CLASSIFICATION OF EUROPEAN UNION COUNTRIES FROM DATA MINING POINT OF VIEW, USING SAS ENTERPRISE GUIDE

Ana Maria Mihaela Iordache[1]
Ionela Catalina Tudorache[2]
Mihai Tiberiu Iordache[3]

Abstract

With the development of computers and the increasing the amount of data it appeared the need of identifying new acquaintances, unknown until that moment in a relatively short time. The term of data mining involves the analysis of data from different views (aspects) in order to extract the knowledge to use them further in the foundation of decisions at microeconomic or macroeconomic levels. In this article we apply data analysis techniques such as principal component analysis and cluster analysis in order to group the countries of the European Union based on the financial indicators registered at the end of 2009 year.

Keywords: classification, cluster, data mining, European Union, financial indicators

Introduction

The development of computers and implicitly the computing power and the large volume of data it led the necessity to discover new information in a relatively short time. In this context, since 2000 it has developed a new technology called data mining. The main disciplines that intersect with data mining are: statistics, the computer programming, the computer assisted learning, the database technology, the digital technology, the information technology, etc. From this point of view the data mining methods come from statistics, database management and artificial intelligence (neural networks, data analysis, image processing, learning assisted by computer, genetic algorithms, etc.). Among data mining techniques, the most common are: exclusion, classification (clustering), discrimination and prediction.

In this context, data mining, also known as "discovery of knowledge in the large databases" is a modern and a powerful information and communication technology, a tool that can be used to extract the useful information but still unknown. This automates the discovery of relations and combinations in the raw data and the results found could be placed in an automated decision support.

The results obtained using data mining techniques may vary and they are specific to each type of user. In general, the data mining techniques are applied in a database for two reasons: validation of assumptions about data, their correctness, and statistical assumptions and the discovery of new features from data. The discovery, in its turn, can be divided into description and prediction. Data description is achieved either by calculating the various statistical indicators (the elementary ones) of the raw data such as

[1] academic assistant Ph.D. at Romanian-American University, iordache.ana.maria.mihaela@profesor.rau.ro
[2] student at Romanian-American University
[3] engineer at IMSAT Bucharest

the average, the variance, the standard deviation, etc. or by the application of advanced data analysis techniques such as the cluster analysis, the principal component analysis, the discriminant analysis, etc. The prediction aims to accurate the expression of some future values data under the analysis.

The most used data mining techniques are: the exclusion of which data processing means in terms of information; classification (clustering) is the operation through the objects in a given lot shall be divided into subsets called "classes" based on similarities and differences between them; the discrimination is different from classification because of the need, in applications of some prior knowledge related classes; the predicting which is attainable based on the trend.

In conclusion, data mining has evolved as a technology because of two complementary events: first of all the continuously expanding amount of data after the database development technologies and the instruments for collecting data and, the second, the need for knowledge concretized by the need to filter and interpret all these volumes of data stored in databases, data warehouses or data banks.

Application of data mining techniques on European Union countries

The analysis of the country risk is a dynamic and current subject. In the context of globalization, we cannot speak about a company action on the individual national markets (regarded as independent entities), but to the company action on global market efficient portfolios constructed on the basis of geographical location business. Even if in the economic theory, to a higher risk it is assigned a higher potential earnings in the country risk theory this principle does not work, because of uncontrollable, by the company, of the country risk factors. All these elements justify the importance of granting at international level the country risk. [8]

In the application presented below we made a classification of European Union countries (Austria, Denmark, Sweden, Finland, Germany, Cyprus, Latvia, Poland, Lithuania, Netherlands, Spain, Portugal, France, Italy, Romania, Greece, United Britain, Bulgaria, Belgium, Hungary, Malta, Slovakia, Czech Republic, Estonia, Slovenia, Luxembourg, Ireland) according to the financial indicators recorded at the end of 2009. The indicators values we have taken from the sites of the European Institute of Statistics [9] and the International Monetary Fund [10]. To facilitate the work we have coded the names of financial ratios as shown in table 1.

Table 1. Financial indicators used in the application

Indicator	The full name of the indicator
IF1	Average annual rate of inflation (reported to the consumer price index)
IF2	The fiscal government deficits (% of GDP)
IF3	Fixed public investment (% of GDP)
IF4	The public debt (government)% of GDP
IF5	Current account balance (% GDP)
IF6	The external government loans (% of GDP)

To reflect as well the reality the values of the indicators were recorded as percentage of GDP (table 2). This enables the comparative analysis between any two countries.

Table 2. The values of financial indicators used in the application

	IF1	IF2	IF3	IF4	IF5	IF6
Austria	0.4	3.50	1.1	67.5	5.332	1,373
Belgia	0.0	6.00	1.9	96.2	-0.267	-5,799
Bulgaria	2.5	4.70	4.9	14.7	-9.465	0
Cipru	0.2	6.00	4.1	58	-9.345	-6,077
Cehia	0.6	5.80	5.2	35.3	-0.997	-5,961
Danemarca	1.1	2.70	2	41.4	3.998	-3,042
Estonia	0.2	1.70	5.1	7.2	4.603	0
Finlanda	1.6	2.50	2.8	43.8	1.384	-2,373
Franta	0.1	7.50	3.3	78.1	-1.451	-7,874
Germania	0.2	3.00	1.6	73.4	4.791	-3,285
Grecia	1.3	15.40	3.4	127	-11.217	-12,866
Ungaria	4.0	4.40	3.1	78.4	0.405	0
Irlanda	-1.7	14.40	4.7	65.5	-2.944	-11,446
Italia	0.8	5.30	2.4	116	-3.365	-5,313
Letonia	3.3	10.20	4.3	36.7	9.438	0
Lituania	4.2	9.20	3.9	29.5	3.821	0
Luxemburg	0.0	0.70	3.5	14.5	5.735	-1,110
Malta	1.8	3.80	2.2	68.6	-3.892	-4,050
Olanda	1.0	5.40	3.9	60.8	8.714	-4,914
Polonia	4.0	7.20	5.2	50.9	-1.646	0
Portugalia	-0.9	9.30	2.4	76.1	-10.057	-9,334
Romania	5.6	8.60	5.3	23.9	-4.398	0
Slovacia	0.9	7.90	2.3	35.4	-3.195	-6,300
Slovenia	0.9	5.80	4.6	35.4	-0.295	-6,127
Spania	-0.2	11.10	4.4	53.2	-5.064	-11,447
Suedia	1.9	0.90	3.6	41.9	6.359	-2,213
Marea Britanie	2.2	11.40	2.7	68.2	-1.321	-10,892

The classification was performed using Enterprise Guide program, being a part of the Statistical Analysis Software (SAS). According to SAS working diagram (figure 1) on the original data matrix we have applied the techniques of principal components analysis, factor analysis and cluster analysis in order to group countries according to financial indicators.

Because where IF6 financial indicator (the foreign government loans) in 2009, some countries have registered no value for analysis as close to the reality we have standardized the data. Consequently, in the working diagram from SAS there is an additional procedure, namely the standardization of data (figure 1).

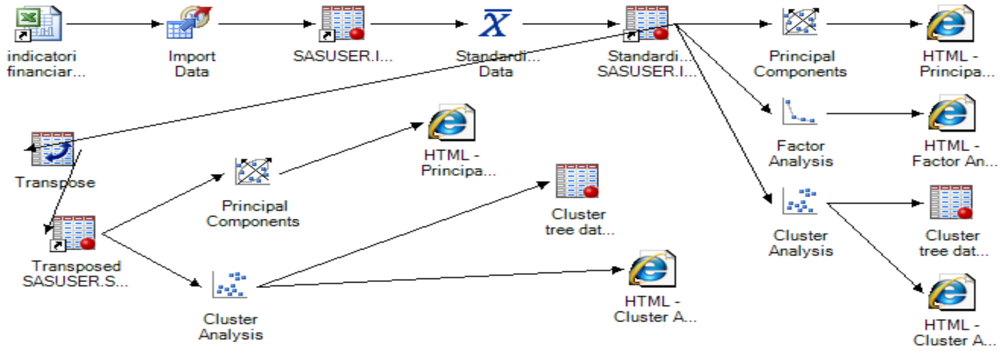

Figure 1. The working diagram from SAS Enterprise Guide for the financial indicators

The principal component analysis is intended for a matrix X the identifying of some new synthetic variables that explain the old variables so the amount of information provided by the cloud of points to be lost in a controlled way. [4]

The first step in the principal components analysis is the determining of the correlation matrix. This matrix indicates the intensity of relations between all pairs of variables.

Table 3. The correlation matrix of financial indicators

Correlation Matrix						
	IF1	IF2	IF3	IF4	IF5	IF6
IF1	1	0.0081	0.2833	-0.2474	0.1062	0.1218
IF2	0.0081	1	0.2298	0.3681	-0.4776	-0.8643
IF3	0.2833	0.2298	1	-0.5132	-0.1206	-0.336
IF4	-0.2474	0.3681	-0.5132	1	-0.3308	-0.3338
IF5	0.1062	-0.4776	-0.1206	-0.3308	1	0.5999
IF6	0.1218	-0.8643	-0.336	-0.3338	0.5999	1

In the correlation matrix (table 3) it can be observed that IF6 (the foreign government loans) is correlated with IF5 (the current account), positively, and in the negatively correlated the indicators are IF6 (the foreign government loans) to IF2 (the deficit tax) and IF4 (the public debt - government) with IF3 (the fixed public investment).

Table 4. The values and the amount of information recovered from the cloud of points

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.583245	0.8882181	0.4305	0.4305
2	1.6950269	0.8724143	0.2825	0.713
3	0.8226126	0.2621495	0.1371	0.8501
4	0.5604631	0.3181991	0.0934	0.9436
5	0.242264	0.1458758	0.0404	0.9839
6	0.0963883		0.0161	1

Eigenvectors						
	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6
IF1	0.104624	0.449673	0.866867	-0.12472	-0.03604	-0.13614
IF2	-0.5529	0.119073	0.141637	0.428815	-0.37784	0.577427
IF3	-0.11438	0.687915	-0.28078	0.024766	0.62521	0.208214
IF4	-0.3292	-0.53922	0.372063	0.084074	0.66997	0.080676
IF5	0.471959	0.003258	0.050214	0.863393	0.11369	-0.12787
IF6	0.582375	-0.14008	0.093213	-0.21779	0.05713	0.762778

The information about the quality of adjustment is expressed by the eigenvalues of the correlation matrix and their properties. Regarding the quantity of information recovered from the cloud of points (table 4) it is observed that after the first four eigenvalues we are recovered the largest amount of information (94.36%) and this was confirmed by chart of their own values (figure 2) .

The Eigenvalue column indicates the coefficients of eigenvalues associated with the principal components. Because the analysis is based on the calculation of the correlation matrix the data are standardized (each variable has the variance equal to 1 and the total variance is equal to 11, in this case). The Difference column expresses the difference between their values and it shows the way of their decreasing. The column Proportion shows the proportion of each eigenvalue in the sum of all eigenvalues. In the Cumulative Column it is the calculated the sum of Proportion column and it indicates the quantity of information recovered. Thus, after the first four axes we stop because we recovered 94.36% of the information.

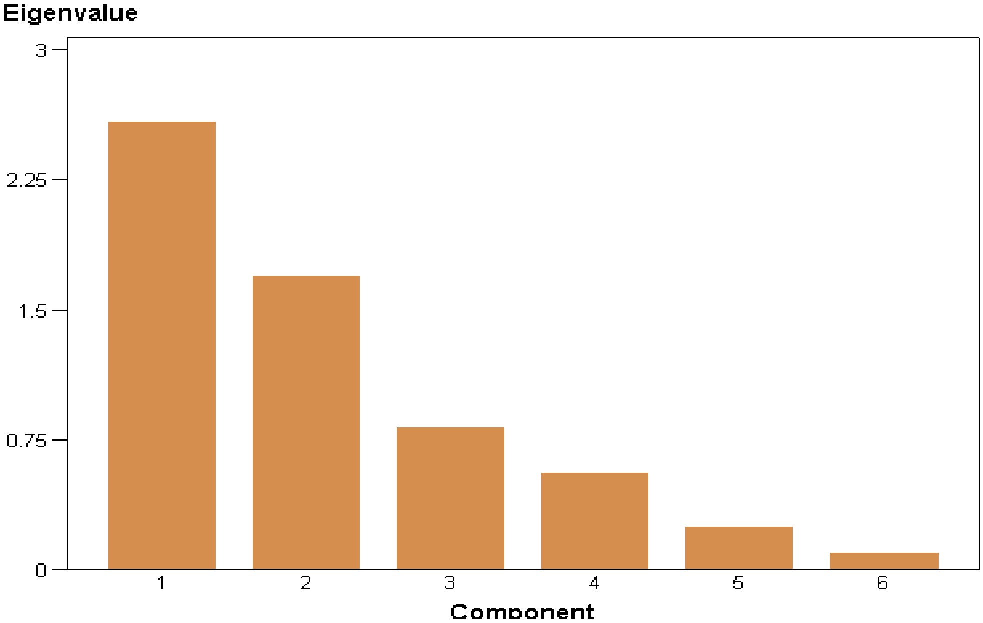

Figure 2. The graph of eigenvalues for the financial indicators

The graph named scree plot (figure 2) shows the graph of eigenvalues coefficients from table 4 and it highlights the best informational contribution brought by each component. Based on this chart it shall be determine the optimal number of principal components.

Information about the principal axes is shown in table named Factor Pattern (table 5). The column of a factor provides information about the weights ("coefficients") that each financial indicator participates to the description of a particular factor. This factor can be expressed, therefore, as a linear combination of financial indicators, with coefficients given in table 5.

Table 5. The correlation coefficients

Factor Pattern		
	Factor1	Factor2
IF1	0.16816	0.58544
IF2	-0.8886	0.15502
IF3	-0.1838	0.89562
IF4	-0.5291	-0.702
IF5	0.75855	0.00424
IF6	0.93602	-0.1824

Standardized Scoring		
	Factor1	Factor2
IF1	0.0651	0.34539
IF2	-0.344	0.09146
IF3	-0.0712	0.52838
IF4	-0.2048	-0.4142
IF5	0.29364	0.0025
IF6	0.36234	-0.1076

Variance Explained by Each Factor	
Factor1	Factor2
2.58325	1.695027

The variation is explained by two factors as follows: the first factor explained 2.58325 (43.05% of the total information), and the second factor explains 1.695027 (it has a contribution of 28.25% from total information). In the structure of the two indicators aimed to identifying the factors that influence positive and negative over the data (table 5). Thus it is noted that IF6 (foreign government loans) and IF3 (fixed public investment) influence positively and IF2 (fiscal deficit) influences negatively the analysis.

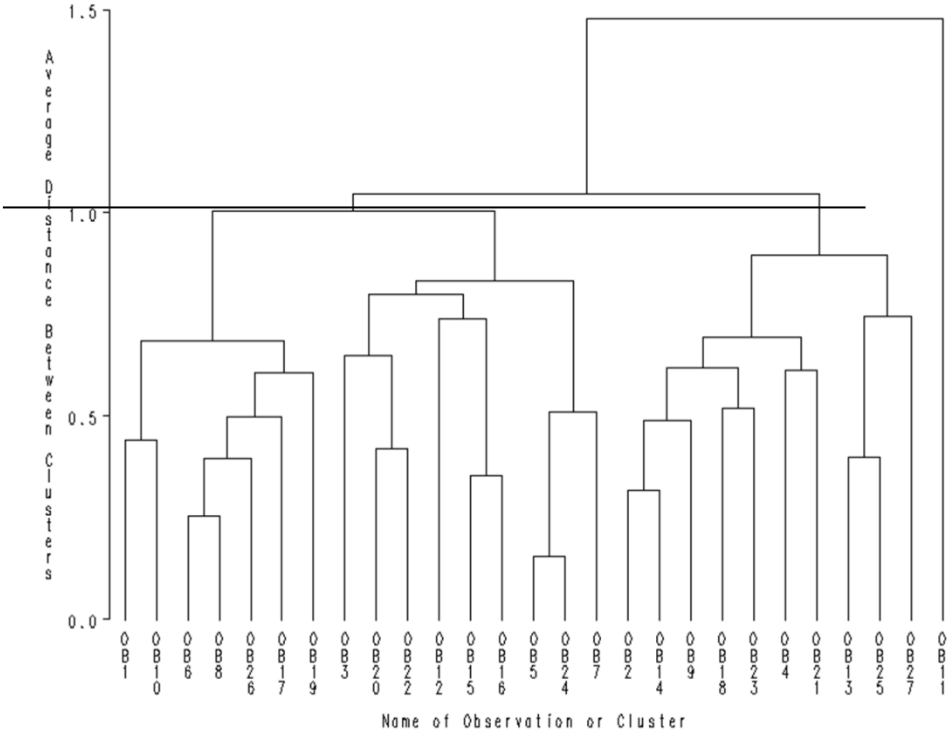

Figure 3. The dendrogram of the financial indicators

From the dendrogram (figure 3) it is observed that the object 11 (Greece) aggregates later, at an average distance between clusters of more than 1.25. For that reason Greece is considered in terms of data mining, as outlier and it should be removed from the analysis. After removing the object 11, if the average distance between the clusters is considered less than 1 then it will be three classes of countries, as follows:

- Class 1 consists of countries: Austria, Germany, Finland, Sweden, Luxembourg and the Netherlands;

- Class 2 consists of countries: Bulgaria, Poland, Romania, Hungary, Latvia, Lithuania, Czech Republic, Slovenia and Estonia;
- Class 3 consists of countries: Belgium, Italy, France, Malta, Slovakia, Cyprus, Portugal, Ireland, Spain and Great Britain.

In the Table 6 there are presented the countries scores for each factor of 11 resulted from the analysis, one for each studied indicator. Further it will retain only two factors because they best explain the variance (they have the highest values in table 6). For a better view of the group of the countries these should be represented in terms of factors 1 and 2.

Table 6. The scores of countries in the composition of each factor

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
Austria	1.40612	-1.6926	0.41066	0.16462	-0.7599	2.75108
Belgia	-0.173	-1.5412	0.20614	0.21604	0.43258	-0.279
Bulgaria	-0.0565	1.37892	-0.4311	-2.3105	-0.4899	-0.0936
Cipru	-0.5259	-0.0101	-0.8675	-1.7103	0.53808	0.97654
Cehia	0.04751	0.85985	-1.1603	-0.1515	0.98397	0.65637
Danemarca	1.07223	-0.6673	0.0267	0.02618	-1.1278	-0.8131
Estonia	0.89313	1.03766	-1.8664	0.30163	0.63988	-1.933
Finlanda	0.98595	-0.2843	0.1282	-0.6132	-0.2544	0.11537
Franta	-0.5515	-0.5594	-0.3598	0.3444	0.60824	-0.4639
Germania	0.82376	-1.4595	0.0815	0.45432	-0.0497	-0.6108
Grecia	-2.6249	-0.6116	1.04141	0.06101	0.7671	0.46515
Ungaria	0.19668	0.00691	1.84315	-0.3012	1.01265	-1.7577
Irlanda	-1.7094	0.1364	-1.7488	1.55878	-0.0257	1.39639
Italia	-0.3529	-1.4783	0.76504	-0.5507	1.86668	0.21674
Letonia	0.33923	1.11223	0.88695	2.34123	-0.4522	0.86472
Lituania	0.24644	1.19919	1.29778	0.92401	-1.2671	0.27281
Luxemburg	1.61715	0.02247	-1.3502	-0.0483	-0.2433	0.41958
Malta	0.27747	-0.7678	0.68765	-1.2807	-0.2637	-0.3023
Olanda	0.61491	-0.0179	-0.139	1.69118	1.32405	-0.0328
Polonia	-0.0895	1.36357	1.02344	-0.3487	1.25661	0.64935
Portugalia	-1.2939	-1.0332	-0.7814	-0.9376	-1.1334	-0.3243
Romania	-0.1129	2.14635	1.5507	-0.9566	-0.3575	0.92956
Slovacia	-0.1141	-0.2651	-0.2073	-0.3609	-2.5392	-0.0796
Slovenia	0.11074	0.66771	-0.8375	-0.0391	0.37907	0.07592
Spania	-1.3621	0.40726	-1.1565	0.43273	-0.3978	-0.7115
Suedia	1.3833	0.10582	0.05767	0.12885	0.9984	-0.5587
Marea Britanie	-1.048	-0.0562	0.89887	0.96445	-1.4458	-1.8293

From the matrix scores of countries from the composition of each factor (table 6) the first two factors best explain the variance (table 5). Further the countries will be represented in terms of two factors to identify a possible classification better than the one offered by the dendrogram. From the representation it should observed that Greece is an outlier (as shown in the dendrogram), many countries are located very close to the axis, but there are other countries that cannot be associated with any class. In conclusion the dendrogram provides an effective image on the classification of countries than representation of the countries in terms of factors.

If we take the principal axes as a reference point when we will obtained a grouping of the countries into four distinct classes, but at the same time it will be some countries in the area of uncertainty (they cannot be associated to any class), so the classification based on dendrogram is more conclusive. For the future scenarios of improving the financial situation of a country it will be retain the classification results based on dendrogram.

Conclusion

Any human activity, both at microeconomic and macroeconomic levels takes place in conditions of risk and uncertainty. Some of these can be avoided easier or harder depending on the level of knowledge, the degree of evaluation or the importance which are given in foundation decisions.

At the microeconomic level, in the context of globalization, the development of business by the economic agents on external markets will be achieved only if there is an incentive powerful enough, able to motivate the companies to take the risks involved carrying out the activity in a particular country. The country rating calculated by different institutions is a very important aggregate indicator when considering the opportunity of investing or not in that area.

At the macroeconomic level, the country risk analysis involves identifying the problems which may arise in a particular state by honoring the obligations from its international commitments taken externally.

Within the European Union it is important for establishing the economic, social and/or the financial policies in order to avoid the macroeconomic imbalances and in achieving the sustainable growth supported, where necessary, the adoption of structural reforms. In order to achieve this, in the last years it is more and more talking about a new concept named macro prudential. But however efficient macro-prudential policies would be they cannot replace the macroeconomic, social or financial policies adopted within each state.

Bibliography

[1] Choudhary A. K., Harding J. A., Tiwari M. K., Data mining in manufacturing: a review based on the kind of knowledge, Springer Science, Vol. 20, Nr. 5 / October, 2009, pg. 501-521, ISSN 1572-8145

[2] Fernandez George, Discriminant Analysis, A Powerful Classification Technique in Data Mining, Statistics and Data Analysis, http://www2.sas.com/proceedings/sugi27/p247-27.pdf

[3] Hallahan Charles, Atkinson Linda, Introduction to SAS Enterprise Guide 4.1 for Statistical Analysis, http://www2.sas.com/proceedings/sugi31/109-31.pdf

[4] Meyers L., Gamst G., Guarino A.J., Data analysis using SAS Enterprise Guide, Cambridge University Press, 2009, ISBN: 9780511601842

[5] Plemmons Howard, Working with a DBMS using SAS Enterprise Guide, SAS Global Forum, Las Vegas, Nevada, 2011
[6] Spircu Liliana, *Analiza datelor. Aplicații economice*, editura ASE, 2005, ISBN: 9735947013
[7] Using the SAS Enterprise Guide (Version 4.1), http://www.stanford.edu/group/ssds/cgi-bin/drupal/files/Guides/Using_the_SAS_Enterprise_Guide_4.1_0.pdf
[8] http://www.businessdictionary.com/definition/sovereign-risk.html
[9] http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/
[10] www.fmi.com