SEGMENTATION OF EU/EA COUNTRIES VIA CLUSTER ANALYSIS OF MACROECONOMICS INDICATORS

Nikolay Kulbakov

Abstract

The base of the research is author's desire to understand the economic position of EU countries better. It's an effort to allocate the states by groups with the help of cluster analysis based on macroeconomics indicators. It's interesting to consider the following characteristics: optimal numbers of groups, basic group forming criterions, distribution of countries by groups. The results of the analysis, based on a sample of EU27 for the 2003-2011 years, reflect adequately the dynamic how the economic situations of the countries are developing. The outcome shows the influence of the crises on a relative position of the Euro Union countries. During the situation prior to the crisis EU countries are easily divided into four clusters. The strongest cluster, from the economical point of view, consists of one country – Luxemburg. The second cluster in 2010,year of the crises, is divided into three groups: the best of the best, the best, the worst of the best. Ireland, which was financially successful, in 2011 year go down and pass to one of the worst cluster together with Portugal. The worst cluster in 2011 year is represented by the only one country – Greece.

Key words: cluster analysis, EU27, macroeconomic segmentation, MATLAB

JEL Code: E01, C38

Introduction

The methodology that is used in the research has been studied for more than half a century ago. Different types of cluster analysis are described in many books (Rezanková and Snásel 2009), scientific papers (Jain, Murty, and Flynn 1999), (Cattinelli et al. 2013), (Baker and Hubert 1975), (Faber 1994), and in others sources. Subject of cluster analysis is relevant to the present day, as used in practice widely and it is still the subject of study nowadays (Löster 2011). The role of technology Data Mining is growing now, and cluster analysis is one of the components of this technology (Berkhin 2006).

The article describes a methodology how to apply clustering analysis for macroeconomic data. Economists use the cluster analysis to study synchronization of business cycles (Papageorgiou, Michaelides, and Milios 2010), and to study the convergence of economies in different countries (Quah 1996). Clustering is used in researches of specialization and concentration in regions (Hallet 2002), to compare the quality of life in different cities (Cernakova and Hudec 2012) and for the construction of convenient ratings.

1 Parameters of the research

1.1 Goal Setting

The goal is to divide the EU countries into groups by cluster analysis. But some questions arose. By what characteristics the countries should be classified? Which of the known clustering methods should be used? How to implement it in an accessible environment MATLAB?

1.2 What kind of data was used for the analysis

Sampled population by countries. I used a set of data for countries from the database of Eurostat (Eurostat 2013), that is why all countries of the EU are included in the sample.

Sampled population by time. I decided to use annual data for several years. Generally, Eurostat provides statistics in a convenient form for the period 2000 – 2011 years, for the indicators that are interesting for me. Some pointers have data for the year 2012, at the time of writing. But all types of data are including information for the period between 2003 and 2011 at the same time. That is why I decided to use time series in the calculation of clusters of annual averages for 2003-2011 years.

Sampled population by data type. I compare the countries by intensive characteristics that reflect the effectiveness and express the problems on a national scale. EU countries are differentiated from each other significantly by the size of territory and by population size. I use the indicators, transformed by population size, using purchasing power parity, and the volume of GDP.

Sampled population by type of macroeconomic indicators.

I decided to split countries into groups according to the following criteria and using the following indicators:

- The effectiveness of the economy:
  - Index of Labor productivity per person employed a value of 100 at the level of EU27
  - Index of GDP per capita in PPS a value of 100 at the level of EU27 countries
- The living standard of the population:
  - Index of GDP per capita in PPS a value of 100 at the level of EU27 countries
  - Index of Annual net earnings a value of 100 at the level of EU27 countries
- The level of living costs:
  - Index of Price level indices a value of 100 at the level of EU27 countries

- The level of economic activity:
  - Index of labour force rate a value of 100 at the level of EU27 countries

$$ILFR = \frac{\mathrm{ILF_i}}{\mathrm{ILF_{EU27}}} * 100 \quad (1)$$

where $ILF = \frac{\text{Active population of country}}{\text{Total population ofcounty}}$, $ILF_{EU27} = \frac{\text{Active population of EU27}}{\text{Total population of EU27}}$

- By the criteria of fiscal discipline and as a result, the confidence of the financial markets:
  - Index of Government deficit where a value of 100 means the worst of EU27 countries and 0 means the best of EU27

$$IGD = 1 - \frac{\mathrm{D_i - D_{min}}}{\mathrm{D_{max} - D_{min}}} * 100 \quad (2)$$

where D*i* –is % Government deficit of GDP

  - Index of bond yields a value of 100 at the level of EU27 countries

1.3 Methods

How to classify the data. It is technically possible to do clustering based on all data containing the entire time series, but because of the contained in the time series autocorrelation, analysis will not have a high quality. Therefore it was decided to carry out the analysis for the years 2003-2011 separately, but including a single set of 7 different indicators listed above.

How to fill the missing data. The data are sets of index values with an average (or at Government deficit maximum) value of 100, for all the countries studied EU27 for each year 2003-2011. Looking at the data collected in the table, it is clear that in each table are missed two or three data cells, and it means that the average analysis lacks 1% of data set.

When I tried to use the arithmetic mean to substitute the missing values this method has been improved, immeasurably reality, the situation of lagging or decreased the leaders. So I wrote a program in the MATLAB software, which complements the missing data, calculating them based on the regression analysis.

What methods are used for cluster analysis.

In the research described in this article, I use two methods:

- The hierarchical approach: hierarchical clustering
- The probabilistic approach: k-means

To study the data set was created a script using these methods. The main result of the hierarchical cluster analysis will be dendrogram reflected some years, and the distribution of

countries into 2-8 groups. In interpreting the dendogramm there is a problem: the lack of unambiguous criteria for selection of clusters. In the literature, they recommend to use two methods - visual analysis and comparison clustering results, performed by other methods.

The results of the hierarchical cluster analysis can be verified iterative cluster analysis by the method of k-means. When comparing the classification of groups of respondents have a share of over 70% of matches (more than 2/3 matches), the cluster solution is adopted. By implementing this method, I was faced with inconsistency assessment of comparison because k-means method has in its algorithm for calculating random composes and clusters based on this algorithm is not constant for the different members of the same data but different calculations. The solution is a multiple launch the program and analysis of the distributions into n 2-8 groups on the basis of the hierarchical analysis (consistent) and k-means (inconsistent). Thus, the comparison of the groups several times can help you determine whether to accept the cluster solution.

1.4 Step by step description of the MATLAB algorithm

To solve the problems of clustering in MATLAB already exist described algorithms for the analysis of hierarchical and k-means (MathWorks site). Therefore, the problem is reduced to the proper implementation of the comprehensive study of the algorithm.

Step 1: Load the data from EXCEL to MATLAB

The cycle "different year": all further steps will be repeated in a cycle for each following year, the first year 2003, the last 2011.

Step 2: Regression analysis on all available complete sets of data for a given year and the addition with the resulting linear function of the missing data. At this stage, the problem may occur, the resulting value may be beyond the allowed values, and it must be controlled manually. Another drawback of this approach is that I do not deal with the statistical significance of each regression analysis, because the volume of added data is less than 2%.

Step 3: Hierarchical analysis: standard algorithm of MATLAB library is used. The algorithm makes calculations on a "nearest neighbor" method with the Euclidean metric. On the way out, we have 7 distributions of countries into 2-8 groups and visualization in the form of a dendrogram of the nearest neighbors.

Step 4: Analysis using the k-means method: standard algorithm of MATLAB library is used. The algorithm computes with set parameters:

- The number of clusters = 2,3..8

- Distance = sqEuclidean - Squared Euclidean distance. Each centroid is the mean of the points in that cluster.
- Replicates = 10 - Number of times to repeat the clustering, each with a new set of initial cluster centroid positions. K-means returns the solution with the lowest value for sumd.

On the out, we have 7 distributions into clusters for groups (2-8 groups), and a description of the centroids, which is just chosen at random, which is the reason of not consistency the results of the k-means method.

Step 5: Comparing the results - which distribution into groups has greater similarity.

This step is an intuitive analog of finding the optimal number of clusters, because it gives the answer how two distributions into clusters are matched, based on various analyses.

The cluster solution is adopted if a match is 70% or more, as mentioned earlier. We're looking for the division that will have the highest percentage of matches. This algorithm is implemented by me. It works on the basis of selection of key containing group number, and based on the matrix variants received by permutation.

Why it is necessary. Imagine two distributions of the same objects A, B, C into two groups. We can get the same distribution based on different methods:

$$(A, B) \cap (C) = 0 \tag{3}$$

That record is clear, but the computer can give the following answers:

Object	HYE1	KMEAN1	HYE2	KMEAN2	HYE3	KMEAN3
A	1	1	2	1	1	2
B	1	1	2	1	1	2
C	2	2	1	2	2	1

That is why we need the key that will compare the data and show whether an object is in the same group or not. In our case, the key is the matrix:

$$\begin{matrix} 1 & 1 \\ 1 & 2 \end{matrix} , \begin{matrix} 2 & 2 \\ 2 & 1 \end{matrix}$$

If we apply this matrix to the comparison above, it will give the answer that YES, the objects belong to the same group, for all the above proposed distributions. Detailed implementation of the algorithm can be found in the software code in the application (Kulbakov 2013).

The output of the algorithm we have the matrix of better matches, which includes data about the percentages of similarities and the number of groups with the best results.

The essence of the algorithm - the matrix of permutations, imposes computational restrictions on the analysis of the number of groups, as the number of different combinations is equal to

n! Where n - the number of groups. That's why I restricted myself by 8 groups. If you take n more than 8, the calculation takes too much time.

The problem with inconsistency of the results received by k-means method, I solved through multiple calculations and comparing the results time and time again. I watch a few times which are issued by the distribution and clustering solution adopted for the sample EU27. So I decided that for the years 2003-2009 could be applied a distribution into four groups, for 2010 year into 8 groups, and for 2011 into five groups.

2 Results of the research

2.1 Output

Tab. 1: Distribution of countries by groups

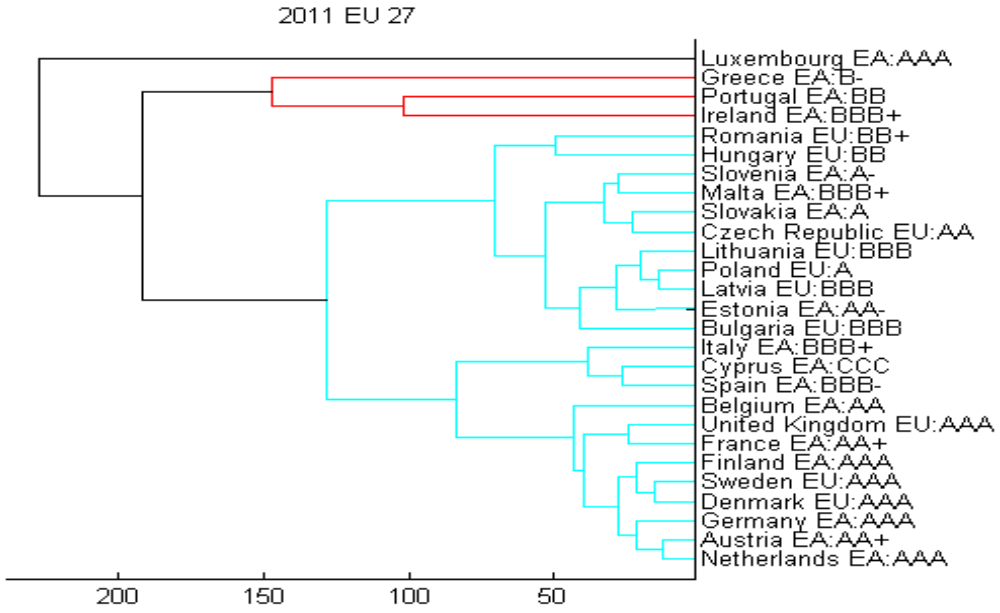
Country	2003 C4	2004 C4	2005 C4	2006 C4	2007 C4	2008 C4	2009 C4	2010 C8	2011 C5	EURO	S&P may2012
Belgium	A	A	A	A	A	A	A	AAA	A	EA	AA
Bulgaria	C	C	C	B	C	C	B	C	C		BBB
Czech Rep.	B	C	C	B	C	B	B	B	C		AA
Denmark	A	A	A	A	A	A	A	AAA	A		AAA
Germany	A	A	A	A	A	A	A	AAA	A	EA	AAA
Estonia	C	C	C	B	C	B	B	B	C	EA	AA-
Ireland	A	A	A	A	A	A	A	A	B	EA	BBB+
Greece	B	B	B	B	B	B	B	GR	GR	EA	B-
Spain	A	A	A	A	B	A	A	AA	A	EA	BBB-
France	A	A	A	A	A	A	A	AAA	A	EA	AA+
Italy	A	A	A	A	A	A	A	AA	A	EA	BBB+
Cyprus	B	B	B	B	B	A	A	AA	A	EA	CCC
Latvia	C	C	C	B	C	C	C	D	C		BBB
Lithuania	C	C	C	B	C	C	C	C	C		BBB
Luxembourg	LU	LU	LU	LU	LU	LU	LU	LU	LU	EA	AAA
Hungary	C	C	C	C	C	C	C	C	C		BB
Malta	B	B	B	B	B	B	B	B	C	EA	BBB+

Netherlands	A	A	A	A	A	A	A	AAA	A	EA	AAA
Austria	A	A	A	A	A	A	A	AAA	A	EA	AA+
Poland	C	C	C	B	C	C	B	C	C		A
Portugal	B	B	B	B	B	B	B	B	B	EA	BB
Romania	C	C	C	C	C	C	C	C	C		BB+
Slovenia	C	B	B	B	B	B	B	B	C	EA	A-
Slovakia	C	C	C	B	C	B	B	B	C	EA	A
Finland	A	A	A	A	A	A	A	AAA	A	EA	AAA
Sweden	A	A	A	A	A	A	A	AAA	A		AAA
UK	A	A	A	A	A	A	A	AAA	A		AAA

Source: own calculations

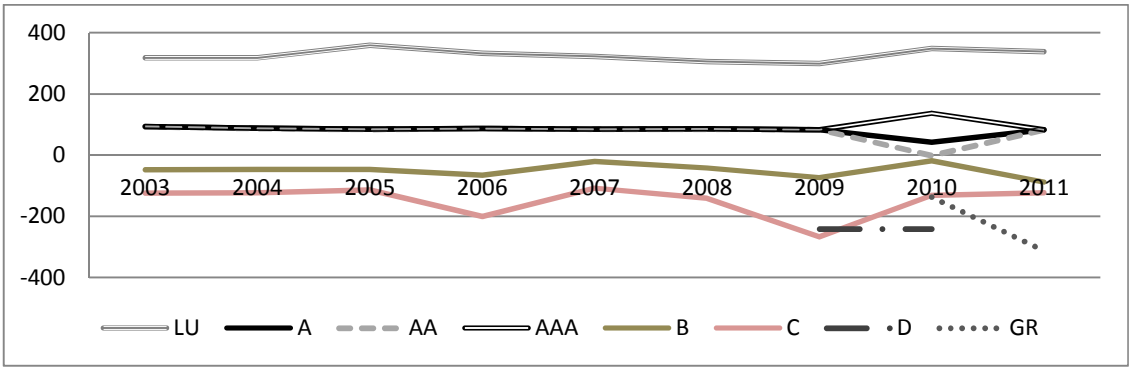
At the output of the research I received better distribution of matrix groups. For convenience, I picked up the names of the groups according to the ratings. Rating has been calculated based on the average values of the indicators for each group given the nature of the indicators, i.e. it is better if the indicator is higher (annual earnings, or vice versa, price level). For a visual analysis of the nearest neighbors, will be given only one dendrogram based on 2011 year. For more informativeness, next to the name of the country is mentioned "EA", if the country belongs to Euro Area, or it is mentioned "EU" if the country belongs to European Union, but it is not in the Euro zone. Also, there is an S&P Sovereigns Rating List next to the name of the country (S&P 2013). Dendrogram for the remaining years is in the attachment.

Fig. 1: Output: dendrogram 2011 EU27



Source: own calculations

Fig. 2: Rating of clusters EU27



Source: own calculations

2.2 Commentary on the results

The logic of the table which shows the distribution on clusters is that the most successful country Luxembourg creates the most successful cluster LU, the indicators of Luxembourg never comes close to the other countries during all the year of the research, and Luxembourg can be combined with any other country.

A cluster of the most developed countries in the EU follows after Luxemburg, with the name of A. Cluster A includes countries such as Belgium, Denmark, Germany, Ireland, Spain, France, Italy, Netherlands, Austria, Finland, Sweden, United Kingdom. During the economic crisis of 2010 year this cluster is divided: on the best of the best -AAA, average -AA, and

worst of the best -A. A cluster is stable, and Spain leaving it in 2007, but then returned. During the crisis, Ireland lost its top position and moved in 2010 to "worst of the best –A", and moved to the worst in 2011, because of the high cost of loan rates on government bonds and high budget deficit.

The third cluster B and the fourth cluster C contain EU countries that are weak and the weakest in economic terms. These two clusters are characterized by low stability of the country and every year is moving between them.

An example of Cyprus is very interesting, the country moves from low developed one to the category " best of the best". It is unfortunate that the research does not reflect the statistics for 2012 and 2013. It is well known that Cyprus in 2013 was in a very difficult economic situation. From the research it is clear that there was a prosperous economic situation even in 2011 year.

Greece is situated in the second group in 2003- 2009, and moves into a single cluster in 2010, after the debt crisis, and the cluster reflects the worst economic of European countries in 2011.

Portugal is stable earlier in the cluster B, but after the economic crisis it's also left the cluster and moved to C. Latvia has improved its position and moves from the weakest countries to the cluster B.

As a result there are such countries as Bulgaria, Czech Republic, Estonia, Latvia, Lithuania, Hungary, Malta, Poland, Romania, Slovenia, and Slovakia in 2011 in cluster B.

Portugal and Ireland are in the weakest cluster in 2011 - in cluster C.

From all the dendrograms for the period 2003 – 2011 is clear that the nearest neighbor for the Czech Republic is Slovakia in the selected indicators. On the dendrogram for 2011 is seen that Greece, Ireland and Portugal are close to each other, these countries have the problems that became apparent on the background of recent global economic crisis. Also from the dendrogram is clear that the most developed economies, with the credit rating of AAA, are together.

Luxemburg, that is situated separately, is the owner of an excellent credit rating, because "Country of banks" surpasses the European countries by level of living.

Conclusion

As a result of the research was created the tools with which it is possible to carry out the cluster analysis with similar parameters quickly and efficiently. The experience that was received can be used in the future for different sets of data and for different sets of countries. After the research it became clear that even on the basis of a small set of data you can construct an adequate picture of the distribution of country groups with different state of macroeconomics. There is a desire to continue the research in this direction and expand the scope. It is possible to increase the complexity in terms of quantity, i.e. it is possible to increase potentially the sample of countries in the region, continent, hemisphere, and the whole world. It is also real to improve the quality of the model, namely, to think over what set of indicators could present more adequate picture of distribution. One might think the best way to process the data to comparing two countries; make an overall rating or a special one, such as credit rating, foreign trade, standard of living, economic strength, etc. The results of this research could be interesting to the public, and could be used as material for journalists and teachers.

Acknowledgment

This article was created with the help of the Internal Grant Agency of University of Economics in Prague No. 6/2013 under the title „Evaluation of results of cluster analysis in Economic problems."

References

Baker, F., & Hubert, L. (1975). Measuring Power of Hierarchical Cluster-Analysis. *Journal of the American Statistical Association*, *70*(349), 31–38. doi:10.2307/2285371

Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques. In J. Kogan, C. Nicholas, & M. Teboulle (Eds.), *Grouping Multidimensional Data* (pp. 25–71). Springer Berlin Heidelberg. Retrieved from

Cattinelli, I., Valentini, G., Paulesu, E., & Borghese, N. A. (2013). A Novel Approach to the Problem of Non-uniqueness of the Solution in Hierarchical Clustering. *Ieee Transactions on Neural Networks and Learning Systems*, *24*(7), 1166–1173. doi:10.1109/TNNLS.2013.2247058

Cernakova, V., & Hudec, O. (2012). Quality of Life: Typology of European Cities Based on Cluster Analysis. *E & M Ekonomie a Management*, *15*(4), 34–48.

Eurostat. (2013). *Eurostat research database*. Retrieved from http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database

Faber, V. (1994). Clustering and the continuous k-means algorithm. *Los Alamos Science*, *22*, 138–144.

Hallet, M. (2002). Regional Specialisation and Concentration in the EU. In P. J. R. Cuadrado-

Roura & P. M. Parellada (Eds.), *Regional Convergence in the European Union* (pp. 53–76). Springer Berlin Heidelberg. Retrieved from

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *Acm Computing Surveys*, *31*(3), 264–323. doi:10.1145/331499.331504

Kulbakov, N.(2013) *Source for MATLAB cluster analysis, input and output data*. Retrieved from http://www.ilovecz.ru/research/matlab01.zip

Löster, T. (2011). *Hodnocení výsledkůmetod shlukové analýzy*. FIS VŠE. Retrieved from https://isis.vse.cz/auth/lide/clovek.pl?id=8340;zalozka=7;studium=92375

MathWorks. (2013). Cluster Analysis. Retrieved from http://www.mathworks.com/help/stats/cluster-analysis.html

Papageorgiou, T., Michaelides, P. G., & Milios, J. G. (2010). Business cycles synchronization and clustering in Europe (1960-2009). *Journal of Economics and Business*, *62*(5)

Quah, D. T. (1996). Regional convergence clusters across Europe. *European Economic Review*, *40*(3-5), 951–958. doi:10.1016/0014-2921(95)00105-0

Rezankova, H., & Loster, T. (2013). Shlukova analyza domacnosti charakterizovanych kategorialnimi ukazateli. *E+M. Ekonomie a Management, 16(3),* 139-147. ISSN: 1212-3609

Rezanková, H., & Snásel, V. (2009). *Shluková analýza dat*. Praha: Professional Publishing.

S&P. (2013, May). Sovereigns Rating. Retrieved from http://www.standardandpoors.com/ratings/sovereigns/ratings-list/en/us?sectorName=null&subSectorCode=39

Contact

Nikolay Kulbakov

University of Economics, Prague

W. Churchill Sq. 4

130 67 Prague 3

Czech Republic

kulbakov@gmail.com