

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