A typology of very small companies using self-organizing maps*

Sylvain Barthélémy
Senior Economist, T-A-C
35140 Saint Hilaire des Landes, France
barth@tac-financial.com

Jean-Baptiste Filippi
UMR CNRS 6134, University of Corsica
20250 Corte, France
filippi@univ-corse.fr

Abstract – *The world of very small companies (VSC) is characterized by a great diversity in terms of quality and behavior. Consequently, the principles for economic evaluation used today on medium to big sized companies cannot directly be used on VSC that must be grouped with companies of similar behavior. The purpose of the present study, conducted for the French Ministry of Finance, was to find an original technique to group these companies into clusters, with an economic theory background. Self Organizing Maps (SOM) were used to create subsets of companies from the original database. The method has been implemented in Java software with a web front-end to perform the computer evaluation of the technique. The method is validated on a data set of 459 companies collected by French students. Using the clustering software we found four subsets, that correspond to the economic theory of 'Worlds of production' : interpersonal, industrial, immaterial and merchant world.*

Keywords: Tracking, filtering, estimation, information fusion, resource management.

1 Introduction

The world of very small company (VSC) is not constituted by one homogeneous group, but is characterized by a very strong diversity of individuals. The common evaluation principles of use for small to medium companies cannot be directly transposed to the VSC. This article presents an original way to identify a better typology to understand the world of very small companies, in order to facilitate their evaluation and their access to credit.

For a long time, economists and bankers have known that a single profile is not sufficient to compare and analyze the entire population of companies of a country. That is why they usually use classification criteria to divide the population into different subsets. These classifications or typologies of companies are traditionally set up using simple criteria of size and sector of activity. Even if it is widely accepted that these traditional methods of classification are not sufficient to identify homogeneous groups of comparable companies, economists and bankers use it because no other classification method is available today. We started to work on this typology in 2001, when the French Ministry of Finance (DECAS) decided to launch a study on the creation of a tool to allow the evaluation of the 'wealth' of very small companies. A tool that would be based on an economic theory and on a new typology of very small companies. This article is part of the research contract made for the french Ministry of Economy and Finance in 2001 (see [1]).

When we started to set up our very small companies evaluation tool, we chose an economic theory and launched a survey with questions tailored to help us identifying a typology. It was important to be sure to explain the groups as an economist would do, because the objective of the evaluation tool is to help companies to evaluate themselves. Our choice has been made on the well recognized economic theory of the 'worlds of production', from [7]. To integrate qualitative and quantitative criteria to create a typology of companies, the authors proposed the notion of 'worlds of production' where companies are differentiated using two main variables : the nature of production (generic or dedicated) and the process of production (specialized or standardized). The four 'worlds of productions' are identified from those two axes: the interpersonal world, the industrial world, the immaterial world and the merchant world.

We also needed a clustering method that would be able to 'detect' a typology and which permits, afterwards, to find the right group for an evaluated company. All these, without any discriminative initial criterion. We can distinguish various objectives for such tool:

- The classification process should not be supervised, as it would be using discriminant analysis or simple neural networks(such as multilayered perceptron). A supervised classification algorithm would use a set of pre-classified VCS (based on an expert reasoning) to learn about the combination of criteria for each group. It would introduce a bias into the classification process as we want to find the criteria

*0-7803-7952-7/03/$17.00

and the group simultaneously.

- Quantitative data are not easy to obtain on small companies, the method of classification must distinguish groups by using a set of qualitative and quantitative criteria which will be extracted from an existing database or from data specifically collected for the study.
- The number of groups should be sufficiently small in order to avoid splitting the data set into too many groups. Too many groups will make the analysis of a single group a difficult task as groups will not represent a great number of companies.
- The tool should be sufficiently robust and flexible to overcome a relatively bad quality of input data (qualitative data collected from a set of companies is often imprecise) and to use incomplete data sets.

From this set of objectives we chose to use Self Organizing Maps (SOM) to perform this typology of very small companies. SOM are a special kind of neural networks, also called Kohonen Nets [4] and they allow to represent multidimensional data in a smaller number of dimensions (usually two), to find clusters in the resulted map and to analyze the input data.

2 Self-Organizing Maps

Self-organizing maps (SOM) [3], also called Kohonen neural nets [4], have the ability to take into account properties of spatial or temporal continuity that enable a non-linear projection of a multidimensional space into a space of reduce dimensions. This kind of network relies onto multidirectional propagation dynamics with high interaction between neurons of the same neighborhood. SOMs are often used in data-mining because they allow to map multidimensional data on a bi-dimensional grid (although the output map can have as many dimension as needed, most studies use only two dimensions). SOMs also allow to reveal more or less homogeneous groups in a data set and are a convenient technique to solve problems with spatial or temporal dimensions.

SOMs present four main advantages for our problem: they are visually explicit, easily understandable, allow an unsupervised classification based on the sample density function and are not highly sensible to the quality of input data. Moreover, traditional hierarchical methods of classification have a greater failure rate when they are used on data sets that are not optimal (with compact and isolated groups) and [6] have shown the superiority of SOMs on empirical data for data sets that contained structural imperfections.

2.1 Using Self Organizing Maps

A SOM is a regular grid, usually bi-dimensional, of cartographical units. Each unit i is represented by a weight vector $w_i = [w_{i1}, ..., w_{in}]$ where n is equal to the dimension of the input vector. The units are connected together with a neighborhood relation, usually using on a rectangular or hexagonal basis. Data that are "alike" are represented by units that are "near". SOM is also a technique for projection of data in a bi-dimensional grid.

A SOM is trained using an iterative procedure. At each iteration, an input vector x is usually chosen by stochastic methods. Every distance between the x vector and the weight vector is then calculated. Then the algorithm searches the unit c the nearest from x, also called *best-matching unit* (BMU) using (1) or (2).

$$||x - w_c|| = min_i\{||x - w_i||\} \tag{1}$$

or

$$c = argmin_i\{||x - w_i||\} \tag{2}$$

The operator $||.||$ gives the Euclidian distance between two points. It is important to note that SOMs can be used on data sets that contain null values, distances are calculated only on samples that are available.

Once the BMU c is found, all vectors that are in the neighborhood of the BMU must be updated using the formula (3):

$$w_i(t+1) = w_i(t) + \alpha(t)h_{c,i}(t)[x(t) - w_i(t)] \tag{3}$$

where t represents time, $\alpha(t)$ the learning rate and $h_{c,i}(t)$ (4) is the neighborhood centered on the BMU c.

$$h_{c,i} = \exp\left(-\frac{||r_c - r_i||^2}{2\sigma^2(t)}\right) \tag{4}$$

where r_c et r_i are the positions of the units c and i on the grid. $\alpha(t)$ and $\sigma^2(t)$ decreasing monotonously over time.

2.2 Performance measurement

For each learning phase of a SOM there is a large number of possible parameters : topologies, neighboring functions, number of steps,... The resulted map is used as measurement to compare trained SOMs together. The mean quantification error is a common measure of SOM precision. If we take the BMU of the input vector x_i by $w_{c,i}$, then the mean quantification error is given by:

$$EMQ = \frac{1}{n}\sum_{i=1}^{n}(||x_i - w_{i,c}||) \tag{5}$$

In addition, the preservation of the data sample typology previously used can be calculated by:

$$ETO = \frac{1}{n}\sum_{i=1}^{n}u(x_i) \tag{6}$$

We write $N_0(a)$ the direct neighborhood of unit a, c_i the BMU of the input vector x_i and d_i the second BMU of the input vector x_i:

$$u(x_i) = \begin{cases} 1 & : \quad d_i \in N_0(c_i) \\ 0 & : \quad d_i \notin N_0(c_i) \end{cases} \tag{7}$$

These two measures are complementary, even in a presence of small cartographical error, the topographical error can be important.

2.3 Visualisation

Many methods of visualization can be used to represent the results obtained from SOMs. The four widely used methods are: Sammon mapping, u-matrices, plane and histograms of data density.

Sammon mapping [8] is mainly used during the learning phase of the SOM. It is an iterative method of multidimensional data representation in two dimensions. Sammon's mapping has been developed before and independently of SOMs. Nevertheless, representing weights vectors in a plane by connecting neighboring neurons by lines allows to obtain an ordered graph reproducing the "shape" of the input data used.

U-matrix, from [9], allows to visualize implicit groups contained in the data sample. The principle is to calculate a matrix of distances between weight vectors and adjacent neurons. Those distances are then represented in a bi-dimensional space. Different shades of grey are usally used to separate neurons that are "near" (light gray) to the neurons that are "far" or "distant" (dark grey).

Planes allow to visualize the relatives values of one of the component of the weight vector in a bi-dimensional map. Like u-matrices, different shades of grey are used in the plane representation.

Histograms of data density represent on a plane the number of units for a given BMU. The number of "hits" per neurons will provide its color, its size or its height (in case of a tri-dimensional representation).

For our classification problem, we tried to find an easy and understandable way to represent the results obtained using the SOM and we used a kind of u-matrix with companies represented as point, over a background on which dark color represents the density.

3 Company classification

Classification of companies has been set up in four successive phases. The first phase was to constitute the company database. In the second phase we conducted three company classification based on different data sets. In the third phase we compared the results with two other well known clustering methods Ward hierarchical analysis and K-Means analysis. Finally, we tried to interpret the resulted groups in terms of 'Worlds of Production'.

3.1 Input data

The test database is constituted of anonymous records extracted from a database selected by the "EM Lyon" (Lyon Management school, France). The sample concerns 459 VSC of less than 50 employees and with a turnover of less than 7 million Euros. As shown in table 1, for each company we have 6 financial variables and 6 qualitative variables.

Variable name	Description
VF1	debt / equity
VF2	EBITDA / sales
VF3	debt / cashflow
VF4	tangible assets / fixed assets
VF5	working capital requirements /
-	(equity + debt)
VF6	cashflow / equity
VQ1	number of competitors (multiple
-	choices field - recoded)
VQ2	market share of the new products
-	(multiple choices)
VQ3	technical innovations less that
-	2 years ago (multiple choices)
VQ4	share of specialized products over
-	standard products (multiple choices)
VQ5	share of sales on the major customers
-	(multiple choices)

Table 1: Description of the database variables.

The so-called 'financial' variables illustrate the behavior and performance of a company according to a standard financial analysis (indebtedness, profits, cashflow,...). The so-called 'qualitative' variables sharpen the evaluation of the VSC. Qualitative variables have a double importance : on one side they allow to get rid of the view of a VSC that behaves like a big company and on the other side they contribute to show that some elements related to financial analysis are missing to assess the company quality and the cluster membership.

3.2 Experiments with Self Organizing Maps

We first tried to use standard self organizing maps algorithms using the whole data set and different number of cells. We obtained the results presented on figure 1. We clearly observed that with more than 10 cells, the distortion measure starts decreasing and the quantization error falls from 13 to 11. We did a lot of experiments and with a number of cells between 9 and 50 and we never obtained more than 6 clusters with more than 20 companies in each. With more than 50 cells, the number of small clusters grows rapidly.

After these experiments, we decided to create a modified SOM algorithm, which seemed to better suit our issue and to be easier for the staff and the employees of very small companies to understand. We created a Java

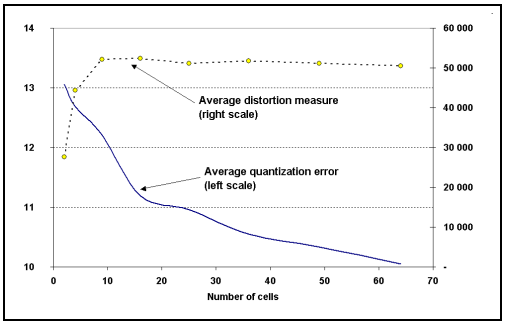

Figure 1: Self Organizing Map experiments with different number of cells.

program that automatically produces the classification and the visualisation map using the following algorithm. At each vector we assign a set of two weights (one for each dimension) and assign to those weights a random value. Our algorithm is iterative. At each iteration, we select a random point on the map and find the nearest neuron (the one that has the nearest weights from the random point). This neuron is considered to be our BMU, then we update the weights for all the neurons using the neighborhood formulas (3) and (4). Instead of a Vorono tessellation, we can consider our algorithm like a cook whose mashed potatoes would go more and more lumpy, where each lump is a group of similar companies.

We performed many experiments of classification by selecting groups of different variables to reveal more or less distinct groups. Each experiment resulted in a map in which companies are organized, a list of each cluster of companies and the most representative company of the cluster. On those maps the euclidian distance expresses the difference (proximity shows similarity).

The different classifications have been conducted on different data sets including the same companies but with different variables : a set with financial data, a set on qualitative data, and a set on all data. From those experiments we have always obtained a small amount of groups (maximum 4). The following Figures (1,2,3) show the graphical results of the clustering. According to the algorithm, similar companies moved together and using this visual representation each dot corresponds to a company and the background corresponds to the companies density (darker for higher density).

The experiment performed using the financial data set is represented in figure 2-1. It does not provide sufficiently dissociated clusters to distinguish clearly groups of companies. A single cluster is revealed (in arrow shape). This result confirms the weaknesses of a clustering by using only financial data. The experiment performed using the qualitative data set clearly shows three clusters (see figure 2-2). Those three groups show that clustering on qualitative data is possible, even though the amount of variables is too small to keep the resulted

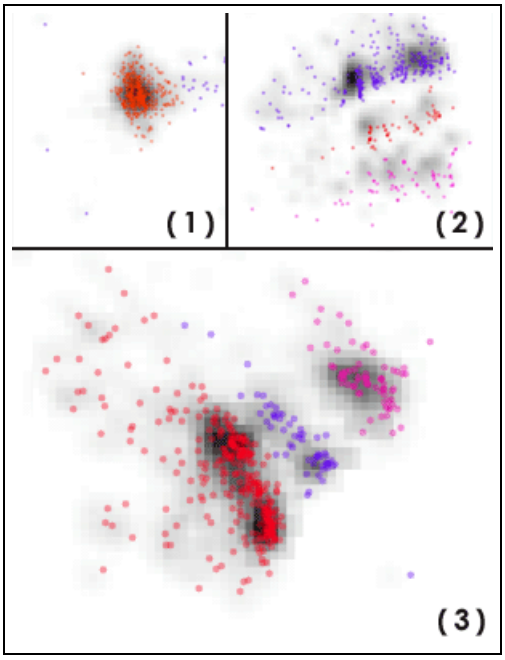

Figure 2: Map of companies after SOM clustering. (1) on financial data, (2) on qualitative data, (3) on all data.

clusters as final ones. We then proceeded to a classification on the whole database, including the 11 financial and qualitative variables (see figure 2-3). The resulted map clearly reveals three groups and a fourth group of companies that seems more isolated but with a position on the map that reflects similar behavior.

On the three tests, only the first one gave poor results. Using the qualitative and the overall data set we obtained more that one cluster. We then performed a comparison of the results with other methods of clustering.

3.3 Comparing SOM with other classification methods

To complement the study made with SOMs, we used two other well known clustering methods : the standard Ward hierarchical method of classification and the K-means method of clustering. We did those tests to compare the results we obtained with more traditional methods in order to verify that the groups that have been formed are not inconsistent. We did not try to obtain identical groups but we simply analyzed the correspondances and checked that the constituted groups were not uniformly distributed among groups constituted by the use of other methods.

For the hierarchical analysis we used the Ward methodology [10] which is well known for its efficiency even if this method usually creates groups of relatively small size. This method permits to obtain spherical and compact groups of companies using variance to evaluate

the distance between "clusters" (Euclidian distance). For the classification of companies using the K-Means algorithm we used the algorithm from [2]. The K-Means method is a non-hierarchial one, introduced by [5], that minimizes the inner-group variance. The principle of this method is to choose K points of the individual space that will serve of reference location to the future classes.

The clustering of the sample of 459 companies, made by the SOMs, gives four groups, for a total of 423 companies classified. We associated a group to each company found with SOMs and compared the results to the results obtained with hierarchical and k-means clustering. The result is presented in table 2.

	G1	G2	G3	G4	Total
G1	3/**14**	51/**5**	0/**3**	2/**34**	56
G2	36/**147**	14/**3**	106/**39**	133/**100**	289
G3	23/**2**	1/**23**	0/**2**	3/**0**	27
G4	1/**17**	46/**1**	4/**1**	0/**32**	51
Total	63/**180**	112/**32**	110/**45**	138/**166**	423

Table 2: Comparaison of groups created by the SOM (lines), hierarchical Ward analysis (columns) and k-means algorithm (columns, **bold**)

Hierarchical clustering confirms groups 1 and 3 found by the SOMs and in general fits at 75% with the results obtained with SOMs. The major difference is that the Ward analysis identifies two different groups into G2 and that it merges G1 and G4 into a single group. With the K-Means algorithm, the results obtained are less clear but we observe that, like in the Ward analysis, G3 is clearly identified but the K-Means algorithm divides G1, G2 and G4 into two groups.

The comparison of the results obtained between SOM, Ward and K-means shows that the typology depends on the method used, but with results not fundamentally different. Further investigation should be done to find a better way to combine these different algorithms of classification. Nevertheless in our case, SOM are efficient, visually explicit, flexible and less sensible to data imperfection (which is not always the case with other clustering methods).

3.4 Economical analysis of the results

Even though we have a relatively large amount of companies in the data sample that we used for the study (459 companies), the number of variables (only 12) reduced the possibilities of variables selection to perform the classification. Nevertheless the economic analysis of the results presented in table 3, shows the efficiency of SOMs in respect to the notion of "worlds of production" [7].

After some economic investigations about the similarities between 'worlds of production' and our groups, we identified the following correspondances: group 1

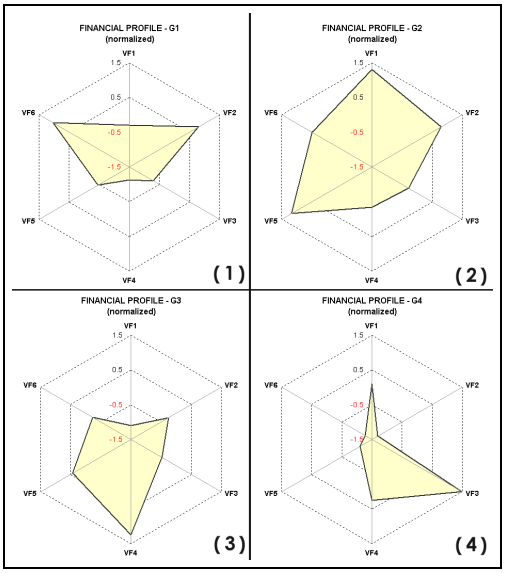

Figure 3: Web charts of the normalized financial profiles of the four identified clusters: (1) Immaterial World, (2) Industrial World, (3) Interpersonal World, (4) Merchant World.

matches the immaterial world, group 2 the industrial world, group 3 the interpersonal world and group 4 the merchant world. The web charts of these groups on the variables used for the classification are presented in figure 3 and the values of the qualitative variables in table 3.

Group	VQ1	VQ2	VQ3	VQ4	VQ5
1	5 to 10	$< 20\%$	2	tailored	15%
2	5 to 10	$< 20\%$	0	equivalent	25%
3	5 to 10	$> 80\%$	0	equivalent	20%
4	5 to 10	$< 20\%$	1	tailored	15%

Table 3: Qualitative profiles of groups identified by SOM

- The first group, with 56 companies, can be considered the immaterial world. It presents a trend for major innovations in terms of production processes of the main activity, mainly tailored fabrication, diversified customers, a relatively high part of fixed investments (22%) and a high auto-financing capacity that corresponds to the idea of a financing that favors the immaterial more than the own ressources. The relative weakness of product innovation (less than 20%) is explained by the small size of the companies that have been considered which rarely have technical (laboratories, ...), human (researchers, engineers,...) and financial ressources.
- The second group, with 289 companies, can be assimilated to the industrial world. It is character-

ized by a high corporate investment level which in general is explained by the generalization of subcontracting based on highly functional plans, high debt which corresponds to the investment effort required from those companies and a good economical performance that can be explained by the scale factor. Unsurprisingly this is also the group that exhibits the weakest product innovation.

- The third group, with 27 companies, corresponds to the interpersonal world which plans its products for a high number of customers while producing for various classes. It is characterized by a high intensity of product innovation (probably incremental), a broad range of customers (less than 20% of income generated with the main clients) and dominating corporate investments. We observe a poor financial profitability and an average economic profitability which can be explained by the competition of bigger units (supermarkets as opposed to small shops).
- The fourth group, with 51 companies, corresponds to the merchant world where dedicated products are exchanged on a standardized market. Innovation has a small effect, tailored production corresponds to the adaptation of the products where the needs of customers are the highest and the main clients are not the major source of income.

The results obtained by the economic analysis of the financial and qualitative profiles of the groups, and even though the input database is relatively small, show that the results are coherent with the economic theory of the 'worlds of productions'. We obtain 4 groups of very small companies, with very different economic profiles but each of these can be assimilated to a 'world of production'.

4 Conclusion

For this study we created a Java software to perform an unsupervised classification methodology, based on self-organizing map, for very small companies. We also created a complete company evaluation system with a Web front-end architecture, on which companies could authentify, fill-out a questionnaire and obtain an evaluation. Our results are very promising and confirm the efficiency and robustness of clustering methods of very small companies using quantitative and qualitative data with a self organizing maps algorithm. Moreover the results have been interpreted easily and match economic theories, which was one of our main goals. We are currently trying out the technique on a specialized database (with more companies and more data) that will allow to experiment the method at a higher scale and obtain more robust results. We also work on some projects of a larger implementation of the online evaluation software for very small companies in France.

References

[1] T. Apoteker, S. Barthélémy, M. Delhom, J.B. Filippi, N. Levratto, L. Mahérault, V. Revest, D. Rivaud-Danset, J.F. Santucci, "L'évaluation des entreprises afin de faciliter l'accès au crédit: quelle intermédiation informationnelle", performed by TAC under the supervision of N. Levratto (CNRS) for the french Ministry of Economy and Finance - Secrétariat d'Etat aux PME - Direction des Entreprises du Commerce, de l'Artisanat et des Services, May 2001.

[2] J. A. Hartigan and M. A. Wong, "A k-means clustering algorithm", *Applied Statistics*, Vol 28, pp. 100–108, 1979.

[3] S. Kaski, "Data Exploration Using Self-Organizing Maps", *Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series*, No 82, Mar. 1997.

[4] T. Kohonen, *Self-organizing maps*, Springer, New-York, 1995.

[5] J. MacQueen, "Some methods for classification and analysis of multivariate observations", Proc. Fifth Berkeley symposium on mathematical statistics and probability, 1967.

[6] P. Mangiameli and S. K. Chen and D. West, "A Comparison of SOM Neural Network and Hierarchichal Clustering methods", *E*uropean Journal of Operation Research, Vol. 93, No. 2, Sept. 1996.

[7] R. Salais and M. Storper, *Les Mondes de Production*, Ed. de l'Ecole des Hautes Etudes en Sciences Sociales, 1993.

[8] J. W. Sammon, "A nonlinear mapping for data structure analysis", *IEEE Transactions on Computers*, IEEE Trans Computer, vol. C-18, pp. 401–409, May. 1969.

[9] A. Ultsch and H. P. Siemon, "Kohonen's self organizing feature maps for explanatory data analysis", *Proc. Int Neural Network Conf*, Dordrecht, The Netherlands, pp. 305–308, 1990.

[10] J.H. Ward, *Hierarchical grouping to optimize an objective function*, Springer-Verlag, Berlin, 1963.