DEPARTAMENTO DE ECONOMIA
PUC-RIO

TEXTO PARA DISCUSSÃO
Nº. 412

The application of clustering analysis to international private indebtedness

André Monteiro D'Almeida Monteiro [1, 3]
andremonteiro@icatu.com.br

Dionísio Dias Carneiro [2]
dionisio@econ.puc-rio.br

Carlos Eduardo Pedreira [3]
pedreira@ele.puc-rio.br

Dezembro 1999

[1] Research Unit, Banco Icatu, Rio de Janeiro, Brazil;
[2] Economics Department, Catholic University of Rio de Janeiro;
[3] Electrical Engineering Department, Catholic University of Rio de Janeiro;

Resumo: O principal objetivo deste texto é aplicar uma combinação de métodos estatísticos não-paramétricos tradicionais e redes neurais para examinar, através de segmentação, a dívida privada em diferentes países. São examinadois trinta e nove países. A relação entre dívida privada e algumas variáveis macroeconômicas são discutidas com mais detalhes. O desempenho da segmentação é melhorado pelas vantagens de propriedades específicas de cada um dos métodos. Os procedimentos são também aplicados em um exemplo numérico controlado.

Abstract: The main goal of this paper is to apply a combination of statistical and connectionist schemes to examine, via clustering analysis, private indebtedness in different countries. Thirty-nine such experiences are used. The relationship between private debts and some macroeconomic variables are discussed in some detail. The clustering performance is improved by taking advantage of specific properties and capacities of each method. The procedures are also applied to a controlled numerical example.

JEL: C 450

1. INTRODUCTION

The Nineties witnessed an explosion of international flows to Emerging Markets. Yet, the wide variety of macroeconomic performances of different countries turned out to defy usual classifications based on the evaluation of fiscal deficits and government debt. Besides the difficulties associated with comparing different measures of fiscal variables, direct causality between fiscal deficits and overall macroeconomic performance is uncertain. Dispersion of private indebtedness and its apparent lack of direct association with usual measures of macroeconomic performance (e.g. economic growth, per capita income, external deficit and inflation) suggests it might be useful as an additional variable in the medium run classification of the macroeconomic performance of countries**.** The fact that Asian countries exhibiting high levels of private debt and relatively low fiscal deficits was at the center stage of the financial turmoil experienced at the end of the Nineties adds interest to international comparisons regarding private debt.

In this paper, a combination of Kohonen's Self-Organizing Feature Map [2][3] with statistical schemes is proposed to investigate, via clustering analysis, international experiences with the private debt problem. The main goal is to explore the relationship between a selection of thirty-nine countries' private debts and some macroeconomic variables.

One of the main targets of segmentation procedures is to obtain well-defined and compact groups. Although statistical clustering schemes [1] are, in general, quite sensitive to initial conditions, they have been successfully applied to some segmentation problems. From the connectionist viewpoint, Kohonen's neural networks [2][3][4] produce selective tuned units to create a topographic map of the input patterns. Although this neural network paradigm was not originally proposed for pattern classification or clustering applications, it is possible to take advantage of its self-organization properties. The overall performance is improved by taking advantage of the specific properties and capacities of each of the procedures.

In Section 2 we present some considerations on the segmentation tools used. Section 3 is dedicated to a segmentation-controlled numerical-experiment. An international private-debts segmentation analysis is presented in Section 4, followed by our concluding remarks.

2. SOME CLUSTERING PROCEDURES AND COMBINATIONS

One of the most common strategies applied to solve segmentation problems is to enforce a cost function based on the minimization of the clusters' sum dispersions. It is in general useful to associate to each group a prototype localized at its center. Kohonen's Self- Organizing Feature Map (SOFM) prototypes (or weight vectors) converge to a set of weight vectors that do not, in general, satisfy this property [5]. Moreover, the minimization of the clusters' dispersions is not one of the explicit objectives of this algorithm. On the other hand, its capacity to provide a preliminary approximation of the data probability distribution function can be very useful when dealing with clustering problems.

Kohonen's Self-Organizing Feature Map possesses two quite useful properties [3]. First, it has the capability to approximate the input patterns' probability density-function. A greater number of neurons are positioned to take care of higher probability density-regions. Second, the mapping preserves the topology of input patterns, in same sense.

Success in segmentation problems is in general associated with obtaining well-defined groups. In this sense an important stage in clustering algorithms should be related to the minimization of the dispersion of the groups.

The main goal of the procedure is to take advantage of the SOFM's capacity to provide a preliminary approximation of the data probability distribution function, and subsequently minimize the average intra-cluster dispersion. Essentially, the idea is to cluster the data in two stages: First run the SOFM, and subsequently minimize the average intra-cluster dispersion. Since the goal is just to get a selected initial approximation, a fine-tuning of the SOFM convergence is not pursued. Two methods are used to reduce dispersion: the K-means algorithm [1] and Global Dispersion Minimization, GDM, a heuristic developed by the authors in sub-section 2.2.

We will compare the procedures' performance against SOFM and K-means in isolation. We point out that the comparison with the Kohonen algorithm is made in full awareness that this method was not originally proposed for classification purposes.

We define dispersion as the average of each cluster's standard deviation weighted by the number of elements in each group. With this linear weighting, a stronger weight is ascribed to denser groups.

2.1. A Brief Description of Kohonen's Self-Organizing Feature Map (SOFM)

The main goal of Kohonen's Self-Organizing Feature Map [2][3] is to gather input patterns in grid-distributed units, or neurons. The prototype candidates, or weights, accomplish communication between this output grid and the input patterns' sub-space. Each unit, or neuron, is associated to a weight, or prototype candidate.

The SOFM algorithm may be divided into four basic stages: (i) Computation of the distances between a randomly chosen input pattern and the prototypes; (ii) Choice of the closest prototype (to the given pattern); (iii) Activation and adaptation of the winner neuron (the one closest to the given pattern) and its neighborhood; (iv) Gradual decrease of the size of the affected neighborhood.

Two metrics are needed: one for the input space, d, and another for the grid, d*. A neighborhood V, centered at neuron j, with radius R, may be defined as:

$$V = V(R) = \{C_j \text{ such that } d^*(C_j, C_i) \leq R_v\}$$

The set of neurons V(R=r) is said to be the r-th neighborhood. Although different metrics may generate different neighborhoods, there is strong experimental evidence that the final map is not affected by the choice of the metric d*.

A Neighborhood Function centered at the winner neuron, $\Im_w$, is responsible to fire this neuron and all neurons belonging to its neighborhood. It also commands an approximation policy of these neurons to the presented pattern. The winner neuron's weight vector (also called prototype) is updated with a unity factor, while that of each

neighbor is updated with decreasing factors proportional to their distance from the winner, measured by d*.

An important parameter of the neighborhood function is the radius R_a. It establishes to what distance from the winner neuron updating is performed. These radius parameters vary, in general, with the iterations of the algorithm. The Neighborhood Function, at iteration t, centered at a winning neuron v, can be written as:

$$\Im_w = \Im_w (X(t), R_a(t))$$

Let us gather the prototypes (or weight vectors) in a matrix U, by placing the prototype associated to neuron j in column j. The training procedure can now be described by the following four basic steps:

S1. Randomly choose an input X_k, such that $X(t)=X_k$

S2. Find the winner neuron v such that:

$v = \arg_j \min d(X(t), U_j(t))$, j = 1,2, ... J;

S3. Update the prototype matrix U by:

$$U(t+1) = U(t) + \gamma(t).\Im_w (X_k, R_a(t)).[X_k - U_v(t)]$$

where $\gamma(t)$ is the learning rate in t;

S4. Stop the algorithm when no significant changes in the prototypes can be detected.

2.2 A Heuristic to Reduce Dispersion

In this sub-section a description of the proposed dispersion-minimization algorithm is given. We define Local and Global Loopings. The Local Looping has the objective of finding the "best" position estimator for one group in accordance to a standard deviation criteria. In this looping some prototype candidates are generated, while the position estimators for all the other groups are kept clamped. The Global Looping comprises one run of the Local Looping for each of the clusters. In Global iterations, the prototype candidates identified as the best for each of the groups are installed to consolidate the process. The algorithm comprises the following steps:

S1: Produce clusters using as initial prototype candidates the SOFM weights, by allocating each point of the data set to a group corresponding to the smallest distance between this data point and the prototype candidate group;

S2: Choose one of these groups;

S3: Calculate a position estimator (e.g. mean, median), and set this as a new prototype candidate;

S4: Reallocate all the data points, by using the nearest prototype candidate criteria (as in Step 1);

S5: Calculate the standard deviation for the chosen group. Store this result to check stability in Step 7.

S6: Calculate the segmentation dispersion. Store this result for future comparison (Step 8).

S7: If Local Stabilization was not reached, return to Step 3. Local Stabilization means that the standard deviation calculated in Step 5 is the same for two consecutive iterations.

S8: Choose the prototype candidate corresponding to the smallest segmentation dispersion calculated in Step 6.

S9: Return to Step 2 without changing any of the prototype candidates. Restart this step until all of the groups have been visited.

S10: Choose new prototype candidates for all groups using the segmentation dispersion-minimization criterion (Step 6) until Global Stabilization is reached. Global Stabilization means that, for two consecutive global runs, there has been no change to any of the prototype candidates calculated in Step 7.

Note that the Local Looping phase can be processed in parallel mode since the order in which the groups are chosen does not affect the final result. The use of combinations of more than one position estimator (median, mean, trimmed mean, etc.) can be directly implemented in the Local Looping. The possibility of employing position estimators other than the mean may be useful to capture the specific characteristics of some clusters. For example, the center of a cluster with fat-tails distribution (compared to Gaussians) may be more efficiently estimated by a trimmed mean than by the standard mean.

Although this algorithm and the K-means retain some similarities, Step 8 constitutes a major difference: the prototype substitution is conditional and associated with a reduction of total dispersion.

3. SEGMENTATION: A CONTROLLED NUMERICAL EXPERIMENT

In this section we present numerical simulations with synthetic data. The main purpose is to show, in a controlled mode, the potential of combined procedures - SOFM in addition to GDM; and SOFM in addition to K-means - versus the pure application of SOFM or K-means. The clusters generated by SOFM that initialized the two procedures are the same. We reiterate that the comparison with SOFM was made in full awareness that this method was not originally proposed for segmentation purposes.

A numerical experiment was designed by generating 6 synthetic clusters - a total of 4075 points. Each cluster has its points generated through a uniform distribution inside a polygon. The size and the densities vary from one polygon to another. A graphic representation of these artificial clusters can be found in Figure 1.

The results of these experiments can be found in the Figures 2-5. In these Figures, the solid lines are decision boundaries between the clusters (divisions, in fact) generated by each of the three methods. The metric used was the Euclidean distance. The SOFM associated with GDM (Figure 2) recognizes all six groups. The SOFM plus K-means procedure (Figure 3) was unable to distinguish between two groups and perceived one segment as two different groups. The SOFM algorithm (Figure 4) performed quite poorly. The dispersions were 0.0476, 0.0702 and 0.0894, respectively. Consequently, the K-means achieved a 21% gain compared with SOFM in terms of dispersion; and the GDM, a 47% gain.

FIGURE 1
The synthetic clusters

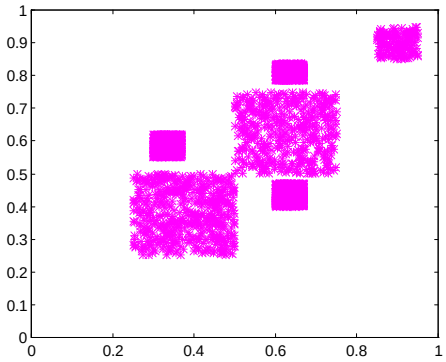


FIGURE 2
SOFM plus GDM segmentation

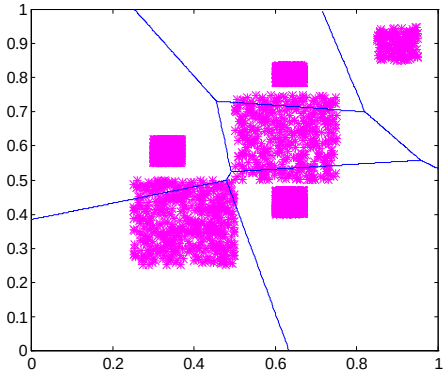


FIGURE 3
SOFM plus K-means segmentation

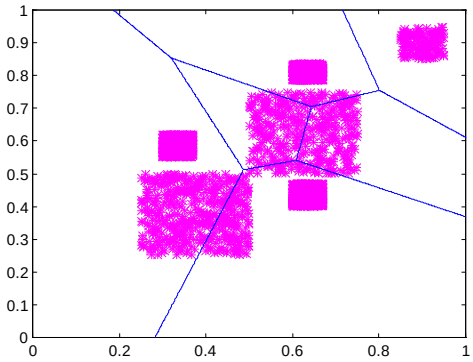
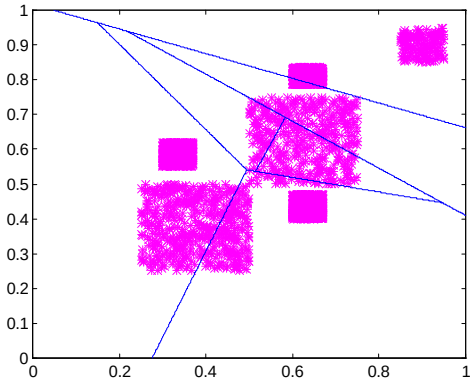
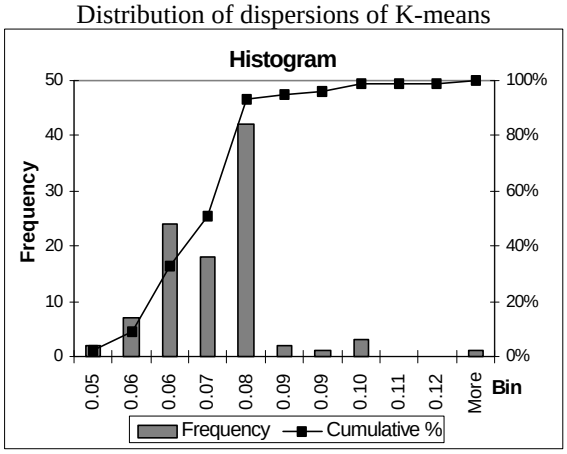


FIGURE 4
SOFM in full segmentation



The comparison between K-means and the proposed procedures has to be made using the former *ex-post* output probabilities. Targeting this goal, the K-means algorithm was run independently one hundred times for this data. Fifty percent of the total segmentations presented dispersion below SOFM plus K-means. In comparison with the SOFM associated with GDM procedure, 9% generated exactly the same segmentation and 91% provided a worse performance.

FIGURE 5
Distribution of dispersions of K-means



Note that both procedures, the SOFM plus GDM and SOFM plus K-means, performed better than either K-means or SOFM alone. This result seems to indicate the potential of combining, in sequence, preliminary approximation of the data probability distribution function provided by SOFM with minimization of segmentation dispersion.

4. CLUSTERING OF INTERNATIONAL PRIVATE-DEBTS ANALYSIS

The objective of this section is to analyze the relationship between private debts, measured as a ratio of GDP, and overall macroeconomic performance. The macroeconomic performance is defined by the following variables[1]: Inflation Rate, Current Account Surplus as a percentage of GDP, Per Capita Income and Economic Growth Rate. The variables were expressed in standardized annual means for 1991 to 1995.

Attention must be called to two aspects relevant in the investigation of the performance of macroeconomic variables related to private-debts. First, this problem is intrinsically multivariate, so its visualization is non-trivial. Second, the countries' details and idiosyncrasies such as specific characteristics of the financial sector, nature of private debt, existence of public guarantees and so forth would render global analysis less appealing, should the diversity of each one of the countries be scrutinized in isolation. Thus, segmentation seems to be a proper procedure in order to search for readily available indicators that might add information to a pre-defined set of measures. In this present case, a successful segmentation means grouping countries that exhibit economic characteristics in common that make sense in the context of macroeconomic experience. Note that the goal is neither to generate a function to forecast values for the private debts as a percentage of GDP based upon four variables, nor to decompose the multivariate variance into components.

Thirty-nine countries were selected from a total of 160 available in the IFS. The criterion to select the countries was based on their relevance in the world scenario and the quality of the data. The chosen countries are: Argentina, Australia, Austria, Belgium, Bolivia, Canada, Chile, China, Colombia, Denmark, Egypt, Finland, France, Germany,

[1] Source: International Financial Statistics database, IMF- March 1997.

Greece, Holland, Hungary, India, Indonesia, Israel, Italy, Japan, Malaysia, Morocco, Mexico, Norway, Paraguay, Peru, Portugal, Singapore, South Korea, Spain, Switzerland, Thailand, Turkey, UK, Uruguay USA, and Venezuela.

The distribution of countries by continent is as follows (in parentheses we show the number of available countries in the IFS): Asia, 9 (38); Europe, 16 (53); North America, 2 (2); Latin America, 9 (17); Africa, 2 (53); Middle East 1 (17). Eighteen (out of 39) are developed countries. [2]

The descriptive statistics of the five variables are shown in Tables 1 and 2. In the first, we have the individual statistics, and in the second, the correlation coefficients.

TABLE 1
Variables (without standardization) Individual Descriptive Statistics

Statistic \ Variables	**Debt/ GDP**	**Inflation Rate (%)**	***Per capita* Income (US$)**	**Growth Rate (%)**	**CAS/ GDP (%)**
Mean	0.649	9.4	11,005	2.7	-1.0
Median	0.539	4.3	7,067	2.1	-1.2
Standard deviation	0.433	12.6	9,977	2.4	3.9
Kurtosis	1.95	7.81	-1.22	0.60	3.03
Skewness	1.22	2.66	0.50	0.99	1.00
Range	1.978	60.7	32,194	10.5	21.7
Minimum	0.109	0.7	384	-0.4	-9.0
Maximum	2.087	61.4	32,578	10.1	12.7

TABLE 2
Correlation Coefficients

Variables	**Debt/ GDP**	**Inflation Rate**	***Per capita* Income**	**Growth Rate**	**CAS/ GDP**
Debt/GDP	1				
Inflation Rate	-0.51	1			
***Per capita* Income**	0.57	-0.51	1		
Growth Rate	0.04	0.04	-0.41	1	
CAS/GDP	0.34	-0.16	0.51	-0.04	1
R^{2*}	*0.46*	*0.36*	*0.66*	*0.33*	*0.32*

* R^2 from equation in which a variable is explained by the others.

As can be noted from Table 2, there are no highly correlated variable pairs. The highest correlation coefficient involves Debt/GDP and Per Capita Income; the lowest, Debt/GDP and Inflation Rate. The Economic growth Rate is only correlated with Per Capita Income. Except for Per Capita Income, all R^2 are small. Therefore, there is no reason to exclude any variable due to linear dependence with other variables.

The methodology described in the previous section is now applied to investigate the international private debt problem. Considering the number of countries in focus, preliminary data analysis shows that four groups are sufficient to represent the data structure. The organization of the groups (G1 to G4) corresponds to the unidimensional

[2] In fact, this Section is part of a research project on private debts covering data since 1981. The East European country data for early the Eighties were not available in the IFS, so they do not appear in this work. Brazil does not appear because of its huge inflation rate during this period.

distances: G1 is closer (or more similar) to G2 than to G3. The SOFM segmentation provided the following result:

G1: China, Colombia, Egypt, Peru, Turkey, Uruguay, and Venezuela;
G2: Argentina, Bolivia, Chile, Greece, Hungary, India, Indonesia, Israel, Malaysia, Morocco, Mexico, Paraguay, Thailand;
G3: Australia, Canada, Italy, Korea, Portugal and Spain;
G4: Austria, Belgium, Denmark, Finland, France, Germany, Holland, Japan, Norway, USA, UK, Singapore, and Switzerland.

TABLE 3
Number and means of SOFM segmentation groups

Groups	No. Countries	Debt/ GDP	Inflation Rate	*Per capita* Income	Growth Rate	CAS/ GDP
G1	7	-0.76	1.60	-0.93	0.53	0.11
G2	13	-0.57	-0.01	-0.78	0.15	-0.72
G3	6	0.15	-0.48	0.29	-0.27	-0.28
G4	13	0.78	-0.59	1.16	-0.42	0.84

All the variables discriminated G1+G2 from G3+G4, except Current Account as a Percentage of GDP. The first two groups present the lowest Debt/GDP ratio and Per Capita Income, and the highest economic growth Rate. The most pronounced division line is the Per Capita Income: in G1+G2, the only country with above average Per Capita Income was Israel. In G3+G4, only Portugal and Korea have below average Per Capita Income. It seems that G1+G2 and G3+G4 show a good division between developed and underdeveloped countries. Inflation is the main factor separating G1 from G2 and Per Capita Income is the best discrimination variable for G3 and G4.

The following result was obtained from the association of the SOFM and K-means methods:

C1: Peru, Turkey, Uruguay, and Venezuela;
C2: Argentina, Bolivia, Chile, China, Colombia, Egypt, Greece, Hungary, India, Indonesia, Israel, Korea, Malaysia, Morocco, Mexico, Paraguay, and Thailand;
C3: Australia, Austria, Canada, Denmark, Finland, France, Germany, Italy, Portugal, Spain, UEA and UK;
C4: Belgium, Holland, Japan, Norway, Singapore and Switzerland.

TABLE 4
Number and means of SOFM plus K-means segmentation groups

Groups	No. countries	Debt/ GDP	Inflation Rate	*Per capita* Income	Growth Rate	CAS / GDP
C1	4	-1.08	2.56	-0.87	0.01	-0.17
C2	17	-0.33	0.00	-0.81	0.48	-0.51
C3	12	0.27	-0.55	0.81	-0.54	-0.04
C4	6	1.12	-0.61	1.26	-0.28	1.64

In general, this second segmentation preserves the differences between the first and the last two groups in a very similar to those observed in the former grouping. The only country that moved from G3+G4 to C1+C2 was Korea. This movement increases the discrimination capacities of the Per Capita Income and Economic growth Rate. The C1 Inflation characterization is very bold: indeed, it represents the highest inflation in the period. However, Groups G3 and G4 were strongly modified such that the discriminating variable for C3 and C4 is now Current Account Surplus as a percentage of GDP: all six counties have the highest values of the sample (not considering Egypt)

Finally, the only difference between the SOFM-plus-K-means and SOFM-plus-GDM procedures appeared in relation to Denmark, which moved from C3 to C4. This can be understood as a consequence of its high Current Account surplus as a percentage of GDP.

The increased fragility of Korea in the early Nineties and strengthening of Denmark, were both captured by the SOFM plus GDM while the application of SOFM plus K-means captured the differences between Korea and OECD countries. None of the procedures were able to distinguish Singapore from the OECD countries, suggesting that private indebtedness was not a fundamental characteristic of the Singapore economy, perhaps due to its role as a financial center.

In terms of total dispersion, which may be considered a good measure of segmentation quality, both procedures improved the SOFM segmentation quality. SOFM-plus-K-means and SOFM-plus-GDM produced an improvement of 9.3 % and 10%, respectively.

It is important to note that one has to consider not only the improvement of about 10% in the dispersion measure but also the correct economic mean of the change promoted by the proposed procedures. Randomly initialized K-means was tried ten times independently. No run was able to produce a reasonable segmentation from the economic and statistical viewpoint.

Since the best segmentation results were generated by the combination of SOFM and GDM, we restricted the final analysis to these groups. Table 5 shows the means of the variables in each segment generated by SOFM-plus-GDM.

TABLE 5
Number and means of SOFM plus GDM segmentation groups

Groups	No. countries	Debt/ GDP	Inflation Rate	*Per capita* Income	Growth Rate	CAS/ GDP
C1	4	-1.08	2.56	-0.87	0.01	-0.17
C2	17	-0.33	0.00	-0.81	0.48	-0.51
C3	11	0.35	-0.54	0.74	-0.56	-0.12
C4	7	0.87	-0.61	1.29	-0.30	1.52

The presence of the consistent groups from the economic and statistical viewpoint indicates the existence in the early Nineties of similarities in terms of countries' experiences with private debts and usual macroeconomic indicators. Besides, all the applied statistical tools were able to identify those similarities among the countries, despite the existence of some differences in terms of segmentation. Furthermore, the presence of a regional component can be noticed in the groups. It is interesting to observe that this information was not supplied explicitly.

Table 2 indicates the level of linear association between Debt/GDP and the other four variables. There is association between Debt/GDP and Inflation Rate (negative) and Debt/GDP and Per Capita Income (positive). The linear association involving Debt/GDP and CAS/GDP is not clear. However, by comparing the groups' means, one may conclude that the focus of two big groups, C1+C2 and C3+C4, enables the identification of a relationship between Debt/GDP and Economic growth Rate. It is worth noting that, in this case, the correlation coefficient did not provide any indication. Based on the groups' means, one can observe that the countries for which the Debt/GDP variable are greater (smaller) than their group means, have an Economic growth Rate above (below) the sample mean.

5. FINAL REMARKS

In this paper, we analyze the private indebtedness as another option of measure for macroeconomic performance of countries in the medium run by using clustering procedures. The procedures are based on a two-step-strategy: preliminary approximation of the data probability distribution function provided by Kohone's SOFM, and minimization of segmentation dispersion.

Through the bidimensional controlled-experiment, we could illustrate the potentials of the procedures. The segmentation generated by the procedures outperformed both K-means and SOFM in isolation. The SOFM plus GDM performed better than SOFM plus K-means.

We applied segmentation techniques with the objective of identifying similarities in international experiences, in the early Nineties, concerning private debts, and well-accepted parameters to measure macroeconomic performance such as: Inflation Rate, Per Capita Income, Current Account as a percentage of GDP and Economic Growth Rate.

The procedures were able to allocate the 39 countries in four statistically and economically consistent groups. The results suggest that the debt to private sector as a

percentage of GDP variable can be used as another macroeconomic performance measure, although it is by no means clear from the macroeconomic viewpoint what separates quality from the size of private debt. The segmentation outcome provides enough sensitivity to capture the differences in private indebtedness indicators, as part of a possible grouping of countries according to sovereign risk, suggesting that adding this measure as part of the usual set of simplified performance-variables may enhance the evaluation of macroeconomic risk.

REFERENCES

Dillon, W.R. and M. Goldstein. "Multivariate Analysis", John Wiley & Sons, 1984.

Likhovidov, V. "Variational Approach to Unsupervised Learning Algorithms of Neural Networks", Neural Networks, vol. 10, no. 2, 1997, pp. 273-289.

Kaski, S. and T. Kohonen. "Exploratory Data Analysis by the Self-Organizing Map: Structures of Welfare and Poverty in the World", i*n*: *Neural Networks in the Capital Markets*, World Scientific, 1996.

Kohonen, T. "Self-Organized Formation of Topologically Correct Feature Maps", Biological Cybernetics 43, 1982, pp. 59-69.

Kohonen, T. "Self-Organizing Maps", Springer Verlag, 1995.