \left(z_{1ij} - \bar{Z}_{1j} \right)^c \left(z_{2ij} - \bar{Z}_{2j} \right)^d \right]$; ((a, b, c, d) ≥ 0 are integers), $j = 1, 2, - - -, p$.

Proof: $B(\Delta) = E\left[\Delta - \bar{Y} \right] = \psi B(\Delta_1) + (1 - \psi) B(\Delta_2)$

where

$$B(\Delta_1) = E\left[\Delta_1 - \bar{Y} \right]$$

$$= E\left[\bar{Y}(1 + e_1) - \sum_{j=1}^{p} \beta_{yz2j} \; \bar{Z}_{2j} \; e_{11j} \left(1 + e_{7j} \right) \left(1 + e_{8j} \right)^{-1} - \bar{Y} \right]$$

Expanding the right-hand side of the above expression binomially, taking expectations and collecting the terms up to the order $o(n^{-1})$, we have

$$B(\Delta_1) = - \left(\frac{1}{u} - \frac{1}{N} \right) \sum_{j=1}^{p} \beta_{yz2j} \left[\frac{C_{0102}}{S_{yz2j}} - \frac{C_{0003}}{S^2_{z2j}} \right] \tag{25}$$

Similarly,

$$B(\Delta_2) = E\left[\Delta_2 - \bar{Y} \right]$$

$$= E\Big[\bar{Y}(1+e_2) + \beta_{yx}\bar{X}(e_3 - e_4 + e_3e_5 - e_4e_5 - e_3e_6 + e_4e_6)$$

$$+ \sum_{j=1}^{p}\Big\{-\beta_{yz2j}\bar{Z}_{2j}\left(e_{12j} - e_{12j}e_{8j}^{*} + e_{7j}^{*}e_{12j}\right)$$

$$+ \beta_{xz1j}\ \bar{Z}_{1j}\ \beta_{yx}\left(e_{13j} - e_{10j}e_{13j} + e_{9j}e_{13j} - e_{14j} - e_{9j}^{*}e_{14j}\right.$$

$$\left. + e_{14j}e_{10j}^{*} + e_5e_{13j} - e_5e_{14j} - e_6e_{13j} + e_6e_{14j}\right)\Big\} - \bar{Y}\Big]$$

Again, expanding the right-hand side of the above expression binomially, taking expectations and retaining the terms up to the first order of approximations, we have

$$B(\Delta_2) = \left(\frac{1}{m} - \frac{1}{N}\right)\sum_{j=1}^{p}\beta_{yz2j}\left[\frac{C_{00003}}{S_{z2j}^2} - \frac{C_{0102}}{S_{yz2j}}\right] + \left(\frac{1}{m} - \frac{1}{n}\right)\left[\beta_{yx}\left\{\frac{C_{3000}}{S_x^2} - \frac{C_{2100}}{S_{yx}}\right\}\right.$$

$$\left. + \beta_{yx}\sum_{j=1}^{p}\beta_{xz1j}\left\{\frac{C_{1110}}{S_{yx}} - \frac{C_{2010}}{S_x^2} - \frac{C_{0030}}{S_{z1j}^2} + \frac{C_{1020}}{S_{xz1j}}\right\}\right] \tag{26}$$

Ignoring fpc, for sufficiently large population size in equations (25) and (26) we get the expressions for the bias of the estimators Δ, Δ_1 and Δ_2 up to the first order of approximations as shown in equations (22), (23) and (24) respectively.

Theorem 4.2: The mean square error of the estimator Δ of the population mean $\bar{Y}$, to the first order of approximations and ignoring fpc, is given by

$$M(\Delta) = \psi^2 M(\Delta_1) + (1-\psi)^2 M(\Delta_2) \tag{27}$$

where $M(\Delta_1) = \frac{1}{u} A\, S_y^2$ (28)

and $M(\Delta_2) = \left[\frac{1}{m}A + \left(\frac{1}{m} - \frac{1}{n}\right)B\right]S_y^2$ (29)

Proof: By the definition of mean square error we have

$$M(\Delta) = E\left[\Delta - \bar{Y}\right]^2 = E\left[\psi\left(\Delta_1 - \bar{Y}\right) + (1-\psi)\left(\Delta_2 - \bar{Y}\right)\right]^2$$

$$= \psi^2 M(\Delta_1) + (1-\psi)^2 M(\Delta_2) + 2\psi(1-\psi)E\left[\left(\Delta_1 - \bar{Y}\right)\left(\Delta_2 - \bar{Y}\right)\right] \tag{30}$$

where $M(\Delta_1) = E\left[\Delta_1 - \bar{Y}\right]^2$ and $M(\Delta_2) = E\left[\Delta_2 - \bar{Y}\right]^2$

Now, using the expressions given in equations (20) and (21), expanding binomially, taking expectations, retaining the terms up to the first order of approximations, we have the following results

$$\mathrm{MSE}(\Delta_1) = \left(\frac{1}{u} - \frac{1}{N}\right) A\, S_y^2 \tag{31}$$

$$\mathrm{MSE}(\Delta_2) = \left[\left(\frac{1}{m} - \frac{1}{N}\right) A + \left(\frac{1}{m} - \frac{1}{n}\right) B\right] S_y^2 \tag{32}$$

Now, ignoring fpc, we have

$$\mathrm{MSE}(\Delta_1) = \frac{1}{u} A\, S_y^2 \tag{33}$$

$$\mathrm{MSE}(\Delta_2) = \left[\frac{1}{m} A + \left(\frac{1}{m} - \frac{1}{n}\right) B\right] S_y^2 \tag{34}$$

Since the estimators Δ_1 and Δ_2 are biased estimators based on two independent samples of sizes u and m respectively, for large population size we have

$$E\left[\left(\Delta_1 - \bar{Y}\right)\left(\Delta_2 - \bar{Y}\right)\right] = 0 \tag{35}$$

Substituting the values of $\mathrm{MSE}(\Delta_1)$, $\mathrm{MSE}(\Delta_2)$ and $E\left[\left(\Delta_1 - \bar{Y}\right)\left(\Delta_2 - \bar{Y}\right)\right]$ from equations (33), (34) and (35) into equation (30) we get the value of $\mathrm{MSE}(\Delta)$ as shown in equation (27).

Remark 4.1 From equation (27), it is visible that the mean square error up to the first order of approximations of the estimator Δ is exactly similar to that of the variance of the estimator T in equations (5). This is one of the positive aspects of the estimator Δ. The estimator Δ is based on sample estimates and up to the first order of approximations; it is equally precise to that of estimator T.

5. Special Case

When the p-auxiliary variates are mutually uncorrelated, i.e., $\rho_{zijzik} = 0\ \forall\ i = 1, 2; j \neq k = 1, 2, - - -, p$ then the expression for the optimum value of μ and $V(T)_{opt.^*}$ reduces to

$$\hat{\mu} = \frac{-A^* \pm \sqrt{A^*(A^* + B^*)}}{B^*} \tag{36}$$

and

$$V(T)_{opt.^*} = \frac{A^*\left[A^* + \mu_0 B^*\right]}{n\,[A^* + \mu_0^2 B^*]} S_y^2 \tag{37}$$

where $A^* = 1 - \sum_{j=1}^{p} \rho_{yz2j}^2$

and $B^* = -\rho_{yx}^2\left(1 + \sum_{j=1}^{p} \rho_{xz1j}^2\right) + 2\rho_{yx}\sum_{j=1}^{p}\left(\rho_{xz1j}\,\rho_{yz1j} + \rho_{yz2j}\,\rho_{xz2j} - \rho_{yz2j}\,\rho_{xz1j}\,\rho_{z2jz1j}\right)$.

6. Efficiency Comparison

The percent relative efficiencies of T with respect to (i) $\bar{y}_n$, when there is no matching and (ii) $\hat{\bar{Y}} = \varphi^* \bar{y}_u + \left(1 - \varphi^*\right)\bar{y}'_m$, when no additional auxiliary information was used at any occasion, where $\bar{y}'_m = \bar{y}_m + \beta_{yx}\left(\bar{x}_n - \bar{x}_m\right)$, have been obtained for different choices of the correlations involved. Since, $\bar{y}_n$ and $\hat{\bar{Y}}$ are unbiased estimators of $\bar{Y}$, following on the line of Sukhatme et al. (1984) the variance of $\bar{y}_n$ and the optimum variance of $\hat{\bar{Y}}$ for large N are respectively given by

$$V(\bar{y}_n) = \frac{S_y^2}{n} \tag{38}$$

$$V\left(\hat{\bar{Y}}\right)_{opt.^*} = \left[1 + \sqrt{\left(1 - \rho_{yx}^2\right)}\right]\frac{S_y^2}{2n} \tag{39}$$

The percent relative efficiencies E_1 and E_2 of T (under optimal condition) with respect to $\bar{y}_n$ and $\hat{\bar{Y}}$ respectively are given by

$$E_1 = \frac{V(\bar{y}_n)}{V(T)_{opt.^*}} \times 100 \quad \text{and} \quad E_2 = \frac{V\left(\hat{\bar{Y}}\right)_{opt.^*}}{V(T)_{opt.^*}} \times 100.$$

6.1.Empirical Study

The expressions of the optimum μ (i.e. μ_0) and the percent relative efficiencies E_1 and E_2 are in terms of population correlation coefficients. Therefore, the values of μ_0, E_1 and E_2 have been computed for different choices of positive correlations. For empirical studies, cases of p = 1 and 2 auxiliary variates have been considered.

Case 1: For p = 1, the values of A and B takes the form $A = 1 - \rho^2_{yz21}$ and $B = -\rho^2_{yx}\left(1 + \rho^2_{xz11}\right) + 2\rho_{yx}\left(\rho_{xz11}\rho_{yz11} + \rho_{xz21}\rho_{yz21} - \rho_{yz21}\rho_{xz11}\rho_{z21z11}\right)$. For convenience we assume $\rho_{xz11} = \rho_{xz21} = \rho_{yz11} = \rho_{yz21} = \rho_1$. Hence, the values of A and B becomes $A = 1 - \rho^2_1$ and $B = -\rho^2_{yx}\left(1 + \rho^2_1\right) + 2\rho_{yx}\rho^2_1\left(2 - \rho_{z21z11}\right)$, which is the work of Priyanka (2008).

Case 2: For p = 2 and assuming that the two auxiliary variates are mutually correlated i.e., $\rho_{zijzik} \neq 0$. The values of A and B are given by

$$A = 1 - \rho^2_{yz21} - \rho^2_{yz22} + 2\rho_{yz21}\rho_{yz22}\rho_{z22z21}$$

and

$$B = -\rho^2_{yx}\left(1 + \rho^2_{xz11} + \rho^2_{xz12}\right) + 2\rho_{yx}\left(\rho_{xz11}\rho_{yz11} + \rho_{yz21}\rho_{xz21} - \rho_{yz21}\rho_{xz11}\rho_{z21z11}\right.$$

$$\left. + \rho_{xz12}\rho_{yz12} + \rho_{yz22}\rho_{xz22} - \rho_{yz22}\rho_{xz12}\rho_{z22z12}\right)$$

$$+ 2\rho^2_{yx}\rho_{xz11}\rho_{xz12}\rho_{z11z12} - 4\rho_{yx}\rho_{yz21}\rho_{xz12}\rho_{z21z12}$$

For convenience we assume

$$\rho_{xz11} = \rho_{xz21} = \rho_{yz11} = \rho_{yz21} = \rho_2,$$

$$\rho_{xz12} = \rho_{xz22} = \rho_{yz12} = \rho_{yz22} = \rho_3$$

and

$$\rho_{z11z12} = \rho_{z21z22} = \rho_{z11z22} = \rho_{z12z21} = \rho_4.$$

In the light of these assumptions, the values of A and B are given as

$$A = 1 - \rho^2_2 - \rho^2_3 + 2\rho_2\rho_3\rho_4$$

and

$$B = -\rho^2_{yx}\left(1 + \rho^2_2 + \rho^2_3 - 2\rho_2\rho_3\rho_4\right) + 2\rho_{yx}\left(2\rho^2_2 - \rho^2_2\rho_{z21z11} + 2\rho^2_3 - \rho^2_3\rho_{z22z12} - 2\rho_2\rho_3\rho_4\right).$$

Substituting these values of A and B in equations (18) and (19), we have the values of optimum μ and $V(T)_{opt.^*}$ respectively. For different choices of correlations, Tables 1-3 show the optimum values of i.e., μ_0 and percent relative efficiencies E_1 and E_2 of T (under optimal condition) with respect to $\bar{y}_n$ and $\hat{\bar{Y}}$ respectively.

Table 1. Optimum values of μ and Percent Relative Efficiencies of T with respect to $\bar{y}_n$ and $\hat{\bar{Y}}$

$\rho_{z21z11} = 0.3$, $\rho_{z22z12} = 0.4$											
ρ_4			0.1			0.3			0.5		
ρ_3	ρ_2	ρ_{yx} ↓	μ_0	E_1	E_2	μ_0	E_1	E_2	μ_0	E_1	E_2
0.4	0.5	0.3	0.46	144.9	141.7	0.47	131.4	128.3	0.47	120.1	117.4
		0.6	0.45	144.4	29.92	0.47	131.8	118.6	0.48	121.4	109.2
		0.9	0.49	156.1	112.0	0.51	142.6	102.4	0.52	131.5	**
	0.7	0.3	0.40	199.1	194.5	0.43	164.6	160.8	0.44	140.8	137.5
		0.6	0.38	187.7	168.9	0.41	157.6	141.8	0.41	136.5	122.8
		0.9	0.39	193.3	138.7	0.42	162.8	116.9	0.45	141.4	101.5
	0.9	0.3	0.25	500.6	489.1	0.34	275.3	269.0	0.38	196.6	192.1
		0.6	0.22	430.1	387.1	0.31	245.5	220.9	0.35	179.8	161.8
		0.9	0.21	419.8	301.4	0.30	241.6	173.5	0.35	178.1	127.9
0.6	0.5	0.3	0.42	185.6	181.3	0.44	153.9	150.3	0.45	131.8	128.7
		0.6	0.40	177.3	159.6	0.43	149.2	134.3	0.45	129.4	116.5
		0.9	0.42	184.8	132.7	0.44	156.0	112.0	0.47	135.7	**
	0.7	0.3	0.34	291.9	285.2	0.39	195.9	191.4	0.43	149.5	146.0
		0.6	0.30	262.7	236.4	0.36	181.7	163.5	0.40	141.7	127.5
		0.9	0.31	262.7	188.6	0.37	182.86	131.3	0.41	143.4	103.0
	0.9	0.3	*	–	–	0.28	370.7	362.1	0.37	200.1	195.5
		0.6	*	–	–	0.25	319.7	287.7	0.33	180.2	162.2
		0.9	*	–	–	0.24	308.2	221.3	0.33	175.8	126.2
0.8	0.5	0.3	0.32	338.1	330.3	0.38	217.5	212.5	0.42	163.1	159.4
		0.6	0.28	301.3	271.2	0.35	200.0	180.0	0.39	153.5	138.2
		0.9	0.29	300.7	215.9	0.35	200.9	144.3	0.39	155.1	111.4
	0.7	0.3	*	–	–	0.31	307.7	300.6	0.39	180.6	176.4
		0.6	*	–	–	0.28	270.2	243.2	0.35	165.1	148.5
		0.9	*	–	–	0.27	263.1	188.9	0.35	162.4	116.6
	0.9	0.3	*	–	–	*	–	–	0.33	243.7	238.1
		0.6	*	–	–	*	–	–	0.29	212.7	191.4
		0.9	*	–	–	*	–	–	0.27	203.0	145.7

Note: ** denotes μ_0 does not exist and ** indicate no gain.*

Table 2. Optimum values of μ and Percent Relative Efficiencies of T with respect to $\bar{y}_n$ and $\hat{\bar{Y}}$

$\rho_{z21z11} = 0.7$, $\rho_{z22z12} = 0.5$											
ρ_4			0.1			0.3			0.5		
ρ_3	ρ_2	$\rho_{yx} \downarrow$	μ_0	E_1	E_2	μ_0	E_1	E_2	μ_0	E_1	E_2
0.4	0.5	0.3	0.47	148.2	144.8	0.48	134.1	131.0	0.48	122.5	119.6
		0.6	0.48	150.9	135.8	0.49	137.6	123.8	0.50	126.5	113.8
		0.9	0.54	170.7	122.6	0.55	155.6	111.7	0.57	143.1	102.8
	0.7	0.3	0.42	208.4	203.6	0.44	171.6	167.6	0.46	146.2	142.8
		0.6	0.41	203.7	183.3	0.44	170.1	153.1	0.46	146.7	132.0
		0.9	0.45	223.1	160.2	0.48	186.5	133.9	0.51	161.0	115.6
	0.9	0.3	0.29	552.2	539.5	0.37	298.1	291.2	0.41	210.5	205.6
		0.6	0.25	497.2	447.5	0.34	278.4	250.6	0.39	201.5	181.4
		0.9	0.27	525.1	377.0	0.36	294.3	211.3	0.42	213.6	153.4
0.6	0.5	0.3	0.43	190.9	186.5	0.45	157.9	154.2	0.46	134.9	131.8
		0.6	0.42	186.7	168.0	0.45	156.6	140.9	0.47	135.5	121.9
		0.9	0.46	202.4	145.3	0.48	170.0	122.1	0.51	147.3	105.8
	0.7	0.3	0.36	308.9	301.7	0.41	205.4	200.7	0.44	155.7	152.1
		0.6	0.34	287.6	258.8	0.40	196.9	177.2	0.43	152.5	137.3
		0.9	0.35	303.6	218.0	0.42	208.5	149.7	0.46	162.0	116.3
	0.9	0.3	*	–	–	0.31	405.1	395.8	0.40	214.5	209.5
		0.6	*	–	–	0.28	364.4	328.0	0.37	201.5	181.4
		0.9	*	–	–	0.29	373.8	268.4	0.38	208.0	149.4
0.8	0.5	0.3	0.33	352.4	344.3	0.39	225.1	219.9	0.43	168.1	164.2
		0.6	0.31	321.3	289.2	0.37	211.7	190.6	0.41	161.7	145.5
		0.9	0.32	332.3	238.6	0.38	219.9	157.9	0.43	168.6	121.1
	0.7	0.3	*	–	–	0.34	326.7	319.1	0.41	189.4	185.1
		0.6	*	–	–	0.30	295.8	266.2	0.38	178.5	160.6
		0.9	*	–	–	0.31	300.8	215.9	0.39	182.8	131.3
	0.9	0.3	*	–	–	*	–	–	0.36	263.1	257.0
		0.6	*	–	–	*	–	–	0.32	238.4	214.6
		0.9	*	–	–	*	–	–	0.32	238.6	171.3

Note: ** denotes* μ_0 *does not exist.*

Table 3. Optimum values of μ and Percent Relative Efficiencies of T with respect to $\bar{y}_n$ and $\hat{\bar{Y}}$

$\rho_{z21z11} = 0.9$, $\rho_{z22z12} = 0.8$											
ρ_4			0.1			0.3			0.5		
ρ_3	ρ_2	ρ_{yx} ↓	μ_0	E_1	E_2	μ_0	E_1	E_2	μ_0	E_1	E_2
0.4	0.5	0.3	0.48	151.1	147.7	0.48	136.6	133.4	0.49	124.6	121.7
		0.6	0.49	157.5	141.8	0.51	143.3	129.0	0.52	131.6	118.4
		0.9	0.59	189.3	135.9	0.61	171.9	123.4	0.62	157.6	113.1
	0.7	0.3	0.44	215.8	210.9	0.46	177.0	173.0	0.47	150.4	146.9
		0.6	0.44	218.2	196.4	0.47	181.4	163.2	0.49	155.8	140.2
		0.9	0.53	259.1	186.0	0.56	214.6	154.1	0.58	183.9	132.0
	0.9	0.3	0.30	594.9	581.2	0.39	316.0	308.8	0.43	221.1	216.0
		0.6	0.29	562.0	505.8	0.38	308.9	278.0	0.43	221.2	199.1
		0.9	0.34	666.6	478.6	0.44	361.0	259.1	0.50	256.9	184.4
0.6	0.5	0.3	0.44	197.9	193.3	0.46	163.0	159.2	0.48	138.8	135.6
		0.6	0.46	200.3	180.3	0.48	167.2	150.5	0.50	144.1	129.7
		0.9	0.53	234.7	168.5	0.56	195.5	140.3	0.58	168.1	120.7
	0.7	0.3	0.38	327.3	319.8	0.43	215.4	210.4	0.46	162.2	158.4
		0.6	0.37	318.6	286.7	0.43	215.4	193.9	0.47	165.4	148.9
		0.9	0.43	370.4	265.9	0.50	248.8	178.6	0.54	190.6	136.8
	0.9	0.3	*	–	–	0.34	439.8	429.7	0.42	228.2	222.9
		0.6	*	–	–	0.32	417.0	375.3	0.42	225.4	202.9
		0.9	*	–	–	0.37	477.0	342.5	0.47	255.7	183.6
0.8	0.5	0.3	0.36	378.1	369.3	0.42	238.4	232.9	0.45	176.5	172.4
		0.6	0.34	361.6	325.4	0.41	234.8	211.3	0.45	177.5	159.8
		0.9	0.39	412.6	296.2	0.47	266.4	191.3	0.51	201.1	144.4
	0.7	0.3	*	–	–	0.36	353.7	345.6	0.43	201.4	196.8
		0.6	*	–	–	0.35	337.5	303.8	0.43	199.6	179.6
		0.9	*	–	–	0.39	379.4	272.4	0.48	223.5	160.4
	0.9	0.3	*	–	–	*	–	–	0.39	286.6	280.0
		0.6	*	–	–	*	–	–	0.37	275.0	247.5
		0.9	*	–	–	*	–	–	0.41	303.7	218.1

Note: ** denotes* μ_0 *does not exist.*

Case3: For p = 2 and assuming that the two auxiliary variates are mutually uncorrelated i.e., $\rho_{zijzik} = 0$. The values of A^* and B^* are given by

$$A^* = 1 - \rho_2^2 - \rho_3^2$$

and $B^* = -\rho_{yx}^2\left(1+\rho_2^2+\rho_3^2\right)+2\rho_{yx}\left(2\rho_2^2-\rho_2^2\rho_{z21z11}+2\rho_3^2-\rho_3^2\rho_{z22z12}\right)$

Substituting A^* and B^* in equations (36) and (37), Tables 4 – 6, show the optimum values of μ, percent relative efficiencies E_1 and E_2 for different choices of correlations under the present assumption.

Table 4. Optimum values of μ and Percent Relative Efficiencies of T with respect to $\bar{y}_n$ and $\hat{\bar{Y}}$ when auxiliary variates are uncorrelated

$\rho_{z22z12} = 0.3$											
ρ_{z21z11}			0.2			0.5			0.8		
ρ_3	ρ_2	$\rho_{yx}\downarrow$	μ_0	E_1	E_2	μ_0	E_1	E_2	μ_0	E_1	E_2
0.4	0.5	0.3	0.45	151.8	148.3	0.45	154.0	150.5	0.46	156.4	152.7
		0.6	0.44	149.5	134.5	0.45	153.6	138.3	0.47	158.3	142.5
		0.9	0.47	159.5	114.5	0.50	167.9	120.5	0.53	178.4	128.1
	0.7	0.3	0.38	219.9	214.9	0.40	227.0	221.8	0.41	235.0	229.6
		0.6	0.36	203.7	183.3	0.38	214.8	193.3	0.40	228.5	205.7
		0.9	0.36	206.6	148.4	0.39	225.3	161.8	0.44	252.1	181.0
	0.9	0.3	0.16	1033.8	1010.0	0.17	1114.3	1088.6	0.18	1219.0	1191.0
		0.6	0.13	852.1	766.9	0.14	943.0	848.7	0.16	1073.9	966.5
		0.9	0.12	814.6	584.8	0.14	940.9	675.5	0.17	1161.8	834.1
0.6	0.5	0.3	0.40	204.8	200.1	0.41	208.0	203.2	0.41	211.4	206.6
		0.6	0.37	192.2	173.0	0.39	197.5	177.7	0.40	203.2	182.9
		0.9	0.38	197.3	197.3	0.40	206.2	148.0	0.42	216.8	155.6
	0.7	0.3	0.29	389.6	380.7	0.30	404.4	395.0	0.32	421.2	411.5
		0.6	0.26	340.4	306.3	0.27	359.5	323.6	0.29	382.9	344.6
		0.9	0.25	334.2	239.9	0.27	362.5	260.2	0.30	400.9	287.8
	0.9	0.3	*	–	–	*	–	–	*	–	–
		0.6	*	–	–	*	–	–	*	–	–
		0.9	*	–	–	*	–	–	*	–	–
0.8	0.5	0.3	0.26	478.4	467.4	0.27	487.9	476.7	0.27	498.1	486.6
		0.6	0.23	412.8	371.5	0.23	424.6	382.1	0.24	437.5	393.8
		0.9	0.22	404.0	290.1	0.23	420.9	302.2	0.24	440.3	316.1
	0.7	0.3	*	–	–	*	–	–	*	–	–
		0.6	*	–	–	*	–	–	*	–	–
		0.9	*	–	–	*	–	–	*	–	–
	0.9	0.3	*	–	–	*	–	–	*	–	–
		0.6	*	–	–	*	–	–	*	–	–
		0.9	*	–	–	*	–	–	*	–	–

Note: ** denotes* μ_0 *does not exist.*

Table 5. Optimum values of μ and Percent Relative Efficiencies of T with respect to $\bar{y}_n$ and $\hat{\bar{Y}}$ when auxiliary variates are uncorrelated

$\rho_{z22z12} = 0.5$											
ρ_{z21z11}			0.2			0.5			0.8		
ρ_3	ρ_2	$\rho_{yx} \downarrow$	μ_0	E_1	E_2	μ_0	E_1	E_2	μ_0	E_1	E_2
0.4	0.5	0.3	0.45	152.7	149.2	0.46	154.9	151.4	0.46	157.3	153.7
		0.6	0.45	151.2	136.1	0.46	155.6	140.0	0.47	160.5	144.4
		0.9	0.48	162.9	116.9	0.51	172.1	123.6	0.54	183.9	132.0
	0.7	0.3	0.39	221.4	216.3	0.40	228.7	223.4	0.41	236.9	231.4
		0.6	0.36	205.9	185.3	0.38	217.6	195.8	0.41	232.0	208.8
		0.9	0.37	210.2	150.9	0.40	230.2	165.3	0.45	259.8	186.5
	0.9	0.3	0.16	1043.3	1019.3	0.17	1126.4	1100.5	0.19	1235.3	1206.9
		0.6	0.13	862.5	776.3	0.14	957.4	861.7	0.16	1095.9	986.3
		0.9	0.12	828.2	594.6	0.14	962.6	691.1	0.18	1205.3	865.4
0.6	0.5	0.3	0.41	207.9	203.1	0.41	211.3	206.4	0.42	214.9	210.0
		0.6	0.38	197.2	177.5	0.40	203.0	182.7	0.41	209.4	188.5
		0.9	0.40	205.8	147.8	0.42	216.3	155.3	0.45	229.1	164.5
	0.7	0.3	0.30	396.6	387.5	0.31	412.3	402.8	0.32	430.4	420.5
		0.6	0.26	349.3	314.4	0.28	370.4	333.3	0.30	396.5	356.8
		0.9	0.26	347.1	249.2	0.28	379.7	272.6	0.32	425.7	305.6
	0.9	0.3	*	–	–	*	–	–	*	–	–
		0.6	*	–	–	*	–	–	*	–	–
		0.9	*	–	–	*	–	–	*	–	–
0.8	0.5	0.3	0.27	495.0	483.6	0.28	505.7	494.1	0.28	517.2	505.3
		0.6	0.24	433.6	390.2	0.25	447.6	402.8	0.25	463.1	416.8
		0.9	0.24	434.3	311.8	0.25	455.9	327.3	0.26	481.6	345.8
	0.7	0.3	*	–	–	*	–	–	*	–	–
		0.6	*	–	–	*	–	–	*	–	–
		0.9	*	–	–	*	–	–	*	–	–
	0.9	0.3	*	–	–	*	–	–	*	–	–
		0.6	*	–	–	*	–	–	*	–	–
		0.9	*	–	–	*	–	–	*	–	–

Note: ** denotes* μ_0 *does not exist.*

Table 6. Optimum values of μ and Percent Relative Efficiencies of T with respect to $\bar{y}_n$ and $\hat{\bar{Y}}$ when auxiliary variates are uncorrelated

$\rho_{z22z12} = 0.9$											
ρ_{z21z11}			0.2			0.5			0.8		
ρ_3	ρ_2	$\rho_{yx} \downarrow$	μ_0	E_1	E_2	μ_0	E_1	E_2	μ_0	E_1	E_2
0.4	0.5	0.3	0.46	154.6	151.1	0.46	156.9	153.3	0.47	159.4	155.7
		0.6	0.46	154.9	139.4	0.47	159.7	143.7	0.49	165.1	148.6
		0.9	0.50	170.6	122.5	0.54	181.9	130.6	0.58	197.2	141.6
	0.7	0.3	0.39	224.5	219.3	0.41	232.1	226.8	0.42	240.8	235.2
		0.6	0.37	210.7	189.6	0.39	223.4	201.1	0.42	239.4	215.5
		0.9	0.38	218.1	156.6	0.42	241.5	173.4	0.49	278.2	199.8
	0.9	0.3	0.16	1063.3	1038.8	0.17	1152.1	1125.6	0.19	1270.2	1240.9
		0.6	0.13	884.6	796.2	0.15	988.5	889.6	0.17	1144.5	1030.1
		0.9	0.13	857.9	615.9	0.15	1011.3	726.1	0.20	1311.0	941.2
0.6	0.5	0.3	0.42	214.6	209.7	0.43	218.5	213.5	0.43	222.6	217.5
		0.6	0.41	208.9	188.0	0.42	216.1	194.5	0.44	224.3	201.8
		0.9	0.44	228.0	163.7	0.48	244.0	175.1	0.52	265.4	190.6
	0.7	0.3	0.31	412.0	402.5	0.32	430.0	420.1	0.34	451.1	440.7
		0.6	0.28	369.9	333.0	0.30	395.9	356.3	0.32	429.4	386.4
		0.9	0.28	378.9	272.0	0.32	424.6	304.8	0.37	497.8	357.4
	0.9	0.3	*	–	–	*	–	–	*	–	–
		0.6	*	–	–	*	–	–	*	–	–
		0.9	*	–	–	*	–	–	*	–	–
0.8	0.5	0.3	0.29	535.1	522.8	0.30	549.2	536.5	0.31	564.6	551.6
		0.6	0.27	488.6	439.8	0.28	509.8	458.9	0.29	534.5	481.1
		0.9	0.27	528.1	379.2	0.31	572.3	410.9	0.35	632.6	454.2
	0.7	0.3	*	–	–	*	–	–	*	–	–
		0.6	*	–	–	*	–	–	*	–	–
		0.9	*	–	–	*	–	–	*	–	–
	0.9	0.3	*	–	–	*	–	–	*	–	–
		0.6	*	–	–	*	–	–	*	–	–
		0.9	*	–	–	*	–	–	*	–	–

Note: ** denotes μ_0 does not exist.*

7. Discussion and Conclusion

Following conclusions can be made from the Tables 1–6:

1. For fixed values of ρ_{yx}, ρ_{z21z11}, ρ_{z22z12}, ρ_3 and ρ_4, the values of μ_0 decrease with the increase in the values of ρ_2 and the values of E_1 and E_2 increase with the increase in the values of ρ_2. Similar trends are visible for increasing values of ρ_3 and ρ_4 provided other correlations are fixed. This pattern indicates that smaller fresh sample is required at current occasion, if highly correlated auxiliary variates are available. The increase in precision indicates the efficient utilization of available auxiliary variates.

2. For fixed values of ρ_2, ρ_{z21z11}, ρ_{z22z12}, ρ_3 and ρ_4, the values of μ_0 first decrease and then increase with the increase in the values of ρ_{yx}, however the values of E_1 and E_2 decrease uniformly with the increase in the values of ρ_{yx}.

3. For fixed values of ρ_2, ρ_3, ρ_4 and ρ_{yx}, it is observed that the values of $\mu_{0,}$ E_1 and E_2 increase with the increase in the values of ρ_{z21z11} and ρ_{z22z12}. This indicates that higher mutual correlation between the auxiliary variates results in increase in precision of estimates at current occasion.

4. A close perusal of Tables 1–3, revels that μ_0 attains the minimum value of 0.21 if two dynamic auxiliary variates are used. However, if only one dynamic auxiliary variate is considered, the minimum value acquired by μ_0 is 0.2729 (Priyanka (2008)).This indicates that use of more than one dynamic auxiliary variate is fruitful in sampling on two occasions.

5. From Tables 4–6, i.e., when the auxiliary variates are mutually uncorrelated, it is observed that the precision of estimates increases much more as compared to the situation when the auxiliary variates are mutually correlated. The minimum value attained by μ_0 for this case is 0.12, which is very low as compared to the previous situation. Hence, the cost of survey will reduce to a greater extent.

Maximum gain and minimum μ_0 is observed for higher values of ρ_2. Hence, the use of varying (dynamic) p-auxiliary variates is highly advantageous in terms of the proposed estimators. They not only provide estimates with increased accuracy but also tremendously reduce the cost of the survey. Thus, the proposed estimators may be recommended for its further practical use.

Acknowledgement

Authors are thankful to the honourable referee for his inspiring comments. Authors are also thankful to the Indian School of Mines, Dhanbad for providing the financial assistance to carry out the present work.

REFERENCES

BIRADAR, R. S. and SINGH, H. P. (2001). Successive sampling using auxiliary information on both occasions. Cal. Stat. Assoc. Bull. 51: 243–251.

CHATURVEDI, D. K. and TRIPATHI, T. P. (1983). Estimation of population ratio on two occasions using multivariate auxiliary information. Jour. Ind. Statist. Assoc., 21: 113–120.

COCHRAN, W. G. (1977): Sampling Techniques. New York: John Wiley & Sons.

DAS, A. K. (1982). Estimation of population ratio on two occasions. Jour Ind. Soc. Agr. Statist. 34: 1–9.

ECKLER, A. R. (1955). Rotation sampling. Ann. Math. Stat., 26, 664–685.

FENG, S. and ZOU, G. (1997). Sample rotation method with auxiliary variable. Commun. Statist. Theo-Meth. 26, 6: 1497–1509.

GUPTA, P. C. (1979). Sampling on two successive occasions. Jour. Statist. Res. 13: 7–16.

JESSEN, R. J. (1942). Statistical investigation of a sample survey for obtaining farm facts. In: Iowa Agricultural Experiment Station Road Bulletin No. 304: 1–104, Ames, USA.

PATTERSON, H. D. (1950). Sampling on successive occasions with partial replacement of units. Jour. Royal Statist. Assoc., Ser. B, 12: 241–255.

PRIYANKA, K. (2008). On the use of model based techniques and development of estimation procedures in search of good rotation patterns on successive occasions and its applications. Unpublished Ph. D. thesis submitted to the Indian School of mines, Dhanbad.

RAO, J. N. K. and GRAHAM, J. E. (1964). Rotation design for sampling on repeated occasions. Jour. Amer. Statist. Assoc. 59: 492–509.

SEN, A. R. (1971). Successive sampling with two auxiliary variables. Sankhya, Ser. B, 33: 371–378.

SEN, A. R. (1972). Successive sampling with p $(p \geq 1)$ auxiliary variables. Ann. Math. Statist. 43: 2031–2034.

SEN, A. R. (1973). Theory and application of sampling on repeated occasions with several auxiliary variables. Biometrics 29: 381–385.

SINGH, V. K., SINGH, G. N. and SHUKLA, D. (1991). An efficient family of ratio-cum-difference type estimators in successive sampling over two occasions. Jour. Sci. Res. 41 C: 149–159.

SINGH, G. N. and SINGH, V. K. (2001). On the use of auxiliary information in successive sampling. Jour. Ind. Soc. Agric. Statist. 54: 1–12.

SINGH, G. N. (2003). Estimation of population mean using auxiliary information on recent occasion in h occasions successive sampling. Statistics in Transition 6: 523–532.

SINGH, G. N. (2005). On the use of chain-type ratio estimator in successive sampling. Statistics in Transition 7: 21–26.

SINGH, G. N. and PRIYANKA, K. (2006). On the use of chain-type ratio to difference estimator in successive sampling. IJAMAS 5, S06: 41–49.

SINGH, G. N. and PRIYANKA, K. (2007). On the use of auxiliary information in search of good rotation patterns on successive occasions. BSE 1, A07: 42–60.

SINGH, G. N. and PRIYANKA, K. (2008 a). Search of good rotation patterns to improve the precision of estimates at current occasion. Commun. Statist. Theo-Meth 37, 3: 337–348.

SINGH, G. N. and PRIYANKA, K. (2008 b). On the use of several auxiliary variates to improve the precision of estimates at current occasion. Journal of Indian Society of Agricultural Statistics, 62, 3, 253–265.

SUKHATME, P. V., SUKHATME, B. V., SUKHATME, S. and ASOK, C. (1984). *Sampling theory of surveys with applications*. Iowa State University Press, Ames, Iowa (USA) and Indian Society of Agricultural Statistics, New Delhi (India).

STATISTICS IN TRANSITION-new series, July 2010
Vol. 11, No. 1, pp. 127—144

ROBUST ESTIMATION OF FINITE POPULATION TOTAL

Manoj Kumar Srivastava[1], Namita Srivastava[2]

ABSTRACT

The present paper deals with the robust prediction of finite population total under the superpopulation model G_R . The design is de-emphasized while developing these predictors under the superpopulation model and making comparison among all resistant estimators. The suggested proposals involve reweighed iterative algorithm for Robust Prediction. The discussion also involves the calculation of asymptotic bias and variance in terms of the influence function computed for these predictors. Two populations have been considered for simulation study to judge the performance of proposed predictors with conventional and model based existing alternatives.

Key words: Influence function, Prediction approach, M-estimator, Superpopulation models.

1. Introduction

Sample survey data often contains outliers. Initially according to literature, applied survey practitioners ignored the problem of outliers in survey sampling probably treating this as perennial problem. Initial advises have been made by Kish (1965, sec.11.4B) in survey sampling literature under the topic "Skew populations" in economic survey and survey of individuals. Hidiroglou and Srinath (1981) and Glasser (1962) form a separate stratum for outlying units and then combined the class means. Ernst (1980), Fuller (1991), and Searls (1966) showed that for skewed population, the mean squared error of the winsorised sample mean is smaller than that of the sample mean. Rivest (1993) studied the behavior of winsorization schemes under simple random sampling. Shoemaker and Rosenberger (1983) calculated the exact formula for the expected value and variance of trimmed mean and of median under the simple random sampling.

[1] Department of Statistics, Institute of Social Sciences, Dr. BRA University, Agra-282002, U.P. India. E-mail: *mks_iss@yahoo.co.in.*
[2] Department of Statistics, St. John's College, Agra-282002, U.P. India. E-mail: *drnamita.sjc@gmail.com.*

Fuller (1991) suggested estimators which are so constructed that the impact of the largest y-value is reduced only when a test for extreme values is significant. Smith (1987) highlighted and showed the importance of detecting and treating outliers under the superpopulation model.

Ichan (1984) suggested predictors of population mean in the presence of spurious observations under a superpopulation model and studied sensitivity and robustification aspects. Chambers (1986) distinguishes between representative and non representative outliers in a sample and suggested ϵ-contaminated model based outliers resistant predictors of population mean by using M-estimation technique. Gwet and Rivest (1992) discussed the problem of estimating the population mean using auxiliary information in the presence of outliers. Beat Hulliger (1995) suggested outlier robust alternatives for Horvitz and Thompson estimator.

In Sample survey the finite population is denoted by $U = \{U_1, U_2, \cdots, U_N\}$; the study variable y on U be denoted by $\mathbf{y} = \{y_1, y_2, \cdots, y_N\}$ and assumed to be unknown. Let an unordered sample s be drawn from F under a given design $p(s) > 0$ and the data $d = \{(k, y_k) : k \in s\}$ has been observed. For any given $s \in F$, we can write

$$T = \sum_{s} y_k + \sum_{\tilde{s}} y_k \tag{1}$$

where $\tilde{s}$ denote the set of labels not in s, *i.e.,* $\sum_{\tilde{s}} y_k$ is the total of non-sampled units. The problem of estimating T is recognized as one of predicting the sum of the unobserved random variable $\sum_{\tilde{s}} y_k$. This leads to the prediction of T. For this, the true value of the parameter $\mathbf{y}$ be a realization of $\mathbf{Y}$ under the superpopulation model given by

$$Y_k = \beta x_k + e_k; \quad k = 1, 2, \cdots, N$$

such that $E(Y_k | x_k) = \beta x_k, \quad V(Y_k | x_k) = \sigma^2 v(x)$

and $\operatorname{cov}(Y_k, Y_l | x_k, x_l) = 0 \quad \forall \; k \neq l = 1, 2, \cdots, N.$ (2)

where $x_k > 0 \; (k = 1, 2, \cdots, N)$ are known, auxiliary values β and σ^2 are unknown and $v(\cdot)$ is known.

The sample s is used to estimate the parameter β in the model ξ in equation (2) which is then used to predict values of y_k for $k \notin s$, *i.e.,* for $N - v(s)$ unobserved coordinates of $\mathbf{y} = (y_1, y_2, \cdots, y_N)$. Therefore,

$$T(d) = \sum_{s} y_k + U(d)$$

is called the predictor of T, where U is considered as predictor of $\sum_{\tilde{s}} y_k$ the total of $N - v(s)$ unobserved coordinates of **y** (Cassel, et al. (1977); Chambers(1986)).

The predictor $T(d)$ is highly sensitive to the presence of outliers both in x's and residuals r's. In the present paper the objective is to obtain a robust optimal predictor of T.

2. Proposed outlier robust predictor of *T*

Brewer (1963) and Royall (1970) gave $UM\xi UMSEE$ (Uniformly minimum ξ -unbiased mean square error estimator) in the class of all homogeneous model unbiased predictors of T

$$T_{BR} = Y_s + \hat{\beta} x_{\tilde{s}}$$

$$\xi MSE_p(T_{BR}) = E\left\{\left(\sum_{\tilde{s}} x_k\right)^2 V_\xi(\hat{\beta}) + \sigma^2 \sum_{\tilde{s}} V_\xi(x_k)\right\} \tag{3}$$

where $x_{\tilde{s}} = \sum_{\tilde{s}} x_k$, $Y_s = \sum_s y_k$, $\hat{\beta} = \dfrac{\sum_s x_k y_k / v(x_k)}{\sum_s x_k^2 / v(x_k)}$ and $V_\xi(\hat{\beta}) = \dfrac{\sigma^2}{\sum_s x_k^2 / v(x_k)}$

Chambers (1986) among others noted the high changes in the distribution of $\hat{\beta}$ when sample data deviated from the Gauss Markov assumptions resulting into high sensitivity in $\hat{\beta}$. This sensitivity is carried over to the predictor T_{BR}. We have proposed an outlier robust predictor alternative to $T_{BR}(d)$ which is less sensitive to sample outliers.

To replace $\hat{\beta}$ by an outlier robust alternative, using general robust procedure developed by Maronna and Yohai (1981) (c.f. Hampel et al. (1987)) for hypothetical populations, and by taking its finite population analogous for superpopulation model (Gwet and Rivest. (1992)). The M-estimator of the parameter β in G_R is obtained by solving the equation implicitly

$$\sum_{k=1}^{n} \eta\left(x_k, \frac{y_k - x_k\beta}{\sigma v(x_k)}\right) x_k' = 0 \tag{4}$$

where $x'_k = \frac{x_k}{\sigma\sqrt{v(x_k)}}$ and $\eta : R^1 \times R^1 \to R^1$ which satisfies some regularity conditions (Hampel, et al. (1987), p 315). The solution of the above equation gives regression *M*-estimator. η-function is a function ψ of x_k, y_k and β which preserves the sign as of x_k, and depends on β only through $r = y - x'\beta$. These restrictions derive their requirements from equivariance considerations. If one writes η as

$$\eta(x,r) = w(x).\psi\big(r, v(z)\big) \tag{5}$$

where $\psi: R^1 \to R^N$, $w: R^P \to R^+$, $v: R^P \to R^+$ (w and v are known as weight functions). In particular, this paper uses Merrill and Schweppe's proposal (Hampel, et al. (1987), p 315) which gives $\eta(x,r) = w(x)\psi\left(r, \frac{1}{w(x)}\right)$ *i.e.,* $v(x) = 1/w(x)$ in equation (5). Large *x* is assigned a low weight and if this results into large residual *r* it assigns high $\frac{1}{w(x)}$ *i.e.* high $r.\frac{1}{w(x)}$, ψ is defined as diminishing or nearly zero outside a range and identity near the origin. Such a $\psi(\cdot)$ function contribute much less for large argument value *i.e.* $r.\frac{1}{w(x)}$. The above proposal suggests using redescending ψ - function.

3. Estimating equation for proposed predictor

Let the residuals *r* and *x* be standardized by dividing by $\sqrt{V(y/x)} = K.\sigma\sqrt{v(x)}$. Using this transformation the present set up of the model G_R is changed to the Markovian set up. The defining equation can now be written as

$$\sum_s \frac{x_k}{K\sigma\sqrt{v(x_k)}}.\frac{1}{w(x_k)}\psi\left(\frac{y_k - \beta x_k}{K\sigma\sqrt{v(x_k)}}.w(x_k)\right) = 0 \tag{6}$$

For a proper choice of ψ, the following conditions are imposed in the defining equation

1. $\psi(t)$ is real and skew-symmetric, i.e. $\psi(-t) = -\psi(t)$.
2. $\psi(t)$ is bounded.

3. $\frac{\psi(t)}{t} \to 1$ as $|t| \to 0$, $\psi(t) \to t$ as $|t| \to 0$, an identity transformation for small values of *t*.

4. $w(t)$ is positive and non-decreasing function of *t*.

we get, for $v(x_k) = x_k$, $\hat{\beta}_{LS} = \frac{\sum_s y_k/n}{\sum_s x_k/n} = \frac{\bar{y}_s}{\bar{x}_s}$ a ratio estimator of the population ratio R and if $v(x_k) \propto x_k^2$, $\hat{\beta}_{LS} = \frac{1}{n}\sum_s \frac{y_k}{x_k}$ and if $v(x_k) \propto$ constant, $\hat{\beta}_{LS} = \frac{\sum_s y_k x_k}{\sum_s x_k^2}$. They all are sensitive to the presence of sample outliers. The proper choice of $\psi(\cdot)$ and weight functions, reduces the effect of presence of sample outliers in *x* and in residual *r,* to make the $\hat{\beta}$ predictors robust. If $\psi(\cdot)$ is bounded and $w(\cdot)$ is constant, the *M*-estimator obtained may still be sensitive to large values of $x's$ (note the factor $x/v(x)$ in defining equation). Such an *M*-estimator, although being insensitive/robust against the presence of outliers among the residuals, is still non-robust against the presence of outliers in *x*. (refer Hampel et al., 1986, pp 314; also the example given there.)

In order to obtained a robustified estimator of β when sample outliers are present both in residuals and among $x's$, a bounded $\psi(\cdot)$ is considered and the weight is taken as $w(x) = \frac{x/\bar{x}}{\sqrt{v(x_k)}}$ to get the estimating equation of β as

$$\sum_s \psi\left(\frac{y_k - \beta x_k}{K\sigma\sqrt{v(x_k)}} \cdot \frac{x_k}{\bar{x}\sqrt{v(x_k)}}\right) = 0.$$

3.1.Solving estimating equation

To solve the estimating equation (6) in order to get the robust estimate of β against the presence of outlying observations in the sample data we adopted Tukey's (1970, 71) and Andrews et al., (1972, pp 239) proposal on the estimating equation. After some mathematical simplification the solution algorithm could be obtained which is termed as computational iteratively re-weighted least squares algorithm (a generalized M-estimator). The results are being stated as under:

Result 1

Under the model ξ using Andrews (1972) and Andrew's et al. (1972) procedures, the solution of defining implicit equation is

$$\beta_n = \frac{\sum_s x'_k y'_k . w\left(r'_k . w\left(x_k\right)\right)}{\sum_s x'^2_k . w\left(r'_k . w\left(x_k\right)\right)} \tag{7}$$

where $x'_k = \dfrac{x_k}{K\sigma\sqrt{v\left(x_k\right)}}$ and $r' = \dfrac{y_k - \beta x_k}{K\sigma\sqrt{v\left(x_k\right)}}$

Starting with $\beta_n^{(0)} = \dfrac{med\left(x_k y_y \big/ S_n^{(0)2} v\left(x_k\right)\right)}{med\left(x_k^2 \big/ S_n^{(0)2} v\left(x_k\right)\right)}$

where $S_n^{(0)} = 1.483\ MAD\left(y_k\right) = 1.483\ med_s \mid y_k - med_s\left(y_l\right)\mid$ (Andrews et al., 1972, p. 239), and weights $w^{(0)}(\cdot) = w^{(0)}\left(r_k'^{(0)} \cdot w(x_k)\right) = \dfrac{\psi\left(r_k'^{(0)} \cdot w\left(x_k\right)\right)}{r_k'^{(0)} \cdot w\left(x_k\right)}$.

Iteration are continued by recalculating the weights, and modifying the value of β_n by computing

$$\beta_n^{(j+1)} = \frac{\sum_s y'_k x'_k \cdot w^{(j)}\left(r_k'^{(j)}\right)}{\sum_s x'^2_k \cdot w^{(j)}\left(r_k'^{(j)}\right)} \tag{8}$$

from the second iteration and onwards $S_n^{j+1} = sqrt\left\{\sum_s w\left(r_k'^{(j)}\right)\left(r_k'^{(j)} - \bar{r}_s\right)^2\right\}$, the square root of weighted residual sum of squares. The iterations are continued until $\mid \beta_n^{(k)} - \beta_n^{(k-1)} \mid \le \varepsilon, \beta_n^{(k)}$ is called k-step M-predictor of β. The rate of convergence depends on the proper robust choices of scale $S_n^{(0)}$ and the slope $\beta_n^{(0)}$. The behavior of one step and fully iterated M-predictor of β essentially the same when ψ is odd and underlying distribution F is symmetric.

Result 2

When the function $\psi(\cdot)$ be differentiable once in the neighborhood $\mid r_k - r_k^{(0)} \mid \le k$ the M-predictor calculated by using

$$T_n^{(j+1)} = T_n^{(j)} + \sum_s \psi\left(\frac{y_k - \beta_n^{(j)} x_k}{KS_n^{(j)}\sqrt{v(x_k)}}\right) \Bigg/ \frac{x_k}{KS_n^{(j)}\sqrt{v(x_k)}} \cdot \psi'\left(\frac{y_k - \beta_n^{(j)} x_k}{KS_n^{(j)}\sqrt{v(x_k)}}\right)$$

the iterations are continued until $T_n^{(\cdot)}$ converge, i.e. $\mid T_n^{(k)} - T_n^{(k-1)} \mid \le \varepsilon$.

In the following discussion the above result have been applied for different bounded and redecending ψ functions

1. Considering Huber ψ -function

$\psi(t) = \mathrm{sgn}(t) \ \min(k_0, |t|) = t \ \min\left(1, \frac{k_0}{|t|}\right)$, weight becomes

$$w_k(\beta, S_n, k_0) = \min\left\{1, \frac{k_0}{\left|\frac{y_k - \beta x_k}{K S_n \sqrt{v(x_k)}}\right| \cdot w(x_k)}\right\}$$

And $\beta_n = \sum_s \frac{y_k x_k}{K^2 S_n^2 v(x_k)} \cdot w_k(\beta, S_n, k_0) \Big/ \sum_s \frac{x_k^2}{K^2 S_n^2 v(x_k)} \cdot w_k(\beta, S_n, k_0)$

where k_0 is some $+ve$ constant, S_n is some predictor of σ, and robustification constant K in the scale parameter is assumed to be known.

Using iteratively reweighted least square algorithm and starting with appropriate robust predictors of σ and β as $S_n^{(0)}, \beta_n^{(0)}, S_n^{(0)} = 1.483 \ \mathrm{MAD}(y_k)$ and

$$\beta_n^{(0)} = \mathrm{med}_k \left\{\frac{y_k x_k}{K_{(0)}^2 S_n^{(0)^2} v(x_k)}\right\} \Big/ \mathrm{med}_k \left\{\frac{x_k^2}{K_{(0)}^2 S_n^{(0)^2} v(x_k)}\right\}$$

One computes

$$\beta_n^{(j+1)} = \frac{\sum_s \frac{y_k x_k}{K_{(0)}^2 S_n^{(j)^2} v(x_k)} \cdot w_k\left(\beta_n^{(j)}, S_n^{(j)}, k_0\right)}{\sum_s \frac{x_k^2}{K_{(0)}^2 S_n^{(j)^2} v(x_k)} \cdot w_k\left(\beta_n^{(j)}, S_n^{(j)}, k_0\right)} \tag{9}$$

until β_n converges. This predictor features the bounded or limited impact of outliers on the predictor calculated under the model $\xi(G_R)$.

2. Cauchy ψ -function which is a typical redescending ψ function defined as

$$\psi_c(t) = \frac{t}{1 + t^2}; \tag{10}$$

The shape of this Cauchy function is a pure redescending function without any tuning constant. For this ψ - function the weight function is calculated as

$$w_k(\beta, \sigma) = \frac{1}{\left\{1 + \left\{\frac{y_k - \beta x_k}{K \sigma \sqrt{v(x_k)}} \cdot w(x_k)\right\}^2\right\}} \tag{11}$$

and the iteratively reweighing least square algorithm gives

$$\beta_n^{(j+1)} = \frac{\sum_s \frac{y_k x_k}{K_{(0)} S_n^{(j)^2} v(x_k)} \cdot w_k\left(\beta_n^{(j)}, S_n^{(j)}\right)}{\sum_s \frac{x_k^2}{K_{(0)} S_n^{(j)^2} v(x_k)} \cdot w_k\left(\beta_n^{(j)}, S_n^{(j)}\right)} \tag{12}$$

taking $S_n^{(0)}$ and $\beta_n^{(0)}$ same as has been considered earlier and iterative procedure is continued until $\beta_n^{(j)}$ converges.

Cauchy ψ function is an identity function near the origin and it converges to zero when $|t| \to \infty$. This is used when impact of outliers on the residual need to be controlled.

3. Chamber type $\psi(\cdot)$ function which is a redescending ψ -function

$$\psi(t) = t \exp\left\{-\frac{a}{2}\{|t| - b\}^2\right\} \tag{13}$$

where b is the location and a is scale constant or shape constant.

For the above ψ -function the weight function is calculated as

$$w_k\left(r_k'(\beta), \sigma, a, b\right) = \exp\left\{-\frac{a}{2}\left(|r_k'| w(x_k) - b\right)^2\right\} \tag{14}$$

using these weights, the iterative least squares reweighted algorithm suggests

$$\beta_n^{(j+1)} = \frac{\sum_s x_k' y_k' . w_k\left(r_k'\left(\beta_n^{(j)}\right), S_n^{(j)}, a, b\right)}{\sum_s x_k'^2 w_k\left(r_k'\left(\beta_n^{(j)}\right), S_n^{(j)}, a, b\right)} \tag{15}$$

until $\beta_n^{(\cdot)}$ converges.

If $|r_k'| w(x_k) \le b$, the weight in the estimating process is $w_k(\cdot) \propto x_k' = x_k \Big/ \left(K \sigma \sqrt{v(x_k)}\right)$. If $|r_k'| w(x_k)$ is low then the weight $w_k \propto x_k'$ and if $|r_k'| w(x_k)$ is large and further becomes larger, $w_k(\cdot)$ becomes lower and lower contributing almost zero or asymptotically insignificant to the process of estimation.

4. If ψ is Biweight function due to Beaton and Tukey (1974) (Hampel (1986), section 2.6, p 151) which is a smooth (in the sense of continuity) function defined as $\psi_{bi,(k_0)} = t\left(k_0^2 - t^2\right)^2 \mathrm{I}_{[-k_0, k_0]}(t)$ the weights are calculated as

$$w_k\left(r_k'(\beta),\sigma,b\right)=\left[k_0^2-\left(r_k'.w\left(x_k\right)\right)^2\right]^2 \mathrm{I}_{[-k_0,k_0]}(r_k'.w_k) \tag{16}$$

and iterative least squares reweighted algorithm be given by

$$\beta_n^{(j+1)}=\frac{\sum_s x_k' y_k' \cdot w_k\left(r_k'\left(\beta_n^{(j)}\right),S_n^{(j)},k_0\right)}{\sum_s x_k'^2 \cdot w_k\left(r_k'\left(\beta_n^{(j)}\right),S_n^{(j)},k_0\right)} \tag{17}$$

until β_n converges. Note that in estimation process outlying observations $\left|r_k' w\left(x_k\right)\right|>k_0$ are totally rejected by assigning them zero weights and well behaved observations are given full weight.

The constant k_0 in ψ is determined such that $\rho^*(\psi)\le k_0$ where

$$\rho^*=\inf\left\{c>0:\left(\psi\left(y,\xi\left(G_R\right)\right)=0\right) \text{ when } |\mathrm{y}|>\text{constant}\right\} \tag{18}$$

which is based on the rejection policy due to Daniel Bernoulli in 1769 (Stigler, 1980) of rejecting contaminated values of y entirely from the estimation process. This corresponds to $IF\left(x,\psi(\cdot),\xi\right)=0$. The condition $\psi(\cdot)=0$ in (18) comes from $IF\left(\cdot,\cdot,\cdot\right)=0$.

The choice of robust constant K should be such that it balances the choice of S_n so that contaminations in residuals are properly downweighed instead rejected entirely leading to wrong conclusions. In other words, K allows the choice of level of robustness in ψ-based predictors. For example, when ψ is redescending selecting large K amounts to $\psi(t)=t$, leading to *LS*-predictors. For bi-weight ψ-robustness procedure K = 2 is recommended for $\left|r_k'\right|\le w\left(x_k\right)$ so that $\frac{r_k'}{w\left(x_k\right)}=\pm 1$ falls roughly at $\pm 2.7\sigma$, other values of K = 3,4, or 5 are also recommended in the literature.

In case of Cauchy ψ-function the robust constant K is calculated by the relation

$$K_{\text{Cauchy}}^{\text{GM}}=K_{\text{Hub}}^{\text{GM}}\ \frac{K_{\text{Cauchy}}^{\text{M}}}{K_{\text{Hub}}^{\text{M}}} \tag{19}$$

Notionally, all ψ-functions fall in between Cauchy and Huber ψ-functions with their corresponding robustifying constant K values 2.34 and 4.54. The choice of K by different statisticians for their ψ-function could be made between this range by using some rationality or optimality criterion.

4. Asymptotic properties

4.1. Influence function for the defined GM- estimator

The effect of small departure of the distribution of $\frac{(y-\beta(F)x)}{\left(K\sigma\sqrt{v(x)}\right)}$ from Gauss-Markov model due to the presence of outliers is measured by the influence function of β at *F(x, y)* is defined as

$$IF\left((x,y);\beta,F(x,y)\right)=\frac{\partial}{\partial t}\left[\beta\left(F_t(x,y)\right)\right]_{t=0} \tag{20}$$

where $F_t(x,y)=(1-t)F(x,y)+t\Delta_{(x,y)}$, differentiating the functional defining equation with respect to *t,* then, making use of the above definition we get

$$\frac{\partial}{\partial t}\left[\beta\left(F_t(x,y)\right)\right]_{t=0}=\frac{x'\cdot\frac{1}{w(x)}\cdot\psi\left(r'.w(x)\right)|_{\beta(F(x,y))}}{\int x'^2\cdot\frac{1}{w(x)}\cdot\psi'\left(r'.w(x)\right)|_{\beta(F(x,y))}.dF(x,y)}$$

4.2. Asymptotic bias and variance

Theorem 4.1. Under the model ξ, the ξ-bias in the predictor T_n is given by

$$\xi\left(T_n-T\right)=\frac{\sum_{\bar{s}}x_k}{\lambda_0}.\frac{1}{n}\sum_s\lambda_k-\sum_{\bar{s}}\mu_k \tag{21}$$

where $\lambda_0=\xi\left[x'^2\frac{1}{w(x)}\psi'\left(r'.w(x)\right)\right]_{\beta(F(x,y))}$

and $\lambda_k=\xi\left[\frac{x'_k}{w(x_k)}\psi\left(r'_k.w(x_k)\right)\right]_{\beta(F(x,y))}$

Proof. If some distribution *G* is considered near *F,* then the first-order Von Mises expansion of β at *F* (which is derived from a Taylor Series) evaluated in *G* is given by

$T(G)=T(F)+\int IF(x;T,F)d(G-F)(x)$ + Remainder (refer Hampel et al. page 85)

$$\beta(G(x,y))=\beta(F(x,y))+\int IF((x,y);\psi,F(x,y))\times d\left(G(x,y)\right)$$

$$-F(x,y))+\text{Remainder}$$

If we assume that (x_k, y_k) are iid from $F(x, y)$ then the empirical distribution function $F_n(x,y) \xrightarrow{£} F(x,y)$, by Glivenko-Cantelli theorem for large n. Also, note that $\int IF((x,y);\psi;F(x,y))dF(x,y)=0$. Under these conditions, we could now replace $G(x, y)$ by the sample empirical distribution function $F_n(x,y)$, we get

$$\beta_n\left(F_n(x,y)\right) \approx \beta\left(F(x,y)\right) + \int IF\left((x,y);\psi,F(x,y)\right)\times dF_n + \text{remainder}$$

If we further assume that $\beta(\cdot)$ is a functional, $\beta_n(F_n)$ then may be approximated by $\beta(F_n)$. Therefore,

$$\beta\left(F_n(x,y)\right) - \beta\left(F(x,y)\right) \approx \frac{1}{n}\sum_{k=1}^{n} IF\left((x_k,y_k);\psi,F(x_k,y_k)\right) + \text{ remainder}$$

In most cases remainder becomes negligible for $n \uparrow \infty$. Rewriting this yields

$$\sqrt{n}\left(\beta\left(F_n(x,y)\right) - \beta\left(F(x,y)\right)\right) \approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n} IF\left((x_k,y_k);\psi,F\right) + R_n$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} IF\left((x_k,y_k);\psi,F\right) \to N\left(0, V_\xi\left(\beta,F(x,y)\right)\right) \text{ as } n\to\infty$$

where $V_\xi\left(\beta,F(x,y)\right) = \int_{(x,y)} IF^2\left((x,y);\psi,F(x,y)\right)dF(x,y)$

Therefore, the asymptotic bias in β_n shall be given by

$$(\beta_n - \beta) \approx \frac{1}{n}\sum_s \frac{\frac{x_k'}{w(x_k)}\psi\left(r_k'\cdot w(x_k)\right)|_{\beta(F(x,y))}}{\int x_k'^2\cdot\frac{1}{w(x)}\psi'\left(r'\cdot w(x)\right)|_{\beta(F(x,y))}\cdot dF(x,y)}$$

and ξ- bias is

$$B(\beta_n) = \xi(\beta_n - \beta) = \frac{1}{n}\sum_s \frac{\xi\left[\frac{x_k'}{w(x_k)}.\psi\left(r_k'.w(x_k)\right)\right]_{\beta(F(x,y))}}{\xi\left[x'^2.\frac{1}{w(x)}\psi'\left(r'.w(x)\right)\right]_{\beta(F(x,y))}}$$

$$= \frac{1}{\lambda_0}\cdot\frac{1}{n}\sum_s \lambda_k$$

The mean of the prediction error $(T_n - T)$ under the model $\xi(G_R)$, *i.e.* $\xi(T_n - T)$, also termed as ξ-bias in T_n, is given by

$$\xi\left(T_n - T\right) = \xi^{\xi}\xi^{\beta/\xi}\left\{-\sum_{\tilde{s}}\left(y_k - \beta_n x_k\right)\right\} = \frac{\sum_{\tilde{s}} x_k}{\lambda_0}.\frac{1}{n}\sum_{s}\lambda_k - \sum_{\tilde{s}}\mu_k$$

Theorem 4.2. *Under the model* ξ in the ξ*-mean square error in the predictor* T_n *is given by*

$$V_{\xi}\left(T_n - y_u\right) = \left(\sum_{\tilde{s}} x_k\right)^2 . \frac{\xi\left[\frac{x'}{w(x)}.\psi\left(r'.w(x)\right)\right]^2_{\beta(F(x,y))}}{\xi^2\left[x'^2.\frac{1}{w(x)}\psi'\left(r'.w(x)\right)\right]_{\beta(F(x,y))}}$$

$$+\left(K\sigma\right)^2\sum_{\tilde{s}}\xi^2 v(x_k) - 2\beta\sum_{\tilde{s}}\operatorname{cov}(x_k, y_k)$$

Proof. The asymptotic variance of the predictor β at distribution function $F(x, y)$ is given by,

$$V_{\xi}\left(\beta, F(x,y)\right) = \int_{(x,y)} IF\left((x,y);\psi, F(x,y)\right)^2 dF(x,y)$$

$$V_{\xi}\left(T_n - y_u\right) = V_{\xi}\left\{\left(\beta_n - \beta\right)\sum_{\tilde{s}} x_k - \sum_{\tilde{s}}\left(y_k - \beta x_k\right)\right\}$$

Here $$V_{\xi}\left(\beta_n - \beta\right) = \frac{\xi\left[\frac{x'}{w(x)}.\psi\left(r'.w(x)\right)\right]^2_{\beta(F(x,y))}}{\xi^2\left[x'^2.\frac{1}{w(x)}\psi'\left(r'.w(x)\right)\right]_{\beta(F(x,y))}}$$

$$V_{\xi}\left(y_k - \beta x_k\right) = \left(K\sigma\right)^2 \xi^X v(x_k)$$

and $$\operatorname{cov}\left\{\beta_n\sum_{\tilde{s}} x_k, \sum_{\tilde{s}} y_k\right\} = \beta\sum_{\tilde{s}}\operatorname{cov}(x_k, y_k)$$

$$\therefore V_{\xi}\left(T_n - y_u\right) = \left(\sum_{\tilde{s}} x_k\right)^2 . \frac{\xi\left[\frac{x'}{w(x)}.\psi\left(r'.w(x)\right)\right]^2_{\beta(F(x,y))}}{\xi^2\left[x'^2.\frac{1}{w(x)}\psi'\left(r'.w(x)\right)\right]_{\beta(F(x,y))}}$$

$$+\left(K\sigma\right)^2\sum_{\tilde{s}}\xi^2 v(x_k) - 2\beta\sum_{\tilde{s}}\operatorname{cov}(x_k, y_k)$$

5. Empirical study

Two populations have been considered for simulating the performance of robust predictors for different ψ functions discussed above. The first population shows the area under wheat in 1936 for 34 village in Lucknow subdivision (Sukhatme and Sukhatme (1970, p 185)), which is also used by Som (1973, p 73) and Gwet and Rivest (1992, p 1175). This population has a single outlier when superpopulation parameter is estimated by the Royall's prediction approach. The second population has been taken from Appendix E of Kish (1965) with $N = 235$. This population has 13 outliers. By simple examination of the data, the relationship between the two variable confirms the validity of the assumed kernel superpopulation model G_R as has been defined in early section, with variance function $v(x) = x$.

All the simulation results are based on over 400 repeated samples of sizes 4, 8, 12 and 20, 30, 40 from population 1 and 2 respectively. For the calculation of the superpopulation parameter β for ψ_{BT}, ψ_{CAU}, ψ_{HUB} in the estimating equation, the respective standard residuals were being calculated. On the basis of ψ-plots for these standardized residuals, the tuning constants have accordingly been chosen.

In Chambers predictor

	ψ_{BT}	ψ_{Hub}	ψ_{Cau}	ψ_{BT}	ψ_{C}
pop. 1	$c = 0.3$	$k = 0.2$	–	$a = 2.5, b = -0.09$	$a = 5, b = -0.09$
pop. 2	$c = 0.5$	$k = 0.3$	–	$a = 100, b = 0.0251$	$a = 100, b = 0.006$

In this performance evaluation study the following conventional and model based estimators and predictors have been included:

1. Simple expansion estimator $T_n = N\bar{y}_s$;
2. least square estimator, which is same as ratio estimator under the model $T_n(RAE) = \dfrac{y_s}{x_s} x_u$;
3. 1-winsorized estimator $T_n(WINS) = \dfrac{y_s^1}{x_s^1} x_u$
4. Robust Predictors $T_{RB}(BT) = T_1 + b_n \sum_2 x_k$,
   $T_{RB}(HUM) = T_1 + b_n(HUM)\sum_2 x_k$, $T_{RB}(HUMG) = T_1 + b_n(HUMG)\sum_2 x_k$,
   $T_{RB}(CAM) = T_1 + b_n(CAM)\sum_2 x_k$, $T_{RB}(CAGM) = T_1 + b_n(CAGM)\sum_2 x_k$;
5. Chambers predictor, $T_{CHAM}(BT) = T_1 + b_n(BT)\sum_2 x_k + \dfrac{\sigma}{n}\sum_2 x_k \left[\sum_1 \psi_c(t)\right]$,
   $T_{CHAM}(CAM) = T_1 + b_n(CAM)\sum_2 x_k + \dfrac{\sigma}{n}\sum_2 x_k \left[\sum_1 \psi_c(t)\right]$

It has been observed that by adopting the biweight reiterative algorithm, the above procedure resulted into the convergence of b_n, only after second to third iteration. This observation falls in agreement to the concept of one step M-estimator procedure. The various parametric values for population 1 and population 2 are summarized in table 5.

Table 4.1. Summary of Population Characteristics

	N	Y_u	X_u	β_u	Outliers
Population 1.	34	7426	26022	0.285374	1
Population 2.	235	4713	7463	0.6315	13

The performance of the predictor has been judged on the basis of relative bias (RB) = {average bias over large number of samples/ true value(T)} × 100, and relative coefficient of variation (RCV)= {square root of mean square error of the predictor over large number of samples/T} × 100. The simulation results for the suggested resistant alternatives with the existing conventional and those model based are summarized in the table 4.2 and table 4.3.

Table 4.2. Simulation results

	$\beta_n(LS)$					$\beta_n(BT)$				
n	$E(\beta_n)$	Bias	RB	MSE	CV	$E(\beta_n)$	bias	RB	MSE	CV
4	.29789	.012	4.39	.006	27.45	.28124	–.004	–1.45	.003	19.51
8	.28657	.001	.43	.001	13.29	.27431	–.01	–3.88	.0009	10.43
12	.27955	–.006	–2.04	.0007	9.68	.27007	–.02	–5.3	.0006	8.89
20	.62684	–.005	–.74	.004	10.59	.63611	.004	.73	.008	14.43
30	.62680	–.005	–.75	.003	8.08	.64149	.01	1.58	.005	10.83
40	.63038	–.001	–.179	.002	6.82	.64147	.01	1.58	.005	11.39
	$\beta_n(HUB)$					$\beta_n(CAU)$				
n	$E(\beta_n)$	Bias	RB	MSE	CV	$E(\beta_n)$	bias	RB	MSE	CV
4	.02805	–.005	–1.70	.003	19.66	.297852	.01	4.37	.007	28.81
8	.27638	–.009	–3.15	.0008	9.66	.286705	.001	.47	.001	12.98
12	.27217	–.01	–4.62	.0006	8.74	.280722	–.005	–1.6	.0007	9.56
20	.64066	.009	1.45	.005	11.37	.647632	.02	2.55	.007	13.53
30	.64047	.009	1.41	.003	8.16	.646652	.02	2.40	.004	10.48
40	.64698	.02	2.45	.003	8.40	.653527	.02	3.49	.003	9.25

The table reveals the conventional way of outlier treatments (RAE, WINS) results into higher negative bias and increased coefficient of variation, comparing T_n (*WINS*) and T_n (*RAE*) for both the populations and for all sample sizes. This suggests the need to use the model based robust predictors.

Table 4.3. Simulation Results

N	T_n (RAE)		T_n (WINS)		T_n (RAE)		$T_n^{(5)}$ (HUM]		$T_n^{(5)}$ (HUGM)	
	RB	CV	RB	CV	RB	CV	RB	CV	RB	CV
4	1.38	22.39	–6.97	26.42	–2.75	15.55	–3.48	14.65	–3.03	15.59
8	–1.73	14.03	–3.46	16.82	–4.86	11.55	–4.53	11.35	–4.38	11.80
12	1.75	10.67	0.27	11.07	–2.42	7.37	1.83	7.73	–1.41	7.93
20	–1.58	11.50	–2.63	12.74	–1.67	13.50	0.04	11.02	0.075	10.59
30	–.49	8.20	–2.21	9.40	–4.19	11.76	1.03	7.23	0.76	7.55
40	–.85	7.58	–2.05	8.18	–2.47	9.11	1.31	7.05	0.83	6.90

	$T_n^{(5)}$ (CAU)		$T_n^{(5)}$ (CAUGM)		$T_n^{(5)}$ (BT)		$T_n^{(5)}$ (CAU)	
n	RB	CV	RB	CV	RB	CV	RB	CV
4	1.36	22.36	1.37	22.38	–2.23	16.59	0.77	21.26
8	–1.77	13.98	–1.74	14.01	–4.71	11.47	–2.68	12.84
12	1.69	10.60	1.73	10.65	–2.38	7.36	0.15	8.89
20	–1.23	11.33	–1.36	11.25	–1.56	13.67	–0.22	10.91
30	–0.05	7.88	–0.24	7.87	–4.09	11.89	0.60	7.45
40	–0.26	7.38	–.41	7.33	–2.46	9.20	0.38	7.03

The M-predictors (HUM and CAU) generally have smaller coefficient of variation as compared to GM-predictors (HUGM and CAGM) for population 1. Therefore, the use of down-weight functions for x results in loss in efficiency. On the contrary the *CV* of GM-predictors are small for population 2.

This has intuitive justification in case of population 2 since this contains 13 outliers in x's and GM predictors downweight large x's by suitable choice of downweight functions $w(x)$.

Practically, $T_n^{(5)}$ *(BT)* and $T_n^{(5)}$ *(HUM)* perform equally well whereas $T_n^{(5)}$ (*CAU*) is better in terms of bias at the cost of high coefficient of variation, in case of population 1. This confirms the non-demanding use of typical redescending type- $\psi(\cdot)$ function, when there is only one outlier in residuals, results in loss in efficiency. This is the reason why $T_n^{(5)}$ *(CAUGM)* and $T_n^{(5)}$ *(HUGM)* perform better as compared to $T_n^{(5)}$ (BT) in terms of RB and CV in case of population 2,

where the use of such predictors are justified. $T_n^{(6)}(\cdot)$ the Chambers predictors do not gain substantially over $T_n^{(5)}(\cdot)$. It has been shown that Chambers predictors are one-step $T_n^{(5)}(\cdot)$ predictors. The robust estimator of $\beta(F)$ in the model G_R uses *M*-estimation procedure by adopting iteratively reweighted least square algorithm (Beaton and Tukey, 1974). Therefore, $T_n^{(5)}(\cdot)$ predictors effectively converge to $T_n^{(6)}(\cdot)$ predictors. That is why the performance of these predictors are usually so similar. In these simulation studies biases are not the important part of the mean squared errors. We note small bias even if the sampled outliers are given weight close to 1.

5.1. Conditional bias analysis

Large biases of $T_n^{(5)}$ estimators as compared to ratio estimator, RAE is observed in the simulation results, this is somewhat deceptive. The relative conditional biases of the resistant predictors have been calculated in the following table. Samples of size n = 30 are drawn from the population 2, conditioned on number of outliers sampled is 0, 1, 2, 3, 4, 5 respectively. Comparison figures RB and CV have been calculated for each of conditional situation.

Tabl 5.1. Absolute Relative Conditional Bias for of size 30 drawn from population 2. The unconditional probabilities of getting 0, 1, 2, 3, 4, and 5 outliers are given by 0.16, 0.32, 0.29, 0.15, 0.05, and 0.01.

	Outliers	*(RAE)*	*(WINS)*	*(BT)*	*(HUM)*	*(HUGM)*
0	RB	8.37	6.68	3.92	12.33	12.51
	CV	(10.38)	(9.96)	(12.14)	(14.17)	(14.11)
1	RB	3.31	3.36	3.72	8.68	8.04
	CV	(7.53)	(8.55)	(11.55)	(11.54)	(10.99)
2	RB	–2.00	–3.08	0.99	3.55	2.59
	CV	(6.79)	(7.76)	(11.09)	(8.52)	(7.79)
3	RB	–5.77	–7.61	–0.65	–0.45	–1.53
	CV	(9.12)	(1.99)	(14.62)	(8.71)	(7.99)
4	RB	–11.37	–12.81	–3.34	–7.63	7.89
	CV	(13.39)	(15.16)	(13.75)	(11.83)	(11.44)
5	RB	–14.25	–15.68	–4.22	–11.95	10.82
	CV	(15.85)	(17.43)	(14.19)	(15.37)	(14.64)

	Outliers	CAU	CAGM	(CHM)BT	(CHM)CAU
0	RB	12.51	12.50	3.98	10.29
	CV	(14.21)	(14.08)	(11.42)	(12.20)
1	RB	6.85	6.59	3.48	6.31
	CV	(10.57)	(10.33)	(10.91)	(9.46)
2	RB	0.57	0.34	0.66	1.03
	CV	(8.15)	(7.96)	(10.44)	(7.01)
3	RB	–4.00	–4.25	–1.05	–2.98
	CV	(9.90)	(9.84)	(13.79)	(8.38)
4	RB	–10.82	–10.96	–4.01	–9.29
	CV	(11.01)	(14.02)	(13.21)	(12.19)
5	RB	–11.82	–14.64	–4.73	–12.83
	CV	(17.37)	(17.36)	(13.33)	(15.16)

The probabilities of getting 0, 1, 2, 3, 4, 5, outliers in the sample are 0.16, 0.32, 0.29, 0.15, 0.05 and 0.01 respectively. Therefore, expected number of outliers in a sample of size n = 30 would be approximately 2. Occurrences of more than 5 outliers in the sample is too low. The table 5.1 shows large biases of ratio and WINS estimators as compared to their resistant competitors, when too many outliers are sampled. In case when only two outliers are sampled, the conditional biases of $T_n^{(5)}$ predictors and the difference of biases grow rapidly as number of outliers sampled increases. Therefore, from the conditional point of view, ratio and WINS are most biased predictors.

REFERENCES

CASSEL, C. M. and SARNDAL C. E. (1977). Foundation of Inference in Survey Sampling, New York : John Wiley.

CHAMBERS, R. L. (1986). Outlier Robust Finite Population Estimation. *Journal of the American Statistical Association,* **81,** 1063–1069.

ERNST, L.R(1980). Comparison of estimator of the Mean which Adjust for large Observations, *'Sankhya, Ser.C.***42***,* 1–16.

FULLAR, W.A.(1991). Simple estimators of the mean of Skewed Populations, *Statistica Sinica,* **1**, 137–158.

GLASSER, G.J.(1962). On the complete coverage of large units in a statistical study. *International Statistical Review,* **30,** 28–32.

GWET, J. P. and RIVEST, L.P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association,* **87**, 1174–1182.

HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J. and STAHEL, W.E. (1986): *Robust Statistics: The Approach Baaed on Influence Functions,* New york: John Wiley.

HIDIROGIOU, M. H., and SRINATH, K. P. (1981): Some Estimators of the Population Total from simple Random Samples Containing Large Units, *Journal of the American Statistical Association,* **76**, 690–695.

HULLIGER, B. (1995): Outlier Robust Horvitz-Thompson Estimators. Survey Methodology 21(1), 79–87.

ICHAN, R.(1984): Sampling strategies, robustness and efficiency: the state of art. *International Statistical Review,* **52**, 209–218.

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

RAO, J.N.K (1985). Conditional Inferences in Survey Sampling, *Survey Methodology,* **11**, 15–31.

RIVEST, L.P. (1993). Winsorization of Survey Data, *Proceedings of the 49th Session, International Statistical Institute*.

SCOTT, A. J., and BREWER, K. R. W., and HO, E. W. H. (1978). Finite Population Sampling and Robust Estimation, *Journal of the American Statistical Association,* **73** , 359–361.

SEARLS, D.T. (1966). An estimator for a population mean which reduces the effect of large observations, *Journal of the. American Statistical Association,* **61,** 1200–1204.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics,* New York : John Wiley.

SHOEMAKER, L.H. and ROSENBERGER, J.L. (1983). Moments and efficiency of the median and trimmed mean for finite population, *Communication in Statistics, Simulations and computation,* **12(4)**, 411–422.

SMITH, T.M.F. (1987). Influential observation in survey sampling, *Journal of Applied Statistics,* **14**,143–152.

SUKHATME, P. V., and SUKHATME, B. V. and ASOK, C. (1984). *Sampling Theory of Surveys With Applications (2nd ed.),* Rome: Food and Agriculture: Organization.

SARNDAL, C. E. and SWENSSON, J. W.(1991). *Model Assisted Survey Sampling,* Springer-Verlag, New York.

STATISTICS IN TRANSITION-new series, July 2010
Vol. 11, No. 1, pp. 145—154

THE GENERALIZED FORMULA FOR AGGREGATIVE PRICE INDICES

Jacek Białek[1]

ABSTRACT

In the paper we propose a generalized formula for aggregative price indexes which satisfies most of the postulates coming from axiomatic price index theory. It is shown, that the ideal Fisher index and Lexis index are special cases of the proposed formula. Moreover, using the generalized formula we can easily define new indexes, which also satisfy given postulates.

Key words: aggregative price indexes, Laspeyres index, Paasche index, Fisher index, Lexis index.

1. Introduction

The contemporary economics makes use many statistical indexes in order to calculate the dynamics of growth of prices or quantities. The indexes make it possible to compare two periods of observations of the given economic processes. For example, the production in next periods depends on the former growth of prices, quantities and the profitability. Hence, it is very important to use the proper and well constructed indexes (see Fisher (1972)). The history of aggregative indexes is quite long - Laspeyres and Paasche indexes have been known since 19-th century (see Diewert (1978, Shell (1998)). Depending on the type of an economic problem we may also use one of the following indexes: Fisher ideal index (see Fisher (1922)), Törnqvist index (Törnqvist (1936)), Lexis index and other indexes (see Zając (1994), von der Lippe (2007)). The statistical indexes are very useful in economy (see Moutlon (1999), Seskin (1998)). Balk (1995) wrote about axiomatic price index theory, Diewert (1978) showed that the Törnqvist index and Fisher ideal index approximate each other. But it is really hard to indicate the best one of the statistical indexes (Dumagan (2002)). The choice of index depends on the information we want to get. If we are interested in dynamics of money in time we should use Fisher or Lexis indexes (see Zając (1994)).

[1] University of Lodz, Chair od Statistical Methods. Lodz, Poland.

From the theoretical point of view, a good index should satisfy a group of postulates (tests), coming from the axiomatic index theory (see Luszniewicz (1984), Balk (1995)). In this place we only mention about the following postulates: *linear homogeneity, proportionality*, i*dentity, time reversibility, circularity, commensurability* and *determinantness test* (see Luszniewicz (1984), Białek (2005), von der Lippe (2007)). Not every statistical index satisfies the mentioned tests. A system of minimum requirements of the index comes from Marco Martini (1992). According to the mentioned system, the price index should at least satisfy the following three conditions: *identity*, *commensurability* and *linear homogeneity.*

Let us consider a group of N components observed at times s, t and let us denote:

$P^s = [p_1^s, p_2^s, ..., p_N^s]$ - the vector of components prices at time s;

$P^t = [p_1^t, p_2^t, ..., p_N^t]$ - the vector of components prices at time t;

$Q^s = [q_1^s, q_2^s, ..., q_N^s]$ - the vector of components quantities at time s;

$Q^t = [q_1^t, q_2^t, ..., q_N^t]$ - the vector of components quantities at time t.

We consider only price indexes but all the discussion can be generalized to the case of quantity indexes. One of the most popular and often used price indexes are Paasche and Laspeyres indexes. Using the above notations the Paasche price index can be defined as follows:

$$I_{Pa}^p(Q^t, P^s, P^t) = \frac{Q^t \circ P^t}{Q^t \circ P^s}, \tag{1}$$

and Laspeyres price index:

$$I_L^p(Q^s, P^s, P^t) = \frac{Q^s \circ P^t}{Q^s \circ P^s}, \tag{2}$$

where "$\circ$" denotes a scalar product of two vectors.

As it is known, the indexes (1) and (2) do not satisfy some of the postulates for price indexes.

For example, the *time reversibility* for the price index I^p, described by formula (3)

$$I^P(Q^s, Q^t, P^s, P^t) = \frac{1}{I^P(Q^t, Q^s, P^t, P^s)}, \tag{3}$$

is not satisfied. Fisher proposed another definition based on Paasche and Laspeyres formulas (see Fisher (1922)). His formula I_F^p is a geometric mean of Paasche and Laspeyres indexes, it means:

$$I_F^p(Q^s,Q^t,P^s,P^t)=\sqrt{I_L^p(Q^s,P^s,P^t)I_{Pa}^p(Q^t,P^s,P^t)}\,. \tag{4}$$

The Fisher definition is called an "ideal formula", because it satisfies the tests mentioned above including *time reversibility*. Now we present a general class of price indexes.

2. The generalized formula for aggregative price indexes

Let $f_j : R_+^N \times R_+^N \to R_+^N$, for $j=1,2,...,m$ with some established $m \in N$, be such functions that for any $X,Y,Z \in R_+^N$ it holds

$$\sum_{j=1}^{m} \ln \frac{f_j(X,Y)\circ Z}{f_j(Y,X)\circ Z}=0\,, \tag{5}$$

or equivalently

$$\prod_{j=1}^{m} \frac{f_j(X,Y)\circ Z}{f_j(Y,X)\circ Z}=1\,. \tag{6}$$

Certainly the set of functions satisfying (6) is not empty – for example we could assume $f_j(X,Y)=c_j(X+Y)$ for $j=1,2,...,m$ and some $c_j \in R_+$.

Let us also notice that the condition (6) means equivalently that for any $Q^s,Q^t,P^s,P^t \in R_+^N$ it holds

$$\prod_{j=1}^{m} \frac{f_j(Q^s,Q^t)\circ P^t}{f_j(Q^t,Q^s)\circ P^t}=1\,,\ \prod_{j=1}^{m} \frac{f_j(Q^s,Q^t)\circ P^s}{f_j(Q^t,Q^s)\circ P^s}=1\,. \tag{7}$$

We propose the following, generalized formula for price indexes:

$$I^p(Q^s,Q^t,P^s,P^t)=[\prod_{j=1}^{m} \frac{f_j(Q^s,Q^t)\circ P^t}{f_j(Q^s,Q^t)\circ P^s}]^{\frac{1}{m}}\,. \tag{8}$$

Let us signify by $\Lambda(m)$ the class of price indexes defined by (6) and (8) for some established $m \in N$. It can be easily proved, that for any $k \in N$ we have (compare (28)-(31)):

$$I^P(Q^s,Q^t,P^s,P^t) \in \Lambda(m) \Rightarrow I^P(Q^s,Q^t,P^s,P^t) \in \Lambda(km)\,. \tag{9}$$

Moreover, each index from the class $\Lambda(m)$ satisfies most of the mentioned postulates. In particular, the following theorem is true:

Theorem 1

Let I^P be the aggregative price index with the structure described by (8) with additional condition (6). Then, the I^P index satisfies: *linear homogeneity*, i*dentity, proportionality, time reversibility* and *determinantness test***.** If we assume additionally that for each f_j it holds $f_j(\lambda X, \lambda Y) = \lambda f_j(X, Y)$, then we also have *weak commensurability* [1] satisfied.

The proof is quite easy. For example let us notice, that for some $\lambda > 0$ we have:

$$I^p(Q^s, Q^t, P^s, \lambda P^t) = [\prod_{j=1}^{m} \frac{f_j(Q^s, Q^t) \circ (\lambda P^t)}{f_j(Q^s, Q^t) \circ P^s}]^{\frac{1}{m}} =$$

$$= [\lambda^m \prod_{j=1}^{m} \frac{f_j(Q^s, Q^t) \circ P^t}{f_j(Q^s, Q^t) \circ P^s}]^{\frac{1}{m}} = \lambda I^p(Q^s, Q^t, P^s, P^t). \tag{10}$$

Thus, the *linear homogeneity* is satisfied. The *determinantness test* is also satisfied because we always have $I^p > 0$. Moreover, let us notice that using (7) and (8) we get

$$I^p(Q^t, Q^s, P^t, P^s) \cdot I^P(Q^s, Q^t, P^s, P^t) =$$

$$= [\prod_{j=1}^{m} \frac{f_j(Q^t, Q^s) \circ (P^s)}{f_j(Q^t, Q^s) \circ P^t}]^{\frac{1}{m}} [\prod_{j=1}^{m} \frac{f_j(Q^s, Q^t) \circ (P^t)}{f_j(Q^s, Q^t) \circ P^s}]^{\frac{1}{m}} =$$

$$= [\prod_{j=1}^{m} \frac{f_j(Q^t, Q^s) \circ P^s}{f_j(Q^s, Q^t) \circ P^s} \cdot \frac{f_j(Q^s, Q^t) \circ P^t}{f_j(Q^t, Q^s) \circ P^t}]^{\frac{1}{m}} =$$

$$= [\prod_{j=1}^{m} \frac{f_j(Q^t, Q^s) \circ P^s}{f_j(Q^s, Q^t) \circ P^s}]^{\frac{1}{m}} [\prod_{j=1}^{m} \frac{f_j(Q^s, Q^t) \circ P^t}{f_j(Q^t, Q^s) \circ P^t}]^{\frac{1}{m}} = 1^{\frac{1}{m}} = 1 \tag{11}$$

Thus, *the time reversibility* (3) is satisfied.

Certainly, if $P^s = P^t$ then $I^p(Q^s, Q^t, P^s, P^t) = 1$ - thus the *identity* is satisfied (by implication also the *proportionality*). The proof of *weak commensurability* is omitted.

[1] It means that $I^p(\lambda^{-1} Q^s, \lambda^{-1} Q^t, \lambda P^s, \lambda P^t) = I^p(Q^s, Q^t, P^s, P^t)$.

3. Special cases of the generalized formula

First of all let us notice, that in the special case of $m = 2$ and for

$$f_1(X,Y) = X\,, \tag{12}$$

$$f_2(X,Y) = Y\,, \tag{13}$$

where $X = (x_1, x_2, .., x_N)$, $Y = (y_1, y_2, .., y_N)$, we have for any $Z \in R_+^N$

$$\prod_{j=1}^{2} \frac{f_j(X,Y)\circ Z}{f_j(Y,X)\circ Z} = \frac{f_1(X,Y)\circ Z}{f_1(Y,X)\circ Z}\cdot\frac{f_2(X,Y)\circ Z}{f_2(Y,X)\circ Z} = \frac{X\circ Z}{Y\circ Z}\cdot\frac{Y\circ Z}{X\circ Z} = 1\,. \tag{14}$$

Thus, the functions described in (12) and (13) satisfy the assumption (6).

Using the formula (8) for $m = 2$, f_1 and f_2 we get the following structure of price index:

$$I^p(Q^s, Q^t, P^s, P^t) = [\prod_{j=1}^{2} \frac{f_j(Q^s, Q^t)\circ P^t}{f_j(Q^s, Q^t)\circ P^s}]^{\frac{1}{2}} = \sqrt{\frac{Q^s\circ P^t}{Q^s\circ P^s}\cdot\frac{Q^t\circ P^t}{Q^t\circ P^s}} =$$

$$= \sqrt{I_L^p(Q^s, P^s, P^t) I_{Pa}^p(Q^t, P^s, P^t)} = I_F^p(Q^s, Q^t, P^s, P^t)\,. \tag{15}$$

The formula (15) leads to the following conclusions: the Fisher index is a special case of the generalized formula defined in (8), it means $I_F^p(Q^s, Q^t, P^s, P^t) \in \Lambda(2)$.Let us also notice that if we assume $m = 1$ and $f_j(X,Y) = \frac{1}{2}(X+Y)$,where $X = (x_1, x_2, .., x_N)$, $Y = (y_1, y_2, .., y_N)$, we have condition (6) satisfied and from (8) we get

$$I^p(Q^s, Q^t, P^s, P^t) = \frac{f_j(Q^s, Q^t)\circ P^t}{f_j(Q^s, Q^t)\circ P^s} = \frac{\frac{1}{2}(Q^s + Q^t)\circ P^t}{\frac{1}{2}(Q^s + Q^t)\circ P^s}\,. \tag{16}$$

The formula (16) is a definition of Lexis[1] index (see Zając (1994)), which is presented in the literature of the subject as

[1] The formula (17) is also called the Marhall-Edgeworth index.

$$I_{Lex}^{p}(Q^s,Q^t,P^s,P^t)=\frac{\sum_{i=1}^{N}\frac{q_i^s+q_i^t}{2}p_i^t}{\sum_{i=1}^{N}\frac{q_i^s+q_i^t}{2}p_i^s}. \tag{17}$$

Thus for (16) and (17) we have the additional conclusion: the Lexis index is a special case of the generalized formula defined in (8).

Taking $m=1$ and $f_1(X,Y)=<1,1,...,1>$ we get the unweighted index of Dutot (1738).

The formula (8) allows to create new definitions of aggregative price indexes. For example taking $m=4$, $f_1(X,Y)=\frac{1}{2}(X+Y)$, $f_2(X,Y)=f_1(X,Y)$, $f_3(X,Y)=X$ and $f_4(X,Y)=Y$ we get the following structure of price index (assumption (6) is certainly satisfied):

$$I_{New}^{p}(Q^s,Q^t,P^s,P^t)=[\prod_{j=1}^{4}\frac{f_j(Q^s,Q^t)\circ P^t}{f_j(Q^s,Q^t)\circ P^s}]^{\frac{1}{4}}=$$

$$=\sqrt[4]{\frac{0.5(Q^s+Q^t)\circ P^t}{0.5(Q^s+Q^t)\circ P^s}\cdot\frac{0.5(Q^s+Q^t)\circ P^t}{0.5(Q^s+Q^t)\circ P^s}\cdot\frac{Q^s\circ P^t}{Q^s\circ P^s}\cdot\frac{Q^t\circ P^t}{Q^t\circ P^s}}=$$

$$\sqrt[4]{(I_{Lex}^{p})^2(I_F^{p})^2}=\sqrt{I_{Lex}^{p}I_F^{p}}. \tag{18}$$

Thus we have the conclusion, that the geometric mean of Fisher and Lexis indexes creates the new index, which also can be described by (8). Moreover, the following theorem is true:

Theorem 2

Let us assume that $I_1^P\in\Lambda(m_1)$ and $I_2^P\in\Lambda(m_2)$.

Let us also assume that $\exists k\in N\ \ m_2=km_1$.

Then we have $\sqrt{I_1^P I_2^P}\in\Lambda(2m_2)$.

Proof

The first assumption that $I_1^P\in\Lambda(m_1)$ means that

$$I_1^{p}(Q^s,Q^t,P^s,P^t)=[\prod_{j=1}^{m_1}\frac{f_{j1}(Q^s,Q^t)\circ P^t}{f_{j1}(Q^s,Q^t)\circ P^s}]^{\frac{1}{m_1}}, \tag{19}$$

and

$$\prod_{j=1}^{m_1} \frac{f_{j1}(Q^s, Q^t) \circ P^t}{f_{j1}(Q^t, Q^s) \circ P^t} = 1, \tag{20}$$

$$\prod_{j=1}^{m_1} \frac{f_{j1}(Q^s, Q^t) \circ P^s}{f_{j1}(Q^t, Q^s) \circ P^s} = 1. \tag{21}$$

Analogically, from the assumption $I_2^P \in \Lambda(m_2)$ we have

$$I_2^p(Q^s, Q^t, P^s, P^t) = [\prod_{j=1}^{m_2} \frac{f_{j2}(Q^s, Q^t) \circ P^t}{f_{j2}(Q^s, Q^t) \circ P^s}]^{\frac{1}{m_2}}, \tag{22}$$

and

$$\prod_{j=1}^{m_2} \frac{f_{j2}(Q^s, Q^t) \circ P^t}{f_{j2}(Q^t, Q^s) \circ P^t} = 1, \tag{23}$$

$$\prod_{j=1}^{m_2} \frac{f_{j2}(Q^s, Q^t) \circ P^s}{f_{j2}(Q^t, Q^s) \circ P^s} = 1. \tag{24}$$

Let us notice that in the case of $m_1 = m_2$ $(k = 1)$, defining $\tilde{f}_j(X, Y) = f_{j1}(X, Y)$ and $\tilde{f}_{j+m_2}(X, Y) = f_{j2}(X, Y)$ for $j = 1, 2, 3, .., m_2$, we have

$$\sqrt{I_1^p I_2^p} = \sqrt{[\prod_{j=1}^{m_2} \frac{f_{j1}(Q^s, Q^t) \circ P^t}{f_{j1}(Q^s, Q^t) \circ P^s}]^{\frac{1}{m_2}} [\prod_{i=1}^{m_2} \frac{f_{i2}(Q^s, Q^t) \circ P^t}{f_{i2}(Q^s, Q^t) \circ P^s}]^{\frac{1}{m_2}}} =$$

$$= \sqrt{[\prod_{j=1}^{2m_2} \frac{\tilde{f}_j(Q^s, Q^t) \circ P^t}{\tilde{f}_j(Q^s, Q^t) \circ P^s}]^{\frac{1}{m_2}}} = [\prod_{j=1}^{2m_2} \frac{\tilde{f}_j(Q^s, Q^t) \circ P^t}{\tilde{f}_j(Q^s, Q^t) \circ P^s}]^{\frac{1}{2m_2}}, \tag{25}$$

where, from (20), (21), (23) and (24)

$$\prod_{j=1}^{2m_2} \frac{\tilde{f}_j(Q^s, Q^t) \circ P^t}{\tilde{f}_j(Q^t, Q^s) \circ P^t} = 1, \tag{26}$$

$$\prod_{j=1}^{2m_2} \frac{\tilde{f}_j(Q^s, Q^t) \circ P^s}{\tilde{f}_j(Q^t, Q^s) \circ P^s} = 1. \tag{27}$$

Thus, for $m_1 = m_2$ we have $\sqrt{I_1^P I_2^P} \in \Lambda(2m_2)$.

Let us consider the case when $m_2 = km_1$ for some $k \in N$. From (19) we get:

$$I_1^p(Q^s, Q^t, P^s, P^t) = [\prod_{j=1}^{m_1} \frac{f_{j1}(Q^s, Q^t) \circ P^t}{f_{j1}(Q^s, Q^t) \circ P^s}]^{\frac{1}{m_1}} =$$

$$= \{[\prod_{j=1}^{m_1} \frac{f_{j1}(Q^s, Q^t) \circ P^t}{f_{j1}(Q^s, Q^t) \circ P^s}]^k\}^{\frac{1}{km_1}} =$$

$$= [\prod_{j=1}^{km_1} \frac{\hat{f}_{j1}(Q^s, Q^t) \circ P^t}{\hat{f}_{j1}(Q^s, Q^t) \circ P^s}]^{\frac{1}{km_1}}, \tag{28}$$

where

$$\hat{f}_{j1} = f_{j1 - \left\lfloor \frac{j1-1}{m_1} \right\rfloor m_1} \text{ for } j1 = 1, 2, ..., km_1. \tag{29}$$

It is easy to check that then we have

$$\prod_{j=1}^{km_1} \frac{\hat{f}_{j1}(Q^s, Q^t) \circ P^t}{\hat{f}_{j1}(Q^t, Q^s) \circ P^t} = 1, \tag{30}$$

$$\prod_{j=1}^{km_1} \frac{\hat{f}_{j1}(Q^s, Q^t) \circ P^s}{\hat{f}_{j1}(Q^t, Q^s) \circ P^s} = 1. \tag{31}$$

From (28), (30) and (31) we have $I_1^p \in \Lambda(km_1)$.

Thus, taking into consideration that $I_2^p \in \Lambda(m_2)$ and $m_2 = km_1$, we have the thesis:

$$\sqrt{I_1^P I_2^P} \in \Lambda(2m_2).$$

Remark 1

By using the theorem 2 we can explain, that $\sqrt{I_{Lex}^p I_F^p} \in \Lambda(4)$ - see the formula (18).

In fact we have $I_{Lex}^p \in \Lambda(1)$ and $I_F^p \in \Lambda(2)$.

Taking $m_1 = 1$, $m_2 = 2$ and $k = 2$ we get from the theorem 2: $\sqrt{I_{Lex}^p I_F^p} \in \Lambda(2m_2)$.

The immediate consequence of the theorem 2 is the next theorem:

Theorem 3

Let us assume that $I_1^P \in \Lambda(m_1)$ and $I_2^P \in \Lambda(m_2)$. Then $\sqrt{I_1^P I_2^P} \in \Lambda(2m_1m_2)$.

The proof is immediate – it is enough to notice, that the following implication is true:

$I_2^p \in \Lambda(m_2) \Rightarrow I_2^p \in \Lambda(m_1m_2)$ (compare formulas (28) and (29)). Taking $k = m_2$ we can use the theorem 2, because $I_1^P \in \Lambda(m_1)$ and $I_2^P \in \Lambda(\tilde{m}_2)$, where $\tilde{m}_2 = km_1$.

Remark 2

The theorem 3 is described by the following implication:

$$I_1^P \in \Lambda(m_1) \ \wedge \ I_2^P \in \Lambda(m_2) \ \Rightarrow \ \sqrt{I_1^P I_2^P} \in \Lambda(2m_1m_2).$$

Let us notice that the opposite implication is not true: from (15) we know that $I_F^p(Q^s, Q^t, P^s, P^t) = \sqrt{I_L^p(Q^s, P^s, P^t) I_{Pa}^p(Q^t, P^s, P^t)} \in \Lambda(2)$, but Paasche and Laspeyres indexes do not satisfy the *time reversibility,* thus $I_L^p(Q^s, P^s, P^t) \notin \Lambda(1)$ and $I_{Pa}^p(Q^t, P^s, P^t) \notin \Lambda(1)$.

4. Conclusions

The presented, generalized formula for aggregative price indexes satisfies almost all postulates coming from axiomatic price index theory. Thus, we have a practical conclusion: it is easier to prove that a given index belongs to the considered class than verify all mentioned postulates. It is shown, that the ideal Fisher index and Lexis index are special cases of the proposed formula. Moreover, using the generalized formula we can easily define new price indexes, which also satisfy given postulates. The geometric mean of the indexes coming from the considered class generates the new index, also belonging to this class.

REFERENCES

BALK M. (1995), *Axiomatic Price Index Theory: A Survey*, International Statistical Review 63, p. 69–95.

BIAŁEK J. (2005), *Indeks Törnqvista jako alternatywa dla idealnego indeksu Fishera,* Wiadomości Statystyczne, Wydawnictwo GUS, Warszawa, p. 9–21.

DIEWERT W. (1978), *Superlative Index Numbers and Consistency in Aggregation*, „Econometrica" 46, p. 883–900.

DOMAŃSKI CZ.., red. (2001), *Metody statystyczne. Teoria i zadania*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

DUMAGAN J. (2002), *Comparing the superlative Törnqvist and Fisher ideal indexes*, Economic Letters 76, p. 251–258.

FISHER I. (1922), *The Making of Index Numbers*, Boston: Houghton Mifflin.

FISHER F. M.(1972), *The Economic Theory of Price Indexes*, Academic Press, New York Krzysztofiak M., Luszniewicz A. (1997), *Statystyka,* PWE, Warszawa.

MARTINI M. (1992), *A General Function of Axiomatic Index Numbers*, Journal of the Italian Statistics Society, 1 (3), p. 359–376.

MOUTLON B., SESKIN E. (1999), *A preview of the 1999 comprehensive revision of the national income an product accounts*, Survey of Current Business No. 79.

SHELL K. (1998), *Economics Analysis of Production Price Indexes*, Cambridge University Press, UK.

TÖRNQVIST L. (1936), *The Bank of Finland's consumption price index*, Bank of Finland Monthly Bulletin 10, p. 1–8.

VON DER LIPPE P. (2007), *Index Theory and Price Statistics*, Peter Lang, Frankfurt, Germany.

ZAJĄC K. (1994), *Zarys metod statystycznych*, PWN, W-wa.

STATISTICS IN TRANSITION-new series, July 2010
Vol. 11, No. 1, pp. 155—160

POLISH STATISTICS DAY

Czesław Domański

In response to the initiative of 29th Session of the United Nations, March 9 has been declared the NATIONAL DAY OF POLISH STATISTICS. The Main Council of the Polish Statistical Society in agreement with the Committee for Statistics and Econometrics of the Polish Academy of Sciences and the Governor of the Central Statistical Office agreed that the Day was to commemorate the anniversary of the first General Census ordered by the Four Year Sejm on 9 March, 1789.

The first national population censuses, which appeared in Poland in the second half of the 18th century, were connected with a growing interest in population issues in the times of Enlightenment. In 1777 the first census of urban population was held and it was repeated several times since then. Ten years later a one-day census of Warsaw population took place; it was supposed to provide data on the number of inhabitants and their socio-occupational structure. The census encompassed residents of particular houses who registered by name.

In his speech delivered on 9 March, 1789 a deputy for the Sejm – Fryderyk Józef Moszyński (1737–1817) pointed at inequalities in tax burden among voivodships, and disproportions in the representation of the country in the Sejm. His thesis was illustrated by a precisely compiled calculation table which was presented to the deputies and which „showed the overall picture of the Republic of Poland, including not only smokes but also a proportion of souls, a multitude of incomes and distances of the state".

Moszyński claimed that "the wealth of the state cannot be measured by the affluence of several aristocratic families and several thousand of rich citizens but it rather should be measured by settlements and wealth of towns and country, prosperous trade and flourishing crafts". The statistical measure of imposing tax for military purposes, which was proposed by Moszyński in the Sejm, "was absolutely unique in character and it was used nowhere else either before or after that time„. In order to provide objective calculation of value of goods in a given poviat (district) Moszyński proposed a statistical method based on the following data:

- value of land and property in the poviat on the basis of deeds of sale recorded in district books in the last 11 years which was representative enough for making calculations

- number of smokes obtained from treasure tariffs both alienated in the last 11 years and those which were not subject to purchase-sale transactions.

The information allowed the Treasury Commission to make calculations based on the value of the alienated goods and provide a precise estimation of the value of goods in a given poviat taking into account the proportion between the alienated goods and the total.

Fryderyk Józef Moszyński was a keen supporter of the so-called co-equation, that is a fair fiscal system which imposed taxation of profits obtained from landed and church properties.

Thanks to Moszyński's efforts the Four Year Sejm (1788–1792) passed an important legal act, which was a milestone for the development of public statistics in Poland. The document of 22 June, 1789 known as "Inspection of smokes and population register" became in fact the first ever national census in Poland. It encompassed the rural and urban population whereas landed gentry and clergy were excluded. The categories included in the register were: sex, occupation and social status which differentiated between sons (below and above the age of 15) and daughters in the family.

As it was mentioned above, regular population censuses started to be carried out in the 18th century. The national census was conducted in Sweden in 1749 and Poland followed in 1789. The Polish census preceded the first one organized in the United States in 1790, the census of 1795 in the Netherlands and in many other European countries.

The census of 1789 included not only the number of population but also its social and occupational structure. Although the census intended for tax and military purposes did not encompass „the privileged classes", i.e. the gentry and the clergy, it became the basis for estimating population of the Republic of Poland at the end of the 18th century. The Four Year Sejm ruled that the natural movement of population should be constantly registered as of 1 January 1790.

The census lasted for nine months and it continued well into the year 1790. The urban and rural populations were registered separately; they were also divided into followers of Christian and Jewish religions. The representatives of landed gentry took an active part in the census; they gave overall numbers of inhabitants of particular village in register papers. Young single men were divided into two groups – one included the persons up to the age of 15, and the other – men of the age 15 and above. The division was connected with the conscription to be held at that time. However, the conscription combined with insufficient experience of census takers prevented proper execution of orders of the state authorities. The results of register of smokes and population, which were presented by Fryderyk Józef Moszyński at the forum of the Parliament in April 1790, were published as a special supplement to the "Trade Daily" later the same year. The census was of summary character, that is it showed the number of inhabitants populating towns and villages of the Republic of Poland. In 1790 a register by name was also taken on the whole territory of Poland.

Simultaneously with the work conducted on the population census, Moszyński presented and implemented his idea of making the church authorities responsible for establishing and keeping records of births and deaths, in order to update the figures obtained in the census. It is worth stressing that those registers were also to encompass gentry.

The statistical tables presented by Moszyński acquired him the name of „an outstanding statesman „not only in Poland but also abroad. T. Korzon, who was quite reluctant to praise Moszyński, had to admit that ”his numbers of space and population were of highest possible level of reliability; not only the ones obtained from official registers but also those based on speculation” (1 smoke – 6 souls) and which he accepted as the basis for the estimation of the population of Poland in the period of the Partitions.

Moszyński was a pioneer of the Polish statistical thought of the second half of the 18th century. His numerous initiatives led to establishing a well-functioning system of registering population and non-cash elements of the national economy. The general census of 1789, keeping constant records of the natural movement of population, and the inspection of smokes provided a sound foundation for national registering and statistics. It was much needed during his lifetime, and much appreciated in the years to come.

It seems worth reminding here that in 1789 the territory of Poland was equal to 520 square kilometres, and according to the data taken down, Poland had 1,434,919 smokes (the notion of smoke can be compared to the notion of household understood as household of either the peasantry or the bourgeoisie), and it was populated by 7,660,787 inhabitants. However, the census did not include the gentry and the clergy, whose estimated numbers were equal to 750–800 thousand.

Polish census of 1789 is ranked among one of the most pioneer undertakings when modern censuses are considered. Therefore, Fryderyk Jozef Moszyński – an economist, a statistician, a political activist and a statesman – fully deserves to be called the father of the public statistics. He was also the author of a project according to which organizations representing different religions were obliged to register births and deaths in the whole population, including the gentry, and to compile and send in annual reports on natural movement of population to the bodies of the state administration (Cf. Konferowicz (1961)).

The next two censuses were taken in 1808 and 1810 in the Duchy of Warsaw. They were the general (full) population censuses which encompassed the condition of population and its socio-occupational structure. In order to examine and evaluate the census data the Statistical Office, directly subordinate to the Ministry of Internal Affairs, was established in Warsaw in 1810. The population censuses were not organized in the first half of the 19th century.

A real breakthrough in population censuses came in 1853 when the International Congress of Statistics in Brussels recommended that general population censuses should be based on uniform rules and conducted every 10 years.

Not always were those rules followed by many of the European countries. In the second half of the 19-th century only a few towns and provinces organized their local population censuses in the Kingdom of Poland – for instance the Radom Province (1868) and Warsaw (1882). The first general population census in the Kingdom of Poland was conducted as late as 1897, and it was simultaneously the census of the whole Russian Empire. Demographic characteristics such as religious affiliation, marital status, education and socio-occupational structure were investigated.

In Galicia of the first half of the 19th century mainly the conscription censuses were carried out; they were meant to gather and examine information related to men of military age. The first ever population census was conducted there in 1857 to be followed by the subsequent censuses of 1869, 1880, 1890, 1900 and 1910. The censuses conducted in Galicia were taken by name, and they included a wide spectrum of demographic, socio-occupational and religious characteristics, as well as physical and mental defects. Moreover, the first two censuses also examined nationality.

On the territory of Poland under the Prussian rule the censuses were taken every three years, starting from 1816 until 1867. The next census was held on 1 December, 1871 and since that time they were organized every five years.

The results of population censuses conducted on the territory of Poland under the supervision of statistical offices of the three Partitioning Countries (Austria, Prussia and Russia) at the end of 19th and the beginning of 20th century were in many respects incomparable due to various statistical methods and ambiguous terminology used by them. It was extremely difficult to make comparisons of numerical data on different social and economic issues related to those territories.

After Poland had regained its independence in 1918, the Central Statistical Office conducted two general population censuses: on 30 September,1921, and 9 December, 1931. The basic characteristics investigated in those two censuses were: sex, age, marital status, nationality (only in the census of 1921), religion, mother tongue, literacy, education level, primary and secondary source of income, place of residence and register of housing buildings.

The results of the population census of 1921 provided information not only related to demographic and socio-occupational structure of the population, but also to national and religious structure of each particular town or village, what was presented in the form of register in alphabetical order. Another register, compiled with the use of data gathered during the 1931 census, was already ready to go to press but it was destroyed in 1939.

The Central Statistical Office had planned to carry out a population census in 1941 yet those plans were never put into operation because of the outbreak of the World War II.

When Poland was under the Nazi occupation the German authorities conducted the general summary population census on 1 March, 1943.

In the post-war period seven general population censuses were conducted on: 14 February,1946; 3 December, 1950; 6 December,1960; 8 December,1970; 7 December,1978; 6 December,1988; and 21 May–8 June, 2002.

The 1946 census was the summary census and it only examined the general condition and structure of population according to three age groups (below 18 years of age, 18–59 years, and 60 years or above). The results of the census presented the age structure of the Polish population after the World War II. The first three censuses encompassed a range of problems similar to the problems included in the pre-war censuses. The socio-occupational characteristics were examined in a more detailed way, however, there were no questions about nationality, religion or mother tongue.

When compared with 1950, the census carried out in 1960 investigated the marital status of the population in more detail. In the census of 1970 the questions about place of birth, and field of education were introduced, and for the first time fertility rate was examined with the use of the representative method. The fertility study was repeated in the census of 1988. Disability, both in the biological and legal sense, was investigated in 1978 and 1988.

The census of 2002 encompassed not only residents living permanently and temporarily but also houses and buildings. Fertility rate among Polish women was examined again.

Due to the fact that a lot of information gathered in general population censuses became outdated very quickly, the Central Statistical Office conducted additional population censuses with the use of the representative method (the so called „microcensuses") on 30 March, 1974, 6 December 1984 and in 1995. They contained all the main characteristics (including disability), which were used in the previous censuses to describe population structure.

The year to come, 2011, will be the year when the next National Census of Population and Housing will be conducted. It is worth stressing that it will be the first general census since Poland became the member state of the European Union. A number of obligations which result directly from this membership include – among other things – the necessity to provide socio-demographic and socioeconomic information at the time and range defined by the European Commission.

The General Census provides the most detailed information on the number of population, its territorial distribution, socio-demographic and occupational structure, as well as socioeconomic characteristics of households and families, housing conditions and resources at all levels of the territorial division of the country: national, regional and local.

REFERENCES

BERGER J. (1993) Badania statystyczne na ziemiach polskich do 1918 r., Wiadomości Statystyczne nr 7,

KONFEROWICZ S. (1961), Fryderyk Józef Moszyński – Statystyka doby Sejmu Czteroletniego, Wydawnictwo SGPiS Warszawa,

KOZŁOWSKI Cz. (1951), Powszechne spisy ludności, Wydawnictwo Gospodarcze Warszawa,

ROMANIUK K. (1975), Udział Polski w pracach Międzynarodowego Instytutu Statystycznego, Wiadomości Statystyczne nr 8,

RZEPKOWSKI A. (2005), Spisy ludności na ziemiach polskich w latach 1789–1939, Przegląd Nauk Historycznych R. IV nr 2 (8).

STATISTICS IN TRANSITION-new series, July 2010
Vol. 11, No. 1, pp. 161—175

THE ECONOMIC ASPECTS OF CONTRADICTION BETWEEN GENERATIONS IN POLAND

Izabella Kudrycka[1], Małgorzata Radziukiewicz[2]

1. Introduction

The conflicts between generations exist in every time and in every country, and concern many domains of human existence. The main reasons for conflicts may be determined as the differences in psychological and behavioral characteristics of each generation. These differences have important influence on the principles, preferences, choices and culture of the particular generation. For example, the generation of flowers – children ("hippies") was quite different from the next generation – generation of "yuppies". For hippies the happiness, lazy life with friends in commune, and narcotics were the crucial targets, while the economical success and individual carrier at job were the most important for "yuppies", however for that two generations the conflict among them and older generations was common.

The conflicts between generations are very interested matter from the point of view of psychology, sociology and culture, but our study is devoted to the economic aspects of that phenomenon.

The paper consists of five main parts and conclusions. The first one is devoted to the presentation of general topics of the paper, that is the identification of contradiction areas, while next four contain the analysis of the particular contents.

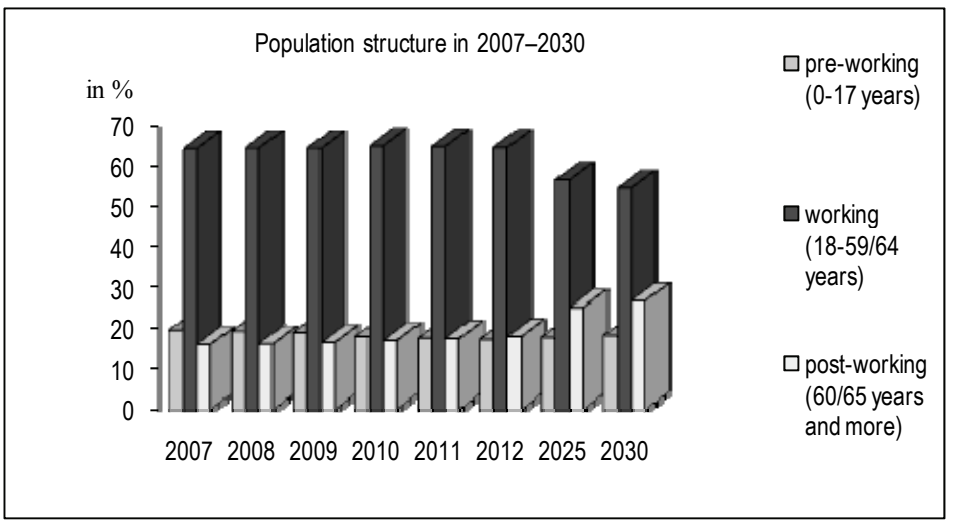
2. The areas of contradictions between generations

First, it is necessary to determine what the term "generation" means. In our opinion, generation is the group of people, or households living in that same finite period of time, and sharing the time with some part of previous generations. Each generation can be divided by separate three groups: people belonging to the educational period, people active on the labour market and retired people. The distribution by age of any particular generation and law determining the minimum level of education, together with regulations concerning the minimum age for retirement, have influence on the structure of generation. The impact of economy

[1] University of Finance and Management in Warsaw.
[2] Institute of Market, Consumption and Business Cycle Research in Warsaw.

takes also meaningful part in this partition, although the demographic structure in previous time periods has the conclusive role. Using the age limits 0–17 for pre-working time, 18–59/64 for working and post-working, the structure of mentioned above three groups in Poland, in last and future years is presented on Figure 1.

Figure 1. Population structure in last 4 years and projection on 2010–2030 (in percentage)



Source: on the basis: "Rocznik Demograficzny Polski", GUS, Warszawa, 2008.

Such dynamic changes of structure – steady growing share of post-working population and diminishing shares of people in working activity age – must be the reason of the contradiction between generations due to some of the economical aspects. The indices of population growth of the third group in 2015, related to number for 2007 is 121.3% and 152.7% in 2030. The appropriate indices for the first group are equal to 92.4 % and 83.5%, what means the reduction of the population in pre-working period, what will provide to the diminishing of the birth rate in future, and diminishing the total number of population in Poland.

For the purpose of our consideration we divided the economical aspects of contradiction between generations on the following parts:

- Social Security System,
- Credit and Debts on macroeconomic level,
- Welfare, Income, Consumption and taxation,
- Revenue and Expenditure of the State Budget.

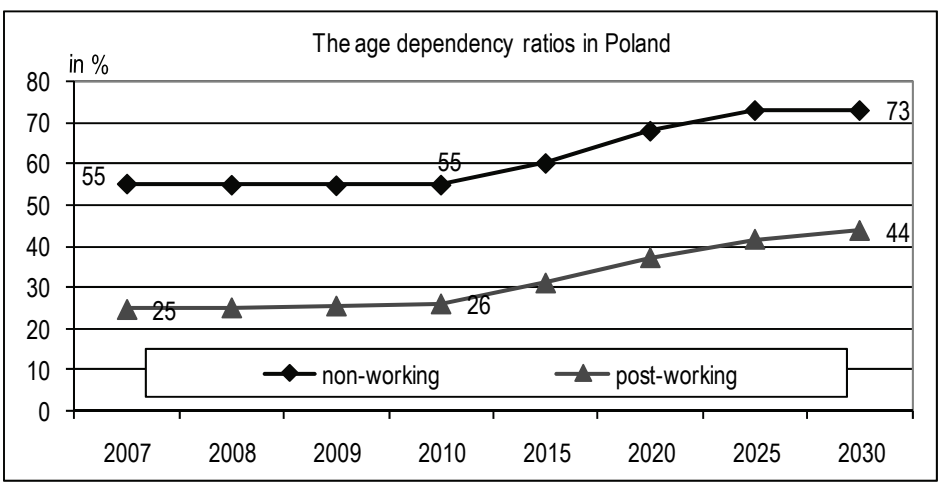
All above areas of contradictions between generations will be presented in detail in the next parts of the paper, and illustrated by the statistical data.

3. Social Security System

In Poland, and also in almost all European countries social security system is so called "pay as you go", what means that pensions of the retired persons are funded from the contributions of working group of each generation. When the social security system does not own any capital, the benefits received by the third group of generation are equal to the contributions of the second generation group. So, now it is obvious that such demographic structure of generations as we can observed in Poland, must be the reason of contradiction between generations. Additionally, the life expectancy for women in age 60 was in 2008 – 23.1 (more 1.6 years than in 2000), and for man in age 65 was in 2008 – 17.9 (more than 1.2 in 2000), and will be increasing in future.

The age dependency ratios are growing in Poland in last 3 years and will be growing in future (see Figure 2).

Figure 2. The age dependency ratios in Poland, 2007–2025



Source: on the basis: "Rocznik Demograficzny Polski", GUS, Warszawa,2008.

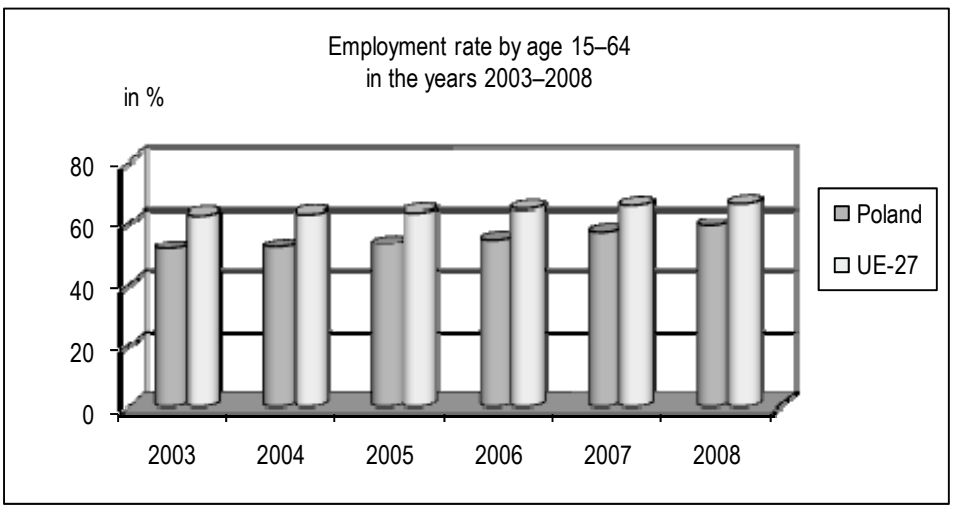
As we can observe, for 100 working persons the number of post-working is growing from 25 in 2007 and 2008 to 42 in 2025. Adding the number of non-working persons we obtain 73 persons dependent on 100 persons in working age in 2025. In comparison with average dependency ratios for the European Union, which was and is predicted for future 2030, it is 36 (for 100 working persons the number of post-working).

Such growth of age dependency ratios means that young generation must pay too much to the social security system, what leads to intergenerational reallocation of resources and may be the source of conflict between generations.

The age dependency ratios may be worsening if we take into account the real number of working persons and also the form of employment. The real number of

working persons depends on the situation on labor market and also on balance of migration for persons in working age. In the last year we can observe increasing unemployment rate, and in 2004–2008 increasing emigration of young people to old European countries, what means the reduction of payments to the social security system in Poland.

Figure 3. Ratio of employment in Poland and in the EU

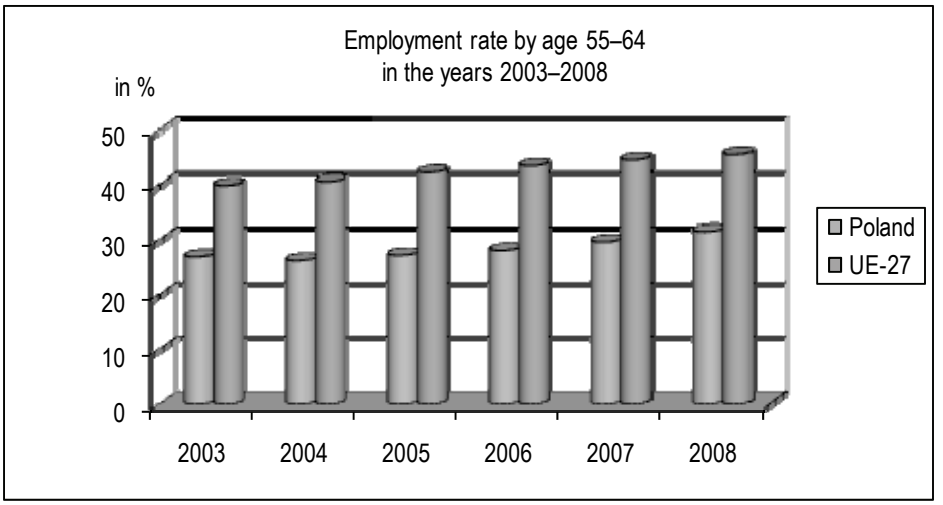


Source: on the basis: "Pracujący w gospodarce narodowej", GUS, Warszawa, 2009.

We also observed the special forms of unemployment which enable to avoid payments to the social security system. Share of employees in total number of working is about 74%[1]. Generally, the ratio of employment in Poland is lower than in the EU (see Figure 3 and Figure 4) and has negative impact on the inflow of funds to social security system.

[1] It means that 26% people are working on part time or without payments to social security system.

Figure 4. Ratio of employment in Poland and in the EU



Source: on the basis: "Pracujący w gospodarce narodowej", GUS, Warszawa, 2009.

The reasons of such low employment rate in Poland are: the large number of disabled people among the population at activity age, the law regulations, which establish the relatively early age for retirement (60 for women and 65 for men) and high unemployment rate, especially in some regions of Poland.

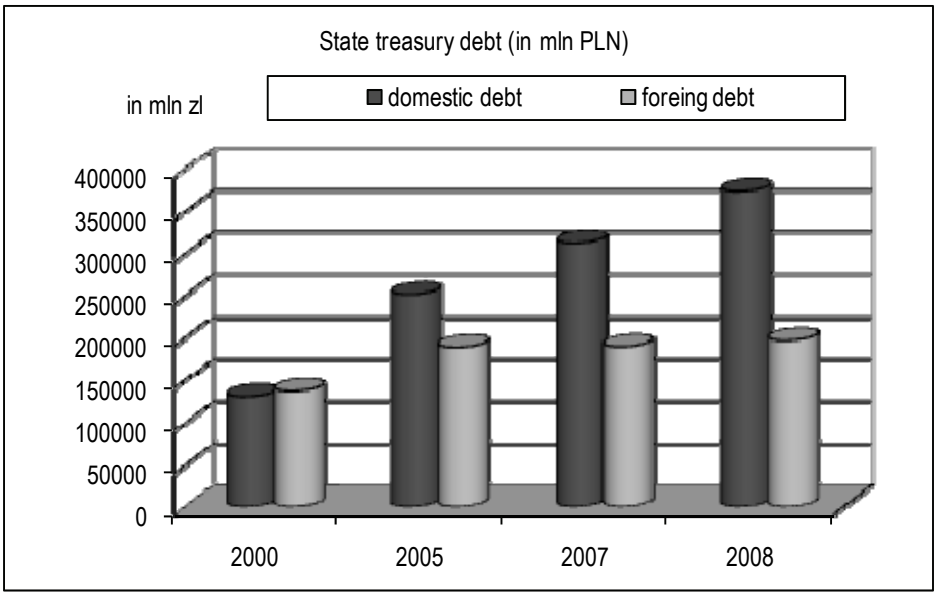
4. Credits and Debts on Macroeconomic Level

The debts are also a source of contradiction between generations, because the credits which were taken by one generation become the debts of the next generations. For example, the last rate of credits which were taken in the 70's (so called the Gierek`s debts) was paid in 2009. So, enlargement of welfare of one generations is the reason of "tightening belt" for the next generations.

It is necessary to recognize the differences between the sources of credits and allocation of the credit funds. The extremely worse is such situation, when credits are provided only for consumption[1] and there will not be any profits from it in future. Better, if the investment will be the destinations of credits and there are positive expectations concerning the future profits. It takes place when the investments are allocated in high technologies or in infrastructure, which has impact on economic growth.

[1] However growth of consumption has positive influence on economy but only in short time horizon.

Figure 5. Credits and cumulated debts in Poland on macroeconomic level



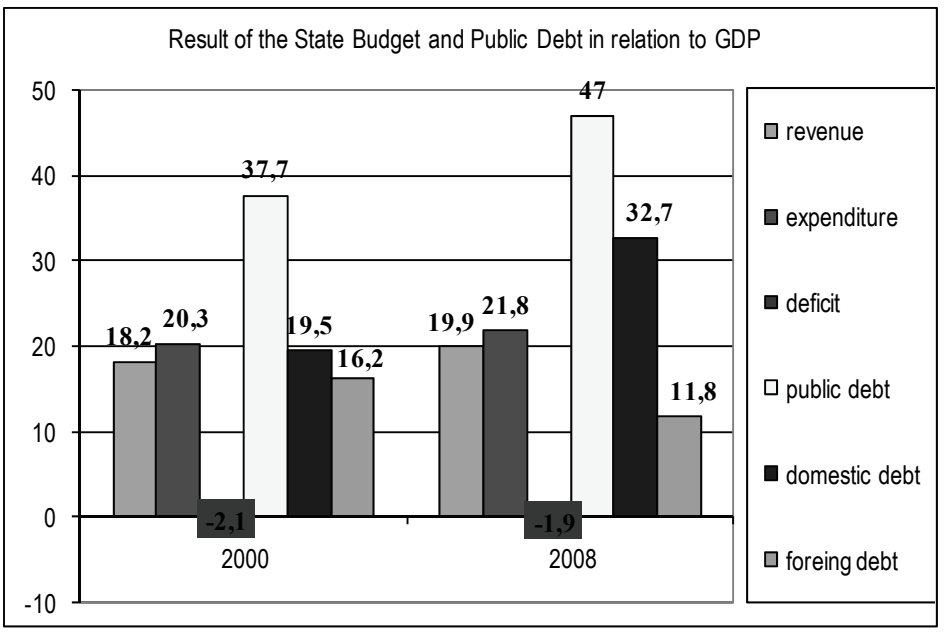
Source: on the basis: "Statistical Yearbook of the Republic of Poland 2009", GUS, Warszawa, 2010.

We can observe the growth of debts, and also their shares in GDP, especially the domestic debt. In 2008 the share of total debt was greater than in 2000 about percentage points.

The negative impact of previous generation debts can be neutralized by enlarging the taxation or intergenerational altruism, if old generations take care about the welfare of descendant generations, and be able to gather some welfare and transfer it for their descendants. It concerns not only government, but also households.

The sources of credits may be internal and external. The internal credits depend on the possibility of economy to generate savings, and such source of credits allows to neglect the exchange rate of foreign currencies in case of Poland not belonging to common currency – Euro, and also the changes of interest rates in countries borrowing the money. But on the another hand, the foreign credits may mainly enlarged the investment, and be additional source of growth in the future, with positive influence on welfare of descendant generations.

Figure 6. The Shares of debts in GDP



Source: on the basis: "Statistical Yearbook of the Republic of Poland 2009", GUS, Warszawa, 2010.

The debt situation in Poland is not satisfying, because of increasing debts of the government budget and low economic growth, expected in few next years. It means that next generations will be debit with that debts, and the balance of intergenerational justice may be weakened.

5. Welfare, Income, Consumption and Taxation

The comparison of incomes of old and young generations is rather complicated because of lack of the appropriate information. Some knowledge may be based on the income per head, or better equivalent income in households of retirees & pensioners, and employees` households. The average monthly per head available income in 2003–2008, achieving by those households is presented in the Table 1.

Table 1. Average monthly available and equivalent income in households

Households	2003		2007		2008	
	income (in PLN)					
	available	equivalent	available	equivalent	available	equivalent
Total	712,0	969,7	937,4	1294,6	1045,5	1460,3
Employees	760,6	1059,3	916,6	1313,0	1049,8	1507,7
Retirees	852,0	1046,1	999,9	1240,8	1096,9	1351,3
Pensioners	608,1	750,8	758,4	950,8	802,4	1006,6

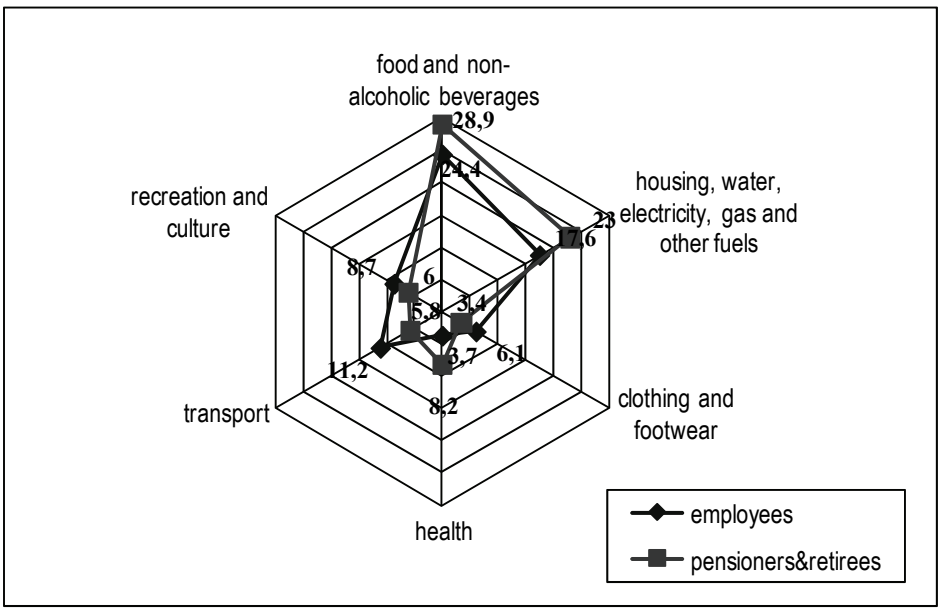
Source: own calculation on the Household Budgets, GUS, Warszawa, 2009.

As we can observe in the period 2005–2008, per head incomes were slightly lower in employees households than in retirees & pensioners households. In 2008 per head income of employees was significantly higher than income of retirees & pensioners.

The equivalent income takes into account the number of consumer units instead of number of persons in households, and it is a better indicator of available income. In 2008 the average monthly equivalent income for all households was 1460.3 zl, while in households of retirees was about 7.5 % and of pensioners about 31% lower (in 2007 was lower about 4.2 % and 26.6%). In the case of non-manual employees households, equivalent income was higher than average – about 1.4 % in 2007 and 3.2% in 2008.

Such crucial differences between per head and equivalent incomes are induced by differences in number of persons in households of employees and in households of retirees & pensioners, in which generally one or two persons live in one household. That characteristic and also different consumer preferences of young and old are the reasons of great differences in structure of consumer expenditures. The position of some chosen expenditures of young and old are presented on Figure 7

Figure 7. The shares of chosen expenditure groups in employees` households and households of retirees & pensioners in 2008



Source: on the basis of the Household Budgets, GUS, Warszawa, 2009.p

The differences in the expenditure structure are meaningful. The expenditure shares on food, housing (including gas, water, electricity and another fuels), and health are lower in employees households than these shares in households of retirees & pensioners. For the rest expenditures, the shares of employees` households are significantly higher than appropriate shares of households of retirees & pensioners, especially in case of transportation.

The stocks of all durables are significantly higher in the employees` households than in households of retirees & pensioners. The differences between equipment level are usually greater than 20 percentage points as it takes place in case of DVD, a digital camera, a personal computer, a mobile telephone and a car. Excluding cars and PC`s, the rest of durables exist on the consumer market for a relatively short time, and it may be partly the reason of such low stocks in the households of old people, who are less interested in new technology gadgets. However, the most important reason is the lack of money.

The welfare of two types of households may be presented by equipment of some durables see Table 2.

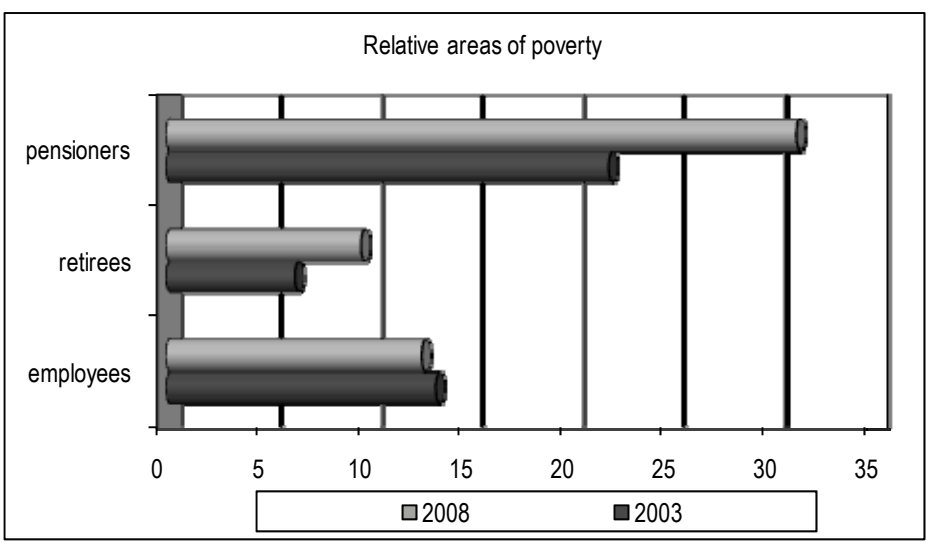
Table 2. Households furnished with selected durable goods

Specification	Households (in % of given group of households)					
	employees		retirees		pensioners	
	2005	2008	2005	2008	2005	2008
Satellite or cable television equipment	57.0	63.6	43.7	48.5	34.9	41.9
DVD player	34.7	61.6	8.9	26.0	8.5	25.3
Hi-fi stereo music system	62.0	57.2	20.5	20.9	22.6	21.5
Video camera	8.7	11.7	2.7	3.5	1.4	2.5
Mobile phone	86.1	97.7	36.5	59.8	40.3	60.5
Home theatre system	–	21.6	–	5.2	–	4.4
Personal computer	57.0	76.0	14.9	23.4	15.4	24.5
of which with access to the Internet	33.7	62.5	8.9	18.4	7.3	17.2
Car	60.1	67.4	31.9	33.0	22.1	22.5

Source: on the basis: "Konsumpcja w Polsce. Raport roczny.", IBRKiK, Warszawa, 2010.

The proof of such reason is the subjective area of poverty determined by the households of retirees, households of pensioners and employees` households in 2007 (see Figure 8).

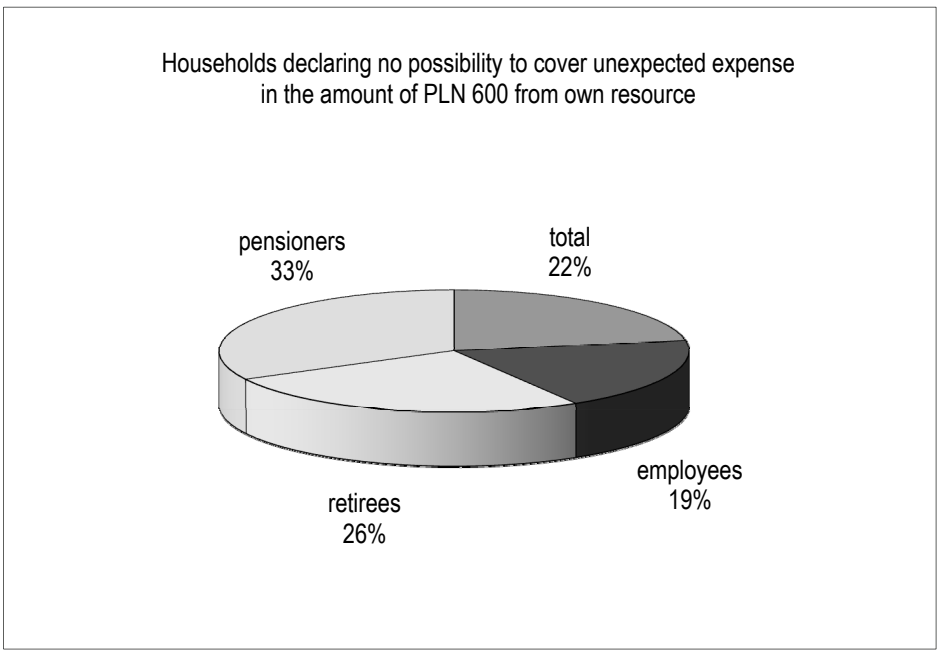
Figure 8. The relative area of poverty in 2003–2008 (percent of households)



Source: own calculation on the basis of the Household Budgets, GUS, Warszawa, 2009.

It is worth to stress the significant difference between relative and subjective areas of poverty, because the first one is based on the per head incomes and in the case of households of retirees & pensioners it does not fit the real poverty area.

Figure 9. The subjective area of poverty in 2008 (percent of households)



Source: on the basis of the Household Budgets, GUS, Warszawa, 2009.

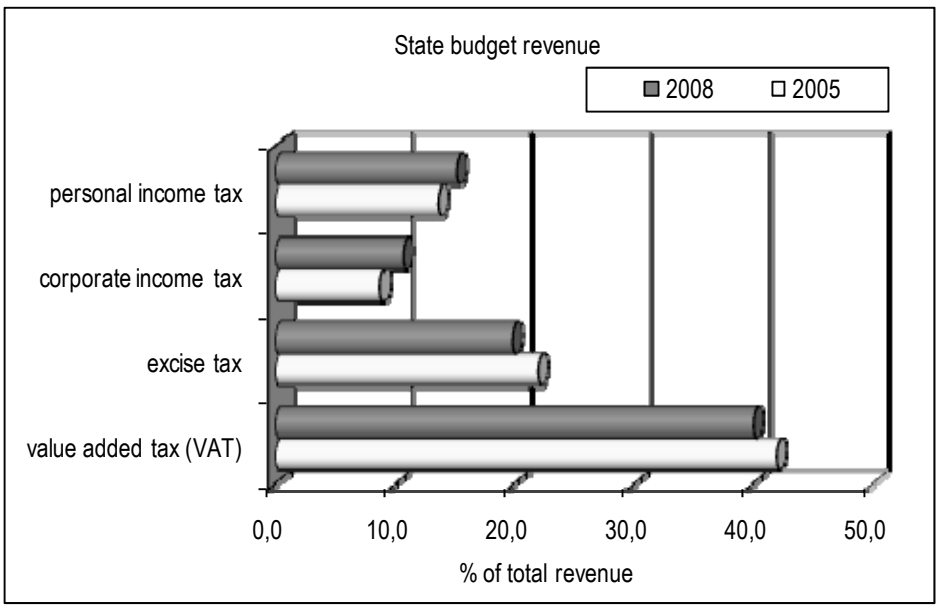
Concluding, there exists a great gap between welfare of employees` households and households of retirees & pensioners, on the gain of the first, and it may be also the reason of contradiction between old and young generations.

6. Revenue and Expenditure of The State Budget

Budget revenue

It is not easy to determine what is the participation of generations in revenue of state budget. We can only generally recognize what is the impact of young and old generation on the main sources of state budget revenues. Looking at the structure of the state budget revenues (see Figure 10) we can suppose that participation of young generations in VAT and excise tax is greater than old generations in this more important source of revenue.

Figure 10. Structure of The State Budget Revenue in 2005 and 2008



Source: on the basis: "Statistical Yearbook of the Republic of Poland 2009", GUS, Warszawa, 2010.

Young people are characterized by greater propensity to consumption and take the greater part in demand for durables, especially new modern goods, which have the higher VAT rates than food. Food with low VAT rates is more important for elderly (most households of retirees & pensioners spend their low incomes for food and housing).

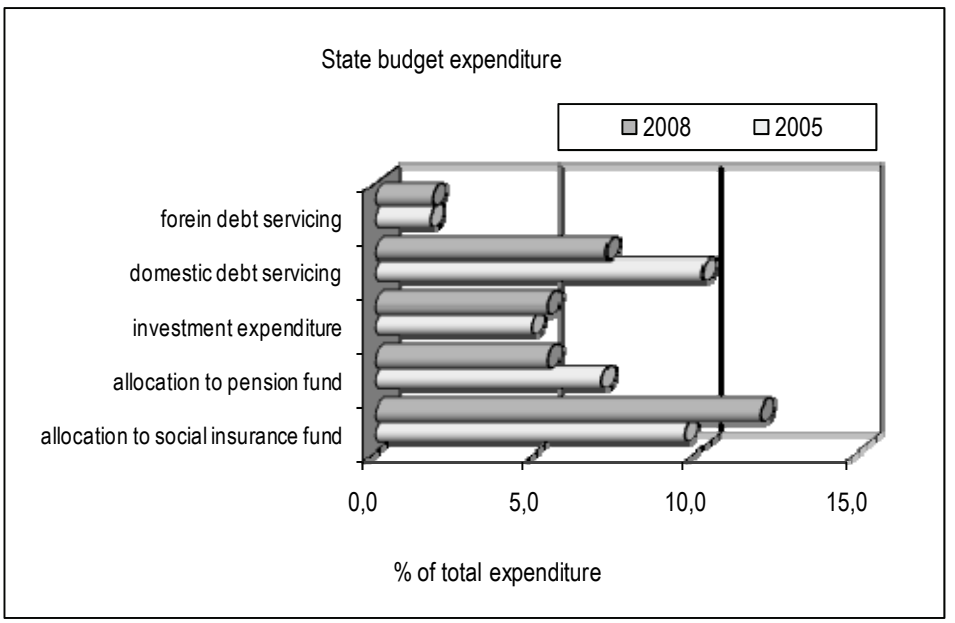
Another important source of revenue is income tax including personal and corporate income taxes. In this case we also can suppose that generations 30+ have the greatest part in generating income personal taxes, and also corporate income taxes. Old generations consist mostly of retirees persons who are not the owners of firms, and receiving low incomes from Pension Funds.

Budget expenditure

This part of paper is devoted to the directions of spending funds from The State Budget, especially how structure of the expenditure affects the well-being of the generations in present time, and in the future. It is obvious that there is the strong competition between the purposes of financing the current targets – usually connected with public consumption, and future targets, which have investment character. The young generations are more interested in such expenditures which will be fruitful in the future, however the current public consumption is also important for them – for example outlays on pre-school education, all level of

school and university education, environment and infrastructure investment, and partly healthcare.

Figure 11. The Structure of State Budget Expenditures in 2005 and 2008 (in per cent)

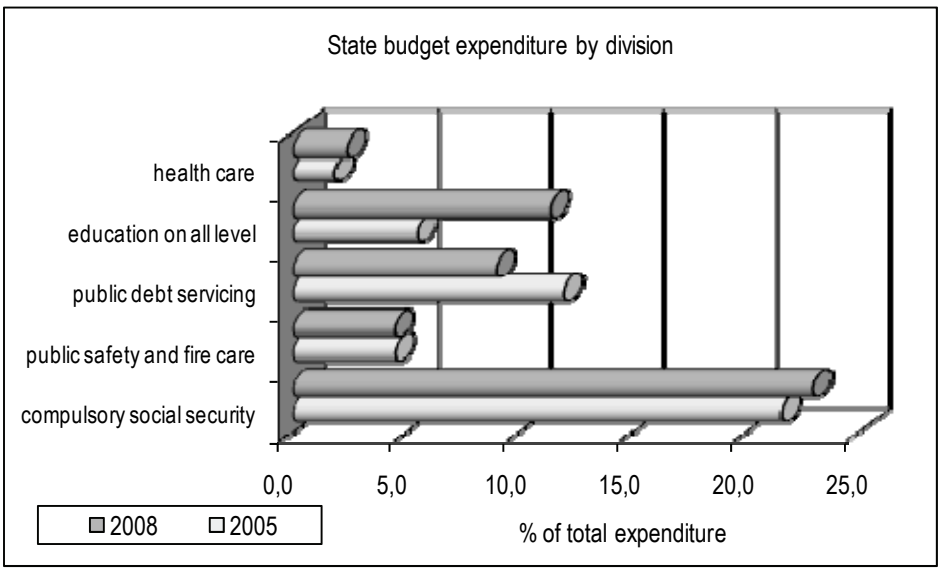


Source: on the basis: "Statistical Yearbook of the Republic of Poland 2009", GUS, Warszawa, 2010.

For the old generations the most important is the stability of Social Insurance Fund, specialized healthcare, and sufficient number of places in homes and facilities for elderly. As we could see before the share of expenditure for health is significantly large in households of retirees & pensioners, containing mostly money for medicine, so also the support for that expenses is important. It is also obvious that old generations are not so much interested in allocation of expenses for long-term investment.

Taking under consideration some kind of the state budget expenditure we can divide it into three main groups: I – expenditure mainly connected with young generations, especially expenditure on education on all levels; II – expenditure mainly connected with old generations – subsides to the Pension Fund; III – neutral – for example expenditure on Justice or National Defense, and IV – Public Debt Servicing.

Figure 12. The Structure of State Budget Expenditures by the group of Expenditure in 2005 and 2008 (percentage of total expenditure)



Source: on the basis: "Statistical Yearbook of the Republic of Poland 2009", GUS, Warszawa, 2010.

7. Conclusions

*The main aspects of economical contradictions between young and old generations concerns:

- Social Security System,
- Credit and Debt on macroeconomic level,
- Welfare, Income, Consumption and Taxation,
- Revenue and Expenditure of the State Budget.

** The demographical structure of Polish society, worsening meaningfully in future, and actual social security system (especially retired and pensions system) are the reasons for excessive intergenerational reallocation of resources. The reformation of the social security system is necessary to avoid such negative consequences of the above.

***Actual debts and necessity of further credit contracts induce the growth of public debts, and costs of servicing those debts in future. The consequence of that is very high share of cumulative debt in GDP (in 2009 – 47%), what may be the restriction of economy growth, and intergenerational reallocation of resources. The negative impact of previous generation debts may be (in some part) neutralized by enlarging the tax and intergenerational altruism.

****The equivalent income is lower in households of pensioners & retirees than in households of employees in the last three years, what means that the situation of old generation is worse than the young's one, especially when analyzing the structure of consumption in that two types of households (in retirees & pensioners the share of necessity expenditure – housing and food – is meaningfully higher than in employees households). The equipment of young households is also significantly better than the one of old households. It is obvious that also subjective poverty area is greater in old households than in young ones.

*****Analysis of the state budget revenue and expenditure leads to the conclusion, that the current input of young generations to the budget revenues is greater than the input of the old ones (but old generations were young in past) , while shares of budget expenditures connected with transfers to the old generations are greater than transfers to the young.

STATISTICS IN TRANSITION-new series, July 2010
Vol. 11, No. 1, pp. 177—186

STATISTICAL CHOICE BETWEEN RATING AND RANKING METHOD OF SCALING CONSUMER VALUES

Piotr Tarka[1]

ABSTRACT

The article describes how many market researchers go wrong with statistical assumption as for the right choice in rating and ranking method application within consumers' values evaluation. Many mistakes still arise owing to ineffective comparison of one scale to another, which in fact constitutes a completely different method of analysis and usage. The author discusses advantages and disadvantages of both ranking and rating methods pertaining to further statistical possibilities in data processing against the theoretical background and later on given brief example. Next stage of description emphasizes the importance of both rating and ranking method, depending on type of consumer (subject) being interviewed, hypothetical type of question being asked, communication with subject or even data results to be obtained.

1. Key sources on the theory of value

Rokeach (1968) defined a value as 'a type of belief about how one ought or ought not to behave, or about some end-state of existence worth or not worth attaining'. The theory defines values as desirable, transsituational goals, varying in importance, that serve as guiding principles in people's lives. The crucial content aspect that distinguishes among values is the type of motivational goal they express. Different contents of values has three universal requirements of human existence: biological needs, requisites of coordinated social interaction, and demands of group survival and functioning. Groups and individuals represent these requirements cognitively as specific values about which they communicate in order to explain, coordinate, and rationalize behaviour. In fact, theoretical interest in values is spread across many disciplines including economics, sociology, psychology, political science and history. Values, then, are the most

[1] University of Economics in Poznań, Department of Strategy and Policy of International Competiveness Policy, al. Niepodległości, POLAND, e-mail: piotr.tarka@ue.poznan.pl.

abstract type of socio-economic cognition that people use to store and guide general responses to classes of stimuli.

Basically, the challenging questions belong to the branch of philosophical inquiry known as *axiology* or *the theory of value* (Hilliard, 1950; Taylor, 1961; Hartman, 1967). A crucial point is that one can understand a given type of value only by considering its relationship to other types of value. We can understand one type of value only by comparing it with other types of value to which it is closely or not-so-closely related.

Typical definition of consumer value in the market context can be taken in the following way [….]. *Value represents an interactive relativistic preference experience.* Typically, such consumer value refers to the evaluation of some *object* by some *subject.* Here, for our purposes, the "subject" in question is usually a consumer or other consumer, whereas the "object" of interest could be any product – a manufactured good, a service, a political candidate, a vacation destination, a musical concert, a social cause, etc.

2. Rating or ranking method for values measurement – what choices to make?

Researchers are still divided over the preference for rating or ranking methods applied for measuring values. However, the debate originates from the false premise that one method must be used to the exclusion of the other. A conceptualization of the value structure that uses characteristics of both rating and ranking systems opens up theory and research to a more complex understanding of values (Ovadia, 2004).

One way that research on values differs substantially from research on other science topics is the inexact definition of the object of analysis. While quantitative items such as wages and fertility are (largely) agreed upon in their definition and how they are to be measured, values do not have a single, unified definition nor a universal model of measurement. The choices that researchers make in defining what a value is not only determining the nature of the item of analysis (the value structure) but also the method of measurement (ratings or rankings). Therefore, a meta-analysis of values research must address two related questions. First, how are values structured in an individual's mind? Second, how does one measure a value?

In a ranking method, the respondent is simply asked to place a list of values in order of importance. This has been the method used by Rokeach in his Value Survey (1968), and many other studies. The other method of measuring values is a rating system, in which values are rated by the respondent independently of one another, typically on a Likert scale or some variant thereof. This approach has been used in the General Social Survey and modified versions of the Rokeach Values Survey and the Schwartz Values Survey. From the statistical point of view, rating scales are important instruments used to measure people's attitudes

towards a variety of stimuli including organizations, products, services, places, institutions, and advertisements. Like all measuring devices, the rating scale is only useful if it provides reliable and valid measurements. There is a considerable amount of research dealing with the problems of constructing questions and rating scales that are reasonably objective and relatively free of measurement errors.

3. Quantitative and qualitative variables differences

It is not a brand new discovery that quantitative variables assume numerical values, whereas qualitative variables assume nominal values indicated usually by some names. However, it sometimes happen that numbers are practiced in order to either denote the value of quality variable (e.g. shopping centre called M1 and measured as number 1) or mark and simply symbolize a name in itself "M1". In world of science, based on empirical research, undertaken analysis on different variables is becoming even more interesting, provided the increase in intensity level among variables can be measured. In this sense measurement is defined as a procedure and process of numbers assignment to different values stated on variable (in psychometrics literature – "constructs") with commonly accepted rules. These rules are set and remain usually arbitrary, which means all measurement units (e.g. length measured in: centimetres, kilometres or miles etc.) are imposed symbolically by community members. Therefore, when human values are considered, we have at our disposal the background of multiple **theoretical constructs**, or so called ideas (Francuz, Mackiewicz 2005). For them a set of rules of measurement and their comparison operation on all their levels as for selection of perfect unit should be initially proposed. Obviously, this measurement differs considerably from subjects or variables expression on natural or real units (e.g. length), where for example each unit falling on "length" variable is simply summed up according to assigned measurement unit (centimetres or inches). In case of constructs one needs to decide on selecting the observable rate or evaluation system to measure constructs in order to proceed further.

One way of considering constructs (subjects' values) as quantitative variables and conducting computations based on their structure or state of preference is to perceive constructs and process of their measurement according to some degree of intensity and their appearance on given numerical value. Construct (value) can be formed if there are higher (based on ascending intensity) numbers allocated (to construct) by subject (respondent). In other words, the greater allocation as for some construct, the better creation of final value and reference point from which any comparisons can be started (Sagan, 2004).

4. Statistical option in data analysis and field work (data collection) within Rokeach (RVS) and Schwartz scales

Despite its widespread use, there have been critiques of the RVS that are relevant to most ranking ones, even though Rokeach developed multiple methods for administering the survey in attempts to simplify the task for the ranking tool. First, a ranking question cannot be easily asked over the telephone, which has become the main mode of survey research in the last three decades. Secondly, requiring a respondent to place 18 items (on questionnaire) in an order of preference is a lengthy process when compared to asking for individual ratings (Szreder, 2004). Rokeach himself admitted that 'many respondents report the ranking task to be a very difficult one – one they have little confidence in having completed in reliable manner and one they are often sure they had completed more or less randomly'. Thirdly, rankings result in a data set that cannot be analyzed with standard statistical methods because of the interdependence of the ranks. If the rankings of 17 of the values in the RVS have been indicated by the respondent, the value of the final value is predetermined. This characteristic of the data (sometimes referred to as being 'ipsative') requires researchers to make complex adjustments in order to perform statistical analyses of the results.

In contrast, the rating system is advocated for its simplicity in execution and analysis, as the respondent can quickly and easily respond to each item without having to compare the value with others on the list of questionnaire. A number of researchers have also found that using a method in which each item is independently evaluated does not lead to a significantly different rank ordering of the items on the list. However, rating measurement have also been subject to considerable criticism. Researchers have found that rating tools allow low-and nonmotivated respondents to give largely uniform responses ('non-differentiation') and thereby understate the differences among values. Rating systems often yield statistical ties between values, which may be the result of indifference to the question, rather than a true equivalence of importance. Because the ranking procedure forces the respondent to give each value a different assignment, it prevents nondifferentiation, irrespective of motivation level. However, it is also possible that ranking procedures force individuals to overstate the differences between values and make comparisons of values that the respondent considers non-comparable or potentially make random responses to meet the requirements of the survey.

Generally respondents using a rating system often cluster their responses within a narrow subset of the range of options, but no such 'underdispersion' is possible in a ranking system where the respondent must use the entire range of the scale. Also rating systems cannot ensure that the scale is used consistently, either between or within respondents. It is possible that one respondent considers a score of 70 (out of 100) to be a high level of importance, while another respondent considers 70 to be a moderate level of importance. Although a rating system is clearly better than a ranking system for assessing the distance between values for

a single respondent, ratings cannot completely ensure that within-respondent ratings are consistent.

5. The pros and cons pertaining to either methods application

Since it is clear that both methods for measuring the consumers' values have strengths and weaknesses, one should ask what each method assumes about the object of analysis itself, the value system. In other words, what is the value system and how is it organized? By looking more closely at the methods, one can see how the conclusions about value systems are not only residuals of the method of data collection, but are expressions of the characteristics of the overall value system that are assumed to exist in the individual. Therefore, the question of ratings or rankings is not only the question of a method, but of the nature of values and their organization in the mind. In an ipsative system, such as the Rokeach Value Survey or any other ranking system, values are arranged in a zero-sum structure by definition. If the ranking of one value increases by one rank, another value must decline by one rank. For Rokeach and other advocates of the ranking method, values represent mutually exclusive choices. In situations that call multiple values into possible action, one must be prioritized over the other(s). This means that as value structures change over time or differ across groups, the higher importance of one value must come at the expense of the importance of another value.

In contrast, a rating method does not require changes in the importance of some values to be compensated for by changes in other values. Since the respondent has the ability to give each value any rating without regard to the ratings given to other items, the ratings of the values have no restrictions. The value system that is assumed in a rating system is one in which the importance of values are independent; there is no limit on the total amount of importance that is distributed among the values.

In general, when researchers are interested in measuring values belonging to subjects (respondents), they often use a rating rather than a ranking method because it is easier and faster to administer and yields data that are amenable to parametric statistical analyses. However, because subjects' values are inherently positive constructs, subjects often exhibit little differentiation among the values and end-pile their ratings toward the positive end of the scale. Such lack of differentiation may potentially affect the statistical properties of the values and the ability to detect relationships with other variables.

6. Brief example

An example of how these differences can exist were considered in a hypothetical value structure containing only four values: **accomplishment, happiness, pleasure,** and **salvation** (Tarka, 2008). Each of these values can be

associated with a specific amount of importance that is later scaled from 0 (no importance) to 100 (maximum importance). Now, consider two individuals (subjects) with the following distributions of importance.

Graph 1. Answers importance by two subjects making judgments

1 subject:	**Accomplishment - 90,**	**Happiness - 80,**	**Pleasure - 70,**	**Salvation - 40**
2 subject:	**Accomplishment - 70,**	**Happiness - 60,**	**Pleasure - 50,**	**Salvation - 30**

Source: own construction

In the very first view ranking survey would show that these two people have the same value structure. Each respondent would report that accomplishment is most important, followed by happiness, pleasure, and salvation, respectively. However, a rating method would show that there are substantial differences between subject (1) and subject (2) in their values judgement. Firstly, subject (1) places more importance on these values overall than subject (2). Second, (s.1) places more importance on any specific value than (s.2), even though they rank the value identically.

Depending on the question being asked in regard to the market values of these individuals, conclusions about the values of these individuals may differ. If the question is about which values will be acted upon in situations where choices must be made, the ranking and rating systems will both point us in the same direction. However, if the questions of interest are about the relative importance of a value for (s.1) as compared to (s.2), the method will lead to different conclusions: ranking will indicate no difference whereas rating will show a difference. In particular, we would conclude that while happiness has the same importance for each individual relative to the other values in the structure, the amount of importance that (s.1) assigns to happiness is greater than the value assigned to it by (s.2). Neither a ranking nor rating measure by itself would reveal this pattern. Both of these conclusions are equally valid and therefore, the 'most correct' answer about their values is one that reveals both facts.

A second advantage in this approach is shown when one introduces change in the value system of (s.1). Assume that over time or in response to some stimulus, the importance of accomplishment for (s.1) declines from 90 to 87, whereas the importance of happiness increases from 80 to 83. The ranking approach would conclude that there has been no change in the value system of (s.1) over time. This is true, as we would expect that in circumstances where accomplishment and happiness were to come into conflict, (s.1) would choose accomplishment at both times of measurement. However, the rating system also tells us that the importance of accomplishment has declined for (s.1) over time, while the importance of happiness has increased. If these observations were part of a study to determine whether (s.1) had been affected by some external stimulus, such as an buying a new car or clothes, **then ranking and rating approaches would**

lead us to two completely different conclusions. The best conclusion uses both pieces of information: the stimulus affected a change, but it was not large enough to change the ordering of the values of interest.

7. Still ongoing transformation dilemmas on ranking and rating scale measure

In market research, the assumption that many statistical (especially multivariate) procedures require at least is intervally scaled data (Green, Tull, 1978). This supposed relationship between measurement scales and statistical techniques can be traced to Stevens (1966) and represents the representational theory of measurement. In contradiction to this view, it is sometimes said that in mathematical statistics literature one will not find scale properties as a requirement for the use of various statistical procedures. The point being made here is that the assumptions for the use of a particular statistical procedure are based solely on the mathematical aspects underlying that procedure and nothing else.

Borgatta and Bohrnstedt (1980) argue that the question of measurement assumptions in statistical applications is often treated inappropriately by market researchers. This occurs because unobserved constructs like values are assumed to be continuous variables (approximately) normally distributed in the population of interest – by definition a metric (interval) variable. When such constructs are measured as discrete variables, they are interval scales with (sometimes considerable amounts of) error. This was a classical theory view of measurement where three primary types of ordinal data were identified (Kim, 1975): [(1) ordered categories, (2) ranking, (3) test and summated rating scales construction], and also a set of simple transformations used by many scientists: [(a) assign each category an evenly spaced numerical value that preserves the original order, (b) convert ranks into corresponding numbers, and (c) consider the given values as valid metric scores].

In literature, procedure of transformation from ranks into numbers is nothing else than just assigning meaning categories to definite numbers with expressions: *better, worse, worst*. A positive number is assigned and starting point falls on lowest number (for a rank) with number 1 – *better*, continuing through 2 – *worse* (next rank) 3, 4, 5 and so on. Problems associated with this sort of transformation appear on the gaps between numbers. In ranking scales we hardly differentiate accurate distances between numbers which are contrary to interval, rating scales where distances are easily measured and most of arithmetic computations can be performed. For that reason, in practice, subjects' characteristics are measured on rating scales where they can express their attitudes and opinion easily using numbers ranging from 1 to 7 or 1 to 5, at which subjects can point up human value as the most important for them (7) or as the least important (1). Because subjects' answers given on ratings (based on this type of scale) are perceived and treated as part of scale numbers, neighbouring numbers adjoining from both sides

to chosen number by subject are sometimes summed up and later single numerical value is retrieved. This numerical value makes up complementary and indicates importance of value (Walesiak, 1996), (Zaborski, 2001).

Virtually transformations make the unrealistic assumption of equal distances between each pair of scale categories and probably do not produce "true" interval scales under most conditions. To use this approach the researcher should have at least assumed that the errors and distortions introduced by the transformation are minor and that they do not affect conclusions drawn from the data analysis. Some researchers have already developed a wide variety of more complicated models which transform paired comparisons or rankings into unidimensional and multidimensional measures. These scales are thought to produce a better fit between the operationalized variable and the latent construct. The Thurston Case procedure – the normalized method of rank order, and similar method of successive categories – (Guilford, 1954) are examples of methods used to derive unidimensional (interval) scales from order-type comparisons. For example, let us assume the following experiment conducted on sample consisting of 250 respondents (men) ranking seven cars on two dimensions: 1). stated preference and 2). perceived quality of the car (Tarka, 2008). Guilford explained that a complex transformation of the rank data can be performed by using the method of successive categories to obtain an interval scale. In this sense each rank can be treated as a category of an underlying unidimensional continuum for which we derive an interval level value for the midpoint of each category. These values then replace the individual responses. A comparison of the original rank values and their rescaled equivalents can show that these methods effectively compress the scale (rank values = 1 to 7; above mentioned “preference” = 1.00, 1.86, 2.41, 2.84, 3.25, 3.72, 4.52; “quality” = 1.00, 1.86, 2.39, 2.81, 3.22, 3.68, 4.47. The new scale values differ by no more than 0.05 across the two dimensions. This implies that respondents are using comparable scaling categories to evaluate the set of quality elements associated with a car on each dimension. In order to prove it, one can further perform a goodness-of-fit test between the new scale values and the original data, indicating that the method of successive categories (or similar one: normalized method of rank order) cannot be rejected. The rest of nonmetric scaling methods (generating multidimensional metric scales from order data) can be found in works of Green and Rao (1972).

8. Conclusions

Comparison of ranking and rating scales and methods of evaluation in human values suggests that there are no perfect types of statistical instruments in evaluation genuine and actual state of human values. This is by reason of values, being hidden in subjects’ mind, which are deeply taking hold of subjective and mental expressions given by subjects (e.g. consumer’s evaluation of product or service). “Values” are still obscure and hard to express for people. For statistical

researchers they are sometimes inscrutable or impenetrable, no matter what type of scale is applied. In fact. this is subject to acceptance and we all have to leave with it. However, the truth is also that ranking and rating scales can be a valuable measurement instrument provided they are applied appropriately. Ultimately, each scale and method has its own unique way of values measurement (in approximate way), depending on research objective or subject being explored. Any transformation procedures from ranking to rating / interval scale (if ever undertaken) should be implemented very carefully. Finally, one can conclude that human values and their expressions are formed (or rather generated from value-items) not only under specific scale or techniques of statistical analysis but also under the influence of a number of factors such as: subject's determination to answer a posed question, time of interview or type of questionnaire, communication with them or even subjects' situational humour.

REFERENCES

BORGATTA, E.F., BOHRNSTEDT. G.W. (1980), *Level of measurement: Once over again,* Sociological Methods and Research, Vol. 9, No. 2, pp. 148–160.

FRANCUZ, P., MACKIEWICZ, R., (2005), *Liczby nie wiedzą skąd pochodzą.* Wyd. KUL Lublin

FRONDIZI, R. (1971), *What is value. An introduction to axiology,* 2nd edition, La Salle, IL, Open Court Publishing Co.

GREEN, P.E., TULL. D.S. (1978), *Research for marketing decisions,* Englewood Cliffs, Prentice – Hall.

GREEN, P.E., RAO. V.R. (1972), *Multidimensional scaling: a comparison of approaches and algorithms,* New York, Holt, Rinehart and Winston Publishing Co.

GUILFORD, J.P. (1954), *Psychometric methods,* New York, McGraw – Hill.

HARTMAN, R.S. (1967), *The structure of values,* Carbondale, IL, Southern Illinois University Press.

HILLIARD, A.L. (1950), *The forms of value: the extension of hedonistic axiology,* New York, Columbia University Press.

KIM, J. (1975), *Multivariate analysis of ordinal variables,* American Journal of Sociology, Vol. 8, No. 2, pp. 261–298

ROKEACH, M. (1968), *The nature of human values,* New York, The Free Press.

OVADIA, S. (2004), *Ratings and rankings: reconsidering the structure of values and their measurement*, Journal Social Research Methodology, Vol. 7, No. 5, pp. 403–414

SAGAN, A. (2004), *Badania marketingowe.* Wyd. AE Kraków

STEVENS, S.S. (1966), *Handbook of experimental psychology,* New York, John Wiley and Sons.

SZREDER, M. (2004), *Metody i techniki sondażowych badań opinii*, Warszawa, PWE

TAIWO, A., HERSHEY H. F., (200), *Overall evaluation rating scales: an assessment,* International Journal of Market Research Vol. 42, Issue 3, pp. 301–312

TARKA, P., (2008*), From ranking (Rokeach – RVS) to rating scales evaluation – some empirical observations on multidimensional scaling Polish and Dutch youth's values* – Innovative Management Journal, Vol. 1, No 2, pp. 24–42

TAYLOR, P.W. (1961), *Normative discourse,* Englewood Cliffs, New York, Prentice – Hall.

WALESIAK, M., (1996), *Metody analizy danych marketingowych*. PWN Warszawa

ZABORSKI, A., (1991), *Skalowanie wielowymiarowe,* Wyd. AE we Wrocławiu.

STATISTICS IN TRANSITION-new series, July 2010
Vol. 11, No. 1, pp. 187—203

CONTIGUITY MATRIX OF SPATIAL UNITS AND ITS PROPERTIES ON EXAMPLE OF LAND DISTRICTS OF PODKARPACKIE PROVINCE

Wiesław Wagner, Andrzej Mantaj[1]

ABSTRACT

Spatial units are an important carrier of numerical data on tested statistical characteristics of selected social-economic phenomena. They are in the relation of contiguity to one another, which can be expressed by the binary matrix of contiguity. Its 1-elements indicate occurrence of common boundary, and zeros – no boundary. In the work various divisions of spatial units have been presented as well as their properties and description of contiguity allowing determination of the contiguity matrix. It stands out with many analytical properties. Its illustration has been shown on land districts of Podkarpackie Province.

Key words: Spatial units, non-directed graphs, contiguity matrix, land districts of Podkarpackie Province

1. Introduction

Spatial units as statistical objects frequently occur in taxonomical analyses due to the tested various complex social-economic phenomena (e.g. Zeliaś 1991, 2000, Grabiński 2003). These units are carriers of multidimensional numerical data about a tested set of statistical characteristics and are evaluated in respect of their intensity. It allows creation of various types of classification of units and their illustration on statistical maps.

Spatial units are in relation of contiguity to one another, sharing the common boundary. This relation for some two units can refer to the very fact of sharing the common boundary, which is expressed binary-wise by 1 – contiguity occurs and 0 – no contiguity, or concerns occurrence of common territorial boundary expressed in its length, which is specified by real values, where contiguity occurs (e.g. km) or zero, when there is no such contiguity. Spatial contiguity can be also considered by appurtenance of common areas of forest complexes, nature reserves

[1] Wyższa Szkoła Informatyki i Zarządzania w Rzeszowie, wwagner@wsiz.rzeszow.pl, amantaj@wsiz.rzeszow.pl.

or protected landscape reserves, and the like, as well as common areas of economic activity which are expressed in surface units (e.g. ha, sq km) for each of the units.

The paper presents the characteristics of spatial units, their properties and common contiguity description giving the basis for determination of the contiguity matrix. For these matrixes various properties were given. It was presented for land districts of Podkarpackie Province.

2. Land districts of Podkarpackie Province

Statistical units used for illustration of contiguity matrix in this work are land districts of Podkarpackie Province (PDK). It is divided into 21 land districts and 4 town districts (Rzeszów, Krosno, Przemyśl and Tarnobrzeg) (fig. 1).

Figure 1. Land and town districts of Podkarpackie Province

Data on area and population number of land districts of PDK province have been given in table 1. They are arranged in ascending order according to their area. The table presents also population density per 1 sq km and the percentage

participation of the area and population of land districts in their general size in the province.

Table 1. Numerical data on districts. State on 31.12.2008

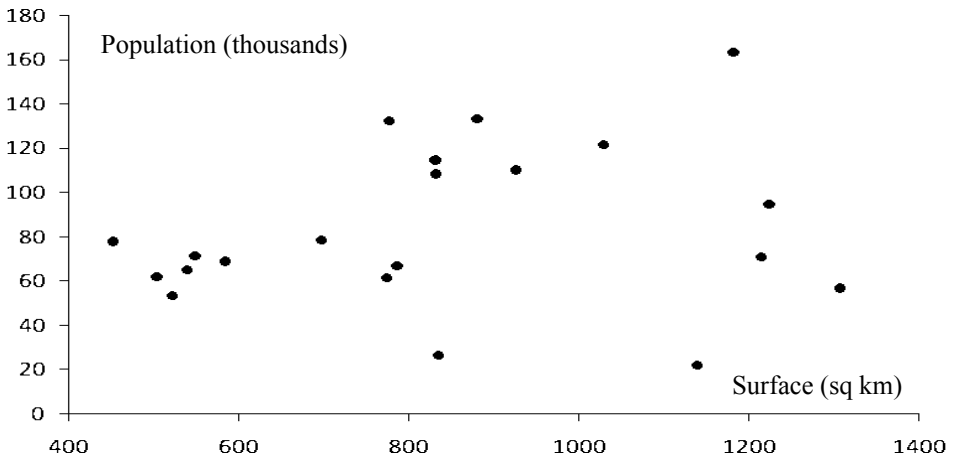
No.	Districts	Code	Area	Population	Population density	Percentage participation of	
						Area	Population
1	Łańcucki	ŁAN	452	77.9	172	2.6	4.4
2	Strzyżewski	STRZ	503	61.9	123	2.9	3.5
3	Tarnobrzeski	TAR	521	53.6	103	3.0	3.0
4	Brzozowski	BRZ	539	65.1	121	3.1	3.7
5	Ropczycko-Sędziszowski	R-S	548	71.3	130	3.1	4,0
6	Leżajski	LEŻ	584	69	118	3.3	3.9
7	Przeworski	PRZW	697	78.6	113	4.0	4.5
8	Kolbuszowski	KOL	774	61.4	79	4.4	3.5
9	Dębicki	DĘB	777	132.6	171	4.4	7.5
10	Niżański	NIŻ	786	67	85	4.5	3.8
11	Jasielski	JAS	831	114.8	138	4.7	6.5
12	Stalowoski	STAL	832	108.5	130	4.7	6.2
13	Leski	LES	835	26.5	32	4.7	1.5
14	Mielecki	MIEL	880	133.3	151	5.0	7.6
15	Krośnieński	KROŚ	926	110.3	119	5.3	6.3
16	Jarosławski	JAR	1,029	121.8	118	5.9	6.9
17	Bieszczadzki	BIESZ	1,139	22.1	19	6.5	1.3
18	Rzeszowski	RZESZ	1,182	163.4	138	6.7	9.3
19	Przemyski	PRZ	1,214	71.1	59	6.9	4.0
20	Sanocki	SAN	1,224	94.7	77	7.0	5.4
21	Lubaczowski	LUB	1,307	56.9	44	7.4	3.2
Sum			17,580	1,761.8	100	100	100

Source: The authors' elaboration on the basis of: „Powierzchnia i ludność w przekroju terytorialnym w 2008 r.", GUS, Warszawa (Area and Population in Territorial Section in 2008)

The area of land districts constitutes 98.52% of the area, and population 84.0% of number of inhabitants of PDK province. Five districts, i.e.. BIESZ, RZESZ, PRZ, SAN and LUB occupy 34.5% of the province area, but they are inhabited by only 23.2% of population (fig. 2). In respect of area the biggest one is LUB district to which belongs 7.4% of general area of the province, and the

most populated is RZESZ district in which lives almost every 10th person of PDK province.

Figure 2. Correlation chart of area and population number of land districts



Source: The authors' elaboration

Districts in fig. 2 create 5 concentrations given in table 2.

Table 2. Concentrations of land districts

Concen-trations	Num-ber	Districts	Size scale	
			surface	population
1	9	ŁAN, STRZ, TAR, BRZ, R-S, LEŻ, PRZW, KOL, NIŻ	low	medium
2	6	DĘB, STAL, JAS, MIEL, KROS, JAR	medium	high
3	1	RZESZ	high	high
4	3	SAN, PRZ, LUB	high	medium
5	2	LES, BIESZ	high	low

Source: The authors' elaboration

3. Spatial units and their division

It is assumed that for a given connected area a set of n spatial units $\mathbf{J} = \{J_1, J_2, ..., J_n\}$ meeting the conditions of disjointing and covering is considered. It constitutes a complete system of units. In the formal respect, from the point of view of the theory of sets, it means fulfilment of the following

conditions: covering $\bigcup_{i=1}^{n} J_i = U$ and disjointing $\bigcap_{i=1}^{n} \operatorname{int} J_i = \varnothing$, where int T means the interior of set T, and U is the divided big (basic) set.

It is assumed that for each unit there exists at least one adjacent unit, i.e. one of contiguity relations is determined:

a) with one unit

$$J_i \to \{J_{i_k}\}, \;\; i_k \in I - \{i\},$$

b) with m units

$$J_i \sim \{J_{i_1}, J_{i_2}, ..., J_{i_m}\}, \; i = 1,2,...,n,$$

where $i_1, i_2, ..., i_{m_i}, m_i \in \{1,2,...,n\}$, which means that $J_i \cap \bigcap_{p=1}^{m_i} fr\, J_{i_p} = \varnothing$, where $fr\, T$ expresses the boundary (edge) of set T.

For the pair of units $J_i, J_k \in \mathbf{J}$ we determine their common boundary of contiguity, which we denote by the symbol $K(J_i, J_k)$. It means that for the set $\mathbf{J}$ a non-empty set of boundaries (edges) $\mathbf{K}$ occurs**.** We express the measure of common boundary by $\# K(J_i, J_k)$ (e.g. length in km, the common boundary of forest complex in km). The fact of occurrence of common boundary means that $\# K(J_i, J_k) > 0$, otherwise such a boundary does not exist and then we assume $\# K(J_i, J_k) = 0$.

For the set of edges $\mathbf{K}$ of their given measures one may carry out the following operations:

a) determining the shortest edge $K_{\min}$ and the longest edge $K_{\max}$ and the total length of all edges,
b) arranging the edges according to their lengths into a non-decreasing sequence,
c) establishing the units of the lowest and the highest number of edges,
d) selecting for each unit its longest edges, and then arranging the units in respect of these edges.

The set of spatial units $\mathbf{J}$ is divided into two disjoint subsets of boundary units $\mathbf{J}_B$ and internal units $\mathbf{J}_W$ in order that $\mathbf{J} = \mathbf{J}_B \cup \mathbf{J}_W$ and $\mathbf{J}_B \cap \mathbf{J}_W = \varnothing$. Their sizes are expressed respectively by n_B, n_W at $n = n_B + n_W$. According to the given division, subsets of boundary edges $\mathbf{K}_B$ and internal edges $\mathbf{K}_W$ are singled out.

Attachment of units to the mentioned sets is determined by assuming that all units $J_{i_1}, J_{i_2}, ..., J_{i_{m_i}}$ are neighbours of unit J_i. Unit J_i is called boundary unit, when $J_i \cap \bigcap_{p=1}^{m_i} fr\, J_{i_p} \neq \varnothing$, and then $J_i \in J_B$. Thus, it is obvious that when $J_i \cap \bigcap_{p=1}^{m_i} fr\, J_{i_p} = \varnothing$, then J_i is the internal unit, i.e. $J_i \in J_W$.

Spatial units of the set $\mathbf{J}_B$ have the embedding (depth) of zero degree. Higher degrees of embedding for the units $\mathbf{J}_W$ are defined by minimum number if units separating them in respect of at least one unit of the set $\mathbf{J}_B$. For example, if the units from the set $\mathbf{J}_W$ have common boundary with at least one unit from the set $\mathbf{J}_B$, then they determine the embedding of first degree.

Lets introduce the measure of distance of the units $J_i \in \mathbf{J}_B$ and $J_k \in \mathbf{J}_W$ by $D(J_i, J_k)$, which is the lowest number of boundaries (edges) of contiguity between the units J_i, J_k, which is formally expressed by

$$D(J_i, J_k) = \underset{m=1,2,...,n}{\arg\min} (|\{J_1, J_2, ..., J_m : J_1 \sim J_i, J_m$$

$$= J_k, J_p \sim J_{p+1} \forall p = 1,2,...,m-1\}|) - 1,$$

where $|T|$ means the size of set T.

Then all units from the set $\mathbf{J}_W$ for which $D(J_i, J_k) = 1$ have the embedding of 1st degree, and when $D(J_i, J_k) = 2$, then they determine the embedding of 2nd degree, etc. We denote corresponding sets of spatial units, determined in this way, by $\mathbf{J}_0 = \mathbf{J}_B, \mathbf{J}_1, \mathbf{J}_2, ..., \mathbf{J}_m$, where m is the maximum degree of embedding. It is obvious that $\mathbf{J}_1, \mathbf{J}_2, ..., \mathbf{J}_m \in \mathbf{J}_W$. For the unit $J_k \in \mathbf{J}_W$ reaching the biggest embedding it is assumed that it constitutes the centre of the set of spatial units, but there can be more central units. In the contiguity of these units there frequently is the highest number of neighbours. Similar contiguities can be considered for any unit from the set $\mathbf{J}_W$.

To any two spatial units J_i, J_k, separated by at least one unit, one or more itineraries (routes) are assigned. These itineraries constitute a set of pairs of adjacent units starting from the pair $(J_i, *)$, and ending in the pair $(*, J_k)$, where * means the unit adjacent to the one given in the pair. We denote the set of these

pairs by $M(J_i, J_k)$, and its size by $\# M(J_i, J_k)$. In consequence for the earlier mentioned measure $D(J_i, J_k)$ we have

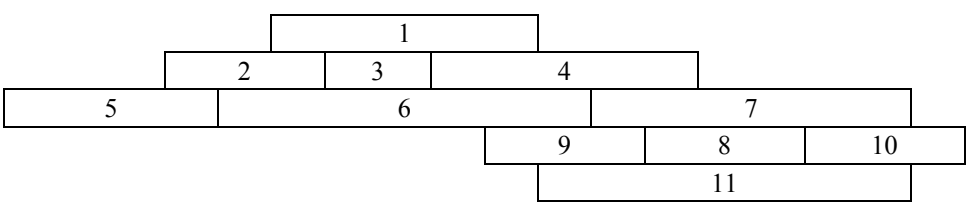
$$M(J_i, J_k) = \{(J_{i_p}, J_{p+1}) : p = 0,1,...,m-1, J_{i_0} = J_i, J_{i_m} = J_k\}.$$

Among pairs of units there are itineraries containing the lowest number of boundaries (edges) which should be overcome, coming through other units, and we will deal with such ones later on. The longest itineraries are always connected with the units of the set $\mathbf{J}_B$ and we denote them by $M^*(J_i, J_k)$. For $M(J_i, J_k)$ the inequality $M^*(J_i, J_k) \geq M(J_i, J_k)$ is fulfilled. The itineraries $M^*(J_i, J_k)$ constitute the diameter of the set of spatial units. There can be a few of such diameters. The units coming into the composition of a given itinerary without boundary units constitute transitive units. Itineraries containing at least 3 units, starting and ending in the same units, are defined as cycles, i.e. when $i = k$ and $|M(J_i, J_k)| \geq 3$.

A number of adjacent units belongs to each unit. It means the assignment to the unit J_i of n_i other units respectively, at $n_i \geq 1$. Such an assignment creates a list of contiguity. The contiguity of units is expressed with the use of a non-directed graph (e.g. Wilson 2004). It is determined by the pair $\mathbf{G} = (\mathbf{J}, \mathbf{K})$ of the set of units (knots) $\mathbf{J}$ and edges (boundaries) $\mathbf{K}$.

We will consider the presented notions on the set of 11 spatial units arranged in the area shown in fig. 2.

Figure 2. Connected area with singled out spatial units



Source: the authors' elaboration

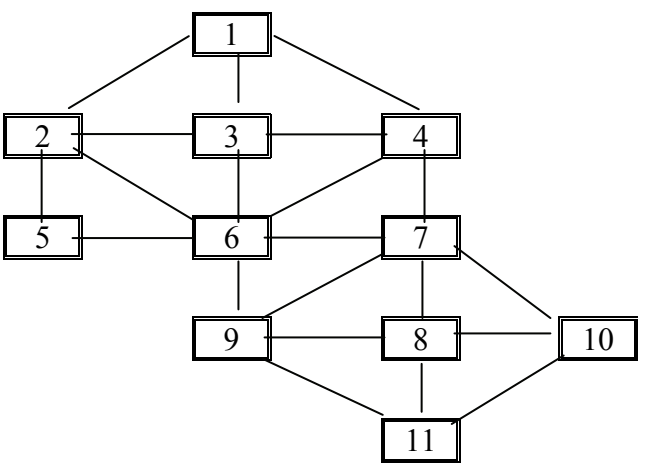
For the units presented in fig. 2, we have respectively:

- $n = 11$, $\mathbf{J} = \{J_1, J_2, ..., J_{11}\}$,
- $\mathbf{J}_B = \{J_1, J_2, J_4, J_5, J_6, J_7, J_9, J_{10}, J_{11}\}$, $q_B = 9$,
- $\mathbf{J}_W = \{J_3, J_8\}$, $q_W = 2$,
- units J_3, J_8 are the units of embedding of 1st degree,

- contiguities for units J_3 and J_8 constitute sets $\{J_1, J_4, J_6, J_2\}$ and $\{J_7, J_{10}, J_{11}, J_9\}$, whose elements are noted in clock-wise order,
- $M(J_1, J_6) = \{(J_1, J_2), (J_2, J_6)\}$, besides there exist other itineraries, then $\{(J_1, J_3), (J_3, J_6)\}$ or $\{(J_1, J_4), (J_4, J_6)\}$,
- $M(J_1, J_{11})$ maximum itinerary composed of two pairs of units, e.g. $\{(1,4),(4,6),(6,9),(9,11)\}$, $\{(1,4),(4,7),(7,9),(9,11)\}$, $\{(1,4),(4,7),(7,8),(8,11)\}$, etc.
- model non-directed graph of contiguity of units has been presented in fig. 3.

For the non-directed graph the pairs (*u, v*), (*v, u*) are equivalent, therefore at specification of the edge of set **K** only pairs (*i, k*) for $1 \le i < k \le n$ are given. For the mentioned graph the set of edges **K** = {(1,2), (1,3), (1,4), (2,3), (2,5), (2,6), (3,4), (3,6), (4,6), (4,7), (5,7), (6,7), (6,9), (7,8), (7,9), (7,10), (8,9), (8, 10), (8,11), (9,11), (10,11)} contains their 21 pairs.

Figure 3. Non-directed graph of contiguity of spatial units from fig. 2



Source: The authors' elaboration

3. Determining contiguity matrix and its properties

The matrix of contiguity **S** of the non-directed graph **G** is a binary square matrix. Its rows and columns correspond to the numbers of spatial units. 1-element occurs for the pair of units sharing common edge (boundary), otherwise 0-element appears. Such an element occurs also for the pair of identical units. Thus, the elements of matrix $\mathbf{S} = (s_{ij}), i, j = 1,2,...,n$ are determined by

$$s_{ij} = s_{ji} = \begin{cases} 1, & (i,j) \in \mathbf{K} \\ 0, & (i,j) \notin \mathbf{K} \end{cases}.$$

For the units in fig. 2 the contiguity matrix takes the form

$$\mathbf{S} = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}.$$

The presented matrix contains *42* elements of the value *1*, i.e. it is the doubled product of the number of edges (size of the set **K)**. For the matrix of contiguity **S** the following properties occur:

a) it is a symmetrical matrix,
b) diagonal elements equal zero,
c) the number of 1-elements equals $q = 2 \cdot \#\mathbf{K}$, where #**K** expresses the size **K,**
d) $\mathbf{S1} = \mathbf{m}$ – the product of matrix **S** and unit vector **1** gives vector **m** whose components express the number of neighbours of each spatial unit**,**
e) the equation $\mathbf{1'S1} = \mathbf{1'm} = q$ occurs,
f) determinant $\det(\mathbf{S})$ can be negative, zero or positive,
g) the value of the determinant of matrix **S** is invariant at change of numeration of spatial units,
h) **SS =T =** (t_{jk}), for *j, k = 1, 2, ..., q* – the product of contiguity matrix per se, being the matrix of elements t_{jk} corresponding to the number of neighbours of *j*-th spatial unit, when *j* = *k*, the number of common neighbours for two different spatial units being in direct contiguity or distant from each other by one unit (transitive by one), when $j \neq k$, but $t_{jk} = 0$ for pairs of units distant from each other by two or more number of units,
i) $diag(\mathbf{T}) = diag(\mathbf{m})$ – diagonal elements of matrix **T** correspond to the components of vector of sizes of neighbours **m**,
j) $\mathbf{T} = diag(\mathbf{m}) + \mathbf{B}$ – matrix **T** is expressed by the diagonal matrix of components of vector **m** and the matrix **B** of the number of common direct and transitive neighbours,

k) $tr(\mathbf{T}) = q$ – the trace of the matrix **T** equals the number of 1-elements of the matrix **S**,

l) the equation $\det(\mathbf{T}) = \{\det(\mathbf{S})\}^2$ occurs,

m) all characteristic values of matrix **S** are real,

n) for matrix **S** one can create its submatrix containing a subset of k spatial units $\{i_1, i_2, ..., i_k\} \in \{1, 2, ..., n\}$ meeting the assumptions:

(i) $i_1 < i_2 < ... < i_k$ and $i_j - i_{j-1} = 1$.

(ii) $\# K(J_{i_{j-1}}, J_{i_j}) > 0$,

the conditions (i) and (ii) guarantee that by removing numbers $i_1, i_2, ..., i_k$ of rows and columns from the initial matrix S, a separated from the initial domain, still compact subdomain of spatial units will arise,

o) $\mathbf{S} = \mathbf{S}_0 + \mathbf{S}_1 + ... + \mathbf{S}_k$ – it is possible to present contiguity matrix in the form of the matrix of the sum of k contiguity matrixes for the created k subsequences of spatial units, where all matrixes are of dimensions $n \times n$,

p) $\mathbf{S} = \mathbf{S}_0 + \mathbf{S}_1 + ... + \mathbf{S}_m$ – contiguity matrix is expressed by the sum of contiguity matrixes of units of embedding of zero, first, …, m-th degrees, but all matrixes on the right side are created for the full set of n spatial units,

q) principal minors $\mathbf{M}_1, \mathbf{M}_2, .\mathbf{M}_3, ..., \mathbf{M}_n$ of matrix **S** are contiguity matrixes of the set of successively first unit, two first units, three first units, etc. and of all units since in the last case $\mathbf{M}_n = \mathbf{S}$.

Upper-triangular matrix **B** for the considered example takes the form:

	1	2	3	4	5	6	7	8	9	10	11
1	0	1 (3)	2 (2, 4)	1 (2)	1 *(2)	3 *(2, 3, 4)	1 *(4)	0	0	0	0
2		0	2 (1, 6)	3 *(1, 3, 6)	1 (6)	2 (3, 5)	1 *(6)	0	1 *(6)	0	0
3			0	2 (1, 6)	2 *(2, 6)	2 (2, 4)	2 *(4, 6)	0	1 *(6)	0	0
4				0	1 *(6)	2 (3, 7)	1 (6)	1 *(7)	2 *(6, 7)	1 *(7)	0
5					0	1 (2)	1 *(6)	0	1 *(6)	0	0
6						0	2 (4, 9)	2 *(7, 9)	1 (7)	1 *(7)	1 *(9)
7							0	2 (9, 10)	2 (6, 8)	1 (8)	3 *(8, 9, 10)
8								0	2 (7, 11)	2 (7, 11)	2 (9, 10)
9									0	3 *(7, 8, 11)	1 (10)
10										0	1 (8)
11											0

where in the notation two variants for the pair of spatial units have been applied: number (numbers of adjacent units) and number *(numbers of transitive adjacent units):

To spatial units there is attributed the variant of density of contiguity $\mathbf{g} = \frac{1}{n-1}\mathbf{m}$ and their average density of contiguity of all units $\bar{g} = \frac{\mathbf{1'm}}{n(n-1)}$.

4. Contiguity matrix for land districts of Podkarpackie Province

Fig. 1 presents the map of $n = 21$ land districts of PDK province. Its simplified version is presented in fig. 4.

Figure 4. Diagram of localizations of land districts of PDK province

<table>
<tr><td rowspan="2">MIEL</td><td>TAR</td><td>STAL</td><td>NIŻ</td><td rowspan="3">PRZW</td><td>LUB</td></tr>
<tr><td>KOL</td><td rowspan="2">RZESZ</td><td>LEZ</td><td rowspan="2">JAR</td></tr>
<tr><td>DĘB</td><td>R-S</td><td>ŁAŃ</td></tr>
<tr><td rowspan="2">JAS</td><td>STRZ</td><td>BRZ</td><td colspan="3">PRZ</td></tr>
<tr><td>KRO</td><td>SAN</td><td>LES</td><td>BIESZ</td><td></td></tr>
</table>

Source: The authors' elaboration.

Table 3 gives the list of contiguity according to the arrangement of land districts in fig. 1. It also includes the vector of contiguity density (Density) and the degree of embedding (Embedding), but the districts have been listed in the order of degree of embedding and number of contiguity.

Table 3. List of contiguity of districts together with density and degree of embedding

Codes	Number	Density	Embedding	Districts of contiguity								
PRZW	6	0.30	0	LUB	JAR	PRZ	RZESZ	ŁAŃ	LEŻ			
PRZ	6	0.30	0	PRZW	JAR	BIESZ	SAN	BRZ	RZESZ			
SAN	5	0.25	0	BRZ	PRZ	BIESZ	LES	KROŚ				
DĘB	4	0.20	0	MIEL	R-S	STRZ	JAS					
KROŚ	4	0.20	0	STRZ	BRZ	SAN	JAS					
LEŻ	4	0.20	0	NIŻ	PRZW	ŁAN	RZESZ					
MIEL	4	0.20	0	TAR	KOL	R-S	DĘB					
NIŻ	4	0.20	0	STAL	LEŻ	RZESZ	KOL					
BIESZ	3	0.15	0	PRZ	LES	SAN						
JAR	3	0.15	0	LUB	PRZ	PRZW						
JAS	3	0.15	0	DĘB	STRZ	KROŚ						
STAL	3	0.15	0	NIŻ	KOL	TAR						
TAR	3	0.15	0	STAL	KOL	MIEL						
LES	2	0.10	0	BIESZ	SAN							
LUB	2	0.10	0	JAR	PRZW							
RZESZ	9	0.45	1	KOL	NIŻ	LEŻ	ŁAN	PRZW	PRZ	BRZ	STRZ	R-S
KOL	6	0.30	1	TAR	STAL	NIŻ	RZESZ	R-S	MIEL			
STRZ	6	0.30	1	R-S	RZESZ	BRZ	KROŚ	JAS	DĘB			
BRZ	5	0.25	1	RZESZ	PRZ	SAN	KROŚ	STRZ				
R-S	5	0.25	1	KOL	RZESZ	STRZ	DĘB	MIEL				
ŁAN	3	0.15	1	LEŻ	PRZW	RZESZ						

Source: The authors' elaboration.

The districts show various degree of embedding. Zero embedding is shown by boundary districts whose number is *15,* and the remaining *6* districts have embeddings of *1* degree, which was marked in table 2 with digital symbols *0* and *1*. The neighbours of each of the districts were given in clock-wise direction, always starting from the northern side.

The division of spatial units, presented in table 3, into boundary and internal ones, can be supplemented with various numerical characteristics, given as an example in table 4.

Table 4. Selected numerical characteristics for boundary and internal districts

Specification	Boundary districts		Internal districts		Total	Quotient
	Value	%	Value	%		
Number of districts	15	71.43	6	28.57	21	0.40
Number of neighbours	56	62.22	34	37.78	90	0.61
Average number of neighbours	3.73	–	5.70	–	4.29	–
Area [sq km]	13,582	77,26	3,998	22.74	17,580	0.29
Population [thousand]	1,261	71,56	501	28.44	1,761.8	0.40

Source: The authors' elaboration

The last column of table 4 shows the quotients of values of characteristics of internal and boundary units. For *15* boundary districts *56* boundaries of contiguity have been determined, which gives the average of *3.73* neighbours per district, and 6 internal districts have *34* boundaries of contiguity with the average of *5.7* neighbours per one district. Boundary and internal districts constitute respectively *71.43*% and *28.57*% of their general number, and the percentages of the number of neighbours for these groups are 62.22% and 37.78% respectively.

At a more general approach of the conducted comparative analysis of the given system of statistical characteristics for boundary and internal spatial units, one can apply statistical tests of significance (e.g. Student's t-test) for individual characteristics or Hotelling's test for the set of characteristics (e.g. Morrison 1990).

The list of contiguity given in table 3 allows creating the square contiguity matrix **S** of dimensions 21 × 21, or the triangular matrix (most often lower-triangular, without marking zero diagonal elements). Table 5 presents the contiguity matrix of districts of PDK province.

Table 5. Contiguity matrix S for districts of PDK province

No.	Dist.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	Sum
1	TAR	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
2	STAL	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
3	NIŻ	0	1	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	4
4	MIEL	1	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	4
5	KOL	1	1	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	6
6	LEŻ	0	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	4
7	LUB	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	2
8	DĘB	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	4
9	R-S	0	0	0	1	1	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	5
10	RZESZ	0	0	1	0	1	1	0	0	1	0	1	1	0	1	1	1	0	0	0	0	0	9
11	ŁAN	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	3
12	PRZW	0	0	0	0	0	1	1	0	0	1	1	0	1	0	0	1	0	0	0	0	0	6
13	JAR	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	3
14	STRZ	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	0	6
15	BRZ	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	1	1	0	0	5
16	PRZ	0	0	0	0	0	0	0	0	0	1	0	1	1	0	1	0	0	0	1	0	1	6
17	JAS	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	3
18	KROŚ	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	1	0	0	4
19	SAN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	1	1	5
20	LES	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	2
21	BIESZ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	3
Sum		3	3	4	4	6	4	2	4	5	9	3	6	3	6	5	6	3	4	5	2	3	90

Source: The authors' elaboration.

The given contiguity matrix is characterised by the following properties:

- there were 90 contiguity boundaries in total,
- numbers of neighbours of individual districts are contained in the last row and column in table 5,
- the highest number of neighbours has RZESZ (9 neighbours), and the lowest LUB (2),
- the distribution of number of neighbours for districts is given in the statement

Number of neighbours	2	3	4	5	6	9	Sum
Number of districts	2	6	5	3	4	1	21
Percentage part	9.52	28.57	23.81	14.29	19.05	4.76	100.00

- probability distribution of number of neighbours

Number of neighbours	2	3	4	5	6	9
Probabilities	0.0952	0.2857	0.2381	0.1429	0.1905	0.0476

and selected numerical characteristics of the distribution

Expected value	Standard deviation	Asymmetry
4.286	1.637	0.968

- the value of the determinant equals –168, i.e. matrix **S** is negative definite, and the sizes of successive determinants of principal minors are given in the statement:

2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
-1	0	1	-2	0	0	0	0	0	0	-3	16	-16	-20	21	80	-116	80	116	-168

The product of the contiguity matrix per se leads to the matrix **T** in which the order of districts is the same as in matrix **S** given in table 5:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	3	1	2	1	2	0	0	1	2	1	0	0	0	0	0	0	0	0	0	0	0
2	1	3	1	2	2	1	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0
3	2	1	4	1	2	1	0	0	2	2	2	2	0	1	1	1	0	0	0	0	0
4	1	2	1	4	2	0	0	1	2	2	0	0	0	2	0	0	1	0	0	0	0
5	2	2	2	2	6	2	0	2	2	2	1	1	0	2	1	1	0	0	0	0	0
6	0	1	1	0	2	4	1	0	1	3	2	2	1	1	1	2	0	0	0	0	0
7	0	0	0	0	0	1	2	0	0	1	1	1	1	0	0	2	0	0	0	0	0
8	1	0	0	1	2	0	0	4	2	2	0	0	0	2	1	0	1	2	0	0	0
9	2	1	2	2	2	1	0	2	5	2	1	1	0	2	2	1	2	1	0	0	0
10	1	2	2	2	2	3	1	2	2	9	2	3	2	2	2	2	1	2	2	0	1
11	0	0	2	0	1	2	1	0	1	2	3	2	1	1	1	2	0	0	0	0	0
12	0	0	2	0	1	2	1	0	1	3	2	6	2	1	2	2	0	0	1	0	1
13	0	0	0	0	0	1	1	0	0	2	1	2	3	0	1	1	0	0	1	0	1
14	0	0	1	2	2	1	0	2	2	2	1	1	0	6	2	2	2	2	2	0	0
15	0	0	1	0	1	1	0	1	2	2	1	2	1	2	5	2	2	2	2	1	2
16	0	0	1	0	1	2	2	0	1	2	2	2	1	2	2	6	0	2	2	2	1
17	0	0	0	1	0	0	0	1	2	1	0	0	0	2	2	0	3	1	1	0	0
18	0	0	0	0	0	0	0	2	1	2	0	0	0	2	2	2	1	4	1	1	1
19	0	0	0	0	0	0	0	0	0	2	0	1	1	2	2	2	1	1	5	1	2
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	1	1	2	1
21	0	0	0	0	0	0	0	0	0	1	0	1	1	0	2	1	0	1	2	1	3

On the main diagonal of matrix **T** the number of neighbours of individual districts can be read, and the elements in the intersection of a given row (*i*-th district) and column (*j*-th district) indicate their joint number of neighbours. In the case given they assume only the values 0, 1, 2 and 3. For example, for NIŻ district (3rd column) joint neighbours are presented in the statement:

J	District	Number	Joint neighbours
1	TAR	2	STAL, KOL
2	STAL	1	KOL
4	MIEL	1	KOL
5	KOL	2	STAL, RZESZ
6	LEŻ	1	RZESZ
9	R-S	2	KOL, RZESZ
10	RZESZ	2	KOL, LEZ
11	ŁAN	2	RZESZ, LEZ
12	PRZW	2	LEZ, RZESZ
14	STRZ	1	RZESZ
15	BRZ	1	RZESZ
16	PRZ	1	RZESZ

RZESZ district, which sometimes functions as a transitory (transitive) district, appeared most often in the role of a neighbour. The number of joint neighbours of individual districts is contained in the statement

Districts	TAR	STAL	NIŻ	MIEL	KOL	LEŻ	LUB	DĘB	R-S	RZESZ	ŁAN
Number	10	10	18	14	22	18	7	14	24	36	16
Districts	PRZW	JAR	STRZ	BRZ	PRZ	JAS	KROŚ	SAN	LES	BIESZ	
Number	21	11	24	25	25	11	15	15	6	10	

RZESZ districts appears in the role of joint neighbour as many as 36 times, and STRZ and BRZ districts by 25 and 24 times respectively. Hence, the conclusion is that the district with the highest number of neighbours (RZESZ – 9) appears most frequently as a joint neighbour of other districts (RZESZ – 36).

For the given order of spatial units (fig. 4) the elements in matrix **T** are grouped mainly along the diagonal.

5. Summary

Contiguity matrix appears in the theory of graphs (see: e.g. Kulikowski 1986, Wilson 2004, http://pl.wikipedia.org/wiki/Macierz sąsiedztwa 10.07.2007). Its function is, among other things, to describe graphs, and its elements express then a number of edges between vertices.

In the work the notion of contiguity matrix has been used to determine mutual positions of spatial units. Its structure is simple, and the only inconvenience is its considerable dimension when it refers to a numerous set of units. The considered matrix is characterised by many properties, of which particularly interesting is the product of this matrix per se.

Contiguity matrixes allow broadening the interpretation of researched social-economic phenomena. Its elements can be used to determine connected domains, constituting some subsets of spatial units, when their considered set is grouped into disjoint classes, which is met in taxonomical methods.

REFERENCES

KULIKOWSKI J., (1986): *Zarys teorii grafów* (An Outline of Theory of Graphs). PWN, Warszawa.

MORRISON D.F., (1990): *Wielowymiarowa analiza statystyczna. (Multidimensional Statistical Analysis)* PWN, Warszawa.

WILSON R., (2004): *Wprowadzenie do teorii grafów. (An Introduction into Theory of Graphs)* PWN, Warszawa.

ZELIAŚ A., (1991): *Ekonometria przestrzenna. (Spatial Econometrics)* PWE, Warszawa.

http://pl.wikipedia.org/wiki/Macierz sąsiedztwa.

STATISTICS IN TRANSITION-new series, July 2010
Vol. 11, No. 1, pp. 205—208

REPORT

2010 International Conference on Comparative EU Statistics on Income and Living Conditions Warsaw, Poland, 25–26 March 2010

The Conference on Comparative EU Statistics on Income and Living Conditions was held in Warsaw from 25th to 26th March 2010. This Conference, hosted by the Central Statistical Office of Poland, was organised by the Statistical Office of the European Communities (Eurostat) and the Network for the analysis of EU-SILC (Net-SILC). The papers presented at the Conference were the first results of the research carried out by Net-SILC partners.

Net-SILC is funded by Eurostat and consists of a group of researchers from 18 institutions using the comparative EU data source Community Statistics on Income and Living Conditions ("EU-SILC"). It brings together expertise from European Statistical System bodies and academics, and it is coordinated by the Luxembourg-based research institute CEPS/INSTEAD. The aims of Net-SILC are to develop methodologies for the analysis of EU-SILC and to carry out in-depth comparative research on incomes and living conditions.

Altogether there were 168 participants from 31 countries (30 from European countries). Concerning the papers, 21 contributed papers were presented in 5 sessions. The last part of conference (Session 6) was dedicated to panel discussion about future of EU-SILC in the wider context of EU statistics. This part was supported by background paper "*Beyond GDP, measuring well-being and EU-SILC*" prepared by Anthony B. Atkinson (Nuffield College, Oxford and London School of Economics, UK) and Eric Marlier (CEPS/INSTEAD).

The conference was opened by Józef Oleński, the President of Central Statistical Office of Poland. The opening session speakers were Carin Lindqvist-Virtanen (Chair of the EU Social Protection Committee's Indicators Sub-Group), Antonia Carparelli (Directorate-General "*Employment, Social Affairs and Equal opportunities*") European Commission, and Inna Steinbuka (Eurostat, European Commission).

The first two sessions were dedicated to ***Income, Poverty and Deprivation*** topics.

First speech was given by **Anthony B. Atkinson** (Nuffield College, Oxford and London School of Economics, UK). In his presentation, the author stressed the importance of EU-SILC to the analysis of the financial dimensions of poverty and inequality in Europe. Other papers presented in this part of the conference

focused on such issues as: *income poverty in regions of Europe; inequality, growth and mobility; income poverty and material deprivation in European countries; gross change in European's living conditions, income from own consumption and in-work poverty*. They were critically discussed by

Third session was devoted to ***Methodological Issues***.

At the next session **(Session 4)** presented papers were concentrated on **problems of health, social participation and households**. Fifth session was devoted to the ***labour market*** while in sixth session a wide spectrum of questions related to ***taxes, government spending and current policy issues*** were in the centre of interest. More information is available at http://www.stat.gov.pl/eusilc/.

Below is the detailed information on the sessions and topics discussed by paper givers and by invited discussants:

Session 1: Income, Poverty and Deprivation I

Chair: Pascal Wolff, Eurostat, European Commission

- Income distribution and financial poverty: EU-SILC in national and international context*, Anthony B. Atkinson (Nuffield College, Oxford and London School of Economics, UK), Eric Marlier and Anne Reinstadler (CEPS/INSTEAD Research Institute, Luxembourg)*;
- Macro determinants of individual income poverty in 93 regions of Europe, *Anne Reinstadler (CEPS/INSTEAD) and Jean-Claude Ray (University of Nancy, France)*;
- Inequality, growth and mobility: The inter-temporal distribution of income in European countries, *Philippe Van Kerm and Maria Noel Pi Alperin (both CEPS/INSTEAD)*;
- Income poverty and material deprivation in European countries, *Alessio Fusco (CEPS/INSTEAD), Anne-Catherine Guio (IWEPS, Belgium) and Eric Marlier (CEPS/INSTEAD)*.

Discussant: Brian Nolan, University College Dublin, Ireland

Session 2: Income, Poverty and Deprivation II

Chair: Eric Marlier, CEPS/INSTEAD, Luxembourg

- Towards an inclusion balance – Accounting for gross change in European's living conditions, *Matthias Till and Franz Eiffe (Statistics Austria)*;
- Income from own consumption, *Merle Paats and Ene-Margit Tiit (Statistics Estonia)*;
- In-work poverty in the EU, *Sophie Ponthieux (French Statistical Office (INSEE))*.

Discussant: Stephen Jenkins, ISER, University of Essex, UK

Session 3: Methodological Issues

Chair: Jean-Marc Museux, Eurostat, European Commission

- The EU-SILC rotational design and annual trends of cross-sectional indicators, *Martina Mysíková (Institute of Economic Studies, Charles University and Institute of Sociology of the Academy of Sciences, Czech Republic) and Martin Zelený (Czech Statistical Office)*;
- Sampling and non-sampling errors in EU-SILC, *Vijay Verma and Gianni Betti (University of Siena, Italy)*;
- The distributional impact of imputed rent in EU-SILC, *Veli-Matti Törmälehto, Anneli Juntto, Marie Reijo, Hannele Sauli (Statistics Finland)*;
- Robustness of some EU-SILC based indicators at regional level, *Vijay Verma and Gianni Betti (University of Siena, Italy)*.

Discussant: Olympia Bover, Bank of Spain, Spain

Session 4: Health, social participation and households

Chair: Rudi Van Dam (Federal Public Service Social Security, Belgium)

- Analysing the socio-economic determinants of health in Europe: new evidence from EU-SILC, *Cristina Hernández-Quevedo, Cristina Masseria, Elias Mossialos (London School of Economics, UK)*;
- Social participation in European countries, *Orsolya Lelkes (European Centre for Social Welfare Policy and Research, Austria)*;
- Household structure in the UE, *Maria Iacovou and Alexandra Skew (ISER, University of Essex, UK)*.

Discussant: Conchita D'Ambrosio, University of Milano-Bicocca, Italy

Session 5: Labour Market

Chair: Janusz Witkowski, Central Statistical Office of Poland

- Employment and social inclusion in the EU, *Marco Di Marco (Italian Statistical Office (ISTAT))*;
- Labour market returns to education in Europe – the potentials of analysis with EU-SILC, *Johannes Giesecke, Kathrin Leuze, Rita Nikolai (Social Science Research Centre Berlin (WZB), Germany)*;
- Distribution of earnings in the Euro-area and the EU, *Andrea Brandolini, Alfonso Rosolia and Roberto Torrini (Bank of Italy)*;
- Educational intensity of employment and polarization in Europe and the U.S., *Donald R. Williams (Kent State University, USA)*.

Discussant: John Micklewright, University of London, UK

Session 6: Taxes, Government Spending and Current Policy Issues
Chair: Anthony B. Atkinson, Nuffield College, Oxford and London School of Economics, UK

- The impact of public services on the distribution of income in European countries, *Rolf Aaberge, Audun Langørgen and Petter Lindgren (Statistics Norway)*;
- Distributional effects of direct taxes and social transfers (cash benefits), *Vaska Atta-Darkua and Andrew Barnard (Office for National Statistics (ONS), UK)*;
- Economic downturn and stress testing European welfare systems, *Francesco Figari, Andrea Salvatori and Holly Sutherland (ISER, University of Essex, UK)*.

Discussant: André Decoster, Catholic University of Leuven, Belgium

Final Panel: The future of EU-SILC in the wider context of EU statistics
Chair: Jean-Louis Mercy, Eurostat, European Commission

- *A.B. Atkinson, Nuffield College, Oxford and London School of Economics;*
- *Isabelle Engsted-Maquet, Directorate-General "Employment, Social affairs and Equal opportunities", European Commission and EU Social Protection Committee's Indicators Sub-Group;*
- *Michael Forster, Organisation for Economic Cooperation and Development (OECD);*
- *Stefan Lollivier, Statistics France (INSEE);*
- *Inna Steinbuka, Eurostat, European Commission.*

Conference Discussion and Conclusions

In discussion, apart from appreciation for the project in general – for its policy relevance and the scope in terms of geographical coverage and informational richness – some important methodological issues were raised too, along with suggestions on dealing with them. One concerned the need to constant improvements of applied research methodology, with suggestion to establish a special group of survey experts taking care of that (V. Verma). Another addressed the problem of an inconsistency between cross-sectional and panel elements of income measurement, with suggestion for better elaboration (incorporation) of the longitudinal module in the data collection and processing systems (S. Jenkins)

Prof. Olenski, in his concluding remarks on behalf of the host organization (GUS), expressed gratefulness to the Eurostat and Net-SILC authorities for choosing GUS as his co-organizer, and for collaboration in arranging the meeting in Warsaw.

Prepared by the *SiTns* Editorial Office (GUS)
and Agnieszka Zięba (SGH)

STATISTICS IN TRANSITION-new series, July 2010
Vol. 11, No. 1, pp. 209—210

REPORT

On the Berlin Meeting

On 29–31 March 2010, the already 15th annual conference of SCORUS (the Standing Committee on Regional and Urban Statistics; SCORUS is a part of IAOS, the International Association of Official Statisticians, which is a part o ISI, the International Statistical Institute) **was held in Berlin**. The Berlin meetings are titled ”Youth Assistance in Big Cities and Statistics”, and each year they are devoted to a specific subject matter related to the main topic. **The March meeting was devoted to “Relations between Generations and Challenges of an Ageing Society”.** The organiser and host of the meeting was, as each year, Dr Eckart Elsner, Professor. The conference is being continuously held at the Training Centre called “Clara Sahlberg” Bildungs-und Begegnungszentrum, Koblanckstrasse 14109 Berlin-Wannsee.

On behalf of the Central Board of the Polish Statistical Society, Dr Ewa Bulska, treasurer of the PSS CB, and Hanna Szenderowska-Czarnocka, M. A., member of the arbitration by fellow-workers, were participating in the meeting. Moreover, the Polish party was represented by researchers from the Institute for Market, Consumption and Business Cycles Research in Warsaw, employees of the Statistical Office in Krakow and the Economic University in Wrocław. The conference gathered, apart from representatives from Germany, also many guests from entire world, among others, from Kyoto, Japan; Hong Kong, China; Detroit, USA; Helsinki, Finland; Lagos, Nigeria; Ljubljana, Slovenia; Budapest, Hungary; Stockholm, Sweden; and Prague, the Czech Republic.

Prof. Eckart Elsner opened the conference. Each of the lecturers had 20 minutes for delivering their papers, and 10 minutes were devoted to questions and discussion. It is worth to emphasise professionalism, a high quality and an interesting form of the topic presented. After each paper, there was a heated discussion evidencing that the main topic of the conference had been selected properly as it proved to be very much up-to-date, and the issues addressed in the papers were the subject matter of interests and willingness to share experience among the representatives of the circles of statisticians and scientists. Many conference participants have been coming to Berlin for years at Prof. Dr Eckart Elsner’s invitation, and thus the atmosphere of those meetings is awfully nice, home and unique, what is also related to Professor’s personality.

The meeting organiser prepared, as each year, a very interesting social programme of the meeting, i.e. an excursion to Schoneberg City Hall at John F.

Kennedy Platz in Berlin, and visits to Berliner Fernsehturm and Internationales Congress Centrum (ICC) in Berlin. Owing to those visits, one might admire a magnificent panorama of the city and see the place which witnessed the visit and famous speech of the 36th US President, John F. Kennedy, in Berlin in June 1963.

The conference programme together with the list of titles of papers and names of their authors, and social attractions are presented below.

Polish Statistical Society