



HAL
open science

The evaluation of the impact of the technology transfers from public laboratories to private firms: the case of the french nuclear authority (CEA)

Vincent Vigneron, Serge Petit

► To cite this version:

Vincent Vigneron, Serge Petit. The evaluation of the impact of the technology transfers from public laboratories to private firms: the case of the french nuclear authority (CEA). 2008. hal-00202719

HAL Id: hal-00202719

<https://hal.science/hal-00202719>

Preprint submitted on 29 Jan 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

**THE EVALUATION OF THE IMPACT OF THE TECHNOLOGY TRANSFERS
FROM PUBLIC LABORATORIES TO PRIVATE FIRMS:
THE CASE OF THE FRENCH NUCLEAR AUTHORITY (CEA)**

VINCENT VIGNERON^{1,2}, SERGE PETIT³

¹Equipe MATISSE-SAMOS CES CNRS-UMR 8173
Université Paris 1
75634 Paris cedex 13, France
{[vigneron.cottrell](mailto:vigneron.cottrell@univ-paris1.fr)}@univ-paris1.fr

²IBISC CNRS FRE 2494
Université d'Evry
91020 Courcouronnes, France
vincent.vigneron@ibisc.univ-evry.fr

³Infineon Technologies AG, CPS EB BI,
St. Martin Straße 53
81669 München, Germany
serge.petit@infineon.com

THE EVALUATION OF THE IMPACT OF THE TECHNOLOGY TRANSFERS FROM PUBLIC LABORATORIES TO PRIVATE FIRMS: THE CASE OF THE FRENCH NUCLEAR AUTHORITY (CEA)

VINCENT VIGNERON^{1,2}, SERGE PETIT³

¹Equipe MATISSE-SAMOS CES CNRS-UMR 8173
Université Paris 1
75634 Paris cedex 13, France
{[vigneron.cottrell](mailto:vigneron.cottrell@univ-paris1.fr)}@univ-paris1.fr

²IBISC CNRS FRE 2494
Université d'Evry
91020 Courcouronnes, France
vincent.vigneron@ibisc.univ-evry.fr

³Infineon Technologies AG, CPS EB BI,
St. Martin Straße 53
81669 München, Germany
serge.petit@infineon.com

Abstract. Research laboratories are increasingly coming under pressure to evaluate the economic impact of their research. After the multiplication of technology discrimination devices bound to foster the innovative efforts in firms, public policy makers are seeking to evaluate the impact of these means on the economy.

Existing evaluation indicators are poorly equipped to grasp the very characteristics of innovative activities. A set of impact indicators has been developed on the basis of an empirical analysis of twenty-two collaborations. The links with some determinants such as the profile of the firm, the characteristics of the project and the type of innovation have been investigated. Since intrinsic structure of the data is not at all linear, our data analysis extensively uses Kohonen maps which are encountered to realize a rapid and robust classification and to visualize the classes, their differences and homogeneity. Policy maker and R&D (Research and Development) program administrators should be especially sensitive to those impacts that depend on manageable determinants.

Key words: innovative technology, Kohonen maps, correspondence Analysis.

1 Theoretical background

The theoretical argument that underlies this paper is the acknowledgement that technology cannot be reduced to a piece of formalized information, but that technology is made of three constituents that can exist independently of each other. Each one can impact at different levels in a firm. The “artifact” dimension is the aspect that evaluation is generally most concerned about, as improvements in production techniques are perceived as a condition *sine qua non* in the ongoing search for industrial competitiveness. But the other dimension, namely skills and knowledge, are often left out although technological development draws heavily upon them. In fact, disembodied knowledge spills over more easily (to other activities) than knowledge embodied in use-specific products. This legitimates the idea that the impact of technology is likely to trespass the scope of traditional evaluation. Add on top of that the possibilities of technological recombinations, and then technology ends up as a potentially pervasive phenomenon that is improperly captured by the prevailing indications. The paper suggests complementary evaluation indicators.

2 The empirical research

2.1 Data collection and methodology

It is clear that this renewed conception of evaluation cannot be thought independently of new methods of *data collection* and processing. A purely confirmatory evaluation checking whether predefined objectives have been met appears largely insufficient for a couple of reasons:

- There is a lack of clearly formulated collaboration objectives. Official program objectives are too general (contribution to industrial wealth creation and unemployment reduction) to be used for individual project evaluations. The laboratories working guidelines are exclusively formulated in technical terms. The collaboration contracts between the laboratories and the contracting firms do not mention any specific objectives either. Thus, it is unclear which criteria projects could be evaluated against.
- Lacking the experience of former evaluations and considering the properties of technology, it would be peremptory to assume that no surprising results could emerge from R&D collaborations. Hence reduce the theoretical impact space to predefined criteria seems a questionable way of proceeding.

Hence, an exploratory evaluation proved to be necessary. It has been carried out on the basis of a sample of 22 R&D collaborations between CEA laboratories and firms of different sizes. All collaborations have taken place between 1975 and 1994, allowing for at least four years of commercial exploitation before evaluation takes place. The technological domains covered are very diverse (microelectronics, optronics, material sciences, industrial process optimization, etc.). For details on the samples, see **table 1**.

Sample	Partner firm	Technology firm	R & D collaboration
1	ALC1	Engraving processes	1984-88
2	ALC2	Reactors	1984-87
3	BRU	Cytofluorimetry	1982-84
4	CEN	Massively parallel processing	1991-94
5	COR	Hygrometer	1980-81
6	IBA	Electron accelerator	1991-93
7	IC	X-tomography	1983-86
8	SGM1	Massively parallel processing	1993-97
9	SGM2	Bubble memories	1976-80/ 1988-89
10	SCA	Strain gauges	1970-80
11	SES	Polishing process	1994-95
12	SOP1	Location electronics	1986-87
13	SOP2	Rectangular detector	1987-88
14	SOP3	Specialized electronics	1989-93
15	TCS	Single-layer CCD	1986-92
16	DEN	Wood identification	1987-95
17	NIP	PVD coating	1982-87
18	SGI	PVD	1986-88
19	SICN1	PVD	1990-92
20	SICN2	PVC	1990-92
21	T+C	Bimetal junctions	1981-85
22	VIC	Plate junctions	1978-84

Table 1. The sample contents (code of the company name, transferred technology, data of R&D collaborations)

Empirical data are gathered through a series of monographs that includes document analysis and face-to-face interviews with project leaders. Instead of relying on an exclusively pre-set formal methodology, the logical requirements of the task have been explored. General topics (motivation to collaborate, merits of CEA technology, induced effects, etc.) have been investigated along with answers to precise preconceived questions. The face-to-face interviews allow grasping contextual – often qualitative – elements that are necessary to understand the relationship between the laboratory and the firm. As a general feature, filed data on R&D collaborations are scarce: beside the archived contract, there are hardly any data available. The memory of the early collaborations is progressively vanishing as relevant interlocutors retired or moved to new jobs. Retrospectively, access to data has proved to be a labor-intensive and time-consuming exercise.

The quality of insight into the different companies is uneven, objective on the number of people that were possible to participate (between 2 and 7 per monograph). In order to have the most reliable information possible, the interlocutors were asked to validate the written report of the interview in the aftermath. As a general rule, the targeted interlocutors were the project leaders themselves, be it in the laboratory of in the firm. Generally speaking, their cooperation was good considering that the collaborations often lay for back in time.

2.2 Variable selection

Informations provided by the interviews had been categorized. One can distinguish two groups of variables:

- *Impact indicators* (or *objective* variables), that account for the impact of R&D collaborations have been evaluated.
- *Structural indicators* (or *explanatory* variables) such as the profile of the firm, the configuration of the collaboration, and the type of innovation that the firm wished to launch with the laboratory's support have been identified.

The first level of results concerns functional relationships between impact and structural indicators.

2.3 Construction of the objective variables (impact indicators)

On the basis of a heap of empirical observations, variables have been elaborated that are not the result of an aggregation of individual effects into single metric, but conceptual contractions of innovation-induced effects into a set of evaluation criteria. The procedure rests on the vicarious experience and on the "naturalistic generalization" of accumulated tacit knowledge as evaluator that "permit" to recognize faces, to understand metaphors. Tacit knowledge includes a multitude of inexpressible associations that give rise to new meanings, new ideas, and new applications of the old ideas. "Naturalistic generalization" is arrived at by *recognizing the similarities of objects and issues in and out of context and by sensing the natural co-variations of happenings*. Of course, this way of proceeding leaves out a couple of individual effects, but as the elaborated criteria are meant to serve as methodological basis for future evaluations, a certain level of generality has to be aimed at. The variables are constructed throughout a sort of *dialogue* between the empirical evidence gathered and the more conceptual and theoretical requirements of the evaluation literature. Modalities associated to each project are determined from the interview.

There are two subsets of *objective variables* (see **Table 2.**).

- I. The first subset is supposed to reflect the project contribution to the achievement of the mission overall objectives. The following variables have been retained:
 1. the *innovation induced turnover* related to the turnover of the concerned business unit (CAI) composed of 4 modalities
 2. the *durability of the average innovation-related turnover* (PERCAI), composed of 4 modalities
 3. the *effect of innovation* on the firm's employment level (EMB) composed of 2 modalities
- II. The second subset takes into account recurrent effects of R&D collaborations that are not accounted for by the first set. Four indicators have been retained:
 1. the *level of capitalization* on transferred knowledge (CAP) composed of 3 modalities
 2. the *level of intrafirm diffusion* (EXT) composed of 3 modalities
 3. the *reputation effects* (REPUT), composed of 2 modalities
 4. the *effect on the firm's competence building* (FONDS), composed of 2 modalities.

	Explanatory variables								Objective variables (impact indicators)						
	Firm profiles		Collaboration			Innovation			achievement of objectives		Induced effects				
Variable	Siz	Compe	Rel	Config	Pos	CompX	Dyn	Delai	Cai	Percai	Emb	Cap	Ext	Reput	Fonds
Modalities	4	4	4	4	4	2	3	3	4	4	2	3	3	2	2

Table 2. Description of the variable set.

In practice, the values of the modalities are gathered from document analysis or face-to-face interviews. This way of proceeding can be contested: *first*, one can argue that it is not the evaluator role to set the evaluation criteria; *second*, the construction of monograph-based variables set on "naturalistic generalization" depends on the unique experience of the evaluator. Considering the lack of data, there is hardly any way around.

2.4 Construction of the explanatory variables

Each R&D collaboration is described through three bricks of variables that account:

- The *profile of the firm*, i.e. its size (variable SIZ composed of 4 modalities) and its skill base (COMPE with 4 modalities).
- The *collaboration* that supposedly intervenes in the type of relationship (REL with 4 modalities), the organizational configuration (CONFIG with 4 modalities), the position of the collaboration in the innovation process (POS with 4 modalities) and its complexity (COMPX with 2 modalities).
- The launched innovation, which is commercial objectives (DYN with 3 modalities) and the time lag between the end of the collaboration and the launch of the innovation, identified by the variable DELAY.

3 A refresher on Multiple Correspondence Analysis

Correspondence analysis is a statistical technique, however, seldom thought of as a model, but rather as a data descriptive procedure with scaled parameter estimates as the descriptive tools. The data on the firms are collected such that N individuals answer to K “questions”. Each question has a finite number of modalities (see **Table 2**). In the classical correspondence analysis, the purpose is to see the relations between the modalities and to project them onto a factorial subspace. We can also do the same projection for the individual and even a simultaneous projection including the modalities of the variables¹. The third operation we can consider in this context is the segmentation of the individual space into homogeneous classes by means of a hierarchical classification algorithm (see [11] for more details).

When we deal with qualitative variables, we seek to highlight the typology of the modalities and we try to emphasize the relations between the relations existing between the modalities of the variables. For instance, if there is a considerable proportion of individuals that chose the modality “a” of the question “1” and the modality “b” of the question “2”, then we will say that the modalities (1,a) and (2,b) attract each other. We would like to get them very close in the representation. Conversely, if there is an important proportion of individuals who choose (1,a) and reject (2,b), then these modalities repulse each other, and we would like to observe distant representation. The goal here is to represent this kind of observations in a global manner which takes into account all the modalities of all the variables.

Suppose we have the answers of N individuals to K qualitative variables. Each question $k, 1 \leq k \leq K$ has m_k modalities and we denote by $M = \sum_{k=1}^K m_k$ the total number of modalities. The data table can be a *complete disjunctive* table (so-called Burt table) and defined as follows.

Let us denote by $D_{(N \times M)}$ the matrix with N rows and M columns which corresponds to the complete disjunctive table:

$$D = \{d_{ij}\} \text{ with } d_{ij} = \begin{cases} 1 & \text{if the individual } i \text{ chooses the modality } j \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

We have $\forall i, 1 \leq i \leq N, d_{i+} = \sum_{j=1}^M d_{ij} = K$. The marginal on the row is constant equal to the number of questions.

The data can be summarized in a Burt table $B_{(M \times M)}$ defined by

$$B_{(M \times M)} = \{b_{ij}\} = D^T D \quad (2)$$

In this case, we loose a part of the information about the individuals answers but we keep the information regarding the relations between the modalities of the qualitative. It is a generalized contingency table made up with $K \times K$ blocks, each block $B_{kr}, 1 \leq k, r \leq K$ representing the contingency table which crosses the question k and the question r . We briefly point out (see as an example [14]) that the Multiple Correspondence Analysis (MCA) is equivalent to a weighted Principal Component Analysis (PCA) performed on the row-profiles or column-profiles, obtained with a particular metric known as the Chi-square metric.

Indeed, for a MCA performed on the complete disjunctive table, we consider the matrix of the row-profiles (which is equal to the matrix $\frac{1}{K} D$), on which we carry out a PCA, by using the metric defined by the matrix NKW^{-1} and the

¹ Although this double projection is not always justified and accepted in the litterature.

weight of the row-profiles defined by the matrix $\frac{1}{N}I_N$ (I_N is the identity matrix). W is the diagonal matrix made up by the elements of the column margin, i.e. $W = \text{diag}(d_{+j}, \dots, d_{+M})$, where $d_{+j} = \sum_{i=1}^N d_{ij}$. In the case of the MCA performed on the Burt matrix, we use the matrix of the row-profiles equal to $\frac{1}{K}W^{-1}B$, the metric defined by the matrix NKW^{-1} and the weight matrix defined by $\frac{1}{KN}W$.

We have finally to perform a factorization of two inertia matrices [11]:

$$I_D = \frac{1}{K}D^T D W^{-1} = (D^C)^T D^C, \text{ with } D^C = \{d_{ij}^C\}, d_{ij}^C = \frac{d_{ij}}{\sqrt{d_{i+}}\sqrt{d_{+j}}}. \quad (3)$$

and

$$I_B = \frac{1}{K^2}B W^{-1}B W^{-1} = (B^C)^T B^C, \text{ with } B^C = \{b_{ij}^C\}, b_{ij}^C = \frac{b_{ij}}{\sqrt{b_{i+}}\sqrt{b_{+j}}}. \quad (4)$$

The two input matrices D^C and B^C are introduced to use uniform weighting and Euclidean distance, because in this way we can use the standard Kohonen algorithm. Moreover, although the theoretical demonstrations are made for an unspecified input distribution and any distance, the adaptive parameters are tested and controlled in the case of the uniform distribution and the Euclidean distance.

A second remark has to be made at this stage that concerns the relations existing between the rows of D^C and the rows of B^C : in the set of modalities, two modalities will be close if there is a large proportion of individuals that choose them simultaneously. In the same time, these individuals will be grouped in the same region. Further, a subset of modalities will be closer to a subset of individuals that have chosen these modalities than a group of individuals that have not.

Actually, there remains a difficulty which is not completely solved in the literature, concerning the simultaneous projection of individuals and modalities in the classical MCA. Indeed, the Burt table is a symmetric $M \times M$ matrix whose rows and columns represent the modalities, while the complete disjunctive table is a $N \times M$ matrix where the rows represent the individuals and the column represent the modalities. In both cases, the column vectors represent the same modalities, except that in one case, they are vectors of \mathbb{R}^M and, in the other one, they are vectors of \mathbb{R}^N . Then, if the dimensions are distinct, the superposition of the new systems of axes (principal components) is not completely justified.

In our methods, this conflict is not present since only the distances of the rows of these two matrices (which are both vectors of \mathbb{R}^M) are taken into account. Knowing that the sample is too small to rely on traditional statistics, we turned towards stochastic statistics, and in particular to a data processing technique based on Kohonen's algorithm.

4 The statistical method and material

4.1 The Self-Organizing Map algorithm

Graphical tools are elaborated in order to facilitate the visual inspection of the data set and discover its intrinsic structure. An interesting review of mapping techniques can be found in Biwas et al. [3] and Siedlecki et al. [17]. When linear mappings give confusing results due to the non-linear structure of the data, one can try to use neural models. The Kohonen algorithm, also called Self-Organizing Maps (SOM), is a widely used algorithm that has the property to reproduce the topology of the input space on a network made up by U units arranged following a grid or a line [13]. Each unit $u, 1 \leq u \leq U$, is characterized by a "code-vector" (also called "weight-vector") y_t^u with the same dimension as the input space. We define also a neighborhood function $V_{r(t)}(u, .)$ and depending on a radius $r(t)$ which is a time-decreasing function and which is defined in a homogeneous way around each unit u .

4.2 Training by the complete disjunctive table

We present now the original algorithm proposed by Ibbou [12] to deal with categorical variables. See also Cottrell *et al.* [6, 9], Cottrell and Rousset [7], Cottrel and Ibbou [8], where the method is fully described and explained. Let us suppose that the network is chosen, *i.e.* the number of units and its architecture (line or grid) are chosen. The purpose of the algorithm is to map the N data samples onto the network. The code-vectors of the units are vectors of \mathbb{R}^M and initialized at random. With the rows of the matrix D^C , we proceed to the training of the network: at each iteration t , we choose uniformly among all the rows of the matrix D^C one row d_i^C :

$$d_i^C = \left(\frac{d_{i1}}{\sqrt{d_{i+}}\sqrt{d_{+1}}}, \dots, \frac{d_{iM}}{\sqrt{d_{i+}}\sqrt{d_{+M}}} \right)^T. \quad (5)$$

We then look for the winner unit u_0 among all the units in the lattice, which is the unit realizing the minimum of the usual distance

$$u_0 = \arg \min_u \|y_t^u - d_i^C\| \quad (6)$$

We update the code vectors of the unit u_0 and its neighbors by the standard formula

$$y_{t+1}^u = y_t^u - \mathcal{E}(t)V_{r(t)}(u_0, u)(y_t^u - d_i^C). \quad (7)$$

These steps are repeated about 4 or 5 times over the total number of the inputs. At the early stage of the learning procedure, the size of the neighborhood is set to a large value. The adaptation parameter $\mathcal{E}(t)$ is also a time-decreasing and satisfies the Robbins Monroe conditions $\sum_t \mathcal{E}(t) = \infty$ and $\sum_t \mathcal{E}(t)^2 < \infty$. Then $\mathcal{E}(t)$ and $V_{r(t)}$ decrease toward zero with learning iterations. At the last stage of the learning process, only the winner unit is updated, which corresponds to a classical *competitive learning* procedure.

At the end, each individual is classified in the network by assignation to his winning unit. We then obtain a Kohonen map where only the individuals are classified. To represent the modalities on the same Kohonen lattice, we classify the rows of the matrix B^C in the map by assigning each one of them to the closest unit. For instance, the modality $p, 1 \leq p \leq M$, corresponding to the row vector b_p^C will be allocated to the unit u_p such that

$$u_p = \arg \min_u \|y_T^u - b_p^C\| \quad (8)$$

where y_T^u is the final value of the code-vector y^u after the training stage. Although the training was not made with the rows corresponding to the modalities, the network obtained can be used to classify the modalities despite it is not a very good quantizer for the modalities.

4.3 Training by the Burt table

The principal is similar to the one detailed in the previous section, but this time, we only use the rows of the matrix B^C to train the lattice. We classify then the modalities represented by the rows of the matrix B^C . To classify the individuals on the same map, we use the rows of the matrix D^C .

Generally, the number of modalities is not very large; consequently the training of the networks is very fast. Since neighboring units in the network are updated during the learning stage, there is a tendency that neighboring units represent neighboring locations in the data space. Once the algorithm has converged, the map displays a classified organization of the modalities such that close modalities are associated to the same class or to close classes, according to the arrangement of the corresponding units in this network.

The main advantages of this algorithm are its simplicity and its timesaving for very large data files, for instance in marketing or insurance companies: it happens that with this type of data base, the complete disjunctive table has several hundred thousands (even more) of individuals and thousands of variables. It can take several hours to classify the individuals into groups by a hierarchical classification. Using this algorithm, it is sufficient to compute the Burt matrix and to train a Kohonen network with its rows. It produces one single map, contrarily to classical analysis that gives

many representations with decreasing information. It is then easy to classify the individuals in the lattice by seeking the winner unit related to the rows of the matrix D^C . Kohonen mapping is poorer but allows a fast interpretation. See Ibbou and Cottrell [11], Cottrell *et al.* [9] for more details and for financial applications. Let us now illustrate this method with our application.

5 Application to the technology transfer

5.1 Global representation

Let us consider the dataset of 22 R&D collaborations between laboratories and industrial firms. Clearly, the sample size is too small with respect to the number of variables (110) contained in the complete disjunctive table, even for a SOM algorithm. Then the first step consists to transform the data into a *distance matrix* or “*dissimilarity*” matrix in order to reduce the data dimension. During the last decade, several propositions have extended the vector SOM model to pairwise dissimilarity data, the so-called DSOM algorithm (Dissimilarity SOM) [19]. Let us note $M = \{m_{ij}\}$ $i, j = 1, \dots, N$, all pairwise dissimilarities between the data, denoted $\|d_i - d_k\|_{i,k}$. The dissimilarity is symmetric and positive (with $\|d_i - d_i\| = 0$). As the standard SOM, the DSOM maps input data from an input space to a low dimensional organized set of U units arranged via a prior structure, but these dissimilarities describe a non-metric space. The major difficulty of the DSOM is the constraint on the output data representation. For (vector) SOM algorithm, there is a latent data model for each output prototype, *i.e.* spherical distribution whose vector-code is the barycentre. For DSOM, there is no data model, so one referent observation is explicitly associated to each vector-code instead of the barycentre processing. This reference is usually chosen (optimization process) among the input observations.

The algorithm uses the initial SOM framework to adapt to dissimilarity data. During the initialisation step, the N individuals are randomly assigned to the prototypes. At each iteration t , an *affectation* step and a *representation* step are sequentially processed. The role of the former is to assign to each observation vector $m_i = (m_{i1}, \dots, m_{iN})^T$ the best matching unit y_i^u . A partition of the whole set is then realized. In the latter, a new observation $m_{f(u)}$ is assigned to the code-vector y^u minimizing the following cost function

$$f(u) = \arg \min_j E(u, j), \quad (9)$$

where $E(u, j)$ is the weighted local distortion if m_j is the reference of the unit y^u :

$$E(u, j) = \sum_{i=1}^N V_{r(t)}(u_0, u) \|m_j - m_i\|^2, \quad (10)$$

with u_0 as defined in (6). The global cost function which is then minimized is the global distortion over the units $\sum_{u=1}^U E(u, f(u))$. But, since the set of referent candidates is finite, distortions occur if several nodes share the same reference. One can find in [19] different propositions made to overcome this difficulty.

All the subroutines are developed with the SAS software, version 9.1. As mentioned above, the first and raw result we get after learning is a classification of the N individuals into 6×6 classes (see **Fig. 1**). Eventually some classes may be empty. At the same time, we get the code-vectors. See in **Fig. 2**, a standard representation of the code vectors inside their own unit.

KACM1 : grille 6x6 et 1500 iterations

TCS		ALC1 T+C			SOP1 SOP2
BRU			IBA		
DEN		CEN			SGM1
	SES		SOP3 NIP		
SCA COR		SICN1 VIC		SICN2	
	SGI		IC		ALC2 SGM2

Fig. 1. The two-dimensional map with the result of MCA for a 36 units grid.

Representants des classes (Poids Finaux)



Fig. 2. In each sell, the final 22 dimensional code-vector is drawn. Note that neighboring cells correspond to similar vectors.

The choice of the number U of units is arbitrary, and there does not exist any method to better choose the size of the network. We can guess that it is often difficult to give relevant interpretation of a too large number of classes. So we propose to reduce the number of classes by means of a hierarchical classification of the U code vectors using the Ward distance for example. In this way, we defined two embedded classifications, and can distinguish the Kohonen classes (or “micro-classes”) and the macro-classes, that grouped together micro-classes. To make visible this two-level classification, we affect to each “macro-class” some color. See in **Fig. 3** a representation of the “micro-classes” put together to constitute 4 “macro-classes”.

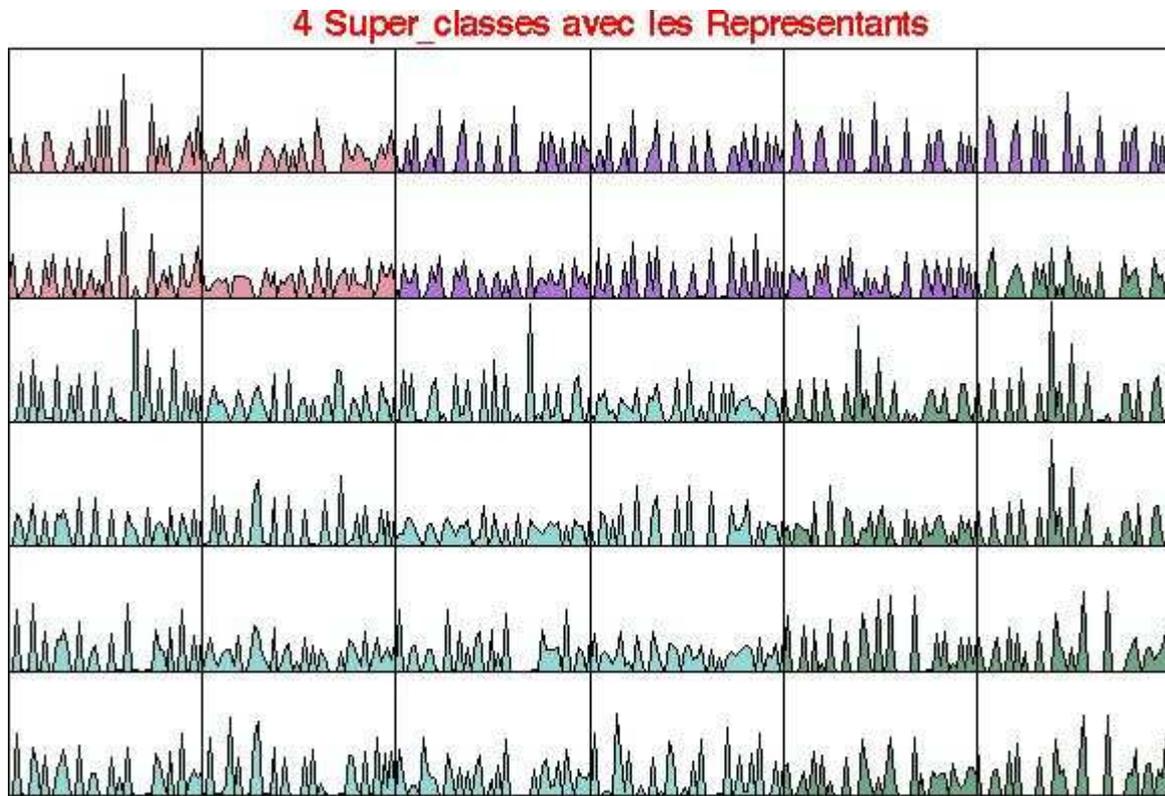


Fig. 3. The 4 “macro-classes” resulting of the hierarchical classification of the 36 code-vectors superposed on the grid.

The advantage of this double classification is the possibility to analyze the data set at a “macro” level where general features emerge and at a “micro” level to determine the characteristics of more precise phenomena and especially the paths to go from one class to another one [5]. Connected components in a macro-class form areas in the grid where units are topologically near.

In order to have a better idea of the overall geometry of the map (the standard representation is a grid, but this can introduce a confusion in the interpretation) it is valuable to visualize the distances between classes. We use the method proposed by Cottrell and De Bodt [9]: each unit is represented by an octagon. The bigger it is, the closer the unit is to its neighbors. So the clusters appear to be regions in which octagons tend to be big and frontiers are regions uncolored. See an example in **Fig. 4**.

We can observe that in general the “macro-classes” boundaries coincide with the most important distances between classes. Conversely, if a boundary between two classes occurs with small distance that means that the second level classification splits a large group into two groups and that the path from one to the other is continuous: we should perhaps consider a hierarchical classification with fewer classes.

4 Super_classes avec les Proches Voisins

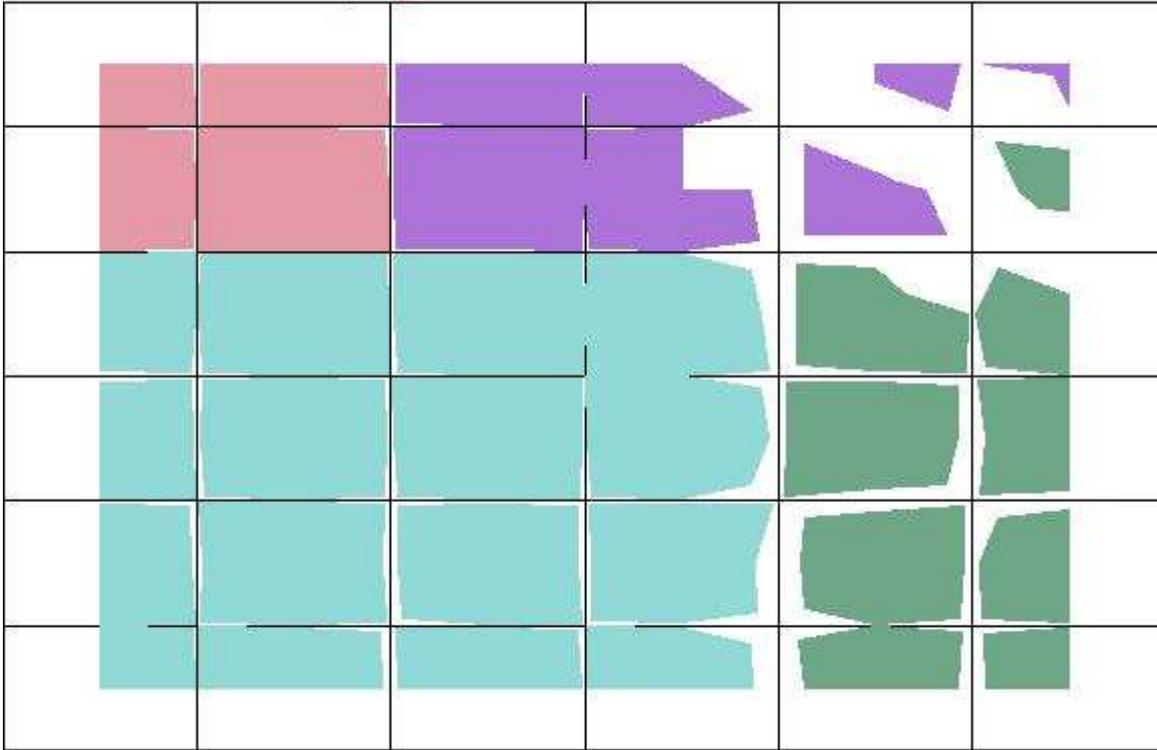


Fig. 4. Representation of the distances: the gaps coincide with the frontiers of the “macro-classes”.

5.2 Piecewise classification

Knowing that the sample is too small, a better idea is to limit the analysis to small number of variables.

The links between the different variables are currently being tested for their significance. The perspective is twofold:

- Try to identify (our primary goal) the *most influencing variables* in the genesis of innovation-induced effects (*i.e.* is it the profile of the firm or the configuration of the collaboration that *matters most?*).
- Identify those cases where the effects lie beyond a purely commercially impact.

Only *interdependences* between a number of variables are of interest here. The study of interdependence leads to the theory of correlation. The correlation parameter ρ may be used as a measure of interdependence when the joint distribution is approximately binormal. But if a variable is correlated with a second variable, this may be merely incidental due to the fact that both are correlated with another variable or a set of variables when others are held constant, *i.e.* conditionally upon those other variables taking some certain fixed values.

If we find that holding another variable fixed reduces the correlation between two variables, we infer that interdependence arises in part through that over variable; and if the partial correlation is zero, we infer that their interdependence is entirely attributable to that variable. Conversely, if the partial correlation is larger than the original correlation between the variables we infer that the other variable was masking the correlation.

Let us assuming that the variables are jointly normally distributed. Such an approach is not necessary, but simplifies the development. Let x denote a vector of M random variables that are multinormally distributed. Without loss of generality, we standardize the variables so that the covariance matrix becomes the correlation matrix C .

Let us examine the resulting correlation expressions: direct correlation is $\text{corr}(x_1, x_2) = \rho_{12}$.

$$\text{corr}(x_1, x_2 | x_3) = \rho_{12.3} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\left[(1 - \rho_{13}^2)(1 - \rho_{23}^2) \right]^{1/2}}. \quad (11)$$

If the subscript m is used to denote any subset of conditioning variables, we arrive at the partial correlation expression:

$$\text{corr}(x_1, x_2 | x_3, x_m) = \rho_{12.3m} = \frac{\rho_{12.m} - \rho_{13.m}\rho_{23.m}}{\left[(1 - \rho_{13.m}^2)(1 - \rho_{23.m}^2) \right]^{1/2}} \quad (12)$$

The subscripts $(3, 4, \dots, M)$ may, of course, be permuted into any order. When $\rho_{ij.m} = 0$, we say that x_i and x_j are *conditionally independent*, given x_m . When the joint distributions are not normal, we may say only that x_i and x_j are *conditionally uncorrelated*.

The diagram represented **Fig. 5** was obtained by computing every statistics following (12). A line is drawn between two variables if (12) put in evidence an interaction between these two variables such that p-value lies between 5 and 10% (black) and such that p-value is less than 5% (red).

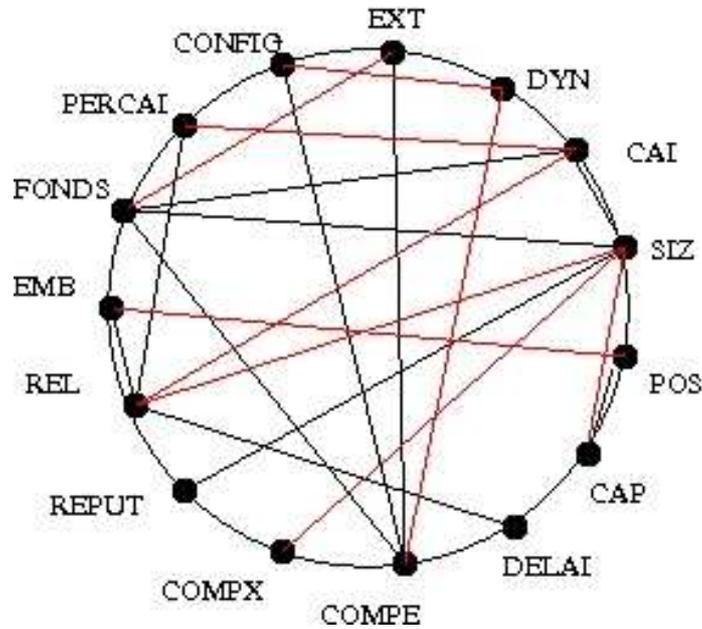


Fig. 6. Dependence graph for CEA data variables.

From this representation, it can be observed that:

1. 3 variables (COMPE, DELAI, REPUT) have only a single link. These “terminal vertices” are carrying only margin information.
2. 5 particular (triangular) associations appear, $\{\text{PERCAI, CAI, REL}\}$, $\{\text{CAI, REL, SIZE}\}$, $\{\text{CAI, FONDS, SIZE}\}$, $\{\text{FONDS, EXT, COMPE}\}$, $\{\text{COMPE, DYN, CONFIG}\}$ that can be analysed further with a three-way contingency table. Such an arrangement is called by convention a “clique”. That should occupy a significant place in the chart.
3. the variables (SIZE, COMPE, REL, CAI, FONDS) are implicated in 2 cliques.

The result of this research is not an aggregated impact indicator, but a heuristic model that allows identifying those cases in which the innovation-induced turnover is likely to be a “sufficient” impact measure and those where this impact measure is not informative enough. In this perspective, aggregation is not absolutely necessary.

As an example, we will analyse in details the case of the clique $\{\text{CAI, FONDS, SIZE}\}$: from direct inspection of the numbers in the Table 5, some association between the size of the firm (SIZE) and the firm’s competence building (FONDS) seems apparent, and both of these seems also positively associated to having an innovation-induced turnover (CAI). But this is all vague and relationships are clearly somewhat complex.

For this classification, we compute the contingency table. To analyse the dependence between these variables, we use a Kohonen MCA algorithm. See in **Fig. 7**, **Fig. 8** and **Fig. 9** the results.

The conclusions are simple. In terms of impact on the firm, a strong CAI is more constructive than a weak CAI. Moreover, for a strong, one usually observe an effects on firm’s competence building (FONDS). We cannot assume a causal connection in the induced effects: claims of causality must, ultimately, rely upon extra-statistical criteria. However, on the basis of the acquired knowledge, it is reasonable to suppose the order of appearance $\text{CAI} \rightarrow \text{FONDS}$. Under this hypothesis, FONDS is regarded as an *indirect induced effect* of the commercial exploitation.

Note that the implication of causality comes from the prior specification of the system, and the role of statistical analysis is to examine whether or not the data support the hypothesis. Cai is an adequate measure of impact under this hypothesis.

KACM1 : grille 5x5 et 1300 iterations

CAI_2 EBU TCS		CAI_5 DEN	SIZ_4	CAI_4 ES OP1 OP2 OP3
	VIC		FONDS_1	
SIZ_2 K SICN1	CAI_1	SIZ_1 FONDS_0 SGM1 NIP		ALC1 T+C
SICN2			IBA	CAI_3
CAL0 ALG2 SGM2		CEN SG1	SIZ_3	COR SCA

Fig. 9. Projection of the modalities of {CAI, FONDS, SIZE} and the 22 individuals on the Kohonen 5x5 grid.

3 Super_classes avec les Representants

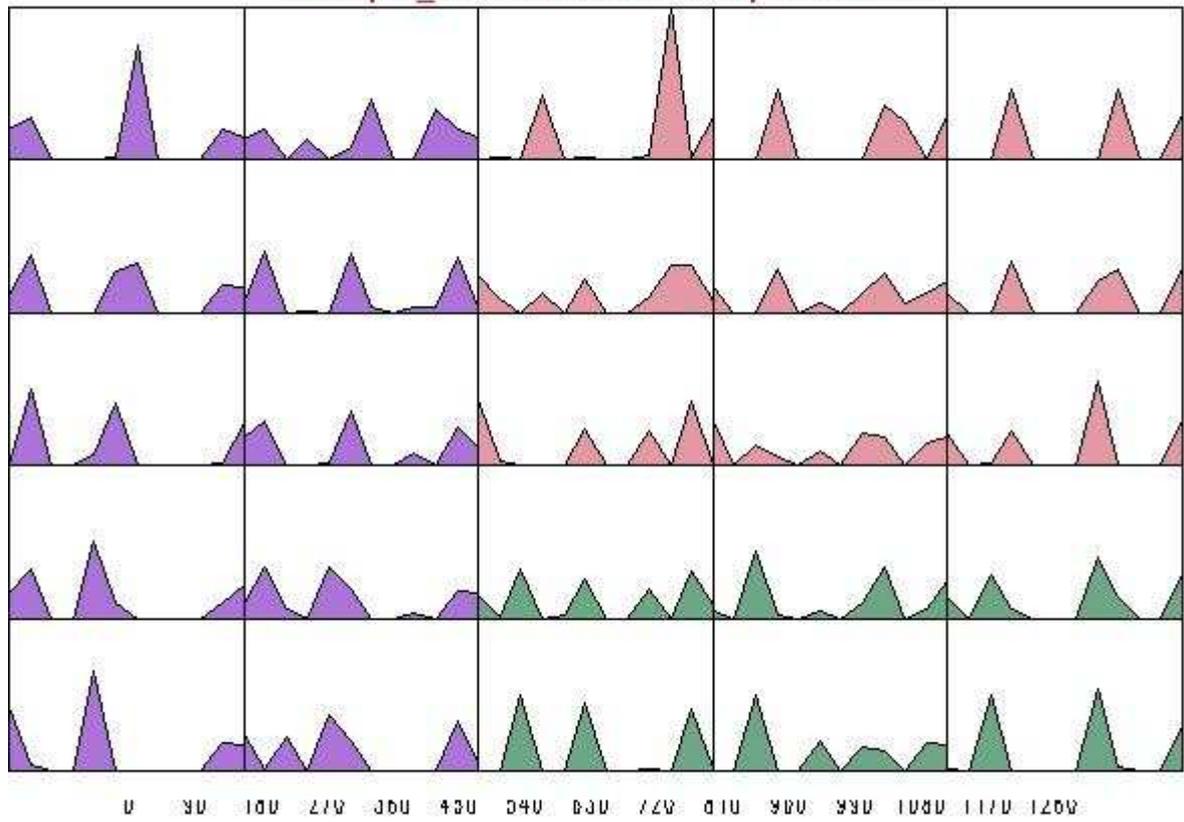


Fig. 10. The contents of the classes and the 3 macro-classes resulting from the classification.

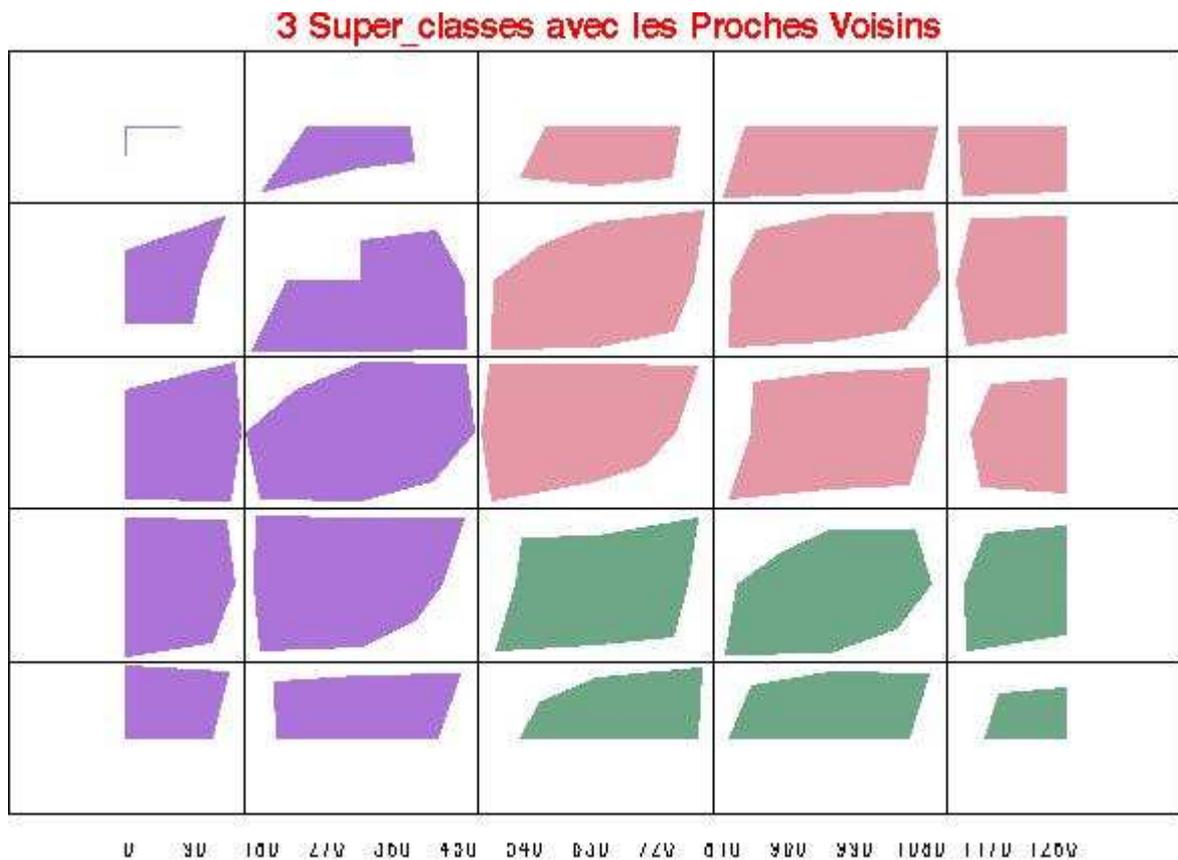


Fig.9. Representation of the distances that gives an idea of the discrimination between the 3 classes.

The map **Fig. 7** can then be read as a non-linear projection of the observations. By highlighting the distance between the different classes (**Fig. 9**), the algorithm makes the interpretation of the map relatively easy.

The first results that we get are very promising: on the samples that we study, our method provides very quickly good representations of the relations between several qualitative variables, and in particular the links between

- the profile of the firm (the “absorptive capacity”),
- the characteristics of the collaboration (the “process”),
- the characteristics of the launched innovation (the “outcome”),
- the innovation-induced effects,
- the contribution of the public laboratory in these effects.

The links between explanatory variables and their weights (profile of firms, configuration of the collaboration and characteristics of the innovation) are currently being evaluated. It shows that the construction of an evaluation tool is constantly torn between the requirement for *operational* indicators, the complexity of the produced empirical knowledge and the *access* to reliable and high-quality information.

Reflowing already well-broken ground, this paper reminds that technology is a phenomenon that is improperly treated by traditional economics and it suggests a more appropriate way to look at it. It shows that traditional expectations towards evaluation are overstretched and it suggests a renewed conception with more realistic expectations.

Rather than executing a confirmatory evaluation (lack of easily-accessible indicators), an exploratory evaluation has been carried out on the basis of twenty two R&D collaborations between CEA laboratories and industrial firms. With the empirical observations that emerge from observation, a set of evaluation criteria have been constructed by “naturalistic generalisation”. The relationships and links between these variables and explanatory variables (profile of the firm, configuration of the collaboration, characteristics of the innovation) are currently being tested by a neuronal technique, namely the algorithm of Kohonen.

It shows that the construction of an evaluation tool is constantly torn between the requirement for *operational* indicators, the *complexity* of the produced empirical knowledge and the *access* to reliable and high-quality information. This research should open up perspectives for policy makers on how to make a programme such as the technology dissemination program “manageable for results” (see Georghiou [10]).

6 Conclusions

We propose a general methodology to analyse multidimensional data when the observations are described by qualitative variables. Many other techniques can be combined in order to improve this analysis and the interpretation, in particular a hierarchical classification to compose macro-classes. In this way, we can propose a typology of the classes easy to interpret. The maps we obtain are less precise than classical MCA projections, but they summarize very well the various relationships (attractions, repulsions) between modalities. Even if the simultaneous representation of the various modalities and the individuals does not have a rigorous justification, it gives nevertheless good results. These methods have the advantage of producing in the one hand a Kohonen map comparable with a traditional MCA projection and on the other hand a classification of the individuals compared to the modalities. Moreover, the method is particularly cheap in computing time, which is a considerable advantage compared to the classical classification methods such as a hierarchical classification.

Furthermore, the research from which this paper is drawn is among the very first to analyse the CEA R&D collaborations from the economic point of view. The particularly rich empirical material has remained largely unexplored by the economists so far. It aims at developing evaluation criterion that allows grasping the impact of technological collaborations between the CEA's laboratories and industrial firms. Within the landscape of existing impact evaluation models, it is meant as a contribution to the “*market-impact*” model.

Bibliography

- [1] Andersen, E.B. (1989). *Introduction to the analysis of categorical data*. Springer.
- [2] Anderson, T.W. (1984). *An introduction to multivariate statistical analysis*. Wiley, 2nd edition, New-York.
- [3] Biwas, G. Jain, A.K. and Dubes, R.C. (1991). Evaluation of projection algorithms. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. PAMI-3, No 6, pp. 701-708.
- [4] Blayo, F. and Demartines P. (1991). Data analysis : How to compare Kohonen neural networks to other techniques ? In *Proc. Of IWANN 91*, Prieto Ed., LNCS Springer, 469-476.
- [5] Cottrell, M. and Rousset, P. (1997). The Kohonen algorithm : a powerful tool for analysing and representing multidimensionnal quantitative and qualitative variables, *IWANN'97*, Lanzarote, juin 1997, J. Mira, R. Moreno-Diaz, J. Cabestany (eds), *Lecture notes in Computer Sciences*, No 1240, Springer, pp. 861-871.
- [6] Cottrell, M. Letremy and Roy, P. (1993). Analyzing a contingency table with Kohonen map: a factorial correspondence analysis. In *Proc. of IWANN'93*, J. Cbestany, J. Mary, A. Priesto (eds), Spriner, pp.305-311.
- [7] Cottrell, M. Fort J.C. and Pages P. (1998). Theoretical aspects of the SOM algorithm. *Neurocomputing*, N0 21, pp. 119-138.
- [8] Cottrell, M. Fort J.C. and Pages P. (1998). Two or three things that we know about the Kohonen algorithm. *Proceedings of ESANN'94*, M.Verleysen ed., Editions D Facto (ISBN 2-9600049-0-6), Bruxelles, p.235-244.
- [9] Cottrell, M. de Bodt, E. and Henrion E.F. (1996). Understanding the Leasing Decision with the Help of a Kohonen Map. *An Empirical Study of the Belgian Market*, *Proc. ICNN'96 International Conference - Washington*, vol. 4, p. 2027-2032.
- [10] Gorghiou, L. (1998). Issues in the evaluation of innovation and technology policy. *Evaluation*, Vol. 4, No 1, pp. 37-51.
- [11] Ibbou, S. and Cottrell, M.(1995). Multiple correspondence analysis of a cross-tabulation matrix using the Kohonen algorithm. *Proc. of ESANN'95*, Avril, Editions D Facto, Bruxelles, p. 27-32.
- [12] Ibbou, S. (1998) *Classification, analyse des correspondances et méthodes neuronales*. Thèse de doctorat.
- [13] Kohonen, T. (1995). *Self-organizing maps*, Springer series in Information Sciences, Vol. 30, Springer.
- [14] Lebart, L. morineau, A. and Piron, M. (1995). *Statistique exploratoire des données*. Dunod, Paris.
- [15] Pearl, J. (1988). *Probabilistic inference in intelligent systems*. Morgan Kaufmann, San Mateo, California.
- [16] Saporta, G. (1992). *Probabilité, analyse de données et statistiques*. Technip, Paris, 1992.
- [17] Siedlecki, W Siedlecka, K. and Slansky, J. (1988). An overview of mapping techniques for exploratory pattern analysis. *Pattern Recognition*, 21(5), pp. 411-429.
- [18] Zuscovitch, E. (1986). The economic dynamics of technologies development. *Research Policy*, Vol.15, p.175-186.
- [19] Ho-Phuoc, T. and Guérin-Dugué, A. (2007) A new adaptation of Self-Organizing Map for dissimilarity data, *IWANN 2007 - LNCS 4507*, pp. 219-226.