THE INFLUENCE OF THE ACCURACY OF STATISTICAL DATA ON THE RESULTS OF A CLASSIFICATION OF EU COUNTRIES IN TERMS OF INNOVATION

Małgorzata Stec, Ph.D., Associate Professor
University of Rzeszów
Institute of Economics and Finance
Department of Quantitative Methods and Economic Informatics
M. Ćwiklińskiej 2 St., 35-601 Rzeszów, Poland
e-mail: malgorzata.a.stec@gmail.com
ORCID: 0000-0003-0185-4510

Received 30 October 2019, Accepted 19 March 2020

Abstract

Research background: The article attempts to include the accuracy of statistical data in a synthetic evaluation and classification of EU countries in terms of innovation.

Purpose: The aim of the article is to evaluate an influence of the accuracy of statistical data on a classification of EU countries in terms of innovation.

Research methodology: The research employed diagnostic variables determining the innovation of EU countries and a methodology proposed by the European Commission in the European Innovation Scoreboard 2019. The influence of the uncertainty of the measurement of the diagnostic variables on the Summary Innovation Index of EU countries was evaluated. In order to do this, a procedure employing the Monte Carlo method was proposed.

Results: Taking into account the uncertainty of the measurement of variables in the evaluation of the innovation of EU countries resulted in qualifying one of the countries to another innovation group.

Novelty: The article draws attention to an important but often neglected problem related to the accuracy of statistical data used in research, and the evaluation of their influence on the calculation of a value of synthetic measure (based on the innovation of EU countries).

Keywords: innovation of EU countries, Summary Innovation Index, accuracy of statistical data, uncertainty of statistical data, the Monte Carlo method

JEL classification: C01, C15, O30, O52, O57
Introduction

Nowadays, innovation is one of the most important factors of economic development in the world. The ability of countries’ economies and economic entities to create, implement and adopt innovations affects their competitiveness. That is why the European Commission pays a lot of attention to innovation policy, treating it as a tool for strengthening the EU economies. A methodology of measurement and evaluation of the innovation of EU countries is quite difficult and complex. The problem is credible, comparable and also accurate innovation indicators for individual countries.

The aim of the article is to evaluate the influence of the accuracy of statistical data on the classification of EU countries in terms of innovation. The evaluation of innovation of EU countries was based on the Summary Innovation Index (SII) which was calculated with 27 indicators presented in the European Innovation Scoreboard 2019.

The evaluation of the influence of the uncertainty of measurement of diagnostic variables on SII values for EU countries and their classification was made using the Monte Carlo method.

1. Innovations – review of definitions

The word ‘innovation’ from the beginning is derived from the Latin word ‘innovare’ which means into new (Costello, Prohaska, 2013). Many definitions of innovation can be found in literature which pertain to a product, entrepreneurship, organization, the whole economy, etc. That indicates a variety and many-sidedness of a definition of innovation (Baregheh, Rowley, Sambrook, 2009; Cumming, 1998; Fischer, 2001; Gault, 2018; Godin, 2008; Gust-Bardon, 2012; Kogabayev, Maziliauskas, 2017).

Innovation is production or adoption, assimilation, and exploitation of a value-added novelty in economic and social spheres; renewal and enlargement of products, services, and markets; development of new methods of production; and the establishment of new management systems. It is both a process and an outcome (Edison, Bin Ali, Torkar, 2013).

According to G.D. Sardana (2016) innovation is more than just gaining knowledge it is continuous learning and that the knowledge also has to be translated into actions.

The most popular definition of innovation is presented in the Oslo Manual.

The Oslo Manual distinguishes between innovation as an outcome (an innovation) and the activities by which innovations come about (innovation activities). This edition defines an innovation as “a new or improved product or process (or combination thereof) that differs
significantly from the unit’s previous products or processes and that has been made available to potential users (product) or brought into use by the unit (process)” (Manual, 2018).

Yet the innovation of economy is an ability and willingness of the economic entities to constantly search, and use in practice, the results of scientific research, research and development works, new concepts as well as ideas and inventions; an ability to improve and develop the technologies of material and non-material (services) production; to implement new methods and techniques in organization and management; to improve and develop an infrastructure and body of knowledge (Frankowski, Skubiak, 2012).

2. Indicators and method of research of the innovation of countries

For 18 years, the European Commission has been conducting research on the innovation of EU countries in order to present the results in a form of the European Innovation Scoreboard (EIS).

The EIS distinguishes between four main types of indicators – Framework conditions, Investments, Innovation activities and Impacts – and ten innovation dimensions, capturing in total 27 indicators (Table 1).

Table 1. Measurement framework of the European Innovation Scoreboard

| FRAMEWORK CONDITIONS |  |
|-----------------------|---|
| Human resources       |  |
| 1.1.1 New doctorate graduates per 1,000 population aged 25–34 |  |
| 1.1.2 Percentage population aged 25–34 having completed tertiary education |  |
| 1.1.3 Percentage population aged 25–64 involved in lifelong learning |  |
| Attractive research systems |  |
| 1.2.1 International scientific co-publications per million population |  |
| 1.2.2 Scientific publications among the top 10% most cited publications worldwide as a percent of total scientific publications of the country |  |
| 1.2.3 Foreign doctorate students as a percent of all doctorate students |  |
| Innovation-friendly environment |  |
| 1.3.1 Broadband penetration |  |
| 1.3.2 Opportunity-driven entrepreneurship (Motivational index) |  |
| INVESTMENTS |  |
| Finance and support |  |
| 2.1.1 R&D expenditure in the public sector (% of GDP) |  |
| 2.1.2 Venture capital (% of GDP) |  |
The methodology for calculating the Summary Innovation Index (SII) comprises the following steps of the procedure (Hollanders, 2019):

1. Identifying and replacing outliers.
2. Setting reference years.
3. Imputing for missing values.
4. Determining Maximum and Minimum stores.
5. Transforming data that have highly skewed distributions across countries.
6. Calculating re-scaled stores (after correcting for outliers and a possible transformation of the data) for all years are calculated by first subtracting the Minimum score and then dividing by the difference between the Maximum and Minimum score. The maximum
re-scaled score is thus equal to 1, and the minimum re-scaled score is equal to 0. For positive and negative outliers, the re-scaled score is equal to 1 or 0, respectively.

7. Calculating composite innovation indexes $SII_i$.

For each year, a composite Summary Innovation Index is calculated as the unweighted average of the rescaled scores for all indicators where all indicators receive the same weight (1/27 if data are available for all 27 indicators).

8. Calculating relative-to-EU performance scores on the basis of a formula:

- **Group I (Innovation Leaders):** $SII_i > 1.2 \overline{SII}$
- **Group II (Strong Innovators):** $0.9 \overline{SII} < SII_i < 1.2 \overline{SII}$
- **Group III (Moderate Innovators):** $0.5 \overline{SII} < SII_i < 0.9 \overline{SII}$
- **Group IV (Modest Innovators):** $SII_i < 0.5 \overline{SII}$

where: $\overline{SII}$ – average of the EU in 2018

3. **Quality of statistical data – theoretical basis**

There is no doubt that the statistical data used in the research should be of high quality. It influences, in the first place, the reliability of the results in such analyses. Furthermore, the precision and adequacy of employed research methods are important (Grabiński, 2003; Grabiński, Farbaniec, Woźniak-Zapór, Zając, 2016; Karr, Sanil, Banks, 2006; Kolonko, Wywiał, 2002).

In the literature, three characteristics of statistical data quality are mentioned. These are: usefulness of data for users’ needs, validity and accuracy. The most important is the accuracy of data, expressed by the approximation of statistical information to a true value, which is the one which could be obtained if the data for all units of a researched group was collected and processed without errors (Kordos, 1987; Kordos, 1988).

At present, the analysis of the influence of errors resulting from the inaccuracy of statistical data is overlooked in the related literature, and data taken from the Statistical Yearbook or appropriate data bases is treated by users as accurate. To limit data errors, statistical offices perform corrective calculations, but still such data cannot be deemed as accurate. It is bound to a way of obtaining statistic information. Random errors (connected mainly with sample choice) and non-random errors (e.g. related to data processing, or discretization errors, i.e. truncating significant digits after the decimal point) are the source of errors in statistical data.
In engineering sciences the issue of measurements and their accuracy was precisely specified in relevant documents (JCGM/WG 1, 2008; *International Vocabulary*...).

According to a definition, measurement error is an arithmetical difference between a measured value and a true value. Error value can be written down as an absolute (a) or a relevant value (b):

\[ a) \quad \Delta X = X_m - X_p \]
\[ b) \quad \delta X = \frac{X_m - X_p}{X_p} \]  

where:
- \( \Delta X \) – absolute measurement error,
- \( \delta X \) – relative measurement error,
- \( X_m \) – set value,
- \( X_p \) – real value.

The accurate values are known in case of a model study or a simulation test, where the variable’s value is set, and its change results from conversions or other mathematical calculations, or interactions of simulated external factors.

If a true value is unknown (as in the case of all measurements) and only an estimate of a true value is known, therefore, the uncertainty of determining measured values, expressed by the dependency provided below, is being analyzed (JCGM/WG 1, 2008):

\[ X_R = X_n \pm u_c \]  

which can be also written as:

\[ X_n - u_c \leq X_R \leq X_n + u_c \]  

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 a variable’s uncertainty.

Uncertainty of measurement is parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand. The parameter an 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).
In Figure 1 the essence of a variable’s uncertainty or synthetic measure is presented.

![Diagram: Setting the uncertainty of a statistical value]

Figure 1. The essence of setting the uncertainty of a statistical value
Source: own elaboration.

If the overall uncertainty of the analyzed quantity depends on some elements of uncertainty, thus the overall uncertainty value is calculated from the dependency (White, 2008):

$$u_c = \sqrt{u_1^2 + u_2^2 + \cdots + u_n^2}$$  \hfill (4)

In research where the variable’s value is set on the basis of a functional dependency of some elements, a value of combined uncertainty $u_c (y)$ is set by the propagation (transfer) of errors. Such a case occurs when a statistical quantity is calculated into a surface area, number of population, GDP, etc. The value of combined uncertainty $u_c (y)$ is calculated from the dependency (Lisiecki, Klysz, 2007):

$$u_c(y) = \sqrt{\sum_{i=1}^{N} \left[ \frac{\partial f(x_i)}{\partial x_i} \right]^2 \cdot u^2(x_i)}$$  \hfill (5)

where:

- $u (x_i)$ – standard uncertainties of measuring input quantities,
- $u_c (y)$ – combined standard uncertainty,
- $\frac{\partial f(x_i)}{\partial x_i}$ – derivative of a function.

In economic research the most occurring dependencies are the product or quotient of two quantities. For the product or quotient of two quantities, the uncertainty of a determined variable comes down to the sum of the uncertainties of these variables.
4. Results

During the investigation on the level of innovation in EU countries some questions were posed: Could the statistical data, which is the basis for the Summary Innovation Index (SII), for individual EU countries, be deemed as accurate, i.e. devoid of any errors and inaccuracies? And if the data is not ideal, how do these errors influence the final results? Might there be a case that due to the inaccuracy of data, one of the objects could be incorrectly allocated to a group of a certain level of innovation?

In order to calculate with what accuracy the value of the variable has been determined, it is necessary to estimate the uncertainty range in which it is located. The uncertainty value of each of the variables making up the synthetic measure (SII) was calculated (or estimated) on the basis of knowledge about the manner in which the given variable was measured. The most accurate statistical data are those that come from registers maintained by government institutions, hence relatively low uncertainty values have been adopted for such variables. However, when estimating the total uncertainty, the uncertainty related to the sample selection error and the uncertainty resulting from the discretization or truncation of significant digits (e.g. to two decimal places) were taken into account and calculated from the relationship (4). The uncertainty of variable units consists of two uncertainties. For example, the variable: 2.2.1 R&D expenditure in the business sector (% of GDP) consists of the uncertainty of determining the variable: R&D expenditure in the business sector (1%) and GDP (1.5%). The total uncertainty value is determined from the dependency (5). The following values of total uncertainty were adopted for individual variables: 1.1.1 (0.011%); 1.1.2 (0.011%); 1.1.3 (0.31%); 1.2.1 (0.011%); 1.2.2 (0.101%); 1.2.3 (0.002%); 1.3.1 (0.6%); 1.3.2 (3.0%); 2.1.1 (2.0%); 2.1.2 (2.5%); 2.2.1 (2.5%); 2.2.2 (2.0%); 2.2.3 (2.1%); 3.1.1 (2.0%); 3.1.2 (0.51%); 3.1.3 (0.51%); 3.2.1 (0.51%); 3.2.2 (0.11%); 3.2.3 (2.0%); 3.3.1 (1.501%); 3.3.2 (1.501%); 3.3.3 (1.501%); 4.1.1 (0.2%); 4.1.2 (0.2%); 4.2.1 (1.0%); 4.2.2 (1.0%); 4.2.3 (1.0%).

From the essence of the uncertainty of measuring variables, it appears that the actual values of innovation indicators for the analyzed objects (countries) are within the range around the nominal value of the variable (read from the statistical year). This situation makes it uncertain whether a given object has been correctly classified. It may be that a true value of a variable (or synthetic measure) differs so much from the nominal value that a given country should be assigned to a different group in terms of innovation, than was done on the basis of data from the statistical yearbook (without taking into account the uncertainty of the measurement of the variable). Using the relationship (6), it is possible to calculate the probability of a situation in
which due to a change in the variables’ value (the true value differs from the estimated one), there will be a change in the country’s assignment to another group.

\[
f(X) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(X-X_n)^2}{2\sigma^2}}
\]

where:
- \(f(X)\) – value in the function at point \(X\),
- \(\sigma\) – standard deviation,
- \(X_n\) – nominal value of the variable.

Due to the fact that the employed method to calculate the synthetic measure leads to a change of the measuring scale, the calculation of the uncertainty of the synthetic measure (Summary Innovation Index) by the analytic method would provide false results. That is why, the Monte Carlo\(^1\) method was employed and the calculations were performed in the R application. Based on the estimated total uncertainty of each of the 27 variables for each object (country), 1,000 values were drawn at random in such a way that they met the conditions of normal distribution. Using such determined data, 1,000 synthetic measures were calculated for each object, which were used to calculate the standard deviation. The value of the standard deviation was considered as the uncertainty value of the synthetic measure \(u_c\), and after multiplying by the extension factor of 1.96 it allowed to obtain uncertainty at the significance level of 0.05. The nominal value of the synthetic measure was considered to be the measure calculated in the “classical” way – based on nominal variable values.

A problematic situation was found in one case (Figure 2). Originally, Portugal with a Summary Innovation Index of 0.4708 ±0.0023 was qualified to group III (Moderate Innovators). The limit separating group II from III is 0.4937.

The nominal value of the SII (treated as an estimate of a true value) is at the point of inflection of the normal distribution curve and is 0.4708 ±0.0023. On this basis, Portugal qualifies to group III, i.e. Moderate Innovators. The limit value dividing Group III from II is 0.4724. The absolute difference that divides Portugal from the limit between group II and III is 0.0016 and is less than the uncertainty value (0.0023). In view of the above, it is likely that Portugal’s actual SII value may be greater than the limit value (0.4724), which would be the

---

\(^1\) In the case of the Monte Carlo method, random variables are used. It is most often employed when one cannot use “classical” methods based on formulas, estimators, etc. (Liu, 2008).
basis for including Portugal in Group II (Strong Innovators). The probability of such an event (calculated from formula 6) is 0.29.

Figure 2. Density function of normal distribution for Portugal for SII
Source: own elaboration.

Figure 3 shows the ranking and final classification of EU countries in terms of a Summary Innovation Index. In addition to the SII value, the designated SII uncertainty interval is marked for each country.
In 2018, the leaders in terms of innovation are: Sweden, Finland, Denmark and the Netherlands. Eight EU countries (Luxembourg, Belgium, the United Kingdom, Germany, Austria, Ireland, France, and Estonia) qualify for the ‘Strong Innovators’ group. The most numerous group containing 14 countries is the ‘Moderate Innovators’ group. Bulgaria and Romania have shown a low level of innovation, creating the ‘Modest Innovators’ group. It can also be seen that taking into account the uncertainty about the actual SII values this creates a chance for Portugal to move to a higher group with a probability of 0.29.  

Conclusions

Innovation is currently an important determinant of the socio-economic development of individual EU countries. The European Commission attaches great importance to the research of the innovativeness of countries, in order to develop the European Innovation Scoreboard annually. In 2018, the highest values of the Summary Innovation Index were obtained by: Sweden, Finland, Denmark and the Netherlands forming the Innovation Leaders group. Luxembourg, Belgium, the United Kingdom, Germany, Austria, Ireland, France and Estonia were included in the Strong Innovators group which also had a very good situation in terms of innovation. 14 EU countries created the Moderate Innovators group; whereas Bulgaria and Romania achieved a low level of innovation by creating the Modest Innovators group.

In the study of complex phenomena, which also includes the innovation of countries, it is worth paying attention to the variables’ accuracy underlying the construction of the Summary Innovation Index. The method of obtaining statistical data causes that they are burdened with uncertainty as to their true values. Taking into account the uncertainty of variables for SII does not change the ordering of objects, but only allows one to specify the confidence interval for the obtained results.

Analyzing SII values and uncertainty intervals, it was found that there was only one case of moving a country between the created innovation groups. This was influenced by the use of multiple variables and the methodology of the Eurostat index calculation.

---

2 In addition to examining the impact of variable uncertainty on general SII, its impact on partial innovation indicators was also calculated, taking into account variables determining 10 dimensions of innovation. The most cases, for which there is a likelihood of group change by EU countries, were observed for the following dimensions: Innovation-friendly environment, Finance and support, Sales impacts (the number of cases was 7 and 3 respectively).

3 In the case of studying other complex phenomena, there were many more cases of countries changing the ranking made in terms of the value of the synthetic measure. It depends on the variables included in the research and the
of data with low uncertainty values and their large numbers (27), which resulted in making an average of aggregate uncertainty. For a smaller number of variables (studies for 10 dimensions of innovation) there were already more collision cases.

It seems that it is worth being aware of the impact of variable uncertainty on the final value of the synthetic measure, which quite often influences decision making.

References

Balazs, A. (2008). International Vocabulary of Metrology-Basic and General Concepts and Associated Terms. CHEMISTRY International, 25.

Baregheh, A., Rowley, J., Sambrook, S. (2009). Towards a multidisciplinary definition of innovation. Management decision, 47 (8), 1323–1339. DOI: 10.1108/00251740910984578.

Costello, T., Prohaska, B. (2013). Innovation. IT professional, 15 (3), 62–64.

Cumming, B.S. (1998). Innovation overview and future challenges. European journal of innovation management, 1 (1), 21–29. DOI: 10.1108/14601069810368485.

Edison, H., Bin Ali, N., Torkar, R. (2013). Towards innovation measurement in the software industry. Journal of Systems and Software, 86 (5), 1390–1407. DOI: 10.1016/j.jss.2013.01.013.

Fischer, M.M. (2001). Innovation, knowledge creation and systems of innovation. The annals of regional science, 35 (2), 199–216. DOI: 10.1007/s001680000034.

Frankowski, P., Skubiak, B. (2012). Innowacyjność w teorii ekonomii i praktyce gospodarczej. Studia i Prace WNEiZ US, 30, 271–283.

Gault, F. (2018). Defining and measuring innovation in all sectors of the economy. Research Policy, 47 (3), 617–622. DOI: 10.1016/j.respol.2018.01.007.

Godin, B. (2008), Innovation: The History of a Category, Institut National de la Recherche Scientifique, Montreal, Project in the Intellectual History of Innovation, 1. Working Paper No. 1.

Grabiński, T. (2003). Analiza taksonometryczna krajów Europy w ujęciu regionów. Kraków: Wydawnictwo AE w Krakowie.

Grabiński, T., Farbaniec, M., Woźniak-Zapór, M., Zając, W. (2016). Metody i narzędzia weryfikacji rzetelności danych liczbowych. Kraków: Oficyna Wydawnicza AFM.
Gust-Bardon, N.I. (2012). Innowacja w myśli ekonomicznej od XVIII do XX wieku: analiza wybranych zagadnień. *Acta Universitatis Nicolai Copernici Ekonomia*, 43 (1), 105–120.

Hollanders, H., Es-Sadki, N., Merkelbach, I. (2019). *European innovation scoreboard 2019*. European Commission. DOI: 10.2873/877069.

Hollanders, H. (2019). European Innovation Scoreboard 2019 – Methodology Report. *International Vocabulary of Metrology- Basic and General Concepts and Associated Terms, JCGM 2000:2012*.

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)*.

Karr, A.F., Sanil, A.P., Banks, D.L. (2006). Data quality: A statistical perspective. *Statistical Methodology*, 3 (2), 137–173.

Kogabayev, T., Maziliauskas, A. (2017). The definition and classification of innovation. *HOLISTICA – Journal of Business and Public Administration*, 8 (1), 59–72. DOI: 10.1515/hjbpa-2017-0005.

Kolonko, J., Wywiał, J. (2002). Pomiar jakości danych statystycznych. *Wiadomości Statystyczne*, 10, 19–27.

Kordos, J. (1987). Dokładność danych w badaniach społecznych. *Biblioteka Wiadomości Statystycznych*, 35, 16–18.

Kordos, J. (1988). *Jakość danych statystycznych*. Warszawa: PWE.

Lisiecki, J., Kłysz, S. (2007). Szacowanie niepewności pomiaru. *Research Works of Air Force Institute of Technology*, 22 (1), 47–79. DOI 10.2478/v10041-008-0003-5.

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

Manual, O.O. (2018). Guidelines for collecting, reporting and using data on innovation.

Sardana, G.D. (2016). Innovation and growth. *South Asian Journal of Business and Management cases*, 5 (1), 13–20. DOI: 10.1177/2277977916634255.

White, G.H. (2008). Basics of estimating measurement uncertainty. *The Clinical Biochemist Reviews*, 29 (Suppl 1), S53.