European Research Studies,
Volume XII, Issue (1) 2009

A Cluster Analysis Approach for Banks' Risk Profile: The Romanian Evidence

By

Nicolae DARDAC [1] Iustina Alina BOITAN [2]

Abstract:

Cluster analysis, as an exploratory technique, by gathering together those credit institutions sharing similar features in terms of financial intermediation activity, proves to be a complementary tool for the peer group analysis, accomplished at the off-site supervision level. The aim of our study was to include a representative sample of Romanian credit institutions into smaller, homogenous clusters, in order to assess which credit institutions have similar patterns according to their risk profile and profitability. We found that, over the period 2004-2006, the clusters remained relatively stable in terms of similarity of exposure to risks and profitability.

Keywords: *banking system, supervision, risk profile, resemblance coefficient, and cluster analysis.*

JEL Classification: C44, E58, G21

1. Introduction

The implementation of Basel II provisions has generated effects for both individual credit institutions, which must rigorously identify, quantify and manage risks, and for supervisory authorities. Thus, the passing from conformity with the prudential banking regulations approach to a risk-based approach requires the update of the traditional surveillance methods, by adding up new quantitative robust techniques, in order to assess in real time the adverse changes occurred in banking activity, which can increase banks' risk exposure.

[1] *Ph.D., Academy of Economic Studies, Bucharest and professor at the Faculty of Finance, Insurance, Banks and Stock Exchanges, Prof. Romana square, no.2, Bucharest, nicolae.dardac@fin.ase.ro.*
[2] *Assist. Prof., Academy of Economic Studies, Bucharest and professor at the Faculty of Finance, Insurance, Banks and Stock Exchanges, Teleajen street, no.54, Bucharest; tel. 0744403356; iustinaboitan@yahoo.com.*

In our opinion, this evolution will be reflected preponderant in the off-site supervision activity, because its basic role is to centralize accurate, high quality, timely information's and to process them by means of several techniques, which range from simpler statistic ones to more sophisticated and complex ones, namely stress-tests and early warning systems. The first category of techniques includes analyses that compute financial ratios, purpose to duplicate the results of the on-site exams. These ratios will be then introduced in *financial ratio* and *peer group* methods, to give a global picture on the activity of a credit institution, to detect tendencies in the banking system and to signal a potential impairment of the banking activity.

Having as purpose the identification of those credit institutions which are similar in terms of their risk profile, we have applied a methodology that has attracted the interest of different economic entities (both economic agents and institutions of the financial market), namely *data mining*. The generous topic of *data mining* consists in algorithms providing classifications, estimates, predictions and groupings. The most frequently used techniques for data exploration are: neural networks, decision trees, genetic algorithms, cluster analysis and case reasoning. Most authors argue that the selection of the most appropriate *data mining* technique, for being applied in a particular situation, it's an art, being conditioned by the analyst's experience.

The present study implements the cluster analysis technique, in order to examine the number and structure of Romanian banking groups that share similar features of the risk exposure, profitability and intermediation activity costs.

2. Methodology- An Overview

Cluster analysis is, by excellence, an unsupervised learning technique, that identifies the complex relationships between variables, without imposing any restriction. Consequently, the initial dataset doesn't need the distinct specification of a target variable (the dependent variable) and respectively, of predictor ones (independent variables). All variables have the same importance, because the analysis's goal is not to predict a certain value, but instead, to identify the presence of specific patterns or correlations among variables, to include the different variables or cases into more homogenous groups. Unlike other data mining techniques, we don't have to establish a predetermined set of classes, or to introduce a training stage based on a collection of past data. The entities' clustering is based exclusively on the similarities identified in the variables' structure. Yet, the results obtained are valid only for the ex-ante defined sample; they cannot be generalized to the entire sector/economy. According to Romesburg (2004), this technique represents "*a mathematical microscope for looking at the relations of similarity among a given set of objects. It cannot be used for making statistical inferences about these relations to a larger population. Any inferences a researcher makes by studying the tree are made by using reasoned analogy rather than by using formal statistical methods*".

Cluster analysis focuses on the examination of the interdependencies between variables, its finality consisting in gathering similar entities into more homogenous groups, named *clusters*. Therefore, there must be completed several stages:

- Definition of the analysis' goal, of the assumption to be tested and the selection of the most significant variables;
- Processing numerical values, by applying a standardization procedure. Standardization is imposed when the variables are expressed in different units of measure, in order to lower the risk of misrepresentation of the resemblance relationships between the entities in the sample. Therefore, the variables will become dimensionless. Another advantage of the standardization procedure consists in the uniformization of the variables' influence, by eliminating extreme values, which are susceptible of generating biased results. Failing standardization, if one variable's values range between a large interval than the other ones, then this particular variable will benefit of a greater importance in establishing the similarities between entities, denaturizing the results.
- Selecting a clustering procedure. Economic literature has consecrated three main procedures:
    - K means clustering (non-hierarchical clustering) needs the specification of a pre-established number of clusters. It is recommended when the number of observations exceeds 1000.
    - Hierarchical clustering, which groups the entities into a hierarchical structure.
    - Two step clustering applied mainly for large data sets or for text variables.
- Selecting an appropriate method for data aggregation. The most frequent applied methods are *single linkage* (nearest neighbor, min distance), *complete linkage* (furthest neighbor, max distance) and *centroid clustering*.
- Choosing a unit of measure or an algorithm for the distance/similarity between entities, according to data type (interval, count, binary variables). It is important to mention that, in this case, the distance isn't measured in physical units, but in terms of resemblance between the intrinsic characteristics of the entities considered. One must compute a *resemblance coefficient*, whose meaning can be interpreted in terms of a *similarity coefficient*, or as a *dissimilarity coefficient*. Therefore, the bigger the similarity coefficient' value, the more resembling the two entities. Instead, a high value of the dissimilarity coefficient indicates a low resemblance.
- Interpretation of the dendrogram and identification of optimal number of clusters. The establishment of the correct number of clusters is, however, a subjective process, depending on the decident's experience.

3. Research Premises

The goal of the present study is to identify resembling credit institutions, which can be included into homogenous groups, according to a series of prudential and profitability indicators. Our study aims to provide an alternative to the peer group techniques, implemented by supervisory authorities in the process of off-site surveillance. According to this technique, credit institutions are, firstly, grouped by size or volume of activity, and then, for each group, are made comparative analyses between the current values of financial indicators and the previous ones. The disadvantage stems from the fact that this method cannot signal the impairment in the financial condition of the whole group, but only the distress of a particular credit institution in that group.

Unlike it, cluster analysis, as an exploratory technique, allows the comparisons between all credit institutions in the sample, classifying them into a certain group, according to the similarities identified. The core principle of this technique is that of minimization of the variance between the components of a group, simultaneously with the maximization of the variance across groups. In this way, one can notice the degree of group stability over time.

The study had been conducted for the period between 2004 – 2006 years, and includes data collected on an annual basis from 16 Romanian credit institutions, classified as universal banks. We have excluded from the sample the specialized banks, implied preponderantly in financing the SMEs, the car acquisitions or the building activity, and the subsidiaries of foreign banks. At the end of 2006, the credit institutions included in our sample concentrated a share of assets into total Romanian banking system's assets of 75.37%.

Cluster analysis had been applied for each of the three years considered, with the aim of examining the clusters evolution over time, the measure in which they remain stable. We have computed 8 financial indicators, based on stock data gathered from banks' balance sheet, in order to assess the intermediation activity's main characteristics, in terms of profitability, costs and risk exposure. Below we provide our list of indicators:

- *Capital and reserves to total assets* indicates the degree of a bank's risk aversion. The higher the ratio is, the bigger the credit institution's risk aversion. This ratio statutes the role of banking capital as a main cushion against financial losses.
- *Cash holdings, securities holdings to total assets* measure the liquidity risk. A higher value indicates that the bank is prepared to withstand a suddenly, significant deposit withdrawal.
- *Loans to deposits ratio* is considered to give a clue concerning the occurrence of a credit boom. It also statutes the degree in which internal resources are adequate to cover credit demands, in order to allow the sustainable expansion of the credit activity.

- *Loans to non-financial institutions and households to total assets* indicate bank's exposure to credit risk.
- *Operational expenses to total assets* measure the efficiency in terms of banking activity costs.
- *Return on assets ROA* reflects the net profit brought by a unity of asset.
- *Return on equity ROE* is the most significant expression of the banking profit, from the shareholders' standpoint. It is known that a higher profitability may also imply riskier practices.
- *Profit margin,* computed as net profit to total income. The bigger its value, the more profitable is the bank.
- *Customers' deposits to total liabilities* it's an important indicator because a rise in its value reflects an intensification of the saving process on the domestic market. It is of particular importance in the actual international framework, characterized by uncertainty and financial turbulences, because it indicates a shift from foreign borrowed capital to the domestic one.

By selecting the above mentioned set of variables, our aim was to capture the information incorporated into the main financial ratios and to aggregate them, by means of cluster analysis, so that to obtain more homogenous groups according to their attitude towards risk. Nevertheless, we have excluded from our dataset those variables that proved to be highly correlated with other variables, to avoid biased results. This was the case for ROE, with correlation coefficients exceeding 0, 8. Once we have defined the data matrix, we proceeded to its standardization.

We have chosen a Z scale conversion, known also as normal standardization, determined as (The Current Value of a variable- Average Value)/(Standard Deviation) (1).

As a measure for the distance between credit institutions, we have decided to employ the *squared Euclidean distance* because, in the process of group building, the distinction between them is made according to the characteristics of the outlier banks. The studies of Wolfson (2004), Gutierrez and Sorensen (2006) propose the same approach. The clustering procedure we chose was the *agglomerative hierarchical clustering;* because it allows the grouping of resembling banks, without specify ex-ante a pre-established number of clusters. The agglomerative technique places, firstly, each credit institution into a distinct group, then proceeds to their merger into successively larger clusters, according to the agglomerative method chosen. In this study we have applied, comparatively, three methods:

- *Single linkage* determines the distance between two clusters by the distance of the two closest objects in the different clusters (nearest neighbor).
- *Complete linkage* joins two clusters characterized by the greatest distance between any two objects in the different clusters (furthest neighbor). This method is usually employed when the entities actually form naturally distinctive groups.

- *Centroid clustering* statutes that the distance between two clusters is determined as the difference between their centroids, the centroid being the average point in the multidimensional space of a cluster.

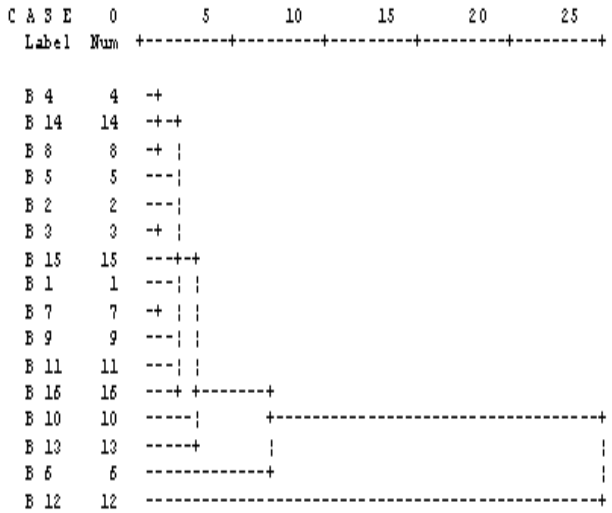
4. Results and Interpretations

As we have previously mentioned, cluster analysis is an exploratory technique which organizes large amounts of observed data into a reduced-size meaningful structure. In order to discover the hidden information in our set of financial indicators, we have applied the clustering technique for each of the three years, taking into account three different methods for computing distance functions.

Although we have formulated our conclusions starting from the single linkage agglomerative method, the other two methods served as a goodness-of-fit test. Table 1 illustrates the output obtained for the year 2006.

At stage 1 appears the first cluster, constituted by credit institutions 4 and 14, because they registered the smallest value of the *squared Euclidean distance coefficient* (0,193). Remember that the *squared Euclidean distance coefficient* is a dissimilarity coefficient. The bigger its value, the more pronounced the discrepancies between the entities analyzed. At stage 2 credit institutions 1 and 7 merge into a new cluster, having a proximity coefficient of 1,126. The clustering algorithm labels each group with the lowest number of the component banks. For instance, the cluster from stage 1 will further be encountered as cluster 4. As one can observe, at stage 3, cluster 4 merges with the bank 8, creating a larger cluster, labeled 4, according to the rule mentioned above. As the value in the coefficients column increases, the distance between groups, expressed as a resemblance measure, increases too. The last credit institution that joins the unique group is 12, with a proximity coefficient of 41,761. Therefore, this particular credit institution is characterized by distinctive banking activity parameters relative to the other banks in the sample. Maybe this is due to the fact that its banking products and services are directed mainly to support the exporter/importer activity. The entire process of mergers is automatically summarized by the dendrogram (hierarchical tree) below (see Graph 1).

Graph 1. Dendrogram using Single Linkage (year 2006)
Rescaled Distance Cluster Combine



At this point of the analysis, we face with two main drawbacks of the clustering algorithm. First of all, the identification of the optimum number of clusters proves to be a difficult, subjective choice. Yet, the dendrogram and the proximity coefficients' value in the agglomerative schedule may help, indicating sudden, large jumps in the level of similarity as more dissimilar banks (groups) are merged. The second drawback derives from the even goal of cluster analysis, that of discovering hidden, latent structures in data, without providing an explanation of their existence or an interpretation.

As we have mentioned before, we have chosen single linkage method because we didn't had any strong a priori expectation concerning the potential number of clusters in the sample, or the presence of some natural groups. Table 2 synthetizes an evolution of groups' components across the 2004-2006 periods, under the assumption of several agglomerative methods.

In order to assess the reliability and validity of the classifications obtained, we repeated the analysis by using each time a different clustering method (single linkage, complete linkage, and centroid), a different distance measure (squared Euclidean distance and Euclidean distance) and a different order of banks in the sample. The results remained unchanged, which means that we can trust their significance and proceed to their interpretation.

In 2004 all three methods identified the same outliers, namely banks 10, 11 and the cluster of banks 12 and 13. This means that, undoubtedly, these banks have distinct patterns concerning their risk profile and profitability. In 2005 the common elements of the three methods were bank 12 and the cluster of 10 and 13. Also, it might be a

certain similarity between credit institutions 7 and 11. 2006 provided a new classification, the outliers being now banks 6 and 12.

Turning to the raw data set, before the application of the standardization procedure, we are able to draw some useful observations. For instance, in 2006, bank 12 registered the biggest value for the loans to deposits ratio (591%) from all the banks in the sample, suggesting an aggressive increase of the credit activity, and implicitly, a higher exposure to credit risk. Also, it indicates that the credit activity relies heavily on borrowed funds from the interbank market, and not on customer's deposits. This finding is emphasized by customers' deposits to total liabilities ratio, whose value is only 1.06%. The ratio of liquid assets to total assets has a value of only 12.63%, being the smallest one. This implies a relatively high exposure to the liquidity risk. Instead, the profitability and operational expenses indicators prove an efficient activity and a good profitability. Consequently, the presence of these specific features can justify the classification of this credit institution as an outlier. The discrimination between the other banks in the sample is much more difficult to be made, because their indicators' values range in a smaller interval and the discrepancies are less obvious. Therefore, this is the appropriate and recommended framework for developing a cluster analysis.

However, this approach isn't able to provide a clear picture on the degree of risk faced by individual banks or by a cluster. Thus, one cannot assess which clusters are riskier than anothers. If we correlate the fact that most credit institutions were gathered in the same cluster with the analyses in the Romanian financial stability report, which statute that the banking system is stable, well capitalized, capable to withstand shocks, then we can affirm that banks from this big cluster are sound, with a moderate exposure to financial risks. All in all, the banking system's main concern consists in managing credit risk.

To conclude with, over the period 2004-2006, the clusters remained relatively stable in terms of similarity of attitude towards risk and profitability. The groups identified are unbalanced, with a big one gathering the high and medium sized banks, and some outliers, represented by small banks, with a market share of 1-1,7%. These small banks are oriented to the retail segment and proved to be very dynamic, especially in 2006. They operate in a flexible, adaptive manner, in order to gain new customers and increase their market share. Yet, cluster analysis doesn't provide a hierarchy of the riskier entities or an explicit reason for their grouping. That's why, in our opinion, the analysis must be extended and completed with several quantitative robust techniques.

5. Conclusions

Cluster analysis, as an exploratory data analysis technique, proves to be valuable not only for assessing homogene banking groups in terms of risk profile and profitability, but also it can identify groups sharing similar features of the financial intermediation activity, large and complex banking groups, as a potential source of

systemic risk (see Financial Stability Review, december 2006), or the degree of financial integration in the euro area banking industry (see Gutierrez, Sorensen 2006).

Reference:

1) Anglim J., 2007, *"Cluster analysis and factor analysis"*, 325-711 Research methods.
2) Chernatony L. 1988, *"Issues to be addressed prior to undertaking hierarchical cluster analysis"*, Cranfield School of Management, School Working Paper 25/88.
3) Costa C.A. 1998, *"Banking Strategies in Portugal-A Cluster Analysis Approach to the Portuguese Banking Activity between 1988-1997"*, Universidade do Minho, Escola de Economia e Gestao.
4) Financial Stability Review 2006, Special features: *"Identifying large and complex banking groups for financial system stability assessment"*, ECB December 2006.
5) Goulet M. and D. Wishart, 1996,*"Classifying a bank's customers to improve their financial services"*, Conference of the Classification Society of North America CSNA, University of Massachusetts, USA, June 1996.
6) Gutierrez J.M.P and C.K. Sorensen, 2006, *"Euro area banking sector integration using hierarchical cluster analysis techniques"*, ECB Working Paper Series no. 627.
7) Romesburg C., 2004, *"Cluster analysis for researchers"* (Lulu Press).
8) Tan P.N., M. Steinbach and V. Kumar, 2006, *"Introduction to Data Mining"*, chapter 8: Cluster Analysis: basic concepts and algorithms (Addison-Wesley).
9) Thalassinos E., Kyriazidis Th., Thalassinos J., 2006, *"The Greek Capital Market: Caught in Between Corporate Governance and Market Inefficiency"*, European Research Studies Journal, Vol. IX, issue 1-2.
10) Wolfson M., M. Zagros and P. James, 2004, *"Identifying national types: a cluster analysis of politics, economics and conflict"*, Journal of Peace Research, vol 41, no. 5, pp.607-623.
11) http://www.statsoft.com/textbook/stcluan.html *"Cluster analysis"*.

Table 1. Agglomeration Schedule (year 2006)

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	4	14	.193	0	0	3
2	1	7	1.126	0	0	5
3	4	8	1.501	1	0	6
4	2	3	1.529	0	0	7
5	1	9	2.001	2	0	10
6	4	5	2.024	3	0	7
7	2	4	2.139	4	6	8
8	2	15	2.805	7	0	11
9	11	16	2.837	0	0	10
10	1	11	3.403	5	9	11
11	1	2	3.406	10	8	12
12	1	10	4.390	11	0	13
13	1	13	4.582	12	0	14
14	1	6	11.228	13	0	15
15	1	12	41.761	14	0	0

Table 2. Group's evolution over time

2004	**Single linkage**	**Complete linkage**	**Centroid clustering**
	2,3,4,1,5,6,16,7, 14,9,8,15	2,3,4,1,7,9,5,6, 16,14	2,3,4,1,7,5,6,16,14,9,8,1 5
	10	8,15	10
	12,13	12,13	12,13
	11	10	11
		11	
2005	**Single linkage**	**Complete linkage**	**Centroid clustering**
	14,16,4	14,16,4,6,8	14,16,4,6,8,1,2,3,5,9,15
	6,8,7,11,1,2,3,5,9,15	10,13	7,11
	10,13	7,11	10,13
	12	1,2,3,5,9,15	12
		12	
2006	**Single linkage**	**Complete linkage**	**Centroid clustering**
	4,14,8,5,2,3,15,1,7,9, 11,16,10,13	4,14,5,8,1,7,9,13,2, 3,11, 16	4,14,8,5,2,3,15,1,7,9,11, 16, 13
	6	10,15	10
	12	6	6
		12	12