Journal of International Scientific Publications:
Materials, Methods & Technologies, Volume 6, Part 2

AN OVERVIEW OF FREQUENTLY USED ALGORITHMS TO BUILD CLUSTERS

Raluca-Mariana Stefan[1]

[1] Academy of Economic Sciences, 6 Piata Romana, Bucharest 010374, Romania

Abstract

Unsupervised classification of data is a technique used to determine the natural tendency of data to group and can be applied to predict economic phenomena which are, more or less, stable. Therefore, cluster techniques constitute an area of great interest.

Unlike the supervised classification, unsupervised classification proposes an optimal partition of feature space in terms of a certain mathematical criterion, without using apriori information (for example, a reference partition). The advantage of these methods is given that is fully automatic and requires no user intervention; it can be used for classification of data for which we do not have information relative to their content (number of classes, the prototype class etc.).

Key words: *data classification, unsupervised learning, cluster analysis, clustering algorithm*

1. INTRODUCTION

Having a sample or a population of economic data we are interested to make an optimized or predictive model considering the yet unknown test data.

The continuously growing large and complex volume of electronic data requires numerously improvements to the methodology used to automatically understand, process, and summarize data (Jain, 2009). We do not have only sure and specific information and if we want to group the data in order to become useful we need to use an object recognizing system.

The systems which are representing series of activities that are necessary in a process to recognize objects, sets of information, procedures, algorithms and techniques used to do this are known as an object recognizing system. All the informational system parts are used to identify the belonging of an object to a certain class extracted from the analyzed population. The input data is represented by the properties of the objects and the result is the data output meaning the object classes determined by some specific classification rules. There are two types of object recognizing systems: unsupervised and supervised recognizing systems. Either can be used regarding the specific goals, the nature of the information involved, the specificity of the methods and tools, and the nature of the data output.

An unsupervised learning system is used to systematize, group and summarize the information in case of large volume of analyzing data that are highly heterogeneous. So, the unsupervised learning techniques are very useful and efficient in the preliminary analysis of data.

In most of the cases a distinction is made between the two learning methods (unsupervised – clustering; supervised – classification). Unsupervised learning, also known as cluster analysis, involves only unlabeled data (Duda et al., 2001). The main characteristic of the unsupervised learning systems is that we do not know the belonging of the analyzed objects to a certain class or another. That means that the number of classes is also unknown until the final phase of classification.

Unlike the supervised classification, unsupervised classification proposes an optimal partition of feature space in terms of a certain mathematical criterion, without using apriori information (for

example, a reference partition). The advantage of these methods is given that is fully automatic and requires no user intervention; it can be used for classification of data for which we do not have information relative to their content and the number of classes or the prototype class or any other prior information is unknown.

2. USING DATA CLUSTERING TO GROUP INFORMATION

Using cluster analysis is important because it efficiently groups data and allows the user to make consistent interpretations based on that.

There are many articles and research studies that have shown beginning with the second half of the last century that clustering techniques are used in varied and complex domains in order to classify huge volumes of data.

Data analysis contains methods and techniques that are considered the most appropriate instruments used in order to identify similar or dissimilar structures from different fields of activity. Economic field is essential to the other fields that involve people's life. In this paper, we are interested in the economic field.

The process of data analysis is defined as an informational process. This kind of process has primary data as input data and summarizing data as output data. We give simple data and through this process we can obtain, using the right method, complex data that can help us making decisions.

Applying cluster analysis procedures to a dataset of object the result will be that any object will be assigned to just one class and the set of formed classes have a meaningful and generalizing significance based on what we can make series of important conclusions for the knowledge process.

In 2007, Tabachnick and Fidell offered an overview for the numerous statistical techniques and methods proposed to use for data analyses (Jain, 2009). Cluster analysis was considered one of them. Cluster analysis is a classification technique having the property that the objects form the classes to whom they are belonging. The number of classes is not known. Cluster analysis is considered by some authors a statistical method and a „collection of different algorithms", a non-typical statistical technique, by others.

Some papers consider that there are two fundamental criteria that have to be accomplished: the similarity of the objects from a class and the dissimilarity of any two objects belonging to two different classes to be maximum.

These two can also be found as a single general criterion that implies maximum dissimilarity between object classes and minimum dissimilarity between any two objects from a class.

The outputs of the cluster analysis are not reduced to a simple and unique configuration of the analyzed objects classes and that involves choosing a certain number of classes.

The possibility of making analysis or prediction in the economic area is conditioned by the existence of the available data regarding level, structure or time evolution of the investigated processes and phenomena. Finding relevant information from massive volumes of data that were structured with the help of cluster techniques becomes easier and the consistency of interpreted data is higher.

Identifying classes, categories or informational groups out of a large amount of data represents the main goal for a cluster analysis.

The following fundamental challenges are associated with clustering and they still are relevant (Jain and Dubes, 1988):

- What is a cluster?
- What features should be used?
- Should the data be normalized?
- Does the data contain any outliers?
- How do we define the pair-wise similarity?
- How many clusters are present in the data?
- Which clustering method should be used?
- Does the data have any clustering tendency?
- Are the discovered clusters and partition valid? (Jain and Dubes, 1988).

Being known the complexity of solving economic activity problems there have been many research papers and numerous studies written on the subject of the recognizing forms systems. The interest in the unsupervised classification has grown for the last decades because it does not need any a priori information to form groups of similar objects.

A generalization of cluster analysis cannot be done considering the vast and diverse number of clustering methods that have been developed in different fields having different definitions.

The variety of clustering techniques is reflected by the variety of terms used for cluster analysis: bryology, classification, clumping, competitive learning, morphometric, nosography, nosology, numerical taxonomy, partitioning, Q-analysis, systematics, taximetrics, taxonomic, typology, unsupervised pattern recognition, vector quantization, and winner-take-all learning, aciniformics and agminatics, the latter being proposed by Good in 1977 (SAS Institute Inc. 2008. SAS/STAT® 9.2 User's Guide. Cary, NC: SAS Institute Inc., 2009 Electronic Edition).

3. SPECIFIC TERMS AND NOTIONS USED IN CLUSTER ANALYSIS

Cluster analysis is a set of varied unsupervised classifying procedures and algorithms. Unsupervised classification of data is a technique used to determine the natural tendency of data to group and can be applied to predict economic phenomena which are, more or less, stable. Therefore, cluster algorithms constitute an area of great interest.

The most important term and the most used term in this analysis is *data.* Data is a quantitative and qualitative expression of some processes and phenomena met in reality. *Cluster analysis* is a term that has been used for the first time by R. C. Tyron in his work with the same title.

Cluster analysis aims to look and identify classes, groups or clusters within the sets of forms or objects so that the elements which are belonging to the same class to be more similar and the elements belonging to different classes to be less similar as possible. That means that cluster analysis is a way to examine the objects in order to see similarities and dissimilarities and to cluster them in classes that are distinct and have homogeneous inside.

The *feature* or *characteristic* is another frequently used term in cluster analysis because it represents the main way to group some observed objects into clusters.

Measure is a number computed in many ways in order to establish the common features between the studied objects of a dataset.

Class is a group of objects. *Cluster* is a list of objects which have some minimum common features. A cluster may be a class or a group and it can be a subset of similar objects that have enough similar features to be considered similar. A cluster can be defined as a subset of objects that have the property that their degree of dissimilarity between any two objects of the cluster is less than the degree of dissimilarity between any two objects that belong to two different clusters.

A cluster can be described as a multidimensional space region that can have a relatively high density of objects.

The *distance* between two clusters is computed based on the clustering method.

The *similarity* indicator is determined based on a similarity measure that can be computed considering the objects that we need to group.

The *dissimilarity* can be computed as indicator or coefficient and it is defined as the distance between two clusters.

4. ALGORITHMS USED TO BUILD CLUSTERS

Algorithms that can be used for carrying out the construction of clusters, know a great variety including heuristic algorithms, optimization algorithms and fuzzy algorithms. The differences between how to build clusters by an algorithm or another are determined mainly by the nature of the method used to evaluate distances between clusters. Moreover, even the type of cluster analysis results from the nature of the algorithm used to build clusters.

In terms of their nature, mode of operation and type of solutions they provide, cluster analysis methods can be divided into two broad categories: hierarchical methods and iterative methods or the partition type. The first category includes agglomerative and divisive clustering methods.

For every hierarchical clustering method there are some specific procedures and we can mention a few: simple agglomerative method, complete agglomerative method, medium agglomerative method, Ward's method etc. Partitioning or iterative clustering methods include a series of algorithms: k-means algorithm, k-medoids algorithm, CLARA algorithm, fuzzy algorithm etc.

The algorithms frequently used to build cluster group in two main categories: *partitioning methods* and *hierarchical methods*. We can easily notice that two clusters can be build out of the clustering techniques. A partitioning algorithm tries to find an optimal clustering considering applying a criterion for a fixed considered number of clusters.

Cluster analysis represents the study of methods and algorithms used for grouping or clustering objects according to measured or perceived intrinsic characteristics or similarity (Jain, 2009). Despite the fact that thousands of algorithms were developed beginning with the second half of the last century there is one that passed the test of time and efficiency: k-means algorithm.

It is the most used algorithm, ever since it was proposed by MacQueen in 1967. It is still used to present. K-means algorithm uses the Euclidean distance to determine the distance between objects and cluster centers; when Mahalanobis distance was used (Mao, Jain, 1996) the computational cost increased.

Hierarchical algorithms have two groups: agglomerative and divisive methods. Agglomerative algorithms start considering a number *c* of clusters and then merge them. Divisive methods start with only one cluster and then divide it successively.

The hierarchical agglomerative clustering methods are most commonly used. The construction of a hierarchical agglomerative classification can be achieved by the following general algorithm:

- find the two closest objects and merge them into a cluster and find and merge the next two closest points, where a point is either an individual object or a cluster of objects;
- if more than one cluster remains then return to the point where it can be found and merge the next two closest points.

Generally, the partitioning methods generate a number of clusters and each object belonging to one cluster that can be represented by a centroid or a cluster representative. Depending on the clustered object type it can be the arithmetic mean if the data is numeric, or a list of some frequently met words for document data. Often there are cases when the big number of the clusters determines for the centroids to be clustered themselves in order to form hierarchy in a set of data.

The most frequently used methods to evaluate the distance between two clusters are: nearest-neighbor method, furthest-neighbor method, weighted pair-group average, unweighted pair-group average, unweighted pair-group centroid, weighted pair-group centroid (median), Ward's method, k-means etc. (Ruxanda, 2009).

Ward's method is different because in order to evaluate the distance between two clusters is using an analysis of the objects variance. It makes the sum of squares of any two considered clusters minimum at each step of the clustering process. The advantage of using such a method is that is very efficient and the disadvantage is that it can create small size clusters.

Based on Ward's method there have been proposed another algorithm, Ward2 that is discussed in Kaufman and Rousseeuw (1990), and Legendre and Legendre (2012) (Murtagh, Legendre, 2011).

Some papers suggest that it could be useful to use both of the clustering algorithms, hierarchical or partitioning in order to make a data classification. Because it is an automatic process, the relevance of classes tends to be lower than for supervised classification, being dependent on the method used and the discriminatory power of features of the space that was used.

5. BUILDING CLUSTERS OUT OF DATA

5.1. Using Statistical Data to Build Clusters

Performing a cluster analysis to classify a set of objects supposes the following phases:

• choosing the characteristics based on which a classification will be made;

• choosing the measure type that will be used to evaluate the distance between the objects;

• setting the rules to build the classes or clusters

• building the classes;

• checking the classifying consistence and significance;

• choosing an optimal number of clusters, in terms of classifying problem nature and the scopes;

• interpreting the clusters significance.

Based on the cluster analysis output results certain rules regarding the evolution of some phenomena population can be deduced, useful principles for the knowledge process can be identified or a series of general scientific conclusions can be formulated.

Many computer science products offer clustering procedures which can be used to cluster the variables of a data set. As a result, hierarchical clusters and disjoint clusters are obtained. If the data are numerical variables then they will be analyzed directly by the procedures.

Studying the most important works from the past and recent literature regarding data clustering it can be noticed the fact that in the economic field only marketing and financial bankruptcy were interested in using the advantages offered by cluster analysis. We consider that we can at least try to apply cluster analysis to another economic field and hope for some good results.

In the context of population growth and increasing natural resources consuming, sustainable development is a model that aims for an equilibrium between economic growth, the quality of life and the preserving of the environment without unlimited use of the natural resources beyond the power of the Earth.

The most known and cited general definition of the concept of sustainable development is to be found in Brundtland Report of the World Commission on Environment and Development (WCED): "Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs".

Natural dimension of the sustainable development is based on the fact that the environment is limited as far as the resources outputs that can be attracted in the economic cycle are considered. It is also based on the absorption capacity for the toxic waste resulted from the economic activity. All these considering that the quality of environment factors is increased. Both of the mentioned limits have a dynamic character, continuously evolving. That thing happens because increased economic growth rhythms and limited resources make nature a restrictive factor for economic activity.

The evolution of life and civilization standards and economic development are influenced by the diminished capacity of the nature to regenerate. So, the natural environment can no longer be considered an endless reservoir.

Social-human dimension of the sustainable development is more than obvious. Countries economic evolution together is strictly influenced by a number of macroeconomic indicators like: resource productivity, population education level, employment rate, unemployment rate by gender or by age etc.

GDP (Gross Domestic Product) is relevant if we can determine the groups of indicators that could modify it. An analysis is presented considering a series of different level relevant indicators.

For a better perspective over the chosen theme we present the necessity of studying the behavior and influence of some considered countries or groups of countries based on chosen macroeconomic indicators.

The statistical data used for this analysis were taken from the European Commission site www.epp.eurostat.ec.europa.eu that offers an annual statistical bulletin about the most important macroeconomic indicators.

The European Commission gives the necessary databases on which one could run a variety of economic analyses and prognoses. The data used refers to some of the level indicators which influence the economy of a country: Real GDP, Employment Rate, Unemployment Rate, Energy Consumption etc. Data is represented by percentage. There are many studies over the annual increasing or decreasing rates of macroeconomic indicators and they allow comparisons of the economic development dynamic between economies of different dimensions.

We propose a test of data clustering made for the data mentioned above in order to find out a good number of clusters. In the two tables numeric variables, macroeconomic indicators, have the following

meanings: GDP – Gross Domestic Product, ER – Employment Rate, EnC – Energy Consumption, URE – Unemployment Rate By Education Level, URG – Unemployment Rate By Gender, RP – Resource Productivity, URA – Unemployment Rate By Age.

Table 1. Primary statistical data used to build clusters (part 1)

Country	GDP	ER	EnC	URE	URG	RP	URA
EU(27)	1.53	16.09	95.04	2.34	0.06	7.03	0.08
EU(25)	1.97	39.08	2472.27	1.42	0.13	3.64	0.35
EU(15)	2.17	8.59	69.28	1.11	0.03	2.23	0.01
Bulgaria	1.14	24.94	505.41	3.35	0.04	1.61	0.10
Czech Rep.	2.31	2.58	9.18	1.49	0.04	7.94	0.04
Denmark	4.69	0.76	6.50	1.71	0.14	14.41	0.00
Germany	2.92	21.88	419.43	1.28	0.01	19.22	0.57
Estonia	1.26	10.96	20.87	1.02	0.02	53.29	0.09
Ireland	3.11	18.40	4.40	1.35	0.00	47.28	0.00
Greece	1.45	9.17	88.90	2.02	0.06	4.59	0.05
Spain	4.45	3.76	9.82	1.19	0.02	15.38	0.01
France	2.69	11.12	27.36	1.38	0.01	13.91	0.00
Italy	2.39	2.41	1.04	1.19	0.00	31.28	0.05
Cyprus	1.06	2.49	4.03	1.17	0.00	58.04	0.00
Latvia	1.52	1.60	82.59	2.91	0.05	0.76	0.04
Lithuania	1.11	35.13	1063.31	1.64	0.10	1.17	1.54
Luxembourg	2.74	6.11	1.65	2.26	0.01	36.57	0.03

5.2. Running clustering procedures to build clusters

In this case, a relevant classification is the one of the objects meaning the countries. We are interested in this kind of classification because of the events that have taken place following the large globalization.

Together we can preserve better our world but only if all of the involved countries work side by side. Initially, the hierarchical classification method is used so the number of clusters is unknown and the applied algorithm offers multilevel solutions that differ by the cluster number and by the agglomerative degree. Choosing the optimum number of classes is realized regarding the proposed objectives for the obtained statistical results.

The variance of a variable is a statistical measure for the dispersion of the variable values around its mean value. The covariance matrix is a generalization for the concept of variance for a number of variables.

The Eigenvalues of the Covariance Matrix table has on the Cumulative column the proportion of the associated variance to each eigenvalue. We notice that the value of standard deviation is 1 and for the distance between variables is computed 3.464157.

Table 1. Primary statistical data used to build clusters (part 2)

Country	GDP	ER	EnC	URE	URG	RP	URA
Hungary	1.35	12.73	10.26	0.86	0.01	90.86	0.00
Malta	1.68	2.99	10223.88	1.48	0.06	10.03	0.02
Netherlands	6.61	2.79	51.51	7.59	0.08	24.25	0.14
Austria	1.26	7.78	154.04	1.87	0.12	2.94	1.35
Poland	1.23	1.08	0.52	7.16	0.00	327.58	0.00
Portugal	1.42	3.73	3.10	3.67	0.02	15.56	0.00
Romania	1.12	49.18	2390.07	1.72	0.21	6.01	20.30
Slovenia	0.59	6.26	76.14	10.79	0.13	7.22	0.02
Slovakia	0.19	5.40	193.22	2.12	0.10	6.17	0.06
Finland	2.94	9.64	97.94	0.63	0.03	15.85	0.02
Sweden	5.89	3.47	55.33	2.16	0.09	3.85	7.80
UK	5.54	8.44	96.40	1.08	0.17	8.75	0.02
Norway	3.92	5.89	118.12	8.07	0.30	9.48	0.06
Switzerland	1.46	22.98	10.95	0.85	0.01	19.36	0.01
Croatia	0.87	2.43	21.38	1.92	0.02	15.19	0.03
Turkey	1.15	9.89	57.43	1.32	0.05	6.43	0.01

5.2. Running clustering procedures to build clusters

In this case, a relevant classification is the one of the objects meaning the countries. We are interested in this kind of classification because of the events that have taken place following the large globalization.

Together we can preserve better our world but only if all of the involved countries work side by side. Initially, the hierarchical classification method is used so the number of clusters is unknown and the applied algorithm offers multilevel solutions that differ by the cluster number and by the agglomerative degree. Choosing the optimum number of classes is realized regarding the proposed objectives for the obtained statistical results.

The variance of a variable is a statistical measure for the dispersion of the variable values around its mean value. The covariance matrix is a generalization for the concept of variance for a number of variables.

The Eigenvalues of the Covariance Matrix table has on the Cumulative column the proportion of the associated variance to each eigenvalue. We notice that the value of standard deviation is 1 and for the distance between variables is computed 3.464157.

Table 2. Covariance Matrix Eigenvalues

Eigenvalues of the Covariance Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	3.01094821	1.43669959	0.4301	0.4301
2	1.57424862	0.31404618	0.2249	0.6550
3	1.26020244	0.41331133	0.1800	0.8351
4	0.84689111	0.56796502	0.1210	0.9560
5	0.27892609	0.26094675	0.0398	0.9959
6	0.01797933	0.00717513	0.0026	0.9985
7	0.01080421		0.0015	1.0000
Root-Mean-Square Total-Sample Standard Deviation				1
Root-Mean-Square Distance Between Observations				3.464157

The next graph presents three statistics used to estimate the number of clusters: the values of the cubic clustering criterion (CCC), pseudo F (PSF) and t^2 (PST2) statistics.

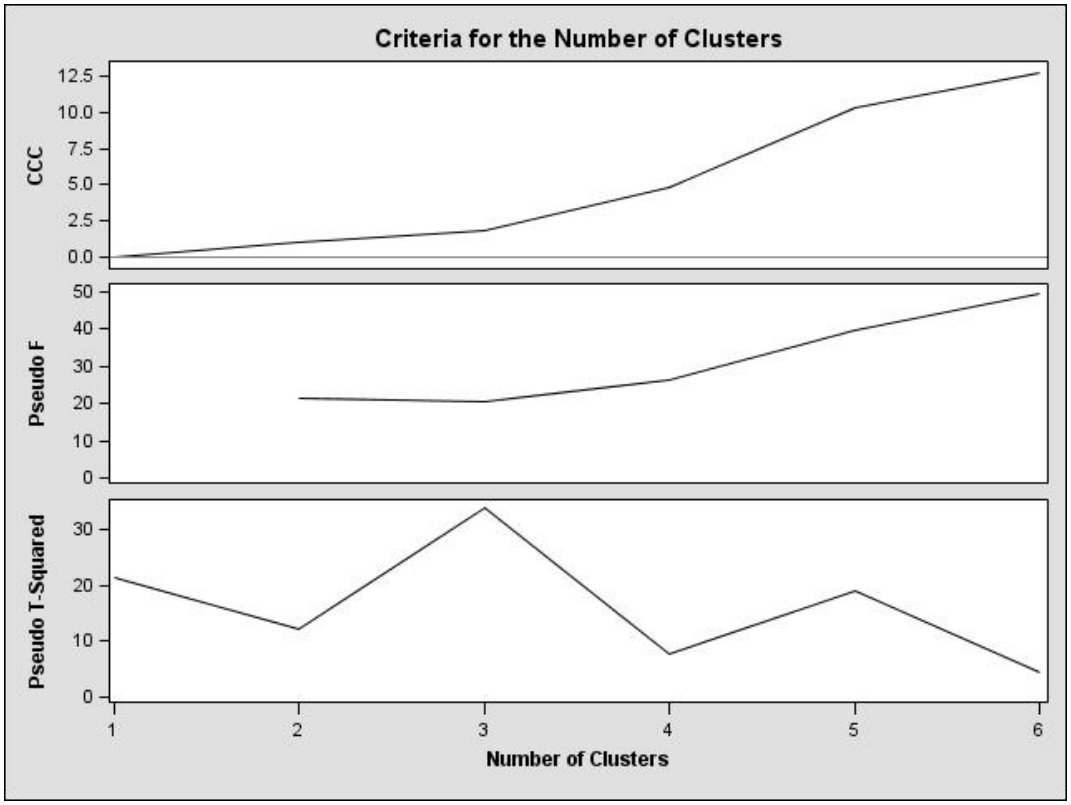


Fig. 1. Three statistics results used to estimate the number of cluster

The values of CCC indicate the existence of a good number of clusters. The relatively large value of Pseudo F statistic points out that it could be a number of 4 clusters or 2 clusters. The values of the Pseudo T-squared indicate that there can be a number of 5, 4, 3 or 2 clusters.

Taking all three statistics in consideration we can say that the data we have analyzed can be clustered in 4 or 2 clusters. Considering the large number of countries that have been studied, that is 34, we consider that the chosen indicators have a strong connection and together they group to influence the global economy. The countries clusters have been built in terms of macroeconomic indicators.

Hierarchical clustering method is used to group the data in a tree structure named dendogram or phenogram. The dendogram below shows the graph containing the clustered observations.

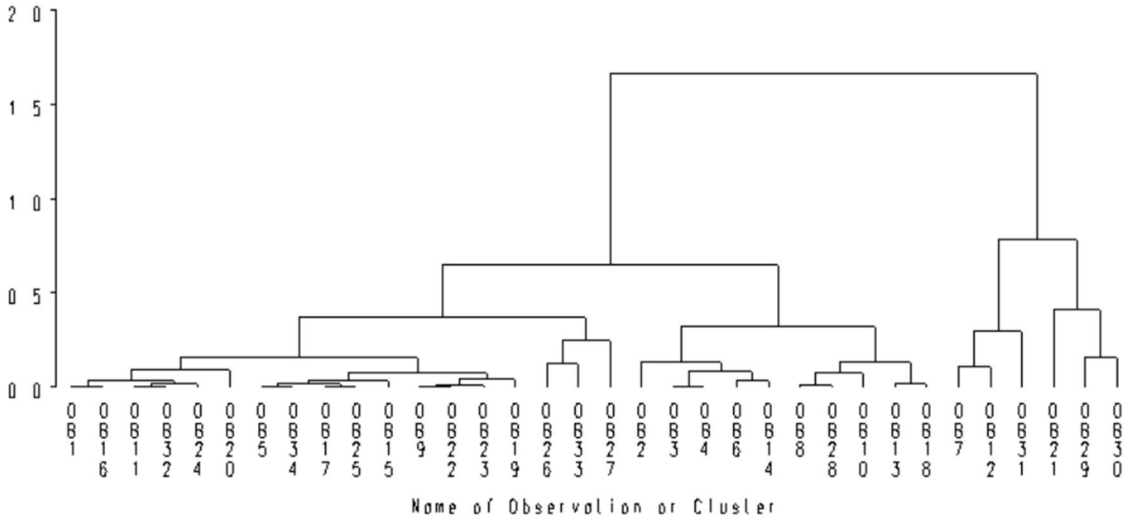


Fig. 2. Resulted dendogram (OB number is observation number) with data clustering levels

The default tree plot given above shows two large clustering groups. The left group has 28 observations while the right group has only 6 observations. There are several sub clusters within these two clusters and they correspond to the next levels of classification. The second level of classification has 4 sub clusters. So, after we have been used some raw data the graph showed us that those data can form two large groups or four smaller groups based on their found common characteristics.

6. CONCLUSIONS

Comparative analyses and cluster analysis in this field present an increasing importance over the last years. The results that are obtained can help specialists from all over the world to consider having the same hypotheses after running a cluster procedure over the data and maybe take the same decision that would be the best for everyone. After all we have to participate to the global economy.

To make a research about the global economy based on the macroeconomic indicators we usually take a sample of objects from the studied population.

Cluster analysis offers us a way to perform a study considering a large amount of data. The data considered may not have obvious common features so we want cluster analysis to find some. The researcher must found as much as significant information as he can and to summarize it having scientific arguments.

We propose further study and research in this direction to obtain relevant results in the economic field. Some future analyses would prove to be interesting another study comparison over the clusters build from data regarding national economic indicators for a country. The results are relevant to latter research and adapted to current and projected economic reality. The results will make digital information so accessible and they will support all users of digital information that need their business to align the standard of European information.

REFERENCES

1. Ciupagea C., Manoleli D., Niţă V., Papatulică M., Stănculescu M. (2006) The 3rd Study of the European Institute from Romania – The Strategic Directions of Sustainable Development in Romania, 178 pp.

2. World Commission on Environment and Development (1987), Our common future, Oxford University Press.

3. Jain, A.K. (2010) Data clustering: 50 years beyond K-means, Pattern Recognition Letters, 35 Volume, 8 Issue, Elsevier B.V., pp. 651-666.

4. Ruxanda, G. (2009) Multidimensional data analysis – Doctoral School Course Notes, 133 pp.

5. Murtagh F., Legendre P., Ward's (2011) Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm, http://arxiv.org/abs/1111.6285

6. Jain, A.K., Dubes, R.C. (1988) Algorithms for Clustering Data, Prentice Hall Inc., 320 pp.

7. Demyanyk, Y., Hasan, I. (2009) Financial crises and bank failures: a review of prediction methods, Published by Bank of Finland Research 35.

8. Van der Heijden F., Duin R.P.W., de Ridder D., Tax D.M.J. (2004) Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB, Wiley, 423 pp.

9. Pedrycz W. (2005) Knowledge-Based Clustering: From Data to Information Granules, Wiley, 316 pp.