STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

Sampling methods and estimation

Research articles

Other articles:

Multivariate Statistical Analysis 2014, Łódź. Conference Papers

Other:

Volume 16, Number 2, Summer 2015

Editorial Office **ISSN 1234-7655**

Marek Cierpiał-Wolan, Ph.D.: Scientific Secretary
m.wolan@stat.gov.pl
Beata Witek: Secretary
b.witek@stat.gov.pl. Phone number 00 48 22 — 608 33 66
Rajmund Litkowiec: Technical Assistant

Address for correspondence

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

STATISTICS IN TRANSITION new series, Summer 2015
Vol. 16, No. 2, pp. 157−160

FROM THE EDITOR

This issue of *Statistics in Transition new series* is a mix of three types of papers. Two articles on sampling methods and estimation are followed by two research papers and five papers selected from an international conference. They are briefly characterized below.

The paper by **W. B. Molefe, D. K. Shangodoyin** and **R. G. Clark** presents ***An Approximation to the Optimal Subsample Allocation for Small Areas*** with focus on methods of allocation for stratified sample surveys which in a way incorporate small area estimation. Stratified sampling with small areas are assumed as the strata. Similar to Longford (2006), the authors seek efficient allocation that minimizes a linear combination of the mean squared errors of composite small area estimators and of an estimator of the overall mean. Unlike Longford, they define the mean squared error in a model-assisted framework, allowing a more natural interpretation of results using an intra-class correlation parameter. This allocation has an analytical form for a special case, and has the unappealing property that some strata may be allocated no sample. They derive a Taylor approximation to the stratum sample sizes for small area estimation using composite estimation giving priority to both small area and national estimation.

Zoramthanga Ralte's and **Gitasree Das's** article ***Ratio-to-Regression Estimator in Successive Sampling Using One Auxiliary Variable*** discusses the problem of estimation of a finite population mean on the current occasion based on the samples selected over two occasions. A chain ratio-to-regression estimator is employed to estimate the population mean on the current occasion in two-occasion successive (rotation) sampling using only the matched part and one auxiliary variable, which is available in both the occasions. The bias and mean square error of the proposed estimator are obtained. The authors propose another estimator, which is a linear combination of the means of the matched and unmatched portion of the sample on the second occasion. The bias and mean square error of this combined estimator are also obtained. The optimum mean square error of this combined estimator was compared with: (i) the optimum mean square error of the estimator proposed by Singh (2005), (ii) the mean per unit estimator, and (iii) the combined estimator suggested by Cochran (1977) when no auxiliary information is used on any occasion. Comparisons are made both analytically as well as empirically, using real life data. In conclusion, it was stressed that the proposed estimator is better than any of the other two estimators, therefore, the authors recommend it for further application in a similar context.

Research section starts with the paper on ***Multinomial Logistic Regression Approach for the Evaluation of Binary Diagnostic Test in Medical Research*** by **Alok Kumar Dwivedi, Indika Mallawaarachchi, Juan B. Figueroa-Casas, Angel M. Morales** and **Patrick Tarwater** Stressing the importance of evaluation of the effect of variables on diagnostic measures in clinical researchers, the authors propose to use logistic regression (LR) models to predict diagnostic measures of a screening test. A marginal model framework using generalized estimating equation (GEE) with logit/log link can be used to compare the diagnostic measures between two or more screening tests. These individual modelling approaches to each diagnostic measure ignore the dependency among these measures that might affect the association of covariates with each diagnostic measure. The diagnostic measures are computed using joint distribution of screening test result and reference test result which generates a multinomial response data. Multinomial logistic regression (MLR) has been shown to be a better approach to modelling these diagnostic measures. The authors compare the validity of LR and GEE approaches to MLR model for the case of modelling diagnostic measures. LR and GEE methods produced more biased estimates as compared to MLR approach, especially for small sample size studies. Since the proposed MLR model for diagnostic measures is simple and available as a part of common statistical software, the authors recommend that MLR method could be used as an alternative for modelling diagnostic measures.

Marek Obrębalski's and **Marek Walesiak's** paper on ***Functional Structure of Polish Regions in the Period 2004-2013 − Measurement via HHI Index, Florence's Coefficient of Localization and Cluster Analysis*** addresses the problems associated with measurement and identification issues which are discussed in reference to particular social and economic areas (referred to as functions) in the regions of the country, using the employment structure analysis and assessment by the sectors of the economy. The Herfindahl-Hirschman index was applied to measuring sectoral concentration and Florence's coefficient of localization to determine regional functional specialization. Finally, cluster analysis was conducted to produce the functional typology of regions. Summarizing their findings, the authors stress that several regions show a significant polyfunctionality, although each of them is characterized by a dominant function. The studied regions, however, show distinct functional specialization (in terms of field and level). However, each region has individual and diversified potential, regional identity and the level of economic competitiveness.

The rest of the issue is composed of papers based on presentations at the Multivariate Statistical Analysis conference held in Lodz (November 2014).

Justyna Wilk's paper on ***Using Symbolic Data in Gravity Model of Population Migration to Reduce Modifiable Areal Unit Problem (MAUP)*** addresses some challenges posed to spatial analyses by modifiable areal unit problem (MAUP). This occurs in operating on aggregated data determined for high-level territorial units (e.g. official statistics for countries) since generalization process deprives the data of variation and excluding territorial distribution of a phenomenon affects the results. The paper proposes to use symbolic data analysis (SDA) to reduce MAUP. SDA proposes an alternative form of individual data aggregation and deals with multivariate analysis of interval-valued, multi-valued and histogram data. Symbolic interval-valued data was used to determine the economic distance between regions which served as a separation function in the model. The proposed approach revealed that economic disparities in Poland are lower than official statistics show but they are still one of the most important factors of domestic migration flows.

In the paper ***Analysis of Convergence of European Regions with the Use of Composite Index*****, Joanna Górna** and **Karolina Górna** discuss the issue of convergence of the regions in the European Union while searching for the appropriate composite index which could capture the heterogeneity among the compared territorial units (such as voivodships in Poland). Several factors may be responsible for differentiation among the regions, such as: expenditure on R&D, HRST, quantity of patents, employment, participation of people in tertiary education among all employees. In empirical analysis some methods and models offered by the spatial statistics and econometrics were used, providing that geographical location has a great impact on the processes of economic growth. However, it has been shown that the spatial dependencies were not significant in each of the cases considered, but omitting them could result in spatial autocorrelation of residuals. In conclusion, the authors suggest further research agenda, pointing to Spatial Durbin Model as a solution for omitted variables.

The paper by **Magdalena Homa** and **Monika Mościbrodzka** on ***Application of Multifactorial Market-Timing Models to Assess Risk and Effectiveness of Equity-Linked Insurance Funds in Poland*** presents an application of traditional models developed by Treynor and Mazuy (T-M) and also by Henriksson-Merton (H-M) − which are called market-timing models − to assessing effectiveness of investment funds. In particular, the authors use some modifications of these models (T-M-FF and H-M-FF) with additional Fama-French factors to assess effectiveness and the risk of equity insurance connected with unit-linked insurance. Estimation and verification of the models for the subject group of equity funds were performed and the significance of the impact of particular factors on returns on reference portfolios was discussed.

Małgorzata Markowska's paper ***A Measure for Regional Resilience to Economic Crisis*** addresses the issue of measuring resilience to crisis, one that may be applied to regional data. In principle, such measure can take either positive or negative values − a positive value indicates resilience to crisis while a negative one the absence of resilience (vulnerability). The proposed measure uses growth rates referred to the previous year under the assumption that crisis results in a slowdown in growth, or even in a decline in values of important economic indicators. Growth rates are standardized by dividing the values of original change rates by medians specified based on spatio-temporal data modules. The measure of resilience to crisis is calculated as an arithmetic mean of the values of characteristics included into comparison. The results of application of the proposed measure to assessing the resilience to crisis during the period 2006-2011 are presented for regions of the European Union NUTS2 units, using six variables: changes in GDP, salaries, investments, household income, employment and unemployment.

The paper by **Germanas Budnikas**, ***Computerised Recommendations on E-Transaction Finalisation by Means of Machine Learning*** starts with an observation that a vast majority of business transactions is supported or executed online. This paper is devoted to the research on user online behaviour and making computerised advice. Several problems and their solutions are discussed: to know user behaviour online pattern with respect to business objectives and estimate a possible highest impact on user online activity. The approach suggested in the paper uses the following techniques: Business Process Modelling for formalisation of user online activity; Google Analytics tracking code function for gathering statistical data about user online activities; Naïve Bayes classifier and a feedforward neural network for a classification of online patterns of user behaviour as well as for an estimation of a website component that has the highest impact on the fulfilment of business objective by a user and which will be advised to be looked at. The technique is illustrated by an example.

This issue is concluded with **Risto Lehtonen's** and **Imbi Traat's** note in memoriam **Gunnar Kulldorff** (1927–2015), who passed away last June.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, Summer 2015
Vol. 16, No. 2, pp. 161

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://stat.gov.pl/en/sit-en/guidelines-for-authors/

STATISTICS IN TRANSITION new series, Summer 2015
Vol. 16, No. 2, pp. 163–182

AN APPROXIMATION TO THE OPTIMAL SUBSAMPLE ALLOCATION FOR SMALL AREAS

W. B. Molefe[1], D. K. Shangodoyin[2], R. G. Clark[3]

ABSTRACT

This paper develops allocation methods for stratified sample surveys in which small area estimation is a priority. We assume stratified sampling with small areas as the strata. Similar to Longford (2006), we seek efficient allocation that minimizes a linear combination of the mean squared errors of composite small area estimators and of an estimator of the overall mean. Unlike Longford, we define mean-squared error in a model-assisted framework, allowing a more natural interpretation of results using an intra-class correlation parameter. This allocation has an analytical form for a special case, and has the unappealing property that some strata may be allocated no sample. We derive a Taylor approximation to the stratum sample sizes for small area estimation using composite estimation giving priority to both small area and national estimation.

Key words: composite estimation, mean squared error, sample design, small area estimation, sample size allocation, Taylor approximation.

1. Introduction

Sampling designs, and sample sizes in particular, are chosen in practice so as to provide reliable estimates for large geographical regions or broad demographic groups. Budget and other constraints usually prevent the allocation of sufficiently large samples to each of the small areas. It is not possible to anticipate and plan for all possible areas (or domains) of applications as “the client will always require more than is specified at the design stage" (Fuller, 1999). The increased emphasis on small area estimation raises the question of how best to design surveys when the precision of small area estimates is a priority. If small area data needs are to be served using survey data then there is a need to develop an overall strategy that involves careful attention to satisfy these needs at the planning, sample design and estimation stages of the survey process (Singh et al., 1994). Singh et al. (1994) presented an illustration of compromise sample size allocation

[1] Department of Statistics, University of Botswana. E-mail: molefewb@mopipi.ub.bw.

[2] Department of Statistics, University of Botswana. E-mail: shangodoyink@mopipi.ib.bw.

[3] National Institute for Applied Statistics Research Australia, University of Wollongong. E-mail: rclark@uow.edu.au.

to satisfy reliability requirements at the provincial level as well as sub-provincial level in Canada.

Assume that small areas are identified in advance, and that stratified sampling is used with H strata defined by the small areas, indexed by $h \in U^1$. The population of units, indexed by j, denoted U is of size N. The population of N_h units in stratum h is U_h and the sample of n_h units selected by simple random sampling without replacement (SRSWOR) from stratum h is denoted by s_h. Let Y_j be the value of the characteristic of interest for the j^{th} unit in the population. The small area population mean is $\bar{Y}_h$ and the national mean is $\bar{Y}$. The corresponding sample estimators are $\bar{y}_h$ and $\bar{y}$, respectively; $\bar{y}_h = n_h^{-1} \sum_{j \in s_h} y_j$ and $\bar{y} = \sum_{h \in U^1} P_h \bar{y}_h$, where $P_h = N_h N^{-1}$. Let the sampling variances be $v_h = \text{var}_p(\bar{y}_h)$ and $v = \text{var}_p(\bar{y})$

Longford (2006) considers the problem of optimal sample sizes for small area estimation for this design. The approach is based on minimizing the weighted sum of the mean squared errors of the planned small area mean estimates and an overall estimate of the mean, with the weights specified to reflect the inferential priorities. An analytical solution exists when no weight is attached to estimating the overall mean but it has undesirable practical properties. For example, the optimal sample allocation is arrived at iteratively and some stratum sample sizes may be zero. When the overall mean is also important Longford does not find an exact or approximate analytical solution to the optimization problem. He suggests that the equation can be solved by numerical methods, such as the Newton method which interpolates between or extrapolates from a pair of provisional solutions, but that solving these equations iteratively may involve a considerable amount of computing when there are thousands of small areas.

The aim of this paper is to find the best allocation to strata for a linear combination of the mean squared errors of small area composite estimators and of an overall estimator of the mean, similar to Longford (2006). In section 2 we reformulate the objective in model-assisted terms, and derive the model-assisted composite estimator. Section 3 is devoted to optimizing the design. In subsection 3.1 we derive the optimal allocation for this objective when national estimation has no priority (G = 0) (similar in form to Longford but with different interpretation due to the explicit use of a model). Longford (2006) did not give an analytical solution when both national (overall) and small area estimates are a priority (G > 0). A numerical algorithm was given but may be computationally intensive, and its iterative nature makes it less transparent. In subsection 3.2 we derive two Taylor series approximations to the optimum. Unfortunately, the optimal allocations (both when G = 0 and when G > 0) have some undesirable properties.

2. Composite estimation

Royall (1973), in a discussion of papers by Gonzalez (1973) and Ericksen (1973), suggested that a choice between direct and synthetic approaches need not be made but that '... *a combination of the two is better than either taken alone*'. A natural way to balance the potential bias of a synthetic estimator $\overline{y}$ for $\bar{Y}_h$ against the instability of a direct estimator $\overline{y}_h$, is to use a composite estimator $\overline{y}_h^C$.

Composite estimators for small areas are defined as convex combinations of direct (unbiased) and synthetic (biased) estimators. A simple example is the composition $\overline{y}_h^C = (1-\phi_h)\overline{y}_h + \phi_h \overline{y}$ of the sample mean $\overline{y}_h$ for the target area h and the overall sample mean $\overline{y}$ of the target variable. The (area-specific) coefficients ϕ_h and $1-\phi_h$ in this composition are set with the intent to minimize its mean squared error (MSE), see for example, Schaible (1978); Brock et al. (1980) and Rao (2003). The coefficients for which minimum MSE would be attained depend on some unknown parameters which have to be estimated.

The design-based MSE of the composite estimator is given by:

$$MSE_p\left(\overline{y}_h^C;\overline{Y}_h\right) = (1-\phi_h)^2 v_h + \phi_h^2\left\{v + B_h^2\right\} + 2\phi_h\left(1-\phi_h\right)C_h \tag{1}$$

where C_h is the sampling covariance of $\overline{y}_h$ and $\overline{y}$, v_h is the sampling variance of the direct estimator $\overline{y}_h$, v is the sampling variance of the synthetic estimator $\overline{y}$ for $\overline{y}_h$ and $B_h = \overline{Y}_h - \overline{Y}$ is the bias of using $\overline{y}$ to estimate $\overline{y}_h$. Further,

$$MSE_p\left(\overline{y}_h^C;\overline{Y}_h\right) \approx (1-\phi_h)^2 v_h + \phi_h^2 B_h^2 \tag{2}$$

because $C_h \ll v_h$ and $v \ll v_h$ when the sample for area h is not a large part of the national sample. Auxiliary variables x_j are assumed to be available for the full population $j \in U^1$.

The following model ξ will be assumed:

$$\left.\begin{aligned} &E_\xi\left[Y_j\right] = \beta^T x_j \\ &\operatorname{var}_\xi\left[Y_j\right] = \sigma^2\left(j \in U_d\right) \\ &\operatorname{cov}_\xi\left[Y_i, Y_j\right] = \rho\sigma^2\left(i \neq j; i, j \in U_d\right) \\ &\operatorname{cov}_\xi\left[Y_i, Y_j\right] = 0\left(i \in U_d, j \in U_k, d \neq k\right) \end{aligned}\right\} \tag{3}$$

where i and j are units and h and g are small areas. Under the model (3),

$$E_\xi\left[v_h\right]=E_\xi\left[\operatorname{var}_p\left(\bar{y}_h\right)\right]=E_\xi\left[n_h^{-1}S_{hw}^2\right]=n_h^{-1}\sigma^2\left(1-\rho\right)$$

and

$$E_\xi\left[B_h^2\right]=E_\xi\left[\left(\bar{Y}_h-\bar{Y}\right)^2\right]\approx \operatorname{var}_\xi\left[\bar{Y}_h\right]=\operatorname{var}_\xi\left(N_h^{-1}\sum_{j\in U_h}Y_j\right)=\sigma^2N_h^{-1}\left[1+\left(N_h-1\right)\rho\right]$$

While the areas may have small sample sizes, their population sizes are substantial, so that $E_\xi\left[B_h^2\right]\approx\sigma^2\rho$. Also,

$$E_\xi\left[v\right]=E_\xi\left[\operatorname{var}_p\left(\bar{y}\right)\right]=E_\xi\left(\sum_{h\in U^1}P_h^2n_h^{-1}S_{hw}^2\right)=\sigma^2\left(1-\rho\right)\sum_{h\in U^1}P_h^2n_h^{-1} \qquad (4)$$

Following Molefe and Clark (2015), we assume a small-area composite estimator which is a weighted mean of an approximately design unbiased estimator

$$\bar{y}_{hr}=\bar{y}_h+\beta^T\left(\bar{X}_h-\bar{x}_h\right)$$

recommended by Hidiroglou and Patak (2004) for small domains, and a model-based synthetic estimator $\bar{Y}_{h(syn)}=\beta^T X_h$. The composite estimator which approximately minimizes the anticipated MSE is

$$\bar{y}_h^C=(1-\phi_{h(opt)})\bar{y}_{hr}+\phi_{d(opt)}\bar{Y}_{h(syn)}=\beta^T\bar{X}_h+(1-\phi_{h(opt)})\left(\bar{y}_h-\beta^T\bar{x}_h\right)$$

where $\phi_{h(opt)}=(1-\rho)\left[1+(n_h^*-1)\rho\right]^{-1}$, assuming that n , N_h and H are all large (Molefe and Clark, 2015). Under the same assumptions, the approximate anticipated MSE of the optimal composite estimator of $\bar{Y}_h$ conditional on n_h^* is

$$\begin{aligned}&E_\xi MSE_p\left(\bar{y}_h^C\left[\phi_{h(opt)}\right];\bar{Y}_h\mid n_h^*\right)\approx\\&\left(n_h^*\rho\left[1+(n_h^*-1)\rho\right]^{-1}\right)^2(n_h^*)^{-1}\sigma^2(1-\rho)+\left((1-\rho)\left[1+(n_h^*-1)\rho\right]^{-1}\right)^2\sigma^2\rho\\&=\sigma^2\rho(1-\rho)\left[1+(n_h^*-1)\rho\right]^{-1}\end{aligned}$$

(5)

3. Optimizing the design

3.1. Area-only optimal design

Provision of precise survey estimates for domains of interest requires that samples of adequate sizes be allocated to the domains. Conflicts arise when equal precision is desired for domains with widely varying population sizes. If estimates of means are desired at the same level of precision for all domains, then an equal allocation may be the most efficient strategy. However, such an allocation can cause a serious loss of efficiency for national estimates.

One way of measuring the performance of designs for small area estimation is with a linear combination of the anticipated MSE's of the small area mean and overall mean estimates. Following Longford (2006), the weights (called inferential priorities) N_h^q for $0 \leq q \leq 2$ are used. The approximate weighted total of the anticipated MSE's for the areas is given by

$$F = \sum\nolimits_{h \in U^1} N_h^q E_\xi MSE_\xi \left(\overline{y}_h^C \left[\phi_{h(opt)} \right]; \overline{Y}_h \mid n_h^* \right) + GN_+^{(q)} E_\xi v \qquad (6)$$

where $N_+^{(q)} = \sum\nolimits_{h \in U^1} N_h^q$

The quantity G is a relative priority coefficient. Ignoring the goal of national estimation corresponds to G = 0 and ignoring the goal of small area estimation corresponds to large values of G, in that case the second component in (6) is dominant. If G is positive, the priority coefficient has to be large because v would generally be much smaller than v_h, where v_h is the sampling variance of the direct estimator $\overline{y}_h$, v is the sampling variance of the synthetic estimator $\overline{y}$, so that G has to be large if the last term of (6) is to have any influence on the outcome. The factor $N_+^{(q)}$ is introduced to appropriately scale for the effect of the absolute sizes of N_h^q and the number of areas on the relative priority G. Criterion (6) is similar to the criterion in Longford (2006), however unlike this paper we adopt the model-assisted approach which treats the design-based inference as the real goal of survey sampling, but employs models to help choose between valid randomization-based alternatives (Särndal et al., 1992). The minimization is subject to a fixed sample size constraint. It would be straightforward to extend this to a fixed cost constraint with cost coefficients specific to the strata.

When national estimation has no priority (G = 0), the solution for the number of units to be sampled from each strata is found by optimizing (6) subject to a fixed total sampling cost function. The stationary point for this optimization is

$$n_{h,opt} = \frac{n\sqrt{N_h^q}}{\sum\nolimits_{h \in U^1} \sqrt{N_h^q}} + \frac{1-\rho}{\rho} \left(\frac{\sqrt{N_h^q}}{H^{-1} \sum\nolimits_{h \in U^1} \sqrt{N_h^q}} - 1 \right) \qquad (7)$$

Equation (7) is the optimal design if it gives a feasible solution ($0 \le n_{h,opt} \le N_h$ for all h); if not, the optimal design must be obtained numerically. An approximate solution can be found by setting the non-feasible solutions to $n_{h,opt} = 0$ when $n_{h,opt} < 0$ or $n_{h,opt} = N_h$ when $n_{h,opt} > N_h$ and then reallocating the remaining small areas (Longford, 2006).

In practice it will always be appropriate to set $0 \le q \le 2$, with q = 0 corresponding to all areas being equally important regardless of size, and q = 2 being the best choice for national estimation. In many cases q = 1 would be a sensible compromise.

The first term in (7) above is the optimal allocation for the direct estimator and corresponds to power allocation (Bankier, 1988). The second term will be positive for more populous areas (large N_h) and negative for less populous areas. Therefore, the allocation optimal for composite estimation has more dispersed subsample sizes $n_{h,opt}$ than the allocation that is optimal for direct estimators.

3.2. Compromise design

To incorporate priority for national estimation in optimizing design for small area estimation, we set the relative priority G to a positive value. Unfortunately, this optimization has no simple closed form solution (Molefe, 2012). The solution can be expressed as a quartic equation. Analytic solutions can be found to quartic equations but finding the solution would be convoluted and difficult to interpret. Also, there are up to 4 real-valued solutions. Another approach would be to find a Taylor series approximation based on ρ close to 0 and then minimize this with respect to n_h. The optimal n_h depends on ρ; one could consider n_h to be a function of this quantity and write $n_h = n_h(\rho)$.

The approximate weighted total of the anticipated MSE's for the areas is given by

$$F = \sum_{h \in U^1} N_h^q \sigma^2 \rho(1-\rho)\left[1+(n_h-1)\rho\right]^{-1} + GN_+^{(q)}\sigma^2\rho(1-\rho)\sum_{h \in U^1} P_h^2 n_h^{-1} \quad (8)$$

Replacing σ^2 by 1 as this value does not affect the optimal design, the minimum of (8) when G>0 satisfies the condition

$$N_h^q \rho^2 \left[1+(n_h-1)\rho\right]^{-2} + GN_+^{(q)} P_h^2 n_h^{-2} = \lambda \quad (9)$$

where λ is the Lagrange multiplier.

This needs to be solved with respect to n_h, but there is no simple closed form solution. One approach would be to find a Taylor series approximation based on

ρ close to 0 to the LHS of (9) and then minimize this with respect to n_h. The objective function is

$$F = \sum\nolimits_{h\in U^1} N_h^q \rho(1-\rho)\left[1+(n_h-1)\rho\right]^{-1} + GN_+^{(q)}(1-\rho)\sum\nolimits_{h\in U^1} P_h^2 n_h^{-1} \tag{10}$$

The first derivative with respect to ρ is:

$$F'(\rho) = \sum\nolimits_{h\in U^1} N_h^q \left\{(1-\rho)\left[1+(n_h-1)\rho\right]^{-1} - \rho\left[1+(n_h-1)\rho\right]^{-1} \right.$$
$$\left. -\rho(1-\rho)n_h\left[1+(n_h-1)\rho\right]^{-2}\right\} - GN_+^{(q)}\sum\nolimits_{h\in U^1} P_h^2 n_h^{-1}$$

Evaluated at $\rho = 0$, we get:

$$F(0) = GN_+^{(q)}\sum\nolimits_{h\in U^1} P_h^2 n_h^{-1}$$
$$F'(0) = \sum\nolimits_{h\in U^1} N_h^q - GN_+^{(q)}\sum\nolimits_{h\in U^1} P_h^2 n_h^{-1}$$

Hence, the first order Taylor series approximation around $\rho = 0$ is:

$$\begin{aligned} F(\rho) &\approx F(0) + F'(0)\rho \\ &= GN_+^{(q)}\sum\nolimits_{h\in U^1} P_h^2 n_h^{-1} + \rho\sum\nolimits_{h\in U^1} N_h^q - GN_+^{(q)}\rho\sum\nolimits_{h\in U^1} P_h^2 n_h^{-1} \\ &= \{(1-\rho)GN_+^{(q)} + \rho\}\sum_{h\in U^1} P_h^2 n_h^{-1} + \rho\sum_{h\in U^1} N_h^q \end{aligned}$$

But it is clear that minimizing this approximation of $F(\rho)$ is equivalent to minimizing $\sum\nolimits_{h\in U^1} P_h^2 n_h^{-1}$, which is equivalent to just ignoring the first term of (10). Therefore, this gives no priority to small area estimation. Hence, a first order Taylor series approximation of F with respect to ρ is not sufficient approximation for the purpose of designing sample for both small areas and national mean.

The second order Taylor series approximation is:

$$F''(\rho) = \sum\nolimits_{h\in U^1} N_h^q \left\{ -\frac{1}{\left[1+(n_h-1)\rho\right]} - \frac{(1-\rho)n_h}{\left[1+(n_h-1)\rho\right]^2} - \frac{1}{\left[1+(n_h-1)\rho\right]} + \right.$$
$$\left. \frac{\rho n_h}{\left[1+(n_h-1)\rho\right]^2} - \frac{(1-\rho)n_h}{\left[1+(n_h-1)\rho\right]^2} + \frac{\rho n_h}{\left[1+(n_h-1)\rho\right]^2} + \frac{2\rho(1-\rho)n_h^2}{\left[1+(n_h-1)\rho\right]^3} \right\}$$

Evaluated at $\rho = 0$ we get:

$$F''(0) = \sum\nolimits_{h \in U^1} N_h^q \left\{-1 - n_h - 1 - n_h\right\} = -2\sum\nolimits_{h \in U^1} N_h^q (n_h + 1)$$

The second order Taylor series approximation is then expressed as:

$$F(\rho) \approx F(0) + F'(0)\rho + \frac{1}{2}F''(0)\rho^2$$
$$= \{(1-\rho)GN_+^{(q)} + \rho\}\sum\nolimits_{h \in U^1} P_h^2 n_h^{-1} + \rho\sum\nolimits_{h \in U^1} N_h^q - \rho^2\sum\nolimits_{h \in U^1} N_h^q (n_h + 1)$$

We now consider minimizing the second order Taylor series approximation with respect to n_h subject to the cost constraint. The Lagrangian is:

$$L = \{(1-\rho)GN_+^{(q)} + \rho\}\sum\nolimits_{h \in U^1} P_h^2 n_h^{-1} + \rho\sum\nolimits_{h \in U^1} N_h^q - \rho^2\sum\nolimits_{h \in U^1} N_h^q (n_h + 1) + \lambda(\sum\nolimits_{h \in U^1} n_h - n)$$

To obtain the solution for the optimal within-strata sample size, we use partial derivatives with respect to n_h and λ, respectively. These are given by equations (A1) and (A2) in the Appendix. The solution for the optimum within-strata sample size n_h is given by

$$\begin{aligned} n_h &\approx n_h(0) + \rho n_h'(0) + \frac{1}{2}\rho^2 n_h''(0) \\ &= nP_h + \frac{1}{2}\rho^2 n^3 P_h (GN_+^{(q)})^{-1}\left\{N_h^q - N^{-1}\sum\nolimits_{h \in U^1} N_h^{q+1}\right\} \qquad (11) \\ &= nP_h\left(1 + \frac{1}{2}\rho^2 n^2 (GN_+^{(q)})^{-1}\left\{N_h^q - N^{-1}\sum\nolimits_{h \in U^1} N_h^{q+1}\right\}\right) \end{aligned}$$

The approximate solution is a function of G, ρ and q. When $G \uparrow \infty$ the approximate solution for n_h tends to $n_h \approx nP_h$, which is proportional allocation. When G is large, priority is given to estimation of the national mean, hence this is as would be expected, since proportional allocation will be optimal when the focus is on estimating accurately the overall mean. When G = 0 the approximate solution is not defined since division by zero is undefined. The approximate solution is therefore not suitable or appropriate when the only goal is small area estimation. When $\rho \downarrow 0$ the approximate solution is approximately equal to $n_h \approx nP_h$. When $\rho \approx 0$, units within a small area are less similar to each other for the variable of interest. When this happens it is natural for small areas to be represented in proportion to their population sizes.

When q = 1 or 2, it is not clear what the value of the approximate solution will be. The value of n_h depends on the magnitude and sign of $N_h^q - N^{-1}\sum_{h\in U^1} N_h^{q+1}$. We obtain large positive and negative values of n_h depending on the population size of the stratum. For relatively smaller strata, the result is large negative values which would in practice be truncated at zero and the opposite is true for relatively large strata. In practice, these would be truncated to either 0 or the population size.

The approximate analytical optimal design based on $\rho \approx 0$ gave counter-intuitive results, particularly when G is small or zero. Hence, we are now going to approximate n_h based on a different quantity based on both ρ and G rather than on ρ only, say, $n_h = n_h(\alpha)$ where $\alpha = f\left(\rho, G\right) = \rho(GN_+^{(q)})^{-1}N^q$. Our interest is the case where α is small. The problem is to minimize

$$F = \sum_{h\in U^1} N_h^q \rho\left[1+(n_h-1)\rho\right]^{-1} + GN_+^{(q)}\sum_{h\in U^1} P_h^2 n_h^{-1}$$

with respect to n_h subject to $\sum_{h\in U^1} n_h = n$. This is equivalent to minimizing

$$F = \alpha\sum_{h\in U^1} P_h^q\left[1+(n_h-1)\rho\right]^{-1} + \sum_{h\in U^1} P_h^2 n_h^{-1}$$

The corresponding Lagrangian function is

$$L = \sum_{h\in U^1} \alpha P_h^q\left[1+\left\{(n_h(\alpha)-1)\rho\right\}\right]^{-1} + \sum_{h\in U^1} P_h^2 n_h^{-1} + \lambda\left(\sum_{h\in U^1} n_h(\alpha) - n\right) \tag{12}$$

The partial derivatives of equation (12) with respect to n_h and λ are, respectively,

$$0 = L_1 = \frac{\partial L}{\partial n_h} = -\alpha P_h^q \rho\left[1+\left\{(n_h(\alpha)-1)\rho\right\}\right]^{-2} - P_h^2 n_h^{-2}(\alpha) + \lambda \tag{13}$$

$$0 = L_2 = \frac{\partial L}{\partial \lambda} = \sum_{h\in U^1} n_h(\alpha) - n \tag{14}$$

Equations (13) and (14) are easily solved when $\lambda = 0$, or in the limit as λ approaches 0. We will derive an approximation for the solution n_h when $\lambda \approx 0$, as this may often be the case in practice.

Let $n_h(\alpha)$ be the solution of (13) and (14) for any given value of α. We can then approximate n_h by $n_h \approx n_h(0) + n_h'(0)\alpha$.

We use equation (13) to obtain the value of $n_h(0)$ by substituting for $\alpha = 0$ to obtain

$$P_h^2 n_h^{-2}(0) = \lambda(0)$$

Solving for $n_h(0)$ we get

$$n_h(0) = P_h\left(\lambda(0)\right)^{-\frac{1}{2}} \tag{15}$$

We substitute for $n_h(0)$ into equation (14) and make it equal to zero to get

$$\left(\lambda(0)\right)^{-\frac{1}{2}} \sum\nolimits_{h\in U^1} P_h = n = \left(\lambda(0)\right)^{-\frac{1}{2}}$$

Substituting for $\left(\lambda(0)\right)^{-\frac{1}{2}}$ into (15) we obtain the value of $n_h(0)$ as

$$n_h(0) = nP_h \tag{16}$$

We take the derivative of (13) with respect to α :

$$0 = \frac{dL_1}{d\alpha} = \frac{\partial L_1}{\partial \alpha} + \frac{\partial L_1}{\partial n_h}\left(\frac{d}{d\rho} n_h(\alpha)\right) + \frac{\partial L_1}{\partial \lambda}\left(\frac{d}{d\rho}\lambda(\alpha)\right)$$

Therefore

$$\begin{aligned} 0 = \frac{dL_1}{d\alpha} &= -P_h^2 \rho\left[1+\left\{(n_h(\alpha)-1)\rho\right\}\right]^{-2} + \\ &\left\{2\alpha P_h^q \rho^2\left[1+\left\{(n_h(\alpha)-1)\rho\right\}\right]^{-3} + 2P_h^2 n_h^{-3}(\alpha)\right\} n_h'(\alpha) + \lambda'(\alpha) \end{aligned} \tag{17}$$

where $n_h'(\alpha) = \dfrac{d}{d\alpha} n_h(\alpha)$ and $\lambda'(\alpha) = \dfrac{d}{d\alpha}\lambda(\alpha)$. Evaluating (17) at $\alpha = 0$ gives

$$0 = -P_h^q \rho\left[1+\left\{(n_h(0)-1)\rho\right\}\right]^{-2} + 2P_h^2 n_h^{-3}(0) n_h'(0) + \lambda'(0)$$

Substituting for $n_h(0)$ given by (16) we get

$$0 = -P_h^q \rho\left[1+\left\{(nP_h-1)\rho\right\}\right]^{-2} + 2P_h^{-1} n_h^{-3} n_h'(0) + \lambda'(0)$$

and therefore

$$n_h'(0) = \frac{1}{2} P_h n^3 \left(P_h^q \rho \left[1 + (nP_h - 1)\rho\right]^{-2} - \lambda'(0) \right) \qquad (18)$$

Differentiating (14) with respect to α tells us

$$\frac{dL_2}{d\alpha}\Big|_{\alpha=0} = \sum\nolimits_{h \in U^!} n_h'(0) = 0$$

Combined with (18), this implies that

$$\sum\nolimits_{h \in U^1} n_h'(0) = \frac{1}{2} P_h n^3 \sum\nolimits_{h \in U^1} P_h \left(P_h^q \left[1 + (nP_h - 1)\rho\right]^{-2} - \lambda'(0) \right) = 0$$

And therefore $\lambda'(0) = \sum\nolimits_{h \in U^1} P_h^{q+1} \rho \left[1 + (nP_h - 1)\rho\right]^{-2}$

Substituting for $\lambda'(0)$ into (18) gives

$$n_h'(0) = \frac{1}{2} P_h n^3 \rho \left(P_h^2 \left[1 + (nP_h - 1)\rho\right]^{-2} - \sum\nolimits_{h \in U^1} P_h^{q+1} \rho \left[1 + (nP_h - 1)\rho\right]^{-2} \right)$$

Hence, the approximation to n_h is

$$n_h \approx n_h(0) + n_h'(0)\alpha$$

$$= nP_h + \frac{1}{2} \alpha P_h n^3 \rho \left(P_h^q \left[1 + (nP_h - 1)\rho\right]^{-2} - \sum\nolimits_{h \in U^1} P_h^{q+1} \rho \left[1 + (nP_h - 1)\rho\right]^{-2} \right)$$

Substituting for $\alpha = \rho (GN_+^{(q)})^{-1} N^q$ we obtain the general result:

$$n_h \approx nP_h + \frac{1}{2_h} \rho^2 (GN_+^{(q)})^{-1} N^q P_h n^3 \left(P_h^q \left[1 + (nP_h - 1)\rho\right]^{-2} - \sum\nolimits_{h \in U^1} P_h^{q+1} \left[1 + (nP_h - 1)\rho\right]^{-2} \right)$$

Rewritten, this becomes

$$n_h = nP_h \left(1 + \frac{1}{2} \rho^2 n^2 (GN_+^{(q)})^{-1} N^q \left\{ P_h^q \left[1 + (nP_h - 1)\rho\right]^{-2} - \sum\nolimits_{h \in U^1} P_h^{q+1} \rho \left[1 + (nP_h - 1)\rho\right]^{-2} \right\} \right) (19)$$

In the previous approximation based on ρ, we obtained large positive or negative values of n_h when n was large. Here, as $n \uparrow \infty$ the approximate sample size is equal to:

$$n_h \approx nP_h\left(1+\frac{1}{2}\rho^2 n^2 (GN_+^{(q)})^{-1} N^q \left\{P_h^q (nP_h\rho)^{-2} - \sum\nolimits_{h\in U^1} P_h^{q+1}(nP_h\rho)^{-2}\right\}\right)$$
$$= nP_h\left(1+\frac{1}{2}(GN_+^{(q)})^{-1} N^q \left\{P_h^{q-2} - \sum\nolimits_{h\in U^1} P_h^{q-1}\right\}\right)$$

which seems more reasonable.

When q = 0 and n is large, we get

$$n_h \approx nP_h\left(1+\frac{1}{2}(GH)^{-1}\left\{P_h^{-2} - \sum\nolimits_{h\in U^1} P_h^{-1}\right\}\right)$$

where $H = N_+^{(0)} = \sum\nolimits_{h\in U^1} N_h^0$

When q = 1 and n is large, we get

$$n_h \approx nP_h\left(1+\frac{1}{2G}\left\{P_h^{-1} - \sum\nolimits_{h\in U^1} P_h^{0}\right\}\right)$$

When q = 2 and n is large, we get

$$n_h \approx nP_h\left(1+\frac{1}{2}(GN_+^{(2)})^{-1} N^2 \left\{P_h^{0} - \sum\nolimits_{h\in U^1} P_h^{1}\right\}\right) = nP_h$$

A priority exponent of q = 2 implies proportional allocation, hence the result is as expected.

When $G \uparrow \infty$ the approximate sample size is equal to $n_h \approx nP_h$. This result is as expected since very large G implies more priority for national estimation. Proportional allocation will be optimal when the focus is on estimating accurately the overall mean. When $G \downarrow 0$, this corresponds to $\alpha \uparrow \infty$, and the approximate solution is undefined. This means that the alternative approximate analytical optimal design for n_h breaks down as $G \downarrow 0$. Perhaps this is not surprising, as our approximation is based on small α not large α. When $\rho \downarrow 0$ the alternative approximate analytical design is equal to $n_h \approx nP_h$. When $\rho \approx 0$, units within a small area are somewhat similar to each other though the degree of similarity is very low. Hence, it is appropriate for sample sizes within small areas to be in proportion to their population sizes.

4. Numerical example

We use data on the 26 cantons of Switzerland (Longford, 2006); their population sizes range from 15,000 (Appenzell-Innerrhoden) to 1.23 million (Zurich). The population of Switzerland is 7.26 million. We assume that n = 10,000, ρ = 0.025. We allocate a sample to the 26 cantons in Switzerland for q = 1 and a range of values of $G \in \{50, 100, 200, 500\}$ using the approximation in equation (11). The planned overall sample size is n = 10,000. The result of the percentiles of the sample sizes is shown in Table 4.1.

Table 4.1. Canton sample sizes by Taylor approximation when q = 1 and $\rho = 0.025$

	Percentiles of n_h				
Priority Coefficient	**Minimum**	**1st Quarter**	**Median**	**3rd Quarter**	**Maximum**
G = 50	-9322.0	-8620.0	-5648.0	-2226.0	97380.0
G = 100	-4470.0	-4097.0	-2634.0	-1088.0	49540.0
G = 200	-2050.0	-1878.0	-1126.0	-519.2	25620.0
G = 500	-617.0	-584.2	-324.5	-168.2	11260.0

When G > 0 and q = 1, the solution gives negative sample sizes for smaller cantons and very large positive sample size for the largest canton so that the negative sample sizes will be truncated at zero.

In summary, the approximate analytical optimal design based on $\rho \approx 0$ does not seem like a sensible approximation as evidenced by the allocation in Table 4.1.

We similarly allocate the sample sizes using the approximation in equation (19). The planned overall sample size is n = 10,000. The result of the percentiles of the sample sizes is shown in Table 4.2.

Table 4.2. Canton sample sizes by the alternative Taylor approximation when q = 1 and $\rho = 0.025$

	Percentiles of n_h				
Priority Coefficient	**Minimum**	**1st Quarter**	**Median**	**3rd Quarter**	**Maximum**
G = 50	237.0	275.2	296.0	383.8	1152.0
G = 100	129.0	181.5	290.5	426.8	1422.0
G = 200	75.0	139.0	288.0	448.2	1558.0
G = 500	42.0	113.2	286.5	461.2	1639.0

From Table 4.2, we see that when G = 50 the sample sizes of the least populous cantons are boosted in relation to proportional allocation at the expense of the most populous cantons. As G increases the sample size allocation approaches proportional allocation.

In summary, the alternative approximate analytical design seems to be useful especially when G > 0. The design seems sensible when there is priority for national estimation and is not applicable when the only priority is small area estimation.

5. Conclusions

The anticipated MSE is a sensible objective criterion for sample design because the particular sample which will be selected is not available in advance of the survey. Hence, a criterion which averages over all possible samples is appropriate. Sårndal et al. (1992, Chapter 14) base their optimal designs on the anticipated variance, which similarly averages over both model realizations and sample selection, although they consider only approximately design-unbiased estimators.

An analytical solution for the stationary point exists when the only priority is small area estimation. However, there are difficulties in applying it because when the strata have disparate population sizes, the stationary point gives negative sample sizes so that the optimum must be obtained numerically. The numerical optimum then has some strata with $n_h = 0$ which is also not desirable.

When priority is given to national estimation as well as to small area estimation so that G > 0, two approximate solutions were derived, based on $\rho \approx 0$ and $\alpha = f\left(\rho, G\right) = \rho(GN_+^q)^{-1}N^q \approx 0$. Both had undesirable properties, giving very large positive and negative sample sizes in some cases. This approximate solution gives counter-intuitive results, with large negative or positive values when there are unequal priorities for strata. Therefore, the Taylor approximation is not useful. An undesirable property of the second design is that it is not applicable when there is no priority for national estimation (G = 0).

REFERENCES

BANKIER, M. D., (1988). Power Allocations: Determining Sample Sizes for Sub-national Areas. The American Statistician, 42(3):174−177.

BINMORE, K. G., (1982). Mathematical Analysis: A straightforward approach. Cambridge University Press, 2nd edition.

BROCK, D. B., FRENCH, D. K., PEYTON, B. W., (1980). Small Area Estimation: Empirical Evaluation of Several Estimators for Primary Sampling Units. In Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 766−771.

DEMIDOVICH, B., editor, (1964). Problems in Mathematical Analysis. MIR Publishers.

ERICKSEN, E. P., (1973). Recent Developments in Estimation for Local Areas. In Proceedings of the Section on Social Statistics, American Statistical Association, pp. 37−41.

FULLER, W. A., (1999). Environmental Surveys Over Time. Journal of Agricultural, Biological and Environmental Statistics, 4: 331−345.

GONZALEZ, M. E., (1973). Use and Evaluation of Synthetic Estimates. In Proceedings of the Section on Social Statistics, American Statistical Association, pp. 33−36.

HIDIROGLOU, M. A., PATAK, Z., (2004). Domain estimation using linear regression. Survey Methodology, 30: 67−78.

LONGFORD, N. T. (2006). Sample Size Calculation for Small-Area Estimation. Survey Methodology, 32(1): 87−96.

MOLEFE, W. B., (2012). Sample Design for Small Area Estimation. PhD thesis, University of Wollongong, http://ro.uow.edu.au/theses/3495.

MOLEFE, W. B., CLARK, R. G., (2015). Model-Assisted Optimal Allocation For Planned Domains Using Composite Estimation, Survey Methodology (forthcoming).

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

ROYALL, R. M., (1973). Discussion of two Papers on Recent Developments in Estimation of Local Areas. In Proceedings of the Section on Survey Research Methods, American, Statistical Association, pp. 43−44.

SÅRNDAL, C., SWENSSON, B., WRETMAN, J., (1992). Model Assisted Survey Sampling. Springer-Verlag.

SCHAIBLE, W. L., (1978). Choosing Weight for Composite Estimators for Small Area Statistics. In Proceedings of the Section on Survey Research Methods, American Statistical 3 Association, pp. 741−746.

SINGH, M. P., GAMBINO, J., MANTEL, H. J., (1994). Issues and Strategies for Small Area Data. Survey Methodology, 20(1): 3−22.

APPENDIX

$$0 = L_1 = \frac{\partial L}{\partial n_h} = -\{(1-\rho)+\rho\}GN_{+}^{(q)}P_h^2 n_h^{-2} - \rho^2 N_h^q + \lambda \tag{A1}$$

$$0 = L_1 = \frac{\partial L}{\partial \lambda} = \sum\nolimits_{h\in U^1} n_h - n \tag{A2}$$

Equations (A1) and (A2) are easily solved when $\rho = 0$, or in the limit as ρ approaches 0. We will derive an approximation for the solution n_h when $\rho \approx 0$, as this may often be the case in practice.

Let $n_h(\rho)$ be the solution of (A1) and (A2) for any given value of ρ. We can then approximate n_h by $n_h \approx n_h(0) + n_h'(0)\rho + \frac{1}{2}n_h''(0)\rho^2$

It is easily shown that $n_h(0) = nP_h$. To derive $n_h(0)$ we use (A1) to obtain the value of $n_h(0)$ by substituting for $\rho = 0$ to obtain $GN_{+}^{(q)}P_h^2 n_h^{-2}(0) = \lambda(0)$

Solving for $n_h(0)$ we get

$$n_h(0) = P_h\left(\frac{GN_{+}^{(q)}}{\lambda(0)}\right)^{\frac{1}{2}} \tag{A3}$$

Substituting for n_h into (A2) gives $\left(\frac{GN_{+}^{(q)}}{\lambda(0)}\right)^{\frac{1}{2}} \sum\nolimits_{h\in U^1} P_h = n$. Summing through and re-arranging terms gives $\left(\lambda(0)\right)^{-\frac{1}{2}} = n(GN_{+}^{(q)})^{-\frac{1}{2}}$. Substituting for $\left(\lambda(0)\right)^{-\frac{1}{2}}$ into (A3), we obtain the value of $n_h(0)$ as

$$n_h(0) = nP_h \tag{A4}$$

We take the first derivative of (A1) with respect to ρ:

$$0 = \frac{dL_1}{d\rho} = \frac{\partial L_1}{\partial \rho} + \frac{\partial L_1}{\partial n_h}\left(\frac{d}{d\rho}n_h(\rho)\right) + \frac{\partial L_1}{\partial \lambda}\left(\frac{d}{d\rho}\lambda(\rho)\right) \tag{A5}$$

using the result on differentiation of composite functions by, for example, Demidovich (1964) and Binmore (1982). The partial derivative of (A1) with respect to ρ is:

$$\frac{\partial L_1}{\partial \rho} = GN_+^{(q)} P_h^2 n_h^{-2} - 2\rho N_h^q$$

and the partial derivative with respect to n_h gives

$$\frac{\partial L_1}{\partial n_h} = 2(1-\rho) GN_+^{(q)} P_h^2 n_h^{-3}$$

Substituting the partial derivatives in (A5) gives

$$\begin{aligned} 0 &= \frac{dL_1}{d\rho} = \frac{\partial L_1}{\partial \rho} + \frac{\partial L_1}{\partial n_h}\left(\frac{d}{d\rho} n_h\right) + \frac{\partial L_1}{\partial \lambda}\left(\frac{d}{d\rho}\lambda\right) \\ &= GN_+^{(q)} P_h^2 n_h^{-2} - 2\rho N_h^q + 2(1-\rho) GN_+^{(q)} P_h^2 n_h^{-3} n_h'(\rho) + \lambda'(\rho) \end{aligned} \tag{A6}$$

where $n_h'(\rho) = \frac{d}{d\rho} n_h$ and $\lambda'(\rho) = \frac{d}{d\rho}\lambda$.

Evaluating (A6) at $\rho = 0$ gives

$$0 = GN_+^{(q)} P_h^2 n_h^{-2}(0) + 2GN_+^{(q)} P_h^2 n_h^{-3}(0) n_h'(0) + \lambda'(0) \tag{A7}$$

Solving for $n_h'(0)$ gives:

$$\begin{aligned} n_h'(0) &= -2\left\{\lambda'(0) + GN_+^{(q)} P_h^2 n_h^{-2}(0)\right\} (GN_+^{(q)} P_h^2)^{-1} n_h^{-3}(0) \\ &= -2\lambda'(0)(GN_+^{(q)} P_h^2)^{-1} n_h^{-3}(0) - 2GN_+^{(q)} P_h^2 n_h^{-2}(0)(GN_+^{(q)} P_h^2)^{-1} n_h^{-3}(0) \\ &= -\lambda'(0)(GN_+^{(q)})^{-1} n^3 P_h - nP_h \end{aligned} \tag{A8}$$

Differentiating (A2) with respect to ρ tells us

$$\frac{dL_2}{d\rho}\Big|_{\rho=0} = \sum\nolimits_{h\in U^!} n_h'(0) = 0 \tag{A9}$$

Combined with (A8), this implies that

$$\sum_{h\in U^1}\left\{-\lambda'(0)(GN_+^{(q)})^{-1}n^3P_h - nP_h\right\} = 0$$

Consequently, $\lambda'(0) = -GN_+^{(q)}n^{-2}$. Substituting for $\lambda'(0)$ into (A8) gives the result that

$$n_h(0) = 0$$

We now take the second derivative of (A5). Let

$$L_3 = \frac{dL_1}{d\rho} = \frac{\partial L_1}{\partial\rho} + \frac{\partial L_1}{\partial n_h}\left(\frac{d}{d\rho}n_h(\rho)\right) + \frac{\partial L_1}{\partial\lambda}\left(\frac{d}{d\rho}\lambda(\rho)\right)$$

and therefore

$$L_3 = GN_+^{(q)}P_h^2 n_h^{-2} - 2\rho N_h^q + 2(1-\rho)GN_+^{(q)}P_h^2 n_h^{-3} n_h'(\rho) + \lambda'(\rho) \qquad \text{(A10)}$$

We take the derivative of (A10) with respect to ρ:

$$0 = \frac{dL_3}{d\rho} = \frac{\partial L_3}{\partial\rho} + \frac{\partial L_3}{\partial n_h}\left(\frac{d}{d\rho}n_h'(\rho)\right) + \frac{\partial L_3}{\partial\lambda}\left(\frac{d}{d\rho}\lambda'(\rho)\right)$$

The partial derivative of (A10) with respect to ρ is given by

$$0 = \frac{dL_3}{d\rho} = -2N_h^q - 2GN_+^{(q)}P_h^2 n_h^{-3} n_h'(\rho) + 2(1-\rho)GN_+^{(q)}P_h^2 n_h^{-3}\frac{\partial}{\partial\rho}n_h'(\rho) + \frac{\partial}{\partial\rho}\lambda'(\rho)$$

The partial derivative evaluated at $\rho = 0$ is then

$$\frac{\partial L_3}{\partial\rho}\Big|_{\rho=0} = -2N_h^q - 2GN_+^{(q)}P_h^2 n_h^{-3}(0)n_h'(0) + 2GN_+^{(q)}P_h^2 n_h^{-3}(0)n_h''(0) + \lambda''(0)$$
$$= -2N_h^q + 2GN_+^{(q)}P_h^2(nP_h)^{-3}n_h''(0) + \lambda''(0)$$

since $n_h'(0) = 0$.

The partial derivative of (A10) with respect to n_h is given by

$$\frac{\partial L_3}{\partial n_h} = -2GN_+^{(q)}P_h^2 n_h^{-3} - 6(1-\rho)GN_+^{(q)}P_h^2 n_h^{-4} n_h'(\rho) + 2(1-\rho)GN_+^{(q)}P_h^2 n_h^{-3} n_h''(\rho)$$

The partial derivative evaluated at $\rho = 0$ is:

$$\begin{aligned} 0 &= \frac{\partial L_3}{\partial n_h} \\ &= -2GN_+^{(q)}P_h^2 n_h^{-3}(0) + 2GN_+^{(q)}P_h^2 n_h^{-3}(0) n_h''(0) \\ &= -GN_+^{(q)}P_h^2 (nP_h)^{-3} + GN_+^{(q)}P_h^2 (nP_h)^{-3} n_h''(0) \end{aligned}$$

since $n_h'(0) = 0$.

We put the results together to obtain

$$\begin{aligned} 0 &= \frac{\partial L_3}{\partial \rho}\,|_{\rho=0} \\ &= \left\{ \frac{\partial L_3}{\partial \rho} + \frac{\partial L_3}{\partial n_h}\left(\frac{d}{d\rho} n_h(\rho) \right) + \frac{\partial L_3}{\partial \lambda}\left(\frac{d}{d\rho} \lambda(\rho) \right) \right\} \\ &= -2N_h^q + 2GN_+^{(q)}P_h^2 n_h^{-3} P_h^{-1} n_h''(0) + \lambda''(0) \end{aligned}$$

since $n_h'(0) = \frac{d}{d\rho} n_h(\rho)\,|_{\rho=0} = 0$ and $\lambda'(0) = \frac{d}{d\rho}\lambda(\rho)\,|_{\rho=0} = 0$.

Solving for $n_h''(0)$ we find

$$n_h''(0) = (2N_h^q - \lambda''(0))\frac{1}{2}(GN_+^{(q)})^{-1} n^3 P_h \qquad \text{(A11)}$$

Differentiating (A9) with respect to ρ gives:

$$0 = \frac{d}{d\rho}\left(\frac{dL_2}{d\rho} \right)|_{\rho=0} = \frac{d}{d\rho} \sum\nolimits_{h \in U^1} n_h'(\rho)\,|_{\rho=0} = \sum\nolimits_{h \in U^1} n_h''(0)$$

Combined with (A11), this implies that

$$\sum\nolimits_{h\in U^1} n_h''(0) = \sum\nolimits_{h\in U^1} (2N_h^q - \lambda''(0))\frac{1}{2}(GN_+^{(q)})^{-1} n^3 P_h = 0$$

Therefore

$$(GN_+^{(q)})^{-1} n^3 N^{-1} \sum\nolimits_{h\in U^1} N_h^{q+1} - \frac{1}{2}\lambda''(0)(GN_+^{(q)})^{-1} n^3 \sum\nolimits_{h\in U^1} P_h = 0$$

Solving for $\lambda''(0)$ we get $\lambda''(0) = 2N^{-1}\sum\nolimits_{h\in U^1} N_h^{q+1}$. Substituting into (A11) gives

$$n_h''(0) = (N_h^q - N^{-1}\sum\nolimits_{h\in U^1} N_h^{q+1})(GN_+^{(q)})^{-1} n^3 P_h$$

Hence, our approximation to n_h is:

$$\begin{aligned} n_h &\approx n_h(0) + \rho n_h'(0) + \frac{1}{2}\rho^2 n_h''(0) \\ &= nP_h + \frac{1}{2}\rho^2 n^3 P_h (GN_+^{(q)})^{-1}\left\{ N_h^q - N^{-1}\sum\nolimits_{h\in U^1} N_h^{q+1} \right\} \\ &= nP_h \left(1 + \frac{1}{2}\rho^2 n^2 (GN_+^{(q)})^{-1}\left\{ N_h^q - N^{-1}\sum\nolimits_{h\in U^1} N_h^{q+1} \right\} \right) \end{aligned}$$

STATISTICS IN TRANSITION new series, Summer 2015
Vol. 16, No. 2, pp. 183–202

RATIO-TO-REGRESSION ESTIMATOR IN SUCCESSIVE SAMPLING USING ONE AUXILIARY VARIABLE

Zoramthanga Ralte[1], **Gitasree Das**[2]

ABSTRACT

The problem of estimation of finite population mean on the current occasion based on the samples selected over two occasions has been considered. In this paper, first a chain ratio-to-regression estimator was proposed to estimate the population mean on the current occasion in two-occasion successive (rotation) sampling using only the matched part and one auxiliary variable, which is available in both the occasions. The bias and mean square error of the proposed estimator is obtained. We proposed another estimator, which is a linear combination of the means of the matched and unmatched portion of the sample on the second occasion. The bias and mean square error of this combined estimator is also obtained. The optimum mean square error of this combined estimator was compared with (i) the optimum mean square error of the estimator proposed by Singh (2005) (ii) mean per unit estimator and (iii) combined estimator suggested by Cochran (1977) when no auxiliary information is used on any occasion. Comparisons are made both analytically as well as empirically by using real life data.

Key words: ratio-to-regression estimator, auxiliary variable, successive sampling, bias, mean square error, optimum replacement policy.

1. Introduction

The successive method of sampling consists in selecting samples of the same size on different occasions such that some units are common to samples selected on previous occasions. In successive sampling, the ratio estimator is among the most commonly adopted estimators of the population mean or total of some variables of interest of a finite population with the help of an auxiliary variable when the correlation coefficient between the two variables is positive. Patterson (1950) and Cochran (1977) suggested a number of estimation procedures on

[1] Department of Statistics, North-Eastern Hill University, Shillong. India.
E-mail: aramaralte7@gmail.com.

[2] Department of Statistics, North-Eastern Hill University, Shillong. India.
E-mail: gitasree22@gmail.com.

sampling over two occasions. Rao and Graham (1964), Gupta (1979), Das (1982), Sen (1971) developed estimators for the population mean on the current occasion using information on two auxiliary variables available on the previous occasion. Sen (1972, 1973) extended his work for several auxiliary variates. Singh *et al.* (1991) and Singh and Singh (2001) used the auxiliary information on current occasion for estimating the current population mean in two occasion successive sampling. Singh (2003) extended their work for h-occasions successive sampling.

Utilizing the auxiliary information on both the occasions, Singh (2005), Singh and Priyanka (2006, 2007a, 2008) proposed varieties of chain-type ratio, difference and regression estimators for estimating the population mean on the current (second) occasion in two occasion successive sampling. Singh (2005) suggested two estimators for population mean using the information on an auxiliary variable in successive sampling over two occasions. Consider a character under study on the first (second) occasion is denoted by x(y) respectively and the auxiliary variable z is available on both the occasions. A simple random sample (without replacement) of n units is taken on the first occasion. The suggested chain-type ratio estimator based on the sample of size $m\left(=n\lambda\right)$ common to both the occasions is given by

$$T_2 = \frac{\bar{y}_m}{\bar{x}_m}\frac{\bar{x}_n}{\bar{z}_n}\bar{Z} \tag{1}$$

The bias B(.) and mean square error M(.) of T_2 up to the first order of approximation and for large population of size N of equation (1) are derived as

$$B\left(T_2\right) = \bar{Y}\left[\left(\frac{1}{m}-\frac{1}{n}\right)\left(C_x^2 - 2\rho_{yx}C_yC_x\right) + \frac{1}{n}\left(C_z^2 - 2\rho_{yz}C_yC_z\right)\right]$$

and

$$M\left(T_2\right) = \bar{Y}^2\left[\frac{C_y^2}{m} + \left(\frac{1}{m}-\frac{1}{n}\right)\left(C_x^2 - 2\rho_{yx}C_yC_x\right) + \frac{1}{n}\left(C_z^2 - 2\rho_{yz}C_yC_z\right)\right]$$

A classical ratio estimator based on a sample of size $u = n - m = n\mu$ (fraction of a sample) drawn afresh on the second occasion is given by

$$T_1 = \frac{\bar{y}_u}{\bar{z}_u}\bar{Z} \tag{2}$$

Combining T_1 and T_2, the resulting estimator of $\bar{Y}$ is

$$T = \phi T_1 + \left(1-\phi\right)T_2$$

where ϕ is an unknown constant to be determined under certain optimum criterion.

After optimizing ϕ and μ (sampling fraction), the optimum mean square error of the estimator T is given by

$$M(T)_{opt} = \frac{(A+C)^2 + (A+C)(B-C)\mu_0}{n\left[(A+C)+(B-C)\mu_0^2\right]}$$

where $A = \overline{Y}^2 C_y^2$, $B = \overline{Y}^2\left(C_x^2 - 2\rho_{yx}C_xC_y\right)$, $C = \overline{Y}^2\left(C_z^2 - 2\rho_{yz}C_yC_z\right)$ and

$$\mu_0 = \frac{-\left(1-\rho_{yz}\right) \pm \sqrt{\left(1-\rho_{yz}\right)\left(1-\rho_{yx}\right)}}{\left(\rho_{yz} - \rho_{yx}\right)}$$

Following the chain-type ratio estimator proposed by Singh (2005), in the present work we propose a chain ratio-to-regression estimator for estimating population mean on the current occasion using auxiliary information that is available on both the occasions. The behaviour of the proposed estimator have been examined analytically and also through empirical means of comparison.

2. Proposed estimator

Consider a population consisting of N units. Let a character under study on the first (second) occasion be denoted by x(y), respectively. It is assumed that the information on an auxiliary variable z is available on the first as well as on the second occasion. We consider the population to be large enough, and the sample size is constant on each occasion. Using simple random sampling without replacement (SRSWOR) we select a sample of size n on the first occasion. Of these n units, a sub-sample of size $m = n\lambda$ is retained on the second occasion. This sub-sample is supplemented by selecting SRSWOR of $u = (n-m) = n\mu$ units afresh from the units that were not selected on the first occasion.

By modifying the estimator in (1), we propose a chain ratio-to-regression estimator for $\overline{Y}$ on the second occasion which is based on a sample of size m common to both the occasions and is given by

$$T_p = \frac{\overline{y}_m}{\overline{x}_m}\left[\overline{x}_m + b_{xz}\left(\overline{z}_n - \overline{z}_m\right)\right]\frac{\overline{Z}}{\overline{z}_n} \qquad (3)$$

where

$\overline{Z}$: Population mean of z.

$\overline{x}_m, \overline{y}_m, \overline{z}_m, \overline{z}_n$: Sample means of the respective variates with sample sizes as shown in the subscript.

b_{xz} : Regression coefficient of x on z.

The bias of the proposed estimator T_p up to the second order of approximation as obtained in Appendix A is:

$$\text{Bias}(T_p) = \frac{1}{\bar{X}}\left\{\begin{array}{l} \beta_{xz}\bar{Y}\left(-\frac{\text{Cov}\left(\bar{z}_n, s_z^2\right)}{S_z^2} + \frac{\text{Cov}\left(\bar{z}_n, s_{xz}\right)}{S_{xz}} + \frac{\text{Cov}\left(\bar{z}_m, s_z^2\right)}{S_z^2} + \frac{\text{Cov}\left(\bar{z}_m, s_{xz}\right)}{S_{xz}} \right) \\ +\beta_{xz}\left[\text{Cov}\left(\bar{y}_m, \bar{z}_n\right) - \text{Cov}\left(\bar{y}_m, \bar{z}_m\right)\right] - \frac{\bar{Y}}{\bar{X}}\text{Var}\left(\bar{x}_m\right) \\ -\beta_{xz}\frac{\bar{Y}}{\bar{X}}\left[\text{Cov}\left(\bar{x}_m, \bar{z}_n\right) - \text{Cov}\left(\bar{x}_m, \bar{z}_m\right)\right] - \bar{X}\text{Cov}\left(\bar{y}_m, \bar{z}_n\right) + \frac{\bar{X}\bar{Y}}{\bar{Z}^2}\text{Var}\left(\bar{z}_n\right) \end{array}\right\} \tag{4}$$

And the mean square error of T_p considering $E(T_p)$ up to the first order of approximation obtained in Appendix B is:

$$M\left(T_p\right) = \bar{Y}^2\left[\frac{C_y^2}{m} + \left(\frac{1}{m} - \frac{1}{n}\right)\left(\rho_{xz}^2 C_x^2 - 2\rho_{xz}\rho_{yz}C_x C_y\right) + \frac{1}{n}\left(C_z^2 - 2\rho_{yz}C_y C_z\right)\right] \tag{5}$$

Following Singh (2005), we combine the estimators T_1 and T_p and the final estimator of $\bar{Y}$ on the second occasion is given as

$$T_{p_1} = \psi T_1 + (1-\psi)T_p \tag{6}$$

where ψ is an unknown constant to be determined under certain criterion.

3. Bias and mean square error of a combined estimator T_{p_1}

Since both T_1 and T_p are biased estimators of $\bar{Y}$, therefore, the resulting estimator T_{p_1} is also a biased estimator of $\bar{Y}$. Using the notations in Section 2, the bias B (T_{p_1}) and mean square error M (T_{p_1}) up to the first order of approximation are given below:

$$B\left(T_{p_1}\right) = \psi B(T_1) + (1-\psi)B\left(T_p\right) \tag{7}$$

and $$M\left(T_{p_1}\right) = \psi^2 M(T_1) + (1-\psi)^2 M\left(T_p\right) \tag{8}$$

where, from Cochran (1977)

$$B\left(T_1\right)=\frac{\overline{Y}}{u}\left(C_z^2-\rho_{yz}C_yC_z\right) \tag{9}$$

and

$$M\left(T_1\right)=\frac{\overline{Y}^2}{u}\left(C_y^2+C_z^2-2\rho_{yz}C_yC_z\right) \tag{10}$$

4. Minimum mean square error of T_{p_1}

Since M (T_{p_1}) in (8) is a function of unknown constant ψ, therefore, it is minimized with respect to ψ and subsequently the optimum value of ψ obtained in Appendix C is

$$\psi_{opt}=\frac{M\left(T_p\right)}{M\left(T_1\right)+M\left(T_p\right)} \tag{11}$$

Now, substituting the value of ψ_{opt} in equation (8) we obtain the optimum mean square error of T_{p_1} **(**given in Appendix C) as

$$M\left(T_{p_1}\right)_{opt}=\frac{M\left(T_1\right)M\left(T_p\right)}{M\left(T_1\right)+M\left(T_p\right)} \tag{12}$$

Using equation (5) and (10), let

$$A'=\overline{Y}^2C_y^2, \qquad B'=\overline{Y}^2\left(\rho_{xz}^2C_x^2-2\rho_{xz}\rho_{yz}C_xC_y\right), \quad C'=\overline{Y}^2\left(C_z^2-2\rho_{yz}C_yC_z\right)$$

Then, from Appendix C

$$M\left(T_{p_1}\right)_{opt}=\frac{\left(A^{'}+C^{'}\right)^2+\left(A^{'}+C^{'}\right)\left(B^{'}-C^{'}\right)\mu}{n\left[\left(A^{'}+C^{'}\right)+\left(B^{'}-C^{'}\right)\mu^2\right]}$$

or

$$M\left(T_{p_1}\right)_{opt}=\frac{\left(\alpha_1^{'}\right)^2+\alpha_1^{'}\alpha_2^{'}\mu}{n\left[\alpha_1^{'}+\alpha_2^{'}\mu^2\right]} \tag{13}$$

where $\alpha_1^{'}=A^{'}+C^{'}$, $\qquad \alpha_2^{'}=B^{'}-C^{'}$

5. Optimum replacement policy

To determine the optimum value of μ (fraction of a sample to be taken afresh at the second occasion) so that population mean $\overline{Y}$ may be estimated with maximum precision, we minimize mean square error of T_{p_1} with respect to μ which results in quadratic equation in μ and solution of μ (from Appendix D) is given below:

$$\begin{aligned}\mu_0^{'} &= \frac{-\alpha_1^{'} \pm \sqrt{\left(\alpha_1^{'}\right)^2 + \left(\alpha_1^{'}\alpha_2^{'}\right)}}{\alpha_2^{'}} \\ &= \frac{-2\left(1-\rho_{yz}\right) \pm \sqrt{2\left(1-\rho_{yz}\right)\left\{2\left(1-\rho_{yz}\right) + \left(1-\rho_{xz}\right)\left(2\rho_{yz}-\rho_{xz}-1\right)\right\}}}{\left(1-\rho_{xz}\right)\left(2\rho_{yz}-\rho_{xz}-1\right)}\end{aligned} \tag{14}$$

From equation (14), it is obvious that the real value of $\mu_0^{'}$ exists if the quantities under square root are greater than or equal to zero. To choose the admissible value of $\mu_0^{'}$, it should be remembered that $0 \le \mu_0^{'} \le 1$. All other values of $\mu_0^{'}$ are inadmissible. Substituting the value of $\mu_0^{'}$ from (14) in (13), we have

$$M\left(T_{p_1}\right)_{opt} = \frac{\left(\alpha_1^{'}\right)^2 + \alpha_1^{'}\alpha_2^{'}\mu_0^{'}}{n\left[\alpha_1^{'} + \alpha_2^{'}\left(\mu_0^{'}\right)^2\right]} \tag{15}$$

6. Comparison of mean square error and efficiency

Now we compare the optimum MSE of the proposed estimator T_{p_1} with (i) the optimum MSE of the estimator proposed by Singh (2005), (ii) $\overline{y}_n$, i.e. mean per unit estimator, and (iii) $\overline{y}_2^{'}$ (Cochran, 1977) when no auxiliary information is used at any occasion. In each case, we obtain conditions under which the estimator T_{p_1} is better than the three estimators mentioned above.

(i) Comparison with Singh’s (2005) estimator.

First, we consider the estimator $T = (\phi)T_1 + (1-\phi)T_2$ which is due to Singh (2005). The variance of this estimator, to the first order of approximation and for large population, is given as

$$M(T) = \frac{1}{n\mu(1-\mu)}\left[(\phi)^2\alpha_1(1-\mu) + (1-\phi)^2\left(\alpha_1\mu + \alpha_2\mu^2\right)\right]$$

where

$$\alpha_1 = A + C,\ \alpha_2 = B - C,\ A = \overline{Y}^2 C_y^2,\ \ B = \overline{Y}^2\left(C_x^2 - 2\rho_{yx}C_yC_x\right),\ C = \overline{Y}^2\left(C_z^2 - 2\rho_{yz}C_yC_z\right)$$

The variance of T is minimized for

$$\phi_{opt} = \frac{\mu\left(\alpha_1 + \alpha_2\mu\right)}{\left(\alpha_1 + \alpha_2\mu^2\right)}$$

Thus, the resulting minimum variance of T is

$$M\left(T\right)_{opt} = \frac{\alpha_1^2 + \alpha_1\alpha_2\mu}{n\left[\alpha_1 + \alpha_2\mu^2\right]}$$

Under the assumption of $C_x = C_y = C_z = C$, minimizing $M\left(T\right)_{opt}$ with respect to μ, the optimum value of μ (fraction of a sample to be taken afresh at the second occasion) is given by

$$\left[\frac{-\alpha_1 \pm \sqrt{\alpha_1^2 + \alpha_1\alpha_2}}{\alpha_2}\right] = \frac{-\left(1-\rho_{yz}\right) \pm \sqrt{\left(1-\rho_{yz}\right)\left(1-\rho_{yx}\right)}}{\left(\rho_{yz} - \rho_{yx}\right)} = \mu_0 \text{(say)}$$

Therefore,

$$M\left(T\right)_{opt} = \frac{\alpha_1^2 + \alpha_1\alpha_2\mu_0}{n\left[\alpha_1 + \alpha_2\mu_0^2\right]}$$

The proposed combined estimator T_{p_1} is better than the combined estimator proposed by Singh (2005) if

$$M\left(T\right)_{opt} - M\left(T_{p_1}\right)_{opt} > 0$$

$$\Rightarrow \frac{\alpha_1^2 + \alpha_1\alpha_2\mu_0}{n\left[\alpha_1 + \alpha_2\mu_0^2\right]} - \frac{\left(\alpha_1^{'}\right)^2 + \alpha_1^{'}\alpha_2^{'}\mu_0^{'}}{n\left[\alpha_1^{'} + \alpha_2^{'}\left(\mu_0^{'}\right)^2\right]} > 0$$

$$\Rightarrow \left(1-\rho\right)\left(1-\mu_0^{'}\right) > 0$$

(See Appendix E(i) for derivation of the above result)

Since $0 \le \mu_0^{'} \le 1$, $\left(1-\mu_0^{'}\right)$ is always positive and $\left(1-\rho\right)$ is always positive, which indicates that the above condition is always valid. That is, T_{p_1} is a better estimator than the estimator T proposed by Singh (2005).

(ii) Comparison with mean per unit estimator.

Next, we consider the mean per unit estimator $\bar{y}_n$ as follows:

$\bar{y}_n = \frac{\sum_{i=1}^{n} y_i}{n}$, under the assumption of $C_y = C$, the variance of mean per unit estimator is given by

$$V\left(\bar{y}_n\right) = \frac{\bar{Y}^2 C_y^2}{n} = \frac{\bar{Y}^2 C^2}{n}$$

The proposed combined estimator T_{p_1} is better than this estimator if

$$V\left(\bar{y}_n\right) - M\left(T_{p_1}\right)_{opt} > 0$$

$$\Rightarrow \left[2-(1-\rho)\left(\mu_0^{'}\right)^2\right] > 2(1-\rho)\left[2-(1-\rho)\mu_0^{'}\right] \tag{16}$$

(See Appendix E(ii) for derivation of the above result)

Hence, T_{p_1} is a better estimator than mean per unit estimator whenever equation (16) is satisfied. And it can be easily verified that the above condition is true for $\rho \geq 0.5$.

(iii) Comparison with Cochran's (1977) estimator.

Now we consider the Cochran's (1977) estimator $\bar{y}_2^{'}$ as follows:

$\bar{y}_2^{'} = \phi_2 \bar{y}_{2u}^{'} + \left(1-\phi_2\right)\bar{y}_{2m}^{'}$, i.e. combined estimator suggested by Cochran (1977) when no auxiliary information is used at any occasion.

Here, $\bar{y}_{2u}^{'}$, $\bar{y}_{2m}^{'}$ are the unmatched and matched portions of the sample at the second occasion respectively, and

$$\phi_2 = \frac{W_{2u}}{W_{2u} + W_{2m}}$$

where W_{2u} and W_{2m} are the inverse variances, i.e.

$$\frac{1}{W_{2u}} = V\left(\bar{y}_{2u}^{'}\right) = \frac{C_y^2 \bar{Y}^2}{u}, \quad \frac{1}{W_{2m}} = V\left(\bar{y}_{2m}^{'}\right) = \frac{C_y^2 \bar{Y}^2 \left(1-\rho_{yx}^2\right)}{m} + \rho_{yx}^2 \frac{C_y^2 \bar{Y}^2}{n}$$

By least square theory, the variance of $\bar{y}_2'$ is

$$V\left(\bar{y}_2'\right)=\frac{1}{W_{2u}+W_{2m}}$$

$$=\frac{C_y^2\bar{Y}^2\left(n-u\rho_{yx}^2\right)}{n^2-u^2\rho_{yx}^2}$$

Minimizing $V\left(\bar{y}_2'\right)$ with respect to u, the optimum value of u is given by

$$u_{opt}=\frac{n}{1+\sqrt{1-\rho_{yx}^2}} \text{ and } m_{opt}=\frac{n\left(\sqrt{1-\rho_{yx}^2}\right)}{1+\sqrt{1-\rho_{yx}^2}}$$

Under the assumption of $C_y = C$, the optimum variance of the above estimator is given as

$$V\left(\bar{y}_2'\right)_{opt}=\left[1+\sqrt{1-\rho_{yx}^2}\right]\frac{\bar{Y}^2C^2}{2n}$$

Then, the proposed combined estimator T_{p_1} is better than this estimator if

$$V\left(\bar{y}_2'\right)_{opt}-M\left(T_{p_1}\right)_{opt}>0$$

$$\Rightarrow\left(1+\sqrt{1-\rho^2}\right)\left[2-(1-\rho)\left(\mu_0'\right)^2\right]>4(1-\rho)\left[2-(1-\rho)\mu_0'\right] \qquad (17)$$

(See Appendix E(iii) for derivation of the above result)

Hence, T_{p_1} is a better estimator than Cochran's (1977) estimator whenever equation (17) is satisfied. And it can be easily verified that the above condition is true for $\rho \geq 0.5$.

Further, conditions (16) and (17) are true for all values of $\rho \geq 0.5$. This is justified since ρ_{yx} expresses the relationship between current occasion and previous occasion study variable, ρ_{xz} for previous occasion study variable and auxiliary variable and ρ_{yz} for current occasion study variable and auxiliary variable.

As an example, the percent relative efficiency of the proposed estimator T_{p_1} with respect to (i) T, the estimator proposed by Singh (2005), (ii) $\bar{y}_n$, i.e. mean per unit estimator, and (iii) $\bar{y}_2'$ (Cochran, 1977) when no auxiliary information is used on any occasion, have been computed for some assumed values of ρ_{yx}, ρ_{yz} and ρ_{zx}. Since $\bar{y}_n$ and $\bar{y}_2'$ are unbiased estimators of $\bar{Y}$, their variances for large N are respectively given by

$$V\left(\bar{y}_n\right) = \frac{\bar{Y}^2 C_y^2}{n}$$

and
$$V\left(\bar{y}_2'\right)_{opt} = \left[1 + \sqrt{1-\rho_{xy}^2}\right] \frac{\bar{Y}^2 C_y^2}{2n}$$

Here, E_1, E_2, E3 are designated as the percent relative efficiencies of the proposed estimator T_{p_1} with respect to Singh's (2005) estimator T , $\bar{y}_n$ and $\bar{y}_2'$ respectively.

Further, the expression of the optimum μ, i.e. μ_0' and the percent relative efficiencies E_1, E_2 and E_3 are in terms of population correlation coefficients. We assumed that all the correlations involved in the expressions of μ, E_1, E_2 and E_3 are equal, i.e. $\rho_{xy} = \rho_{xz} = \rho_{yz} = \rho$. Accordingly, computed values of μ_0', E_1, E_2 and E_3 for different choices of high positive correlations are shown in Table 1.

Table 1. Relative Efficiency (%) of T_{p_1} with .respect to estimators T, $\bar{y}_n$ and $\bar{y}_2'$.

ρ	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
μ_0'	0.5359	0.52786	0.5203	0.513167	0.506411
E_1	107.18	112.194	117.3	122.5148	127.8537
E_2	107.18	131.966	173.435	256.5835	506.4113
E_3	100	118.769	148.646	205.2668	363.5754

From Table 1, it is clear that the proposed estimator is more efficient than T, $\bar{y}_n$ and $\bar{y}_2'$. Also, when the correlation coefficient $\rho \geq 0.5$ is increasing, the gain in precision of the proposed estimator T_{p_1} over T, $\bar{y}_n$ and $\bar{y}_2'$ is also increasing. Further, it may be noticed that the maximum gain in efficiency occurs while comparing it with mean per unit estimator, which is very obvious.

7. Illustration using real life data

In order to illustrate the comparison of relative efficiency of the proposed estimator with respect to (i) T, the estimator proposed by Singh (2005), (ii) $\bar{y}_n$, i.e. mean per unit estimator, and (iii) $\bar{y}_2^{'}$ (Cochran,1977) when no auxiliary information is used on any occasion, using real life approach, the data from the Census of India (2001) and (2011) was considered. We define the variables x(y) as the total number of workers in villages in the state of Mizoram, India in 2001 (2011) and z is defined as an auxiliary variable which is the total number of literate people in villages in the state of Mizoram, India.

Using successive sampling as defined in section 2 for the above data set we take n = 70, m = 35 and u = 35. The following table shows the values of the different estimators as computed from the sample along with their corresponding mean square errors and efficiency of the proposed estimator T_{p_1} with respect to T, $\bar{y}_n$ and $\bar{y}_2^{'}$.

Table 2. Relative Efficiency (%) of T_{p_1} with .respect to estimators T, $\bar{y}_n$ and $\bar{y}_2^{'}$ using real life data.

Estimators	Estimates	MSE	Efficiency %
T_{p_1}	354	249.98	100
T	378	269.39	107.76
$\bar{y}_n$	300	1808.14	723.31
$\bar{y}_2^{'}$	332	1240.26	496.14

The above table shows that the conclusions are the same as those of Table 1, that is the proposed estimator T_{p_1} is more efficient than T, $\bar{y}_n$ and $\bar{y}_2^{'}$ with maximum gain in efficiency occurring while comparing it with mean per unit estimator, which is very obvious.

8. Conclusion

In this study we have proposed a new chain ratio-to-regression estimator in successive sampling. The bias of the proposed estimator was computed up to the second order of approximation. The optimum replacement policy of the sampling fraction was obtained and considering bias up to the first order of approximation, the optimum mean square error of the proposed estimator was also obtained. The optimum mean square error of the proposed estimator was compared with that of

Singh's (2005) estimator and it was found that the proposed estimator is always better than Singh's (2005) estimator. Further, the proposed estimator was compared with mean per unit estimator and Cochran's (1977) estimator when no auxiliary information is used on any occasion. It was found that the proposed estimator is better than both of these estimators for $\rho \geq 0.5$, which is entirely justified. An example was considered by assuming different values of $\rho \geq 0.5$ to illustrate the above facts. At the end, a real life study was also done to demonstrate that the proposed estimator is more efficient than the other three existing estimators. Hence, the proposed estimator is recommended for further use.

REFERENCES

CENSUS OF INDIA, (2001). www.cesusindiagov.in.

CENSUS OF INDIA, (2011). www.cesusindiagov.in.

COCHRAN, W. G., (1977). Sampling Techniques. Third edition. Wiley Eastern Ltd.

DAS, A. K., (1982). Estimation of population ratio on two occasions. Journal of the Indian Society of Agricultural Statistics. 34: 1−9.

GUPTA, P. C., (1979). Sampling on two successive occasions. Journal of Statistical Research. 13: 7−16.

PATTERSON, H. D., (1950). Sampling on successive occasions with partial replacement of units. Journal of the Royal Statistical Society. 12(B): 241−255.

RAO, J. N. K., GRAHAM J., E., (1964). Rotation designs for sampling on repeated occasions. Journal of American Statistical Association. 59: 492−509.

SEN, A. R., (1971). Successive sampling with two auxiliary variables. Sankhya. 33(B): 371−378.

SEN, A. R., (1972). Successive sampling with $p(p \geq 1)$ auxiliary variables, The Annals of Mathematical Statistics. 43: 2031−2034.

SEN, A. R., (1973). Theory and application of sampling on repeated occasions with several auxiliary variables. Biometrics. 29: 381−385.

SINGH, G. N., (2003). Estimation of population mean using auxiliary information on recent occasion in h occasions successive sampling. Statistics in Transition. 6(4): 523−532.

SINGH, G. N., (2005). On the use of chain-type ratio estimator in successive sampling. Statistics in Transition. 7(1): 21−26.

SINGH, G. N., PRIYANKA, K., (2006). On the use of chain-type ratio to difference estimator in successive sampling. International Journal of Applied Mathematics and Statistics. 5(S06): 41−49.

SINGH, G. N., PRIYANKA, K., (2007). On the use of auxiliary information in search of good rotation patterns on successive occasions. Bulletin of Statistics and Economics. 1(A07): 42−60.

SINGH, G. N., PRIYANKA, K., (2008). On the use of several auxiliary variates to improve the precision of estimates at current occasion. Journal of the Indian Society of Agricultural Statistics. 62(3): 253−265.

SINGH, G. N., SINGH, V. K., (2001). On the use of auxiliary information in successive sampling. Journal of the Indian Society of Agricultural Statistics.54(1): 1−12.

SINGH, H. P., VISHWAKARMA, G. K., (2007). A general class of estimators in successive sampling. Metron. LXV(2): 201−227.

SINGH, P., TALWAR, H. K., (1991). Estimation of population regression coefficient in successive sampling. Biometrical Journal. 33: 599−605.

APPENDICES

Appendix A

Bias of the proposed estimator

$$T_p = \frac{\bar{y}_m}{\bar{x}_m}\left[\bar{x}_m + b_{xz}\left(\bar{z}_n - \bar{z}_m\right)\right]\frac{\bar{Z}}{\bar{z}_n}$$

Bias (T_p) = E(T_p) - $\bar{Y}$

Let

$\bar{y}_m = \bar{Y} + \varepsilon_0^{'}$, $\bar{z}_m = \bar{Z} + \varepsilon_1^{'}$, $\bar{z}_n = \bar{Z} + \varepsilon_1$, $\bar{x}_m = \bar{X} + \varepsilon_2^{'}$, $s_{xz} = S_{xz} + \varepsilon_3$, $s_z^2 = S_z^2 + \varepsilon_4$, $\beta_{xz} = \frac{S_{xz}}{S_z^2}$

such that $E\left(\varepsilon_i\right) = 0$ assuming the terms having order higher than 2 in ε 's are negligible, then

$$T_p = \left(\frac{\bar{Y}+\varepsilon_0^{'}}{\bar{X}+\varepsilon_2^{'}}\right)\left[\bar{X}+\varepsilon_2^{'}+\left(\frac{S_{xz}+\varepsilon_3}{S_z^2+\varepsilon_4}\right)\left(\bar{Z}+\varepsilon_1-\bar{Z}-\varepsilon_1^{'}\right)\right]\left(\frac{\bar{Z}}{\bar{Z}+\varepsilon_1}\right)$$

$$= \frac{1}{\bar{X}}\left(\bar{Y}+\varepsilon_0^{'}\right)\left(1-\frac{\varepsilon_2^{'}}{\bar{X}}+\frac{(\varepsilon_2^{'})^2}{\bar{X}^2}\right)\left[\bar{X}+\varepsilon_2^{'}+\left(1+\frac{\varepsilon_3}{S_{xz}}\right)\left(1-\frac{\varepsilon_4}{S_z^2}+\frac{{\varepsilon_4}^2}{S_z^4}\right)\beta_{xz}\left(\varepsilon_1-\varepsilon_1^{'}\right)\right]\left(1-\frac{\varepsilon_1}{\bar{Z}}+\frac{{\varepsilon_1}^2}{\bar{Z}^2}\right)$$

$$= \left\{\begin{array}{l}\frac{1}{\bar{X}}\left(\bar{Y}+\varepsilon_0^{'}-\frac{\bar{Y}}{\bar{X}}\varepsilon_2^{'}-\frac{1}{\bar{X}}\varepsilon_0^{'}\varepsilon_2^{'}+\frac{\bar{Y}}{\bar{X}^2}(\varepsilon_0^{'})^2-\bar{Y}\varepsilon_1-\varepsilon_1\varepsilon_0^{'}+\frac{\bar{Y}}{\bar{X}}\varepsilon_1\varepsilon_2^{'}+\frac{\bar{Y}}{\bar{Z}^2}{\varepsilon_1}^2\right)\\ \left[\bar{X}+\varepsilon_2^{'}+\beta_{xz}\left(\varepsilon_1-\frac{\varepsilon_1\varepsilon_4}{S_z^2}+\frac{\varepsilon_1\varepsilon_3}{S_{xz}}-\varepsilon_1^{'}+\frac{\varepsilon_1^{'}\varepsilon_4}{S_z^2}+\frac{\varepsilon_1^{'}\varepsilon_3}{S_{xz}}\right)\right]\end{array}\right\}$$

$$= \left\{\frac{1}{\bar{X}}\left[\begin{array}{l}\bar{Y}\bar{X}+\beta_{xz}\bar{Y}\left(-\frac{\varepsilon_1\varepsilon_4}{S_z^2}+\frac{\varepsilon_1\varepsilon_3}{S_{xz}}+\frac{\varepsilon_1^{'}\varepsilon_4}{S_z^2}+\frac{\varepsilon_1^{'}\varepsilon_3}{S_{xz}}\right)+\beta_{xz}\left(\varepsilon_0^{'}\varepsilon_1-\varepsilon_0^{'}\varepsilon_1^{'}\right)\\ -\beta_{xz}\frac{\bar{Y}}{\bar{X}}\left(\varepsilon_2^{'}\varepsilon_1-\varepsilon_1^{'}\varepsilon_2^{'}\right)-\beta_{xz}\left(\varepsilon_1^2-\varepsilon_1\varepsilon_1^{'}\right)-\bar{X}\varepsilon_0^{'}\varepsilon_1+\frac{\bar{X}\bar{Y}}{\bar{Z}^2}\varepsilon_1^2\end{array}\right]\right\}$$

Therefore,

$$E(T_p) = \bar{Y} + \frac{1}{\bar{X}}\left\{\begin{array}{l} \beta_{xz}\bar{Y}\left(-\frac{Cov\left(\bar{z}_n, s_z^2\right)}{S_z^2} + \frac{Cov\left(\bar{z}_n, s_{xz}\right)}{S_{xz}} + \frac{Cov\left(\bar{z}_m, s_z^2\right)}{S_z^2} + \frac{Cov\left(\bar{z}_m, s_{xz}\right)}{S_{xz}}\right) \\ +\beta_{xz}\left[Cov\left(\bar{y}_m, \bar{z}_n\right) - Cov\left(\bar{y}_m, \bar{z}_m\right)\right] - \frac{\bar{Y}}{\bar{X}} Var\left(\bar{x}_m\right) \\ -\beta_{xz}\frac{\bar{Y}}{\bar{X}}\left[Cov\left(\bar{x}_m, \bar{z}_n\right) - Cov\left(\bar{x}_m, \bar{z}_m\right)\right] - \bar{X} Cov\left(\bar{y}_m, \bar{z}_n\right) + \frac{\bar{X}\bar{Y}}{\bar{Z}^2} Var\left(\bar{z}_n\right) \\ \end{array}\right\} \qquad (18)$$

Hence, $Bias\left(T_p\right) = E\left(T_p\right) - \bar{Y}$

$$= \frac{1}{\bar{X}}\left\{\begin{array}{l} \beta_{xz}\bar{Y}\left(-\frac{Cov\left(\bar{z}_n, s_z^2\right)}{S_z^2} + \frac{Cov\left(\bar{z}_n, s_{xz}\right)}{S_{xz}} + \frac{Cov\left(\bar{z}_m, s_z^2\right)}{S_z^2} + \frac{Cov\left(\bar{z}_m, s_{xz}\right)}{S_{xz}}\right) \\ +\beta_{xz}\left[Cov\left(\bar{y}_m, \bar{z}_n\right) - Cov\left(\bar{y}_m, \bar{z}_m\right)\right] - \frac{\bar{Y}}{\bar{X}} Var\left(\bar{x}_m\right) \\ -\beta_{xz}\frac{\bar{Y}}{\bar{X}}\left[Cov\left(\bar{x}_m, \bar{z}_n\right) - Cov\left(\bar{x}_m, \bar{z}_m\right)\right] - \bar{X} Cov\left(\bar{y}_m, \bar{z}_n\right) + \frac{\bar{X}\bar{Y}}{\bar{Z}^2} Var\left(\bar{z}_n\right) \\ \end{array}\right\}$$

Appendix B

Mean Square Error of the proposed estimator

$$T_p = \frac{\bar{y}_m}{\bar{x}_m}\left[\bar{x}_m + b_{xz}\left(\bar{z}_n - \bar{z}_m\right)\right]\frac{\bar{Z}}{\bar{z}_n}$$

From (18), we can see that up to the first order of approximation $E\left(T_p\right) = \bar{Y}$. Therefore, mean square error up to the first order of approximation is given by $M\left(T_p\right) = E\left(T_p\right)^2 - \bar{Y}^2$.

Now,

$$T_p = \left(\frac{\bar{Y} + \varepsilon_0'}{\bar{X} + \varepsilon_2'}\right)\left[\bar{X} + \varepsilon_2' + \left(\frac{S_{xz} + \varepsilon_3}{S_z^2 + \varepsilon_4}\right)\left(\bar{Z} + \varepsilon_1 - \bar{Z} - \varepsilon_1'\right)\right]\left(\frac{\bar{Z}}{\bar{Z} + \varepsilon_1}\right)$$

$$= \frac{1}{\bar{X}}\left(\bar{Y}+\varepsilon_0'\right)\left(1-\frac{\varepsilon_2'}{\bar{X}}\right)\left[\bar{X}+\varepsilon_2'+\left(1+\frac{\varepsilon_3}{S_{xz}}\right)\left(1-\frac{\varepsilon_4}{S_z^2}\right)\beta_{xz}\left(\varepsilon_1-\varepsilon_1'\right)\right]\left(1-\frac{\varepsilon_1}{\bar{Z}}\right)$$

Therefore,

$$\left(T_p\right)^2 = \frac{1}{\bar{X}^2}\left\{\left(\bar{Y}+\varepsilon_0'-\frac{\bar{Y}}{\bar{X}}\varepsilon_2'\right)\left[\bar{X}+\varepsilon_2'+\left(1+\frac{\varepsilon_3}{S_{xz}}-\frac{\varepsilon_4}{S_z^2}\right)\beta_{xz}\left(\varepsilon_1-\varepsilon_1'\right)\right]\left(1-\frac{\varepsilon_1}{\bar{Z}}\right)\right\}^2$$

Now again,

$$E\left(T_p\right)^2 = \frac{1}{\bar{X}^2}E\left\{\bar{X}\bar{Y}+\bar{X}\varepsilon_0'+\beta_{xz}\bar{Y}\varepsilon_1-\beta_{xz}\bar{Y}\varepsilon_1'-\frac{\bar{X}\bar{Y}}{\bar{Z}}\varepsilon_1\right\}^2$$

Assuming the terms up to the second order of approximation and neglecting terms higher than 2nd power of ε ’s, we have

$$E\left(T_p\right)^2 = \frac{1}{\bar{X}^2}E\left\{\begin{array}{l}\left(\bar{X}\bar{Y}\right)^2+\bar{X}^2\left(\varepsilon_0'\right)^2+\beta_{xz}^2\bar{Y}^2\varepsilon_1^2+\beta_{xz}^2\bar{Y}^2\left(\varepsilon_1'\right)^2+\left(\frac{\bar{X}\bar{Y}}{\bar{Z}}\right)^2\varepsilon_1^2+2\beta_{xz}\bar{X}\bar{Y}\varepsilon_0'\varepsilon_1\\ -2\beta_{xz}\bar{X}\bar{Y}\varepsilon_0'\varepsilon_1'-2\frac{\bar{X}^2\bar{Y}}{\bar{Z}}\varepsilon_0'\varepsilon_1-2\beta_{xz}^2\bar{Y}^2\varepsilon_1'\varepsilon_1-2\beta_{xz}\frac{\bar{X}\bar{Y}^2}{\bar{Z}}\varepsilon_1^2+2\beta_{xz}\frac{\bar{X}\bar{Y}^2}{\bar{Z}}\varepsilon_1\varepsilon_1'\end{array}\right\}$$

$$= \bar{Y}^2\left\{\begin{array}{l}1+\frac{C_y^2}{m}+\frac{\rho_{xz}^2C_x^2}{n}+\frac{\rho_{xz}^2C_x^2}{m}+\frac{C_z^2}{n}+2\frac{\rho_{xz}\rho_{yz}C_xC_y}{n}-2\frac{\rho_{xz}\rho_{yz}C_xC_y}{m}\\ -2\frac{\rho_{yz}C_yC_z}{n}-2\frac{\rho_{xz}^2C_x^2}{n}\end{array}\right\}$$

$$= \bar{Y}^2\left\{1+\frac{C_y^2}{m}+\left(\frac{1}{m}-\frac{1}{n}\right)\left(\rho_{xz}^2C_x^2-2\rho_{xz}\rho_{yz}C_xC_y\right)+\frac{1}{n}\left(C_z^2-2\rho_{yz}C_yC_z\right)\right\}$$

Therefore,

$$M(T_p) = E(T_p)^2 - \bar{Y}^2 =$$

$$\bar{Y}^2\left\{\frac{C_y^2}{m}+\left(\frac{1}{m}-\frac{1}{n}\right)\left(\rho_{xz}^2C_x^2-2\rho_{xz}\rho_{yz}C_xC_y\right)+\frac{1}{n}\left(C_z^2-2\rho_{yz}C_yC_z\right)\right\}$$

Appendix C

Minimum Mean Square Error of T_{p_1}

From equation (8), we have

$$M\left(T_{p_1}\right)=\psi^2M\left(T_1\right)+\left(1-\psi\right)^2M\left(T_p\right) \tag{19}$$

Since M (T_{p_1}) is a function of an unknown constant ψ, therefore, it is minimized with respect to ψ, i.e. differentiating M (T_{p_1}) with respect to ψ and equating them to zero, we get

$$\Rightarrow \psi_{opt}=\frac{M\left(T_p\right)}{M\left(T_1\right)+M\left(T_p\right)}$$

Then, substituting ψ_{opt} from equation (19), the minimum MSE of T_{p_1} as

$$M\left(T_{p_1}\right)_{opt}=\frac{M\left(T_1\right)M\left(T_p\right)}{\left[M\left(T_1\right)+M\left(T_p\right)\right]}$$

For simplification, let

$$A^{'}=\overline{Y}^2C_y^2,\ \ B^{'}=\overline{Y}^2\left(\rho_{xz}^2C_x^2\text{-}2\rho_{xz}\rho_{yz}C_xC_y\right),\ \ C^{'}=\overline{Y}^2\left(C_z^2\text{-}2\rho_{yz}C_yC_z\right)$$

Therefore,

$$M\left(T_1\right)=\frac{A^{'}+C^{'}}{u}=\frac{A^{'}+C^{'}}{n\mu}$$

$$M\left(T_p\right)=\frac{A^{'}}{m}+\left(\frac{1}{m}\text{-}\frac{1}{n}\right)B^{'}+\frac{C^{'}}{n}=\frac{A^{'}}{n\left(1\text{-}\mu\right)}+\frac{B^{'}\mu}{n\left(1\text{-}\mu\right)}+\frac{C^{'}}{n}\quad \text{since } m=n\text{-}u=n\text{-}n\mu=n\left(1\text{-}\mu\right)$$

Now,

$$M\left(T_{p_1}\right)_{opt}=\frac{\frac{A^{'}+C^{'}}{n}\left[A^{'}+B^{'}\mu+C^{'}\left(1\text{-}\mu\right)\right]}{\frac{\left(A^{'}+C^{'}\right)\left(1\text{-}\mu\right)+A^{'}\mu+B^{'}\mu^2+C^{'}\mu\left(1\text{-}\mu\right)}{\left(1\text{-}\mu\right)}}$$

$$=\frac{\left(A^{'}+C^{'}\right)^2+\left(A^{'}+C^{'}\right)\left(B^{'}\text{-}C^{'}\right)\mu}{n\left[\left(A^{'}+C^{'}\right)+\left(B^{'}\text{-}C^{'}\right)\mu^2\right]}$$

or

$$M\left(T_{p_1}\right)_{opt}=\frac{\left(\alpha_1^{'}\right)^2+\alpha_1^{'}\alpha_2^{'}\mu}{n\left[\alpha_1^{'}+\alpha_2^{'}\mu^2\right]} \tag{20}$$

where

$$\alpha_1^{'}=A^{'}+C^{'},\qquad \alpha_2^{'}=B^{'}\text{-}C^{'}$$

Appendix D

Optimum Replacement Policy

Differentiating $M\left(T_{p_1}\right)_{opt}$ from equation (20) with respect to μ and equating them to zero, we get

$$\frac{1}{n}\left[\frac{\left(\alpha_1^{'}+\alpha_2^{'}\mu^2\right)\left(\alpha_1^{'}\alpha_2^{'}\right)-\left(\left(\alpha_1^{'}\right)^2+\alpha_1^{'}\alpha_2^{'}\mu\right)\left(2\alpha_2^{'}\mu\right)}{\left[\alpha_1^{'}+\alpha_2^{'}\mu^2\right]^2}\right]=0$$

$$\Rightarrow\left(\alpha_1^{'}\alpha_2^{'}\right)\mu^2+2\left(\alpha_1^{'}\alpha_2^{'}\right)\mu-\left(\alpha_1^{'}\right)^2\alpha_2^{'}=0$$

Then, $\mu=\left[\dfrac{-\alpha_1^{'}\pm\sqrt{\left(\alpha_1^{'}\right)^2+\left(\alpha_1^{'}\alpha_2^{'}\right)}}{\alpha_2^{'}}\right]$

Now, assuming $C_x=C_y=C_z=C(\text{say})$

$$\begin{aligned}\alpha_1^{'}&=A^{'}+C^{'}=\overline{Y}^2C_y^2+\overline{Y}^2\left(C_z^2-2\rho_{yz}C_yC_z\right)\\&=2\overline{Y}^2C^2\left(1-\rho_{yz}\right)\\\alpha_2^{'}&=B^{'}-C^{'}=\overline{Y}^2\left[\rho_{xz}^2C_x^2-2\rho_{xz}\rho_{yz}C_xC_y\right]-\overline{Y}^2\left(C_z^2-2\rho_{yz}C_yC_z\right)\\&=\overline{Y}^2C^2\left[\left(1-\rho_{xz}\right)\left(2\rho_{yz}-1-\rho_{xz}\right)\right]\\\alpha_1^{'}\alpha_2^{'}&=\left(A^{'}+C^{'}\right)\left(B^{'}-C^{'}\right)\\&=2\overline{Y}^2C^2\left(1-\rho_{yz}\right)\left\{\overline{Y}^2C^2\left[\left(1-\rho_{xz}\right)\left(2\rho_{yz}-1-\rho_{xz}\right)\right]\right\}\\&=\left(\overline{Y}^2C\right)^2\left[2\left(1-\rho_{yz}\right)\left(1-\rho_{xz}\right)\left(2\rho_{yz}-1-\rho_{xz}\right)\right]\end{aligned}$$

Therefore, $\mu=\dfrac{-\alpha_1^{'}\pm\sqrt{\left(\alpha_1^{'}\right)^2+\left(\alpha_1^{'}\alpha_2^{'}\right)}}{\alpha_2^{'}}$

$$=\frac{-2\left(1-\rho_{yz}\right)\pm\sqrt{2\left(1-\rho_{yz}\right)\left\{2\left(1-\rho_{yz}\right)+\left(1-\rho_{xz}\right)\left(2\rho_{yz}-\rho_{xz}-1\right)\right\}}}{\left(1-\rho_{xz}\right)\left(2\rho_{yz}-\rho_{xz}-1\right)}=\mu_0^{'}(\text{say})$$

Appendix E

Comparison of mean square error

(i). $M(T)_{opt}-M(T_{p_1})_{opt}>0$

$$\Rightarrow \frac{\alpha_1^2+\alpha_1\alpha_2\mu_0}{n\left[\alpha_1+\alpha_2\mu_0^2\right]}-\frac{\left(\alpha_1^{'}\right)^2+\alpha_1^{'}\alpha_2^{'}\mu_0^{'}}{n\left[\alpha_1^{'}+\alpha_2^{'}\left(\mu_0^{'}\right)^2\right]}>0$$

$$\Rightarrow \frac{\left[(1-\rho_{yz})+(\rho_{yz}-\rho_{xz})\mu_0\right]}{\left[(1-\rho_{yz})+(\rho_{yz}-\rho_{xz})\mu_0^2\right]}-\frac{\left[2(1-\rho_{yz})+(1-\rho_{xz})(2\rho_{yz}-1-\rho_{xz})\mu_0^{'}\right]}{\left[2(1-\rho_{yz})+(1-\rho_{xz})(2\rho_{yz}-1-\rho_{xz})\left(\mu_0^{'}\right)^2\right]}>0$$

(21)

Let us assume that $\rho_{xz}=\rho_{yz}=\rho(\text{say})$, equation (21) will become

$$\frac{\left[(1-\rho)+(0)\mu_0\right]}{\left[(1-\rho)+(0)\mu_0^2\right]}-\frac{\left[2(1-\rho)+(1-\rho)(\rho-1)\mu_0^{'}\right]}{\left[2(1-\rho)+(1-\rho)(\rho-1)\left(\mu_0^{'}\right)^2\right]}>0$$

$$\Rightarrow 1-\frac{\left[2-(1-\rho)\mu_0^{'}\right]}{\left[2-(1-\rho)\left(\mu_0^{'}\right)^2\right]}>0$$

Therefore,

$$M(T)_{opt}-M(T_{p_1})_{opt}>0 \;\Rightarrow\; (1-\rho)\left(1-\mu_0^{'}\right)>0$$

(ii). $V(\bar{y}_n)-M(T_{p_1})_{opt}>0$

$$\Rightarrow \left\{1-\frac{2(1-\rho_{yz})\left[2(1-\rho_{yz})+(1-\rho_{xz})(2\rho_{yz}-1-\rho_{xz})\mu_0^{'}\right]}{\left[2(1-\rho_{yz})+(1-\rho_{xz})(2\rho_{yz}-1-\rho_{xz})\left(\mu_0^{'}\right)^2\right]}\right\}>0 \qquad (22)$$

Assuming that $\rho_{xz}=\rho_{yz}=\rho(\text{say})$, equation (22) will become

$$1-\frac{2(1-\rho)\left[2-(1-\rho)\mu_0^{'}\right]}{\left[2-(1-\rho)\left(\mu_0^{'}\right)^2\right]}>0$$

$$\Rightarrow\left[2-(1-\rho)\left(\mu_0^{'}\right)^2\right]-2(1-\rho)\left[2-(1-\rho)\mu_0^{'}\right]>0$$

Therefore,

$$V\left(\bar{y}_n\right)-M\left(T_{p_1}\right)_{opt}>0$$

$$\Rightarrow \left[2-(1-\rho)\left(\mu_0^{'}\right)^2\right]>2(1-\rho)\left[2-(1-\rho)\mu_0^{'}\right]$$

(iii). $V\left(\bar{y}_2^{'}\right)_{opt}-M\left(T_{p_1}\right)_{opt}>0$

$$\Rightarrow \left\{\left[1+\sqrt{1-\rho_{xy}^2}\right]\frac{\bar{Y}^2C^2}{2n}-\frac{2\left(\bar{Y}^2C^2\right)^2\left(1-\rho_{yz}\right)\left[2\left(1-\rho_{yz}\right)+\left(1-\rho_{xz}\right)\left(2\rho_{yz}-1-\rho_{xz}\right)\mu_0^{'}\right]}{n\left(\bar{Y}^2C^2\right)\left[2\left(1-\rho_{yz}\right)+\left(1-\rho_{xz}\right)\left(2\rho_{yz}-1-\rho_{xz}\right)\left(\mu_0^{'}\right)^2\right]}\right\}>0 \tag{23}$$

Assuming $\rho_{xz}=\rho_{yz}=\rho(\text{say})$, equation (23) will become

$$\Rightarrow\left(1+\sqrt{1-\rho^2}\right)\left[2-(1-\rho)\left(\mu_0^{'}\right)^2\right]-4(1-\rho)\left[2-(1-\rho)\mu_0^{'}\right]>0$$

Therefore,

$$V\left(\bar{y}_2^{'}\right)_{opt}-M\left(T_{p_1}\right)_{opt}>0$$

$$\Rightarrow \left(1+\sqrt{1-\rho^2}\right)\left[2-(1-\rho)\left(\mu_0^{'}\right)^2\right]>4(1-\rho)\left[2-(1-\rho)\mu_0^{'}\right]$$

STATISTICS IN TRANSITION new series, Summer 2015
Vol. 16, No. 2, pp. 203–222

MULTINOMIAL LOGISTIC REGRESSION APPROACH FOR THE EVALUATION OF BINARY DIAGNOSTIC TEST IN MEDICAL RESEARCH

Alok Kumar Dwivedi[1], Indika Mallawaarachchi[2], Juan B. Figueroa-Casas[3], Angel M. Morales[4], Patrick Tarwater[5]

ABSTRACT

Evaluating the effect of variables on diagnostic measures (sensitivity, specificity, positive, and negative predictive values) is often of interest to clinical researchers. Logistic regression (LR) models can be used to predict diagnostic measures of a screening test. A marginal model framework using generalized estimating equation (GEE) with logit/log link can be used to compare the diagnostic measures between two or more screening tests. These individual modeling approaches to each diagnostic measure ignore the dependency among these measures that might affect the association of covariates with each diagnostic measure. The diagnostic measures are computed using joint distribution of screening test result and reference test result which generates a multinomial response data. Thus, multinomial logistic regression (MLR) is a more appropriate approach to modeling these diagnostic measures. In this study, the validity of LR and GEE approaches as compared to MLR model was assessed for modeling diagnostic measures. All methods provided unbiased estimates of diagnostic measures in the absence of any covariate. LR and GEE methods produced more biased estimates as compared to MLR approach especially for small sample size studies. No bias was obtained in predicting sensitivity measure using MLR method for one screening test. Our proposed MLR method is robust for modeling

[1] Assistant Professor, Division of Biostatistics & Epidemiology, Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center, El Paso, Texas, USA.
E-mail: alok.dwivedi@ttuhsc.edu.

[2] Research Associate, Division of Biostatistics & Epidemiology, Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center, El Paso, Texas, USA.
E-mail: indika.mallawaarachchi@ttuhsc.edu.

[3] Associate Professor, Division of Pulmonary and Critical Care Medicine, Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center, El Paso, Texas, USA.
E-mail: Juan.Figueroa@ttuhsc.edu.

[4] Assistant Professor, Department of Surgery, Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center, El Paso, TX 79905, USA. E-mail: angel.morales@ttuhsc.edu.

[5] Professor, Division of Biostatistics & Epidemiology, Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center, El Paso, Texas, USA.
E-mail: patrick.tarwater@ttuhsc.edu.

diagnostic measures of a screening test as opposed to LR method. MLR method and GEE method produced similar estimates of diagnostic measures for comparing two screening tests in large sample size studies. The proposed MLR model for diagnostic measures is simple, and available in common statistical software. Our study demonstrates that MLR method should be preferred as an alternative for modeling diagnostic measures.

Key words: multinomial logistic regression, predictive values, sensitivity, specificity, acute appendicitis, pulmonary abnormalities, medical diagnostic test.

1. Introduction

Diagnostic tests are an essential component in medical care for confirming or establishing the disease diagnosis, evaluating disease prognosis, stratifying risk of disease, and screening for early detection. Clinical researchers conduct studies about diagnostic tests mainly for the purpose of either estimating the diagnostic accuracy of a test according to different patient or environmental characteristics or comparing diagnostic accuracy of different tests according to different patient or environmental characteristics. Very limited statistical methods are available to evaluate the diagnostic measures in regression framework (Leisenring et al., 1997). Studies are required to develop robust statistical methods to analyze data from diagnostic studies and assess the properties of available statistical methods. In this study, we proposed a statistical regression method to analyze data from diagnostic studies.

In diagnostic studies, an investigational/new test is often referred to as screening/diagnostic test and a definite diagnostic test is referred to as the reference or gold standard test. When the screening test and reference test are measured in a binary outcome then various measures are required to assess the performance of screening tests in relation to the reference test. Most commonly used diagnostic performance measures are sensitivity (P{positive test result|disease}), specificity (P{negative test result|no disease}), positive predictive value (P{disease|positive test result}), and negative predictive value (P{no disease|negative test result}) (Leisenring et al., 2000). Sensitivity and specificity are probabilities of the test result measured through a screening test, conditional on disease status measured through a reference test while a predictive value is the probability of disease conditional on the test result measured through a screening test. Clinical researchers are often interested in evaluating these four diagnostic measures of screening tests according to patient and clinical characteristics. Regression approaches are needed to address such clinical questions.

Application of logistic regression (LR) in predicting common diagnostic measures including sensitivity (Se), specificity (Sp), positive predictive value (PPV), and negative predictive value (NPV) of a screening test according to patient or other environmental covariates was proposed by Coughlin et al. (1992). LR models for Se and Sp include reference test result as an independent variable while modeling PPV and NPV include screening test result as a predictor. We refer to this modeling approach as adjusted LR models for diagnostic measures. The adjusted LR models have been used in clinical studies for evaluating diagnostic measures (Coughlin et al., 1992; Elie et al., 2008). Another alternative is the use of LR models for Se and Sp by restricting the analysis to disease and non-disease group respectively. Similarly, LR models can be used to model PPV and NPV by restricting the analysis to positive screening test result and negative screening test result respectively. We refer to the modeling approaches restricted to a group of individuals as subgroup LR models. Subgroup LR models have also been used in clinical studies (Carney et al., 2003; Laya et al., 1996). Recently an application of LR model for predicting likelihood ratio was also developed (Janssens et al., 2005). Ordinary LR models are sensitive to small sample size and rare events (Nemes et al., 2009; King and Zeng, 2001). Thus, LR models may produce biased estimates of diagnostic measures. Therefore, we determined the bias in estimating diagnostic measures using adjusted and subgroup LR models in presence of a binary cofactor in various scenarios.

The diagnostic measures depend on the four cell frequencies generated from a 2x2 table of screening test result and reference test result. The most natural way is to model the joint distribution of screening test result and reference test result. Typically, each diagnostic measure is modelled independently using LR as a function of risk factors. Since the diagnostic measures are computed using the joint distribution of screening test result and reference test result thus these measures are dependent. Independent modeling of these measures ignores dependency among these measures and that subsequently might affect the association of cofactors with these measures (Puggioni et al., 2008). Since the joint distribution of screening test result and reference test result follows a multinomial distribution, thus a multinomial logistic regression (MLR) can be used to estimate the diagnostic measures. We compared the performance of LR models and MLR model for estimating the common diagnostic measures using simulation studies and our published study data (Figueroa-Casas et al., 2014). We can easily extend the MLR model for comparing two or more screening tests. However, studies involving two or more screening tests often provide paired structure data since each patient usually undergoes through each screening test. Thus, such studies require accounting for clustering effects in the analysis. Sandwich error estimation is commonly used to analyze clustered data, repeated measures data, and data obtained through clustered randomized design. Such procedure provides robust variance estimation. Robust variance approach appropriately accounts for correlation structure in the dataset (Leisenring et al., 1997; Liu, 1998). We suggest using a robust variance approach while modeling diagnostic measures using MLR method for two or more screening tests.

A marginal model framework using generalized estimating equation (GEE) with logit link approach has been proposed to compare diagnostic measures between two or more screening tests. It has been advocated to use independent working correlation matrix for fitting marginal models for diagnostic measures with robust variance estimates (Leisenring et al., 1997; Leisenring et al., 2000).

As discussed earlier for LR models, adjusted GEE and subgroup GEE models can be fitted to compare diagnostic measures between two or more screening tests. Further, the individual approaches to modeling each diagnostic measure through GEE models do not account for dependency among these measures. We also compared the individual modeling approach using GEE models and a joint modeling approach using MLR models for estimating diagnostic measures with simulation studies and real study dataset.

The aim of this study was to propose an alternative regression approach to evaluating a binary diagnostic test based on joint distribution of a new test result with reference test result. Specifically, we evaluated the validity of the proposed MLR approach in estimating diagnostic measures and compared with subgroup and adjusted LR models of diagnostic measures. In addition, we extended MLR approach to modeling more than one screening test for comparing diagnostic measures between screening tests and compared it with GEE approach to modeling diagnostic measures for more than one screening test. The applications of MLR approach for estimating and comparing diagnostic measures were illustrated using data from medical research studies.

2. Methods

2.1. Estimating diagnostic accuracy using a logistic regression (LR) model

Suppose a diagnostic study involves a screening test (T) and a reference test (D). If both the screening test and reference test provide binary (positive/negative) results then data can be summarized using a 2x2 table as presented in Table 1. Se, Sp, PPV, and NPV can be estimated as a/(a+c), d/(b+d), a/(a+b), and d/(c+d) respectively. We need regression approaches to estimate these diagnostic measures in presence of significant patient characteristics or other clinical covariates. LR models (Coughlin et al., 1992) can be used to predict Se, Sp, PPV, and NPV in relation to cofactors.

2.2. Multinomial logistic regression (MLR) for estimating diagnostic accuracy

The common diagnostic measures are based on the four cell frequencies obtained from Table 1. The probabilities of these four cells follow a multinomial distribution. Thus, MLR can be used for estimating common diagnostic measures in presence of patient and environmental covariates. Data summarized in Table 1 have unobserved probability p_k corresponding to each of the 4 cells, where $\sum_{k=1}^{4} p_k = 1$.

The joint probabilities for a screening test (T) and a reference test (D) would be:
P(T=1 and D=1) = True positive probability = a/(a+b+c+d)

P(T=1 and D =0) = False positive probability = b/(a+b+c+d)
P(T=0 and D =1) = False negative probability = c/(a+b+c+d)
P(T=0 and D =0) = True negative probability = d/(a+b+c+d)

A new outcome variable with four categories needs to be generated for applying MLR model. The four categories of the new outcome variable (Y=1, 2, 3, and 4) will be true positive (Y=1: T=1 and D=1), false positive (Y=2: T=1 and D=0), false negative (Y=3: T=0 and D =1), and true negative (Y=4: T=0 and D =0) as described in Table 1. We can fit MLR models by considering any one category as a reference category. For example, if we consider the false negative (Y=3) as a referent category then it compares the likelihood of true positive over false negative which is equivalent to fitting LR model for Se of screening test T. At the same time this model also provides comparison of true negative over false negative which is equivalent to fitting LR model for NPV of screening test T. Thus, a single MLR model can be used to predict Se, Sp, PPV, and NPV of a screening test. However, at least two LR models (one for Se and one for Sp) are needed to estimate all four diagnostic measures.

2.3. Comparing diagnostic accuracy using generalized estimating equation (GEE) and MLR methods

The data needs to be reorganized for comparing diagnostic measures of two or more screening tests using GEE or MLR methods. Suppose we have “n” subjects who underwent two screening tests, then it will be 2n records in a reorganized dataset. In this approach, an indicator variable (Z) is defined to classify each record for each specific test. In other words, each subject will have two records corresponding to each test. Suppose a subject has data on three variables: D, T_1 (screening test 1), and T_2 (screening test 2) in an original dataset, then that subject will have two records: D, T, Z=1 with $T=T_1$ and D, T, Z=0 with $T=T_2$ in a reorganized dataset. The logit or log link under the GEE framework can be applied in the reorganized dataset to compare the diagnostic measures between two screening tests (Moskowitz and Pepe, 2006). The equations for developing GEE models of diagnostic measures are published (Leisenring et al., 1997; Leisenring et al., 2000). MLR models using a robust variance approach can be used to compare diagnostic measures between two or more screening tests by modeling a new dependent variable Y (as described in section 2.2) in the reorganized dataset. The details of LR and MLR models can be found in the Appendix.

3. Data analysis

3.1. Simulation studies.

The performance of MLR as compared with LR models for estimating diagnostic measures was evaluated using Monte Carlo simulation studies. We first created a unique ID variable and a variable (X) from a Bernoulli distribution. We then created a random reference test variable (D) from the Bernoulli distribution with a mean equal to probability (p)

$\mathrm{Logit}(p)=a_1+a_2*X$, where $0\leq p\leq 1$

After that we randomly created a binary screening test (T) for each subject from the Bernoulli distribution having a mean p'. The p' was determined using the following function:

$\mathrm{Logit}(p')=b_1+ b_2*D - b_3 *(1-D)$, where $0\leq p' \leq 1$

where a_1 and b_1 are regression intercepts. The a_2, b_2, and b_3 are regression coefficients.

First, we compared the bias in all common four measures of diagnostic accuracy estimated using LR and MLR models in the absence of any cofactor. Then, we focused only on comparing the bias in the estimate of Se of the screening test T. The true Se for screening test T in relation to reference test D was obtained and compared with Se estimated using adjusted LR, subgroup LR, and MLR approaches.

The comparison of MLR and GEE methods for estimating diagnostic measures of two screening tests was also evaluated in various simulation studies as described for a single screening test. We randomly created two binary screening tests (T_1 and T_2) for each subject from the Bernoulli distributions having mean $p^{\dagger}$ and p^{*} respectively. The $p^{\dagger}$ and p^{*} were determined using the following functions:

$\mathrm{Logit}(p^{\dagger})=c_1+ c_2*D - c_3 *(1-D) + u_1$, where $0\leq p^{\dagger} \leq 1$

$\mathrm{Logit}(p^{*})=d_1+ d_2*D - d_3 *(1-D)+ u_2$, where $0\leq p^{*} \leq 1$

where c_1 and d_1 are regression intercepts. The c_2, c_3, d_2, and d_3 are regression coefficients. To introduce correlation between two screening tests, a random effect component (u_1, u_2) was included for the outcome of each test. The u_1 and u_2 were drawn from a bivariate normal distribution with a known correlation structure. The true Se for screening test T_1 and T_2 in relation to reference test D were obtained and compared with the estimated Se for each test obtained using adjusted GEE, subgroup GEE, and MLR approaches.

The percent relative bias in the estimate was reported. Each simulation study was conducted for a small sample size (100) as well as a large sample size (500). Each simulation study was also conducted for low prevalence (<10%) and moderate prevalence (20-30%). The effect of different prevalence of a binary cofactor was also examined. Each simulation study was repeated for 1000 simulations. The percent of relative bias was estimated using average of [(true diagnostic value – estimated diagnostic value)*100/true diagnostic value] from 1,000 random data sets. The choice of regression coefficients in the above models was made according to the simulation study. Statistical package STATA 12.1 was used for data analysis.

3.2. Real data analysis

To demonstrate our proposed strategy, we used data from two studies (study I and study II). In study I (single screening test), we were interested in assessing the accuracy of chest radiographs (chest x-ray) to identify bilateral pulmonary infiltrates consistent with acute respiratory distress syndrome in relation to computed tomography (CT, reference test). We used a subgroup LR model to determine the clinical characteristics associated with diagnostic performance measures of chest radiographs. A total of 90 patients met the inclusion criteria and had near simultaneous chest radiograph and CT results to evaluate for specified pulmonary abnormalities. The prevalence of these pulmonary abnormalities was 74% determined using CT (Figueroa-Casas et al., 2014). In the present study, we compared the results of subgroup LR models with our proposed MLR model to assess factors associated with the diagnostic measures of chest radiograph. For study II (two screening tests), we used our motivating study data on acute appendicitis. In study II, a total of 200 patients were evaluated with computer tomography (CT) for the diagnosis of appendicitis. The prevalence of acute appendicitis was found as 95.5%. The surgery residents and radiologists reviewed independently CT for each patient and made diagnosis for acute appendicitis. For each patient, we have pathological diagnosis for acute appendicitis. In this case, pathological diagnosis was considered as a reference test. The aim of this study was to compare the accuracy of CT readings with surgical residents as compared with radiologists. We compared the Se, Sp, PPV, and NPV of CT reading with surgical residents and radiologists in relation to pathological findings using GEE with logit link and robust variance estimation. MLR was also performed to compare Se, Sp, PPV, and NPV of CT reading with surgical residents with radiologists. The results of subgroup LR, subgroup GEE, and MLR approaches were reported using regression coefficient (RC), standard error (SE), and p-value.

4. Results

We found no bias in estimating Se, Sp, PPV, and NPV using either LR (adjusted or subgroup) or MLR methods in the absence of any cofactors. Table 2 shows the percent bias in estimating Se using subgroup LR, adjusted LR, and MLR methods. Subgroup LR model provided biased estimate of Se in the range of 0.06% to 31% while adjusted LR model provided biased estimate of Se in the range of 0.68% to 38.6% when sample size was 100. There was less bias in the estimate of Se using subgroup LR and no bias using MLR models when sample size was 500. However, we obtained biased estimates of Se in the range of 1.43% to 15.2% using adjusted LR models for sample size 500. The bias in the estimate of Se using LR model was found larger when the prevalence of disease was not similar between the two levels of a cofactor as compared to when the prevalence of disease was similar between the two levels of a cofactor. There was no bias obtained in estimating Se using MLR in any scenario.

Table 3 demonstrates the percent bias in estimating Se using subgroup LR, adjusted LR, and MLR models when the prevalence of disease was moderate (20-30%) for sample size n=100 and n=500. The bias in the estimate of Se using subgroup LR model was less than 8% when the sample size was small while no bias was obtained when the sample size was high (n=500). The bias in the estimated Se using adjusted LR model was obtained from 1% to 19.8% when the sample size was 100 while the bias in the Se using adjusted LR was obtained from 0.35% to 8.47% when the sample size was 500. No bias in any situations was obtained in estimating Se using MLR model.

In summary, the subgroup LR model always provided less biased estimate of Se as compared to adjusted LR model in any scenario. The bias in the estimate of Se was found to be higher when the prevalence of disease was different in different levels of a cofactor. Further, LR model with low prevalent cofactor provided large bias in the estimate of Se as compared to LR model with high prevalent cofactor. The two methods, subgroup LR and MLR, provided unbiased estimate of Se when the disease prevalence was more than 20% and cofactor prevalence was moderate (50%). MLR method always provided an unbiased estimate of Se in any scenario.

Table 4 illustrates the comparison of subgroup LR model and MLR model to evaluate factors associated with the diagnostic measures of chest x-rays in identifying bilateral pulmonary infiltrates consistent with the diagnosis of acute respiratory distress syndrome. Subgroup LR models provided slightly different estimates of regression coefficients and p-values as compared to MLR models. Slightly lower p-values were obtained in the subgroup LR models as compared to MLR models. We further developed LR and MLR models including only gender variable as a cofactor using study I data. We did not find bias in the estimates of diagnostic measures obtained using subgroup LR and MLR models when we included only gender variable. However, less than 3% bias in the estimates of diagnostic measures was obtained using adjusted LR model in study I dataset.

No bias was obtained in estimating any diagnostic measures using different methods for two screening tests in the absence of cofactors. The absolute percent relative bias in the estimate of Se using GEE and MLR methods for different scenarios is shown in Table 5 when the disease prevalence was low. The bias in the estimate of Se was found to be almost similar with subgroup GEE method and MLR method when disease prevalence was low and cofactor prevalence was 50%. However, slightly lower bias in the estimate of Se was obtained using MLR method as compared to subgroup GEE method when disease prevalence varied according to covariate strata. Less than 10% bias in the estimate of Se was obtained using subgroup GEE and MLR methods when cofactor prevalence was 50%. Adjusted GEE approach provided large bias in the estimate of Se as compared with subgroup GEE and MLR methods. Similar bias pattern was obtained across different methods for estimating Se when cofactor prevalence was 20%. The bias was found to be larger with each method when cofactor prevalence was low.

The absolute percent relative bias in the estimate of Se using GEE and MLR for different scenarios is shown in Table 6 when the disease prevalence was greater than 20%. The range of bias in the estimate of Se using subgroup GEE model and MLR model was found to be 1.71%-5.81% for small sample size (n=100) and 0.33%-3.76% for large sample size (n=500) when cofactor prevalence was 50%. The bias was less than 9% with subgroup GEE and MLR models in an equal prevalence scenario and when the cofactor prevalence was 20%. The subgroup GEE and MLR models both produced bias in estimate of Se up to 16% when cofactor prevalence was 20% and disease prevalence was not the same in different strata. Adjusted GEE method provided very large bias in the estimate of Se in most of the scenarios.

Table 7 delineates a comparison of subgroup GEE and MLR models for determining the differences in diagnostic performance of radiologist CT reading for the diagnosis of appendicitis as compared to surgical residents after adjusting cofactors. Both approaches showed that CT reading with radiologist for the diagnosis of appendicitis had significantly higher Se and lower Sp than CT reading with surgical residents. The p-values obtained from MLR models were slightly different than the p-values obtained using GEE models. The p-values for comparison of specificity between two screening tests were obtained as 0.02 and 0.04 using MLR model and GEE model respectively after adjusting other cofactors.

5. Discussion

The diagnostic measures of a screening test depend upon (1) the cell frequencies generated from a cross-tabulation of screening test result and reference test result, and (2) the study population and clinical characteristics. We need a regression approach to modeling diagnostic measures that describe joint distribution of screening test result and reference test result. We proposed MLR model as direct modeling approach to modeling each common diagnostic measure. We further extended our approach to comparing diagnostic measures of two or more screening tests. The validity of available regression approaches in estimating the diagnostic measures in different scenarios was also estimated in this study. We found that our proposed MLR approach provides unbiased estimates of diagnostic measures as compared to LR methods. We also found our proposed MLR approach to be more appropriate for comparing two or more screening tests as opposed to adjusted GEE method.

Adjusted LR method provided bias in the estimate of Se up to 19% for small sample size and 12.4% for large sample size, when disease prevalence was low (10%) and cofactor prevalence was 50%. This bias was increased to 31% when the prevalence of covariate was 20% for a low sample size and a low prevalence study. Coughlin et al. (1992) also found 25% bias in the estimate of Se when the prevalence was unequal across covariate strata. In our study, subgroup LR method provided bias in the estimate of Se up to 8% when the prevalence was unequal across covariate strata and prevalence of covariate was 50%. Coughlin et al. (1992) found 7% bias using subgroup model when the prevalence was unequal across covariate strata. In general, adjusted LR model provided biased estimate of Se in all scenarios. Additionally, subgroup LR model provided bias estimates for large sample size studies when disease prevalence was less than 10% and cofactor prevalence was 20%. Our proposed MLR method produced unbiased estimate of Se in all scenarios.

It has been shown that ordinary LR model produces bias estimates for small sample size studies (Nemes et al., 2009; Bergtold et al., 2011). LR model produces large bias when the sample size is small and the outcomes are rare (King and Zeng, 2001). Thus, obviously utilizing LR models for modeling diagnostic measures in such cases will produce biased estimates. Our study demonstrates that MLR is less sensitive to small sample size as compared with LR models for modeling diagnostic measures. Ye and Lord (2014) also showed that MLR model requires smaller sample size as compared with mixed logit model using crash severity data. Further, modeling diagnostic measures directly through MLR approach avoids the dependency problem that arises through individual modeling of each diagnostic measure using LR approach.

In our real data example of accuracy of chest radiograph for detecting pulmonary abnormalities according to gender status, we found no bias in the estimates of any diagnostic measures using subgroup LR and MLR models while up to 3% bias was observed using adjusted LR model. It was expected to obtain unbiased estimates of diagnostic measures using subgroup LR model because disease prevalence (74%) and male gender prevalence (63%) were very high in the study. Slightly lower P-values were obtained in subgroup LR models as compared to MLR models. It has been observed that binary LR models for each pair of multi-response data underestimate the standard errors of the coefficients as compared with MLR model (Agresti, 2007). In other words, ignoring the dependency among diagnostic measures through individual modeling may provide smaller standard errors for the model parameters. Thus, individual modeling of each diagnostic measure using LR model may provide inappropriate inferences as opposed to our proposed MLR method of modeling diagnostic measures.

Subgroup GEE and MLR approaches provided similar results for modeling diagnostic measures of more than one screening test. Subgroup GEE method produced slightly higher biased estimates as compared with MLR model especially for studies with low sample size and low disease prevalence. Subgroup GEE and MLR approaches provided bias in the estimate of Se up to 9% when disease prevalence was low and up to 6% when disease prevalence was greater than 20% with cofactor prevalence 50%. This bias increased up to 33% when cofactor prevalence was 20%. This bias can be eliminated by restricting the analysis to the specific cofactor strata in MLR or subgroup GEE models. Adjusted GEE approach produced biased estimate of Se in almost all scenarios.

In our real data example for comparison of two screening tests, the Se of CT reading with radiologists was found larger than CT reading with residents. Another study observed no differences in diagnosing acute appendicitis though CT readings by radiology residents as compared with CT readings by radiology faculty (Albano et al., 2001). There were no differences observed between two approaches to comparing each diagnostic measure in the absence of any cofactors in study II dataset. It was expected that the MLR model for multi-response data is similar to modeling two separate logistic regressions in the absence of predictor and interaction (Fidler and Nagelkerke, 2013). Our simulation studies also confirmed these findings that the modeling diagnostic measures through GEE approach and MLR approach provide the same results in the absence of any covariates. Robust variance estimate and independent working correlation matrix were used in GEE models and robust variance estimate was used in MLR models. Despite this, GEE and MLR approaches produced slightly different estimates and p-values for comparing diagnostic measures between two screening tests. This further confirms that ignoring the dependency among diagnostic measures through individual modeling or choosing binomial marginal distribution for estimating diagnostic measures may provide inaccurate results as opposed to joint modeling of screening test result and reference test result using multinomial distribution. Further, it has been demonstrated that the MLR approach is not equivalent to modeling two separate logistic regressions for multi-response data in the presence of an interaction effect. The MLR approach should be preferred over two separate logistic regressions in the presence of a cofactor (Fidler and Nagelkerke, 2013; Miettinen, 1976).

In simulation studies, we have considered only one binary cofactor for the sake of simplicity. The MLR approach can handle both categorical and continuous cofactors. We demonstrated the MLR modeling of common diagnostic measures in presence of a perfect reference test. This approach can also be used in the absence of a perfect reference test. This application is under investigation by us for a future publication. We have shown bias in the estimate of Se measure using different methods. Similarly, we can demonstrate for other diagnostic measures. This study has not provided an inferential comparison in evaluating the association of a cofactor with diagnostic measures using different methods.

6. Conclusions

In this study, we showed MLR model can be used directly for modeling Se, Sp, PPV, and NPV as a function of covariates. We also demonstrated that MLR model can easily be extended for comparing diagnostic measures between more than one screening test. The correlation involved in multiple screening tests can be handled using robust variance approach available in statistical software. Developing MLR models for diagnostic measures is straightforward, simple, and available in common statistical software. In the absence of cofactors, all methods provided unbiased estimates of diagnostic measures. In general, all approaches provided very consistent results in many conditions. The MLR method always produced unbiased estimate of each diagnostic measure of a screening test. Subgroup LR method also produced unbiased estimate of each diagnostic measure in large sample size studies. The results of subgroup GEE and MLR with robust variance estimate for more than one screening test were found consistent. For small sample sizes, subgroup GEE and MLR approaches can produce bias estimates, especially with low prevalent cofactor. In such cases, a restricted analysis of covariate strata can be performed to correct the bias. Adjusted LR and adjusted GEE models should be avoided for predicting diagnostic measures. Subgroup LR and subgroup GEE models can be utilized for estimating diagnostic measures for large sample size studies. However, these methods may provide inaccurate inferences due to ignoring the dependency among the diagnostic measures. We suggest using MLR as an alternative and more appropriate approach to GEE with logit link and LR models for modeling Se, Sp, PPV and NPV.

REFERENCES

AGRESTI, A., (2007). An Introduction to Categorical Data Analysis. John Wiley & Sons, Inc., Hoboken, New Jersey, p. 174.

ALBANO, M. C., ROSS, G. W., DITCHEK, J. J., DUKE, G. L., TEEGER, S., SOSTMAN, H. D., FLOMENBAUM, N., SEIFERT, C., BRILL, P. W., (2001). Resident Interpretation of Emergency CT Scans in the Evaluation of Acute Appendicitis. Academic Radiology, 8, 915–918.

BERGTOLD, J. S., YEAGER, E. A., FEATHERSTONE, A., (2011). Sample Size and Robustness of Inferences from Logistic Regression in the Presence of Nonlinearity and Multicollinearity. The Annual Meeting of Agricultural and Applied Economics Association.

CARNEY, P. A., MIGLIORETTI, D. L., YANKASKAS, B. C., KERLIKOWSKE, K., ROSENBERG, R., RUTTER, C. M., GELLER, B. M., ABRAHAM, L. A., TAPLIN, S. H., DIGNAN, M., CUTTER, G., BALLARD-BARBASH, R., (2003). Individual and Combined Effects of Age, Breast Density, and Hormone Replacement Therapy Use on the Accuracy of Screening Mammography. Annals of Internal Medicine, 138(3), 168−75.

COUGHLIN, S. S., TROCK, B., CRIQUI, M. H., PICKLE, L. W., BROWNER, D., TEFFT, M. C., (1992). The Logistic Modeling of Sensitivity, Specificity, and Predictive Value of a Diagnostic Test. Journal of Clinical Epidemiology, 45, 1−7.

ELIE, C., COSTE, J., THE FRENCH SOCIETY OF CLINICAL CYTOLOGY STUDY, (2008). A Methodological Framework to Distinguish Spectrum Effects from Spectrum Biases and to Assess Diagnostic and Screening Test Accuracy for Patient Populations: Application to the Papanicolaou Cervical Cancer Smear Test. BMC Medical Research Methodology, 8, 7.

FIDLER, V., NAGELKERKE N., (2013). The Mantel-Haenszel Procedure Revisited: Models and Generalizations. PLoS One, 8(3), e58327.

FIGUEROA-CASAS, J. B., CONNERY, S. M., MONTOYA, R., DWIVEDI, A. K., LEE, S., (2014). Accuracy of Early Prediction of Duration of Mechanical Ventilation by Intensivists. Annals of the American Thoracic Society, 11(2), 182−185.

JANSSENS, A. C., DENG, Y., BORSBOOM, G. J., EIJKEMANS, M. J., HABBEMA, J. D., STEYERBERG, E. W., (2005). A New Logistic Regression Approach for the Evaluation of Diagnostic Test Results. Medical Decision Making, 25(2), 168−177.

KING, G., ZENG, L., (2001). Logistic Regression in Rare Events Data. Political Analysis, 9, 137−163.

LAYA M. B., LARSON E. B., TAPLIN S. H., WHITE E., (1996). Effect of Estrogen Replacement Therapy on the Specificity and Sensitivity of Screening Mammography. Journal of National Cancer Institute, 88(10), 643−649.

LEISENRING, W., PEPE, M. S., LONGTON, G., (1997). A Marginal Regression Modelling Framework for Evaluating Medical Diagnostic Tests. Statistics in Medicine, 16, 1263−1281.

LEISENRING, W., ALONZO, T., PEPE, M. S., (2000). Comparisons of Predictive Values of Binary Medical Diagnostic Tests for Paired Designs. Biometrics, 56, 345−351.

LIU, H., (1998). Robust Standard Error Estimate for Cluster Sampling Data: A SAS/IML Macro Procedure for Logistic Regression with Huberization. In: Proceedings of the Twenty-Third Annual SAS Users Group International.

MIETTINEN, O. S., (1976). Stratification by a Multivariate Confounder Score. American Journal of Epidemiology. 104, 609−620.

MOSKOWITZ, C. S., PEPE, M. S., (2006). Comparing the Predictive Values of Diagnostic Tests: Sample Size and Analysis for Paired Study Designs. Memorial Sloan-Kettering Cancer Center, Department of Epidemiology & Biostatistics Working Paper Series. Working Paper 5.

NEMES, S., JONASSON, J. M., GENELL, A., STEINECK, G., (2009). Bias in Odds Ratios by Logistic Regression Modelling and Sample Size. BMC Medical Research Methodology, 9, 56.

PUGGIONI, G., GELFAND, A. E., ELMORE, J. G., (2008). Joint Modeling of Sensitivity and Specificity. Statistics in Medicine, 27(10), 1745−1761.

YE, F., LORD, D., (2014). Comparing Three Commonly Used Crash Severity Models on Sample Size Requirements: Multinomial Logit, Ordered Probit and Mixed Logit Models. Analytic Methods in Accident Research, 1, 72−85.

APPENDICES

Appendix 1.

Table 1. Cross-tabulation of test results (T) with reference test (D)

Test result1	**References test**		**Total**
	Positive	Negative	
Positive	a (True positive)	b (False Positive)	a+ b
Negative	c (False Negative)	d (True Negative)	c+ d
Total	a+ c	b+ d	a+ b+ c+ d

Table 2. The percent relative bias in estimating Se using subgroup LR, adjusted LR, and MLR

Disease prevalence <= 10%		**N=100**			**N=500**		
		Sub-group LR	**Adjusted LR**	**MLR**	**Sub-group LR**	**Adjusted LR**	**MLR**
When x=50%							
Equal prevalence	X=1	2.99	10.27	0.00	0.00	-12.46	0.00
	X=0	0.22	3.11	0.00	0.00	-6.30	0.00
Unequal prevalence	X=1	7.60	19.10	0.00	0.00	-12.41	0.00
	X=0	0.80	6.60	0.00	0.00	-5.71	0.00
When x=20%							
Equal prevalence	X=1	22.85	33.01	0.00	0.46	-1.35	0.00
	X=0	0.00	-1.32	0.00	0.00	-1.20	0.00
Unequal prevalence	X=1	31.00	38.60	0.00	1.37	15.21	0.00
	X=0	0.06	-0.68	0.00	0.00	-1.43	0.00

*Se: 20-30%; Sp: 20-30%; Se: sensitivity; Sp: specificity, LR: logistic regression.

Table 3. The percent relative bias in estimating Se using subgroup LR, adjusted LR, and MLR

Disease prevalence > 20%		N=100			N=500		
		Sub-group LR	Adjusted LR	MLR	Sub-group LR	Adjusted LR	MLR
When x=50%							
Equal prevalence	X=1	0.00	-3.58	0.00	0.00	-1.98	0.00
	X=0	0.00	-6.20	0.00	0.00	-1.59	0.00
Unequal prevalence	X=1	0.19	-1.02	0.00	0.00	-5.23	0.00
	X=0	0.00	-4.34	0.00	0.00	-0.49	0.00
When x=20%							
Equal prevalence	X=1	2.95	14.29	0.00	0.00	-8.47	0.00
	X=0	0.00	-2.70	0.00	0.00	-0.50	0.00
Unequal prevalence	X=1	8.08	19.82	0.00	0.00	-10.69	0.00
	X=0	0.00	-1.58	0.00	0.00	-0.35	0.00

*Se: 20-30%; Sp: 20-30%; Se: sensitivity; Sp: specificity, LR: logistic regression.

Table 4. Models of sensitivity, specificity and predictive values using subgroup LR and MLR approaches

Diagnostic models	Subgroup LR			MLR		
	RC	SE	p-value	RC	SE	p-value
Se						
Female gender	1.722	0.807	0.033	1.683	0.806	0.037
BMI(Kg/m^2)>25	-0.761	0.854	0.372	-0.628	0.850	0.461
Sp						
Female gender	-1.596	1.025	0.120	-1.444	0.985	0.143
BMI (Kg/m^2)>25	-0.860	1.342	0.522	-0.426	1.273	0.738
PPV						
Female gender	-1.294	0.887	0.145	-1.310	0.888	0.140
BMI(Kg/m^2)>25	-0.460	1.156	0.690	-0.522	1.152	0.651
NPV						
Female gender	1.526	0.918	0.096	1.548	0.912	0.090
BMI(Kg/m^2)>25	-0.380	1.050	0.718	-0.532	1.001	0.595

BMI: Body mass index; Se: sensitivity; Sp: specificity; PPV: positive predictive value; NPV: negative predictive value, LR: logistic regression.

Table 5. The absolute percent relative bias in estimating Se using subgroup GEE, adjusted GEE, and MLR

Disease prevalence <= 10%	N=100			N=500		
	Sub-group GEE	**Adjusted GEE**	**MLR**	**Sub-group GEE**	**Adjusted GEE**	**MLR**
When x=50%						
Equal prevalence	1.50-9.07	7.12-22.65	1.63-8.55	2.30-8.23	3.45-36.36	2.24-8.40
Unequal prevalence	2.31-14.79	5.96-28.48	2.19-13.29	2.63-5.91	1.04-31.33	2.61-6.03
When x=20%						
Equal prevalence	0.11-30.05	5.18-45.16	0.38-20.61	0.37-2.54	8.57-21.04	0.53-2.53
Unequal prevalence	0.40-33.27	8.99-50.18	0.54-24.69	0.18-7.17	1.28-20.68	0.35-6.82

*Se: 20-30%; Sp: 20-30%; Se: sensitivity; Sp: specificity, GEE: generalized estimating equation.

Table 6. The absolute percent relative bias in estimating Se using subgroup GEE, adjusted GEE, and MLR

Disease prevalence > 20%	N=100			N=500		
	Sub-group GEE	**Adjusted GEE**	**MLR**	**Sub-group GEE**	**Adjusted GEE**	**MLR**
When x=50%						
Equal prevalence	2.68-5.60	1.35-26.82	2.64-5.81	0.72-2.58	6.14-16.65	0.69-2.62
Unequal prevalence	1.71-4.04	2.51-26.70	1.71-4.37	0.35-3.71	5.18-19.67	0.33-3.76
When x=20%						
Equal prevalence	0.70-8.91	4.45-19.64	0.64-8.31	0.11-9.09	4.01-27.92	0.09-9.14
Unequal prevalence	0.53-15.54	4.17-29.08	0.49-14.11	0.15-11.75	5.88-33.10	0.12-11.77

*Se: 20-30%; Sp: 20-30%; Se: sensitivity; Sp: specificity, GEE: generalized estimating equation.

Table 7. Models of sensitivity, specificity and predictive values of diagnosing acute appendicitis using subgroup GEE and MLR approaches

Diagnostic models	**Subgroup GEE**			**MLR**		
	RC	**SE**	**p-value**	**RC**	**SE**	**p-value**
Se						
Radiologist*	-1.994	0.543	0.000	-1.993	0.543	0.000
Age (years)	0.001	0.014	0.955	0.001	0.014	0.964
Male gender	-0.022	0.503	0.964	-0.021	0.505	0.967
WBC	0.028	0.059	0.634	0.028	0.060	0.636
Sp						
Radiologist*	1.997	0.965	0.038	1.948	0.848	0.022
Age (years)	0.009	0.045	0.838	0.015	0.046	0.741
Male gender	-0.470	1.746	0.788	-0.349	1.049	0.739
WBC	0.049	0.211	0.816	0.030	0.121	0.807
PPV						
Radiologist*	0.765	0.430	0.075	0.749	0.437	0.086
Age (years)	0.031	0.036	0.384	0.029	0.035	0.406
Male gender	-0.211	0.870	0.808	-0.166	0.894	0.852
WBC	0.196	0.057	0.001	0.192	0.057	0.001
NPV						
Radiologist*	-0.888	0.823	0.281	-0.795	0.789	0.314
Age (years)	-0.005	0.028	0.857	-0.013	0.038	0.732
Male gender	-0.413	0.932	0.658	-0.203	0.928	0.826
WBC	-0.137	0.121	0.258	-0.134	0.141	0.343

WBC: white blood cells;*referent: surgical residents; Se: sensitivity; Sp: specificity; PPV: positive predictive value; NPV: negative predictive value, GEE: generalized estimating equation.

Appendix 2.

Estimating diagnostic accuracy using a logistic regression (LR) model

Suppose a diagnostic study involves a screening test (T) and a reference test (D). LR models can be used to predict Se in relation to cofactors:

$\text{Logit}(P(T=1|D=1, X)=\alpha'_{1D}+\alpha'_{2D}*X_1+...+\alpha'_{kD}*X_k$ (1a: sub-group)

$\text{Logit}(P(T=1|D, X)=\alpha_{1D}+\alpha_{2D}*(D=1)+\alpha_{3D}*X_1+...+\alpha_{kD}*X_k$ (1b: adjusted)

The equation (1a) is referred to as a subgroup model and the equation (1b) is referred to as an adjusted model. Substituting D=0 in the above equations will provide models for 1-specifcity. Thus, LR models can also be used to predict Sp in the presence of cofactors:

$\text{Logit}(P(T=0|D=0, X)=\alpha'_{1\bar{D}}+\alpha'_{2\bar{D}}*X_1+...+\alpha'_{k\bar{D}}*X_k$ (2a: sub-group)

$\text{Logit}(P(T=0|D, X)=\alpha_{1\bar{D}}+\alpha_{2\bar{D}}*(D=0)+\alpha_{3\bar{D}}*X_1+...+\alpha_{k\bar{D}}*X_k$ (2b: adjusted)

Possible LR models for predicting PPV and NPV are:

$\text{Logit}(P(D=1|T=1, X)=\beta'_{1T}+\beta'_{2T}*X_1+...+\beta'_{kT}*X_k$ (3a: sub-group)

$\text{Logit}(P(D=1|T, X)=\beta_{1T}+\beta_{2T}*(T=1)+\beta_{3T}*X_1+...+\beta_{kT}*X_k$ (3b: adjusted)

$\text{Logit}(P(D=0|T=0, X)=\beta'_{1\bar{T}}+\beta'_{2\bar{T}}*X_1+...+\beta'_{k\bar{T}}*X_k$ (4a: sub-group)

$\text{Logit}(P(D=0|T, X)=\beta_{1\bar{T}}+\beta_{2\bar{T}}*(T=0)+\beta_{3\bar{T}}*X_1+...+\beta_{k\bar{T}}*X_k$ (4b: adjusted)

In the above equations (1, 2, & 3), α'_{1D},α_{1D} $\alpha'_{1\bar{D}}, \alpha_{1\bar{D}}$ $\beta'_{1T},\beta_{1T}, \beta'_{1\bar{T}}$,and $\beta_{1\bar{T}}$ are the intercepts while $\alpha'_{2D}..\ \alpha'_{kD},\alpha_{1D}...\alpha_{kD}$ $\alpha'_{2\bar{D}}....\alpha'_{k\bar{D}}, \alpha_{2\bar{D}}...\alpha_{k\bar{D}}$ $\beta'_{2T}...\beta'_{kT},\beta_{2T}...\beta_{kT}, \beta'_{2\bar{T}}...\beta'_{k\bar{T}}$,and $\beta_{2\bar{T}}..\beta_{k\bar{T}}$ are the regression coefficients and X (X_1….. X_k) is the vector of k covariates. D and $\bar{D}$ denote the presence and the absence of disease respectively while T denotes positive test result and $\bar{T}$ denotes negative test result.

Multinomial logistic regression (MLR) for estimating diagnostic accuracy

The MLR models for predicting a new outcome variable Y (1=true positive; 2=false positive; 3= false negative; 4=true negative):

$$\log\left[\frac{P(Y=1|X)}{P(Y=3|X)}\right]=\mu_1+\mu_2*X_1+...+\mu_k*X_k \qquad \text{Se model}$$

$$\log\left[\frac{P(Y=4|X)}{P(Y=2|X)}\right]=\pi_1+\pi_2*X_1+...\pi_k*X_k \qquad \text{Sp model} \qquad (5)$$

$$\log\left[\frac{P(Y=1|X)}{P(Y=2|X)}\right]=\rho_1+\rho_2*X_1+...\rho_k*X_k \qquad \text{PPV model}$$

$$\log\left[\frac{P(Y=4|X)}{P(Y=3|X)}\right]=\tau_1+\tau_2*X_1+...\tau_k*X_k \qquad \text{NPV model}$$

where μ_1 π_1, ρ_1, and τ_1 are the regression intercepts and μ_2…, μ_k, π_2,…, π_k, ρ_2,…, ρ_k and τ_2,…, τ_k are the regression coefficients and X (X_1….. X_k) is the vector of k covariates.

Comparing diagnostic accuracy using generalized estimating equation (GEE) and MLR methods

The MLR described in the above equation (5) can be extended for two screening tests as:

$$\log\left[\frac{P(Y=1|Z,X)}{P(Y=3|Z,X)}\right]=\mu_1+\mu_2*Z+\mu_3*X_1+...+\mu_k*X_k$$

$$\log\left[\frac{P(Y=4|Z,X)}{P(Y=2|Z,X)}\right]=\pi_1+\pi_2*Z+\pi_3*X_1+...\pi_k*X_k \qquad (6)$$

$$\log\left[\frac{P(Y=1|Z,X)}{P(Y=2|Z,X)}\right]=\rho_1+\rho_2*Z+\rho_3*X_1+...\rho_k*X_k$$

$$\log\left[\frac{P(Y=4|Z,X)}{P(Y=3|Z,X)}\right]=\tau_1+\tau_2*Z+\tau_3*X_1+...\tau_k*X_k$$

where μ_1, π_1, ρ_1, and τ_1 are the regression intercepts and μ_2…, μ_k, π_2,…, π_k, ρ_2,…, ρ_k and τ_2,…, τ_k are the regression coefficients and X (X_1….. X_k) is the vector of k covariates in equation (6). The μ_2 and π_2 provide the comparison of sensitivities and specificities between two screening tests respectively whereas ρ_2 and τ_2 provide the comparison of positive predictive values and negative predictive values between the two screening tests respectively.

STATISTICS IN TRANSITION new series, Summer 2015
Vol. 16, No. 2, pp. 223–242

FUNCTIONAL STRUCTURE OF POLISH REGIONS IN THE PERIOD 2004-2013 – MEASUREMENT VIA HHI INDEX, FLORENCE'S COEFFICIENT OF LOCALIZATION AND CLUSTER ANALYSIS

Marek Obrębalski[1], **Marek Walesiak**[2]

ABSTRACT

The article addresses the measurement and identification problems covering particular social and economic areas (referred to as functions) in the regions of the country, based on the employment structure analysis and assessment by the sectors of the economy. The Herfindahl-Hirschman index was applied to measure sectoral concentration and Florence's coefficient of localization to determine regional functional specialization. Finally, cluster analysis was conducted to produce the functional typology of regions.

Key words: regional economy, dominating functions, functional specialization, typology of regions.

1. Introduction

Economic base theory remains one of the most popular concepts explaining local and regional development (see Sokołowski, 2006, pp. 33-35; Markowski, 2008; Korenik and Zakrzewska-Półtorak, 2011, pp. 23-35). The socio-economic structure of each area is determined by a system which is both complex and complicated, and which covers social and economic fields of population activity influenced by past and present management status and natural conditions. These fields are referred to as functions of particular territorial units or settlement systems in different spatial scale.

Economic base theory allows for identifying two groups of functions, i.e. exogenous (directed outside a particular territorial unit) and endogenous (related to meeting the needs of the community of this unit). Therefore, it facilitates the identification of those functions which determine the development of particular

[1] Wroclaw University of Economics, Department of Regional Economics, Jelenia Góra. E-mail: marek.obrebalski@ue.wroc.pl.

[2] Wroclaw University of Economics, Department of Econometrics and Computer Science, Jelenia Góra. E-mail: marek.walesiak@ue.wroc.pl.

locations, cities or regions, since its basic assumption is to support the above-mentioned development by export-oriented (exogenous) activities. Hence, external demand for goods or services produced in a given territorial unit area (e.g. a region) is considered the most important incentive of its economic growth.

Both measurement and identification of functions are generally performed based on the employment structure analysis and assessment in accordance with local and regional economy fields (see Obrębalski, 1989, pp. 25-29). An economic base of a territorial unit is reflected by the quantitative proportions of employment in particular activity areas. Its precise and direct measurement remains, however, a complicated and laborious task. It would have to involve a detailed analysis of goods and services sales in terms of their volume and direction with reference to each entity running a business in the area of the studied territorial unit. Therefore, both in theory and practice, the identification and measurement of the economic base is commonly performed using indirect methods. One of them is the method called by R.B. Andrews the macrocosmic method (see Dziewoński, 1971, p. 49). It consists in the identification of the economic base size by comparing the employment structure in the analysed territorial unit against the general employment structure in a larger scale unit, e.g. a country. This method commonly applies two measures, namely Florence's local specialization coefficient (localization quotient) and Hoyt's employment surplus coefficient (Jerczyński, 1973, p. 38). This method is extensively applied, for instance, in functional specialization (see Dacko, 2009, pp. 25-34; Karmowska, 2011, pp. 85-93; Gwosdz, 2012, pp. 21-23) and in the economic base differentiation research (see Sokołowski, 2008, pp. 254-257).

In practice, numerous studies have been conducted regarding the coefficient of localization application to measure the functional specialization level of each region in a country. The specialization index was, among others, applied in the study (Angulo, Mur and Trivez, 2014) to separate sectors in which Spanish regions were specializing in 2010. The study covered 6 sectors of the economy and 47 regions (NUTS-3). The specialization analysis of 13 Greek regions (NUTS-2) in the system of three sectors of the economy in 2007 was performed in the study by (Christofakis and Gkouzos, 2013).

The cognitive and practical purpose of the this paper is to discuss the level and scope of the differentiation between functions with reference to particular regions (NUTS-2 – voivodships). The study of sectoral concentration, specialization and typology of Polish regions in the period 2004-2013 with application of the research method covering the combined application of cluster analysis and Herfindahl-Hirschman index is a pioneering one on Polish market. Identification and measurement of the functional structures of Polish regions in terms of the dynamics is important primarily because of its scope and direction of the socio-economic transformation, as well as the apparent dearth of current research and information in this regard. The results of the study will extend the information base for monitoring national regional policy and developmental policies of individual regions.

2. Sectoral structure of Polish economy

The research covering functional concentration and specialization of Polish regions will be conducted by sectors for the years 2004 and 2013. Due to the fact that Polish Classification of Activities (PCA) was changed in the period under analysis, Table 1 presents Polish economy sectoral structure in accordance with 2004 PCA and 2007 PCA.

Table 1. Polish economy sectoral structure in accordance with 2004 PCA and 2007 PCA

Sectors	2013		2004	
	Sections / name		Sections / name	
S_1. Agriculture	A	Agriculture, hunting, forestry and fishing	A	Agriculture, hunting and forestry
			B	Fishing
S_2. Industry and construction	B	Mining and quarrying	C	Mining and quarrying
	C	Manufacturing	D	Manufacturing
	D	Electricity, gas, steam and air conditioning supply	E	Electricity, gas and water supply
	E	Water supply; sewerage, waste management and remediation activities		
	F	Construction	F	Construction
Market services (S_3 and S_4)				
S_3. Logistic support of the population and companies	G	Trade; repair of motor vehicles	G	Trade and repair
	H	Transportation and storage	I	Transport, storage and communication
	J	Information and communication		
	I	Accommodation and catering	H	Hotels and restaurants
S_4. Entrepreneurship development support	K	Financial and insurance activities	J	Financial intermediation
	L	Real estate activities	K	Real estate, renting and business activities
	M	Professional, scientific and technical activities		
	N	Administrative and support service activities		

Table 1. Polish economy sectoral structure in accordance with 2004 PCA and 2007 PCA (cont.)

Sectors	2013		2004	
	Sections / name		Sections / name	
S_5. Non-market services	O	Public administration and defence; compulsory social security	L	Public administration and defence; compulsory social security
	P	Education	M	Education
	Q	Human health and social work activities	N	Health and social work
	R	Arts, entertainment and recreation	O	Other community, social and personal service activities
	S	Other service activities		
	T	Activities of households as employers and products-producing activities of households for own use	P	Private household with employed persons
	U	Extra-territorial organizations and bodies	Q	Extra-territorial organizations and bodies

Source: for 2013 – Regulation by the Council of Ministers regarding Polish Classification of Activities (PCA) (Journal of Laws from 2007 no. 251, item 1885 and from 2009 no. 59, item 489). For 2004 – Regulation by the Council of Ministers regarding Polish Classification of Activities (PCA) (Journal of Laws from 2004 no. 33, item 289).

PCA sections are grouped in 5 sectors: agriculture, industry and construction, logistic support of the population and companies, entrepreneurship development support and non-market services. The basis for determining market services of two separate sectors in the system was the similarity of types and scope of activities (see Obrębalski, 2012, p. 116).

3. Research methodology for functional structures of regions

The article presents the conducted research covering functional structures of regions referring to the following problems:

- determining the dominant functions of regions,
- identifying the functional specialization of regions,
- conducting the functional typology of regions.

In order to define the dominant functions of regions the percentage of the share employment by sectors of the economy was calculated. Herfindahl-

Hirschman index was applied to measure sectoral concentration (dominance) of regions (Herfindahl, 1950; Hirschman, 1964):

$$HHI_i = \sum_{j=1}^{m} b_j^2 , \tag{1}$$

where: $i = 0, 1, \ldots, n$ – object number (0 refers to a country whereas $1, \ldots, n = 16$ refers to the number of regions)

$b_j = \frac{Z_{ij}}{\sum_{j=1}^{m} Z_{ij}} \cdot 100\%$ – for regions,

$b_j = \frac{Z_{\bullet j}}{\sum_{j=1}^{m} Z_{\bullet j}} \cdot 100\%$ – for a country,

$j = 1, \ldots, m = 5$ – the number of the sector of the economy.

Herfindahl-Hirschman Index (*HHI*) is the most well-known measure of specialization and concentration constructed on the basis of structural data in economics (Calkins, 1983). In Polish literature specialization and concentration indices (with *HHI* index) are presented, among others, in the studies by Szyrmer (1975) and Kukuła (1976).

HHI_i index represented by (1) takes values form $\left[\frac{10{,}000}{m}; 10{,}000\right]$ interval. In the case of five sectors of the economy the index takes values from $[2{,}000; 10{,}000]$ interval. The higher the values from the bottom limit the higher the sectoral concentration in a particular region.

The coefficient of localization (also referred to as specialization ratio) introduced by P. Florence (Florence, 1939; Florence, 1944, p. 96), as presented below, was applied to identify and measure the specialized functions of regions:

$$S_{ij} = \frac{Z_{ij} \Big/ \sum_{j=1}^{m} Z_{ij}}{Z_{\bullet j} \Big/ \sum_{j=1}^{m} Z_{\bullet j}} , \tag{2}$$

where: S_{ij} – specialization coefficient of *i*-th territorial unit (region) in *j*-th sector of the economy,

Z_{ij} – employment in *j*-th sector in *i*-th territorial unit (region),

$Z_{\bullet j}$ – employment in *j*-th sector of the economy in a country,

$i = 1, \ldots, n = 16$ – the number of the region.

In Polish literature it is presented, among others, in the studies by (Jerczyński, 1971, p. 126; Kostrubiec, 1972, p. 25; Runge, 2007).

Florence's coefficient of localization measures the share of employment ratio in *j*-th region sector against the share of employment in *j*-th sector of a country. Values higher than one indicate greater share of employment in a region than in a country for a given sector. It means that a region specializes in a particular sector of the economy.

Cluster analysis was applied to conduct the functional typology of regions (see Walesiak, 2008; Walesiak, 2009). In order to identify the classes of similar regions, in terms of Florence's coefficient of localization values in 2004 and then in 2013, the following research procedure was applied:

- GDM1 distance was used for metric data to determine the distance matrix between regions in each year (see Walesiak, 2011, p. 39);
- hierarchical agglomeration method of the furthest neighbour was applied to divide 16 regions into relatively homogenous clusters. The results of cluster analysis were graphically presented by means of a dendrogram;
- Caliński-Harabasz index for quality assessment of classification results was adopted to determine the number of clusters into which the analysed 16 regions in 2004 and 2013 should be divided (see Walesiak, 2011, p. 61). Moreover, the identified divisions of the regions should remain stable. Replication analysis using *replication.Mod* function of *clusterSim* package was applied for the assessment of stability of the results of cluster analysis (see Walesiak and Dudek, 2015):
- adjusted Rand index was used to calculate agreement between two partitions of 16 regions for the years 2004 and 2013 (Hubert and Arabie, 1985),
- the profiles of the identified typological clusters were specified and the changes characteristic for the period 2004-2013 were assessed.

4. Dominant functions of regions

Each region is characterized by social, economic and spatial diversity. Table 2 presents information about functional diversification of regions in the years 2004 and 2013, identified based on employment structure by sectors.

In the period 2004-2013 the following multidirectional changes occurred in the employment sectoral structure in the national economy:

- the importance of the agricultural sector decreased (the share of employment in this sector field was reduced from 17.29% to 17.11% of the total employment in the national economy),
- the decreasing trend was also observed in the industry and construction sector (the share if this sector in the employment structure was reduced from 28.28% to 26.33%),
- the importance of logistics service for population and companies increased (its share went up from 23.53% to 24.34% of the total employment),
- the importance of the entrepreneurship development support sector went down (the share of employment in this sector decreased from 9.66% to 7.94%),
- the non-market services sector increased (the share of employment in this sector field went up from 21.25% to 24.27%).

Having analysed Herfindahl-Hirschman index values one should conclude that in the analysed period a slight increase in sectoral concentration in Poland was observed (*HHI* value increased from 2197 up to 2231).

Both in the entire country and in every of its regions the significant importance of the service-oriented activity identified according to fields is recognized (S_3, S_4 and S_5). In 2013, 56.6% of total employment was in the service sector. Among the service-oriented fields of population occupational activity the major role was played by commercial operations (15.3%), education, health care and social aid activity types.

The data referring to particular regions also confirm the dominating role of the broadly understood role of the service sector. In 2013 the highest level of the discussed dominance referred to the following regions: Mazowieckie (almost 68% of total employment), Zachodniopomorskie (63.6%), Pomorskie (62.8%) and Dolnośląskie (60.8%). On the other hand, the lowest level of dominance of the service function refers to such regions as: Podkarpackie (43.0%), Lubelskie (44.3%) and Świętokrzyskie (44.9%).

In relation to entities conducting activities in the fields covering logistics service of population and companies, the following regions were characterized by the highest share of employment in 2013: Mazowieckie (over 29.2% of total employment), Zachodniopomorskie (over 27.7%) and Pomorskie (almost 27.5%), whereas the lowest one – Podkarpackie (only 17.5%) and Lubelskie (17.7%).

On the other hand, entrepreneurship development support played a more significant role in the regional labour market structure of the following regions: Mazowieckie (almost 14% of total employment), while a relatively smaller one referred to Podkarpackie (less than 4%) and Świętokrzyskie regions (slightly more than 4%).

Table 2. Employment structure as well as concentration and specialization coefficients by Polish sectors and regions in the years 2004 and 2013

Specification	Total	S_1	S_2	S_3	S_4	S_5	*HHI*
2004							
POLAND	12413284	2145668	3509917	2920913	1198803	2637983	
%	100	17.29	28.28	23.53	9.66	21.25	2197
Dolnośląskie	875865	75070	280775	221000	95851	203169	2396
%	100	8.57	32.06	25.23	10.94	23.20	
S		0.4959	1.1337	1.0723	1.1332	1.0915	
Kujawsko-Pomorskie	640041	118161	189486	141792	52780	137822	2240
%	100	18.46	29.61	22.15	8.25	21.53	
S		1.0680	1.0470	0.9415	0.8539	1.0133	
Lubelskie	724950	278582	131564	125631	38092	151081	2568
%	100	38.43	18.15	17.33	5.25	20.84	
S		2.2232	0.6418	0.7365	0.5441	0.9807	
Lubuskie	282474	27580	87674	72063	25675	69482	2397
%	100	9.76	31.04	25.51	9.09	24.60	
S		0.5649	1.0977	1.0842	0.9412	1.1575	

Table 2. Employment structure as well as concentration and specialization coefficients by Polish sectors and regions in the years 2004 and 2013 (cont.)

Specification	Total	S_1	S_2	S_3	S_4	S_5	*HHI*
2004							
Łódzkie	887833	192391	261680	187647	72295	173820	2235
%	100	21.67	29.47	21.14	8.14	19.58	
S		1.2537	1.0424	0.8982	0.8432	0.9213	
Małopolskie	1011715	184121	271209	237231	92258	226896	2186
%	100	18.20	26.81	23.45	9.12	22.43	
S		1.0529	0.9481	0.9965	0.9442	1.0553	
Mazowieckie	2024968	320826	449008	534272	303658	417204	2088
%	100	15.84	22.17	26.38	15.00	20.60	
S		0.9166	0.7842	1.1213	1.5528	0.9695	
Opolskie	290772	50403	87799	63649	22366	66555	2274
%	100	17.33	30.20	21.89	7.69	22.89	
S		1.0028	1.0679	0.9303	0.7965	1.0771	
Podkarpackie	635569	158887	179289	121908	40238	135247	2282
%	100	25.00	28.21	19.18	6.33	21.28	
S		1.4463	0.9977	0.8152	0.6556	1.0013	
Podlaskie	388691	139540	74070	71839	23667	79575	2450
%	100	35.90	19.06	18.48	6.09	20.47	
S		2.0769	0.6740	0.7855	0.6305	0.9634	
Pomorskie	656222	62582	196192	176256	71111	150081	2347
%	100	9.54	29.90	26.86	10.84	22.87	
S		0.5517	1.0574	1.1415	1.1221	1.0762	
Śląskie	1491783	71369	565094	387078	148891	319351	2689
%	100	4.78	37.88	25.95	9.98	21.41	
S		0.2768	1.3397	1.1027	1.0335	1.0073	
Świętokrzyskie	429552	144126	95412	82407	25008	82599	2391
%	100	33.55	22.21	19.18	5.82	19.23	
S		1.9411	0.7856	0.8153	0.6028	0.9048	
Warmińsko-Mazurskie	386626	67343	110384	86668	29821	92410	2252
%	100	17.42	28.55	22.42	7.71	23.90	
S		1.0077	1.0097	0.9527	0.7987	1.1247	
Wielkopolskie	1209924	210057	398498	274746	110424	216199	2304
%	100	17.36	32.94	22.71	9.13	17.87	
S		1.0044	1.1648	0.9650	0.9450	0.8408	
Zachodniopomorskie	476299	44630	131783	136726	46668	116492	2372
%	100	9.37	27.67	28.71	9.80	24.46	
S		0.5421	0.9785	1.2199	1.0146	1.1509	
2013							
POLAND	13919826	2382129	3665103	3388065	1105776	3378753	2231
%	100	17.11	26.33	24.34	7.94	24.27	
Dolnośląskie	1018172	88433	310822	256211	89768	272938	2437
%	100	8.69	30.53	25.16	8.82	26.81	
S		0.5075	1.1594	1.0339	1.1099	1.1044	
Kujawsko-Pomorskie	676971	107287	195271	157955	46312	170146	2306
%	100	15.85	28.84	23.33	6.84	25.13	
S		0.9261	1.0955	0.9586	0.8612	1.0354	

Table 2. Employment structure as well as concentration and specialization coefficients by Polish sectors and regions in the years 2004 and 2013 (cont.)

Specification	Total	S_1	S_2	S_3	S_4	S_5	*HHI*
2013							
Lubelskie	799820	307911	137488	141646	36980	175795	2596
%	100	38.50	17.19	17.71	4.62	21.98	
S		2.2496	0.6529	0.7276	0.5820	0.9055	
Lubuskie	320293	36780	99339	81211	18871	84092	2461
%	100	11.48	31.02	25.36	5.89	26.25	
S		0.6710	1.1779	1.0417	0.7417	1.0816	
Łódzkie	925303	179190	253262	212338	60387	220126	2259
%	100	19.37	27.37	22.95	6.53	23.79	
S		1.1316	1.0395	0.9428	0.8215	0.9801	
Małopolskie	1259992	272715	295212	302983	95830	293252	2195
%	100	21.64	23.43	24.05	7.61	23.27	
S		1.2648	0.8898	0.9879	0.9574	0.9588	
Mazowieckie	2274610	301358	429915	664813	317861	560663	2190
%	100	13.25	18.90	29.23	13.97	24.65	
S		0.7742	0.7178	1.2008	1.7591	1.0155	
Opolskie	311442	50536	96450	64968	17597	81891	2381
%	100	16.23	30.97	20.86	5.65	26.29	
S		0.9482	1.1762	0.8570	0.7113	1.0833	
Podkarpackie	792771	259686	192221	138789	31316	170759	2447
%	100	32.76	24.25	17.51	3.95	21.54	
S		1.9141	0.9209	0.7193	0.4973	0.8874	
Podlaskie	400090	126790	78881	78580	19396	96443	2383
%	100	31.69	19.72	19.64	4.85	24.11	
S		1.8518	0.7488	0.8069	0.6103	0.9931	
Pomorskie	753429	66394	213948	207036	68362	197689	2410
%	100	8.81	28.40	27.48	9.07	26.24	
S		0.5149	1.0785	1.1290	1.1422	1.0810	
Śląskie	1638657	101963	586968	419282	129360	401084	2638
%	100	6.22	35.82	25.59	7.89	24.48	
S		0.3636	1.3604	1.0512	0.9938	1.0084	
Świętokrzyskie	453970	149635	100598	84001	18424	101312	2434
%	100	32.96	22.16	18.50	4.06	22.32	
S		1.9261	0.8416	0.7602	0.5109	0.9194	
Warmińsko-Mazurskie	419637	70022	118921	89792	24269	116633	2345
%	100	16.69	28.34	21.40	5.78	27.79	
S		0.9751	1.0763	0.8791	0.7280	1.1451	
Wielkopolskie	1367192	213618	420864	347679	94414	290617	2338
%	100	15.62	30.78	25.43	6.91	21.26	
S		0.9130	1.1691	1.0448	0.8693	0.8757	
Zachodniopomorskie	507477	49811	134943	140781	36629	145313	2445
%	100	9.82	26.59	27.74	7.22	28.63	
S		0.5736	1.0099	1.1398	0.9086	1.1797	

S – Florence's coefficient of localization presented as (2).

Source: authors' compilation based on: Pracujący w gospodarce narodowej w 2013 r. [Employment in national economy in 2013] Central Statistical Office, Warsaw 2014, pp. 40-47; Pracujący w gospodarce narodowej w 2004 r. [Employment in national economy in 2004] Central Statistical Office, Warsaw 2005, pp. 34-39.

The fields of non-market services were characterized by their relatively high importance in the employment structure in two regions: Zachodniopomorskie (over 28.6% of total employment) and Warmińsko-Mazurskie (almost 27.8%).

Industry and construction played a significant role in the following regions: Śląskie (over 35.8% of total employment), Lubuskie and Opolskie (31.0% each), Wielkopolskie (30.8%) and Dolnośląskie (30.5%).

Agricultural function is recognized as crucial in regional economy of Lubelskie (38.5% of total employment), Świętokrzyskie (almost 33%), Podkarpackie (32.8%) and Podlaskie (nearly 31.7%).

Following the analysis of Herfindahl-Hirschman index values it should be observed that:

- the highest *HHI* values were recorded for Śląskie region (industry and construction dominate) and Lubelskie region (agricultural function remains the dominant one), whereas the lowest value was true for Mazowieckie region,
- in the analysed period the majority of regions were characterized by higher level of sectoral concentration. In the case of Podlaskie and Śląskie regions only the decrease in *HHI* index values was observed.

5. Functional specialization of regions

The rank of particular regions, in a broader spatial system (e.g. a country), is determined by the so-called specialized functions. The functions are represented by the social and economic activity sectors, the importance of which in the analysed territorial unit is larger than the one typical for its environment.

Specialization levels of *i*-th territorial unit (region) in *j*-th economic sector are defined in the article as follows:

$S_{ij} \leq 1$	no specialization (endogenous function),
$1 < S_{ij} \leq 1.2$	very low level of specialization,
$1.2 < S_{ij} \leq 1.5$	low level of specialization,
$1.5 < S_{ij} \leq 2.0$	medium level of specialization,
$S_{ij} > 2.0$	high level of specialization.

The levels of functional specialization characteristic for particular regions in the country in the years 2004 and 2013 are presented in Table 3.

Table 3. Functional specialization of regions in 2004 and 2013

Sectors of the economy		Functional specialization level in regions		
		high	medium	low
S_1. Agriculture	2004	Lubelskie, Podlaskie	Świętokrzyskie	Podkarpackie, Łódzkie
	2013	Lubelskie	Świętokrzyskie, Podkarpackie, Podlaskie	Małopolskie
S_2. Industry and construction	2004	–	–	Śląskie
	2013	–	–	Śląskie
S_3. Logistic support of the population and companies	2004	–	–	Zachodnio-pomorskie
	2013	–	–	Mazowieckie
S_4. Entrepreneurship development support	2004	–	Mazowieckie	–
	2013	–	Mazowieckie	–
S_5. Non-market services	2004	–	–	–
	2013	–	–	–

Source: authors' compilation.

The analysed economic activity sectors are characterized by the diversified specialization level in the regions of the country.

In 2013 the agricultural sector determined a high functional specialization of Lubelskie region. In 2004 this specialization level in these fields was also recorded in Podlaskie region.

The number of regions characterized by a medium specialization level in the agricultural sector fields saw an increase. In 2004 this level was recorded in Świętokrzyskie region only, while in 2013 this group covered also Podkarpackie and Podlaskie regions. On the other hand, a low level of functional specialization in agriculture in 2013 referred to Małopolskie, whereas in 2004 this group included Podkarpackie and Łódzkie regions.

With reference to functional specialization in the fields of industry and construction the only region with a low specialization level was Śląskie region.

In relation to logistics service of population and companies Mazowieckie region showed a low level of functional specialization. Moreover, Mazowieckie region also showed a medium specialization level in the fields of entrepreneurship development support.

As far as the non-market services are concerned none of the regions under analysis revealed any specialization. It is substantively justified since the non-market services sector remains crucial in reflecting spatial distribution of population since it primarily covers the infrastructure fields focused on meeting the widely felt social needs by local and regional communities in each of the regions (e.g. in terms of education, health care, social aid, culture and recreation).

The sectoral perspective provides the general dimension of the functional structure and specialization in particular regions. However, a more detailed analysis of PCA sections system allows for presenting the field-oriented specialization and therefore:

- a high specialization level was recorded in the following regions: Lubelskie (section A: agriculture, forestry, hunting and fishing – S_{ij} = 2.2496) and Mazowieckie (section J: information and communication – S_{ij} = 2.0749),
- a medium specialization level referred to such regions as: Dolnośląskie (section N: administrative and support service activities – S_{ij} = 1.5507), Mazowieckie (section K: financial and insurance activities – S_{ij} = 1.8539; section M: professional, scientific and technical activities – S_{ij} = 1.8014), Podkarpackie (section A: agriculture, forestry, hunting and fishing – S_{ij} = 1.9141), Podlaskie (section A – S_{ij} = 1.8518), Świętokrzyskie (Section A – S_{ij} = 1.9261) and Zachodniopomorskie (section I: accommodation and catering – S_{ij} = 1.7828).

A clear functional specialization was observed not only in the agricultural sector fields, but also in some fields of market services. It mainly referred to Mazowieckie and Dolnośląskie regions, whereas tourism was recorded as a medium specialization level in Zachodniopomorskie region. It is facilitated not only by attractive natural conditions, but also by extensive tourism-oriented investments used in both summer and winter seasons. A relatively low level of this specialization refers to the following regions: Małopolskie, Pomorskie, Dolnośląskie, Mazowieckie, Warmińsko-Mazurskie and Śląskie.

6. Functional typology of regions

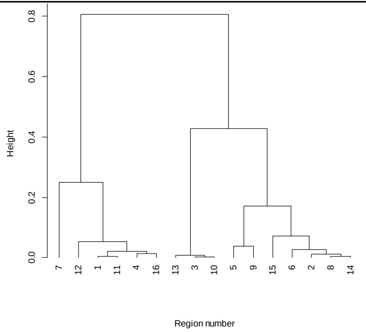
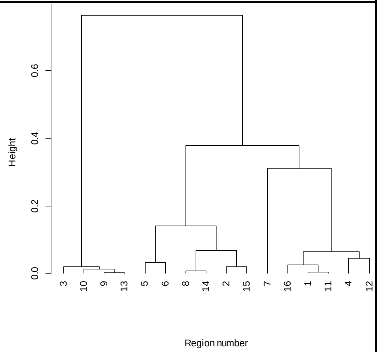
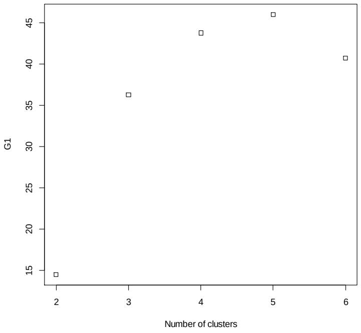
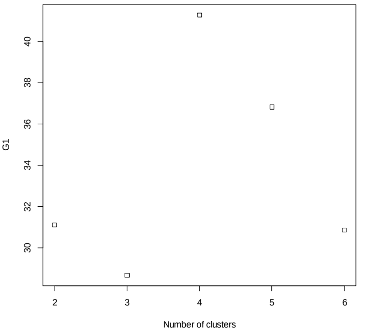
Cluster analysis was applied in conducting the functional typology of regions. Based on the data presented in Table 2 and following the procedure described in point 3 the clusters of regions similar in terms of Florence's coefficient of localization were determined for the years 2004 and 2013. The results of cluster analysis are presented in Table 4.

Table 4. Functional typology of regions in terms of Florence's coefficients of localization values in the years 2004 and 2013

Specification	2004	2013
The results of the division of a set of regions into clusters by applying the furthest neighbour method		

Cluster 1	(1) Dolnośląskie; (4) Lubuskie;
	(11) Pomorskie; (12) Śląskie;
	(16) Zachodniopomorskie
		(1) Dolnośląskie; (4) Lubuskie;
		(11) Pomorskie; (12) Śląskie;
		(16) Zachodniopomorskie

Cluster 2	(2) Kujawsko-Pomorskie;
	(6) Małopolskie; (8) Opolskie;
	(14) Warmińsko-Mazurskie;
	(15) Wielkopolskie
		(2) Kujawsko-Pomorskie;
		(5) Łódzkie;
		(6) Małopolskie; (8) Opolskie;
		(14) Warmińsko-Mazurskie;
		(15) Wielkopolskie

Cluster 3	(3) Lubelskie; (10) Podlaskie; (13) Świętokrzyskie	(3) Lubelskie; (9) Podkarpackie; (10) Podlaskie; (13) Świętokrzyskie
Cluster 4	(5) Łódzkie; (9) Podkarpackie	(7) Mazowieckie
Cluster 5	(7) Mazowieckie	–
Dendrogram		
Graphic interpretation of G1 Caliński-Harabasz index. Criterion of u clusters number selection: $\hat{u} = \underset{u}{\arg\max}\{G1(u)\}$		
Results of replication analysis	0.5212	0.6513
Agreement between two partitions	0.7887	

Source: authors' compilation using ***R*** *(R Development Core Team, 2015).*

The maximum value of Caliński-Harabasz index was obtained following the division into 5 classes (for 2004) and the division into 4 classes (for 2013). Replication analysis was conducted to assess the stability of the obtained cluster division into classes. The purpose of replication analysis is the stability assessment of the conducted classification covering the set of objects. The stability assessment was performed based on the adjusted Rand index value from $[-\infty;1]$ interval. The values obtained as a result of replication analysis for the year 2004 and 2013 confirmed a relatively stable division of regions into classes.

In order to facilitate the obtained results the interpretation medians from Florence's coefficient were specified for each class regarding 5 sectors of the economy:

[1] Medians (2004)

[.1] [.2] [.3] [.4] [.5]

[1.] 0.5421 1.09770 **1.1027** 1.0335 1.0915

[2.] 1.0077 1.04700 0.9527 0.8539 1.0553

[3.] **2.0769** 0.67400 0.7855 0.6028 0.9634

[4.] **1.3500** 1.02005 0.8567 0.7494 0.9613

[5.] 0.9166 0.78420 **1.1213** **1.5528** 0.9695

[1] Medians (2013)

[.1] [.2] [.3] [.4] [.5]

[1.] 0.51490 **1.1594** 1.0512 0.99380 1.08160

[2.] 0.96165 1.0859 0.9507 0.84135 1.00775

[3.] **1.92010** 0.7952 0.7439 0.54645 0.91245

[4.] 0.77420 0.7178 **1.2008** **1.75910** 1.01550

The specialization ratios over 1.10 were marked in bold.

0.7887 value of adjusted Rand index confirms high consistency of the obtained divisions of regional clusters into classes in the years 2004 and 2013. In the analysed period class 4 regions from 2004 moved to class 2 (Łódzkie region) and class 3 (Podkarpackie region). Łódzkie region recorded a significant reduction in specialization level with reference to S_1 sector (agriculture), whereas Podkarpackie region an extensive strengthening of specialization in this area.

Based on the obtained results the following conclusions can be put forward:

- class 3 regions (both in 2004 and in 2013) shows a clear specialization in S_1 sector (agriculture);
- one-element class covering Mazowieckie region (class 5 in 2004 and class 4 in 2013) specializes primarily in S_4 sector (entrepreneurship development support) and highly in S_3 sector (logistic support of the population and companies); in the analysed period the specialization ratio values increased significantly;
- in the case of class 2 regions (both in 2004 and in 2013) the absence of sectoral specialization was observed;
- for class 1 regions a low level of specialization was recorded in S_3 sector in 2004 and in S_2 sector in 2013.

Therefore, having conducted the typology of regions by sectoral specialization level and scope in 2013 the following regions can be determined:

- industry and service-oriented regions (class I: Dolnośląskie, Lubuskie, Pomorskie, Śląskie, Zachodniopomorskie);
- non-specialized regions (class II: Kujawsko-Pomorskie, Łódzkie, Małopolskie, Opolskie, Warmińsko-Mazurskie, Wielkopolskie);
- agricultural regions (class III: Lubelskie, Podkarpackie, Podlaskie, Świętokrzyskie);
- capital region (class IV: Mazowieckie) characterized by market services specialization; this region's individuality in the presented typology results from the developed central service-oriented functions in Warsaw, provided not only for its regional environment (see Obrębalski, 2014, p. 121).

The presented typology confirms limited composition variance of the analysed regional groups by sectoral specialization in the period under analysis. It does not, however, mean that in terms of particular social and economic activity areas within the framework of the identified sectors a relative stability of specialization level was also observed. Functional specialization factors result from many diversified local and regional determinants of demographic and social, natural and cultural, economic, institutional and spatial nature.

7. Final remarks and policy implications

In general, particular regions show a significant polyfunctionality, although each of them is characterized by a dominant function. In every region of the country it takes the form of a service function diversified by fields, but in many regions the significant role is also played by an industrial and agricultural function. The studied regions, however, show distinct functional specialization (in terms of field and level). It is at a high level in the agricultural sector for Lubelskie region only. Apart from agriculture, a medium specialization level is recorded also in the entrepreneurship development support sector, whereas a low level – in industry and construction sectors as well as logistics service of population and companies.

Mazowieckie region, with the dominant Warsaw, is characterized by a high specialization in market services. This region was identified as a result of the conducted typology as one of functional specialization types. This typology also allowed for separating the group of agricultural, industry and service-oriented and also non-specialized regions.

It should be observed, however, that despite many common typological characteristics, each region has individual and diversified potential, regional identity and the level of economic competitiveness. In the context of the national strategy of regional development this will concern the future development of the individual regions and the country (see Krajowa strategia …, 2010).

REFERENCES

ANGULO, A., MUR, J., TRIVEZ, J., (2014). Measure of the resilience to Spanish economic crisis: the role of specialization, Economics and Business Letters, 3(4), 263−275.

CALKINS, S., (1983). The new merger guidelines and the Herfindahl-Hirschman index, California Law Review, Vol. 71, Issue 2, 402−429.

CHRISTOFAKIS, M., GKOUZOS, A., (2013). Regional specialisation and efficiency of the agricultural sector in Greece: the relationship with regional funding allocation, Regional and Sectoral Economic Studies, Vol. 13-1, 119−130.

DACKO, M., (2009). Badanie stanu bazy ekonomicznej i struktury funkcjonalnej gmin województwa małopolskiego metodami pośrednimi [The analysis of the condition of the economic base and the functional structure of the Małopolskie voivodeship gminas using indirect methods], Folia Pomeranae Universitatis Technologiae Stetinensis. Seria Oeconomica, No. 268 (54), 25−34.

DZIEWOŃSKI, K., (1971). Studium rozwoju pojęć, metod i ich zastosowań [A study of the development of concepts, methods and applications], In: Baza ekonomiczna i struktura funkcjonalna miast [Urban economic base and functional structure of cities], Prace Geograficzne IG PAN, No. 87, Warszawa: PWN, 9−110.

FLORENCE, P., (1939). Report of the location of industry, London: Political and Economic Planning, UK.

FLORENCE, P., (1944). The selection of industries suitable for dispersion into rural areas, Journal of the Royal Statistical Society, Vol. 107, No. 2, 93−116.

GWOSDZ, K., (2012). Baza ekonomiczna i specjalizacja funkcjonalna miast konurbacji katowickiej po dwu dekadach transformacji [The economic base and functional specialisation of Katowice Conurbation’s towns and cities after two decades of the transition], Acta Geographica Silesiana, T. 11, 15−29.

HERFINDAHL, O. C., (1950). Concentration in the steel industry. Doctoral thesis, Columbia University.

HIRSCHMAN, A. O., (1964). The paternity of an index, The American Economic Review, Vol. 54, 761−762.

HUBERT, L., ARABIE, P., (1985). Comparing partitions, Journal of Classification, No. 1, 193−218.

JERCZYŃSKI, M., (1971). Metody pośrednie identyfikacji i pomiaru bazy ekonomicznej miast [Indirect methods of identification and measurement], W: Baza ekonomiczna i struktura funkcjonalna miast [Urban economic base and functional structure of cities], Prace Geograficzne IG PAN, No. 87, Wrocław: Zakład Narodowy im. Ossolińskich, 111−142.

JERCZYŃSKI, M., (1973). Zagadnienia specjalizacji bazy ekonomicznej większych miast w Polsce [Problems of specialization of the urban economic base of major cities in Poland], In: Dziewoński K. (ed.), Studia nad strukturą funkcjonalną miast [Studies of functional structure of towns], Prace Geograficzne IG PAN, No. 97, Wrocław: Zakład Narodowy im. Ossolińskich, 9−134.

KARMOWSKA, G., (2011). Badanie i pomiar rozwoju regionalnego na przykładzie województwa zachodniopomorskiego [Research and the measurement of regional development on the example of Zachodniopomorskie province], Roczniki Nauk Rolniczych, Seria G, Tom 98, z. 2, 85−93.

KORENIK, S., ZAKRZEWSKA-PÓŁTORAK, A., (2011). Teorie rozwoju regionalnego – ujęcie dynamiczne [Theories of regional development – dynamic approach], Wrocław: Wydawnictwo Uniwersytetu Ekonomicznego.

KOSTRUBIEC, B., (1972). Analiza zjawisk koncentracji w sieci osadniczej. Problemy metodyczne [Analysis of concentration phenomena in settlement network. Methodical issues], Prace Geograficzne IG PAN, No. 93, Wrocław: Zakład Narodowy im. Ossolińskich.

KRAJOWA STRATEGIA ROZWOJU REGIONALNEGO 2010-2020. REGIONY – MIASTA – OBSZARY WIEJSKIE [NATIONAL STRATEGY FOR REGIONAL DEVELOPMENT 2010-2020. REGIONS – CITIES – RURAL AREAS], (2010). Ministerstwo Rozwoju Regionalnego, Warszawa, https://www.mir.gov.pl/aktualnosci/polityka_rozwoju/Documents/KSRR_13_07_2010.pdf (access, 8.05.2015).

KUKUŁA, K., (1976). Kilka uwag o związkach między wskaźnikami specjalizacji i koncentracji przestrzennej skonstruowanymi na podstawie danych strukturalnych [Some remarks on relations between specialization indices and spatial concentration constructed on the basis of structural data], Przegląd Geograficzny, T. XLVIII, z. 3, 417−429.

MARKOWSKI, T., (2008). Teoretyczne podstawy rozwoju lokalnego i regionalnego [The theoretical basis of local and regional development]. In: Z. Strzelecki (ed.), Gospodarka regionalna i lokalna [Regional and local economy], Warszawa: PWN.

OBRĘBALSKI, M., (1989). Pomiar i identyfikacja wyspecjalizowanych funkcji usługowych regionu jeleniogórskiego [Measurement and identification of specialized service functions of Jelenia Góra region], In: Usługi w strukturze gospodarczej regionu [Services in economic structure of region], Prace Naukowe Akademii Ekonomicznej we Wrocławiu, no. 472, 25−29.

OBRĘBALSKI, M., (2012). Specyfika i zróżnicowanie rynku pracy w przygranicznych regionach Polski, północnych Czech i Niemiec [The specificity and the differentiation of labour market in border regions of Poland, Northern Czech and Germany], Gospodarka i Finanse, z. 2 „Zarządzanie kryzysowe jako element polityki społeczno-ekonomicznej” [Crisis management as element of socio-economic policy], 113−127.

OBRĘBALSKI, M., (2014). Centralność miast wojewódzkich w Polsce w zakresie usług FIRE w latach 2005-2012 – identyfikacja, pomiar i ocena [The centrality of voivodship cities in Poland in the scope of FIRE services in the years 2005-2012 – identification, measurement and evaluation], Gospodarka i Finanse, z. 4 „Rozwój lokalny i regionalny” [Local and regional development], 113−124.

Pracujący w gospodarce narodowej w 2004 r. [Employment in national economy in 2004], (2005). Warszawa: GUS [Central Statistical Office].

Pracujący w gospodarce narodowej w 2013 r. [Employment in national economy in 2013], (2014). Warszawa: GUS [Central Statistical Office].

R Development Core Team, (2015). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, URL http://www.R-project.org.

RUNGE, J., (2007). Metody badań w geografii społeczno-ekonomicznej – elementy metodologii. Wybrane narzędzia badawcze [Methods of research in socio-economic geography – the elements of the methodology. Selected research tools], Katowice: Wydawnictwo Uniwersytetu Śląskiego.

SOKOŁOWSKI, D., (2006). Funkcje centralne i hierarchia funkcjonalna miast w Polsce [Central functions and functional hierarchy of cities in Poland], Wydawnictwo Uniwersytetu Mikołaja Kopernika, Toruń.

SOKOŁOWSKI, D., (2008). Baza ekonomiczna większych miast w Polsce w okresie transformacji systemowej [The economic base of the largest cities in Poland in the period of systemic transformation], Przegląd Geograficzny, T. 80, z. 2, 245−266.

SZYRMER, J., (1975). Stopień specjalizacji rolnictwa [The degree of agricultural specialization], Przegląd Geograficzny, T. XLVII, z. 1, 117−135.

WALESIAK, M., (2008). Procedura analizy skupień z wykorzystaniem programu komputerowego clusterSim i środowiska R [Cluster analysis procedure with clusterSim computer programme and R environment], Taksonomia 15, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 7 (1207), 44−56.

WALESIAK, M., (2009). Analiza skupień [Cluster analysis], In: M. Walesiak, E. Gatnar (eds.), Statystyczna analiza danych z wykorzystaniem programu R [Statistical data analysis with R], Warszawa: Wydawnictwo Naukowe PWN, 407–433.

WALESIAK, M., (2011). Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R [The Generalized distance measure GDM in multivariate statistical analysis with R], Wrocław: Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu.

WALESIAK, M., DUDEK, A., (2015). clusterSim package, URL http://www.R-project.org.

STATISTICS IN TRANSITION new series, Summer 2015
Vol. 16, No. 2, pp. 243–264

USING SYMBOLIC DATA IN GRAVITY MODEL OF POPULATION MIGRATION TO REDUCE MODIFIABLE AREAL UNIT PROBLEM (MAUP)

Justyna Wilk[1]

ABSTRACT

Spatial analyses suffer from modifiable areal unit problem (MAUP). This occurs while operating on aggregated data determined for high-level territorial units, e.g. official statistics for countries. Generalization process deprives the data of variation. Carrying out research excluding territorial distribution of a phenomenon affects the analysis results and reduces their reliability. The paper proposes to use symbolic data analysis (SDA) to reduce MAUP. SDA proposes an alternative form of individual data aggregation and deals with multivariate analysis of interval-valued, multi-valued and histogram data.

The paper discusses the scale effect of MAUP which occurs in a gravity model of population migrations and shows how SDA can deal with this problem. Symbolic interval-valued data was used to determine the economic distance between regions which served as a separation function in the model. The proposed approach revealed that economic disparities in Poland are lower than official statistics show but they are still one of the most important factors of domestic migration flows.

Key words: modifiable areal unit problem (MAUP), symbolic data analysis (SDA), gravity model, population migration, economic distance.

1. **Introduction**

A large number of spatial analyses suffer from modifiable areal unit problem (MAUP), regardless of the research field (e.g. economics, biology, sociology, finance, medicine, etc.) (see Openshaw, 1984; Arbia, 1989, pp. 7-21; King, Tanner and Rosen (Eds.), 2004; Wong 2009). MAUP occurs while operating on aggregated data which is a procedure frequently used to describe higher-level territorial units, e.g. countries, metropolitan areas, regional labour markets, etc. Generalization process deprives the data of variation. Carrying out research excluding territorial

[1] Wrocław University of Economics, Department of Econometrics and Computer Science, 58-500 Jelenia Góra (Poland), Nowowiejska 3 Street. E-mail: Justyna.Wilk@ue.wroc.pl.

distribution and spatial features of a phenomenon affects the analysis results and reduces their reliability.

This problem is mostly seen in socio-economic studies in which a territorial unit is a result of an administrative division or territorial division for statistical purposes (see, e.g. NUTS classification). For example, Poland is administratively divided into 2479 municipalities (LAU 2 units). Each of them is located in the territory of one of 314 districts (LAU 1 units). A set of bordered districts is assigned to one of 16 provinces (NUTS 2). Official (economic, social, environmental, demographical etc.) statistics present aggregated values which generalize the situations of territorial units. They do not show the ranges, densities, distributions, outlier values or spatial variation in data. Then, we cannot infer results from one scale to the other scales of territorial division due to ecological fallacy. The paper deals with MAUP which occurs while modelling of population migrations.

In the era of market economy, domestic population migrations represent an integral part underlying the functioning of societies and economies. They regulate the size and structure of human resources, as well as job market situation, the consumption of goods and services, etc. Thus, an integral part of developing the policy of sustainable regional development is carrying out research studies regarding not only the results of migration flows (e.g. an amount of inflow) but, first of all, the conditions and causes of people's decisions why and where to migrate (see Bunea, 2012; Lucas, 1997; White and Lindstrom, 2006).

The intensity of domestic migration flows is strongly determined by the macroeconomic trends which affect people's propensity to move. But the directions of migrations depend on regional factors such as an economic, social, political, environmental situation, etc., as well as spatial and relational factors, e.g. ethnic differences (Van der Gaag (Ed.), 2003, pp. 1-141). These factors can be examined using an econometric gravity model.

The aim of the empirical study is to examine the determinants of domestic migrations in Poland. The research covers migration flows between 16 Polish NUTS 2 units (provinces) in the years 2011-2013, in which the world economy was overcoming the economic crisis. In terms of relatively stable political and cultural terms, the strongest determinants of migration processes are economic motives, e.g. improving the standard of living (Todaro, 1980; Lucas, 1997; Kupiszewski, Durham and Rees, 1999; Holzer, 2003; White, Lindstrom, 2006; Ghatak, Mulhern and Watson, 2008). Cohesion policy of the European Union directs national policies of regional development to convergence processes. Thus, in this study, the crucial issue is to identify the economic disparities in Poland and examine their impact on domestic migration flows.

The preliminary data analysis showed the occurrence of the scale effect of MAUP. Therefore the objective of this paper is to propose a solution to MAUP. The proposed approach employs symbolic data analysis (SDA) to construct the gravity model of population migration. SDA covers multivariate analysis of interval-valued, multi-valued, modal and dependent data. It is a support to manage data structure and reduction problems (see Bock and Diday (Eds.), 2000; Billard and Diday, 2006; Diday and Noirhomme-Fraiture (Ed.), 2008; Gatnar and Walesiak, 2011).

The first section of the paper discusses the essence of the modifiable areal unit problem. The second part concerns the scale effect of MAUP occurring in the gravity model of population migration. The third section employs symbolic data analysis to reduce this problem. The fourth part discusses the results of the study and shows the influence of MAUP on the spatial interaction analysis results.

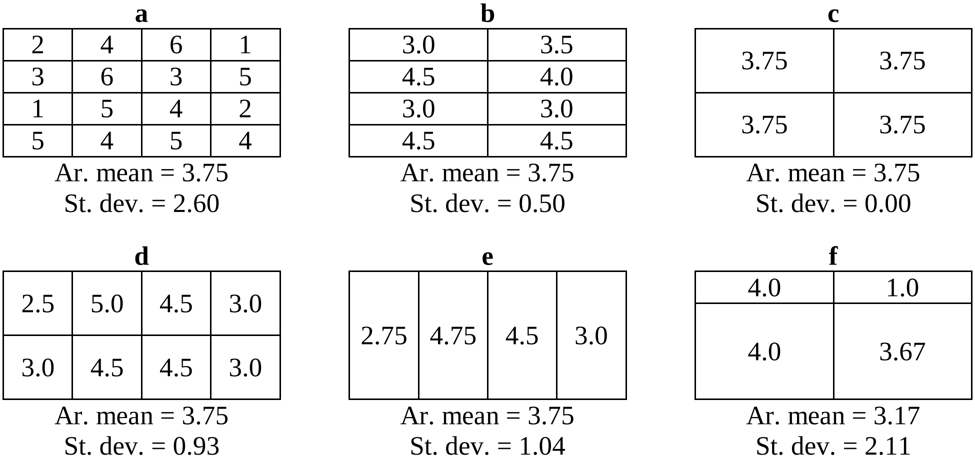
2. Modifiable areal unit problem (MAUP) of spatial data analyses

Yule and Kendall (1950) introduced a fundamental distinction between two different kinds of analysed units: the non-modifiable and modifiable units. Modifiable units differ from non-modifiable units because they can be further decomposed into smaller units and, moreover, this decomposition can be done in a few ways. The relevance of this distinction is that the value of any statistical measure "will, in general, depend on the unit chosen if that unit is modifiable" (Yule and Kendall, 1950).

This problem is known in the literature as the modifiable areal unit problem (MAUP). MAUP results from data generalization and multiscaling of spatial phenomena (see Openshaw, 1984; Arbia, 1989, pp. 7-21; Anselin, 1988, p. 26-28; Suchecka (Ed.), 2014, pp. 56-60; Gotway, Crawford and Young, 2004; Wong, 2009). The problem arises from the fact that areal units are usually arbitrarily determined and modifiable in the sense that they can be aggregated to form units of different sizes or spatial arrangements (Jelinski and Wu, 1996, p. 130).

Openshaw and Taylor (1979) distinguished two aspects of MAUP: the scale effect and zonation effect. The scale (aggregation) aspect refers to different results which can be achieved in statistical analysis with the same set of data grouped at different scale levels (e.g. countries or regions). Thus, the scale effect occurs if a set of areas is considered from the point of view of larger areal units, with each combination leading to different data values and inferences. The problem is "the variation in results that may be obtained when the same data are combined into sets of increasingly larger areal units of analysis" (Openshaw and Taylor, 1979).

The scale effect of MAUP results from a few of reasons, e.g. human society is organized in territorial units usually arranged into nested hierarchies, e.g. town, regions, states, countries (see Moellering and Tobler, 1972; Cliff and Ord, 1981, p. 133). The scale effect of MAUP was proved by Gehlke and Biehl, 1934, pp. 169-170; Jelinski and Wu, 1996; Dark and Bram, 2007, pp. 471-479. Parts a-c of Figure 1 illustrate the scale effect of MAUP.



a

2	4	6	1
3	6	3	5
1	5	4	2
5	4	5	4

Ar. mean = 3.75
St. dev. = 2.60

b

3.0	3.5
4.5	4.0
3.0	3.0
4.5	4.5

Ar. mean = 3.75
St. dev. = 0.50

c

3.75	3.75
3.75	3.75

Ar. mean = 3.75
St. dev. = 0.00

d

2.5	5.0	4.5	3.0
3.0	4.5	4.5	3.0

Ar. mean = 3.75
St. dev. = 0.93

e

2.75	4.75	4.5	3.0

Ar. mean = 3.75
St. dev. = 1.04

f

4.0	1.0
4.0	3.67

Ar. mean = 3.17
St. dev. = 2.11

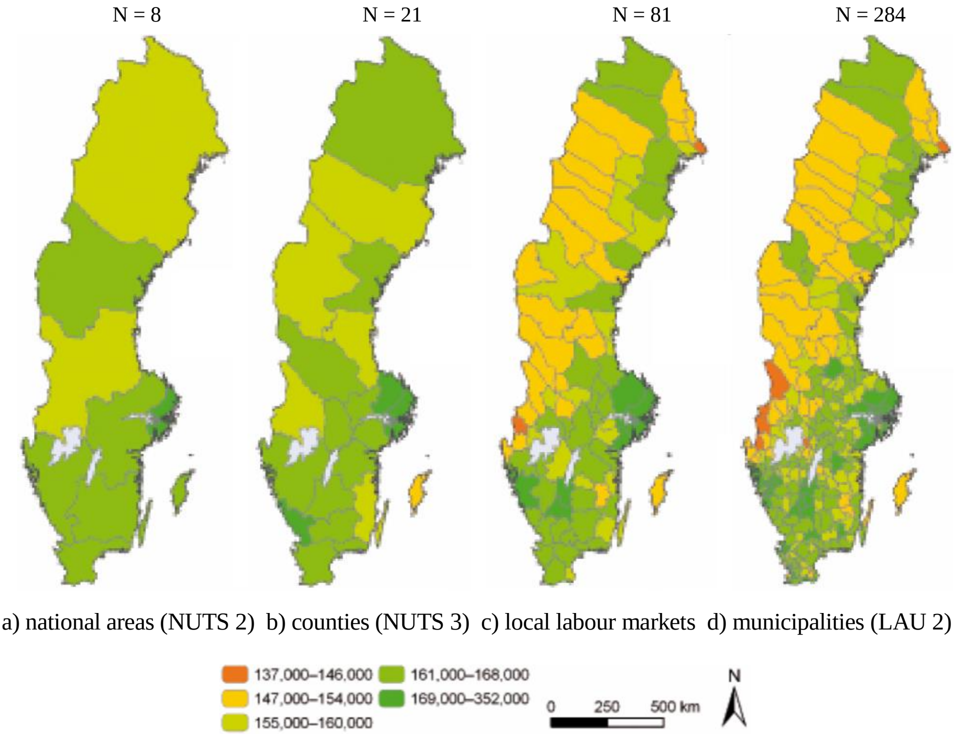
Figure 1. Examples of the scale effect (a-c) and zoning effect (d-f) of MAUP

Source: Jelinski D. E., Wu J., 1996, The modifiable areal unit problem and implications for landscape ecology, Landscape Ecology, Vol. 11, No. 3, pp. 129−140.

The operation of “averaging” data results in smoothing the data and losing information. For example, the disposable income in Swedish NUTS 2 units was between 155 and 168 SEK, whereas the values recorded by 284 Swedish municipalities (LAU 2) are held in [137,000 – 352,000] SEK *per capita* in 2002 (see parts a, b, d of Figure 2). The scale effect has at least two consequences. The data aggregation (shifting from a finer to a coarse scale) results in decreasing the variance (see Moellering and Tobler, 1972), as well as the statistical correlation tends to increase with increasing the size of the areas considered (see Yule and Kendall, 1950).

The zonation (grouping, delimitation) effect concerns the spatial arrangement in zones. It considers the variability of results not due to variations in the size of the areas but rather to their shapes, e.g. metropolitan areas, local labour markets, urban areas, tourist regions, etc.

When dealing with the aggregation problem, no loss of information occurs if we shift from one boundary system to another, rather there is an alternation of information (see Arbia, 1989, p. 18). For example, in parts c, e and f of Figure 1 one can see that even when the number of zones is held constant (N = 4) the mean and variance is affected. A comparison of parts b and d of Figure 2 shows a change in variance when the orientation is altered but the size of the units remains fixed (Jelinski and Wu, 1996). For example, depending on the zone boundaries, the interpretation of disposable income in Sweden changes (see parts c and d of Figure 2).



a) national areas (NUTS 2) b) counties (NUTS 3) c) local labour markets d) municipalities (LAU 2)

Figure 2. Disposable income *per capita* (20-64 years old people) in Sweden in 2002 (SEK)

Source: the modifiable areas unit problem, European Observation Network, Territorial Development and Cohesion, The ESPON Monitoring Committee, Luxembourg 2006, p. 47.

In regional studies based on a set of units resulted from an administrative or statistical division of a territory the zonation problem exists but is omitted due to operating on territorial units which are defined in advance (e.g. NUTS classification of territorial units) and function independently. The empirical study presented in this paper is based on NUTS 2 Polish units (provinces) which function as self-government territorial units. Therefore, in the article, special attention is paid to the scale effect of MAUP and methods which deal with it.

Some solutions to this case are discussed in the literature, for example King (1997) proposed error-bound approach, Tobler (1979) formed scale-insensitive migration model, Tate and Atkinson (2001) proposed to use fractal analysis and geostatistics (kriging and related methods such as variograms), Benali and Escoffier (1990) proposed smooth factorial analysis, Fotheringham, Charlton and Brunsdon (2001) proposed the geographically weighted regression. However, none of these solutions is sufficient and universal. The scale effect of MAUP is still an open issue.

3. The scale effect of MAUP in gravity model of population migration

Migrations occur in territorial space as flows from one area to another. An econometric gravity model is a tool which examines the internal and external conditions of flows, by analogy with Newton's (1687) concept of gravity (see Isard, 1960; Chojnicki, 1966; Anderson, 1979; Fotheringham and O'Kelly, 1989; Grabiński, Malina, Wydymus and Zeliaś, 1988; Sen and Smith, 1995; Zeliaś, 1999, pp. 172-175; Roy, 2004; LeSage and Pace, 2008; Suchecki (Ed.), 2010, pp. 226-230; Chojnicki, Czyż and Ratajczak, 2011, Shepherd, 2013, Beine, Bertoli and Fernández-Huertas Moraga, 2015).

The model typically examines three types of factors to explain mean interaction frequencies (Fischer and Wang, 2011):

a) factors pushing flows from the origin location (outflows) which indicate the ability of the origin location to produce or generate flows,
b) factors pulling flows to the destination location (inflows) which show the attractiveness of the destination location,
c) separation function that reflects the way spatial separation of origins from destinations constrains or impedes interaction such as geographical, time, economic, social, political, cultural, technological distance between locations etc.

The model can also examine the determinants of migration flows within locations (intra-regional flows). In its extended version, the model identifies the nature of spatial dependences between locations (see LeSage and Pace, 2008; Griffith and Fischer, 2013).

A researcher should also consider some problems in the construction and estimation of gravity modelling. Bertoli and Fernández-Huertas Moraga, 2013, pay a special attention to multilateral resistance in a gravity model. Santos Silva and Tenreyro, 2006, consider econometric problems resulting from heteroscedastic residuals, variables bias and the zero problem. This paper discusses the scale effect of modifiable areal unit problem which affects the results of a gravity model.

The following study concerns the economic determinants of population migrations in Poland in the years 2011-2013. The study examines factors pushing and pulling migration flows and the role of distance. The paper intentionally uses a relatively simple version of a gravity model and ignores any other problems with the construction of a gravity model to consider the scale effect of MAUP.

The gravity model used in the study (after logarithmic linearization) takes the form of:

$$\overline{Y}^* = \beta_0^* + X_o\overline{\beta}_o + X_d\overline{\beta}_d + \gamma\overline{d}^* + \overline{\varepsilon} \tag{1}$$

where: $\overline{Y}^* = \ln\overline{Y}$, $\overline{Y}$ – vector of flows from origin to destination locations,

$\boldsymbol{X_o}$ ($\boldsymbol{X_d}$) – matrices of explanatory variables realizations in the origin (destination) locations,

$X_o = [\ln\overline{x}_{o1}, \ln\overline{x}_{o2}, ..., \ln\overline{x}_{ok}]$, $X_d = [\ln\overline{x}_{d1}, \ln\overline{x}_{d2}, ..., \ln\overline{x}_{dk}]$,

$\bar{d}$ – vector containing distances between each pair of locations,

$\bar{\beta}_o, \bar{\beta}_d \ \ \gamma$ – structural parameters,

β_o^* – constant,

$\bar{\beta}_o = [\beta_{o1}, \beta_{o2}, \dots, \beta_{ok}]'$, $\bar{\beta}_d = [\beta_{d1}, \beta_{d2}, \dots, \beta_{dk}]'$,

$\bar{\varepsilon}$ – vector of disturbances.

The intensity of domestic migrations is strongly affected by macroeconomic trends. In respect of the registered migrations for permanent residence, the biggest domestic migration flows occurred just before Poland's accession to the European Union (2001-2004) and in the first years of accession (2005-2007) in which Polish economy was in the economic upturn. A big decrease in migration flows in 2008 was a reaction to the world financial and economic crisis. In subsequent years, the intensities of internal migration flows did not fluctuated. The following study covers the years 2011-2013 in which the economic situation in Poland was going to stabilize and the intensity of domestic migration flows was not changing rapidly.

The aggregated number of migration flows for permanent residence from an origin to destination province (NUTS 2 unit) in the years 2011-2013 in relation to 100 thousand inhabitants of the destination province in these years defines the dependent variable. Statistical data was collected from the Demography Database of the Central Statistical Office of Poland. Migration flows occur in territorial space and each origin is also a destination, thus we form a non-symmetric squared data matrix. This matrix is transformed into a data vector according to the approach presented in LeSage and Pace, 2008. An alternative approach is to use a panel gravity model (see Parikh and Van Leuvensteijn, 2002; Bunea, 2012; Pietrzak, Drzewoszewska and Wilk, 2012). This will allow for including provincial fixed effects and considering the issue of multilateral resistance to migration.

A set of explanatory variables was used to explain the changes of the dependent variable. The first subgroup refers to the factors which push and pull migration flows. People usually migrate to improve their living and working conditions. But their migration decisions are frequently affected by their economic and socio-economic situation and environment. In this paper we use Gross Domestic Product *per capita* (in PLN), which is a popular indicator of regional development level, as an explanatory variable of people propensity to migrate.

In an origin province, the level of regional development indicates the factor pushing migration flows to the other provinces, e.g. a weak access to education in a province may provoke massive emigrations. But for a destination province, the level of regional development is a factor pulling migration flows, e.g. relatively low costs of living may attract people to come and live in the province. The values of GDP *per capita* in 16 Polish provinces refers to 2011, which is the year of opening the studied period (2011-2013). Migrations are a long-term reaction to previous economic situation. Statistical data was provided by the Local Data Bank of the Central Statistical Office of Poland.

Other economic features (e.g. investment outlays, salary and wages, etc.) can be also used in the gravity model. But they were statistically correlated with GDP *per capita* and were excluded from the analysis to avoid multicollinearity. The alternative solution is to employ structural equation modelling in the construction of the gravity model (see, e.g. Pietrzak, Żurek, Matusik and Wilk, 2012).

The second subgroup of explanatory variables includes factors which show statistical distances between provinces. In a typical version of a gravity model of migration flows, the geographical distance is used as a separation function. However, Greenwood (1997, pp. 648-720) noticed that geographical distance elasticity of migration declines over time due to modern information, communication and transport technologies. Therefore, the economic distance is an important area of interests. In gravity models the economic distance is defined in a few ways, e.g. transportation costs, economic disproportions between units, e.g. countries, companies (see Conley and Topa, 2002, Horning and Dziadek, 1987, *Reshaping*…., 2009, p. 75, Pietrzak and Wilk, 2014).

In the following study, the economic distance will indicate the scale of the economic disparities between 16 Polish provinces and serve as the last explanatory variable in the gravity model. Because the economic disproportions result from many issues such as the level of economic activity, economic profile, attractiveness of foreign capital, local society's purchasing power and propensity to invest, labour market absorption, entrepreneurship, productivity, capacity of industry, etc., we determined a set of variables to define it (see Table 1).

Table 1. Set of variables defining the economic distances between 16 Polish provinces in 2011

No	Abbreviation	Definition	Unit
1	Investments	Investment outlays in companies per working-age people	PLN
2	Wages and salaries	Average monthly gross wages and salaries	PLN
3	Unemployment	Registered unemployment rate	%
4	Foreign capital	Companies with foreign capital per 10 thousand people	entity
5	Individual businesses	Natural persons conducting economic activity per 100 working-age people	entity
6	Employment in T&S	People employed in trade and service sectors (PKD 2007 classification) per 1 thousand working-age people	person
7	New entities	New entities of the national economy registered in REGON register per 10 thousand people	entity

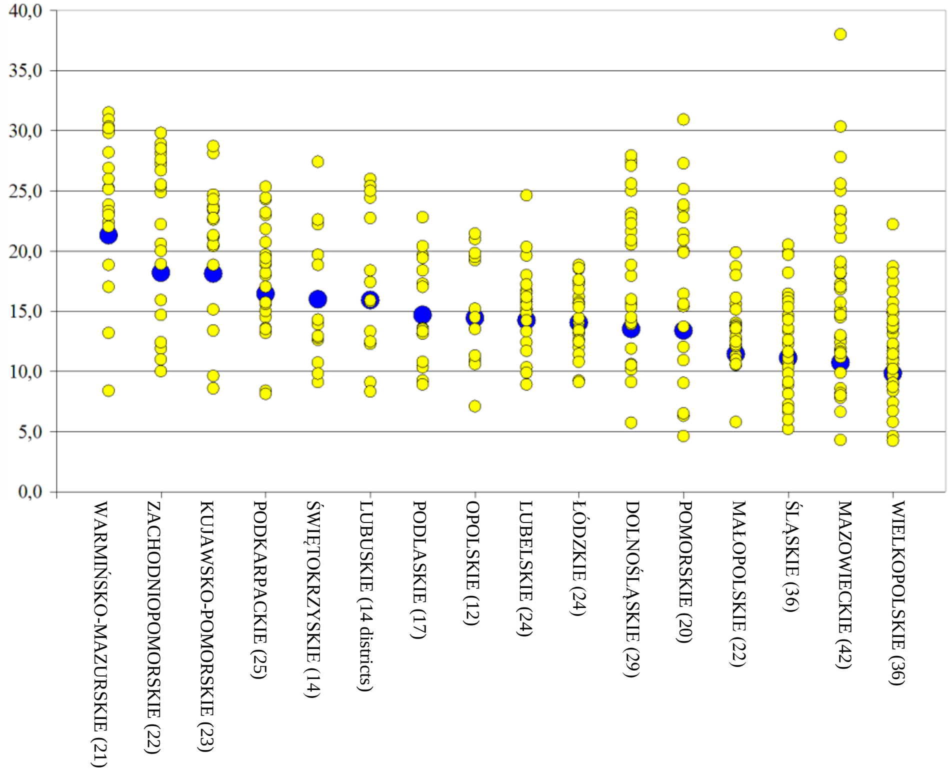
The set of diagnostic variables meet the following application criteria: statistical data availability , comparability, clear definition of the research problem and measurability. High statistical variation and low statistical correlation were also required.

The preliminary data analysis is carried out to examine if the scale effect of MAUP exists. A situation of each province was separately examined according to each variable based on statistical data for its districts (LAU 1 units).

An empirical example of the scale effect of MAUP will be presented based on the Unemployment variable. Figure 3 shows the values of the registered unemployment rate for 16 Polish provinces and 379 Polish districts assigned to provinces they are located. Dark circle tags indicate the values of official statistics for provinces, while grey circle tags show the values of official statistics for districts.

Ranges and spacing show the differences and similarities between provinces in densities, variation and reveal outlier values. For example the Zachodniopomorskie and Kujawsko-pomorskie provinces present the same level of the unemployment rate (approximately 18 %) according to provincial statistics. But in Zachodniopomorskie province the situation is much more serious. Half of its districts note at least 25 % of the unemployment rate, while majority of Kujawsko-pomorskie province’s districts record less than 25 % of the unemployment rate.

One of the lowest values of the unemployment rate is presented in the Mazowieckie province (10.7 %), while above 80 % of its districts note higher level of the unemployment rate. In the Podkarpackie and Warmińsko-mazurskie provinces, the outlier values make the official statistics much lower than they would really be.



Explanation: ● province (NUTS 2 unit) official statistics, ○ district (LAU 1 unit) official statistics.
WARMIŃSKO-MAZURSKIE (21) the name of a province (the number of districts located in the province)

Figure 3. Registered unemployment rate in Polish provinces and districts in 2011 (%)

Source: own elaboration based on Local Data Bank of the Central Statistical Office of Poland.

Table 2 presents province official statistics (POS) of the unemployment rate and basic statistics for provinces based on district data in 2011. In a vast majority of provinces, the coefficient of variation is higher than 20%, which proves that there is a relatively high internal diversification of the unemployment rate. Province official statistics are close to median values. Normal distributions do not exist for any of provinces.

Table 2. Province official statistics (POS) of registered unemployment rate and basic statistics for provinces based on district data in 2011

Name of province (NUTS 2 unit)	POS	Min	Max	Range	Mean	Median	POS per mean	POS per median	Stand. dev.	Coef. of variation (%)	Kurtosis	Skewness
Łódzkie	14.00	9.10	18.80	9.7	14.39	14.00	0.97	1.00	2.78	19.30	-0.82	-0.11
Mazowieckie	10.70	4.30	38.00	33.7	16.56	15.70	0.65	0.68	6.78	40.93	1.13	0.82
Małopolskie	11.40	5.80	19.90	14.1	13.39	13.10	0.85	0.87	3.02	22.59	1.12	0.07
Śląskie	11.10	5.20	20.50	15.3	12.11	11.60	0.92	0.96	4.09	33.80	-0.66	0.37
Lubelskie	14.20	8.90	24.60	15.7	15.46	15.90	0.92	0.89	3.36	21.71	1.34	0.34
Podkarpackie	16.40	8.10	25.30	17.2	17.98	18.90	0.91	0.87	4.49	25.00	-0.11	-0.44
Podlaskie	14.70	8.90	22.80	13.9	15.18	13.60	0.97	1.08	4.05	26.71	-1.03	0.11
Świętokrzyskie	16.00	9.10	27.40	18.3	16.39	14.10	0.98	1.13	5.53	33.74	-0.98	0.49
Lubuskie	15.90	8.30	26.00	17.7	17.61	16.65	0.90	0.95	5.96	33.82	-1.41	0.05
Wielkopolskie	9.80	4.20	22.20	18.0	11.97	11.70	0.82	0.84	4.00	33.39	0.11	0.25
Zachodnio-Pomorskie	18.20	10.00	29.80	19.8	21.84	24.90	0.83	0.73	6.57	30.10	-1.21	-0.56
Dolnośląskie	13.50	5.70	27.90	22.2	17.66	17.90	0.76	0.75	6.35	35.95	-0.94	-0.16
Opolskie	14.40	7.10	21.40	14.3	15.33	14.85	0.94	0.97	4.58	29.85	-1.30	-0.20
Kujawsko-Pomorskie	18.10	8.60	28.70	20.1	21.36	22.60	0.85	0.80	5.19	24.30	1.02	-1.07
Pomorskie	13.40	4.60	30.90	26.3	17.31	18.15	0.77	0.74	7.28	42.05	-0.88	-0.12
Warmińsko-Mazurskie	21.30	8.40	31.50	23.1	24.05	25.10	0.89	0.85	5.76	23.96	1.32	-1.10

Source: own elaboration based on Local Data Bank of the Central Statistical Office of Poland.

In view of the preliminary analysis results, we can conclude that official statistics poorly illustrate the real situations of provinces. Similar situation occurs in the rest of variables from Table 1. Therefore the scale effect of modifiable areal unit problem has been detected and requires a special attention.

4. Symbolic data analysis as a tool to reduce the scale effect of MAUP

As Tobler (1989) and Openshaw (1984) advocated: data aggregation is not only a quantitative process; it changes dramatically the main point of the units as well as the variables. The question of MAUP must be posed before the treatment rather than during it. The approach proposed in this paper deals with this standpoint and employs symbolic data analysis to reduce the scale effect of MAUP.

Symbolic data differs from classical data situation. Conventional data set includes the observations of variables which realizations are in the form of real values or single categories. Symbolic data analysis (SDA) deals with another data form to solve data structure and data reduction problems. The first issue refers to the analysis of fuzzy (e.g. freezing interval of liquids), imprecise (e.g. income intervals of respondents), multinominal data (e.g. foreign languages known by job recruits), data fluctuated in time (e.g. investment outlays). The second problem occurs if a set of data is too large or too complicated to be analysed, e.g. a big set of units, a set of correlated variables, a big set of variables and time series.

SDA distinguishes two types of objects (units) regarding the level of data aggregation: first order and second order symbolic objects. First order objects are indivisible units (e.g. respondents, products, patients, etc.) which cannot result from data aggregation, and are described by symbolic data. Second order objects are the result of an aggregation of first order objects described by classical data. These objects can be seen as more or less homogeneous classes of individuals described by symbolic data.

The observations on symbolic variables take the form of intervals of values (interval-valued variables), sets of categories, values or intervals (multivalued variables), values or intervals with associated weights, frequencies, probabilities, etc., (modal variables), and also taxonomic, logical or hierarchical structures (dependent variables). Furthermore, classical data is a special case of symbolic data, e.g. a classical ratio data is equivalent to the point which is a special case of the interval-valued variable realization, a classical categorical data is equivalent to a single realization of a multivalued variable (see Bock and Diday (Eds.), 2000; Billard and Diday, 2006; Diday and Noirhomme-Fraiture (Ed.), 2008; Wilk, 2011).

Symbolic data occurs in a natural form or results from classical data aggregation. The ways of symbolic data construction in presented in Bock and Diday (Eds.), 2000; Wilk, 2012. In regional research, higher level units (e.g. regions) can be characterized based on situations of lower level units (e.g. towns) using symbolic data.

For example, we gathered a classical data set regarding 16 Polish NUTS 2 units (provinces) and the values of unemployment rate recorded by all Polish LAU 1 units (districts). Each district is located in the territory of one of the provinces. We can distinguish a set of unemployment categories such as low (unemployment rate of less than 10%), medium (unemployment rate of 10-20%) and high unemployment (unemployment rate higher than 20%) levels in sub-regions. Then, for each province we calculate the frequencies of districts which satisfy each category. In this way we have already constructed a set of second order symbolic objects (provinces) described by a symbolic modal variable (see Table 3).

Table 3. Registered unemployment rate in Polish regions in 2010 (%)

NUTS 2 unit (province)		LAU 1 unit (district)		Fraction of LAU 1 units satisfying each level of unemployment*			Symbolic modal variable realizations
Name	Value	Name	Value	low	medium	high	
Mazowieckie	9.7	Warszawa	3.5	0.2	0.5	0.3	{low (0.2), medium (0.5), high (0.3)}
		Warszawski zachodni	5.9				
		⋮	⋮				
		Radomski	30.8				
		Szydłowiecki	36.0				
Kujawsko-Pomorskie	17.0	Bydgoszcz	8.0	0.1	0.3	0.6	{low (0.1), medium (0.3), high (0.6)}
		⋮	⋮				
		Lipnowski	28.9				
⋮	…	…	…	…	…	...	…

* low (under 10%), medium (10-20%], high (over 20%) level of unemployment rate

Source: own elaboration based on Local Data Bank of the Central Statistical Office of Poland.

The aim of the following procedure is to reduce the scale effect of MAUP in the examination of the economic distance between provinces using symbolic data analysis. In the first step of the procedure we construct a set of second order symbolic objects. These objects represent 16 Polish NUTS 2 units (provinces) which are described by 7 symbolic interval-valued variables. We use a set of indicators presented in Table 1.

In this case, the constructions of variables result from the aggregation of district data. For each province we construct an interval of values which consists of minimum and maximum values recorded by the districts located in a province as regards each variable. But the construction of intervals of values required to detect and remove outlier values. We removed the values which were much higher than 80% of observations for districts in a province but not higher than 10% of observations for districts in the province.

In the next step we determine the statistical distances between symbolic objects. We use Ichino-Yaguchi's normalized distance measure (Ichino and Yaguchi, 1994):

$$d_{ijk} = \mu(v_{ik} \oplus v_{jk}) - \mu(v_{ik} \otimes v_{jk}) + \gamma v(v_{ik}, v_{jk}) \tag{2}$$

where: i, j – the number of an object, $i \in [1, m]$,

k – the number of a variable, $k \in [1, p]$,

$v_{ik} \oplus v_{jk} \equiv \left[\min\{\underline{v_{ik}}, \underline{v_{jk}}\}, \max\{\overline{v_{ik}}, \overline{v_{jk}}\}\right]$,

$v_{ik} \otimes v_{jk} \equiv v_{ik} \cap v_{jk}$,

$\nu(v_{ik}, v_{jk}) \equiv 2\mu(v_{ik} \otimes v_{jk}) - \mu(v_{ik}) - \mu(v_{jk})$,

$\underline{v_{ik}}, \underline{v_{jk}}$ ($\overline{v_{ik}}, \overline{v_{jk}}$) – the start point and end point of interval of values observed by *k* variable, accordingly, for *i* and *j* objects,

γ – parameter, $\gamma \in [0.0, 0.5]$.

The distance measure (Equation 2) examines a dissimilarity between *i* and *j* objects regarding *k* variable. It calculates Cartesian meet ($v_{ik} \otimes v_{jk}$) and Cartesian join ($v_{ik} \oplus v_{jk}$). If the intersection $v_{ik} \otimes v_{jk}$ takes the empty value, both intervals of values observed for *i* and *j* objects have no common part. If both intervals of values observed for *i* and *j* objects have the same minimum and maximum values, the Cartesian join $v_{ik} \oplus v_{jk}$ is an interval with minimum and maximum values observed for *i* or *j* unit (see Ichino and Yaguchi, 1994, Wilk, 2006). Parameter γ takes values from 0.0 to 0.5 but for 0.5 the $\nu(v_{ik}, v_{jk})$ is equal to 0 and no intersection is included. Then, in the following study, the parameter γ is equal to 0.4.

We use Minkowski's metric (with λ equal to 2) to determine the total distance between *i* and *j* objects:

$$d_{ij} = \left[\sum_{k=1}^{p} (d_{ijk})^{\lambda} \right]^{1/\lambda} \tag{3}$$

where: λ – parameter, $\lambda \geq 1$.

Minkowski's metric takes the values in $[0.0, \infty]$. If all variables take the same observations for *i* and *j* objects, then the value of the measure is equal to 0. The higher the values, the longer the distance between *i* and *j* objects. If the distance is short, then the similarity of two compared provinces is high. If all pairs of provinces present very short distances then the economic disparities would be very low as well.

The same distance measure was also implemented to determine dissimilarities between 16 Polish provinces described by province official statistics (ratio data). For ratio data, the intersection $v_{ik} \otimes v_{jk}$ takes the values of 0 if $v_{ik} \neq v_{jk}$. If $v_{ik} = v_{jk}$, then $v_{ik} \otimes v_{jk} \equiv v_{ik} \oplus v_{jk}$.

Table 4 presents the comparison of results of economic distance measurements based on province official statistics and symbolic interval-valued data. It includes basic descriptive statistics for both the distance matrices and the sets of pairs of provinces with very long and very short distances.

Table 4. Comparison of the economic distance measurement results based on two data sets (official statistics and symbolic data)

Data set	Descriptive statistics			Coef. of var. (%)	Economic distance	
	Min	Max	Median		Very short	Very long
Province official statistics	0.19	2.46	0.83	52.77	Lubelskie vs. Podlaskie (0.19) Podlaskie vs. Świętokrzyskie (0.24) Dolnośląskie vs. Pomorskie (0.31) Lubelskie vs. Podkarpackie (0.33)	Mazowieckie vs. Warmińsko-Mazurskie (2.46) Mazowieckie vs. Podkarpackie (2.38) Mazowieckie vs. Lubelskie (2.29) Mazowieckie vs. Świętokrzyskie (2.23) Mazowieckie vs. Podlaskie (2.20)
Symbolic interval-valued data	0.16	0.91	0.53	34.20	Lubelskie vs. Podlaskie (0.16) Lubelskie vs. Podkarpackie (0.16) Lubelskie vs. Łódzkie (0.20)	Mazowieckie vs. Świętokrzyskie (0.91) Świętokrzyskie vs. Wielkopolskie (0.88) Świętokrzyskie vs. Pomorskie (0.87) Mazowieckie vs. Podlaskie (0.87)

Source: own estimation in symbolicDA package (Dudek, Pełka and Wilk, 2013) of R-CRAN, based on Local Data Bank of the Central Statistical Office of Poland.

In both cases distances between provinces are higher than 0 which means that there are no two provinces with the same values of all variables and the economic disproportions in Poland exist. The comparison of distance ranges proves that the scale of the economic disparities in Poland is smaller than official statistics show. Moreover, the economic relations between provinces based on symbolic interval-valued data set differ from those presented by province official statistics.

5. Conditions of domestic population migration in Poland

Three gravity models of population migration in Poland in 2011 were used in the following study (Table 5). The aggregated number of migration flows for permanent residence from an origin to destination province (NUTS 2 unit) in the years 2011-2013 in relation to 100 thousand inhabitants of the destination province in the same period defines the dependent variable.

Table 5. Gravity models of domestic population migration in Poland in 2011

Specification	Model 1	Model 2	Model 3
Pulling and pushing factor	GDP *per capita*	GDP *per capita*	GDP *per capita*
Separation function	Geographical distance	Economic distance	Economic distance
Data set	Province official statistics	Province official statistics	Symbolic interval-valued data

Source: own elaboration.

All models include two types of dependent variables: a pushing and pulling factor and a separation function as well. All models employ GDP *per capita* to measure the impact of regional development level in an origin to push migration flows and attractiveness of a destination to pull migration flows.

But the first model examines geographical distance as well. In this study, the number of kilometres in a straight line as regards the centroids served to determine the geographical distances between provinces. This model is based on province official statistics.

In contrast to the first model, the second and third models examine the role of the economic distance. But the second model employs province official statistics (ratio data) to determine the economic distance between provinces. In the third model, the economic distance between each pair of provinces is determined based on symbolic interval-valued data. Economic distances between provinces include their internal situation and economic disproportions between districts.

Table 6 presents the results of gravity models of population migration. All models use GDP *per capita* to explain pushing and pulling factors of flows but the models differ from the type of distance used and the method of its determination. The first model concerns the geographical distance. The other models consider the economic distance. One of them is based on official statistics recorded by provinces. The second one is based on symbolic data set.

Table 6. The estimates of gravity models of domestic population migration in Poland in 2011

Parameter		Model 1		Model 2		Model 3	
		estimate	p-value	estimate	p-value	estimate	p-value
Constant	α_1	-100,095	0,093***	-103,749	0,093*	109,936	0,047**
The impact of economic situation of an origin on population outflow from the origin	β_o	0,231	0,001***	0,005	0,000***	0,0034	0,002***
The impact of economic situation of a destination on population inflow to the destination	β_d	0,396	0,000***	0,005	0,000***	0,0033	0,003***
The impact of distance on population flows	γ	-102,008	0,001***	216,765	0,000***	583,379	0,000***
R^2 coefficient		0,46		0,54		0,51	

Significance level: *** 5%, ** 10%, * 15%

Source: own estimation in Gretl based on Local Data Bank of the Central Statistical Office of Poland.

The least squares method was used in estimation. All estimated values are statistically significant. Estimates of both origin-destination flows parameters (β_o and β_d) are positive. The better the economic situation in a province, the higher the population migration inflows, as well as population outflows.

The geographical distance estimate is negative, whereas the economic distance estimates are positive in both approaches. Therefore, the longer the geographical distance, the lower the migration flows. People migrate on relatively short distances in Poland. But the longer the economic distance, the higher the migration flows. Moreover, migration movement is big when the economic disparities occur in the country.

The estimates of the economic distance in both models are comparable due to using the same distance measure. The estimate in the model based on symbolic data is much higher than the estimate in the model based on official statistics. Therefore, the role of the economic distance in domestic migration flows is in fact more serious than official statistics show.

6. Conclusions

The presented study discussed the modifiable areal unit problem (MAUP) in spatial data analysis and proposed to use symbolic data analysis (SDA) to reduce the scale effect of MAUP in the gravity model of population migration.

Symbolic data was used to measure the economic disparities in Poland. This approach involves a special way of data aggregation. Its main advantage is to include details rather than "averaged" values. In fact, the scale of economic diversification in Poland is smaller whereas the role of the economic distance in domestic migration flows is more serious than official statistics show. The disadvantage of this approach is a highly complicated procedure, which includes preliminary data aggregation and the use of distance measure dedicated to symbolic data analysis.

An open issue is symbolic variables and construction of objects. In the following study, the second order symbolic objects (provinces) included district-level data due to statistical data availability. The ideal situation would be to dispose individual data for non-modifiable units (e.g. towns) as a starting point in data aggregation procedure. But some data is not presented due to statistical confidentiality (e.g. suicide data) and some statistical surveys are not carried out at the local level of territorial division (e.g. GDP *per capita*).

The second open issue is an appropriate construction of symbolic variables, which is an individual case and depends on the properties of a variable. In the study, aggregation of classical ratio data into symbolic interval data required removing outlier values. In this situation we lost some information. Moreover, if symbolic interval data is not a natural form of data, a symbolic interval-valued variable covers some additional values which were not recorded by any analysed units. In these situations we can try to use another type of symbolic variables, e.g. modal variables. This problem will be an objective of further research study.

REFERENCES

ANDERSON, J. E., (1979). A Theoretical Foundation for the Gravity Model, American Economic Review, 69 (1), pp. 106−116.

ANSELIN, L., (1988). Spatial Econometrics: Methods and Models, Kluwer Academic, Dordrecht.

ARBIA, G., (1989). Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems, Advanced Studies in Theoretical and Applied Econometrics, Vol. 14, Kluwer Academic Publishers, Dordrecht-Boston-London.

BEINE, M., BERTOLI, S., FERNÁNDEZ-HUERTAS MORAGA, J., (2015). A Practitioners' Guide to Gravity Models of International Migration, World Economy.

BENALI, H., ESCOFIER, B., (1990). Analyse factorielle lissée et analyse factorielle des différences locales, Revue de statistiques appliquées, XXXVIII (2), pp. 55−76.

BERTOLI, S., FERNÁNDEZ-HUERTAS MORAGA, J., (2013). Multilateral resistance to migration, Journal of Development Economics, 102, May, pp. 79−100.

BILLARD, L., DIDAY, E., (2006). Symbolic Data Analysis. Conceptual Statistics and Data Mining, Wiley, Chichester.

BOCK, H.-H., DIDAY, E. (Eds.), (2000). Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data, Springer-Verlag, Berlin-Heidelberg.

BUNEA, D., (2012). Modern Gravity Models of Internal Migration. The Case of Romania, Theoretical and Applied Economics, Vol. XIX, No. 4(569), pp. 127−144.

CHOJNICKI, Z., (1966). Application of gravity and potential models in spatio-economic research (in polish), PWN, Warsaw.

CHOJNICKI, Z., CZYŻ, T., RATAJCZAK, W., (2011). Potential model. Theoretical basis and applications in spatio-economic and regional research (in polish), Bogucki Wydawnictwo Naukowe, Poznan.

CLIFF, A. D., ORD, J. K., (1981). Spatial processes: models and applications, Pion, London.

CONLEY, T. G., TOPA, G., (2002). Socio-economic distance and spatial patterns in unemployment, Journal of Applied Econometrics, Vol. 17/4.

DARK, S. J., BRAM, D., (2007). The modifiable areal unit problem (MAUP) in physical geography, "Progress in Physical Geography", Vol. 31, No. 5.

DIDAY, E., NOIRHOMME-FRAITURE, M. (Eds.), (2008). Symbolic data analysis and the Sodas software, John Wiley & Sons, Chichester.

DUDEK, A., PEŁKA, M., WILK, J., (2013). symbolicDA package of R-CRAN, http://cran.r-project.org/web/packages/symbolicDA/index.html.

FISCHER, M. M., WANG, J., (2011). Spatial data analysis: models, methods and techniques, Springer Briefs in Regional Science, Springer, Berlin.

FOTHERINGHAM, A. S., CHARLTON, M.E., BRUNSDON, C. F., (2001). Spatial Variations in School Performance: a Local Analysis Using Geographically Weighted Regression, Geographical & Environmental Modelling, Vol. 5, pp. 43−66.

FOTHERINGHAM, A. S., O'KELLY, M. E., (1989). Spatial interaction models: formulations and applications, Kluwer, Dordrecht,

GATNAR, E., WALESIAK, M. (Ed.), (2011). Qualitative and symbolic data analysis using R program (in polish), C.H. Beck, Warsaw.

GEHLKE, C. E., BIEHL, K., (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material, Journal of the American Statistical Association, No. 29.

GHATAK, S., MULHERN, A., WATSON, J., (2008). Inter-regional migration in transition economies. The case of Poland, Review of Development Economics, No. 12(1), Oxford, pp. 209−222.

GOTWAY CRAWFORD, C. A., YOUNG, L. J., (2004). A spatial view of the ecological inference problem, In: G. King, O. Rosen, M. Tanner (Eds.), Ecological Inference: New Methodological Strategies, Cambrige University Press, pp. 233−244.

GRABIŃSKI, T., MALINA, A., WYDYMUS, S., ZELIAŚ, A., (1988). International statistics methods (in polish), PWE, Warsaw.

GREENWOOD, M. J., (1997). Internal migration in developed countries, In: Rosenzweig, M. R. and Stark, O. (Eds.), Handbook of Population and Family Economics, vol. 1B, Elsevier, Amsterdam, pp. 648−720.

GRIFFITH, D. A., FISCHER, M., (2013). Constrained variants of the gravity model and spatial dependence: Model specification and estimation issues, Journal of Geographical Systems, 15(3), pp. 291−317.

HOLZER, J. Z., (2003). Demography (in polish), PWE, Warsaw.

HORNING, A., DZIADEK S., (1987). Outline of land transport geography (in polish), PWN.

ICHINO, M., YAGUCHI, H., (1994). Generalized Minkowski Metrics for Mixed Feature-Type Data Analysis, IEEE Transactions on Systems, Man, and Cybernetics, Vol. 24, No. 4.

ISARD, W., (1960). Methods of regional analysis, MIT Press, Cambridge.

JELINSKI, D. E., WU, J., (1996). The modifiable areal unit problem and implications for landscape ecology, Landscape Ecology, Vol. 11, No. 3, pp. 129−140

KING, G., (1997). A solution to the ecological inference problem: reconstructing individual behaviour from aggregate data, Princeton University Press, Princeton.

KING, G., TANNER M. A., ROSEN O. (Eds.), (2004). Ecological Inference: New Methodological Strategies, Cambridge University Press, New York.

KAPISZEWSKI, M., DURHAM, H., REES, P., (1999). Internal Migration and Regional Population Dynamics In Europe: Poland Case Study, In: P. Rees, M. Kupiszewski (Eds.), Internal Migration and Regional Population Dynamics in Europe: A Synthesis, Collection Demography, Council of Europe, Strasbourg.

LESAGE, J. P., PACE, R. K., (2008). Spatial econometric modeling of origin-destination flows, Journal of Regional Science, Vol. 48, No. 5, pp. 941−968.

Local Data Bank of the Central Statistical Office of Poland, http://www.stat.gov.pl/bdl.

LUCAS, R., (1997). Internal Migration in Developing Countries, In: M.R. Rosenzweig, O. Stark (Eds.), Handbook of Population and Family Economics, Elsevier Science B.V, Amsterdam, pp. 721−798.

MOELLERING, H., TOBLER, W. R., (1972). Geographical variances, Geographical Analysis, Vol. 4, pp. 34−50.

OPENSHAW, S., TAYLOR, P. J., (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem, In: N. Wringley (Ed.), Statistical Applications in the spatial sciences, Pion, London, pp. 127−144.

OPENSHAW, S., (1984). The Modifiable Areal Unit Problem, GeoBooks, CATMOG 38, Norwich.

PARIKH, A., VAN LEUVENSTEIJN, M., (2002). Internal migration in regions of Germany: A panel data analysis, Working Paper, No. 12, European Network of Economic Policy Research Institutes.

PIETRZAK, M., ŻUREK, M., MATUSIK, S., WILK, J., (2012). Application of Structural Equation Modeling for analysing internal migration phenomena in Poland, Przegląd Statystyczny (Statistical Review), No 4, Vol. LIX, pp. 487−503.

PIETRZAK, M. B., DRZEWOSZEWSKA, N., WILK, J., (2012). The analysis of interregional migrations in Poland in the period of 2004-2010 using panel gravity model, Dynamic Econometric Models, Vol. 12(2012), pp. 111−122.

PIETRZAK, M., WILK, J., (2014). The economic distance in spatial phenomena modelling with the use of gravity model (in polish), In: K. Jajuga, M. Walesiak (red.), Taksonomia 22. Klasyfikacja i analiza danych – teoria i zastosowania, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, 37, pp. 177−185.

Reshaping Economic Geography, (2009). World Bank, Washington.

ROY, J. R., (2004). Spatial Interaction Modeling: A Regional Science Context, Springer-Verlag, Berlin.

SANTOS SILVA, J., TENREYRO, S., (2006). The log of gravity, The Review of Economics and Statistics, 88, pp. 641−58.

SEN, A., SMITH, T. E., (1995). Gravity models of spatial interaction behavior, Springer, Berlin-Heilderberg-New York.

SHEPHERD, B., (2013). The Gravity Model of International Trade: A User Guide, United Nations ESCAP & ARTNET.

SUCHECKA, J. (Ed.), (2014). Spatial statistics. Spatial structures analysis methods (in polish), C.H. Beck., Warsaw.

SUCHECKI, B. (Ed.), (2010). Spatial econometrics. Spatial data analysis methods and models (in polish), C.H. Beck, Warsaw.

TATE, N. J., ATKINSON, P. M. (Eds.), (2001). Modelling scale in geographical information sciences, Wiley & Sons, London.

TOBLER, W., (1979). Smooth pycnophylactic interpolation for geographical regions, Journal of the American Statistical Association, Vol. 74, pp. 519−536.

TOBLER, W., (1989). Frame independent spatial analysis, In: Accuracy of Spatial Databases, M. Goodchild S. Gopal (Eds.), CRC Press, pp.115−122.

TODARO, M., (1980). Internal Migration in Developing Countries. A survey, In: R. A. Easterlin, Population and Economic Change in Developing Countries, University of Chicago Press, Chicago, pp. 361−402.

VAN DER GAAG, N. (Ed.), (2003). Study of past and future interregional migration trends and patterns within European Union countries: in search of a generally applicable explanatory model, Report on behalf of Eurostat.

WHITE, M. J., LINDSTROM, D. P., (2006). Internal Migration, In: D.L. Poston, M. Micklin (Eds.), Handbook of Population, Springer, Berlin-Heilderberg.

WILK, J., (2012). Symbolic approach in regional analyses, Statistics in Transition – new series, Vol. 13, No 3, December 2012, pp. 581−600, http://stat.gov.pl/cps/rde/xbcr/pts/SIT_13_3_December_2012.pdf.

WILK, J., (2011). Cluster analysis based on symbolic data (In Polish), In: E. Gatnar, M. Walesiak (Eds.), Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R, C.H. Beck, Warsaw, pp. 262−279.

WILK, J., (2006). Problems of symbolic objects classification. Symbolic distance measures (In Polish), In: J. Garczarczyk (Ed.), Ilościowe i jakościowe metody badania rynku. Pomiar i jego skuteczność, Zeszyty Naukowe Akademii Ekonomicznej w Poznaniu, 71, pp. 69−83.

WONG, D., The modifiable areal unit problem (MAUP), (2009). In: Fotheringham A.S., Rogerson P.A. (Eds.), The SAGE Handbook of Spatial Analysis, SAGE Publications Ltd., pp. 105−123.

YULE, U., KENDALL M. S., (1950). An introduction to the theory of statistics, Charles Griffin, London.

ZELIAŚ, A. (Ed.), (1991). Spatial econometrics (In Polish), PWE, Warsaw.

STATISTICS IN TRANSITION new series, Summer 2015
Vol. 16, No. 2, pp. 265–278

ANALYSIS OF CONVERGENCE OF EUROPEAN REGIONS WITH THE USE OF COMPOSITE INDEX

Joanna Górna[1], **Karolina Górna**[2]

ABSTRACT

Convergence study is related to several crucial issues. One of those problems is an individual character of every region in the selected area, as the regions established accordingly to the European classification system NUTS-2 are not homogeneous. Therefore, while analysing convergence in the European Union, regions with extremely dissimilar characteristics (for example GDP per capita) are taken under consideration. Absolute β-convergence means that all of the investigated regions tend to the same level of economic growth. Thus, among the regions with highly differential amounts of the examined variables the convergence hypothesis can be rejected. Due to the heterogeneity in the conducted investigation a classification based on the composite index will be used so that the convergence clubs could be established. Several approaches to convergence will be used according to those regimes. Moreover, there will be an attempt to indicate the determinants that differentiate the selected regions, such as: expenditure on R&D, HRST, quantity of patents, employment, participation of people in tertiary education among all employees. This will allow the analysis of conditional β-convergence to be conducted. In the investigation some methods and models offered by the spatial statistics and econometrics will be used. There are empirical proofs that geographical location has a great impact on the processes of economic growth. Consequently, spatial dependencies will be analysed as well.

Key words: economic convergence/divergence, spatial autocorrelation, spatial econometric model, composite index.

Introduction

Regional growth and convergence are issues that attract a great deal of attention. Research on this topic is developing in different directions, with β-convergence phenomenon being one of them.

The neoclassical growth theory assumes that economies with initially low level of development tend to have faster productivity growth, due to diminishing returns

[1] The Nicolaus Copernicus University in Toruń. E-mail:gorna.joanna@gmail.com.
[2] The Nicolaus Copernicus University in Toruń. E-mail:gorna.karolina@gmail.com.

on capital. In other words, regions with initial position of a relatively high capital-labour ratio tend to grow relatively slower than economies with low ratio. That induces that low technology regions are able to converge to the steady-state (Fingleton, 2001, p. 117). At its simplest, the convergence phenomenon connected with the neoclassical growth theory implies elimination of dissimilarity between investigated economies. Taking region differences under consideration leads to the expanded convergence model, which accommodates the existence of regionally differentiated steady-state. That leads to conditional convergence, where regions converge to their specific steady-states rather than to one common steady-state level.

Among the absolute convergence studies, the following works can be mentioned: Baumol W. J. (1986), De Long J. B. (1988). Conditional convergence was investigated in Mankiw N. G., Romer D., Weil D. N. (1992), Barro R. J., Sala-i-Martin X. X. (1992).

Apart from the diminishing returns of scale impact on the convergence phenomenon, there are also additional forces that lead regions to their steady-states. Those forces are connected with spatial or regional interactions among investigated economies. Spatial interactions in growth and convergence investigations are conspicuously absent in most empirical convergence investigations. However, the literature on spatial econometrics emphasizes that spatial interactions are crucial in growth and convergence understanding. Those models should acknowledge that changes in one region are able to spill over into other regions. It leads to conclusions that the dynamics of steady-state is affected by interregional interdependencies. Such understanding of convergence can be found, e.g. in: Abreu, M., de Groot, H. L. F., Florax, R. J. G. M. (2005), Rey, S. J., Janikas (2005), Ertur, C., Koch, W. (2007), Fingleton, B., López-Bazo, E. (2006).

The purpose of this investigation is to answer the question whether the phenomenon of convergence of per capita GDP occurs in the area of the European Union countries. Moreover, convergence clubs are established to investigate if convergence has a global or a local character. Convergence clubs are connected with the fact that regions are not strictly homogenous.

The structure of the other part of the paper is as following: in section 1 the subject and the range of investigation is specified. Section 2 briefly characterizes the data used in the investigation and the preliminary analysis. Section 3 presents methodology, including theoretical β-convergence models for cross-section and spatio-temporal regressions, and also points diagnostic statistic tests. In section 4 the results of the investigation are presented. Section 5 presents conclusions and further investigation directions.

1. Subject and range of investigation

According to regional and spatial interactions relevance mentioned above, the main hypothesis of investigation is that spatial dependencies are crucial in investigating the convergence phenomenon. During the investigation several questions will be answered:

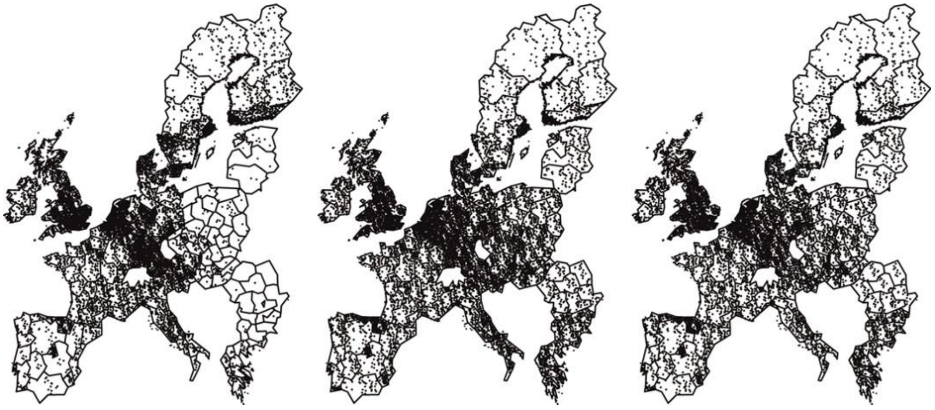
1. How to establish convergence clubs, which will include homogenous regions?
2. What is the relation between the type of convergence (absolute or conditional) and the value of speed of convergence and half-life?
3. Does the approach in convergence investigation (global or local) affect the time in which economies reach the same level of development, understood as per capita GDP growth rate?
4. How can the problem of omitted variables in conditional convergence study be solved?

To achieve the aim mentioned in the introduction it will be demonstrated that empirical spatial models of convergence for cross-section data have better statistical characteristics than the ones that ignore the spatial dependencies among regions. It will be also shown that those models allow more precise economic interpretation of parameters. Empirical models are used for the hypothesis verification that the location of the region towards other regions has crucial impact for growth rate in this region.

2. Data

In the investigation the values of per capita GDP were used. The data refer to the period of 2003-2011. The spatial range of the investigation consists of 249 regions of NUTS-2 classification of the European Union countries.

Besides per capita GDP, some other variables were used in the investigation. Conditional convergence estimation was based on the following variables: general expenditures on R&D (GERD), human resources in science & technology (HRST), participation of people with tertiary education among all employees. Figure 1 presents the spatial density of the above mentioned variables. It is visible on the maps that the established area is not a homogenous group of spatial units, therefore it is reasonable to consider clubs of convergence.



(a) (7) (12)

Figure 1. Density of: (a) GERD, (b) HRST, (c) tertiary education level in 2003

3. Methodology

The classical approach to β-convergence analysis assumes that initial values of per capita GDP are negatively correlated with growth rates. The linear models using data in cross-section take the forms:

1. absolute convergence:

$$\ln\left[\frac{GDP_{iT}}{GDP_{i1}}\right] = \alpha_0 + \beta\ln[GDP_{i1}] + \varepsilon_i,$$

2. conditional convergence:

$$\ln\left[\frac{GDP_{iT}}{GDP_{i1}}\right] = \alpha_0 + \beta\ln[GDP_{i1}] + \sum_{m=1}^{r}\alpha_m x_{im} + \varepsilon_i,$$

where:

β – convergence parameter; when it is negative and statistically significant, the convergence phenomenon is observed,

x_{im} – value of m^{th} characteristic of region i.

In the conditional convergence model, additional variables inform about special characteristics of regions.

While analysing convergence, spatial approach is advisable. So far, the tools and methods offered by spatial econometrics have been rarely used. Yet in the spatial literature it is strongly pointed that in growth investigations spatial connections cannot be obtained.

The essential element in spatial models is the connectivity matrix, usually marked as **W**. This matrix defines the structure of the spatial connections among spatial units (neighbourhood based on a common border). The matrix has as many rows and columns as there are spatial units. **W** is given as:

$$\boldsymbol{W} = \left[w_{ij}\right]_{N\times N}.$$

Each element of the matrix is non-zero for pairs of regions which are neighbours. Because a region cannot be its own neighbour, element $w_{ij} = 0$ for $i = j$.

Spatial dependencies can be provided by using one of following models: spatial autoregressive model or spatial error model. Those models take the following forms:

1. spatial autoregressive model (SAR):
   1) absolute convergence:

$$\ln\left[\frac{GDP_{iT}}{GDP_{i1}}\right] = \alpha + \beta\ln[GDP_{i1}] + \rho\sum_{j\neq i} w_{ij}\ln\left[\frac{GDP_{jT}}{GDP_{j1}}\right] + \varepsilon_i,$$

   2) conditional convergence:

$$\ln\left[\frac{GDP_{iT}}{GDP_{i1}}\right] = \alpha_0 + \beta\ln[GDP_{i1}] + \sum_{m=1}^{r}\alpha_m x_{im} + \rho\sum_{j\neq i} w_{ij}\ln\left[\frac{GDP_{jT}}{GDP_{j1}}\right] + \varepsilon_i,$$

2. spatial error model (SEM):
   1) absolute convergence:

$$\ln\left[\frac{GDP_{iT}}{GDP_{i1}}\right] = \alpha + \beta \ln[GDP_{i1}] + \eta_i,\ \eta_i = \lambda \sum_{j\neq i} w_{ij}\,\eta_j + \varepsilon_i,$$

   2) conditional convergence:

$$\ln\left[\frac{GDP_{iT}}{GDP_{i1}}\right] = \alpha + \beta \ln[GDP_{i1}] + \sum_{m=1}^{r} \alpha_m x_{im} + \eta_i,\ \eta_i = \lambda \sum_{j\neq i} w_{ij}\,\eta_j + \varepsilon_i.$$

When parameters ρ or λ are significant it indicates that spatial dependencies are crucial in the estimated model.

For the verification of estimated models several diagnostics tests were used:

1. the Moran test (Moran's I) – for the consideration of the 1st order spatial autocorrelation of per capita GDP in the established area and for spatial independence of residuals of the β-convergence models,
2. the Lagrange Multiplier tests (LMlag, LMerr) and their robust versions (RLMlag, RLMerr) as spatial dependence diagnostics,
3. the Likelihood Ratio test (LR) for testing the significance of the spatial dependence.

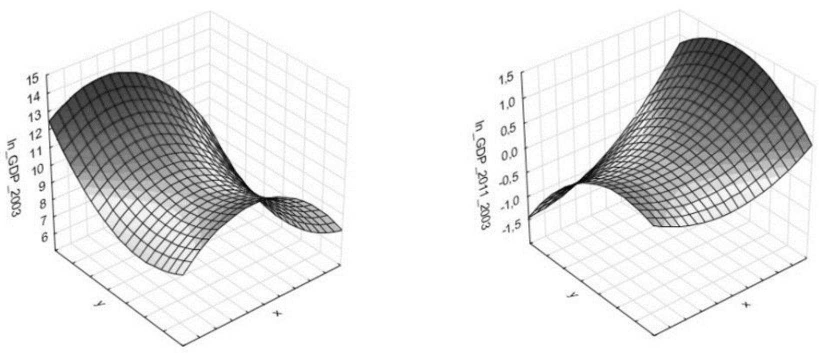
In the conducted investigation the division for convergence clubs was based on the values of the composite index. Stimulants used to calculate the index were the following: general expenditure on R&D, human resources in science & technology, participation of people with tertiary education among all employees and per capita GDP.

4. Results of empirical analysis

The convergence hypothesis was investigated with the use of two approaches. Firstly, the whole selected area was considered. Secondly, spatial regimes were identified and club convergence was considered.

4.1. European Union regions

The first stage in β-convergence investigation was founded on spatial trend surface analysis. Figure 2 compares the spatial distribution of per capita GDP in the initial year to GDP growth rate during the period 2003-2011. The tendencies in Figure 2 part (a) and (b) have inverted surfaces. This leads to the conclusion that the values of per capita GDP in 2003 are negatively correlated to growth rates. This can suggest that the convergence phenomenon occurs in the selected area.



(a) **(b)**

Figure 2. Spatial trends: (a) per capita GDP in 2003, (b) per capita GDP growth rate 2003-2011

Spatial trend analysis was followed by the estimation of econometric models presented in Section 3. Convergence was investigated using the absolute and conditional approach, respectively. Table 1 contains the results of the estimation and verification for the models mentioned above.

Table 1. Results of estimation and verification LM, SAR and SE models

	LM		SAR		SE	
	absolute	conditional	absolute	conditional	absolute	conditional
β	-0.2412 (0.0000)	-0.3245 (0.0000)	-0.1102 (0.0000)	-0.1984 (0.0000)	-0.1660 (0.0000)	-0.2757 (0.0000)
ρ / λ	–	–	0.5954 (0.0000)	0.4518 (0.0000)	0.6640 (0.0000)	0.5294 (0.0000)
α_0	2.5971 (0.0000)	2.9760 (0.0000)	1.1732 (0.0000)	1.7749 (0.0000)	1.8757 (0.0000)	2.5572 (0.0000)
α_1	–	0.00003 (0.2010)	–	0.00002 (0.4200)	–	0.00002 (0.3020)
α_2	–	0.0204 (0.0000)	–	0.0132 (0.0000)	–	0.0170 (0.0000)
α_3	–	-0.0122 (0.0000)	–	-0.0072 (0.0000)	–	-0.0091 (0.0006)
Moran I	0.4430 (0.0000)	0.3044 (0.0000)	-0.0445 (0.8146)	-0.0035 (0.4954)	-0.0555 (0.8725)	-0.0191 (0.6304)
LR ratio	–	–	109.3400 (0.0000)	63.2340 (0.0000)	88.9480 (0.0000)	43.9070 (0.0000)
LM_{err}	92.2574 (0.0000)	43.5694 (0.0000)	–	–	–	–
LM_{lag}	113.6818 (0.0000)	62.3821 (0.0000)	–	–	–	–
RLM_{err}	1.2502 (0.2635)	1.1563 (0.2822)	–	–	–	–
RML_{lag}	22.6746 (0.0000)	19.9690 (0.0000)	–	–	–	–

The convergence parameter (β) is negative and statistically significant in every estimated model. This confirms our preliminary conclusions based on the scatterplots. Diagnostics for the linear models (absolute and conditional) are unsatisfying. In both classical models spatial residuals autocorrelation appears. Also Lagrange Multipliers inform that spatial dependencies must not be omitted. So, linear models cannot be treated as the final tools to verify the convergence phenomenon hypothesis. Therefore, models with spatial dependencies are proposed (SAR and SE). In the models augmented with connectivity, matrix residuals have better characteristics – spatial autocorrelation is eliminated. Robust Lagrange Multipliers inform that SAR model should be applied.

4.2. Regimes based on composite index

Originally, convergence clubs were supposed to be determined by the values of the composite index. So, regimes presented in Figure 3 consist of regions with similar values of the index (classification based on the mean value and standard deviation of the composite index). Group 1 includes economies with the lowest values of the index, whereas group 4 the ones with the highest values.

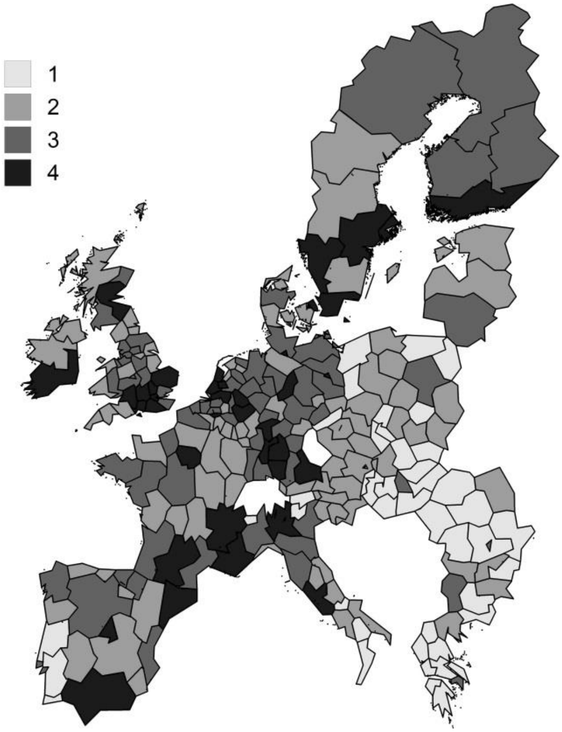


Figure 3. Composite index regimes

One significant issue is strictly connected to the appointed division. For the proposed connectivity matrix (based on a common border) spatial coherence should be maintained. This assumption is not fulfilled as it can be noticed on the map above

(Figure 3). So, for those regimes only the classical models (without spatial dependencies) were estimated.

The results are presented in Table 2. β parameter in every model is negative and statistically significant. Absolute magnitude of convergence parameter in each pair of models is higher for the conditional approach than for the absolute. Although clubs are incoherent, Moran *I* statistics inform that spatial dependencies should be included in estimation, which confirms that none of economies is independent from its neighbours.

Table 2. Estimation and verification of LM models for composite index regimes

	regime I		regime II		regime III		regime IV	
	absolute	conditional	absolute	conditional	absolute	conditional	absolute	conditional
β	-0.3235 (0.0000)	-0.3541 (0.0000)	-0.2465 (0.0000)	-0.3297 (0.0000)	-0.3613 (0.0000)	-0.4080 (0.0000)	-0.0200 (0.7530)	-0.1834 (0.0052)
α_0	3.2615 (0.0000)	3.6430 (0.0000)	2.6413 (0.0000)	3.0098 (0.0000)	3.8196 (0.0000)	3.6780 (0.0000)	0.3495 (0.5930)	1.6050 (0.0086)
α_1	–	0.0003 (0.6131)	–	0.0003 (0.0086)	–	0.00004 (0.4626)	–	0.00005 (0.0641)
α_2	–	0.0114 (0.0326)	–	0.0212 (0.0000)	–	0.0219 (0.0000)	–	0.0167 (0.0001)
α_3	–	-0.0329 (0.0001)	–	-0.0149 (0.0000)	–	-0.0085 (0.0034)	–	-0.0114 (0.0006)
Moran I	0.4086 (0.0051)	0.2327 (0.0579)	0.4409 (0.0002)	0.4017 (0.0008)	0.3587 (0.0019)	0.1811 (0.0621)	0.6617 (0.0004)	0.3264 (0.0428)
LM_{err}	6.8879 (0.0087)	2.2353 (0.1349)	21.5509 (0.0000)	17.8900 (0.0000)	11.8241 (0.0006)	3.0150 (0.0825)	20.1451 (0.0000)	4.9020 (0.0268)
LM_{lag}	2.8715 (0.0902)	0.8292 (0.3625)	11.4128 (0.0007)	3.4171 (0.0645)	7.7959 (0.0052)	2.8343 (0.0923)	11.8122 (0.0000)	6.8703 (0.0088)
RLM_{err}	4.0280 (0.0448)	1.4145 (0.2343)	10.3467 (0.0013)	14.9031 (0.0001)	4.0562 (0.0440)	0.5855 (0.4442)	8.5418 (0.0035)	0.3812 (0.5370)
RML_{lag}	0.0115 (0.9146)	0.0084 (0.9271)	21.7595 (0.0000)	0.4302 (0.5119)	0.0279 (0.8671)	0.4048 (0.5246)	0.2089 (0.6477)	2.3494 (0.1253)

4.3. Spatio-composite regimes

In order to obtain spatial coherence, the modification of the prior division was introduced. As a result three regimes were established, as shown on Figure 4. For those groups of regions the classical models were estimated as well as the spatial models.

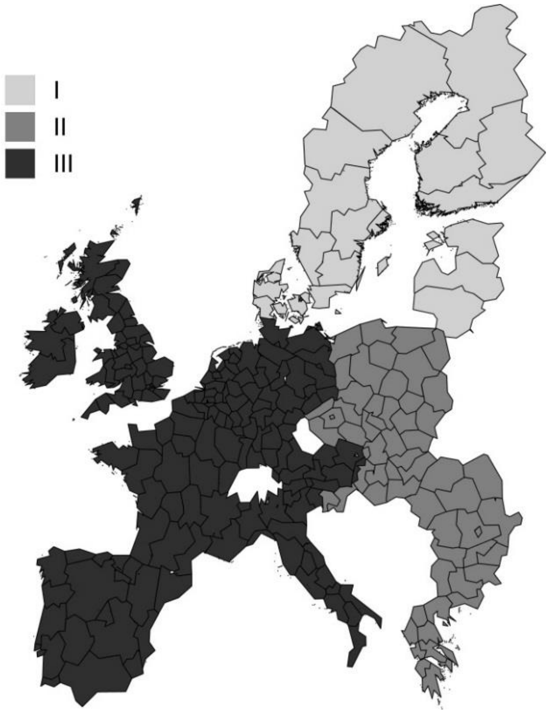
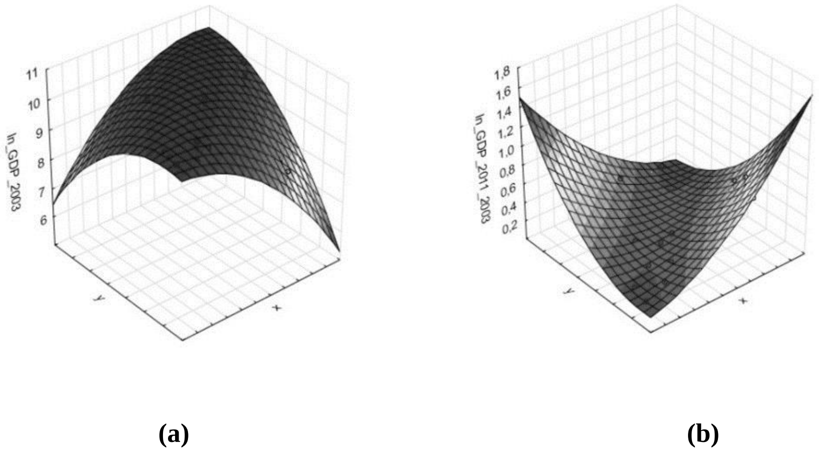


Figure 4. Spatio-composite regimes

For each regime, the analysis of trends and all proposed models was conducted.

The spatial trends of per capita GDP in the initial year and of growth rate in regime 1 are presented in Figure 5. The negative correlation between those two processes is noticeable, therefore it can be preliminary concluded that convergence occurs. The estimated models are included in Table 3.



(a) **(b)**

Figure 5. Spatial trends: (a) per capita GDP in 2003, (b) per capita GDP growth rate 2003-2011 for regime 1

In accordance with the initial conclusion β-convergence occurs in this regime. Moran statistics for linear classical models inform about the non-existence of spatial autocorrelation. It is confirmed by Lagrange Multiplier. A lack of spatial dependencies can be a consequence of regions selected to this regimes and the structure of their neighbourhood. As a result spatial parameters are not statistically significant.

Table 3. Estimation and verification of models for regime 1

	LM		SAR		SE	
	absolute	conditional	absolute	conditional	absolute	conditional
β	-0.2685 (0.0000)	-0.3108 (0.0000)	-0.1740 (0.0000)	-0.2335 (0.0000)	-0.2774 (0.0000)	-0.3106 (0.0000)
ρ / λ	–	–	0.3221 (0.0337)	0.2441 (0.1244)	-0.1876 (0.3647)	0.1043 (0.6131)
α_0	3.0028 (0.0000)	3.2760 (0.0000)	1.9541 (0.0000)	2.5231 (0.0000)	3.0915 (0.0000)	3.2770 (0.0000)
α_1	–	0.00004 (0.2721)	–	0.0001 (0.0673)	–	0.00005 (0.1016)
α_2	–	0.0090 (0.0728)	–	0.0063 (0.1479)	–	0.0091 (0.0317)
α_3	–	-0.0084 (0.1068)	–	-0.0086 (0.0351)	–	-0.0088 (0.0482)
Moran I	-0.1427 (0.6557)	0.0330 (0.3430)	-0.3374 (0.8969)	-0.0417 (0.4796)	-0.0104 (0.4249)	0.0049 (0.3939)
LR ratio	–	–	3.0968 (0.0784)	1.5355 (0.2153)	0.5582 (0.4550)	0.0719 (0.7885)
LM_{err}	0.3612 (0.5479)	0.0193 (0.8895)	–	–	–	–
LM_{lag}	1.9357 (0.1642)	0.9867 (0.3205)	–	–	–	–
RLM_{err}	6.1547 (0.0131)	1.0691 (0.3012)	–	–	–	–
RML_{lag}	7.7292 (0.0054)	2.0365 (0.1536)	–	–	–	–

The analogous analysis was conducted for regime 2. Figure 6 presents trend surfaces. In this case, the negative correlation is also noticeable.

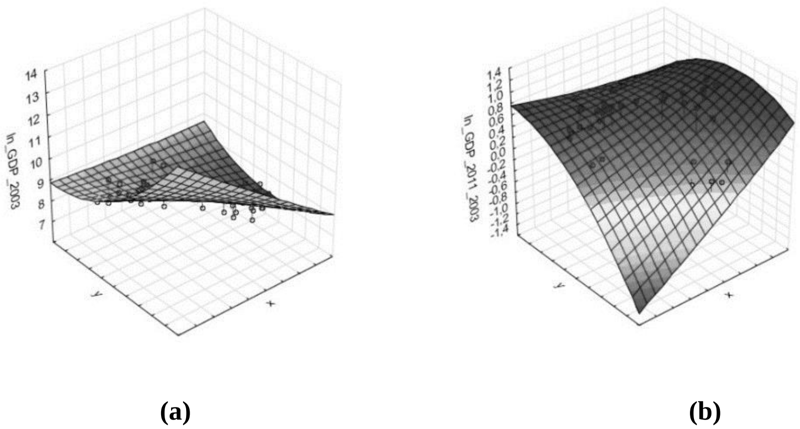


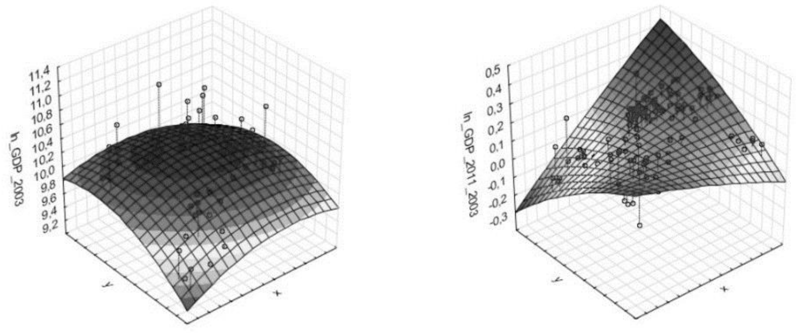
Figure 6. Spatial trends: (a) per capita GDP in 2003, (b) per capita GDP growth rate 2003-2011 for regime 2.

The estimated models confirm the convergence phenomenon. For this regime spatial dependencies are crucial. It is shown by Moran statistics and Lagrange Multiplier for linear models. Parameters ρ / λ are significant in all models. Thanks to the connectivity matrix the spatial autocorrelation is eliminated.

Table 4. Estimation and verification of the models for regime 2

	LM		SAR		SE	
	absolute	conditional	absolute	conditional	absolute	conditional
β	-0.2712 (0.0000)	-0.3331 (0.0000)	-0.1448 (0.0008)	-0.2389 (0.0000)	-0.1824 (0.0014)	-0.3133 (0.0000)
ρ / λ	–	–	0.5166 (0.0000)	0.3408 (0.0080)	0.5383 (0.0000)	0.4399 (0.0014)
α_0	2.8722 (0.0000)	3.0200 (0.0000)	1.5078 (0.0003)	2.0846 (0.0000)	2.1199 (0.0000)	2.8226 (0.0000)
α_1	–	0.00008 (0.8979)	–	0.00002 (0.9686)	–	0.00001 (0.9831)
α_2	–	0.0241 (0.0002)	–	0.0181 (0.0023)	–	0.0232 (0.0010)
α_3	–	-0.0167 (0.0251)	–	-0.0097 (0.1587)	–	-0.0128 (0.1656)
Moran I	0.2481 (0.0017)	0.2156 (0.0051)	-0.0397 (0.5997)	0.0707 (0.1671)	0.0008 (0.4223)	0.0334 (0.2895)
LR ratio	–	–	13.6820 (0.0002)	5.5269 (0.0187)	8.8727 (0.0029)	6.2843 (0.0122)
$\mathrm{LM}_{\mathrm{err}}$	6.9655 (0.0083)	5.2605 (0.0218)	–	–	–	–
$\mathrm{LM}_{\mathrm{lag}}$	13.4603 (0.0002)	5.0877 (0.0241)	–	–	–	–
$\mathrm{RLM}_{\mathrm{err}}$	2.3121 (0.1284)	0.5587 (0.4548)	–	–	–	–
$\mathrm{RML}_{\mathrm{lag}}$	8.8070 (0.0030)	0.3859 (0.5345)	–	–	–	–

The results for regime 3 are not unambiguous. However, for the classical models the convergence parameters are negative and significant, the values of Moran *I* and Lagrange Multiplier suggest that spatial dependencies should be introduced to estimation. After the connectivity matrix has been added to the models, β parameters are positive, which can suggest that divergence can be observed in the established area.



(a) **(b)**

Figure 7. Spatial trends: (a) per capita GDP in 2003, (b) per capita GDP growth rate 2003-2011 for regime 3.

Table 5. Estimation and verification of the models for regime 3

	LM		SAR		SE	
	absolute	conditional	absolute	conditional	absolute	conditional
β	-0.0196 (0.4970)	-0.1540 (0.0000)	0.0105 (0.45890)	-0.0199 (0.3211)	0.0227 (0.1991)	0.0157 (0.4981)
ρ / λ			0.8194 (0.0000)	0.7765 (0.0000)	0.8189 (0.0000)	0.8142 (0.0000)
α_0	0.3434 (0.2380)	1.3970 (0.0000)	-0.0792 (0.5770)	0.1596 (0.3943)	-0.0854 (0.6357)	-0.0553 (0.8010)
α_1	–	-0.00001 (0.6230)	–	-0.00001 (0.5428)	–	-0.00001 (0.5079)
α_2	–	0.0154 (0.0000)	–	0.0033 (0.0065)	–	0.0013 (0.5439)
α_3	–	-0.0105 (0.0000)	–	-0.0017 (0.0753)	–	0.00001 (0.9958)
Moran I	0.8155 (0.0000)	0.5967 (0.0000)	-0.0517 (0.8024)	-0.0584 (0.8351)	-0.0501 (0.7938)	-0.0513 (0.7997)
LR ratio	–	–	200.3200 (0.0000)	153.7800 (0.0000)	201.4200 (0.0000)	149.35 (0.0000)
LM$_{err}$	222.2529 (0.0000)	118.9891 (0.0000)	–	–	–	–
LM$_{lag}$	224.2186 (0.0000)	155.5310 (0.0000)	–	–	–	–
RLM$_{err}$	1.7846 (0.1816)	1.9274 (0.1650)	–	–	–	–
RML$_{lag}$	3.7503 (0.0528)	38.4694 (0.0000)	–	–	–	–

For the purpose of comparing the prior results, the speed of convergence and half-life for the whole area were calculated, and the regimes established by the spatio-composite approach. Those measures are presented in Table 6. Excluding regime 3 some successive conclusions can be deduced. The values of half-life are lower for the conditional models than for the absolute ones – because they include specific region characteristics. In general, the time of reducing the differences of per capita GDP is longer in models containing spatial dependencies.

Table 6. Speed and half-life for whole area and established regimes

	LM		SAR		SEM	
	249 regions of the EU					
	absolute	conditional	absolute	conditional	absolute	conditional
β	-0.2412	-0.3245	-0.1102	-0.1984	-0.1660	-0.2757
speed	0.0307	0.0436	0.0130	0.0246	0.0202	0.0358
half-life	22.60	15.90	53.43	28.21	34.37	19.34
	Regime I					
β	-0.2685	-0.3108	-0.1740	-0.2335	-0.2774	-0.3106
speed	0.0347	0.0414	0.0212	0.0295	0.0361	0.0413
half-life	19.95	16.76	32.63	23.46	19.20	16.77
	Regime II					
β	-0.2712	-0.3331	-0.1448	-0.2389	-0.1824	-0.3133
speed	0.0352	0.0450	0.0174	0.0303	0.0224	0.0418
half-life	19.72	15.40	39.88	22.85	30.98	16.60
	Regime III					
β	-0.0196	-0.1540	0.0105	-0.0199	0.0227	0.0157
speed	0.0022	0.0186	–	0.0022	–	–
half-life	315.15	37.30	–	310.35	–	–

5. Conclusions

As it has been shown, the trend surface analysis is helpful for initial conclusions if convergence occurs. The convergence phenomenon is strictly connected to the time and the spatial range of the investigation – for the presented area and regimes the convergence hypothesis was not confirmed in every case. However, it is noticeable that the speed of convergence is higher for clubs of regions than for the whole selected area. Regimes are more homogenous groups (except for club 3), therefore the equalization of the economic development can be of a faster pace. The spatial dependencies were not significant in each of the cases considered, but omitting them could result in spatial autocorrelation of residuals. The time of reducing the differences in the levels of GDP is shorter for the non-spatial models.

The direction for further investigations: dividing regime 3 because of its heterogeneity and diversity of the composite index; analyses with panel data models – changes through investigated period will be exploited; introducing individual fixed effects or Spatial Durbin Model (proposed solutions for omitted variables).

REFERENCES

ABREU, M. H., DE GROOT, L. F., FLORAX, R. J. G. M., (2005). Space and Growth: A Survey of Empirical Evidence and Methods, Région et Développment, Vol. 21, pp. 13−14.

BARRO, R. J., SALA-I-MARTIN, X., (1995). Economic Growth, New York: McGraw-Hill,.

CAURESMA, J. C., FELDKIRCHER, M., (2015).,Spatial Filtering, Model Uncertainty and the Speed of Convergence in Europe, Journal of Applied Econometrics, Vol. 28, Issue 4, pp. 720–741.

ELHORST, P., PIRAS, G., ARBIA, G., (2010). Growth and Convergence in a Multiregional Model with Space-Time Dynamics, Geographical Analysis, Vol. 42, pp. 338–355.

ERTUR, C., KOCH, W., (2007). Growth, Technological Interdependence and Spatial Externalities: Theory and Evidence, Journal of Applied Econometrics, Vol. 22, pp. 1033−1062.

FINGLETON, B., (2001). Equilibrium and economic growth: spatial econometric models and simulations, Journal of Regional Science, Vol. 41, No. 1, pp. 117−147.

FINGELTON, B., (2004). Regional Economic Growth and Convergence: Insight from a Spatial Econometric Perspective, in: L. Anselin, R. J. G. M. Florax, S. J. Rey (eds.), Advanced in Spatial Econometrics. Methodology, Tools and Applications, Springer, pp. 397−432.

FINGLETON, B., LÓPEZ-BAZO, E., (2006). Empirical growth models with spatial effects, Regional Science, Vol. 85, No. 2, pp. 177–198.

GÓRNA, J., GÓRNA, K., SZULC, E., (2014). A β-convergence Analysis of European Regions. Some Re-specification of the Traditional Model, Models and Methods for Analysing and Forecasting Economic Processes. Theory and Practice (edited by Józef Pociecha), Cracow, pp. 83−97.

LE GALLO, J., ERTUR, C., BAUMONT, C., (2003). A spatial econometric analysis of convergence across European Regions, 1980-1995. In B. Fingleton (ed.), European Regional Growth. Advances in Spatial Science. Springer-Verlag, Berlin Heidelberg, pp. 99–129.

MANKIW, N. G., ROMER, D., WEIL, D., (1992). A Contribution to the Empirics of Economic Growth, Quarterly Journal of Economics, Vol. 107, pp. 407−437.

REY, S. J., JANIKAS, (2005). Regional Convergence, Inequality, and Space, Journal of Economic Geography, Vol. 5, pp. 155−176.

REY, S. J., LE GALLO, J., (2009). Spatial analysis of economic convergence, in Mills T. C., Patterson K. (eds.), Palgrave Handbook of Econometrics, Volume II: Applied Econometrics, Palgrave MacMillan, New York, pp. 1251−1290.

STATISTICS IN TRANSITION new series, Summer 2015
Vol. 16, No. 2, pp. 279–292

APPLICATION OF MULTIFACTORIAL MARKET-TIMING MODELS TO ASSESS RISK AND EFFECTIVENESS OF EQUITY-LINKED INSURANCE FUNDS IN POLAND

Magdalena Homa[1], **Monika Mościbrodzka**[2]

Abstract

Traditionally, models developed by Treynor and Mazuy (T-M) and also by Henriksson-Merton (H-M), which are called market-timing models, are applied to assess effectiveness of investment funds. The objective of the presented study is an application of the T-M and H-M models and their T-M-FF and H-M-FF modifications with additional Fama-French factors to assess effectiveness and risk of equity insurance connected with unit-linked insurance. Estimation and verification of the models for the subject group of equity funds were performed and the significance of the impact of particular factors on returns on reference portfolios was discussed.

Key words: market-timing model, Fama-French factor, equity funds.

1. Introduction

According to a classical capital asset pricing model (CAPM) the skill of managers described as microforecasting is assessed. Such skill covers identification of single assets, which are undervalued or overvalued compared to assets in general at a particular market situation. The fund manager will possess it if during the selection of securities to a portfolio s/he considers risk analysis characteristic for particular securities, not focusing only on the risk of the entire market at the same time. Market timing, on the other hand, is understood the skill to forecast short-term increases or inclines in security prices and proper responding to such changes. A proper response of an investor, who uses market-timing techniques, ought to assure proper proportions within the investment portfolio between risk and safe assets in such a manner to obtain a higher level of portfolio risk during the increase periods and a lower risk level on declining markets. In this case we assess the investor's skill, that is a proper forecasting with the difference that it concerns

[1] Uniwersytet Wrocławski Instytut Nauk Ekonomicznych. E-mail: homam@prawo.uni.wroc.pl.
[2] Uniwersytet Wrocławski Instytut Nauk Ekonomicznych. E-mail: m.moscibrodzka@prawo.uni.wroc.pl.

movements of the entire market. The introduction of additional variables, the so-called Fama-French factors (Fama, French, 1996), was proposed as far as classical market-timing models are concerned. The task of such variables was to explain the part of inaccurate indications in a classical capital asset pricing model arising from the property of fundamental companies.

The utility of such models was investigated on the example of the Polish stock market. Attempts to apply a three-factor model for the Polish market was made, among others, by Kowerski (Kowerski 2008), the utility of such a model was verified by Czapkiewicz and Skalna in 2011. Within the scope of selectivity of assets and application of market-timing techniques to investment funds, the models were applied by Olbryś, and also when taking into consideration the construction of the Fama-French factors, she verified hybrid market-timing models along with the assessment of the skill to manage equity investment funds management and stability of parameters (Olbryś 2008a,b,c, 2009, 2011a and Mościbrodzka 2014).

The application of classical and hybrid multifactorial market-timing models to assess the risk and effectiveness of unit-linked insurance (UFK) are proposed in the thesis. Their utility was verified and it was investigated whether managers of unit-linked insurance in Poland possess skills within the scope of:

- forecasting of price changes of single assets, that is selectivity of securities,
- forecasting of changes in the market globally, that is, changes of a market factor (application of market-timing techniques).

Thus, it was demonstrated that market-timing models may constitute a new and supportive tool allowing the insured to make a proper decision concerning investment strategy of resources into specific capital funds.

2. The essence of unit-linked insurance

The unit-linked insurance is a product of hybrid character, the structure of which is based on classical life insurance or pure endowment insurance with investment into selected insurance capital funds. So, this is a contract between the insured and the insurer according to which the insured pays premiums and the insurer in return assures a benefit in the amount equal or greater than the value of:

- G_{Π} - the guaranteed amount,
- $b(S_t)$ - the amount arising from the value of a reference portfolio dependent on the determination of fund price.

So, the unit-link insurance differs fundamentally from classical life insurance and pure endowment insurance in that it is related to investment of resources coming from premiums into segregated funds. In contrast to traditional life insurance within unit-linked insurance, not only payment of the benefit is random, but also is the amount of the benefit paid. At the moment of occurrence of the event the insurer will pay the insured sum which is equal: a minimum of the guaranteed amount insured or the market value of the insurance portfolio. So, the payment is dependent on a development of a certain index or the value of a certain specified insurance portfolio. In accordance with such construction of the unit-linked insurance the payment (future benefits) with the guaranteed insurance sum is equal to:

$$b(t) = \max\{X(t), g(t)\} = g(t) + \max\{0, X(t) - g(t)\} \tag{1}$$

where $X(t) -$ the value of insurance portfolio (a reference one) at the *t*-moment,

$g(t) -$ the guaranteed sum at the *t*-moment.

So, the value of the insurance portfolio connected with insurance is random and it is a proper function of accumulated investment dependant on unit prices of a selected fund arising from the strategy accepted by the insured. The unit-linked insurance contracts, which are offered in Poland, are the products allowing the insured to accumulate savings in insurance capital funds, which are run by external societies independently of investment funds. An additional value of policies of this type is not only the possibility of selection of different funds out of a wide market offer, but also the fact that they are managed by different companies. So, such funds differ within the scope of risk and investment policy, which from the viewpoint of the investment risk diversification are of high significance. Since unit-linked insurance policies are transparent and of open structure, it gives the insured the opportunity to systematically adjust investment strategy depending on a changeable market situation, influencing the final amount of payment at the same time. So, when deciding about the choice of the capital fund, the insured takes responsibility for possible negative consequences of his decisions and is encumbered with financial risk (Homa, 2013).

3. Multifactorial market-timing models

3.1. The CAPM model

The *Capital Asset Pricing Model (CAPM*) warrants explanation of the achieved rates of return on securities as market risk function (Reilly, Brown, 2001). This is founded on an assumption that the formation of rates of return of shares is determined by a factor that reflects changes on the capital market. The equation of such a model can be written in the following form:

$$r_{i,t} = \alpha + \beta \cdot r_{M,t} + \varepsilon_{i,t}, \tag{2}$$

where $r_{i,t}$ - vector of excess rates of return of portfolio at the *t*-moment over a risk-free rate,

$r_{M,t}$ - denotes an excess rate of return on the market index at the t-moment over a risk-free rate.

In practice, we most frequently assume that a risk-free rate is a profitability index of treasury bills or inter-bank market rate (e.g. WIBOR) (Jajuga, Jajuga 2006). Yet, one ought to remember that even treasuries are not riskless, so when speaking of a risk-free rate one ought to mean the rate that is accompanied by the lowest risk possible at a particular time among different classes of financial assets.

The idea of the *CAPM* model is based on the thesis that an additional rate of return ought to arise from the selection of securities, i.e. during the selection of securities to a portfolio the fund manager will consider risk analysis characteristic for particular securities for particular papers not focusing only on the risk of the entire market. So, a positive and significant α parameter means that the manager makes attempts of a detailed market analysis and his expectations of price behaviour of particular securities are accurate. A market portfolio is of key significance for the investment value within the capital asset pricing model. This is a portfolio which consists of all shares and other securities of positive risk occurring on the market and contribution of particular shares in such a portfolio are equal to the contribution of such shares on the market. So, the beta coefficient within this model is treated as a risk measure showing approximately by how many units a rate of return on portfolio will increase if a rate of return of the market index will increase by one unit (Jajuga, Jajuga, 2006). When making the decision concerning the selection of securities to a portfolio the investor is often influenced by the value of the beta coefficient as the value of premium for the risk of the capital involved.

3.2. Classical models of market timing

The market-timing idea concerns identification of market trends and the manager possessing such skills will adapt the composition of the fund managed to a market situation. So, in order to test the skills of the portfolio manager within the scope of the so-called market timing, we should adopt classical parametric market-timing models, with occurring variable represents the market. In practice, most frequently this is a rate of return on market portfolio where a corresponding stock exchange index or excess rate of return on market portfolio over a risk-free rate is a substitute. Classical market-timing models include:

- the Treynor-Mazuy model (T-M),
- the Henriksson-Merton model (H-M).

Both the model developed by Treynor and Mazuy and also the one developed by Henriksson and Merton serve to test the skills within the scope of application of market-timing techniques and selectivity of assets by managers of investment portfolios. They described market sensing as a proper reaction to changes of rate of return on the stock exchange index, but they defined it in a different manner. The regression model (Treynor, Mazuy, 1966), denoted as the T-M model, has the following form:

$$r_{i,t} = \alpha + \beta_1 \cdot r_{M,t} + \beta_2 \cdot r_{M,t}{}^2 + \varepsilon_{i,t} \tag{3}$$

while the model proposed by Henriksson and Merton (1981) denoted as the H-M model, is:

$$r_{i,t} = \alpha + \beta_1 \cdot r_{M,t} + \beta_2 \cdot \max[0; -r_{M,t}] + \varepsilon_{i,t} \tag{4}$$

In both models the β_2 parameter shows the skills of application of market-timing techniques (short-term market trends), the value of which constitutes the adjustment by possible pessimistic expectations of the fund manager as far as future formation of market rate is concerned. If it assumes values greater than zero, then the portfolio managers accurately forecast market movements and the value of such a coefficient shows the level of such a skill. The authors of the model recommend applying it both during the periods of a huge increase and the periods of heavy decrease in stock exchange indices because reactions on small market movements with the use of such a model are not observable. If the β_2 coefficient is close to zero, then the investor does not show any prognostic capabilities concerning the market. A significant fault of a parametric version of the Henriksson-Merton model is the assumption concerning permanency at the time of probability of an accurate forecast, which cannot be fulfilled if an investor forecasts the movements of bigger markets than the movements of smaller markets more easily (Czekaj et al. 2000). A significantly negative value of the β_2 parameter estimator denotes a negative impact of the market-timing technique on the portfolio value.

3.3. Hybrid market-timing models

The T-M and H-M models underwent further modifications arising from the fact that their incorrect indications within the scope of explanation of effective interest rate differential were observed. It was shown that balance indices such as a book/market value and a company's size impact the value of the expected value of a rate of return on equity portfolio (Bhandari, 1988). So, classical market-timing models were extended by additional factors on the basis of which the managers make allocation decisions. Fama and French in their theses (Fama, French, 1992, 1993) investigated monthly rates of return of American companies from 1963 to 1991 quoted on NYSE, AMEX and NASDAQ (since 1972). In the first step they divided such companies in terms of their capitalisation volume into groups of companies above and under median, creating portfolios of big companies (B-Big) and small ones (S-Small). Another criterion of the division of companies was the size of the BV/MV index, that is a quotient of a book value to a company's market value. From an investigated sample, Fama and French sectioned off two groups of companies: companies with growth potential and companies with value potential. The first ones are the companies of a low BV/MV index and within this group market estimate considerably exceeds balance value, which proves that the investors expect very good results and increase in assets from such companies in the future. The companies with value potential are characteristic of a high BV/MV ratio. All the investigated companies were divided into three groups. Namely 30% of companies of the lowest index value within population were counted among the groups of companies with growth potential and created the Low (L) portfolio, 30% of companies of the greatest index value were counted among the groups of companies with value potential and created the High (H) portfolio, the remaining 40% of companies landed in the Medium (M) portfolio. After such division the authors constructed 6 portfolios which are a cross section of a group of sets of big and small companies, and the ones with low and high BV/BM: BL, BM, BH, SL, SM, SH index. The portfolios, which were created according to the above procedure, were used to calculate values of variables within the Fama-French model: *SMB* (Small-minus-Big) and *HML* (High-minus-Low). Namely, the *SMB* factor constituted arithmetic mean of differences between returns on a portfolio of small companies (SL,SM,SH) and big companies (BL,BM,BH), while *HML* constituted arithmetic mean of differences between returns on companies' portfolios with value potential (SH, BH) and growth potential (SL,BL), that is,

$$SMB_t = \tfrac{1}{3} \cdot \left(R_{SL,t} + R_{SM,t} + R_{SH,t} - R_{BL,t} - R_{BM,t} - R_{BH,t} \right) \tag{5}$$

$$HML_t = \tfrac{1}{2} \cdot \left(R_{SH,t} + R_{BH,t} - R_{SL,t} - R_{BL,t} \right) \tag{6}$$

where R denotes weighted average with capitalisation of a rate of return on a corresponding portfolio of the companies.

As a result, the two FF factors were considered within models and in such manner the hybrid models (T-M-FF) and (H-M-FF) were obtained respectively in the following form:

$$r_{i,t} = \alpha + \beta_1 \cdot r_{M,t} + \beta_{SMB} \cdot r_{SMB,t} + \beta_{HML} \cdot r_{HML,t} + \beta_2 \cdot r_{M,t}{}^2 + \varepsilon_{i,t}, \tag{7}$$

$$r_{i,t} = \alpha + \beta_1 \cdot r_{M,t} + \beta_{SMB} \cdot r_{SMB,t} + \beta_{HML} \cdot r_{HML,t} + \beta_2 \cdot \max[0, -r_{M,t}] + \varepsilon_{i,t}, \tag{8}$$

where $r_{SMB,t}$ is an excess rate of return on a portfolio simulating *SMB* over a risk-free rate of return in the *t*-period,

$r_{HML,t}$ is an excess rate of return on a portfolio simulating *HML* over a risk-free rate of return in the *t*-period.

The β_{SMB} and β_{HML} coefficients are measures of sensitivity of a rate of return on investment to changes of rates of return on portfolios simulating respectively *SMB* and *HML*. So, their loads constitute the additional premium for the risk connected with investment in companies of low capitalisation and the ones of a high value of the balance index respectively, which is a quotient of a book value to company's market value.

4. The results of empirical research

4.1. The construction of the Fama-French factors

The investigation covered the period from March 2009 to October 2014 and selected 34 insurance capital funds[3]. The analysis was based on weekly data coming from the period considered. The Insurance Capital Funds were divided into 6 groups: balanced mixed, absolute rate mixed, stable growth mixed, active allocation mixed, shares of small and medium-sized companies (SMEs) and shares. Such division was dictated by a suggested division of funds by insurance companies according to investment risk. Any fundamental data and company quotations

[3] Availability of data determined the selection of unit-linked insurance.

present on the Warsaw Stock Exchange in Warsaw were taken from the Warsaw Stock Exchange Statistical Bulletins and from the stooq.pl portal, data concerning the Insurance Capital Funds was taken from their placing memoranda, reports and also fund cards and from web pages of particular funds.

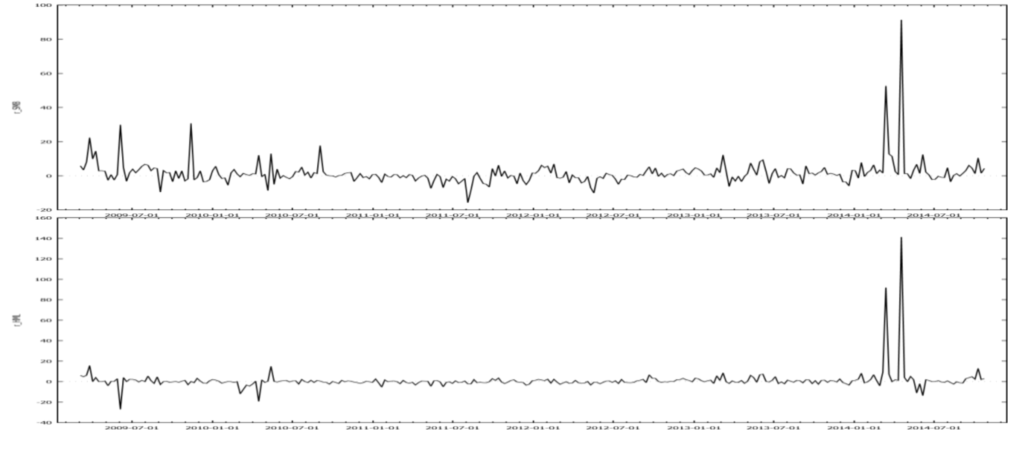
In the first stage of the research coefficient values of the so-called portfolios stimulating SMB and HML were determined. Monthly balance data covering the period from February 2009 until October 2014 was used to construct the Fama-French factors. The research considered quotations of all the companies listed on the Stock Exchange Market, the number of which during the period covered by the research amounted to the level of 266 companies at the end of February 2009 and 463 in October 2014. The construction of factors was achieved analogically as in the Fama-French thesis in the following manner:

STEP 1. At the end of each month the companies were divided into six disjunctive portfolios: BH,BM,BL and SH,SM,SL according to their balance indices.

STEP 2. Such division was maintained during a consecutive month and the *SMB* and *HML* factors were determined during the week of a particular month. Namely, the difference between a rate of return on portfolios of big companies (BL, BM, BH) and a rate of return on portfolios of small companies (SL, SM, SH) was the basis for creation of the *SMB* factor, according to the formula (5), whereas the difference between a rate of return on portfolios of companies with value potential (BH, SH) and a rate of return on portfolios of companies with growth potential (BL, SL) according to the formula (6), served to create the *HML* factor.

STEP 3. Determined *SMB* and *HML* factors were reduced by a risk-free rate.

Formation of the factors within the Fama-French models from the beginning of 2009 until the end of October 2014 was depicted in the drawings below.



Drawing 1. Cleaned Fama-French factors: SMB and HML

Source: own elaboration.

4.2. Risk assessment and selectivity of assets

In the next stage the assessment of unit-linked insurance was carried out as regards the risk and skills of the managers within the scope of forecasting of price behaviour of particular securities, namely the so-called selectivity of assets. Depending on the type of the fund investment zone and its benchmark as a market rate of return, the following were considered respectively: the Warsaw Stock Exchange Index (WIG), a rate on index of small and medium enterprises - InvestorMS or a corresponding benchmark of the WIG rate with the rate of treasury bond market TBSP.Index (Treasury BondSpot Poland). Due to the fact that from 2012 the issue of treasury bonds was stopped, the profitability of which was most frequently indicated as a risk-free rate, WIBOR, which is an inter-bank rate, is assumed as a risk-free rate of interest (see Jajuga, Jajuga, 2006, p.224). The results of estimation of the least squares method for the CAPM model for unit-linked insurance in Poland (ULIP) are presented in Table 1.

Table 1. The estimation results of the CAPM parameters for ULIP

BALANCED MIXED	**α**	**β**	**R^2**
UniKorona Balanced	0.0762	0.6324***	0.8210
Active Investing	0.0661	0.5924***	0.7319
Balanced Portfolio	0.1423*	0.4820***	0.5242
AXA of Business Cycle	0.0655	0.6516***	0.7216
AXA Mixed	0.0860	0.4016***	0.5323
ABSOLUTE RATE MIXED	**α**	**β**	**R^2**
Noble Fund Global	0.0411	0.5352***	0.5305
Quercus Selective	0.1749***	0.4622***	0.5241
STABLE GROWTH MIXED	**α**	**β**	**R^2**
PKO Stable Growth	0.1019*	0.4705***	0.6786
Stable Growth	0.1049*	0.4538***	0.6105
AXA Optimal	0.1115*	0.4543***	0.6117
Stable Growth Portfolio	0.1574**	0.3920***	0.4694
AXA Stable Growth	0.1317*	0.5287***	0.6409
ACTIVE ALLOCATION MIXED	**α**	**β**	**R^2**
Legg Mason of Strategy	0.0079	0.5959***	0.8000
Noble Fund Timing	0.0238	0.7644***	0.7406
SMES SHARES	**α**	**β**	**R^2**
UniAkcje SM	0.0014	0.6782***	0.4579
Noble Fund Stock SM	0.0929	0.8191***	0.6491
SHARES	**α**	**β**	**R^2**
Uni Korona Stock	0.0375	0.8682***	0.8559
Skarbiec Stock	-0.0899*	0.9041***	0.9112
PKO Stock	0.0013	0.7731***	0.9131
Investor Stock BM	-0.0129	0.8224***	0.7337
Legg Mason Stock	0.0157	0.8260***	0.9418
Stock	0.0330	0.8281***	0.7652
Noble Fund Stock	-0.0005	0.8927***	0.9489
AXA Stock Portfolio	0.1175	0.6417***	0.5072
Quercus Agrressive	0.0639	0.8123***	0.7437
AXA Stock BM	0.0186	0.8666***	0.7422
AXA Stock	0.0222	0.8238***	0.7345

*** - relevance at a level of 0.01 ** - relevance at a level of 0.05 * - relevance at a level of 0.10

Source: own elaboration.

One ought to focus on the fact that within all the fund groups, the parameter which is liable for systematic risk was statistically significant and its value explicitly indicates that stable growth funds are distinguished by the lowest risk. The group of share funds was the group of the greatest risk was. However, it should be emphasised that none of the funds within this group belonged to the so-called groups of aggressive funds. From among all the discussed groups, only one fund of a significantly negative α parameter occurred in this group, which implies that the manager of such fund randomly selected assets to a portfolio. But it is worth to take into consideration the fact that this fund was the most risky and share funds are burdened with the greatest financial risk. In the case of the remaining groups, the results are not so explicit any more. When assessing the skills of managers within the scope of selection of assets, the only group where the managers of all the funds took risk analysis characteristic for particular securities, not only focusing on the risk of the entire market and accurately selected assets to the fund, is the group of stable growth funds. Within the remaining groups, only in few situations the parameter concerning assessment of asset selectiveness was positive and statistically significant. The greatest surprise was the results of the active allocation fund group, whose managers ought to accurately forecast price behaviour of particular securities. However, according to Jensen's interpretation (Jensen 1972), a positive but statistically unimportant estimator value of such a parameter may be the result of a positive load of estimator and not necessarily it reflects the skills of the portfolio manager. Within the group of share funds, only in one fund of the greatest risk a significant but negative value of coefficient was observed, which means that the manager of this fund randomly allocates the funds into instruments offered on the market.

4.3. Assessment of abilities within the scope of application of market-timing techniques

In the next stage of research the assessment of skills of the managers of unit-linked insurance within the scope of application of market-timing techniques was made and parameters of classical T-M and H-M[4] market–timing models were estimated. The results of the Treynor-Mazuy model estimation are presented in Table 2.

Table 2. The estimation results of parameters of the Treynor-Mazuy model for ULIP

BALANCED MIXED	α	β_1	β_2	R^2
UniKorona Balanced	0.0073	0.6191***	0.0081***	0.8256
Active Investing	-0.0222	0.5754***	0.0104***	0.7397
Balanced Portfolio	-0.0522	0.4445***	0.0230***	0.5688
AXA of Business Cycle	-0.0744	0.6441***	0.0195***	0.7449
AXA Mixed	-0.0808	0.3694***	0.0197***	0.5805

[4] Due to restrictions connected with the number of pages and the fact that the results of the H-M estimation model were analogous to the results of the T-M model, the results for the H-M model are not included in the thesis.

Table 2. The estimation results of parameters of the Treynor-Mazuy model for ULIP (cont.)

ABSOLUTE RATE MIXED	**α**	**β_1**	**β_2**	**R^2**
Noble Fund Global	-0.0782	0.5287***	0.0166***	0.5502
Quercus Selective	-0.0137	0.4520***	0.0262***	0.5845
STABLE GROWTH MIXED	**α**	**β_1**	**β_2**	**R^2**
PKO Stable Growth	-3.48E-06	0.4509***	0.0120***	0.6947
Stable Growth	-0.0137	0.4310***	0.0140***	0.6317
AXA Optimal	-0.0060	0.4316***	0.0139***	0.6326
Stable Growth Portfolio	-0.0452	0.3529***	0.0239***	0.5356
AXA Stable Growth	-0.0570	0.5185***	0.0262***	0.6968
ACTIVE ALLOCATION MIXED	**α**	**β_1**	**β_2**	**R^2**
Legg Mason of Strategy	-0.0316	0.5883***	0.0047*	0.8012
Noble Fund Timing	-0.0713	0.7592***	0.0132**	0.7494
SMES SHARES	**α**	**β_1**	**β_2**	**R^2**
UniAkcje SM	-0.1352	0.6708***	0.0190**	0.4680
Noble Fund Stock SM	0.0062	0.8144***	0.0121*	0.6555
SHARES	**α**	**β_1**	**β_2**	**R^2**
Uni Korona Stock	0.0551	0.8716**	-0.0021	0.8555
Skarbiec Stock	-0.0459	0.9126***	-0.0052**	0.9121
PKO Stock	0.0144	0.7756***	-0.0015	0.9130
Investor Stock BM	-0.0162	0.8218***	0.0004	0.7328
Legg Mason Stock	0.0285	0.8285***	-0.0015	0.9418
Stock	-0.0365	0.8146***	0.0082*	0.7673
Noble Fund Stock	0.0366	0.8999***	-0.0044**	0.9496
AXA Stock Portfolio	-0.0699	0.6056***	0.0221**	0.5290
Quercus Agrressive	-0.0238	0.8075***	0.0122**	0.7507
AXA Stock BM	-0.0742	0.8615***	0.0129**	0.7490
AXA Stock	-0.0621	0.8192***	0.0117**	0.7408

*** - relevance at a level of 0,01 ** - relevance at a level of 0,05 * - relevance at a level of 0,1

Source: own elaboration.

The obtained results show that the majority of the managers of the unit-linked insurance possess the skill to use short-term market trends. In all the analysed groups, except the group of share insurance funds, a significant and positive value of the coefficient means that the managers try to adapt the composition of a managed fund to a current market situation and appropriately react by increasing or reducing the fund's exposure to market risk, for instance, through decreasing or increasing contribution of security instruments such as bonds or treasury bills. Simultaneously, a positive value of the coefficient indicates that managers benefit from their own expectations as far as changes of a market rate of return in the future are concerned, although to a different extent, which is inversely proportional to the risk of a systematic portfolio.

In the funds of shares of the greatest risk, a significant but negative value of the β_2 coefficient was recorded, which means that the application of market-timing strategies by managers in this particular case has a negative influence on a rate of return of such funds.

4.3. Assessment of premium for the risk of investment into aggravated risk companies

In the last stage of the analysis it was investigated whether the managers of unit-linked insurance make allocation decisions taking into consideration publicised additional information and the ones arising from fundamental property of companies. In order to assess the impact of balance factors on the value of the funds, the parameters of hybrid models were estimated, which considered variables introduced by Fama and French based on the following theses:

1. The shares of companies with a small capitalisation are more risky than the shares of the companies with a high capitalisation (SMB).
2. Companies with value potential are more risky than the companies with growth potential (HML).

The parameters of the T-M-FF and H-M-FF hybrid models were estimated and results for the T-M-FF[5] model are shown in Table 3.

Table 3. The estimation results of the T-M-FF model parameters for the unit-linked insurance in Poland (ULIP)

BALANCED MIXED	α	β_1	β_2	β_{SMB}	β_{HML}	R^2
UniKorona Balanced	0.0169	0.6450***	0.0077***	-0.0255**	0.0185**	0.8271
Active Investing	-0.0223	0.5835***	0.0101***	-0.0076	0.0093	0.7388
Balanced Portfolio	-0.0468	0.4690***	0.0224***	-0.0236	0.0217	0.5698
AXA of Business Cycle	-0.0713	0.6517***	0.0192***	-0.0090	0.0053	0.7451
AXA Mixed	-0.0685	0.4160***	0.0187***	-0.0453***	0.0390***	0.5960
ABSOLUTE RATE MIXED	α	β_1	β_2	β_{SMB}	β_{HML}	R^2
Noble Fund Global	-0.0693	0.5856***	0.0138**	-0.0650*	0.0499**	0.5600
Quercus Selective	-0.0167	0.4686***	0.0252***	-0.0182	0.0187	0.5877
STABLE GROWTH MIXED	α	β_1	β_2	β_{SMB}	β_{HML}	R^2
PKO Stable Growth	0.0130	0.5019***	0.0109***	-0.0495***	0.0432***	0.7136
Stable Growth	-0.0025	0.4725***	0.0131***	-0.0405***	0.0345***	0.6420
AXA Optimal	0.0050	0.4730***	0.0130***	-0.0402***	0.0345***	0.6430
Stable Growth Portfolio	-0.0338	0.3980***	0.0229***	-0.0437**	0.0383***	0.5489
AXA Stable Growth	-0.0497	0.5575***	0.0243***	-0.0447*	0.0334*	0.7021
ACTIVE ALLO-CATION MIXED	α	β_1	β_2	β_{SMB}	β_{HML}	R^2
Legg Mason of Strategy	-1.63E-02	0.6379***	0.0037	-0.0485***	0.0390***	0.8124
Noble Fund Timing	-0.0602	0.7576***	0.0137**	0.0003	-0.0100	0.7517

[5] Due to restrictions connected with the number of pages and the fact that the results of the H-M-FF estimation model were analogous to the results of the T-M-FF model, the results for the H-M-FF model are not included in the thesis.

Table 3. The estimation results of the T-M-FF model parameters for the unit-linked insurance in Poland (ULIP) (cont.)

SMES SHARES	α	β_1	β_2	β_{SMB}	β_{HML}	R^2
UniAkcje SM	-0.1517	0.5937***	0.0227***	0.0886*	-0.0642*	0.4812
Noble Fund Stock SM	-0.0017	0.7274***	0.0166**	0.0986**	-0.0806***	0.6702
SHARES	α	β_1	β_2	β_{SMB}	β_{HML}	R^2
Uni Korona Stock	0.0541	0.8628***	-0.0018	0.0084	-0.0089	0.8549
Skarbiec Stock	-0.0352	0.9217***	-0.0050*	-0.0100	-0.0019	0.9124
PKO Stock	0.0218	0.8021***	-0.0021	-0.0258**	0.0216***	0.9150
Investor Stock BM	-0.0287	0.8351***	-0.0007	-0.0107	0.0293*	0.7381
Legg Mason Stock	0.0321	0.8422***	-0.0018	-0.0133	0.0116*	0.9420
Stock	-0.0591	0.7608***	0.0090**	0.0535***	-0.0355**	0.7716
Noble Fund Stock	0.0381	0.8982***	-0.0043**	0.0013	-0.0036	0.9494
AXA Stock Portfolio	-0.0770	0.5870***	0.0224***	0.0183	-0.0129	0.5266
Quercus Agrressive	-0.0349	0.7473***	0.0151***	0.0690**	-0.0515**	0.7569
AXA Stock BM	-0.0732	0.8317***	0.0146**	0.0331	-0.0304	0.7517
AXA Stock	-0.0662	0.7878***	0.0133**	0.0358	-0.0281	0.7427

*** - relevance at a level of 0,01 ** - relevance at a level of 0,05 * - relevance at a level of 0,1

Source: own elaboration.

The results of estimation of the β_{SMB} and β_{HML} parameters show that additional factors such as capitalisation and market value index to book value are significant in the case of all the funds from the group of stable growth funds and shares of small and medium-sized companies, i.e. in a significant manner they determine their rate of return but investment risk into other companies is compensated. In the case of stable growth funds, a positive value of the β_{HML} coefficient proves that investment risk of companies with value potential is compensated, while in the case of share funds of SMEs a positive value of the β_{SMB} estimator proves that investment risk of low capitalisation companies was compensated by the additional premium. For remaining groups which are considered situation is not clear. It is observed unit-linked insurance for which considering of additional balance factors does not have a significant impact on the value of the rate of return or else only one of considered factors had a significant impact on its rate of return.

SUMMARY

In this thesis, classical and hybrid market-timing models to assess the risk and efficiency of unit-linked insurance were used for the first time. The obtained results confirm that such models may form the basis for assessment of both the risk of unit-linked insurance and also their efficiency through investigation of the managers' skills within the scope of:

- a proper selection of securities,
- analysis of short-term trends dominating the market (market timing),
- using additional information concerning companies.

The results revealed that market-timing hybrid models may constitute an effective tool during the strategic decision-making process the insured go through. In addition, the discussed models can also be successfully used to assess the effectiveness and risks of other financial instruments available on the market, e.g. investment funds or equity.

REFERENCES

CZAPKIEWICZ, A., SKALNA, I., (2010). The CAPM and Fama-French Models in Warsaw Stock Exchange, „Przegląd Statystyczny" 57(4), pp.128–141.

CZAPKIEWICZ, A., SKALNA, I., (2011). Użyteczność stosowania modelu Famy i Frencha w okresach hossy i bessy na rynku akcji GPW w Warszawie (The performance of the Fama-French model for the Warsaw Stock Exchange boom and bust cycles), „Bank i kredyt" 42 (3), pp. 61–80.

CZEKAJ, J., JAJUGA, K., SOCHA, J., (2000). Rynek funduszy inwestycyjnych w Polsce (Investment fund market in Poland). AE, Kraków 2000.

FAMA, E. F., French, K. R., (1996). Multifactor Explanations of Asset Pricing Anomalies, „Journal of Finance", 51(1), pp. 55–84.

FOCARDI, S. M, FABOZZI, F. J., (2004). The Mathematics of Financial Modeling and Investment Management. Wiley Finance.

HOMA, M., (2013). Rozkład wypłaty w ubezpieczeniu na życie z funduszem kapitałowym a ryzyko finansowe (Distribution of the payments in the unit-linked life insurance and financial risk), Prace Naukowe UE nr PN 312 Zagadnienia aktuarialne – teoria i praktyka, Wyd. Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław, pp.78–87.

HENRIKSSON, R., MERTON, R., (1981). On market timing and investment performance. II. Statistical procedures for evaluating forecasting skills, Journal of Business, 54, pp. 513–533.

JAJUGA, K, JAJUGA, T., (2006). Inwestycje. (Investments), Wydawnictwo Naukowe PWN, Warszawa.

KOWERSKI, M., (2008). Trójczynnikowy model Famy i Frencha dla Giełdy Papierów Wartościowych w Warszawie (Fama – French Three – Factor Model for Warsaw Stock Exchange), „Przegląd Statystyczny" 55 (4), pp. 131–145.

MOŚCIBRODZKA, M., (2014). Stabilność czynników ryzyka w modelu Famy-Frencha wyceny kapitału na GPW w Warszawie (Stability of Risk Factors in Fama-French Pricing of Capital Model on Warsaw Stock Exchange), Zeszyty Naukowe Uniwersytetu Szczecińskiego nr 803, „Finanse, Rynki Finansowe, Ubezpieczenia" nr 66, Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Szczecin, pp. 305–319.

OLBRYŚ, J., (2011a). Obciążenie estymatora współczynnika alfa Jensena a interpretacje parametrów klasycznych modeli market-timing (The influence of the bias of the Jensen's alpha coefficient estimator on interpreting the parameters of the classical market-timing models), "Przegląd Statystyczny", pp. 42–59.

OLBRYŚ, J., (2011b). Wieloczynnikowe hybrydowe modele market-timing polskich funduszy inwestycyjnych (Multifactor hybrid market-timing models of Polish mutual funds), Studia Ekonomiczne – Zeszyty Naukowe Wydziałowe Uniwersytetu Ekonomicznego w Katowicach 01/2012; 97, pp. 149–161.

OLBRYŚ, J., (2010a). Czynniki Famy-Frencha w wieloczynnikowych modelach market-timing polskich funduszy inwestycyjnych (Fama and French Factors in Multifactor Market – Timing Models of Polish Mutual Funds), „Zeszyty Naukowe Uniwersytetu Szczecińskiego. Finanse. Rynki Finansowe. Ubezpieczenia" nr 29, pp. 33–48.

OLBRYŚ, J., (2010b). Three-factor market-timing models with Fama and French's spread variables, Operations Research and Decisions, 2/2010, pp. 91–106.

OLBRYŚ, J., (2010c). Ocena efektywności zarządzania portfelem funduszu inwestycyjnego z wykorzystaniem wybranych wieloczynnikowych modeli market-timing (Selected multifactor market-timing models for mutual fund performance evaluation), Optimum. Studia Ekonomiczne, 4(48), pp. 44–61.

OLBRYŚ, J., (2009). Conditional market-timing models for mutual fund performance evaluation, „Prace i Materiały Wydziału Zarządzania Uniwersytetu Gdańskiego"4/2, pp. 519–532.

OLBRYŚ, J., (2008a). Parametryczne testy umiejętności wyczucia rynku – porównanie wybranych metod na przykładzie OFI akcji (Parametric tests of market timing skills – a comparison of selected methods), [w:] Z. Binderman (red.) „Metody ilościowe w badaniach ekonomicznych IX", Wydawnictwo SGGW w Warszawie, pp. 81–88.

OLBRYŚ, J., (2008b). Ocena umiejętności stosowania strategii market-timing przez zarządzających port- felami funduszy inwestycyjnych a częstotliwość danych (Data Frequency Affects Inference Regarding Market Timing Ability of Mutual Fund Managers), „Studia i Prace Wydziału Nauk Ekonomicznych i Zarządzania" Nr 10, Uniwersytet Szczeciński, Szczecin, pp. 96–105.

OLBRYŚ, J., (2008c). Parametric tests for timing and selectivity in Polish mutual fund performance, „Optimum. Studia Ekonomiczne", Wydawnictwo Uniwersytetu w Białymstoku, 3(39)/2008, pp. 107–118.

REILLY, F. K., BROWN, K. C., (2001). Analiza inwestycji i zarządzanie portfelem (Investment analysis and portfolio management) tom II, PWE, Warszawa.

TREYNOR, J., MAZUY, K., (1966). Can mutual funds outguess the market?, Harvard Business Review, 44, pp. 131–136.

STATISTICS IN TRANSITION new series, Summer 2015
Vol. 16, No. 2, pp. 293–308

A MEASURE FOR REGIONAL RESILIENCE TO ECONOMIC CRISIS[1]

Małgorzata Markowska[2]

ABSTRACT

The purpose of the study (presented in this article) was to develop a measure of resilience to crisis, one that may be applied to regional data. In principle, such measure can take either positive or negative values. A positive value confirms resilience to crisis, whereas a negative one confirms the absence of resilience (sensitivity/vulnerability). The measure uses growth rates referred to the previous year under the assumption that crisis results in a slowdown in growth, or even in a decline in values of important economic indicators. Growth rates are standardized by dividing values of original change rates by medians specified based on spatio-temporal data modules. Such division results in each characteristic being brought to equal validity. Simultaneously, the original character is maintained and variables are not "flattened" by the outliers. Changing destimulants into stimulants occurs during growth rates calculation. The measure of resilience to crisis is calculated as an arithmetic mean of the values of characteristics brought to comparability. The measure of resilience can be converted into the measure of sensitivity by multiplying it by (-1).

The application of the proposed measure to assessing the resilience to crisis in the period 2006-2011 is presented for regions meant as the European Union NUTS2 units. The measure is based on comparable data, which allowed for using only six variables measuring changes in GDP, salaries, investments, household income, employment and unemployment.

Key words: economic crisis, aggregate measure, NUTS 2.

1. Introduction

Economic resilience to crisis with reference to a region is defined as its economic capacity to overcome negative external impacts. It depends on macroeconomic factors and internal determinants. Among macroeconomic factors the following can be listed: fiscal, economic and monetary policy. Internal factors

[1] The project has been financed by the Polish National Science Centre, decision DEC-2013/09/B/HS4/00509.

[2] Wrocław University of Economics.

take the form of, e.g.: economic structure, restructuring and modernization level of enterprises, competitiveness and innovation. Among the important internal factors the level of human capital, including entrepreneurship, is also considered (Masik, Rzyski 2014).

The objective of the article is to present the proposal for the construction of a measure of resilience to economic crisis, possible to be applied to regional data.

2. Sensitivity to crisis – research overeview

The assessments of economic reactions to shocks resulting from, e.g. an economic crisis are performed by analyzing macroeconomic sensitivity specified:

- in a more extensive sense as the *"vulnerability to external factors distracting a particular economy from following the desirable trajectory of development"* (Zaucha *et al.* 2014: 208),
- whereas in a narrower sense (sensitivity) in the context of *"economic structures and their tools for weakening negative stimuli and threats, as well as deriving benefits from the occurring opportunities without any structural changes"* (Zaucha *et al.* 2014: 208).

The studies of resilience and sensitivity to macroeconomic impacts, covering especially small countries, have been conducted for twenty years both independently and in a team by L. Briguglio (Briguglio 1995) from the University of Malta. The team's output includes, among other things: methods for the "construction" of economic resilience in small countries (Briguglio, Kisanga 2004, Briguglio, Cordina, Kisanga 2006, Briguglio 2014), developing the concept and measuring both sensitivity and resilience (Briguglio *et al.* 2006a, Briguglio *et al.* 2009), updating and extending the Economic Vulnerability Index (Briguglio, Galea 2003), the proposal of sensitivity and resilience profiles (Briguglio *et al.* 2010), the identification of economic resilience pillars in small countries (Briguglio *et al.* 2008), the analysis of growth problems in terms of resilience (Briguglio, Piccinino 2012), the assessment of economic resilience and adaptation potential (Briguglio, Cordina 2003).

Moreover, the studies focused on regional resilience were carried out by, e.g.: S. Christopherson, J. Michie and P. Tyler (2010) – theoretical and empirical aspects, K. Chapple, M. Belzer (2010) – job market, G. Bristow (2010) - competitiveness, J. Clark, H.-I. Huang and J. Walsh (2010) – innovation districts; R. Hassink (2010) as well as A. Pike, S. Dawley and J. Tomaney (2010) – differences in regional adaptation.

The research team, under the leadership of P. Churski (The National Centre for Science Project entitled: *Socio-economic growth vs. the development of growth and economic stagnation areas (2011-2013*)) conducted research the results of which are available on the project website: www.owsg.pl. The identification and assessment cover growth and stagnation areas based on the set of 49 indicators divided into five blocks (population and settlement, job market and economy structure, technical infrastructure and spatial availability, financial situation and wealth level, innovative economy and business environment), whereas within the framework of blocks – the factors described by means of qualities characteristic for a given factor (Perdał, Hauke 2014: 71). The studies presented by the research team are mainly focused on the territory of Poland (various NUTS levels), with particular emphasis on Wielkopolska region. For the purposes of performing comparisons the data from Slovakia, Lithuania and Latvia were used, among other things. The identification of factors and analyses were carried out with reference to the following groups of spatial units: all units, growth areas, transition areas and stagnation areas, mainly in the period 2000-2010.

The research on resilience to crisis, especially in Pomorskie region, is conducted within the framework of the project: Economic Crisis, Resilience of Regions – ESPON 2013 (partners: Cardiff University (project leader), FTZ-Leipzig, Aristotle University, Tartu University, University of Gdańsk, Manchester University, Experian Plc.), the purpose of which is (Masik 2013): *"the identification of economic crisis impacts on regional economies, the analysis of structural and functional determinants in regions, an attempt to answer the question why some regions* are *more resilient than others, the identification of policies supporting economic resilience"*.

The team under the leadership of J. Szlachta (Zaucha *et al.* 2014: 206-234) conducted the review of the subject literature in terms of approaches to regional sensitivity measurement within the framework of the project – *The sensitivity of Polish regions to challenges of contemporary economy. Implications for regional development policy,* grant from the National Centre for Science 1635/B/H03/2011/40 and within the framework of project implementation supervised by D. Strahl entitled: *"Smart growth vs. sensitivity to economic crisis in regional dimension – measurement methods"* (grant from the National Centre for Science 2013/09/B/HS4/00509) M. Markowska (2014), focused on such areas as: economy, job market and households, listed as the most vulnerable in the context of crisis phenomena assessment.

3. Proposal for measuring regional resilience to economic crisis (RRC)

It has been initially assumed that the suggested measure can take both positive and negative values. Its positive value indicates that a region is resistant to crisis, whereas a negative one informs about the absence of resistance, i.e. sensitivity and vulnerability to crisis phenomena.

The growth rate of variables calculated against previous years (formulas (1) and (2) was used in the construction of the measure. It results from the assumption that the effect of crisis is manifested in a slowdown in growth or even a decline in the values of crucial economic factors. Destimulants are changed into stimulants in the course of growth rates calculation:

$$w_{ijt} = 100\left(\frac{x_{ijt}}{x_{ij,t-1}} - 1\right) \text{ for stimulants,} \tag{1}$$

$$w_{ijt} = 100\left(1 - \frac{x_{ijt}}{x_{ij,t-1}}\right) \text{ for destimulants.} \tag{2}$$

At this point a conclusion can be drawn that in order to calculate an average rate a geometric mean rather than an arithmetic one should be used, however, w* values calculated below represent in fact the ratios of the rate and the median rather than the rate itself. The comparability of characteristics is obtained as a result of dividing the original rate values of variables changes (1) or (2) by the medians determined from spatio-temporal data modules (3). This transformation results in equal validity of the discussed characteristics. Such procedure maintains the original change rate sign and, moreover, the phenomenon of variables "flattening" by outlier values does not occur. Standardization (understood as achieving comparability) of changes of rates is performed by applying the following formula:

$$w^*_{ijt} \; = \; w_{ijt}/Me\left(\left|w_{ijt}\right|\right) \tag{3}$$

The measure of resistance to crisis is calculated as an arithmetic mean of the values of characteristics standardized by formula (3). The suggested measure takes the following form:

$$RRC_{it} = \frac{1}{m}\sum_{j=1}^{m} w^*_{ijt} \tag{4}$$

where:

i – object's number (region),

j – characteristic's number,

t – time unit number,

m – number of characteristics,

*w** - standardized change rate,

RRC – measure for regional resilience to crisis.

The range of measure values does not have either upper or lower limit. It should be assumed that it corresponds to a rational opinion that, on the one hand, it is never so bad that it could not be worse and, on the other, it can always be better than it actually is. The measure of resistance can be transformed into the measure of sensitivity by multiplying it by (-1).

4. Basic characteristics of RRC – preliminary assessment of results

Economy, job market and households represent the areas of regional sensitivity to economic crisis. In order to perform the assessment of regional economic situations, in terms of their resilience or sensitivity to economic crisis, the following indicators were used in the study covering the period 2005-2011 (as of 31st October 2014 the information for 2012 regarding the data presented in values and necessary to calculate change rates was not provided by Eurostat database):

- GDP in million PPS in a region (CR_GDP),
- investments in million Euro in a region (CR_IN),
- employment rate (as a percentage of professionally active population in 15-64 age group) (CR_ER),
- unemployment rate (destimulant) (as a percentage of the total number of professionally active population) (CR_UR),
- salaries in million Euro in a region (globally) (CR_S),
- disposable income per capita in a household in PPS (CR_DI).

The choice of variables was preceded by checking Eurostat database resources in terms of data availability, whereas the preliminary selection of variables was performed by assessing their changes, especially in 2009 against the previous years, among other things.

The EU territorial units at NUTS 2 level constituted the base of regions covered by the assessment – the total of 264 regions (excluding Croatian and overseas Spanish and French regions – due to significant data gaps).

In the dynamic assessment of changes the declines in 2009 against 2008 should be emphasized, since they were recorded in 250 regions (CR_GDP), 173 (CR_S), 212 (CR_IN), 202 (CR_DI), 205 (CR_ER). Moreover, for 171 EU NUTS 2 regions an increase in the unemployment rate was observed. It should also be emphasized that in the case of over 100 regions a decline in investments and GDP values was recorded also in 2008 (against the previous year). Simultaneously, further drops were observed in over 100 regions in the subsequent years for 159 and 110 regions with respect to employment rate and investments (127 and 120), along with an increase in the unemployment rate for 171 and 114 regions.

In performing the assessment of regions in terms of RRC attention was also paid to the EU 15 regions (the so-called “old” EU) and the EU 12 regions – from the accessions in 2004 and 2007.

The modules of medians of variables determined jointly in the entire period under analysis (dynamics in the period 2006-2011) were used in the standardization process and their values are presented in table 1. With reference to job market the attention should be paid to the median module of the unemployment rate which is several times higher than the employment rate.

The preliminary analysis of the obtained results indicates that 2009 represents the main crisis year in the EU NUTS 2 regions – the median is negative (median measure), as well as the mean value, and even the third quartile. In 2011 a group of weak regions was identified (Greek regions, in which a dramatic drop in salaries was recorded, among other things) which resulted in a strong left-sided asymmetry of the measure. The basic characteristics of the Measure for Regional Resilience to Economic Crisis are included in table 2.

Table 1. Medians used in standardization

Variable	The median of change rate module in regions
GDP change rate	4.54
Salaries change rate	4.15
Investments change rate	9.32
Household income change rate	3.22
Employment rate change	1.52
Unemployment rate change (destimulant)	12.50

Source: author's estimations.

Table 2. RRC characteristics

Year	$\bar{x}$	Min	$Q_{0.10}$	$Q_{0.25}$	Me	$Q_{0.75}$	$Q_{0.90}$	Max	SD	As
2006	0.58	-1.55	-0.04	0.24	0.53	0.77	1.28	2.72	0.58	0.82
2007	1.02	-0.76	0.31	0.57	0.89	1.20	1.98	5.06	0.76	1.69
2008	0.48	-1.59	-0.90	-0.10	0.48	0.87	2.03	4.12	1.01	0.46
2009	-0.52	-3.12	-1.31	-0.80	-0.44	-0.15	0.16	1.23	0.60	-0.72
2010	0.36	-1.76	-0.54	-0.01	0.37	0.87	1.24	2.73	0.70	-0.17
2011	0.09	-5.34	-0.47	-0.11	0.27	0.73	0.97	3.29	1.26	-3.04

Source: author's compilation.

Picture 1 presents the distribution of regions in terms of RRC values. The same scale of horizontal axis allows one to "follow the moves" of the measure distribution. The highest diversification is observed for 2008, whereas the highest deviation is true for 2011 (the Greek group visible on the left side of the distribution). The effect of RRC distribution approximation by a normal distribution on one graph is illustrated on fig. 2.

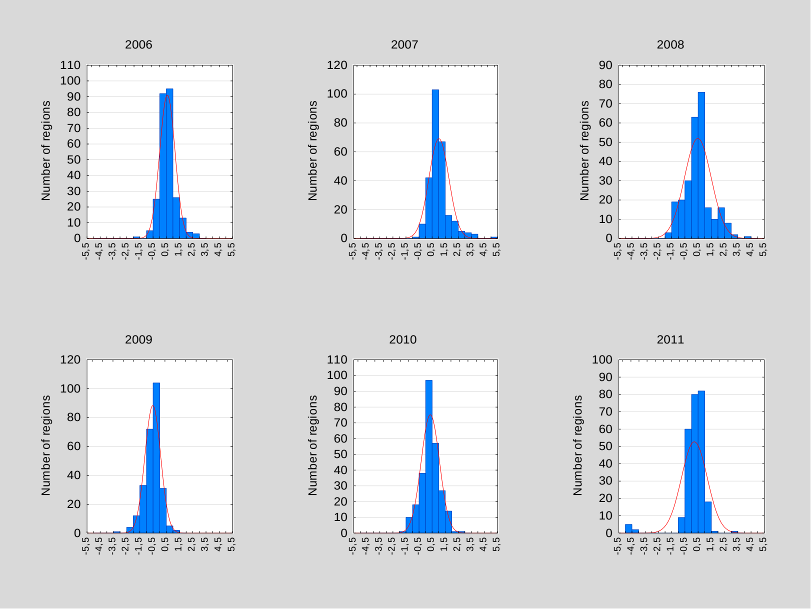


Figure 1. RRC distribution in the period

Source: author's compilation.

Picture 3 presents RRC deciles (the first decile at the bottom and the ninth on the top). The line at the zero level stands for the division of sensitivity and resilience. Line 1 above represents the resistant regions. In the period 2006-2007 about 50% regions were included in this part. The 2009 crisis is well visible. Only slightly less than 10% of regions were placed on the positive side, thus only the best ones were resistant to crisis.

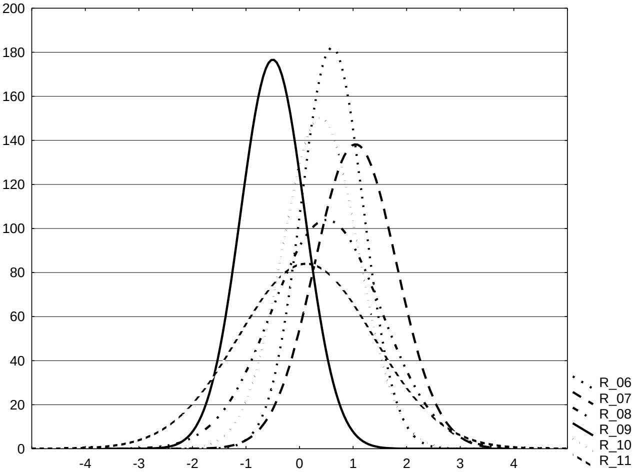


Figure. 2. Approximation of RRC distribution by a normal distribution

Source: author's compilation.

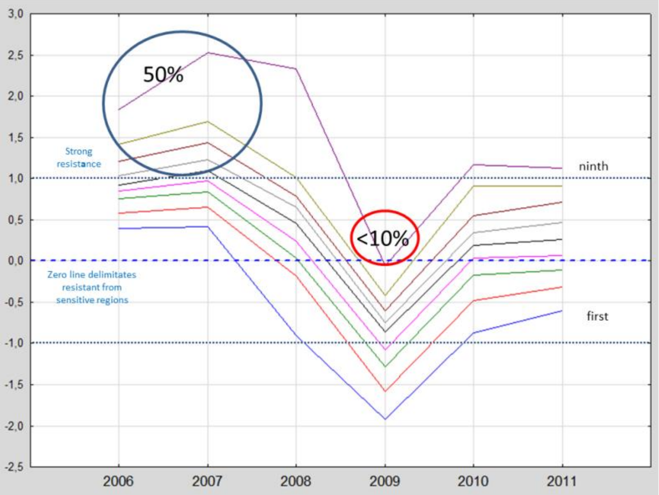


Figure 3. RRC deciles

Source: author's compilation.

On the basis of the analysis of numerical values of characteristics and the distributions of empirical values the division of RRC measure variability range into six classes can be proposed (see tab. 3).

Table 3. The suggested RRC classes

RRC value	Class
below -1	(-3) Strong sensitivity to crisis
from -1 to -0.5	(-2) Average sensitivity to crisis
from -0.5 to 0	(-1) Poor sensitivity to crisis
from 0 to 0.5	(+1) Poor resilience to crisis
from 0.5 to 1	(+2) Average resilience to crisis
above 1	(+3) Strong resilience to crisis

Source: author's compilation.

These classes allow for a more generalized assessment of sensitivity or resilience to crisis, as well as the quantification of these responses to crisis.

5. Results of the EU nuts 2 regions' division into groups based on RRC values

In each year of the study each region was assigned to one of the classes identified before. An even more general assessment than assigning to one of the six classes specified whether a particular region in a given year was resilient to crisis (a class coded with a plus), or sensitive to crisis (one of the classes coded with a minus). The assessment of regional response to crisis in the entire analyzed 6-year period can be easily obtained by counting pluses – from zero to six. The analysis of the distribution of regions into these seven values results in suggesting four classes of resilience to crisis presented in table 4.

The number of regional positive measure values of resilience to crisis in the period 2006-2011 constituted the basis for identifying particular classes – see tab. 4.

The most numerous classes are made up of the “coping” or “fighting” regions – 111 and 97 regions respectively, whereas the least numerous one covers the “resistant” regions – see tab. 4 and 5. Table 4 also summarizes the information about the number of positive measure values in the period 2006-2011.

Table 4. Classes based on RRC values in the period 2006-2011

Number (+) for RRC in the period 2006-2011	Regions	Number of regions in a class	Percentage
0-2	Sensitive	32	12.1
3-4	Fighting	97	36.7
5	Coping	111	42.0
6	Resistant	24	9.2

Source: author’s compilation.

While assessing the distribution of the EU NUTS 2 regions in classes attention should be paid to the fact that the regions from four countries were included in three classes: French, Czech and Belgian regions in “resistant”, “coping” and “fighting” classes, whereas Italian regions in the “coping”, “fighting” and “sensitive” classes.

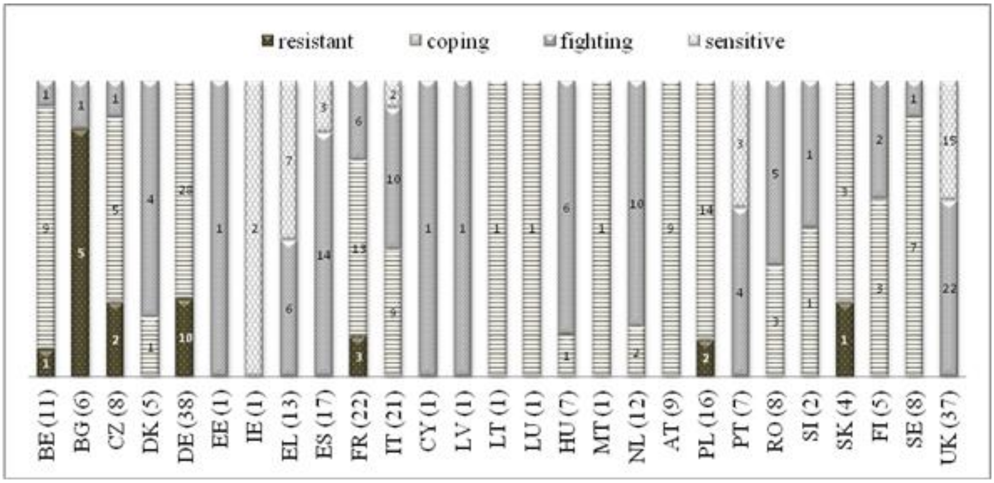


Figure 4. Summary – the number of regions in a class in the EU countries

Source: author’s compilation.

Île de France (FR) was the only region recorded in the group of EU 15 capital regions or the ones including the capital of a country in the “resistant” class. The “coping” regions group covered the following ones: Région de Bruxelles-Capitale (BE), Berlin (DE), Lazio (IT), Wien (AT), Stockholm (SE), and the “fighting” regions class included: Hovedstaden (DK), Noord-Holland (NL), Etelä-Suomi (FI), Outer London and also Inner London (UK), Comunidad de Madrid (ES) and Lisboa (PT), whereas the Greek Attiki and the Irish Southern and Eastern regions were listed in the class of “strongly sensitive” ones (see fig. 4).

The region capital or the ones including the capital of a country from the countries of 2004 and 2007accessions were classified in the following groups:

- “resistant”: Yugozapaden (BG), Mazowieckie (PL) and Bratislavský kraj (SK),
- “coping”: Praha (CZ) and Bucuresti – Ilfov (RO),
- „fighting”: Zahodna Slovenija (SI) and Közép-Magyarország (HU).

Among NUTS 1 regions the following were listed in the “coping” class: Lithuania, Malta and Luxemburg, whereas Eesti, Kypros and Latvia were included in the “fighting” class.

Table 5. Assigning the EU regions to classes

Country (number of regions): regions	Class
BE (1): Prov. Oost-Vlaanderen; BG (5): Severents entralen, Severoiztochen, Yugoiztochen, **Yugozapaden**, Yuzhen tsentralen; CZ (2): Jihozápad, Severozápad; DE (10): Brandenburg, Mecklenburg-Vorpommern, Weser-Ems, Köln, Münster, Detmold, Arnsberg, Trier, Leipzig, Sachsen-Anhalt; FR (3): **Île de France**, Poitou-Charentes, Provence-Alpes-Côted'Azur; PL (2): **Mazowieckie**, Pomorskie; SK (1): **Bratislavský kraj**.	Resistant(24)
BE (9): **Région de Bruxelles-Capitale**, Prov. Antwerpen, Prov. Limburg, Prov. Vlaams-Brabant, Prov. Brabant Wallon, Prov. Hainaut, Prov. Liege, Prov. Luxembourg, Prov. Namur; CZ (5): **Praha**, Strední Cechy, Jihovýchod, Strední Morava, Moravskoslezsko; DK (1): Syddanmark; DE (28): Stuttgart, Karlsruhe, Freiburg, Tübingen, Oberbayern, Niederbayern, Oberpfalz, Oberfranken, Mittelfranken, Unterfranken, Schwaben, **Berlin**, Bremen, Hamburg, Darmstadt, Gießen, Kassel, Braunschweig, Hannover, Lüneburg, Düsseldorf, Koblenz, Rheinhessen-Pfalz, Saarland, Dresden, Chemnitz, Schleswig-Holstein, Thüringen; FR (13): Basse-Normandie, Bourgogne, Nord-Pas-de-Calais, Lorraine, Franche-Comté, Pays de la Loire, Bretagne, Aquitaine, Midi-Pyrénées, Rhône-Alpes, Auvergne, Languedoc-Roussillon, Corse; IT (9): Piemonte, Liguria, Provincia Autonoma di Bolzano, Provincia Autonoma di Trento, Veneto, Friuli-Venezia Giulia, Emilia-Romagna, Toscana, **Lazio**; **LT Lithuania**; **LU Luxembourg**; HU (1): Dél-Dunántúl; **MT Malta**; NL (2): Zeeland, Noord-Brabant; AT (9): Burgenland, Niederösterreich, **Wien**, Kärnten, Steiermark, Oberösterreich, Salzburg, Tirol, Vorarlberg; PL (14): Łódzkie, Małopolskie, Śląskie, Lubelskie, Podkarpackie, Świętokrzyskie,	Coping (111)

Podlaskie, Wielkopolskie, Zachodniopomorskie, Lubuskie, Dolnośląskie, Opolskie, Kujawsko-pomorskie, Warmińsko-mazurskie; RO (3): Nord-Est, **Bucuresti-Ilfov**, Sud-Vest Oltenia; SI (1): Vzhodna Slovenija; SK (3): Západné Slovensko, Stredné Slovensko, Východné Slovensko; FI (3): Länsi-Suomi, Helsinki-Uusimaa, Pohjois- ja Itä-Suomi; SE (7): **Stockholm**, Östra Mellansverige, Smaland med öarna, Västsverige, Norra Mellansverige, Mellersta Norrland, Övre Norrland.	
BE (1): Prov. West-Vlaanderen; BG (1): Severozapaden; CZ (1): Severovýchod; DK (4): **Hovedstaden**, Sjalland, Midtjylland, Nordjylland; **EE Eesti**; EL (6): Dytiki Makedonia, Thessalia, Ipeiros, Ionia Nisia, Dytiki Ellada, Notio Aigaio; ES (14): Galicia, Principado de Asturias, Cantabria, País Vasco, Comunidad Foral de Navarra, Aragón, **Comunidad de Madrid**, Castilla y León, Castilla-la Mancha, Extremadura, Comunidad Valenciana Andalucía, Región de Murcia, Canarias; FR (6): Champagne-Ardenne, Picardie, Haute-Normandie, Centre, Alsace, Limousin; IT (10): Valle d'Aosta, Lombardia, Umbria, Marche, Abruzzo, Puglia, Basilicata, Calabria, Sicilia, Sardegna; **CY Kypros**; **LV Latvia**; HU (6); **Közép-Magyarország**, Közép-Dunántúl, Nyugat-Dunántúl, Észak-Magyarország, Észak-Alföld, Dél-Alföld; NL (10): Groningen, Friesland, Drenthe, Overijssel, Gelderland, Flevoland, Utrecht, **Noord-Holland**, Zuid-Holland, Limburg; PT (4): Norte, Centro, **Lisboa**, Regiao Autónoma dos Açores; RO (5): Nord-Vest, Centru, Sud-Est, Sud–Muntenia, Vest; SI (1) **Zahodna Slovenija**; FI (2): **Etelä-Suomi**, Aland; SE (1): Sydsverige; UK (22): Tees Valley and Durham, Greater Manchester, Lancashire, Cheshire, Merseyside, South Yorkshire, Leicestershire, Rutland and Northamptonshire, Lincolnshire, Bedfordshire and Hertfordshire, Essex, **Inner London, Outer London**, Berkshire, Buckinghamshire and Oxfordshire, Surrey, East and West Sussex, Hampshire and Isle of Wight, Kent, Dorset and Somerset, West Wales and The Valleys, East Wales, Eastern Scotland, North Eastern Scotland, Highlands and Islands.	Fighting (97)
IE (2): Border, Midland and Western, **Southern and Eastern**; EL (7): Anatoliki Makedonia, Thraki, Kentriki Makedonia, Sterea Ellada, Peloponnisos, **Attiki**, Voreio Aigaio, Kriti; ES (3) La Rioja, Cataluna, IllesBalears; IT (2): Molise, Campania; PT (3): Algarve, Alentejo, Regiao Autónoma da Madeira; UK (15): Northumberland and Tyne and Wear, Cumbria, East Yorkshire and Northern Lincolnshire, North Yorkshire, West Yorkshire, Derbyshire and Nottinghamshire, Herefordshire, Worcestershire and Warwickshire, Shropshire and Staffordshire, West Midlands, East Anglia, Gloucestershire, Wiltshire and Bristol, Cornwall and Isles of Scilly, Devon, South Western Scotland, Northern Ireland.	Sensitive (32)

Capital regions or the regions including the capital of a country are marked in bold.

Source: author's compilation.

Figure 5 illustrates the geographical distribution of regions from the established classes.

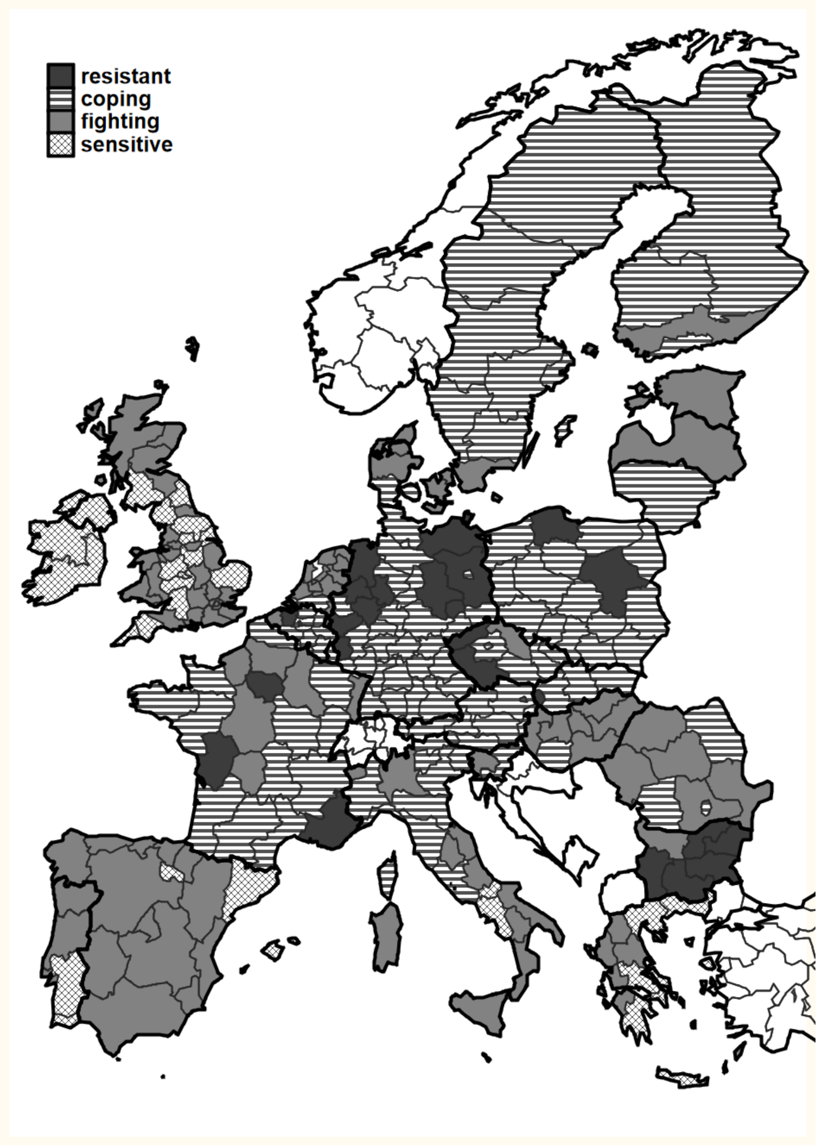


Figure 5. Graphic presentation of the classification results
Source: author's compilation.

Out of 208 EU 15 NUTS 2 regions the "resistant" class included 6.7% of regions (mainly German – 10 out of 14 regions in this class, three French ones and a Belgian region - Prov. Oost-Vlaanderen). The following two classes covered a similar number of regions from the EU 15: “coping” 39.4% and “fighting” 38.5%. The “sensitive” class included 15.4% of regions from the “old” EU, however, none of the EU 12. It is worth emphasizing that more than half of EU 12 regions (51.8%) were listed in the “coping” class, 30% in the “fighting” class and 17.9 in the “resistant” class.

6. Conclusions

The suggested construction of the measure allows for the assessment of resilience (resistance) to crisis in regions. The measure facilitates:

- arranging regions by their resilience (sensitivity) level to crisis,
- dynamic analyses and
- the synthetic identification and interpretation of classes obtained as a result of applying the dynamic taxonomy of regions,
- the generalized assessment of resilience (sensitivity) to crisis by assigning it to the suggested quality classes.

The obtained results indicate that the crisis affected the wealthy regions to a much larger extent, which is related to overproduction (resulting from the lack of moderation in meeting the needs) of banking products, as S. Bartosiewicz emphasizes (Bartosiewicz 2014).

It can be assumed that primarily in the case of the EU 12 regions, included in the group of resilient and coping regions, the cohesion policy carried out by the EU had a decisive impact on economy of these regions. For the regions from the countries of recent accession, structural funds and their influence on many spheres of economic life turned out to be a kind of “catalyst” for resilience to economic crisis. The effect of pre-accession structural funds was observed: (Bulgarian and Romanian regions), prolonged financial activities from the period 2004-2006 and also the period 2007-2013 (e.g. Polish regions).

The above mentioned assumptions were also confirmed by another research (Markowska, Strahl 2015), which assess the relations of variables characterizing smart growth of the European Union regions at NUTS 2 level (in the system of three pillars, i.e. innovation, creative regions and smart specialization described by several variables) with their sensitivity to economic crisis using logit models.

REFERENCES

BARTOSIEWICZ, S., (2014). Głos w dyskusji na temat przyczyn kryzysów gospodarczych [The voice in the discussion about the causes of economic crises], [in:] Modelowanie i prognozowanie zjawisk społeczno-gospodarczych. Teoria i praktyka [Modelling and forecasting of socio-economic phenomena. Theory and practice], B. Pawełek (ed.), Cracow University of Economics Publishing House, Cracow.

BRIGUGLIO, L., PICCININO, S., (2012). Growth with Resilience in East Asia and the 2008-2009 Global Recession, Asian Development Review, vol. 29, No. 2, pp. 183–206.

BRIGUGLIO, L., (1995). Small Island States and their Economic Vulnerabilities, World Development, vol. 23(9), pp. 1615–1632.

BRIGUGLIO, L., (2014). Resilience Building in Vulnerable Small States, Commonwealth Yearbook 2014, http://www.um.edu.mt/_data/assets/pdf_file/0012/205104/Briguglio_Resilience_Article_for_Comsec_Yearbook_13Jan13.pdf [downloaded on May 28, 2014].

BRIGUGLIO, L., CORDINA, C., VELLA, S., VIGILANCE, C., (2010). Profiling Vulnerability and Resilience: A Manual for Small States. Commonwealth Secretariat, Marlborough House, London.

BRIGUGLIO, L., CORDINA, G., (2003). The Economic Vulnerability and Potential for Adaptation of the Maltese Islands to Climate Change. Proceedings of the International Symposium on Climate Change, ISCC, Beijing, pp. 62–65.

BRIGUGLIO, L., CORDINA, G., FARRUGIA, N., VELLA, S., (2009). Economic Vulnerability and Resilience: Concepts and Measurements. Oxford Development Studies, Vol. 37 (3), pp. 227–247.

BRIGUGLIO, L., CORDINA, G., FARRUGIA, N., VIGILNACE, C., (2008). Small States and the Pillars of Economic Resilience of Small States. Islands and Small States Institute, University of Malta and Commonwealth Secretariat, Marlborough House, London.

BRIGUGLIO, L., CORDINA, G., KISANGA, E. J., (2006). Building the Economic Resilience of Small States. Islands and Small States Institute, University of Malta and Commonwealth Secretariat, Marlborough House, London.

BRIGUGLIO, L., GALEA, W., (2003). Updating and Augmenting the Economic Vulnerability Index. https://secure.um.edu.mt/_data/assets/pdf_file/0012/44130/eviar_briguglio_galea_ver4.pdf [downloaded on May 28, 2014].

BRIGUGLIO, L., KISANGA, E. J., (2004). Economic Vulnerability and Resilience of Small States. Islands and Small States Institute, University of Malta and Commonwealth Secretariat, Marlborough House, London.

BRIGULIO, L., CORDINA, G., FARRUGIA, N., VELLA, S., (2006a). Conceptualising and Measuring Economic Resilience, https://secure.um.edu.mt/__data/assets/pdf_file/0013/44122/resilience_index. pdf [downloaded on May 28, 2014].

BRISTOW, G., (2010). Resilient regions: re-`place`ing regional competitiveness. Cambridge Journals of Regions, Economy and Society, 3, 153-167.

CHAPPLE, K., BELZER, M., (2010). The resilient regional labour market: the US case. Cambridge Journals of Regions, Economy and Society, 3, 85–104.

CHRISTOPHERSON, S., MICHIE J., TYLER, P., (2010). Regional resilience: theoretical and empirical perspectives, Cambridge Journals of Regions, Economy and Society, 3, 3–10.

CLARK, J., HUANG, H.-I., WALSH, J., (2010). A typology of `Innovation Districts`: what it means for regional resilience, Cambridge Journals of Regions, Economy and Society, 3, 121–137.

HASSINK, R., (2010). Regional resilience: a promising concept to explain differences in regional economic adaptability? Cambridge Journals of Regions, Economy and Society, 3, 45–58.

MARKOWSKA, M., (2014). Ocena zależności między rozwojem inteligentnym a odpornością na kryzys ekonomiczny w wymiarze regionalnym – przegląd badań [Assessment of dependence between smart growth and resilience to economic crisis in regional dimension – research review, [in:] Gospodarka regionalna w teorii i praktyce [Regional economy in theory and practice], D. Strahl, A. Raszkowski and D. Głuszczuk (eds.), Research Studies of Wrocław University of Economics No. 333, Wrocław, 22–32.

MARKOWSKA, M., (2015). The vulnerability of regions to economic crisis - measurement problems, [in:] Regional Economy and Policy. Territory and Cities, Hlavacek P., Olsova P. (eds.), Jan Evangelista Purkyne University in Usti nad Labem, Usti nad Labem, 104–112.

MARKOWSKA, M., STRAHL, D., (2015). Inteligentny rozwój a podatność na kryzys ekonomiczny regionów Unii Europejskiej – próba oceny z wykorzystaniem modeli logitowych [Smart development and vulnerability to crises in UE regions – evaluation with logistic regression], Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie [Scientific Papers of the University of Economics in Cracow], Cracow (accepted for publication).

MASIK, G., RZYSKI, S., (2014). Resilience of Pomorskie region to economic crisis, Bulletin of Geography. Socio-economic Series, Volume 25, Issue 25, 129–141.

MASIK, G., (2014). Uwarunkowania i instrumenty wspierające odporność gospodarczą miast i regionów, referat na konferencji na temat „Polityka miejska. Wyzwania, doświadczenia, inspiracje" (Warszawa 25-26.06.2013), http://www.euroreg.uw.edu.pl/media/prezentacje_konferencja_polityka_miejska/_2_grzegorz_masik.pdf [downloaded on May 28, 2014].

PERDAŁ, R., HAUKE, J., (2014). Czynniki rozwoju obszarów wzrostu i obszarów stagnacji gospodarczej w Polsce, [in:] Rozwój społeczno-gospodarczy a kształtowanie się obszarów wzrostu i obszarów stagnacji gospodarczej, P. Churski (ed.), Rozwój Regionalny i Polityka Regionalna, 25/2014, Instytut Geografii Społeczno-ekonomicznej i Gospodarki Przestrzennej, Bogucki Wydawnictwo Naukowe, Poznań, 69–88.

PIKE, A., DAWLEY, S., TOMANEY, J., (2010). Resilience, adaptation and adaptability. Cambridge Journals of Regions, Economy and Society, 2010, 3, 59–70.

REGIONS IN THE EUROPEAN UNION. NOMENCLATURE OF TERRITORIAL UNITS FOR STATISTICS NUTS 2010/EU-27, (2011). Series: Methodologies & Working Papers, European Commission, Luxemburg.

ZAUCHA, J., CIOŁEK, D., BRODZICKI, T., GŁAZEK, E., (2014). Wrażliwość polskich regionów na wyzwania gospodarki globalnej, [in:] Gawlikowska-Hueckel K., Szlachta J. (eds.), Wrażliwość polskich regionów na wyzwania współczesnej gospodarki. Implikacje dla polityki rozwoju regionalnego. Oficyna Wolters Kluwer business, Warszawa, 206–244.

STATISTICS IN TRANSITION new series, Summer2015
Vol. 16, No. 2, pp. 309–322

COMPUTERISED RECOMMENDATIONS ON E-TRANSACTION FINALISATION BY MEANS OF MACHINE LEARNING

Germanas Budnikas[1]

ABSTRACT

Nowadays a vast majority of businesses are supported or executed online. Website-to-user interaction is extremely important and user browsing activity on a website is becoming important to analyse. This paper is devoted to the research on user online behaviour and making computerised advices. Several problems and their solutions are discussed: to know user behaviour online pattern with respect to business objectives and estimate a possible highest impact on user online activity. The approach suggested in the paper uses the following techniques: Business Process Modelling for formalisation of user online activity; Google Analytics tracking code function for gathering statistical data about user online activities; Naïve Bayes classifier and a feedforward neural network for a classification of online patterns of user behaviour as well as for an estimation of a website component that has the highest impact on a fulfilment of business objective by a user and which will be advised to be looked at. The technique is illustrated by an example.

Key words: online behaviour, Google Analytics, Naïve Bayes classifier, artificial neural network.

1. Introduction

Practically all of nowadays businesses rely on websites and web services. The structure of their interaction with a customer can be represented as a two-phase process. During the first phase a user gets some information about a service, during the second phase the user finalises his (her) transaction with a website and/or leaves the website. A transaction finalisation is a website content dependent process – it might be a service ordering, commenting, Facebook likes, etc. It is extremely important for business owners to know how website guests behave online and if it

[1] University of Bialystok, Faculty of Economics and Informatics in Vilnius. Kaunas University of Technology, Faculty of Informatics, Lithuania. E-mail: german.budnik@uwb.edu.pl.

is possible to influence their actions. This paper addresses these issues and presents results of the research.

The topic of the paper has a practical value. Analysis and understanding of web user behaviour is a key topic of a behavioural targeting. Behavioural targeting is an evolving area of a web mining that deals with optimisation of web online ads based on an analysis of web user behaviours. The research presented in the paper has some similarities to works in the considered field of the study. Methods of behavioural analysis investigate web surfing data gathered mainly from log files. The topic is actively investigated; examples of similar works include papers by (Angeletou, Rowe and Alani, 2011), (Dembczyński K., 2009), (Robinson D.J., 2008).

Approach by (Robinson D.J., 2008) suggests a method for monitoring user online behaviour. The method is implemented based on data pulled from log files where HTTP/GET requests are saved when a user clicks a hyperlink. These data are gathered using agent devices installed on a user's computer. The approach uses Open Directory Project (Xian, Chen and Wang, 2014) for a categorisation of visited websites. The research emphasises the creation of behaviour profiles with respect to web page visitation event, frequencies and probability distributions, and causality relations or time-dependencies.

The technique by (Dembczyński K., 2009) describes the problem of predicting behaviour of web users based on real historical data. The data are gathered from the user's cookie files. An analysis is performed using a statistical decision theory.

Paper by (Angeletou, Rowe and Alani, 2011) presents a method for modelling and analysis of user behaviour in online communities that include personal profiles, wiki, blogs, file sharing, and a forum. The approach implements behaviour modelling, role mining and role inference and is based on a statistical clustering.

The approach proposed in the current paper differs from the works listed above by its application area – it operates at Internet level, while (Angeletou, Rowe and Alani, 2011) and (Robinson D.J., 2008) approaches operate at Intranet level. The approach proposed is similar to (Angeletou, Rowe and Alani, 2011) because they both use a dynamical update of estimations with respect to new data.

The technique suggested in the paper consists of the following steps. At the beginning, in order to know the actual on-site user behaviour, user browsing activities should be formalised. The paper applies Business Process Modelling Notation (Drejewicz, 2012) for such formalisation. It enables a definition of data to be read off from a website during monitoring user browsing activities by means of Google Analytics (Clifton, 2012) tracking function. This permits to gather statistical data required for an analysis. The aim of such analysis is to build a model of user on-site behaviour (an earlier paper on that topic can be found in (Budnikas, 2015)) – whether a website guest is willing to finalise a transaction or not. The statistical data are handled during the second step. The model used for the analysis is based on classic machine learning techniques – Naïve Bayes classifier and a feedforward artificial neural network – Multi-layer Perceptron model. In the third step, Naïve Bayes classifier is applied to analyse the actual website user browsing activities based on gathered statistical data. In the fourth step, the two already mentioned techniques are used together in order to classify actual user online behaviour with respect to gathered statistical data. Depending on an outcome of the classification, the website may recommend a visit to that web page to an online user, which has the biggest impact on the transaction finalisation to be defined using auxiliary classification.

The paper is structured in the following way. The second section gives a formalisation of browsing activities with respect to a website category as well as data needed to monitor a website. The third section presents a sketch of a procedure to gather statistical data from a website and to handle possible inconsistency cases. Machine learning data analysis methods used in the proposed technique are discussed in the fourth section, namely – Naïve Bayes classifier and Multi-layer Perceptron model, whose structure and parameters are given too. The fifth section illustrates the technique proposed. Conclusions summarise main results achieved and state future work directions.

2. Website formalisation

Surfing on websites usually differs with respect to types of these sites. Open Directory Project (ODP) differentiates the following website types: Arts, Business, Computers, and 13 more instances. These types generalise manually selected websites in different languages and are used in various kind of research including the suggested in this paper. Classification of websites into types helps in understanding possible kinds of behaviour. Specification of sub-types and its instances is actual for understanding behaviour cases. The paper considers an instance of the Consumer Goods and Services sub-type of a business type with respect to ODP classification. Each browsing activity on websites, especially on business sites, can be logically divided into two parts – *introductory* part, which usually includes list of services, descriptions, etc., and (transaction) *finalisation* part, which could be expressed by paying for services, commenting, Facebook likes and so on. According to Figure 1, the introductory logical part of a browsing activity may consist of Product Category Selection, Product Selection, Product Related Information Viewing, Delivery and Company Information Viewing; while Check-out and Payment browsing activity corresponds to the logical part – the transaction finalisation.

Formalisation of browsing activity makes it possible to understand user on-site behaviour that can be monitored by using various techniques, e.g., Google Analytics (Clifton, 2012) tracking function.

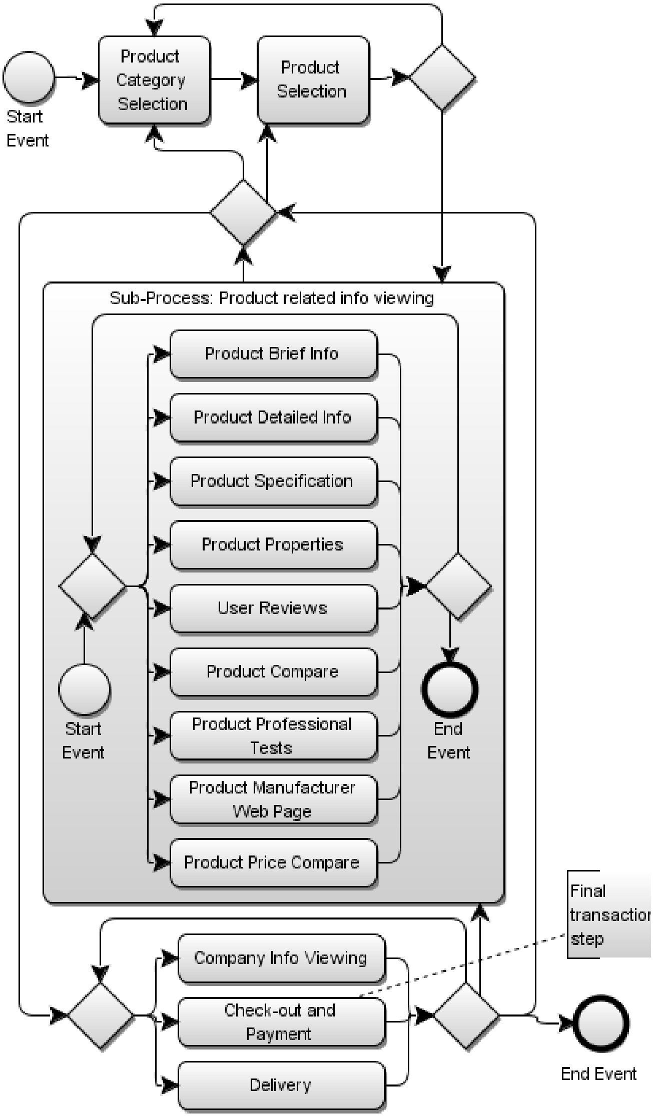


Figure 1. A generalised view of user behaviour on "Consumer Goods and Services" sites using Business Process Modelling Notation.

Source: (Budnikas, 2015).

A 5-tuple

<*e*, *y*, *u*, *t*, *m*>, where

e is user browsing *session* during which website pages are visited;

y is a *category* of a product viewed by a user. As e-commerce website may contain a huge number of products (even of the same category), products are differentiated only if they belong to the different categories;

u is a *user* that is identified by a cookie file. A cookie is a small text file that contains user visiting on-site specific information;

t is a kind of an activity or a *task* performed by a user on the website page like "Product Category Selection", "Viewing Product Price Comparison" (see Figure 1);

m is activity *t* start time *moment* which application is twofold. First, it is used to know a sequence number of a web page visit for the first time during a session. Second, it is used to count revisits to the same web page.

defines data needed for monitoring user online activities. These data also set requirements for database table where browsing activity statistical data are stored.

3. Gathering statistical data

In order to classify user actual on-site behaviour, a training data set should be collected from the site. The technique suggested in this paper uses Event Tracking method, which is a part of Google Analytics tracking code (Clifton, 2012). It enables recording user interactions with website elements, such as web page, embedded AJAX page element, page gadgets, and Flash-driven element and so on. Additionally to tracking function, a cookie file is used for unique user identification (Nikiforakis, Acar and Saelinger, 2014).

During a session of website browsing information about visited pages is collected and stored in the following form

$$\langle e, y, u, t_1, \dots, t_{n-1}, r_{t_1}, \dots, r_{t_{n-1}}, s_{t_1}, \dots, s_{t_{n-1}}, t_f \rangle,$$

where t_f corresponds the final task, s_{t_i} – means a sequence number of the t_i-th web page visit for the first time during the *e*-th session andr_{t_i} is a counter of revisits to the same t_i-th web page. For example, Table 1 record R_1 represents a situation that a user during his/her first session has visited Product Properties (t_4), Product Price Compare (t_9) and Delivery (t_{11}) web pages and has not finalised the transaction – Check-out and Payment task (t_f) has not been accomplished and web page t_4 has been visited first in a sequence (s_{t_4}=1) and was revisited twice. Task designations have the following meanings: 0 means a web page has not been visited (i.e., a task has not been accomplished) and 1 means that a web page has been visited. User next session (see record R_2 of Table 1) consists of visits to the same pages (it is marked by grey background colour in the table) that resulted in the transaction finalisation.

As seen in Table 1, inconsistent data entries with respect to the visited web pages may exist in the gathered statistical data. An inconsistency case is when the same set of accomplished tasks in different data entries is followed by opposite finalisation tasks. A fragment of the pseudo-code of an algorithm used for the inconsistency case handling is presented next (see Figure 2). This fragment excludes variables $s_{t_i}$and r_{t_i}as they have no influence on inconsistency. Note also that if the same user visits a site repeatedly and his/her browsing activity is different, corresponding records do not join since separate records represent a real situation on browsing activities in a database. Such an approach also simplifies computations.

Table 1. Illustration of statistical data fragment read off from a website

Record number	Session	Product category	User ID	Brief Info	Detailed info	Specification	Properties	User Reviews	Compare	Professional Tests	Manufacturer Web page	Price Compare	Company Info	Delivery	Check-out and Payment	…	Revisits# to Properties page	Sequence# of web page 1st visit	…
	e	y	u	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_f	…	r_{t_4}	s_{t_4}	
R_1	1	1	1	0	0	0	1	0	0	0	0	1	0	1	0	…	2	1	…
R_2	2	1	1	0	0	0	1	0	0	0	0	1	0	1	1	…	0	3	…

Source: own elaboration.

Algorithm Handling of inconsistency cases in gathered statistical data

```
1:  $T^I = \emptyset; T = \emptyset$
2:  $T = T \cup t^i | t^i = \{y^i, u^i, t_1^i, \dots, t_{n-1}^i, t_f^i\}$
3:  If $t_f^i \neq t_f^j$, where $\forall i, j, k: i \neq j; y^i = y^j; t_k^i = t_k^j; t^{i,j} \in T \cup T^I$
    then
4:      If $u^i = u^j$, $t_f^i \neq t_f^j$, $t_f^j = 0$, $t_k^i = t_k^j$ then
5:          $T = T \cup t^i \backslash t^j$
6:          $T^I = T^I \backslash t^i \backslash t^j$
7:      End If
8:      If $u^i \neq u^j$, $t_f^i \neq t_f^j$, $t_f^j = 0$, $t_k^i = t_k^j$ then
9:          $T^I = T^I \cup t^i \cup t^j$
10:         $T = T \backslash t^i \backslash t^j$
11:     End If
12: End If
13: GOTO 2
```

Figure 2. A fragment of the algorithm for inconsistent data handling

Source: own elaboration.

The algorithm initialises an inconsistent data set T^I and a statistical data set *T*. Further, the set *T* is supplemented with data about web page visits, a user, and a product category. If the transaction finalisation tasks t_f^i and t_f^j in *i* and *j* data entries from the all data sets are different while the rest of accomplished tasks are the same for the same product category, two inconsistency handling options are available – described in the steps 4-7 and 8-11 respectively. If inconsistency has arisen in the browsing sessions by the same users u^i and u^j, data entry t^i corresponding to the transaction finalisation is added to the statistical data set *T* and excluded from the inconsistent data set T^I, while opposite data entry t^j is excluded from all the sets. If inconsistency has arisen in browsing sessions by different users, inconsistent data entries $t^i$and t^j are added to the set T^I and excluded from the set *T*. The algorithm is repeated starting from the step 2 along with the arrival of data about next browsing session.

4. Machine learning data analysis methods

In spite of recent research in big data analysis that is common for well-known e-commerce sites, less known e-commerce websites still exist, whose customer visits and number of successful transaction finalisations are not so big. As statistical data are being gradually added to a database, the number of training data entries is not sufficient for some classification methods. This fact sets a premise to use a classification technique like Naïve Bayes classifier that works well with a comparatively small set of training data. When the number of statistical data reaches the threshold corresponding to the minimal number of entries in a training data set that is sufficient for a classification with a predefined error level, Multi-layer Perceptron (MLP) technique is applied additionally to Naïve Bayes classifier. If outcomes of the two classification methods are different, a class that represents transaction non-finalisation is regarded as dominating. The threshold is calculated using the rule of thumb

threshold = *number of weights* / *error level*

where *number of weights* and *error level* are parameters of the MLP model.

If estimation of actual user browsing activity shows transaction non-finalisation possibility and a user has visited at least 30% of all pages as described in website activity formalisation step, a website-to-user interaction procedure starts (note that 30% level is set based on experiment outcomes). The purpose of the procedure is to estimate the web page with the highest impact on the transaction finalisation and suggest a user to visit that page. Estimation experiments use the classification technique to find maximal similarity to the desired class while considering distinct unvisited web pages.

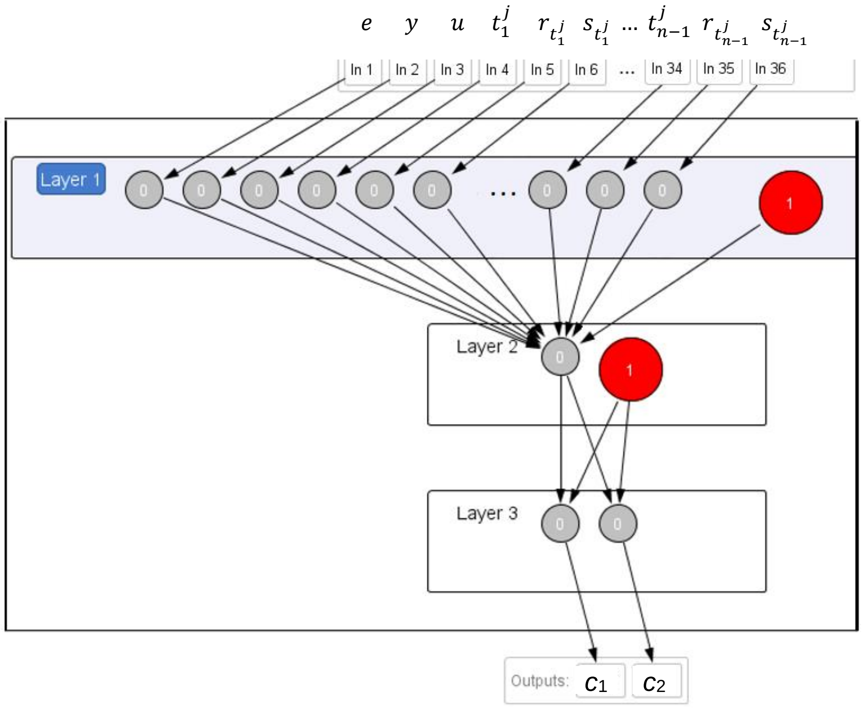


Figure 3. A structure of Multi-layer Perceptron model

Source: own elaboration using NeurophStudio.

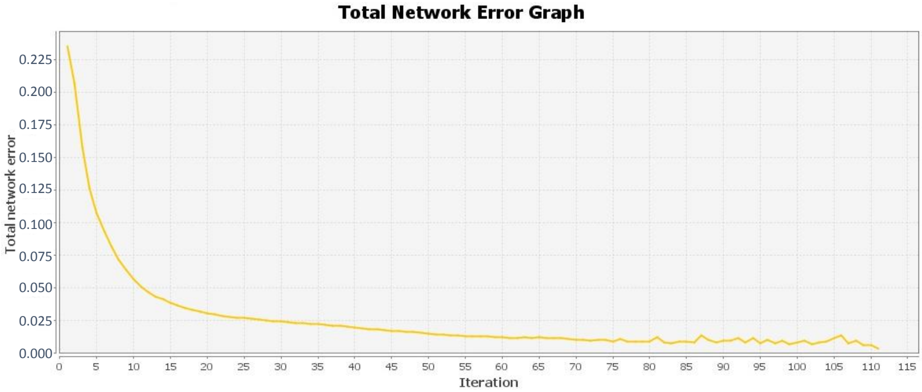


Figure 4. A total error network graph for the MLP model with respect to the number of training data set equals 7800, max error level equals 0.005, and learning rate equals 0.15

Source: NeurophStudio generated graph based on training process of the MLP.

Naïve Bayes classifier is calculated using a classical formula (Russell, 2010):

$$\text{classify}\left(e, y, u, t_1^j, \dots, t_{n-1}^j, r_{t_1^j}, \dots, r_{t_{n-1}^j}, s_{t_1^j}, \dots, s_{t_{n-1}^j}\right) =$$

$$\underset{c\in\{c_1,c_2\}}{\text{argmax}}\, p(C = c) \prod_{i=1}^{n-1} p\left(\langle e, y, u, t_i^j, r_{t_i^j}, s_{t_i^j}\rangle \,|C = c\right), \text{ where}$$

C denotes one of the possible classes representing the transaction finalisation (c_1) or non-finalisation (c_2). Note that t_f is not used in the formula as defined in the classical approach because it corresponds the final task, which occurrence probability is evaluated.

MLP uses data about browsing activity $e, y, t_1^j, \dots, t_{n-1}^j, r_{t_1^j}, \dots, r_{t_{n-1}^j}, s_{t_1^j}, \dots, s_{t_{n-1}^j}$ as inputs and classify them into two opposite classes – c_1 or c_2. A structure of a feedforward neural network corresponding to a general website, which browsing activity is depicted in Figure 1, is presented further (see Figure 3).

The MLP model presented in Figure 3 consists of one hidden layer with a neuron. Input and hidden layers have a bias (denoted by a bigger red circle). MLP uses back-propagation learning algorithm and hyperbolic tangent transfer function. A total error network graph for the considered MLP model (see Figure 4) shows an ability of the neural network to perform classification experiments at the predefined error level.

5. Experiment: recommendations based on analysis of user online behaviour

An abstract website, which browsing activity diagram is presented in Figure 1, was used for an illustration of the proposed technique.

Let us consider the situation when a statistical database contains 30 entries and user online activities form the following new data entry – see Table 2.

Table 2. An example of a fragment of actual browsing activity by a user (record R_{31})

	e	y	u	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	…
…	…	…		…	…	…	…	…	…	…	…	…	…	…	…
R_{31}	1	1	x	0	0	0	1	0	0	0	0	0	1	1	…

Source: own elaboration.

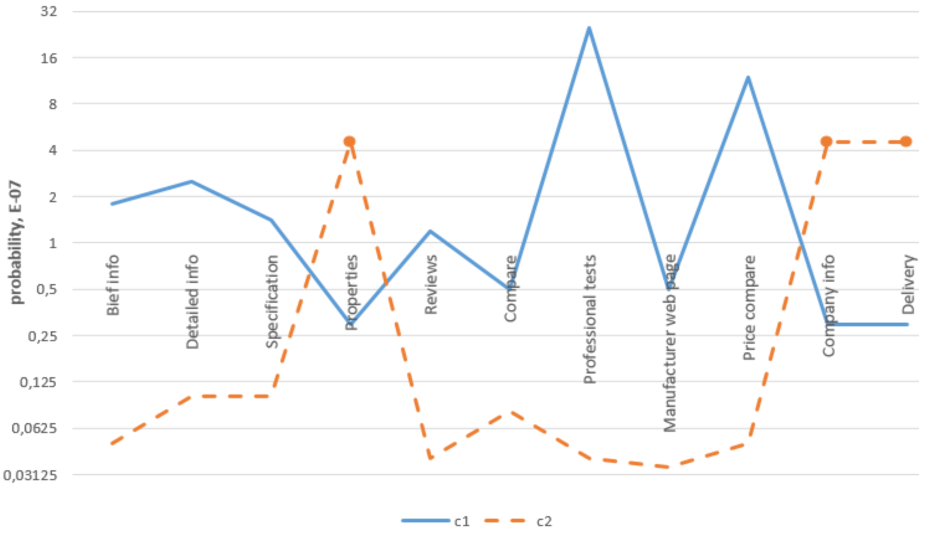


Figure 5. Results of experimental estimations of probabilities in order to forecast and recommend a web page with the highest impact on the transaction finalisation. Data series denoted by markers correspond to existing probability of the actual browsing activity

Source: own elaboration.

Naïve Bayes classifier estimates the probabilities for the class c_1 and c_2:

$$\text{classify}\left(e, y, t_1^{31}, \dots, t_{11}^{31}, r_{t_1^{31}}, \dots, r_{t_{11}^{31}}, s_{t_1^{31}}, \dots, s_{t_{11}^{31}}\right) =$$

$$= \underset{c \in \{c_1, c_2\}}{\text{argmax}}\, p(C = c) \prod_{i=1}^{11} p\left(\langle e, y, u, t_i^{31}, r_{t_i^{31}}, s_{t_i^{31}}\rangle \,|C = c\right) = c_2$$

(0.29E-07 < 4.54E-07)

Next, the procedure is being activated that experimentally estimates a web page to offer a user a visit, which has a maximal impact on the transaction finalisation. Figure 5 presents results of experimental estimations of probabilities of class 1 (solid line) and class 2 (dashed line). Figure 5 vividly shows – website-to-user interaction will advise visiting Professional tests web page as it has a maximal impact on the transaction finalisation:

$$\text{classify}(1,1,\text{x},0,0,0,1,0,0,\mathbf{1},0,0,1,1,\dots) = c_1$$

(24.73E-07 > 0.04E-07)

Let us consider the situation when statistical database contains 7800 entries and user online activities form the following new data entry – see Table 3.

Results of estimations by Naïve Bayes classifier:

$$\text{classify}\left(e, y, t_1^{7801}, \dots, t_{11}^{7801}, r_{t_1^{7801}}, \dots, r_{t_{11}^{7801}}, s_{t_1^{7801}}, \dots, s_{t_{11}^{7801}}\right) =$$

$$= \underset{c \in \{c_1, c_2\}}{\text{argmax}}\, p(C = c) \prod_{i=1}^{11} p\left(\langle e, y, u, t_i^{7801}, r_{t_i^{7801}}, s_{t_i^{7801}}\rangle \,|C = c\right)$$

$$= c_2$$

(4.61E-05 < 14.33E-05)

MLP model has classified given data as class 2 with a score 0.967.

Next, the procedure is being activated that experimentally estimates a web page to offer the user a visit, which has a maximal impact on the transaction finalisation. Figure 6 presents results of experimental estimations of data classification using MLP to class 1 (denoted as c1 (MLP)) or class 2 (denoted as c2 (MLP)). Several unvisited web pages were analysed – Brief info, Price compare and Company info. For comparative purposes, results of the experimental estimations using Naïve Bayes classification are presented too (classes 1 and 2 are denoted as c1 (NB) and c2 (NB) respectively). Figure 6 clearly shows that both classification methods work well and estimate the same outcome. A website-to-user interaction will advise visiting Price compare web page.

Table 3. An example of actual browsing activity by a user (record R_{7801})

	e	y	u	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	…
…	…	…		…	…	…	…	…	…	…	…	…	…	…	…
R_{7801}	1	2	y	0	1	1	1	1	1	1	1	0	0	1	…

Source: own elaboration.

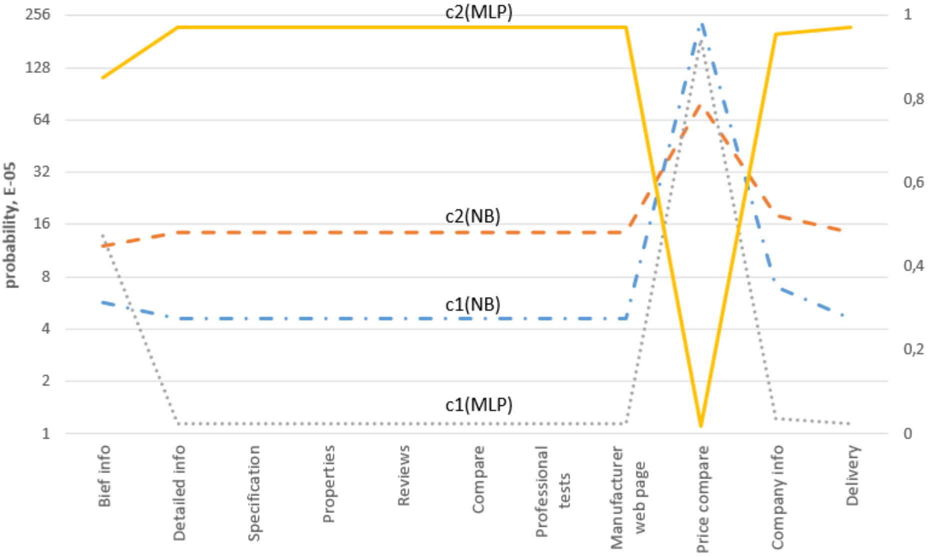


Figure 6. Results of experimental estimations of data classification using Multi-layer Perceptron and Naïve Bayes classifier

Source: own elaboration.

Note that Price compare functionality of a website is quite sensitive to any business. Moreover, as stated in Shopping Cart Abandonment report (Mulpuru, Hult and McGowan, 2010), 27% of e-customers cancel their purchases due to a comparison of prices from different retailers. According to the author's view, to overcome this issue, it is better to deal with this challenge on own website by applying the following policy. In order to offer a customer a product or a service at the lowest price, the mentioned Price compare functionality usually acts as follows. It offers an instant discount in case of an existence of a vendor that offers a lower price for the same product (if such a discount is possible with respect to a company price policy) or it includes vendors with poor customer ratings in the price compare list (if a discount cannot be applied). Obviously, giving a Price compare information, which is not profitable for a company – without any correction with regard to a website company – usually leads to a purchase cancellation, and should be avoided.

6. Conclusions

1. The proposed technique permits one to estimate website user actual on-site behaviour with respect to the transaction finalisation using Naïve Bayes classification and feedforward neural network – Multi-layer Perceptron model that are based on previous visitors' browsing activities.
2. The technique permits one to define actions to recommend a website user who theoretically has an impact on his/her decision to finalise a transaction.

Future works in this direction include widening the application of the technique to other areas as well as deepening the technique by applying other methods of multivariate analysis.

REFERENCES

ANGELETOU, S., ROWE, M., ALANI, H., (2011). Modelling and Analysis of User Behaviour in Online Communities. The Semantic Web – ISWC 2011 (pp. 35−50). Lecture Notes in Computer Science Volume 7031.

BUDNIKAS, G., (2015). Creation of user online behaviour analysis model for increase of an enterprise competitiveness. Rzeszów: In proceedings of VI Ogólnopolska Konferencja Naukowa „Społeczeństwo Informacyjne. Stan i kierunki rozwoju w świetle uwarunkowań regionalnych" (in press).

CLIFTON, B., (2012). Advanced Web Metrics with Google Analytics (3rd Edition ed.). Indianapolis: John Wiley & Sons.

DEMBCZYŃSKI, K., KOTŁOWSKI, W., SYDOW, M., (2009). Effective Prediction of Web User Behaviour with User-Level Models. Journal Fundamenta Informaticae, 89(2−3), 189−206.

DREJEWICZ, S., (2012). Zrozumieć BPMN. Modelowanie procesów biznesowych. Helion.

MULPURU, S., HULT, P., MCGOWAN, B., (2010, May 20). Understanding Shopping Cart Abandonment. Retrieved June 25, 2015, from https://www.forrester.com/Understanding+Shopping+Cart+Abandonment/fulltext/-/E-RES56827

NIKIFORAKIS, N., ACAR, G., SAELINGER, D., (2014). Browse at your own risk. Spectrum, IEEE, 51(8), 30−35.

ROBINSON, D. J. B. V., (2008). Online Behavioural Analysis and Modeling Methodology (OBAMM). Social Computing, Behavioural Modeling, and Prediction, 100−109.

RUSSELL, S. A., (2010). Artificial Intelligence: International Version: A Modern Approach (3 ed.). Pearson.

WHITE, R. W., CHU, W., HASSAN, A., HE, X., SONG, Y., WANG, H., (2013). Enhancing personalized search by mining and modeling task behavior. Proceedings of the 22nd International Conference on World Wide Web (pp. 1411–1420). ACM.

XIAN, X., CHEN, F., WANG, J., (2014). An Insight into Campus Network User Behavior Analysis Decision System. (pp. 537−540). Taichung: IEEE.

STATISTICS IN TRANSITION new series, Summer 2015
Vol. 16, No. 2, pp. 323-324

In Memoriam
Gunnar Kulldorff
06.12.1927 – 25.06.2015

Gunnar Kulldorff, an influential scientist and leading person in several scientific and professional organizations, passed away on 25 June 2015 in Umeå, Sweden. Gunnar was born in 1927 in Malmö, Sweden. He studied at Lund University and defended his PhD thesis on "Estimation from Grouped and Partially Grouped Samples" in 1961 at the same university. He worked as Lecturer of Statistics at Lund University until 1965. Then he moved to Umeå where he lived the rest of his life. Gunnar became Professor in the newly established Umeå University, first in Statistics and later in Mathematical Statistics. He served as the first Dean of the Faculty of Philosophy and further, the Dean of the Faculty of Mathematics and Natural Sciences. Throughout his career he was a highly appreciated teacher, researcher, leader and colleague.

Gunnar considered professional communication, exchange of scientific knowledge, cooperation and consolidation to be of great importance. The international dimension of science was a leading principle for him. As a member, often leader of professional organizations, he worked actively to strengthen this principle in the field of Statistical Science. He served as President of the International Statistical Institute (ISI) in 1989-1991. He was President of the Swedish Statistical Association and board member of the ISI, American Statistical Association and Bernoulli Society. He was elected as honorary member of the Finnish Statistical Society and the Estonian Statistical Society. In 2006 Gunnar was awarded a degree of Doctor Honoris Causa by the University of Vilnius. He was an elected member of the ISI since 1968.

Gunnar was a great visionary. In the position of President of the ISI he travelled in Asia, Africa and Latin America to promote activities in Statistical Science. However, he devoted much of his energy to the countries not so far from his homeland. His major mission was to spread statistical culture, professional education and scientific cooperation in the newly independent Baltic Countries and further, in Ukraine and Belarus. His activity created a vital network of survey statisticians that has now been functioning more than 20 years. Gunnar was the founder and long-term chair of the network, today called the Baltic-Nordic-Ukrainian Network on Survey Statistics. Since 1997, annual Workshops or Summer Schools on Survey Statistics have been organized in the Baltic countries, Ukraine and Belarus. A major effort has been the establishment of the series of regular international conferences. The latest, the Fourth Baltic-Nordic Conference on Survey Statistics, was held in Helsinki in August this year. There were participants from 15 countries (also several from Poland), indicating the strong international nature of the network activities. An article by Gunnar Kulldorff entitled "Statistical Science is International − And Survey Statistics is Cool and Hot", published in the Lithuanian Journal of Statistics in 2014, presents Gunnar's personal summary in one key area of his international activities.

Risto Lehtonen, University of Helsinki

Imbi Traat, University of Tartu

STATISTICS IN TRANSITION new series, Summer 2015
Vol. 16, No. 2, pp. 325−328

ABOUT THE AUTHORS

Budnikas Germanas received his doctor's degree in informatics (Physical Sciences) from Kaunas University of Technology (KUT), Lithuania in the area of formal specification, verification and knowledge engineering. He was a scientific visitor in Katholieke Universiteit Leuven (Belgium) at the Faculty of Business and Economics. He works at KUT and the branch of the University of Białystok in Vilnius. He is an author or co-author over 40 research papers. His current research interest is machine learning for discovering patterns of online behaviour.

Clark Robert is an Associate Professor at the National Institute for Applied Statistics Research Australia (NIASRA) at the University of Wollongong, Australia. He was previously a director in the Methodology Division of the Australian Bureau of Statistics. He has published research and managed major projects on sample design, the analysis of survey data, and ecological and environmental statistics. Together with Ray Chambers, he is the co-author of the textbook "An Introduction to Model-Based Survey Sampling".

Das Gitasree is a Professor of Statistics in North-Eastern Hill University, Shillong, Meghalaya, India. She has been working in the same university for more than 37 years. Her main research interest encompasses preliminary test estimators in double sampling with two auxiliary variables, combined regression preliminary test estimators in double sampling, sampling techniques in forestry with satellite remote sensing inputs, estimators in successive sampling. She is actively involved in projects focused on leaf area estimation of *pinus kesiya,* ordering of shoots with special reference to *Exbucklandia populnea,* nonlinear statistical models for culm growth of bamboos of different species, and timber species diversity hotspots with satellite remote sensing inputs.

Dwivedi Alok Kumar is an Assistant Professor of Biostatistics and Epidemiology at the Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center (TTUHSC), El Paso, Texas. He completed his post-doctoral fellowship in biostatistics from University of Cincinnati and Center for Clinical and Translational Science and Training, Cincinnati. Dr. Dwivedi's broad area of research includes biostatistics, epidemiology, clinical trial, cancer and neurological disorders. His research interests include application and advancement of zero inflated/hurdle models, predictive models, bootstrap methods, quantile regression, and meta-analysis. Currently, Dr. Dwivedi has been involved as co-investigator/biostatistician in 7 funded NIH/CPRIT studies. Dr. Dwivedi has published 50 peer reviewed manuscripts and 1 book chapter. He is a recipient of Dean's Young Investigator award 2013 by TTUHSC, El Paso.

Figueroa-Casas Juan B. is an Associate Professor at the Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center (TTUHSC), El Paso, Texas. He specializes in pulmonary disease, and has over 24 years of experience in the field of medicine. His research interests include mechanical ventilation, critical care of pulmonary disease, pulmonary mechanics measurements. He is President-elect for Faculty Council at the TTUHSC, El Paso.

Górna Karolina is a student of Doctoral Studies in Economics at the Faculty of Economic Science and Management of Nicolaus Copernicus University in Toruń. She has published 14 papers, mostly about economic convergence but also about usage of fuzzy sets in economic issues. In most of her investigations she emphasizes the importance of spatial dependencies among objects. Her area of interest includes tools and methods offered by spatial econometrics. She has participated in 17 conferences (international and national), presenting results of her investigations at most of them.

Górna Joanna is a student of Doctoral Studies in Economics at the Faculty of Economic Science and Management of Nicolaus Copernicus University in Toruń. She has published 14 papers, mostly about economic growth and determinants of convergence but also about usage of fuzzy sets in economic issues. In most of her investigations she points to the importance of location of chosen objects. Her scientific interests concern tools and methods offered by spatial econometrics. She has participated in 17 conferences (international and national), presenting her investigations at most of them.

Homa Magdalena PhD works at the Institute of Economic Sciences of the University of Wroclaw, Poland. Her work is focused on actuarial mathematics and application of stochastic processes in life insurance, with major interest in the possibility of actuarial modelling of complex insurance products, such as unit-linked insurance, participating insurance or multi-state life insurance. In particular, it addresses the issue of risk arising from extended actuarial risk which is observed, and financial risk incorporated into the actuarial framework.

Mallawaarachchi Indika is a Research Associate in the Division of Biostatistics and Epidemiology at Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center (TTUHSC), El Paso, Texas. He graduated from University of Texas at El Paso with his Master's degree in Statistics. He is a SAS certified statistician. He has been co-authored in a number of peer reviewed manuscripts. His research interests are applications of biostatistics and epidemiological methods including longitudinal data analysis, survival analysis, regression methods, statistical modelling and statistical programming.

Markowska Małgorzata is an Associate Professor at the Regional Economics Chair of Wrocław University of Economics. She is a member of Polish Classification Society and Polish section of Regional Studies Association. Her research deals with econometric measurement, evaluation, variability and dynamics of development, competitiveness, knowledge-based economy, smart specializations, convergence and innovativeness in European regional space. As an author or co-author she has published more than 100 scientific papers and 25 chapters in books, and recently her own dissertation "Dynamic Taxonomy of Regions' Innovativeness". She has participated in 10 scientific projects financed by Polish National Centre of Science and the European Union, and in projects for governmental, local administration and business units.

Molefe Wilford is a lecturer in the Department of Statistics at the University of Botswana. He received his PhD degree from the University of Wollongong in Australia, New South Wales in 2011. His main research area is sample survey methodology. He is a member of the Democracy Research Project and AfroBarometer research consortium.

Morales Gonzalez Angel is a surgeon in El Paso, Texas and is affiliated with multiple hospitals in the area, including Del Sol Medical Center and Las Palmas Medical Center. He has been in practice for 13 years. He is a Fellow of the American College of Surgeons and the American Society of Colon and Rectal Surgeons. His main clinical interests involve the management of colorectal cancer, inflammatory bowel disease, and ano-rectal disorders, as well as minimal invasive approaches to these disease processes. Before joining Rio Grande Surgeons, Dr. Morales was Assistant Professor of Surgery at the Paul L. Foster School of Medicine.

Mościbrodzka Monika, PhD in Mathematical Sciences (Polynomially growing pluriharmonic functions on symmetric Siegel domains); she works at the Department of Functional Analysis, Institute of Economic Sciences, the University of Wroclaw, Poland. Her major research interests include the analysis of the capital market, the measurement and evaluation of the effectiveness and risks of investment funds, including using tools of multidimensional comparative analysis.

Obrębalski Marek, PhD, works at the Department of Regional Economics, Wroclaw University of Economics. His main research domains are social, economic and spatial problems of regional and local economy. M. Obrębalski has published more than 180 research papers in international or national journals. For several years his research and practical experiences have been focused on the area of regional statistics. He is a co-author of Local Data Bank – a nationwide database in a cross-section of territorial units, among other things. He is also interested in local and regional self-government, urban policy and local finance. In the period 2006-2010 he was the Mayor of Jelenia Góra, and since 2010 he has been a councillor of the Voivodship Regional Council of Lower Silesia.

Ralte Zoramthanga is a PhD student in statistics in North Eastern Hill University, Shillong, India (supervised by Prof. Gitasree Das), conducting research on the development of new estimators in successive sampling and their applications to real life data, and on the application of successive sampling under stratified random sampling. He participated in the 68th Annual Conference on

Statistics and Informatics in Agricultural Research organized by ICAR-Indian Agricultural Statistics Research Institute in 2015, and was awarded an appreciation certificate for presentation. His major research interests are focused on the area of developing estimators in stratified random sampling and successive sampling.

Shangodoyin D. K. is a full professor of statistics at the University of Botswana, Botswana. His principal strengths are in data mining, econometrics, Bayesian modelling, multivariate analysis and time series in which he has several publications that address a varied set of challenging issues requiring a thorough background in statistical theory and computing.

Tarwater Patrick M. is a Professor of Biostatistics and Epidemiology at the Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center (TTUHSC), El Paso, Texas. He holds the position of Adjunct Professor in the Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland. He is a Fellow of American College of Epidemiology, Raleigh, NC. He has published over 90 peer-reviewed manuscripts. Currently, he is participating as a Co-Investigator & Principal Biostatistician in 3 R01 NIH studies. Dr. Tarwater's experience ranges from the conduct of longitudinal cohort studies in the public health sciences to experimental design in the basic sciences.

Walesiak Marek, Professor at Wroclaw University of Economics, Department of Econometrics and Computer Science in Jelenia Góra (Poland). His main fields of research interest include classification and data analysis, multivariate statistical analysis, and marketing research. Actively involved in academic and international cooperation and research projects, organizational leadership and institutional management: Head of Department of Econometrics and Computer Science (1997 –), Dean of Faculty of Regional Economics and Tourism (since 2012 Faculty of Economics, Management and Tourism), Polish Statistical Association (PTS), The Council of Section on Classification and Data Analysis of Polish Statistical Society (SKAD), International Federation of Classification Societies (IFCS), The Committee of the Statistics and Econometrics of Polish Academy of Sciences, and Vice Editor-in-Chief of Polish Statistical Review „Przegląd Statystyczny".

Wilk Justyna is an Assistant Professor in the Department of Econometrics and Computer Science at Wroclaw University of Economics, Poland. Her major research interests cover the application of symbolic data analysis, multivariate data analysis and spatial econometrics in regional research. She carries out comparative and dependence studies to reveal the determinants and problems of socio-economic development. She is particularly interested in the quality of data and proposed applying symbolic data analysis to reducing the scale effect of modifiable areal unit problem. She is a co-author of symbolic DA-package of R-CRAN for symbolic data analysis.

GUIDELINES FOR AUTHORS

We will consider only original work for publication in the Journal, i.e. a submitted paper must not have been published before or be under consideration for publication elsewhere. Authors should consistently follow all specifications below when preparing their manuscripts.

Manuscript preparation and formatting

The Authors are asked to use *A Simple Manuscript Template (Word or LaTeX) for the Statistics in Transition Journal (published on our web page:* http://stat.gov.pl/en/sit-en/editorial-sit/).

- ***Title and Author(s)***. The title should appear at the beginning of the paper, followed by each author's name, institutional affiliation and email address. Centre the title in **BOLD CAPITALS**. Centre the author(s)'s name(s). The authors' affiliation(s) and email address(es) should be given in a footnote.
- ***Abstract.*** After the authors' details, leave a blank line and centre the word **Abstract** (in bold), leave a blank line and include an abstract (i.e. a summary of the paper) of no more than 1,600 characters (including spaces). It is advisable to make the abstract informative, accurate, non-evaluative, and coherent, as most researchers read the abstract either in their search for the main result or as a basis for deciding whether or not to read the paper itself. The abstract should be self-contained, i.e. bibliographic citations and mathematical expressions should be avoided.
- ***Key words***. After the abstract, *Key words* (in bold italics) should be followed by three to four key words or brief phrases, preferably other than used in the title of the paper**.**
- ***Sectioning***. The paper should be divided into sections, and into subsections and smaller divisions as needed. Section titles should be in bold and left-justified, and numbered with **1.**, **2.**, **3.**, etc.
- ***Figures and tables***. In general, use only tables or figures (charts, graphs) that are essential. Tables and figures should be included within the body of the paper, not at the end. Among other things, this style dictates that the title for a table is placed above the table, while the title for a figure is placed below the graph or chart. If you do use tables, charts or graphs, choose a format that is economical in space. If needed, modify charts and graphs so that they use colours and patterns that are contrasting or distinct enough to be discernible in shades of grey when printed without colour.
- ***References.*** Each listed reference item should be cited in the text, and each text citation should be listed in the References**.** Referencing should be formatted after the Harvard Chicago System – see http://www.libweb.anglia.ac.uk/referencing/harvard.htm. When creating the list of bibliographic items, list all items in alphabetical order. References in the text should be cited with authors' name and the year of publication. If part of a reference is cited, indicate this after the reference, e.g. (Novak, 2003, p.125).