Learning and Nonlinear Models – Revista da Sociedade Brasileira de Redes Neurais, Vol. 1, No. 4, pp. 264-277, 2004 .

A New Clustering Procedure Applied to an International Comparison of Indebtedness

André d'Almeida Monteiro
Risk Department
Gavea Investments
amonteiro@gaveainvest.com.br

Dionísio Dias Carneiro
Department of Economics and IAPUC
PUC-Rio
dionisio@econ.puc-rio.br

Carlos Eduardo Pedreira
Electrical Engineering Department
PUC-Rio
pedreira@ele.puc-rio.br

Abstract

This paper presents a procedure for clustering analysis that combines Kohone's Self-organizing Feature Map (SOFM) and statistical schemes. The idea is to cluster the data in two stages: run SOFM and then minimize the segmentation dispersion. The advantages of proposed procedure will be illustrated through a synthetic experiment and a real macroeconomic problem. The procedure is then used to explore the relationship between private indebtedness and some macroeconomic variables commonly used to measure macroeconomic performance. The experiences of thirty-nine countries in the early nineties are analyzed. The procedure outperformed others clustering techniques in the job of identifying consistent groups of countries from the economic and statistical viewpoints. It found out similarities in different countries concerning their respective levels of private indebtedness when added to well-accepted parameters to measure macroeconomic performance.

Keywords: Vector quantization, Clustering, Self-Organizing Feature Map, Macroeconomic Performance, Private Indebtedness.

1. INTRODUCTION

Vector quantization (VQ) has been extensively explored from theoretical as well as applied exercises , in particular to approach clustering problems. [3] and [8] are classical reviews and, more recently, [1] is an excellent overview of the broad family of learning algorithms based on Stochastic Gradient Descent. Some largely used algorithms may be considered in a VQ framework, among those, the family of the K-means [4] and Kohonen's Self-organizing Feature Map (SOFM) [6], [7].

The goal of this paper is to propose a procedure for clustering analysis that combines SOFM and statistical schemes. One of the main targets of segmentation procedures is to obtain well-defined and compact groups. Although some statistical clustering schemes [4] are sensitive to initial conditions, they have been successfully applied to some clustering (or segmentation) problems. SOFM [5], [6], [7] produces selective tuned units to create a topographic map of the *n*-dimension input patterns on a lower dimension grid and to approximate the input patterns'

probability density function. However, in general, the prototypes produced by SOFM do not coincide with clusters' centers. Therefore, to bring each SOFM's prototype to the center of its associated cluster tend to reduce the overall segmentation dispersion.

The proposed procedure clusters the data in two stages: run SOFM and then minimize the segmentation dispersion. The method used to reduce the segmentation dispersion starting from SOFM's clusters determines a variation of the procedure. Two methods are used: the K-means algorithm [5] and Global Dispersion Minimization (GDM), a heuristic proposed ahead. Therefore, two variations of the procedure are been proposed: SOFM-plus-K-means and SOFM-plus-GDM.

The advantages of proposed procedure will be illustrated through a bidimensional controlled-numerical experiment and a real macroeconomic problem.

The new procedure will be applied to analyze the relationship between private indebtedness and some macroeconomic variables commonly used to measure macroeconomic performance. Segmentation seems to be a proper tool to approach this problem because it is intrinsically multivariate and the countries' details and idiosyncrasies would render global analysis less appealing.

The motivation for this analysis is that 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. Simple indicators of the building-up of macroeconomic difficulties that may arise as a result of excessive accumulation of debt by the private sector are also difficult to define, because of both the 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). But as an indicator of financial fragility of the firms and households, which may precipitate sudden falls of capital investment outlays or consumption expenditures, it has been used as an additional variable in the medium run classification of the macroeconomic performance of countries**.** In the Asian crises of the second half of the nineties, the high levels of private debt side by side with relatively low fiscal deficits were part of the picture and this may be seen as one major flaws in the analysis of the Washington-based multilateral institutions (IMF and IBRD) which led to a general underestimation of the risks of defaults. Recovery following that financial turmoil has given rise to an new interest in international comparisons regarding private debt.

The procedure is used to do an exploratory investigation on the size of private debt because high values of this variable may be associated with several phenomena, both structural, such as the efficiency of the financial system, and cyclical such as experiences of growth spurts, short-lived or not. In some instances, therefore, the depth of private indebtedness may be seen as a favorable characteristic of the economy such as an environment of high confidence and structural improvement, as is the case when institutional improvements make loans more attractive. In other instances, high debt may be a sign of financial fragility, as firms or households' balance sheets may prove to be excessively sensitive to changes in macroeconomic environment, as has been the case when sudden stops in external finance requires drastic changes in the macroeconomic policy stance.

The paper is organized as follows: in Section 2 we present the segmentation tools used; Section 3 is dedicated to a controlled-numerical experiment which helps to understand the outcome of the procedures and helps to interpret the results for the international private-debt segmentation analysis which is presented in Section 4, followed by concluding remarks.

2. THE NEW CLUSTERING PROCEDURE

One of the most common strategies applied to solve segmentation problems is to enforce a cost function based on the minimization of the clusters' dispersions. It is in general useful to associate to each group a prototype located at its center. The SOFM's prototypes converge to a set of weight vectors that do not, in general, satisfy this property [5]. To be sure, the minimization of the clusters' dispersions is not one of the explicit objectives of this algorithm. However, SOFM exhibits two quite useful properties when dealing with clustering problems [7]. First, it has the capability to approximate the input patterns' probability density function: the denser is the probability region, the greater is the number of neurons positioned to take care of it. Secondly, the mapping preserves the topology of input patterns, in same sense.

The idea of the proposed procedure is to cluster the data in two stages. First run SOFM to provide a preliminary clustering of the data, taking advantage of the SOFM's properties. And subsequently minimize the segmentation 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 [5] and Global Dispersion Minimization, GDM, a heuristic procedure which is proposed in sub-section 2.2. Thus, two variations of the procedure are been proposed: SOFM-plus-K-means and SOFM-plus-GDM.

2.1. A BRIEF DESCRIPTION OF SOFM

The main goal of SOFM [6], [7] is to gather input patterns in grid-distributed units, the so-called Kohonen's map. Each unit, or neuron, is associated to a prototype. The prototypes accomplish communication between this output grid, C, and the input patterns space, X. Two metrics are required: one for the input space, d, and another for the grid, d^*, thus, defining two metric spaces: (X, d) and (C, d^*).

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 input pattern (the winner neuron or prototype); (iii) Activation and approximation of the winner prototype and its neighborhood to the input pattern; (iv) Gradual decrease of the size of the affected neighborhood.

A neighborhood V on the grid, centered at neuron j, with radius r, may be defined as:

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

The set of neurons $V(r = n)$ is said to be the *n-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^*, [6], [7].

The approximation of the prototypes to the presented input pattern is controlled by a neighborhood function, $\Im$. It is a smoothing kernel defined over the grid which depends on the distance between the prototypes. It has to fulfill some requirements in order for the algorithm to converge [7].

$$\Im_{i,j} = \Im_{i,j}\left(d^*\left(C_i, C_j\right)\right)$$

Indeed, the neighborhood function centered at the winner neuron is the one which defines the approximation policy. The winner neuron's prototype is updated with a unity factor, while those of each neighbor are 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. It establishes to what distance from the winner neuron updating is performed. The radius r decreases with the algorithm iterations, reducing the size of the neighborhood affected by the chosen input pattern. The neighborhood function, at iteration t, centered at a winning neuron w, can be written as:

$$\Im_{w,j}(t) = \Im_{w,i}\left(d^*\left(C_w, C_j\right); r(t)\right)$$

Let us gather the prototypes 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 pattern X_k, such that $X(t) = X_k$

S2. Find the winner neuron w such that:

$$w = \arg\min_j d\left(X(t), U_j(t)\right), j = 1,2, \ldots J;$$

S3. Update the prototype matrix U by:

$$U_j(t+1) = U_j(t) + \gamma(t)\Im_{w,j}(t)\left[X_k - U_w(t)\right], j = 1,2, \ldots J;$$

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

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

The M input patterns are allocated in the J neurons according to the following competitive rule:

$$C_j = C(U_j) = \left\{ X_k \in X : j = \arg\min_i d\left(X_k, U_i\right) \right\}, \; j = 1,2, \ldots J; \qquad (1)$$

If $M >> J$, the SOFM is performing clustering analysis. In this case, each neuron is a cluster (or group or segment) and its prototype can be seen as a representative of the input patterns allocated in it.

Under certain assumptions, the SOFM Mean-Risk Functional can be written as [5]:

$$R(U) = \sum_{i=1}^{J} \int_{C_i(U)} \left[\sum_{j=1}^{J} \Im_{ji} . d\left(X_k, U_j\right) \right] p(X) dX$$

It is showed [5] that, under some conditions, the sequence $R(t) = R(U(t))$ goes to a local minimum R^* with probability 1 and, if the limit of the sequence $U(t)$ exists, each prototype satisfies the following condition:

$$U_j^* = \frac{\sum_{j=1}^{J} \Im_{i,j} \mu_j(U^*)}{\sum_{j=1}^{J} \Im_{i,j} p_j(U^*)}, \qquad (2)$$

where $\mu_j(U^*) \equiv \int_{C_j(U^*)} Xp(X)dX$ is the mean pattern of cluster j and $p_j(U^*) \equiv \int_{C_j(U^*)} p(X)dX$ is the associated probability. Both are generated by U^* according to (1).

Equation (2) establishes the interdependence relationship among the prototypes caused by the neighborhood function: the lower is the number of input patterns in the cluster and the more on the grid center it is, the greater is the attraction of other clusters' prototypes on its prototype. The attraction forces may bring a prototype away from center of its cluster. Thus, the cluster dispersion may end up being higher because of the allocation of input patterns far way from its center. And finally, the segmentation dispersion may be negatively impacted.

Two statistical schemes are then used to bring the prototypes to the center of their cluster and, consequently, to reduce the segmentation dispersion: the K-means algorithm [5] and Global Dispersion Minimization, GDM, a heuristic proposed ahead.

Fixed the number of clusters J, the k-means algorithm can be briefly described in four steps:

S1: Choose arbitrary J cluster representatives;

S2: Produce clusters by allocating each input pattern to a group according to (1), where U is the matrix containing the clusters representatives;

S3: Update each cluster representative as the mean of the input patterns allocated in the associated cluster produced in S2.

S4: Keep running S2 and S3 until no change in the cluster representatives occurs.

2.2 A HEURISTIC TO REDUCE DISPERSION

Before presenting the heuristic, it is necessary to formally define cluster dispersion and segmentation dispersion. The dispersion of the cluster C_j is defined as the square root of the trace of the covariance matrix computed over the input patterns belonging to C_j:

$$D_j = \sqrt{tr(\Sigma_j)} \qquad (3)$$

Let m_j be the number of input patterns allocated in the cluster C_j. The segmentation dispersion, D, is as the average of each cluster's dispersion weighted by the proportion of number of elements in each group. A stronger weight is ascribed to denser groups.

$$D = \sum_{j=1}^{J} \left(\frac{m_j}{M} \right) D_j \qquad (4)$$

The heuristic is composed by Local and Global loops. The Local loop has the objective of finding the "best" position estimator for one group in accordance to the segmentation dispersion.

In this loop some prototype candidates are generated, while the position estimators for all the other groups are kept clamped. The Global loop comprises one run of the Local loop 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's prototypes, by allocating each input pattern to a group according to (1);
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 input patterns, by using (1);
S5: Calculate the dispersion 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 dispersion 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 loop 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 loop. 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 segmentation dispersion.

3. 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 two variations of the procedure, SOFM-plus-GDM and SOFM-plus-K-means, versus the pure application of SOFM or K-means. The clusters generated by SOFM that initialized the procedure are the same.

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 SOFM-plus-K-means achieved a 21% gain compared with SOFM in terms of dispersion; and the SOFM-plus-GDM, a 47% gain.

FIGURE 1
The synthetic clusters

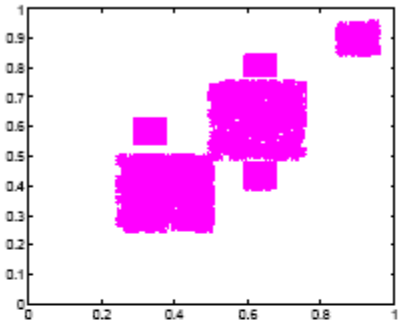


FIGURE 3
SOFM plus K-means segmentation

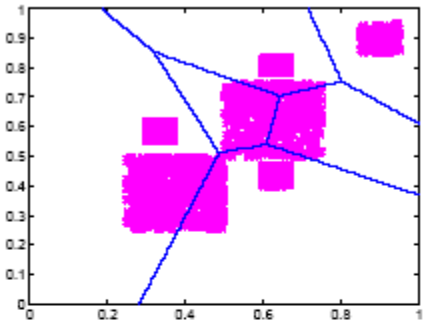


FIGURE 2
SOFM plus GDM segmentation

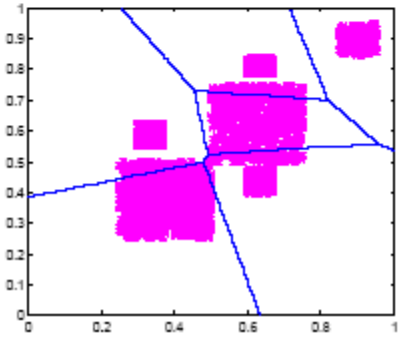
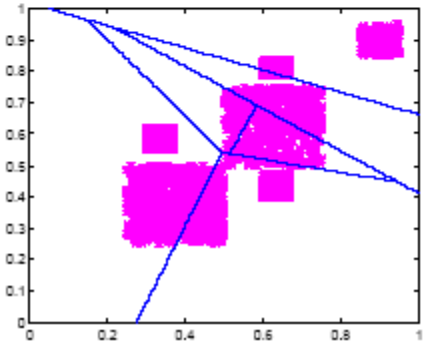
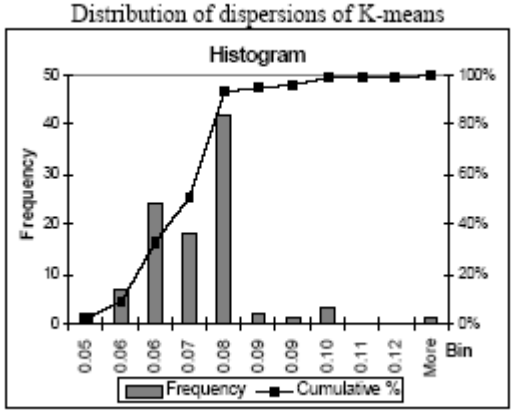


FIGURE 4
SOFM in full segmentation



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 density function provided by SOFM with minimization of segmentation dispersion.

4. CLUSTERING ANALYSIS OF INTERNATIONAL PRIVATE-DEBTS

This section uses the clustering devices described in the previous ones to analyze how private debt, measured as a ratio of GDP, may be related to overall macroeconomic performance. The macroeconomic performance is measured by the following variables: Inflation Rate, Current Account Surplus as a percentage of GDP, Per Capita Income and Economic Growth Rate.[1] The variables were expressed in standardized annual means for 1991 to 1995. Each input pattern is a vector containing the above 5 variables of a country.

Two aspects are relevant when we probe into the relationship between the performance variables and private debt. 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 a search for readily available indicators that might add information to a pre-defined set of macroeconomic measures. In the present case, a successful segmentation means grouping countries that exhibit common economic characteristics 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,

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

Colombia, Denmark, Egypt, Finland, France, Germany, Greece, Holland, Hungary, India, Indonesia, Israel, Italy, Japan, Malaysia, Morocco, Mexico, Norway, Paraguay, Peru, Portugal, Singapore, South Korea, Spain, Switzerland, Thailand, Turkey, UK, Uruguay, US, and Venezuela. Brazil does not appear in the list because of its huge inflation rate during this period.[2]

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.

The descriptive statistics of the five variables are shown in Tables below. Table 1 presents the individual statistics, and Table 2 the correlation coefficients between the selected variables in the sample.

TABLE 1
Variables Individual Descriptive Statistics (without standardization)

Statistic	**Debt/ GDP (%)**	**Inflation Rate (%)**	***Per capita* Income (US$)**	**Growth Rate (%)**	**CAS/ GDP (%)**
Mean	64.9	9.4	11,005	2.7	-1.0
Median	53.9	4.3	7,067	2.1	-1.2
Standard deviation	43.3	12.6	9,977	2.4	3.9
Kurtosis	195	7.81	-1.22	0.60	3.03
Skewness	122	2.66	0.50	0.99	1.00
Range	197.8	60.7	32,194	10.5	21.7
Minimum	10.9	0.7	384	-0.4	-9.0
Maximum	208.7	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
Multivariate correlation Coeficient *	*0.46*	*0.36*	*0.66*	*0.33*	*0.32*

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

[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.

As can be seen in Table 2, there are no highly correlated variable pairs. The strongest positive correlation is between Debt/GDP and Per Capita Income (0.57) and the strongest negative correlation is between Debt/GDP and Inflation Rate (-0.51). Economic Growth is correlated in the sample only with Per Capita Income. Except for Per Capita Income, all R^2 are small. We have, therefore, 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 whether private debt may be meaningfully added to the macroeconomic performance variables as part of a criterion to segment the countries. 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 (S1 to S4) corresponds to the one-dimensional distances: S1 is closer (or more similar) to S2 than to S3. The SOFM segmentation provided the following result:

S1: China, Colombia, Egypt, Peru, Turkey, Uruguay, and Venezuela;
S2: Argentina, Bolivia, Chile, Greece, Hungary, India, Indonesia, Israel, Malaysia, Morocco, Mexico, Paraguay, Thailand;
S3: Australia, Canada, Italy, Korea, Portugal and Spain;
S4: Austria, Belgium, Denmark, Finland, France, Germany, Holland, Japan, Norway, US, UK, Singapore, and Switzerland.

TABLE 3
Number and means of SOFM segmentation groups

Groups	# of countries	Debt/ GDP	Inflation Rate	*Per capita* Income	Growth Rate	CAS/ GDP
S1	4	-1.08	2.56	-0.87	0.01	-0.17
S2	17	-0.33	0.00	-0.81	0.48	-0.51
S3	11	0.35	-0.54	0.74	-0.56	-0.12
S4	7	0.87	-0.61	1.29	-0.30	1.52

All the variables discriminated S1+S2 from S3+S4, 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 S1+S2, the only country with above average Per Capita Income was Israel. In S3+S4, only Portugal and Korea have below average Per Capita Income. It seems that S1+S2 and S3+S4 show a good division between developed and underdeveloped countries. Inflation is the main factor separating S1 from S2 and Per Capita Income is the best discrimination variable for S3 and S4.

The following result was obtained from SOFM-plus-K-means procedure:

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, US and UK;
C4: Belgium, Holland, Japan, Norway, Singapore and Switzerland.

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

Groups	# of 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	0.35	-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 S3+S4 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 S3 and S4 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 clusters and SOFM-plus-GDM clusters (G1 to G4) appeared in relation to Denmark, which moved from C3 to G4. 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 the 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, 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 economic meaning of the changes in classification which resulted from 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 viewpoints.

Since the best segmentation results were generated by SOFM-plus-GDM, we restricted the final analysis to these groups. Tables 5 and 6 show, respectively, the means and the standardized means of the variables in each segment.

TABLE 5
Number and means of SOFM-plus-GDM segmentation groups

Groups	# of countries	Debt/ GDP	Inflation Rate	*Per capita* Income	Growth Rate	CAS/ GDP
G1	4	-1.08	2.56	-0.87	0.01	-0.17
G2	17	-0.33	0.00	-0.81	0.48	-0.51
G3	11	0.35	-0.54	0.74	-0.56	-0.12
G4	7	0.87	-0.61	1.29	-0.30	1.52

TABLE 6
Number and means of SOFM-plus-GDM segmentation groups (without standardization)

Groups	# of countries	Debt/ GDP (%)	Inflation Rate (%)	*Per capita* Income (US$)	Growth Rate (%)	CAS/ GDP (%)
G1	4	18.1	41.7	2,325	2.7	-1.7
G2	17	50.6	9.4	2,924	3.9	-3.0
G3	11	80.1	2.6	18,388	1.4	-1.5
G4	7	102.6	1.7	23,875	2.0	4.9

The presence of the consistent groups from the economic and statistical viewpoints indicates the existence of similarities in terms of countries' experiences with private debts and usual macroeconomic indicators in the early nineties have been identified by all the applied statistical tools, and some interesting differences have been captured by the segmentation devices. Furthermore, the presence of a regional component can be noticed in the groups. It is interesting to observe that this information was not supplied.

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, G1+G2 and G3+G4, 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 proposed a procedure for clustering analysis that clusters the data in two stages: run SOFM and then minimize the segmentation dispersion. The method used determines a variation of the procedure. Two methods were used to reduce the segmentation dispersion starting from SOFM's clusters: the K-means algorithm and Global Dispersion Minimization (GDM), a

heuristic proposed in the paper. Therefore, two variations of the procedure were tested: SOFM-plus-K-means and SOFM-plus-GDM.

After describing the nature of the procedure, we illustrated its performance by means of a bidimensional controlled-experiment. It was found that in the controlled experiment, 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.

The segmentation techniques were then applied to the identification of similarities in different countries in the early nineties, concerning their respective levels of private indebtedness when added to well-accepted parameters to measure macroeconomic performance. Both proposed procedures were able to allocate the 39 countries in four statistically and economically consistent groups, outperforming SOFM and K-means in isolation. SOFM-plus-GDM provided a slightly better segmentation than SOFM-plus-K-means.

The results suggest that the debt to private sector as a percentage of GDP variable can be useful as an additional macroeconomic performance measure, when we examine the fragility of the different economies to changes in the international financial environment, even if we disregard important considerations such as the quality of private debt, the source of finance and so on. The segmentation which resulted from the exercise exhibited enough sensitivity to capture the differences in private indebtedness indicators, and suggests that a possible grouping of countries according to sovereign risk, may be obtained from the use of this method to enhance the usual set of performance variables.

Research in progress by the authors suggest the results described in this paper provide powerful instruments to classify macroeconomic risk, independent of the observed fiscal and monetary variables which occupy a prominent role in the check list of risk graders. This may be illustrated by Table 7. There, we present the number of defaults of private international debt between 1996 and 2000 for the different groups defined by SOFM plus GDM procedure. Consider, for example, G2 cluster of countries. Even if we exclude the USA, (for the obvious reason that the number of defaults in the US economy would unduly overstress our point), the group concentrates 75% of the observed defaults of private external debt, in spite of the fact that macroeconomic data per se seem quite reasonable in the context of the whole sample. It has the second smallest debt/GDP ratio (50.6%), inflation close to the sample mean (9.4%) the higher average rate of growth (3.9%) and the highest current account deficit (3% of GDP). The group includes several of the countries that occupied the center stage of the international turbulence in the second half of the nineties.

.

TABLE 7
Number of private debt defaults during 1996-2000 (without US defaults)
per SOFM-plus-GDM segmentation groups

Groups	# of countries	# of defaults	# of defaults per country
G1	4	0	0
G2	17	113	7
G3	11	29	3
G4	7	9	1

Source: Moody's

REFERENCES

[1] Bottou, L. "Stochastic Learning", Lecture Notes In Artificial Intelligence, 3176, PP. 146-168, 2004.

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

[3] Gersho, A. "Asymptotically Optimal Block Quantization", IEEE Transaction on Information Theory, Vol. 25, pp. 373-380, 1979.

[4] Jonson, R. A. and Wichern, D. W. "Applied Multivariable Statistical Analysis", Prentice Hall, 1998.

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

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

[7] Kohonen, T. "Self-Organizing Maps", Third Edition, Springer Verlag, 2001.

[8] Zador, P. **"**Asymptotic Quantization Error of Continuous Signals and the Quantization Dimension", IEEE Transaction on Information Theory, Vol. 28, pp. 139- 149, 1982.