The Visualization and Grouping of European Union Member States According to Structural Indicators: A Data-Driven Approach*

Jasna Soldic-Aleksic
Belgrade University, Belgrade, Serbia

The main idea presented in this paper concerns the visual exploration of the convergence of European Union member states in the light of structural indicators (SI) defined by the European Commission. The structural indicators list 79 indicators which cover the following six domains: general economic background, employment, innovation and research, economic reform, social cohesion and environment. The relevant data are provided by EU statistical office (EUROSTAT). For the purpose of visualization and grouping of 27 EU member states, the technique of Kohonen self-organizing map (SOM) is applied. This technique belongs to the group of unsupervised neural networks models. Prior to the SOM application an exploratory data analysis for all SI domains is performed. It reveals the presence of a higher variability among EU states in the area of economic reform, general economic background and I&R indicators comparing to the environment, employment and social cohesion indicators. The SOM application is carried out in two stages. In the first stage, the Kohonen SOM algorithm is applied on each group of SI dataset separately, and in the next stage, the general map of European Union states based on all indicators is presented. The general map suggests the “natural” grouping of the European Union countries in three clusters, but, the other cluster solutions are also presented to illustrate the clustering tendency of European Union states and their regional cohesion. It is remarkable that geographic location of many countries is reflected in the final map organization, with some exceptions, notably Ireland and Luxembourg. But this “geography factor” doesn’t apply in the same extent to all SI domains. “Geography” matters the most regarding the general economic background and I&R, but the least regarding the employment area. Finally, the four-cluster profile analysis is generated. It reveals the divisions among EU states in detail.

Keywords: Kohonen self-organizing map, EU states, visualization, clustering, European commission structural indicators, group profile analysis

Introduction

At the Lisbon meeting of the European Council in 2000, the so-called Lisbon strategy was formulated,
aimed at achieving the European Union strategic goal for the next decade “… of becoming the most competitive and dynamic knowledge-based economy in the world capable of sustainable economic growth with more and better jobs and greater social cohesion”\(^1\). As a part of this strategy, the initiative was put forth for the introduction of the specific set of social and economic indicators for European Union (EU) member states. The relevant indicators were defined as structural indicators (SI) and have been used as an objective measure of the progress that EU has made towards the Lisbon objective. Currently, the list of structural indicators comprises 79 indicators which cover the following six domains: general economic background, employment, innovation and research, economic reform, social cohesion and environment.

Considering this on-going historical process and facing the end of the decade when the Lisbon objective was set, the main idea of the paper is to provide a comprehensive image of the current status and the relative positions of EU member states according to the structural indicators. For this purpose, Kohonen self-organizing map (SOM) algorithm is applied. This algorithm belongs to the group of data mining tools (Bozdogan, 2004; Keating, 2008) which are appropriate for visualization and clustering purposes.

The paper is organized as follows. After an introduction, the second part offers a literature review. The third part provides a description of data used for SOM application. The next section deals with exploratory data analysis where multiple box plots and coefficients of variation for particular datasets are presented. The fifth section describes the main characteristics of the applied algorithm. This section is followed by the sixth crucial part of the paper which discusses the empirical results of the visualization and grouping of EU states by SOM algorithm. Finally, the concluding remarks are presented.

### Literature Review

Generally the problem of studying the convergence and grouping structure in the EU area is a topic that attracts great attention of economists and other researchers. One strand of studies relates to the domain of economic growth theory and encompasses a large amount of literature. A great number of authors, to mention only a few, Chatterji (1992), Quah (1996), Corrado, Martin, and Weeks (2005), Mora (2005, 2008), Sassi (2006), Dall’erba and Le Galo (2008) focused on the idea concerning a phenomenon known as the “club convergence hypothesis” and the existence of “convergence clubs” among European countries and regions. In recent years, this idea was empirically investigated by many researchers and a great deal of research effort was devoted to developing appropriate econometric tools for testing the hypothesis and exploring the phenomenon.

The other approach, which brought not only a remarkable contribution to the idea of convergence and cluster structures in EU area, but also provided a contribution to the clustering methodology, is proposed by Gligor and Ausloos (2006, 2007, 2008a, 2008b, 2008c). These authors considered the European countries as interacting agents of a complex system. They recalled the properties of the minimal spanning tree (MST) method, which had been used by Hill (2001) as a methodology for linking countries in a way that international price and quantity indexes can be chained. To research the clustering structures in the EU area, Gligor and Ausloos (2008b) proposed the application of the MAMLP (moving-average-minimal-length-path) method. The principal findings of these authors concerned the empirical verification of the decreasing of the mean statistical distance between EU countries.

On the other side, from the perspective of the idea of visualization, Kohonen SOM algorithm appeared as an appropriate method in many studies. The author indicates here a few studies that employed Kohonen algorithm, as a visualization and cluster-extracting method. One of the early socio-economic applications of the Kohonen map was made by Varfis and Versino (1992), who compared the results of the SOM algorithm to some well-known statistical techniques (principal component analysis and hierarchical clustering). They mainly focused on the techniques’ properties making comparisons between them and noted that the Kohonen algorithm is a worthy alternative to well-established statistical techniques. The authors provided the Kohonen map of European Statistical Territorial Units and discovered that units on the map cluster into distinct European geographic areas. Also, interesting results are presented by Kaski and Kohonen (1996) who used the Kohonen algorithm in creating the “welfare map” of the countries of the world. Among the remarkable findings of this article, concerning the identification of welfare and poverty structures among the countries of the world, one was of particular importance. That was the fact that, although no geographical factors were applied in computing the map, the organization of the countries on the map reflects their geographical positions.

Similar conclusions regarding the influence of “geography” on economic factors can be found in the exploratory study provided by Haughton, Deichmann, Eshghi, Sayek, Teebagy, and Topi (2003) where 25 Central and Eastern European and Central Asian states (“transition economies”) were investigated in order to group according their abilities to attract foreign direct investment (FDI). The authors employed Kohonen map to determine the extent to which the distribution of foreign direct investment in different regions reflects the geographic organization of the countries. The resulting map based on a set of 21 macroeconomic indicators, showed that the distribution of the FDI determinants closely approximates the geographic grouping of the countries in the region.

The other study realized by Mattoscio, Colantonio, Carlei, and Furia (2009) analyzed clusters and distances among the EU’s member states in terms of health standards and economic development. The Kohonen map in this analysis revealed some well-defined groups of countries. Also, the authors derived a conclusion that differences among countries in the identified groups are based on just a few variables of the whole set used in the analysis, which can be summarized in terms of current expenses and investments.

Interesting findings regarding the application of the modified Kohonen algorithm is provided by Aaron, Perraudin, and Rynkiewicz (2003) who studied the convergence of EU countries within the Maastricht criteria framework for the period 1980-2002. They used cross-country time-series data for four economic variables: deficit in GNP, debt in GNP, inflation rate, and long-term nominal interest rate. The approach which they proposed related to the quantification of the trajectory of each country, where each trajectory was considered as a multivariate function of time. They applied the SOM algorithm to obtain a classification of the functions. Further on, they showed that this approach allowed the tracking of the individual evolution of each country with respect to the chosen criteria. Their results “confirmed the convergence of the European countries to Maastricht criteria, since the number of classes has diminished since 1980” (Aaron, Perraudin, & Rynkiewicz, 2003, p. 330).

In the context of the previous broad literature, the objective of this paper is to explore the relative positions and grouping of 27 EU countries in the light of the Lisbon strategy and to provide empirical comparisons to some earlier studies.
Description of the Data

The data considered in this study is provided by EU statistical office (EUROSTAT) and may be found at EUROSTAT web portal\(^2\). The complete SI list is divided in six domains: general economic background (GEB), innovation and research (I&R), economic reform, employment, environment and social cohesion. These indicators were defined as a part of broader set of EU policy indicators. In addition to the structural indicators, EU policy indicators include principal European economic indicators (Euro Indicators PEEIs), sustainable development indicators and employment and social policy indicators. In this paper, only the data on structural indicators for EU member states are analyzed. Therefore, the results of this study should be considered only in that respect.

In the following the list of structural indicators that have been treated in the application is presented.

Economic reform indicators are supposed to reflect the EU member states’ capacity for coherent development of economic environment with regards to: effective competition and trade, price convergence, government’s initiatives for supporting entrepreneurship in the EU and providing the healthy business environment for job creation and poverty reduction, market integration by different types of trade activity (eliminating barriers to the free movement of goods, services and people within the EU). This set of structural indicators includes the following\(^3\):

- Business demography;
- Business investment;
- Comparative price levels;
- Electricity prices by type of user;
- Gas prices by type of user;
- Market integration by type of trade activities;
- Market integration-foreign direct investment (FDI) intensity;
- Market share of the incumbent in fixed telecommunications by types of call;
- Market share of the leading operator in mobile telecommunication;
- Market share of the largest generator in the electricity market;
- Price of telecommunications by type of call;
- Public procurement;
- State aid by type of aid.

In the area of employment the general objective of the EU employment policy has been defined in the form of concrete figures—as the achievement of an average employment rate of 70% for the EU overall and at least 60% for women by 2010. The indicators in this domain address one of the vital issues of prosperous economic and social development of any country, i.e., the issue of adequate and balanced utilization of the human capital as an essential economic resource. The list of 11 indicators is defined as follows\(^4\):

- Employment rate by gender;
- Employment rate of older workers by gender;
- Average exit age from the labor force by gender;
- Gender pay gap in unadjusted form;


THE VISUALIZATION AND GROUPING OF EUROPEAN UNION MEMBER STATES

- Tax wedge on labor cost;
- Fatal accidents at work;
- Implicit tax rate on labor;
- Life-long learning by gender;
- Tax rate on low wage earners by marginal effective tax rates on employment incomes;
- Unemployment rate by gender;
- Serious accidents at work by gender.

In the context of EU economic and social development, special attention has been dedicated to the environmental issues dealing with greenhouse gas emission, renewable energy sources, efficient usage of the disposable resources, waste reduction and recycling, etc. To monitor the progress in this sphere the following indicator list has been proposed:

- Car share of inland passenger transport;
- Combined heat and power generation;
- Electricity generated from renewable sources;
- Energy intensity of the economy;
- Farmland bird index;
- Greenhouse gas emissions;
- Healthy life years at birth by gender;
- Municipal waste generated
- Municipal waste by type of treatment;
- Resource productivity;
- Road share of inland freight transport;
- Sufficiency of sites designated under the EU habitats directive;
- Implicit tax rate on energy;
- Urban population exposure to air pollution by ozone;
- Urban population exposure to air pollution by particulate matter;
- Volume of freight transport relative to GDP;
- Volume of passenger transport relative to GDP.

Structural indicators on general economic conditions provide the basis for structural reform and monitor a country’s economic prosperity. They include the following indicators:

- GDP per capita in purchasing power standards (PPS);
- Employment growth by gender;
- General government debt;
- Inflation rate;
- Labor productivity per hour worked;
- Labor productivity per person employed;
- Public balance;
- Real GDP growth rate;
- Real unit labor cost growth.

---

The area of innovation and research is considered to be the key driver for developing European information or knowledge-based society. The following list of 16 indicators is defined to follow the progress in this domain:

- Broadband penetration rate;
- E-commerce via Internet;
- E-government on-line availability;
- E-government usage by enterprises;
- E-government usage by individuals by gender;
- Gross domestic expenditure on R&D (GERD) by source of funds;
- Gross domestic expenditure on R&D (GERD);
- High-tech exports;
- ICT expenditure by type of product;
- Level of internet access—households;
- Patent applications to the European Patent Office (EPO);
- Patents granted by the United States Patent and Trademark Office (USPTO);
- Science and technology graduates by gender;
- Spending on human resources;
- Venture capital investments by type of investment stage;
- Youth education attainment level by gender.

In the field of social cohesion, the main EU objective is “to significantly reduce the number of persons at risk of poverty and social exclusion by 2010”. The achievement of this objective should be evaluated by monitoring a list of 10 indicators:

- Early school leavers by gender;
- Formal child care by duration and age group;
- Inequality of income distribution;
- Jobless households—children;
- Jobless households by gender;
- Long-term unemployment rate by gender;
- At-risk-of-poverty rate before social transfers by gender;
- At-risk-of-poverty rate after social transfers by gender;
- Dispersion of regional employment rates by gender;
- Persistent-at-risk-of-poverty rate by gender.

The complete SI list includes 79 indicators. However, since the data on three indicators are not available for single states, the application is based on previously cited 76 indicators. The analyzed datasets are yearly data.

One of the problems that frequently appeared in practical data mining analysis, which also appeared in this study, is the problem of missing data (J. W. Grzymala-Busse & W. J. Grzymala-Busse, 2005). This problem

---

9 The following three indicators are not included in the analysis: price convergence between EU member states, convergence of interest rates by type of loan and fish catches from stocks outside of “safe biological limits”.
THE VISUALIZATION AND GROUPING OF EUROPEAN UNION MEMBER STATES

brings severe concerns also when traditional methods are used (Tabachnick & Fidell, 2007). However, due to the fact that the SOM is a very robust algorithm, this problem was overcome in a relatively simple way: as the crucial part of the SOM algorithm is the choice of the best-matching node on the map for each input vector, in the case of missing value(s) for some input vector attribute(s), the process of looking-up the best-matching node on the map is limited to the available values, i.e., the look-up is conducted in a reduced map. After the training process is completed and the map is generated, the missing value(s) for particular attribute(s) can be replaced by the corresponding SOM nodes’ value(s)\(^{10}\).

**Exploratory Data Analysis**

Exploratory data analysis usually appears as an initial phase of many data analysis assignments (Giudici & Figini, 2009). Very often it sets the stage for further analysis and is fundamental for understanding what might be discovered through the application of some sophisticated data analysis technique, such as data mining techniques (Hand, Mannila, & Smyth, 2001). In this study, the basic objective of the initial (explorative) data analysis is to describe the main features of the six SI datasets regarding, in particular, their variability. For this purpose the multiple box plots are used. They are particularly useful for making efficient comparisons between distributions of different univariate datasets (Myatt, 2007; Myatt & Johnson, 2009). Having in mind that particular structural indicators are measured in different scales and that great discrepancy of the original SI data values is dominant, data are standardized prior to the graphical presentation. Figure 1 shows the relative image of SI datasets in each of the six groups.

The box plots presented in Figure 1 reveal the general central tendency and variability of the particular structural indicators datasets, as well as the shape of the distributions. It is evident that some distributions are skewed positively, others negatively, and for some indicators the outliers and/or extreme values are recorded.

In the area of economic reform structural indicators, three of them: namely indicator number 7 (market integration-foreign direct investment intensity), number 9 (market share of the leading operator in mobile telecommunication) and number 12 (public procurement) show strong positive skewness with outliers and extreme values to the right of the distributions.

In the group of employment structural indicators several indicators, such as: number 5 (tax wedge on labor cost), number 8 (life-long learning by gender), number 9 (tax rate on low wage earners by marginal effective tax rates on employment incomes), number 10 (unemployment rate by gender) and number 11 (serious accidents at work by gender) also show moderate or substantial departure from normal distribution.

The environment structural indicators show the most versatile data distribution shapes with a lot of outliers and/or extreme values. In particular, this is the case with the following indicators: number 1 (car share of inland passenger transport), number 2 (combined heat and power generation), number 3 (electricity generated from renewable sources), number 4 (energy intensity of the economy), number 12 (sufficiency of sites designated under the EU habitats directive), number 15 (urban population exposure to air pollution by particulate matter) and number 17 (volume of passenger transport relative to GDP).

In the domain of GEB structural indicators, the presence of outliers can be observed in relation to indicator

---

\(^{10}\) Viscovery SOMine 5.1., User manual and guide (2009). More on the subject of dealing with missing data with neural networks methodology may be found in the papers of Piela (2002, 2003) and related articles based on the research conducted under the EUREDIT project (http://www.cs.york.ac.uk/euredit). The most important method used in this research project is tree-structured self-organizing map, or TS-SOM.
number 1 (GDP per capita in PPS), number 4 (inflation rate) and number 6 (labor productivity per hour worked), as well as the skewness of distributions.

Figure 1. Box plots of six SI datasets (standardized values).

Among the 16 structural indicators on innovation and research a few of them: number 5 (e-government usage by individuals by gender), number 11 (patent applications to the European Patent Office), number 12 (patents granted by the United States Patent and Trademark Office), number 14 (spending on human resources) and number 16 (youth education attainment level by gender)—have outliers and substantial departures from normal distribution.
Regarding the group on social cohesion indicators, the most remarkable is the behavior of the first indicator (early school leavers by gender) with a substantial positive skewness and outliers. Also, indicator number 3 (inequality of income distribution), number 4 (jobless households—children) and number 5 (jobless households by gender) show positive skewness.

The previous results are very indicative and helpful in making decisions on the potential data transformations which are usually considered in the early stage of the Kohonen map processing.

In addition to the box plots, the graphical presentation of coefficients of variation for each SI dataset is examined and the most interesting findings are presented in the Figure 2.

![Figure 2: Coefficients of variation for each SI dataset.](image)

Coefficients of variations for structural indicators concerning the employment, environment and social cohesion are generally in the range of 0.0-1.0, while the coefficients of variation for structural indicators on innovation and research, general economic background and economic reform are spread in wider ranges: 0.131-1.159, 0.324-2.036 and 0.169-3.445 respectively.

It is interesting to note that in the group of economic reform structural indicators, most coefficients of variation take values in the 0.168-0.602 interval. However, there is a peak on the variability of indicator number seven, i.e., on market integration—foreign direct investment (FDI) intensity (average value of inward and outward FDI flows divided by GDP—in percent). The coefficient of variation for this indicator has an
extreme value of 3.445 due to the high value of this indicator for Luxembourg (234%), while the other indicators from this group have moderate variability.

The values of coefficient of variation for employment indicators are spread from 0.025 to 0.804, having the peak on indicator number 8 (life-long learning by gender)—0.804. When this value is excluded the upper limit in this SI group is 0.411.

In the group of environment structural indicators the minimal value of coefficient of variation is found for indicator number 7—healthy life years at birth by gender (0.067) and maximal value for the third indicator—electricity generated from renewable sources (0.986). Also, there is a high variability for the second and the forth indicator: combined heat and power generation and energy intensity of the economy with coefficients of variation equal to 0.830 and 0.731 respectively.

Concerning the general economic background structural indicators, a few of them present extremely high variability (indicator numbers 2, 7, 8, and 9). These are indicators on annual percentage change in total employed population, public balance, real GDP growth rate and real unit labor cost growth. The coefficients of variation for these indicators are: 1.196, 1.635, 2.036, and 1.325 respectively.

In the I&R group, two indicators (numbers 11 and 12: patent applications to the European Patent Office (EPO) and patents granted by the United States Patent and Trademark Office—USPTO) have the coefficient of variation above 1.0, i.e., 1.101 and 1.159. The coefficients of variation for the other I&R indicators take the values from the 0.131-0.886 interval.

Finally, the group of Social cohesion indicators demonstrates the lowest variability across the EU states. The respective coefficient of variation values fall within the 0.146-0.592 interval.

General Characteristics of the Kohonen Self-organizing Map Algorithm

The Kohonen SOM algorithm is a well-known learning one that belongs to the group of unsupervised neural networks models. Historically the SOM method was originally proposed as an explanation for biological phenomena. The fundamental idea of SOM map was introduced by Marlsburg (1973) and Grossberg (1976), but later on Kohonen (1981) proposed the model which was successfully applied to a number of pattern recognition and engineering applications (Cherkassky & Mulier, 1998). The main feature of the SOM algorithm, which makes it helpful for this study, is that the relations between data items from the original data space become explicit on the SOM map due to a non-linear projection from a high-dimensional space onto a two-dimensional display11. Different aspects of the Kohonen algorithm were extensively analyzed by many authors and, also, numerous examples of SOM applications may be found in literature12.

To avoid turning this paper into a technical-oriented one, here only the essentials of the SOM technique are presented (Deboeck & Kohonen, 1998, pp. 163-164). The SOM map consists of units (nodes or neurons) which reflect the general form of the input data space. After net training, each unit represents a group of individuals with similar features, i.e., individuals with similar features correspond to the same unit or to neighboring units on the map. Assuming that the observation space is \( n \)-dimensional, the real sample vectors \( x(t), t = 1, 2, \ldots \) are:

\[
x(t) = [x_1(t), x_2(t), \ldots, x_n(t)]
\]

11 Two-dimensional maps are mostly used in practice. However, one-dimensional and multi-dimensional maps are also possible.
12 The extensive list of the bibliography of SOM algorithm applications can be found at the following web address: http://www.cis.hut.fi/research/refs/.
where $t = 1, 2, \ldots$, refers to the index of the sample.

Similar to the sample vectors, there are $n$-dimensional model vectors:

$$m_i(t) = [m_{i1}(t), m_{i2}(t), \ldots, m_{in}(t)] \tag{2}$$

which are associated with each node on the map. The values of the model vectors are adjusted in response to samples $x(t)$, $t = 1, 2, \ldots$. At each step of the algorithm, primarily the index of the “winner node” is identified in the following manner:

$$\|x(t) - m_c(t)\| = \min_i \|x(t) - m_i(t)\| \tag{3}$$

where the index $c$ refers to the “winner neuron”. This is followed by the adjustment of the model vectors $m_i(t)$:

$$m_i(t+1) = m_i(t) + \alpha(t) h_c(t) [x(t) - m_i(t)] \tag{4}$$

For the purpose of convergence it is necessary that $h_c(t) \to 0$ when $t \to \infty$. This is a unique feature of SOM algorithm that neighborhood width gradually decreases as iterations progress (Cherkassky & Muller, 1998).


**Major Results: The Visualization and Grouping of EU Member States**

In this section major results of SOM application are presented. The section consists of two parts. In the first part the particular solutions for each SI dataset are presented. These particular mapping solutions are followed by the general SOM map of EU states based on all structural indicators.

**Kohonen Maps of EU States—Particular Solutions for Each SI Dataset**

As explained in the previous section, during the SOM training process the nodes of the map gradually adapt to the intrinsic shape of data distribution. But, prior to the training of the SOM algorithm each dataset forwarded into SOM algorithm passes through a pre-processing data procedure (Deboeck & Kohonen, 1998). As SI datasets show substantial differences regarding the appropriate measurement level and variability, indicators are scaled by the appropriate standard deviation (Berry & Lindoff, 1997). The other issue relevant to the data preprocessing phase is the transformation of the particular SI datasets. Data transformations are based on the exploratory data analysis results (presented in the section 4). For substantial positively skewed distributions, with outliers and/or extreme values, the square root or log transformation are considered and applied, while for substantial negatively skewed data with outliers, the relevant transformations for positive skewness are conducted on the reflected original datasets (Tabachnick & Fidell, 2007). Finally, regarding the priority of the particular indicators, all indicators are treated equally, i.e., the same priority factor equivalent to number 1 is assigned to each of them.

The respective maps in Figure 3 show visual images of EU states and the appropriate intrinsic grouping in clusters. Due to the general SOM property of topology preserving, the closer the positions of two countries in a map, the similar their profiles regarding the respective structural indicators.

Along with the projection of the EU states in the map, Figure 3 provides the appropriate clustering solutions. Distinctive clusters are painted in different colors which are automatically generated by the concrete

---

13 The Kohonen maps were created using the software package *Viscovery SOMine* (Version 5.1.) from Eudaptics, Software GmbH, Inc., Austria.

14 If a particular indicator has a priority factor less than 1, it means that such an indicator is less relevant for the training process and, vice versa, if a particular indicator has the priority factor greater than 1, the indicator is more dominant.
program solution\textsuperscript{15}. The specific regions are determined by the number of the clusters, as well as, by the chosen granularity of the map. While the number of the SOM nodes, as an input program value, may be arbitrary chosen and entered by the user at the beginning of the SOM processing, the number of map clusters is not predetermined. Namely, the choice of the concrete cluster solution is based on some objective criteria used in the clustering procedure. Here it is the cluster indicator which is explained in the next part and presented in the Figure 4. The previous presented clustering solutions should be considered as initial options for particular SI structure, and are usually expanded with more in-depth analysis of the concrete map.

\textsuperscript{15} Viscovery SOMine (Version 5.1.), \textit{User manual and guide} (2009), Viscovery Software GmBH. Vienna.
With regards to the clustering procedure, the SOM-Ward-clustering method is applied. This method combines the well-known Ward’s algorithm (standard hierarchical cluster procedure) and the local order information of the map. This combination assumes that the algorithm starts as a Ward’s method (Hair, Black, Babin, Anderson, & Tathan, 2006), where each node forms a cluster, and continues with combining two clusters at each step of the algorithm. The main principle of the Ward’s method is focused on minimizing the total within-cluster sum of squares which is used as the measure of within-clusters homogeneity. Thus, at each algorithm step, clusters are combined in such a way that the resulting cluster solution has the minimal within-cluster sums of square across all variables in all clusters. However, in addition to this classical Ward’s procedure, SOM-Ward-clustering method takes into account the positions of the clusters in the map, allowing only neighboring clusters to be combined.

Having in mind the explorative nature of SOM algorithm, the number of clusters is not predetermined. Actually, it was experimenting with different cluster numbers by fine-tuning the map parameters, to provide, as much as possible, the optimal partition of data items. Also, it is used a specific operative measure for determining the cluster number-cluster indicator which is calculated at each step of the clustering procedure on the base of the difference between two neighboring SOM-Ward sums of square\(^\text{16}\). This indicator is used as a heuristic measure for a possibly good clustering: Its high value for a particular cluster count is a marker of good “natural” clustering, and vice versa, its low value suggests a predominantly “artificial” clustering. Therefore, by following the values of this indicator on the cluster indicator graph, some interesting clustering solutions may be observed. The cluster indicator histograms are given in Figure 4.

The cluster indicator histograms reveal different clustering tendencies in the domains of six SI datasets. Thus, it may be observed that:

- The clustering tendencies of EU states regarding the economic reform and General Economic Background (GEB) are more or less alike. Considering the structural indicators on economic reform, the highest quality of EU states clustering is provided for three clusters and then for 23 and 26 clusters. Regarding the General Economic Background (GEB) the interesting cluster-number may be: four, eight, 22 and 26 clusters. Having in mind that the EU consists of 27 states, solutions with more than 20 clusters have no practical relevance. However, this also reveals that EU states are highly differentiated in the domain of economic reform and

\textsuperscript{16} Ibid., pp. 116-117.
general economic background;

- Considering employment indicators, the optimal cluster number may be 13 and eight. Again, the option with 26 clusters is not considered as a relevant one;
- In the domain of environment and social cohesion, cluster indicator histograms have similar shape: High cluster indicator values are recorded for three and four clusters and very low values for the other cluster solutions;
- Considering the I&R structural indicators, a possible clustering solution for EU member states may be the partition into three or four clusters, then eight or 10 clusters.

For the purpose of understanding the clustering tendencies of EU states, it is experimented with different cluster number solutions in each SI field. Only the case of GEB structural indicators is discussed here. According to the cluster indicator histogram (see Figure 4. (d)) the optimal cluster number is four, but if there is a need for more detailed analysis, the second best solution is eight clusters (partitions into 22 and 26 clusters were excluded for practical reasons).

In addition to the visualization and grouping procedure, the final evaluation of the Kohonen map involves the inspection of the appropriate component planes\(^{17}\) which usually accompany the map. These planes provide information on the relative distribution of a particular input variable (indicator) and its influences on the final map. Furthermore, they provide information on dependencies among the indicators themselves. Here, only one set of component planes is presented, i.e., component planes for GEB indicators, which illustrate the contribution to the respective map (see Figure 6).

The component planes positioned at the right-hand side of the picture in Figure 6 show the contribution of each GEB structural indicator to the respective Kohonen map (left-hand side of the picture). The scale at the bottom of each plane goes from the lowest values to the highest values for each indicator\(^{18}\). Comparing the component planes of the nine contributing economic indicators it can be observed that some of them are demonstrating similar structure and contribution to the final map. Looking at the component planes X5 (labor productivity per hour worked) and X6 (labor productivity per person employed), it is obvious that they are very similarly colored, indicating similar influences of those two components on the final map. Also, the first plane (indicator X1-GDP per capita in PPS-purchasing power standards) shows obvious similarity with

---

\(^{17}\) A component plane is a picture which displays the distribution of values of the respective component (indicator) over the map, thereby representing a cross-section through the map.

\(^{18}\) Different colors are used to indicate lower and higher values. Thus, the contribution of a particular indicator to the map and the correlation between indicators may be easily observed.
planes X5 and X6. These three component planes are very much alike, showing the largest value at the lower left part and generally larger values at the left-hand side, compared to the right side of the plane. Similar behavior of respective indicators may be confirmed with high values of Pearson coefficient of correlation (see Table 1).

![Figure 6. SOM on GEB structural indicators and component planes.](image)

Table 1

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>1</td>
<td>0.283</td>
<td>0.055</td>
<td>-0.493</td>
<td>0.909</td>
<td>0.918</td>
<td>0.380</td>
<td>-0.305</td>
<td>-0.061</td>
</tr>
<tr>
<td>X2</td>
<td>0.283</td>
<td>1</td>
<td>-0.232</td>
<td>-0.063</td>
<td>0.150</td>
<td>0.108</td>
<td>0.523</td>
<td>0.426</td>
<td>-0.078</td>
</tr>
<tr>
<td>X3</td>
<td>0.055</td>
<td>-0.232</td>
<td>1</td>
<td>-0.557</td>
<td>0.282</td>
<td>0.337</td>
<td>-0.250</td>
<td>-0.118</td>
<td>-0.311</td>
</tr>
<tr>
<td>X4</td>
<td>-0.493</td>
<td>-0.063</td>
<td>-0.557</td>
<td>1</td>
<td>-0.639</td>
<td>-0.670</td>
<td>-0.129</td>
<td>-0.068</td>
<td>0.381</td>
</tr>
<tr>
<td>X5</td>
<td>0.909</td>
<td>0.150</td>
<td>0.282</td>
<td>-0.638</td>
<td>1</td>
<td>0.969</td>
<td>0.349</td>
<td>-0.305</td>
<td>-0.161</td>
</tr>
<tr>
<td>X6</td>
<td>0.918</td>
<td>0.108</td>
<td>0.339</td>
<td>-0.670</td>
<td>0.969</td>
<td>1</td>
<td>0.237</td>
<td>-0.343</td>
<td>-0.109</td>
</tr>
<tr>
<td>X7</td>
<td>0.380</td>
<td>0.523</td>
<td>-0.250</td>
<td>-0.129</td>
<td>0.349</td>
<td>0.237</td>
<td>1</td>
<td>0.048</td>
<td>-0.114</td>
</tr>
<tr>
<td>X8</td>
<td>-0.305</td>
<td>0.426</td>
<td>-0.118</td>
<td>-0.068</td>
<td>-0.305</td>
<td>-0.343</td>
<td>0.048</td>
<td>1</td>
<td>-0.523</td>
</tr>
<tr>
<td>X9</td>
<td>-0.061</td>
<td>-0.078</td>
<td>-0.311</td>
<td>0.381</td>
<td>-0.161</td>
<td>-0.109</td>
<td>-0.114</td>
<td>-0.523</td>
<td>1</td>
</tr>
</tbody>
</table>

Although the Pearson coefficient of correlation measures only the linear correlation between variables, it also provides precious knowledge on GEB indicators’ dependences. Remarkable values of the coefficient are observed among several pairs of GEB indicators: indicators X1 (GDP per capita in PPS-purchasing power standards) and X5 (labor productivity per hour worked) and between indicators X1 and X6 (labor productivity per person employed) with a coefficient of correlation equal to 0.909 and 0.918 respectively. Also, there is a large coefficient of correlation between indicators X5 (labor productivity per hour worked) and X6 (labor productivity per person employed)—0.969, which is in compliance with the visual image of the relevant component planes in Figure 6. It is obvious that these indicators have similar distributions. In contrast, the following pairs of component planes: X1 (GDP per capita in PPS-purchasing power standards) and X4 (Inflation rate), X3 (general government debt) and X4 (inflation rate), X4 (Inflation rate) and X5 (labor
productivity per hour worked), as well as the pair X4 (inflation rate) and X6 (labor productivity per person employed), present images of more or less inverse distributions. The visual perception of their distributions on the appropriate component planes can be confirmed with the negative values of the respective correlation coefficients in Table 1.

**An Overall Kohonen Map—An Atlas of EU States Based on All SI Datasets**

Whilst the previous section describes the visualization and grouping of EU states based on each of six SI datasets separately, this section deals with the visualization of EU states based on all 76 structural indicators. Having applied the same procedure of mapping as it was previously described, including the pre-processing data phase. The following results are obtained (see Figures 7 and 8).

![Figure 7. Cluster indicator histogram for a general SOM map.](image)

(a) Three-cluster solution  
(b) Four-cluster solution  
(c) Five-cluster solution  
(d) Six-cluster solution  
(e) Nine-cluster solution  
(f) Eleven-cluster solution

![Figure 8. General map of EU states based on all SI datasets: Different cluster solutions.](image)

The cluster indicator graph in Figure 7 demonstrates that the most relevant grouping of EU states is in three clusters. Apart from this most striking cluster solution, some additional partitions are applied to illustrate the clustering tendency of EU states. The respective results are given in Figure 8.

The map partitions presented in Figure 8 provide a general view of the clustering tendency of EU states...
(from three to 11 clusters) based on all structural indicators.

The proposed groupings of EU member states may be the subject of further analysis, primarily of a group profile analysis. Here presented the results of the profile analysis referring to the four-cluster solution. The appropriate profile analysis for other cluster solutions may be carried out in a similar fashion.

The second map in Figure 8 shows the grouping of EU states in four clusters as follows:

- Cluster 1: Belgium, Ireland, Greece, Spain, France, Italy, Portugal, United Kingdom and Luxembourg;
- Cluster 2: Bulgaria, Estonia, Latvia, Lithuania, Poland, Romania, Slovenia, Czech Republic, Hungary and Slovakia;
- Cluster 3: Denmark, Netherlands, Austria, Germany, Finland and Sweden;
- Cluster 4: Cyprus and Malta.

The graph in Figure 9 shows the relative image of the four clusters with each cluster described by the appropriate SI mean values. The concrete height of a bar in the bar charts is given in the relative terms: It shows the deviation of a particular cluster SI mean value from the entire data set SI mean value.

As all deviations are measured by the same unit, i.e., standard deviation of the SI values for entire dataset, the previous bar charts provide a good solution for the comparison of cluster profiles. Thus, in comparing four clusters, the following points emerge:

- Going from the first to the last cluster there is a strong tendency of increasing deviations of the particular cluster’s SI means from the entire dataset mean values;
- The profile of the first cluster, consisting of Belgium, Ireland, Greece, Spain, France, Italy, Portugal, United Kingdom and Luxembourg can be considered to be an “average” cluster profile, since deviations of its SI mean values from the general average values are the ones closest to zero;
Most of SI means for the second cluster (Bulgaria, Estonia, Latvia, Lithuania, Poland, Romania, Slovenia, Czech Republic, Hungary and Slovakia) show negative deviations from the general SI mean values for all EU states. They fall in the range of -1.0 and 1.0 standard deviations around the zero line;

Unlike in the second cluster, the SI mean values for EU states belonging to the third cluster (Denmark, Netherlands, Austria, Germany, Finland and Sweden) mainly show the positive deviations from the general average values for all EU states;

The fourth cluster comprising just two states—Malta and Cyprus, shows substantial differences on both sides (positive and negative) from the general EU average of all structural indicators. The deviations are spread in the interval of -2.5 to 2.5 standard deviations around the general mean values.

In addition to the previous profile analysis where all seventy six structural indicators are applied and the respective comparisons between clusters are made, in the next part a more in-depth profile analysis of each cluster is proposed. Actually, a four-cluster solution in each group of the structural indicators is profiled, where SI standardized values serve as clustering variables. Figure 10 provides the respective graphical results.

Figure 10. Profile analysis of standardized SI datasets for the four-cluster solution.
Generally, these graphs confirm some of the previous conclusions. For almost all SI groups, except for GEB indicators and economic reform, the first cluster takes the role of the “average” cluster with the lowest variability of SI values. On the other hand, the fourth cluster shows the largest variability of average standardized SI values around the zero line, but again with the exception of GEB indicators. Cluster number three has clearly distinctive standardized values for the group of social cohesion, I&R and employment indicators, which is not the case for the other three SI groups (general economic background, environment and economic reform) where a lot of intersections with other clusters’ SI values are observed. The second cluster consisting of east European states (Bulgaria, Estonia, Latvia, Lithuania, Poland, Romania, Slovenia, Czech Republic, Hungary, and Slovakia), that have gone through the process of economic and social transformations, reveals specific behavior patterns regarding the GEB and economic reform indicators, but follows the first cluster’s values for employment, social cohesion and I&R indicators.

In addition to the above observations, Figure 10 indicates a different level of separation between four clusters for particular structural indicators. On the one side, in the domain of social cohesion, employment and I&R indicators, clear clusters’ separations exist, while on the other side, with regards to GEB, environment and economic reform indicators, a lot of clusters intersections dominate.

Concluding Remarks

This paper demonstrates how the Kohonen SOM algorithm can be used to provide a visual presentation of the relative position of EU member states according to the specific set of economic and social indicators defined by the European Commission as structural indicators (SI). Apart from the general map of EU states (based on all SI datasets), separate maps are presented for each particular SI domain: general economic background, employment, innovation and research, economic reform, social cohesion and environment. Since the SOM algorithm combines the goals of both projection and clustering, the paper also discusses a clustering tendency of the EU states based on all structural indicators.

Prior to the SOM application an exploratory analysis is performed by the usage of multiple box plots and graphical presentations of the coefficients of variation for each particular SI dataset. It reveals the presence of a higher variability for the economic reform, general economic background and I&R indicators comparing to the environment, employment and social cohesion indicators. The principal findings of SOM application can be summarized as follows:

1. The resulting maps provide intuitive and helpful graphical images of complex SI datasets. Regarding the clustering procedure, it is shown how an optimal cluster number is determined by the data itself. For this purpose, a specific operative measure—the cluster indicator—is used. Based on cluster indicator values, the three-cluster solution appears as the most relevant for the general map of EU countries (cluster 1—Belgium, France, United Kingdom, Luxembourg, Italy, Spain, Greece, Portugal, Ireland, Cyprus and Malta; cluster 2—Bulgaria, Estonia, Latvia, Lithuania, Romania, Poland, Hungary, Slovenia, Slovakia, and Czech Republic; cluster 3—Denmark, Netherlands, Austria, Germany, Finland, and Sweden);

2. Considering the clustering solutions for the general map (based on all 76 structural indicators) it is remarkable that geographic location of many countries is reflected in the final map organization, with some exceptions, notably Ireland and Luxembourg. This observation complies with the results of some earlier studies that are discussed in the second part of the paper;
(3) The previous conclusion concerning the relevance of "geography factor" doesn't apply in the same extent to all Kohonen maps in six SI domains. Going through clustering solutions in these areas, "geography" matters more regarding the general economic background, innovation and research, and less in the domain of economic reform, environment and social cohesion and the least regarding the employment area;

(4) Also, it is observed that two country groups—Eastern European countries and Scandinavian countries (Sweden, Finland) with Denmark and Netherlands, appeared as more homogenous and stable groups across all six SI domains comparing to other EU countries;

(5) Finally, the profile analysis (for the four-cluster solution) reveals the main clusters' features and also presents different divisions among clusters regarding particular SI datasets. Whilst in the sphere of employment, I&R and social cohesion, a clear separation between clusters of EU member states is recorded, on the other side, more intertwined trends between clusters dominate in the case of environment, GEB and economic reform indicators.

This paper presents only a static picture of EU countries within the scope of predefined SI datasets. As the Lisbon strategy may be considered as an ongoing historical process which assumes a dynamic dimension, the great challenge for future work is to expand this picture through a time-horizon, and explore the change of individual country positions in that respect.

References


