Comparison of mortality due to critical illnesses in the EU countries

Lucie Kopecká[1], Pavla Jindrová[2]

Abstract

Health is a precondition for economic prosperity in each country and citizens' health is also a core EU priority. Cancer, heart disease, diabetes, respiratory, mental and other chronic diseases represent great suffering to citizens and represent a huge costs for society and the economy. Huge differences in health and health care exist between and within EU countries and regions. The aim of this article is to present the results of applying of multivariate statistical methods, such as correlation analysis, component analysis, cluster analysis and multidimensional comparative analysis, and to provide an overview of the gravity of the situation in mortality from the serious diseases by the selected indicators, their various causal relations and regional differences and similarities in EU countries. The basic source of data is the database of the World Health Organization (WHO) for Europe.

Keywords: *critical illnesses, correlation analysis, component analysis, cluster analysis, multidimensional comparative analysis*

JEL Classification: C38, I15

1 Introduction

Critical illnesses are the most serious causes of death all over the world. The risk of occurrence is not only the thing of health sector but it is also the subject of insurance companies as Jindrová demonstrates (in Jindrová, 2013). The growth of life expectancy is the most common positive indicator of health status and quality of life in individual countries but on the other hand, ageing of population is a problem of many European countries and it brings financial risks such as social, pension, health, etc. (For details see Jindrová and Slavíček, 2012; Kubanová, 2014; Linda et al., 2014). Although the EU is the most developed part of the world huge socio-economic disparities exist among the EU member countries, and they influence health in Europe as described in Staníčková (Staníčková, 2015).

The WHO was founded within OSN as an independent international health organization in 1948. The goal of this Organization is to create better health care for all people all over the world. First of all, the organization focuses on communicable and non-communicable

[1] University of Pardubice/Faculty of Economics and Administration, Institute of Mathematics and Quantitative Methods, Studentská 84, 532 10 Pardubice, Czech Republic, e-mail: Lucie.Kopecka1@seznam.cz.

[2] University of Pardubice/Faculty of Economics and Administration, Institute of Mathematics and Quantitative Methods, Studentská 84, 532 10 Pardubice, Czech Republic, e-mail: Pavla.Jindrova@upce.cz.

diseases, fighting poverty, healthy food and safe air to breathe. On the websides of the WHO there are data which provide information about health status of the citizens in Europe. Specifically, the database contains information about demographic and socio-economic indicators, mortality-based indicators, morbidity, disability and hospital discharges, life style and health care resources, (For details see WHO, 2017).

The main aim of this article is to provide an overview of the health state of citizens in 28 EU member states according to the selected indicators by using multivariate statistical methods.

2 Data and standardization of data file

All data were obtained from the database of the WHO which provides the data from 1970 to 2015. The major problem of this database is missing data because some countries publish data with a considerable delay.

First of all, it is necessary to compile a data matrix. The rows of this matrix are represented by objects (28 EU member states) and the columns are represented by variables which evaluate individual objects (European Health Information Gateway, 2016).

As mentioned above there are selected indicators (variables) which provide information about health status of citizens in 28 EU member states. The first variable X_1 is life expectancy at birth. The next variables are related to standardized death rates (SDR) which are caused by critical illnesses. Variables X_2-X_9 are presented for the age group between 0-64, per 100 000 population. Finally, variables X_{10} and X_{11} are related to standardized death rates which are caused by smoking and drinking alcohol, per 100 000 population. The multivariate statistical methods will use the following quantitative variables: life expectancy at birth (X_1), SDR - diseases of circulatory system (X_2), SDR - ischemic heart diseases (X_3), SDR - cerebrovascular diseases (X_4), SDR - malignant neoplasms (X_5), SDR - trachea/bronchus/lung cancer (X_6), SDR - diseases of respiratory system (X_7), SDR - diabetes mellitus (X_8), SDR - mental disorders (X_9), SDR - selected alcohol-related causes (X_{10}), SDR, selected smoking-related causes (X_{11}) as demonstrated in European Health Information Gateway (European Health Information Gateway, 2016).

The standardization of the dataset is an important part of the methods. The data should be presented in the same measuring units. One of the possibilities of standardization is normalization. According to Řezánková it means to introduce a new normalized variables with mean value 0 and standard deviation 1 (Řezánková et al., 2009).

3 Methods

In this section there are described procedures and possibilities of methods such as correlation analysis, Kaiser-Meyer-Olkin index, component analysis, cluster analysis and multidimensional comparative analysis.

Non-correlation variables are the next important precondition of the method such as a cluster analysis. Pearson and Spearman rank correlations measure the strength of the association between the variables. The values of these simple correlation coefficients are located between -1 and +1. Spearman correlation coefficient is used when the assumption of normality is broken and it is based on the rank of values as demonstrated in Kubanová and Řezánková (Kubanová, 2008; Řezánková et al., 2009).

Kaiser-Meyer-Olkin index (KMO) is another way of measuring association within the group of variables. This index is based on simple and partial correlation coefficients. When the KMO index is close to 1 there is strong association among variables which is a good result for application of the component analysis. Measure of Sampling Adequacy (MSA) is an analogous and simplified rate of KMO index. According to Hebák and Řezánková it is applicable to each variable separately (Hebák et al., 2007; Řezánková et al., 2009).

The component analysis solves the problem of correlation among variables. It creates lower number of non-correlation artificial variables which explain most of the original variables variability. This analysis is able to find the right dimension of a data file. The component analysis is a necessary step for a cluster analysis because it is very sensitive to the association among variables. In case of cluster analysis the normalized component scores will be used. The dispersions of normalized component scores equal 1 (Hebák et al., 2007; Řezánková et al., 2009; Stankovičová and Vojtková, 2007).

The main aim of a cluster analysis is to classify the objects into a group so that the objects are the most similar inside the group and the most different among the groups. One of the most important part of a cluster analysis is to find out distances among objects. The Euclidean distance (1) is the most common possibility for measuring distances.

$$D_E\left(x_i, x_j\right) = \sqrt{\sum_{l=1}^{m}\left(x_{il} - x_{jl}\right)^2}. \tag{1}$$

Methods of the cluster analysis are divided into hierarchical and non-hierarchical methods. The agglomerative algorithm is used in the case of hierarchical methods. In practice hierarchical methods are applied before non-hierarchical methods are used. The main reason for this procedure is to get a prior information about the number of the clusters, which is crucial for the non-hierarchical methods. The hierarchical methods are able to provide this information by using a dendrogram which displays the results of a cluster analysis and provide the information about significant clusters. The hierarchical methods include Ward's method, etc. The mentioned hierarchical method is mostly used. Among the non-hierarchical clustering methods K-means clustering algorithm is included (Hebák et al., 2007; Petr et al., 2010; Řezánková et al., 2009).

The multidimensional comparative analysis can be used for comparing objects (EU member states) which are evaluated by using several variables. First of all, the type of each variable in the data file should be defined because *"great"* values of variables influence the analysis positively (stimulants) or on the other hand, *"small"* values of variables are favourable (destimulants). This is the reason why it is necessary to make variables compatible by using standardization. To standardize the data the formulas for stimulants (2) and destimulants (3) are applied. Formulas for stimulants contain maximal measurement of *j*th variable and for destimulants minimal measurement of *j*th variable as demonstrated in (Pacáková and Papoušková, 2016; Pacáková et al., 2016; Stankovičová and Vojtková, 2007).

$$b_{ij} = \frac{x_{ij}}{x_{\max,j}} \cdot 100 \tag{2}$$

$$b_{ij} = \frac{x_{\min,j}}{x_{ij}} \cdot 100 \tag{3}$$

Finally the score for each country is calculated as the average of the b_{ij}, $i = 1,\ldots,n$.

4 Results of methods and discussion

All presented results of methods were created in MS EXCEL, STATISTICA and STATGRAPHICS programmes by using the methods that are mentioned above.

As mentioned above Spearman correlation coefficients measure the strength of the association between each pair of the variables. The strong associations between variables exist. For example, a significant negative correlation (more than 0.80) between X_1 and X_2, X_4, X_5 and X_{11} is detected. A significant positive correlation is detected between X_2 and X_3 and also X_{11}. On the other hand, a poor correlation between X_6 and X_{11} is found.

These significant correlations between variables can be eliminated by a component analysis. Value of overall MSA rate is 0.80, which points out the adequacy of the data for a component analysis. In Figure 1 there are displayed eigenvalues providing information about number of components which are suitable for the next cluster analysis. For details you can see Hebák (Hebák et al., 2007).

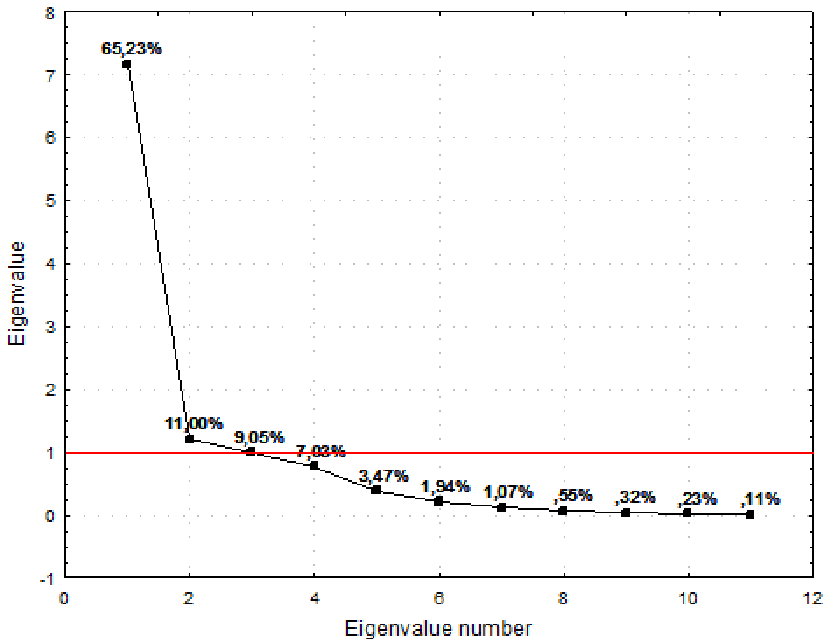

Fig. 1. Scree plot presenting eigenvalues of correlation matrix.

The top two eigenvalue numbers are higher than eigenvalue 1 so this is the reason for using the top two non-correlation components which together explain 76,23% of the original variables variability. This issue is described in Hebák (Hebák et al., 2007).

The factor loadings reveal associations between components and variables. According to these correlation coefficients there are revealed significant correlations between component 1 and variables X_1-X_8, X_{10} and X_{11} and significant correlation between component 2 and variable X_9. It means that the correlation coefficients point out the influence between variables and components. Variable X_1 contributes most of all to the determination of component 1. There is a positive correlation between X_1 and component 1 unlike the others. Component 2 is determined by X_9 and there is a negative correlation between them. For details see Stankovičová and Vojtková (Stankovičová and Vojtková, 2007).

In Figure 2 EU member states are shown according to component 1 and component 2. In this figure the states according to non-correlation components are shown. Here, the outliers can be detected as demonstrated in (Stankovičová and Vojtková, 2007).

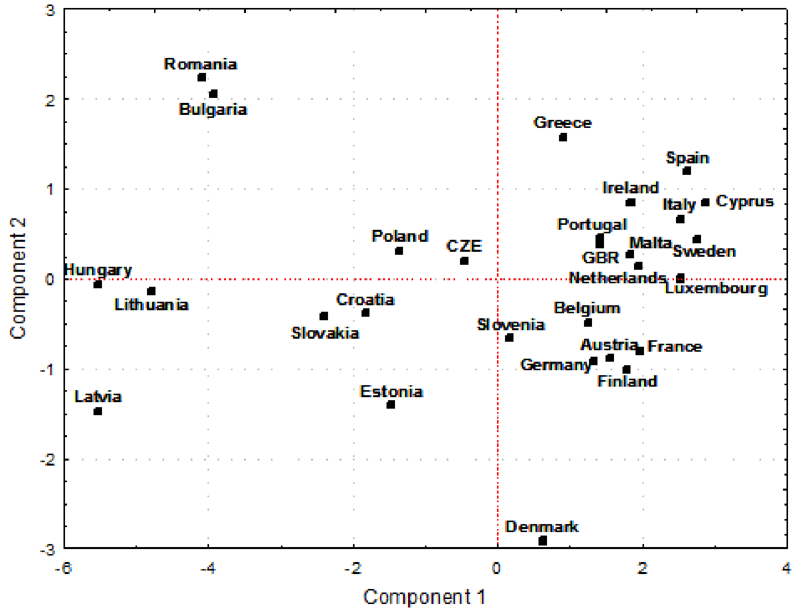

Fig. 2. Graf of the EU member states according to component 1 and component 2.

In Table 1 there are the normalized component scores for each object of the data file which are used for the cluster analysis.

States	*Com. 1*	*Com. 2*	States	*Com. 1*	*Com. 2*
Austria	0.58	-0.80	Ireland	0.69	0.77
Belgium	0.47	-0.44	Italy	0.94	0.60
Bulgaria	-1.46	1.86	Latvia	-2.06	-1.34
Croatia	-0.68	-0.35	Lithuania	-1.79	-0.12
Cyprus	1.08	0.77	Luxembourg	0.94	0.00
CZE	-0.17	0.18	Malta	0.69	0.24
Denmark	0.23	-2.65	Netherlands	0.73	0.12
Estonia	-0.55	-1,28	Poland	-0.51	0.29
Finland	0.66	-0.92	Portugal	0.53	0.41
France	0.74	-0.73	Romania	-1.53	2.03
Germany	0.49	-0.83	Slovakia	-0.89	-0.39
GBR	0.53	0.35	Slovenia	0.06	-0.60
Greece	0.34	1.43	Spain	0.98	1.09
Hungary	-2.07	-0.06	Sweden	1.03	0.39

Table 1. Normalized component scores.

Figure 3 displays the dendrogram by using hierarchical Ward's method and Euclidean distance.

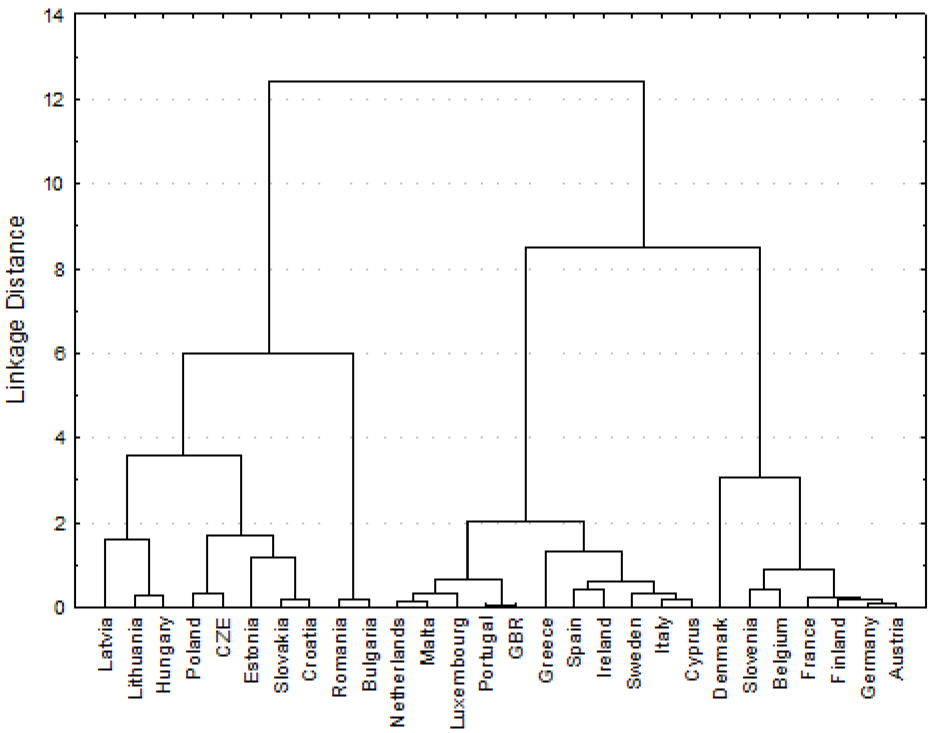

Fig. 3. Representation of similarities within EU countries by using dendrogram.

By using a horizontal cut in this dendrogram it is possible to determine a number of clusters. When the linkage distance equals 4 then the number of significant clusters is 4. In case of non-hierarchical K-means algorithm there are 4 identified clusters as the priori information. This algorithm provides slightly different classification than Ward's method. The first cluster generated by using the K-means algorithm contains Croatia, Hungary, Latvia, Lithuania, Poland, Slovakia, the next one contains Cyprus, the Czech Republic, Great Britain, Greece, Ireland, Italy, Luxembourg, Malta, Netherlands, Portugal, Spain, Sweden. The states which create the third of them include Austria, Belgium, Denmark, Estonia, Finland, France, Germany, Slovenia. The last cluster contains Bulgaria and Romania.

The results of multidimensional comparative analysis are displayed in the Table 2. These countries are arranged by descending order from the best health status of the population into the worst. Score 1 indicates life expectancy at birth and Score 2 indicates standardized death rates of all critical illnesses which are mentioned above. This issue is described in (Hebák et al., 2007; Petr et al., 2010; Řezánková et al., 2009).

States	Score 1	States	Score 1	States	Score 2	States	Score 2
Spain	83.30	Belgium	80.80	Cyprus	77.48	Greece	56.50
Luxembourg	83.00	Greece	80.80	Spain	73.15	Germany	53.38
France	82.60	Denmark	80.40	Italy	73.07	Denmark	52.58
Cyprus	82.50	Slovenia	80.00	Sweden	73.04	Slovenia	49.78
Italy	82.40	CZE	79.00	Luxembourg	71.24	CZE	43.65
Sweden	82.40	Croatia	78.00	France	67.53	Poland	39.37
Malta	82.20	Poland	77.90	Netherlands	66.00	Estonia	38.25
Austria	81.80	Estonia	77.50	Malta	64.48	Croatia	37.83
Netherlands	81.60	Slovakia	77.10	Ireland	63.98	Romania	37.22
Finland	81.30	Hungary	76.00	Finland	62.54	Bulgaria	36.59
Germany	81.30	Romania	75.70	Portugal	60.47	Slovakia	34.96
GBR	81.20	Bulgaria	75.00	Austria	58.06	Lithuania	32.14
Ireland	81.10	Latvia	74.80	Belgium	57.85	Latvia	27.71
Portugal	81.00	Lithuania	74.80	GBR	57.33	Hungary	26.34

Table 2. Ranking of the EU countries according to Score 1 and Score 2.

The countries with the highest life expectancy at birth include Spain, Luxembourg and France, and the countries with the lowest life expectancy at birth include Bulgaria, Latvia and Lithuania. On the other hand, the lowest mortality cause by critical illnesses is in Cyprus, Spain and Italy and the highest is in Lithuania, Latvia and Hungary. There is a strong association between Score 1 and Score 2. The Spearman rank correlation coefficient acquires value 0.96.

Conclusions

The WHO's database provides data files carrying information about health status of citizens in the 28 EU member states. This information is obtained by using multivariate statistical methods. First, there are detected strong associations between variables and among the group of variables by Spearman correlation and KMO index. This significant correlations are eliminated by two non-correlation components which explain 76.23 % of the variables variability where component 1 expresses general component of health status and component 2 reflects SDR caused by mental disorders. In the Figure 2 there are displayed EU member states according to this components. According to general component 1 on the one hand, the countries such as Cyprus, Sweden and Spain belong to the best within health status and on the other hand, the countries such as Latvia and Hungary to the worst. Next according to component 2 which expresses SDR caused by mental disorders, the best mental health is in Romania and Bulgaria and the worst in Denmark. Based on new normalized components Ward's method and K-means algorithm, which both provide different results of classification, are applied. The reason for different classification in case of K-means algorithm are the outliers (Greece and Denmark) which influence location of centroids at the start of this algorithm. Finally, by using multidimensional comparative analysis the countries are arranged according to life expectancy at birth and standardized death rates, which shows a strong association.

Acknowledgements

This paper was supported by SGS of the University of Pardubice, FEA, No. SGS_2017_022.

References

European Health Information Gateway (2016). *WHO.* Available: http://gateway.euro.who.int/en/data-sources/european-health-for-all-database/.

Hebák, P., Hustopecký, J., Pecáková, I., Plašil, M., Průša, M., Řezánková, H., Vlach, P., & Svobodová, A. (2007). *Vícerozměrné statistické metody (3)*. Praha: Informatorium.

Jindrová, P. (2013). Quantification of Risk in Critical Illness Insurance. In *Conference proceedings from 9th international scientific conference Financial Management of Firms and Financial Institutions, VŠB Ostrava*, (pp. 298-306).

Jindrová, P., & Slavíček, O. (2012). Life expectancy development and prediction for selected European countries. In *6-th International Scientific Conference Managing and Modelling of Financial Risk Proceedings* (pp. 303-312).

Kubanová, J. (2008). *Statistické metody pro ekonomickou a technickou praxi.* Bratislava: Statis.

Kubanova, J. (2014). The process of population aging and its impact on country's economy. In *SGEM2014 conference on Political sciences, Law, Finance, Economics and Tourism* (Vol. 4, No. SGEM2014 Conference Proceedings, ISBN 978-619-7105-28-5/ISSN 2367-5659, September 1-9, 2014, Vol. 4, 767-774 pp, pp. 767-774). STEF92 Technology.

Linda, B., Kubanova, J., & Pacakova, V. (2014). Mortality Rate Estimate for the Old Age Population by Bootstrap Method. In *SGEM2014 conference on Political science, Law, Finance, Economics and Tourism* (Vol. 4, No. SGEM2014 Conference Proceedings, ISBN 978-619-7105-28-5/ISSN 2367-5659, September 1-9, 2014, Vol. 4, 481-486 pp, pp. 481-486). STEF92 Technology.

Pacáková, V., Jindrová, P., & Zapletal, D. (2016). Comparison of Health Care Results in Public Health Systems of European Countries. *European Financial Systems 2016*, (pp. 534-541).

Pacáková, V., Papoušková, M. (2016). Multidimensional Comparisons of Health Systems Functioning in OECD Countries. *International Journal of Mathematical Models and Methods in Applied Sciences*, 10, (pp. 388-394).

Petr, P., Křupka, J., & Provazníková, R. (2010, October). Statistical approach to analysis of the regions. In *Conference proceeding of the 10th WSEAS International Conference on Applied Computer Science* (pp. 280-285).

Řezánková, H., Húsek, D., & Snášel, V. (2009). *Shluková analýza dat*. Praha: Professional Publishing.

Stanickova, M. (2015). Classifying the EU Competitiveness Factors Using Multivariate Statistical Methods. *Procedia Economics and Finance*, *23*, (pp. 313-320).

Stankovičová, I., & Vojtková. M. (2007). *Viacrozmerné štatistické metódy s aplikáciami*. Bratislava: IURA Edition.

WHO. *Who we are, what we do*. (2017). Available: http://www.who.int/about/en/.