Statistical evaluation of research and development activity of the EU countries with regard to the accuracy of statistical data

Małgorzata Stec[1]

Abstract

The aim of this paper is to evaluation of research and development activity of the EU countries, including the accuracy of statistical data. 8 diagnostic variables describing this economic phenomenon in years 2010 and 2017 were used for the empirical study. Min-max normalisation was employed to perform a linear ordering of objects. Because a quality of obtained results depends, among others, on a quality of data used for calculations, the study also contains an evaluation ofinfluence of the accuracy of statistical data on a result of the linear ordering of the EU countries with regard to the level of R&D. Due to the fact that the problem of influence of the accuracy of diagnostic data on the results of the taxonomic analyses does not have any thorough methodology in the economic research yet, an original approach for the analyses of the subject was proposed. For this purpose, the uncertain theory of measurements, which is used in technical sciences, was employed and adjusted to the specificity of methods of multidimensional comparative analyses. In view of measuring scales used in taxonomy (among others, permissibility of mathematical operations on these scales), the Monte Carlo method was employed in order to determine the uncertainty ranges of a synthetic measure.

Keywords: *R&D, European Union countries, synthetic measure, Monte Carlo method*
JEL Classification: *O32, O52, C15*

1. Introduction

The characteristic trait of the contemporary highly developed countries is an economy based on knowledge and new technologies (Knowledge-based Economy). Intellectual potential, knowledge and new technologies are the factors that are decisive forcountries and regions to have chances for development and competitiveness. Research and development activity have a crucial role in ability to create knowledge and to remold it into new technologies, products and services (Bilbao-Osorio and Rodríguez-Pose, 2004; Bravo-Ortega and Marin, 2011; Cetenak and Oransay 2017; Coccia, 2012; Cunningham and Link, 2016; Falk, 2007; Grzebyk and Stec, 2015; Hall et al., 2010; Rodríguez-Pose, 2001).

Therefore, construction of ranking of the EU countries in terms of the level of progression of research and development activity, is an interesting research problem. While undertaking a statistical evaluation of the EU countries in matters of the researched phenomenon and other complex phenomena[2] by means of widely available statistical data, one should pay attention to a problem of their "accuracy". Statistical information used in the research will decide about the final results. The way of obtaining statistical data by the institutions collecting such data (e.g.

[1] Corresponding author: University of Rzeszów, Faculty of Economics, Department of Quantitative Methods and Economic Informatics, 2 M. Ćwiklińskiej St., 35-601 Rzeszów, Poland, e-mail: malgorzata.a.stec@gmail.com.

[2] Complex economic phenomena are the phenomena not subjected to the direct measurement of, e.g. the socio-economic development of countries, regions, etc., standard of living, job market situation, financial situation of local government unit, companies, banks, etc. (Pawełek, 2008).

Eurostat and statistical offices of each country) results inproducing errorsthat are impossible to omit. Thus, one should be aware of their existence and, whenever possible, should include their impact on the results of undertaken research and drawn conclusions. The aim of this paper is to evaluate research and development activity of the EU countries, including the accuracy of statistical data. 8 diagnostic variables describing this economic phenomenon in years 2010 and 2017 were used for the empirical study. The linear ordering of EU countries was done by the min-max normalisation. Moreover, an influence of uncertainty of measurement of diagnostic variables on the values of synthetic measure was investigated. Monte Carlo method was adopted for this purpose.

2. Theoretical basis for statistical data accuracy

The quality concept of official statistics is based on the definition of the European Statistical System definition of the quality and defined on the basis of the following 6 criteria: relevance, accuracy, timeliness and punctuality, accessibility and clarity, comparability, coherence. Accuracy denotes the closeness of computations or estimates (after collecting, processing, imputation, estimation of data and the like) to the exactor true values. The difference between these two values is the error (Vademecum of quality in official statistics, 2012).

The way of obtaining statistical data has an impact on their accuracy. The most accurate data are gathered in official national registers, while less accurate data are collected by means of sample surveys. In yearbooks, there is no information about estimated values of uncertainty that burden a given statistical variable. Therefore, the author made such estimations on the basis of the available knowledge about the way of obtaining data used to calculate a precise variable.

In case of statistical data, the real value of measured variable is usually unknown. Then, the accuracy of statistical data (similarly as in technical sciences) can be equated with the measurement uncertainty (JCGM/WG 1, 2008).

The formal definition of the term ‘uncertainty of measurement’ is as follows: uncertainty (of measurement) parameter, associated with the result of a measurement, that characterises the dispersion of the values that could reasonably be attributed to the measurand. The parameter a uncertainty of measurement may be, for example, a standard deviation called standard measurement uncertainty (or a specified multiple of it), or the half-width of an interval, having a stated coverage probability (Balazs, 2008).

A variable read from a yearbook or downloaded from a data base is taken as a nominal value of a variable X_n. Estimated value of uncertainty u_{Bc} determines the upper and lower limit of a nominal value in which, with an established probability, a real value of this variable can be found:

$$X_R = X_n \pm u_{Bc} \tag{1}$$

where: X_R – limits of the range in which the real value of a variable is; X_n – variable’s nominal value; u_{Bc} – estimated value of variable’s uncertainty.

Assuming that the respective variables are burdened by uncertainty (real values are unknown, only expected values – estimates – and distribution are known), one should analyse if the ranges are not too wide to "blur" the difference between researched objects. In the Fig. 1 a graphical representation of a nature of comparison of assigned values to variables is presented for two objects (countries) whose ranges of uncertainty "overlap". Such case will occur if for a variable-booster, an object's upper limit of uncertainty occupying a lower position in a ranking has a higher value, than the lower limit of uncertainty of an object positioned higher.

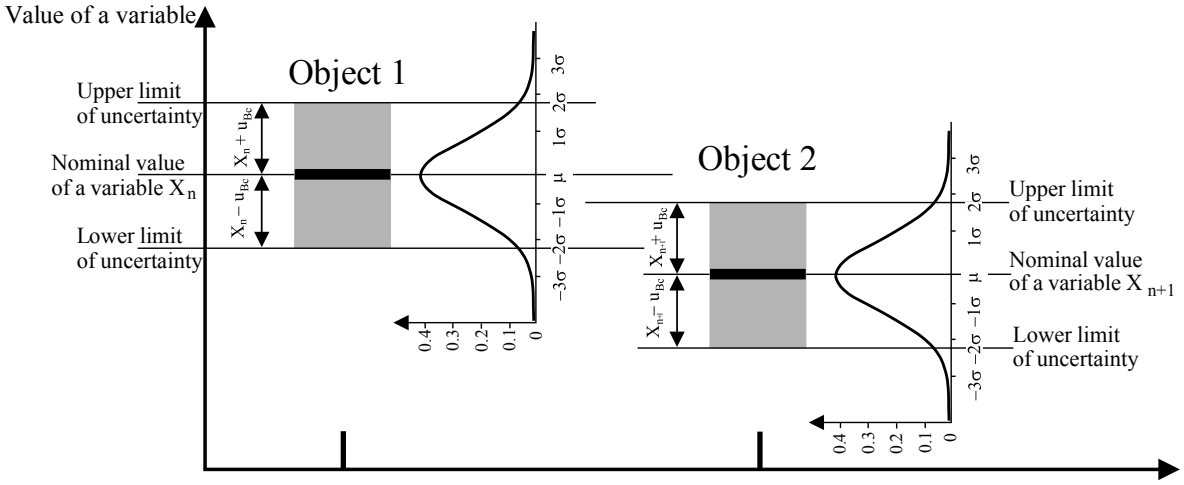

Fig. 1. A nature of estimation of uncertainty of statistical value

The overlapping of ranges of uncertainty can take place if assigned valuesof variables of two or more objects (countries) differ little from each other, while the estimated uncertainties are relatively big. With small differentiations of assigned values of variables of objects, there may be a situation in which a few objects are characterised by similar values of one variable, which may hinder an interpretation of real differences between these objects. Analogical situation also refers to synthetic measures.

The range of uncertainty for determining a diagnostic variable can be described by a density function of normal distribution described by a dependence 2.

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X-m)^2}{2\delta^2}} \tag{2}$$

Area under a curve of density function $f(X)$ is a measure of probability of finding an assigned value to variable in a determined range. It is possible to calculate a probability of event, in which, as a result of change of assigned values of variables, a position of researched objects (countries) changes. Fig. 2 presents such case for a diagnostic variable $X1$ (Research and development expenditure (in % of GDP)), in relation to Hungary and Portugal. If a real value of variable $X1$ for Hungary ($X1$=1.35) was nevertheless lower than a range value X_g, whereas for Portugal ($X1$=1.32) higher than X_g; therefore, these countries would exchange their positions in the ranking. Due to the fact that both events are independent, the total probability would be a product of probabilities for

respective ranges determined by the density function of normal distribution. Change of positions of objects is also possible in other cases which were not described in this article.

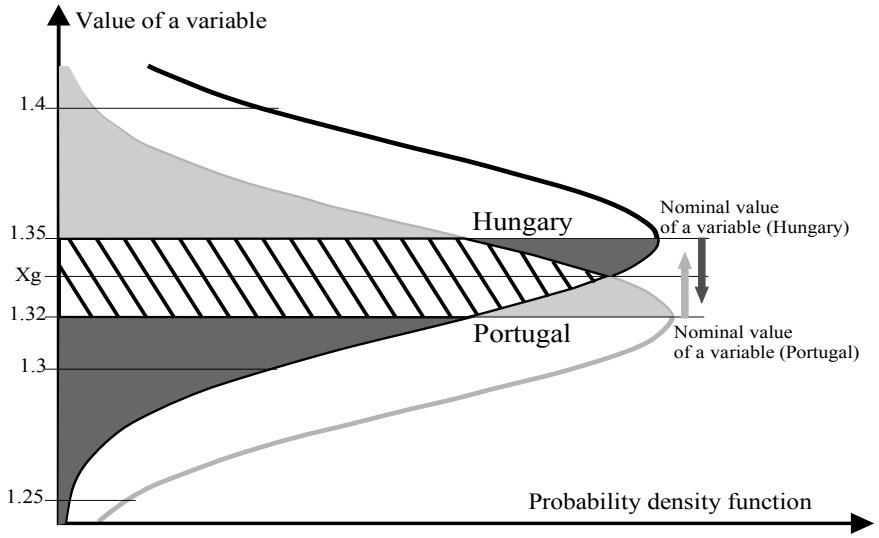

Fig. 2. Density function of normal distribution for Hungary and Portugal for variable *X*1

Analogical situation that takes place for diagnostic variables can also occur in case of an analysis of uncertainty range for a calculated synthetic measure.

3. Methods applied

In this paper, the min-max normalisation was used in order to calculate a value of synthetic measure for all countries in terms of research and development activity in year 2010 and 2017 (Kukuła, 2000).

A normalisation of the variable values was conducted using the following formulas:

$$z_{ij} = \frac{x_{ij} - \min_i\{x_{ij}\}}{R_j} \quad \text{for stimulating factors} \tag{3}$$

$$z_{ij} = \frac{\max_i\{x_{ij}\} - x_{ij}}{R_j} \quad \text{for non-stimulating factors} \tag{4}$$

where:z_{ij} – normalized value of *j*-th variable for the *i*-th object, x_{ij}, – value of *j*-th variable for the *i*-th object, R_j – range for the *j*-th variable.

It should be emphasised that the calculations were done in a dynamic manner, using the so called 'object-periods'. The synthetic measure was calculated as an arithmetic mean of the normalised value of variables:

$$MS_i = \sum_{j=1}^{m} z_{ij} \tag{5}$$

where: MS_i – synthetic measure in *i*-th object, m – number of variables.

Due to the fact that the employed method to calculate the synthetic measure leads to change of the measuring scale, the calculation of the uncertainty of the synthetic measure by analytic method would provide false results. That is why, the Monte Carlo[3] method was employed and calculations were performed in the *R* application (Walesiak and Gatnar, 2009).

In order to calculate the value of uncertainty of the synthetic measure, it was concluded that for a sample big enough (the calculations were performed on a set of data counting 1000 for each object), standard deviation can be considered as a measure of distribution identified with a range of uncertainty of the synthetic measure. The following algorithm of procedure was chosen (Stec, 2017; Stec and Wosiek, 2018):

- for each diagnostic variable, 1000 values were drawn for every object (28 countries) which fulfilled the following conditions:
  - value of each drawn variable was comprised in the assumed range of uncertainty created around the nominal value for this variable,
  - drawn values for each variable had normal distribution,
- from the drawn variables, sets of data were created (1000 sets for each object),
- drawn sets of data underwent normalisation,
- on the basis of the normalised set of data, the synthetic measures were calculated (1000 values of partial measures),
- from 1000 set synthetic, partial measures standard deviation was calculated, which constituted the measure of uncertainty of synthetic measure.

The above mentioned procedure allowed for calculation of nominal values of synthetic measures for each object (country) and their uncertainty.

4. Diagnostic variables employed in the research

The evaluation of research and development activity for 28 EU countries was done with an employment of 8 diagnostic variables: $X1$-Research and development expenditure (% of GDP) (S); $X2$-Intramural R&D expenditure (GERD) by source of funds (Business enterprise sector-% of total GERD) (S); $X3$-Share of government budget appropriations or outlays on research and development (% of total) (S); $X4$-Research and development personnel (Full time equivalent-% of the labour force) (S); $X5$-High-tech exports % of exports (S); $X6$-Employment in high- and medium-high technology manufacturing sectors and knowledge-intensive service sectors (% of total employment) (S); $X7$-Human resources in science and technology (HRST) (% of active population) (S); $X8$-Patent applications to the European patent office (EPO) by priority year per 100 thous. population (S)[4].

[3] The Monte Carlo method solves a numerical problem by performing calculations on random variables, it is a tool for solving quantity problems, when analytical methods based on formulas, estimators, etc., fail. (Kopczewska et al., 2016; Liu, 2008; Niemiro, 2013).

[4] (S)-stimulant.

Empirical data describing the states under study were extracted from the Eurostat database[5]. Also, an influence of the uncertainty of diagnostic variables on the results, regarding ordering of objects in terms of values of proposed variables was assessed. Table 1 shows a compilation of investigated values of uncertainty and a number of cases of overlapping (collisions) ranges of uncertainty for each diagnostic variables caused by too small difference between their nominal values in relation to the calculated uncertainty. The uncertainty of data was estimated on the basis of research sample (for Poland) and the uncertainty resulting from rounding. At the same time, a simplifying assumption was taken in order to acknowledge that for the purposes of the article, it is adequate to apply the same uncertainties for all countries for the analysed years.

Table 1. Comparison of estimated uncertainty values of diagnostic variables and overlapping ranges of uncertainty of diagnostic variables (2017)

	Diagnostic variables							
	***X*1**	***X*2**	***X*3**	***X*4**	***X*5**	***X*6**	***X*7**	***X*8**
Overall uncertainty of a variable	2.5%	1.0%	0.8%	1.0%	0.8%	0.7%	0.5%	0.5%
Overall number of conflicts	21	21	11	12	9	13	19	2
Number of conflicts with probability > 0,05	11	12	7	10	5	10	9	1

5. Empirical results

The statistical evaluation of research and development activity of the EU countries in years 2010 and 2017 was done in a traditional way by analysing value of a synthetic measure and by taking into account uncertainties of a given synthetic measure.

Considering only values of synthetic measure and rankings of the EU countries based on them in terms of research and development activity, it can be noticed that:

- in 2010, as regards the R & D activity, the leading positions in the rank of EU countries were taken by: Finland, Germany, Sweden, Denmark and Luxembourg. Last positions were taken by: Bulgaria, Romania, Latvia, Greece and Poland.
- in 2017, the leading positions in the rank of EU countries were taken by: Germany, Denmark, Sweden Austria and Finland. Last positions were taken by: Latvia, Romania, Bulgaria, Cyprus and Greece.

In 2017 in comparison to 2010, the highest advancement in the ranks of the EU countries in terms of R+D was noted for: Poland (from 24th position in 2010 to 19th in 2017), Austria and Ireland (advancement of 3 places). The same position in both years was held by: Estonia,

[5] https://ec.europa.eu/eurostat/data/database (accessed on 07.01.2019).

Hungary, Romania, Slovenia and Sweden. Countries that lost their position were: Luxembourg (from 5th in 2010 to 13th in 2017) and Finland (from 1st to 5th position).

In the second variant of the study, the aim was to verify if the inclusion of uncertainty as to the value of diagnostic variables used in the research, and inclusion of uncertainty for an aggregate measure can have an influence on a change of a position held by the individual EU countries in terms of research and development activity.

Fig. 3 presents results of ordering of the EU countries in regard to the level of research and development activity in 2010; whereas fig. 4 presents the same ordering for 2017.

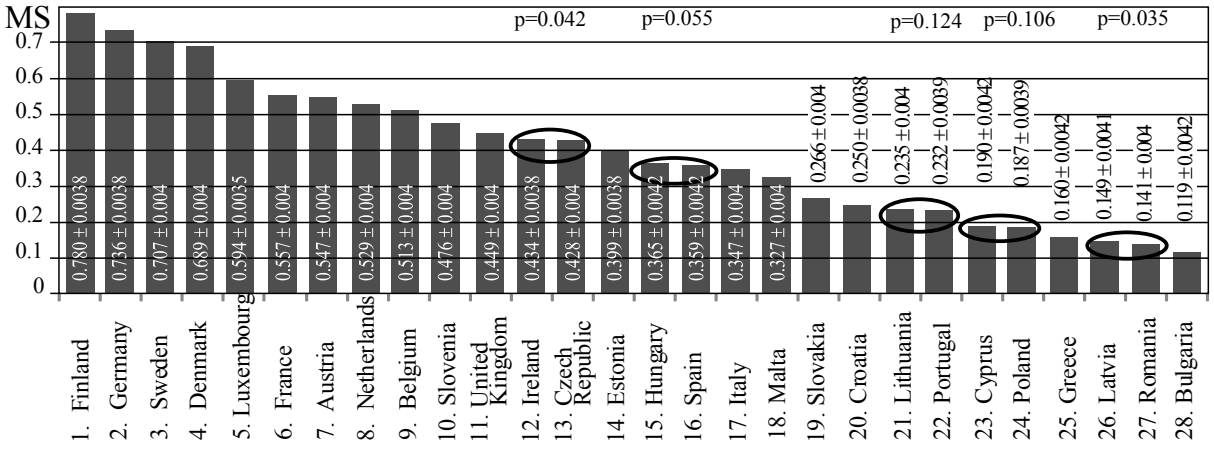

Fig. 3. Results of linear ordering of the EU countries in terms of the level of research and development activity (2010)

The cases in which inclusion of uncertainty resulted in overlapping of uncertainty ranges (collisions) between countries were emphasised. This situation took place in 2010 in: Ireland-Czech Republic, Hungary-Spain, Lithuania-Portugal, Cyprus-Poland, Latvia-Romania. The analysis of probability of position change caused by data errors shows that only in three cases the probability exceeded 0,05.

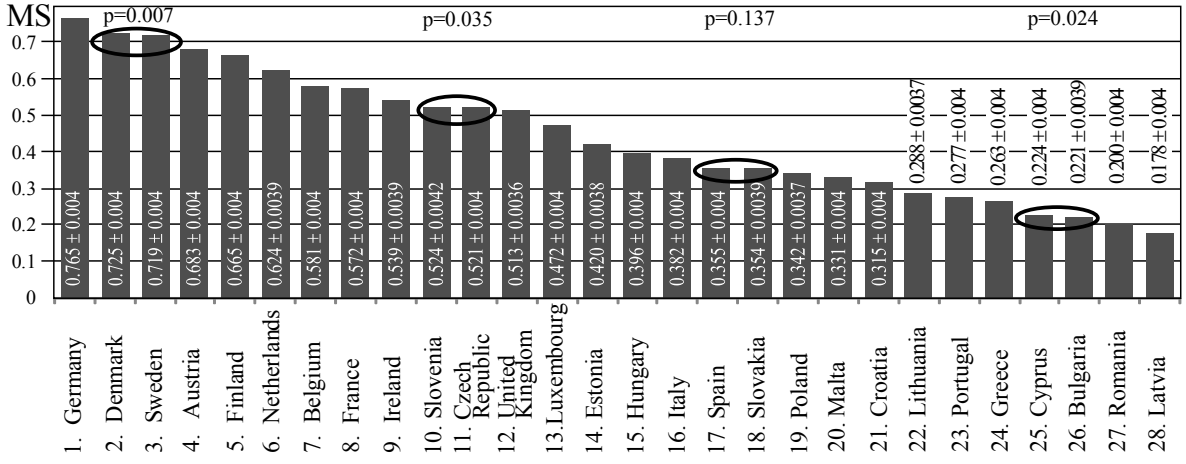

Fig. 4. Results of linear ordering of the EU countries in terms of the level of research and development activity (2017)

For 2017, four collisions were noticed for: Denmark-Sweden, Slovenia-Czech Republic, Spain-Slovakia, Cyprus-Bulgaria. Only in one case, the probability of position change caused by data errors exceeded 0,05.

6. Conclusions

The following conclusions can be drawn based on the carried out research:

- The statistical evaluation of research and development activity of the EU countries was done with an employment of 8 diagnostic variables. The min-max normalisation was employed for the empirical studies.
- The results confirm the diversity of EU countries in terms of research and development activity. In 2017 Germany, Denmark, Sweden Austria and Finland were the leaders interms of research and development activity. The lowest level of the analyzed phenomenon is represented by the following countries: Latvia, Romania, Bulgaria, Cyprus and Greece.
- The conducted research shows that the method of obtaining statistical data influences the value of uncertainty estimation.The method of analysis of diagnostic variables and a synthetic measure proposed in the article, including uncertainty ranges, allows to determine the trust range for the obtained results.The analysis does not change the order of ordering, nor does it verify the actual ordering of objects.Taking into account the uncertainties in the value of synthetic measures may influence the final conclusions resulting from the research.

References

Balazs, A. (2008). International vocabulary of metrology-basic and general concepts and associated terms. *Chemistry International*, 20–1.

Bilbao-Osorio, B., & Rodríguez-Pose, A. (2004). From R&D to innovation and economic growth in the EU. *Growth and Change*, 35(4), 434–455.

Bravo-Ortega, C., & Marin, A.G. (2011). R&D and productivity: A two way avenue?. *World Development*, 39(7), 1090–1107.

Cetenak, O.O., & Oransay, G. (2017). Economic Growth and Dynamic R&D Investment Behavior. In: *Global Business Strategies in Crisis*, 243–259. Cham: Springer.

Coccia, M. (2012). Political economy of R&D to support the modern competitiveness of nations and determinants of economic optimization and inertia. *Technovation*, 32(6), 370–379.

Cunningham, J.A., & Link, A. N. (2016). Exploring the effectiveness of research and innovation policies among European Union countries. *International Entrepreneurship and Management Journal*, 12(2), 415–425.

Eurostat database:https://ec.europa.eu/eurostat/data/database (accessed on 07.01.2019).

Falk, M. (2007). R&D spending in the high-tech sector and economic growth. *Research in Economics*, 61(3), 140–147.

Grzebyk, M., & Stec, M. (2015). Sustainable development in EU countries: concept and rating of levels of development. *Sustainable Development*, 23(2), 110–123.

Hall, B.H., Mairesse, J., & Mohnen, P. (2010). *Measuring the Returns to R&D*. In: *Handbook of the Economics of Innovation* (Vol. 2, pp. 1033–1082). North-Holland.

JCGM/WG 1 2008 Working Group. (2008). Evaluation of measurement data–guide to the expression of uncertainty in measurement. In: *Tech Rep JCGM 100: 2008 (BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML*.

Kopczewska, K., Kopczewski, T., & Wójcik, P., (2016). *Metody ilościowe w R: aplikacje ekonomiczne i finansowe*. Warszawa: CeDeWu, 221.

Kukuła, K. (2000). *Metoda unitaryzacji zerowanej*. Warszawa: Wydawnictwo Naukowe PWN.

Liu, J. S. (2008). *Monte Carlo strategies in scientific computing*. Springer Science & Business Media.

Niemiro, W. (2013). *Symulacje stochastyczne i metody Monte Carlo*. Warszawa: Wydawnictwo Uniwersytetu Warszawskiego.

Pawełek, B. (2008). Metody normalizacji zmiennych w badaniach porównawczych złożonych zjawisk ekonomicznych. *Zeszyty Naukowe/Uniwersytet Ekonomiczny w Krakowie. Seria Specjalna, Monografie*, (187).

Rodríguez-Pose, A. (2001). Is R&D investment in lagging areas of Europe worthwhile? Theory and empirical evidence. *Papers in regional science*, 80(3), 275–295.

Stec, M. (2017). *Taksonomiczna analiza poziomu rozwoju społeczno-gospodarczego województw Polski*. *Studium przypadku-województwo podkarpackie*, Rzeszów: Wydawnictwo Uniwersytetu Rzeszowskiego, 87–99.

Stec, M. & Wosiek, M. (2018). Evaluation of the socio-economic situation of European Union countries, taking into account accuracy of statistical data. In: Papież M. and Śmiech S. (Eds.), *The 12th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomen. ConferenceProceedings*. Cracow: Foundation of the Cracow University of Economics, 483–492.

Vademecum of quality in official statistics. (2012). Warsaw:http://bip.stat.gov.pl/en/activity-of-official-statistics/quality-in-statistics/ (accessed on 04.01.2019).

Walesiak, M., & Gatnar, E. (Eds.). (2009). *Statystyczna analiza danych z wykorzystaniem programu R*. Warszawa: Wydawnictwo Naukowe PWN.