# Evaluation of Scientific and Technological Innovation using Statistical Analysis of Patents*

Martin Rajman[1], Vivi Peristera[2], Jean-Cédric Chappelier[1], Florian Seydoux[1], Antonis Spinakis[2]

[1] EPFL – DI-LIA – INR (Ecublens) – CH-1015 Lausanne – Switzerland
[2] Quantos SARL – 154 Sygrou Ave – 176 71 Athens – Greece

## Abstract

In this paper we tackle the problem of the multidimensional analysis of patents based on the use of textual and statistical analysis techniques. The use of correspondence and cluster analysis permit to identify technological trends and innovation. Furthermore the interactions between the different fields of activities are captured through the use of these statistical methods. Also indicators based on patents can be produced in order to depict in a quantitative way the technological activity in a European level. Finally here are presented the different steps required for the textual and statistical analysis of patent data.

## Résumé

Ce papier présente une étude statistique textuelle multidimentionnelle d'une base de brevets. L'utilisation de l'analyse des correspondances permet d'identifier les grandes tendances innovatrices. De plus, les relations entre divers champs d'activité peuvent être mises en évidence au moyens de techniques de classification. Plusieurs indicateurs basés sur l'analyse de cette base de brevets peuvent être produits et décrire de façon quantitative l'activité technologique au niveau européen. L'article détaille les différentes étapes de l'analyse textuelle et statistique de la base de brevets.

**Keywords:** Textual analysis, Correspondence analysis, Patent classification.

## 1. Introduction

The measurement and assessment of technological innovation is a very specific subject that attracts the interest of many actors. As reflected in national S&T statistical publications, the demand for such indicators has steadily increased over the past decades, both for macro and micro-economic analysis, as well as for policy use. Furthermore, as indicators of the inventive output are mostly based on patents, the development of efficient and innovative methodologies for the analysis of the information related to technological innovation stored in patent databases is a necessity (Comanor and Scherer, 1969; Dou, 1995; Narin, 1995).

In this paper, we present a methodology for multidimensional analysis of the information "hidden" in patents, based on textual and statistical analysis techniques such as correspondence and cluster analysis (Benzécri et al., 1973; Lebart et al., 1997; Lebart L., 1998; Johnson and Wichern, 1998). This methodology was developed and is currently being extended, tested and implemented in the framework of the European IST project STING aiming at the analysis of

*The work described in this paper was funded by European Community and OFES in the framework of the STING project (IST-99.20847/OFES-00.0421).

scientific and technological innovation in Europe, and, more specifically at the production of indicators regarding technological trends and innovation (Guellec and van Pottelsberghe, 1998).

Furthermore the methodology must provide ways to analyze European scientific and technological innovation and progress at various different levels. Thus, it has to provide tools to perform analysis at the sector level in order to identify technological evolution within a specific sector taken in isolation. The same analysis tools should however also be applicable to a whole set of sectors (eventually all sectors) in order to enable the production of information about aggregated scientific and technological activity in relation with sets of specific companies, or for a specific year or range of time. Finally, analysis at the level of countries shall be available as well so that homogeneous groups of countries might be identified to comparatively analyze scientific and technological progress, competitive levels or specific technological specializations for different European countries.

Among the main objectives of the methodology, we will therefore find goals such as: (1) to support multidimensional comparisons between countries, sectors or companies; (2) to identify competition in a given sector; and (3) to capture interactions that might exist between domains of technological activity and poles of innovation inside these domains.

In the rest of this contribution, we will first provide, in section 2, an overall description of the methodology. Then, in section 3, we more specifically focus on the use of correspondence analysis for the processing of the textual content of the patent data, while, in section 4, we concentrate on the application of cluster analysis to the production of relevant groups of technologies or patents, as well as the identification of relevant relationships between them. For both sections, results obtained on real patent data extracted with the MIMOSA software from the European Patent Office[1] ESPACE/EPA database (Vol. 2000/006) are provided. Finally, some conclusions are drawn in section 5.

## 2. Overall description of the proposed methodology

**Technical features of the method** The main innovative character of the methodology consists in the fact that it not only makes use of the full International Patent Classification (IPC) codes (instead of the first digits of these codes as it is usually the case) but also takes into account the additional information that is stored in textual part of the patent data (especially in the titles and the abstracts). To fully exploit this information and derive richer and more reliable results, textual and statistical analysis techniques such as correspondence and cluster analysis are used. Supplementary variables that are not directly involved in the core computations required by these techniques (such as companies submitting the patents, inventors, countries in which the patent is submitted, etc.) may further enrich the result of the analysis.

Figure 1 provides a synoptic overview of the steps involved in the presented methodology.

The first step of the analysis is the linguistic preprocessing of the textual parts of the patent data. In our current experiments, this preprocessing is limited to lemmatization and part-of-speech assignment, along with simple vocabulary filtering based on frequency range and part-of-speech selection.

The second step mainly concerns the possibility in the foreseen analysis environment to perform simple standard analyses and to produce a variety of different types of graphs in order to

[1] www.epo.org

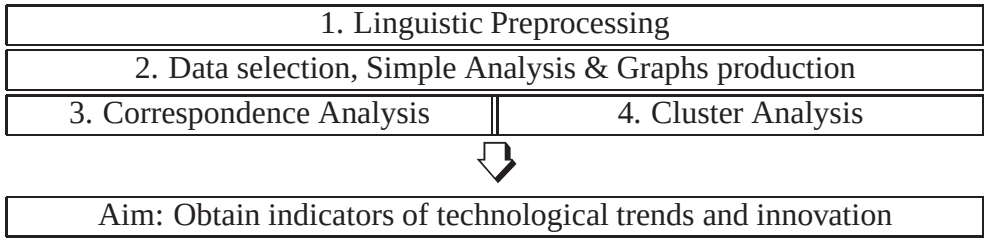


Figure 1: *Steps undertaken in the analysis of patent data*

visualize data and results in the most appropriate way. A second important concern within this step is data selection, as the user also needs to have efficient ways to produce the contingency matrices (capturing the specific information he is interested in) required for the more complex analyses (based on correspondence and cluster analysis) described in the subsequent sections.

**Linguistic preprocessing** As far as the preprocessing of the textual data is concerned, we use the Ingenia SYLEX (Constant, 1995) syntactic analyzer to automatically perform lemmatization and part-of-speech assignment. Lemmatization consists in restricting the morphologic variation of the textual data by reducing each of the different inflections of a given word form to a unique canonical representation (or lemma). To further reduce the vocabulary size, we restricted the analysis to the 3 word categories (as identified by the assigned parts-of-speech) bearing most of the semantic content: nouns, verbs and adjectives.

As the resulting vocabulary still contained over 40,000 different lemmas, additional reduction was necessary to reach a more tractable size; to that end, we computed the impact on vocabulary size of the following upper and lower thresholds for the lemma frequencies (in the whole patent database):

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| fmin | – | 5 | 10 | 10 | **15** | 15 | 30 | 50 | 100 | 200 | 500 |
| fmax | – | – | – | 4000 | **4000** | 2000 | 2000 | 2000 | 2000 | 2000 | 2000 |
| voc. size | 39,939 | 14,778 | 9,877 | 9,848 | **7,853** | 7,807 | 5,113 | 3,670 | 2,252 | 1,271 | 475 |

and finally choose fmin=15 and fmax=4,000. In addition, we also removed some frequent "empty" words such as *be, have, can, may*, as well as some non lexical forms such as *p1, st11, 6-12c*, ..., ending up with a total vocabulary of 7,724 lemmas. The large size of the remaining vocabulary indicates the important lexical variety observed for the patent titles and abstracts.

**Correspondence Analysis** In traditional approaches to patent analysis, knowledge about the content of a patent is usually restricted to (parts of) the IPC codes that have been assigned to that patent during the application process. Although the IPC represent a quite rich hierarchy of codes, not taking into account the textual content of the patent may be considered as a limitation to a fully efficient exploitation of patent data. Therefore, one of the important characteristics of the presented methodology is the integration of textual data analysis techniques for the processing of the textual content of the patents (titles and abstracts). The underlying idea is that the vocabulary that is characteristic for patent classes build on the basis of various descriptive variables associated with the patents (such as country or date of application) provide additional interesting insights for the analysis of the patents themselves. As a case study illustrating

our approach, we applied the described techniques to the collection of 23,750 patents extracted from the ESPACE/EPA database (Vol. 2000/006).

**Cluster Analysis** In the case of textual data, clustering techniques are used for representing proximities between the elements of lexical tables. In the general case, cluster analysis operates on contingency tables to identify relationships between two different nominal variables. In the case of patent data, the aim of the procedure is to identify groups of technologies that share common vocabulary and groups of patents that share common technologies in order to derive conclusions about technological trends and innovation.

More specifically, we apply cluster analysis to contingency tables cross-tabulating full IPC codes and words (resp. patents and words) in order to identify homogeneous groupings of and relevant relationships between grouping of IPC codes (resp. patents).

The information captured in such clusters and inter-cluster relationships can then be directly used for the production of technological indicators, the goal being to identify areas of technology that share common characteristics, as well as innovative areas characterized by isolated clusters.

## 3. Correspondence analysis for patent data

**Defining a contingency table:** After the preprocessing phase, the first step of the methodology was to choose the descriptive variable(s) that shall be used to group the patents into classes to be submitted to textual analysis. For our experiments, we decided to use the variable "country of application", and, in order to have significant patent populations in all the processed classes, we restricted our study to the 16 countries with the highest patent application totals, as listed in table 1.

| US | DE | JP | GB | FR | IT | SE | FI | NL | AU | KR | CH | AT | DK | IL | ES |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8,534 | 4,113 | 2,908 | 1,531 | 1,275 | 545 | 430 | 272 | 240 | 215 | 204 | 187 | 153 | 142 | 125 | 111 |

Table 1: *List of the 16 countries with the highest patent application totals*

Once the above mentioned vocabulary selection and patent grouping were made, the 16 x 7724 contingency matrix that constitutes the raw material for the correspondence analysis was produced. In this contingency matrix, each row is associated with a country, each column with a word forms (lemma), and each cell records the number of occurrences of a given lemma in the title and/or abstract of the patents submitted in a given country. With such a set up, the goal of the correspondence analysis performed on that contingency matrix is to enable an exploration of the non random dependencies that can be observed between the possible application countries and the vocabulary used in the title/abstract of the patents. More precisely, the correspondence analysis produces a new vector space (the factor space) in which similarities between the rows and the column of the contingency matrix (as measured by the $\chi^2$-distance) can be visualized as geometric proximities.

**Selecting the factor axes:** As it is quite difficult to actually perform any visualizing in more than 2 dimensions, it is of a central concern to decide how the dimensions (i.e. factors) of the resulting factor space should grouped to produce plots that are effectively useful for the interpretation. To carry out this important step of the methodology, various contribution measures

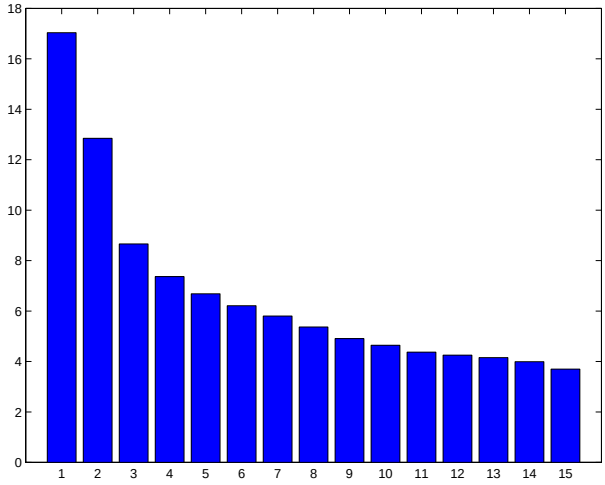


Figure 2: *Explained inertia for Country* $\otimes$ *Vocabulary*

can be used. First, for each of the factors, an overall contribution ("*inertia*") to the resulting factor space can be computed and used to measure the relative importance of each of the factors. The contributions of the 15 factors are shown in decreasing order in figure 2.

The relatively slow decrease rate indicates there is no subset of factors responsible for a major part of the observed dependencies (for example, the 2 first factors only contribute to about 30% of the whole inertia). To provide some help for the selection of the most useful factors to be drawn, a second type of contributions can be used: the $cos^2$ (or relative contribution) measuring for each of the application countries the importance of each of the factors for its representation in the factor space. For example, the $cos^2$ (in decreasing order) of the 4 most contributing countries[2] are the following:
**JP:** Factor 2 (.768), Factor 1 (.209), ...
**US:** Factor 1 (.951), Factor 5 (.021), ...
**DE:** Factor 1 (.476), Factor 2 (.159), Factor 7 (.146), Factor 5 (.106), ...
**FI:** Factor 3 (.852), Factor 5 (.058), ...
where the values in parentheses are the $cos^2$ for the different factors.

These $cos^2$ values can be used to simplify the interpretations by restricting to the groups of (usually 1 or 2) factors that provide the best representations for various groups of countries. For example, on the basis of the above given $cos^2$ (and with a lower threshold value of .150), factor 1 and 2 might be used only for US, JP and DE and factor 3 only for FI. If the same selection procedure is applied to the 16 countries and 15 factors associated with our contingency matrix, we obtain the following results[3]:

factors 1 and 2 should be used to visualize the relative positions of JP (.997), US (.951) and DE (.605)

factors 5 and 6 should be used to visualize the relative positions of DK (.854) and G-B (.719)

factors 7 and 9 should be used to visualize the relative positions of IT (.709) and SE (.501)

[2]By "most contributing countries", we mean here the countries with the highest contribution to the overall inertia.

[3]The values in parentheses are the sums of the $cos^2$ over the corresponding factors

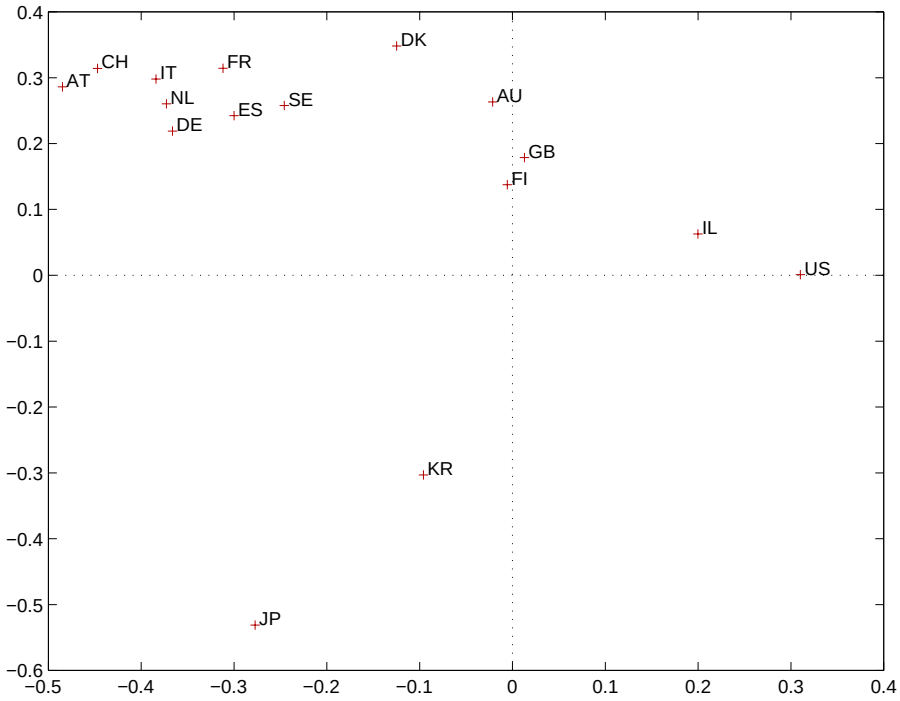


Figure 3: *Projection in the first two factorial axis*

factors 11 and 12 should be used to visualize the relative positions of IL (.764) and AT (.629)

factors 13 and 14 should be used to visualize the relative positions of ES (.824) and AU (.771)

The other factors are all associated with mainly one unique country (factor 3 with FI (.852), 4 with FR (.691), 8 with KR (.727), 10 with NL (.435), and 15 with CH (.826)).

For the rest of the description of the methodology, we will essentially concentrate on the plot associated with factors 1 and 2. The projection in the corresponding factor plane of all the 16 countries is given in figure 3. Notice however that, as already mentioned, this factor plane is mainly useful for the analysis of the relative positions of US, JP and DE.

**Interpreting the factors:** Visualizing relative positions is of course of some interest per se (in particular in the case of proximities). However, a more precise analysis of the factors is required in order to provide an interpretation that really uncovers some of the interesting information that is hidden in the patent data. In particular, it is of central concern to be able to relate the observed relative positions with the content of the corresponding patents, in other words with the underlying vocabulary used in the titles and abstracts.

However, as the number of lemmas (7,724) is very large, an additional effort needs to be made to automate the selection of the lemmas that are the most useful for the interpretation of the plots. To help with this task, for each of the plots and each of the pairs of countries it contains, we compute the variation of the distance (in the plot) between the 2 countries consecutive to the removal of a given lemma in the representations. Then, only the lemmas corresponding to the highest variations are visualized on the plots to help the interpretation of the relative positions

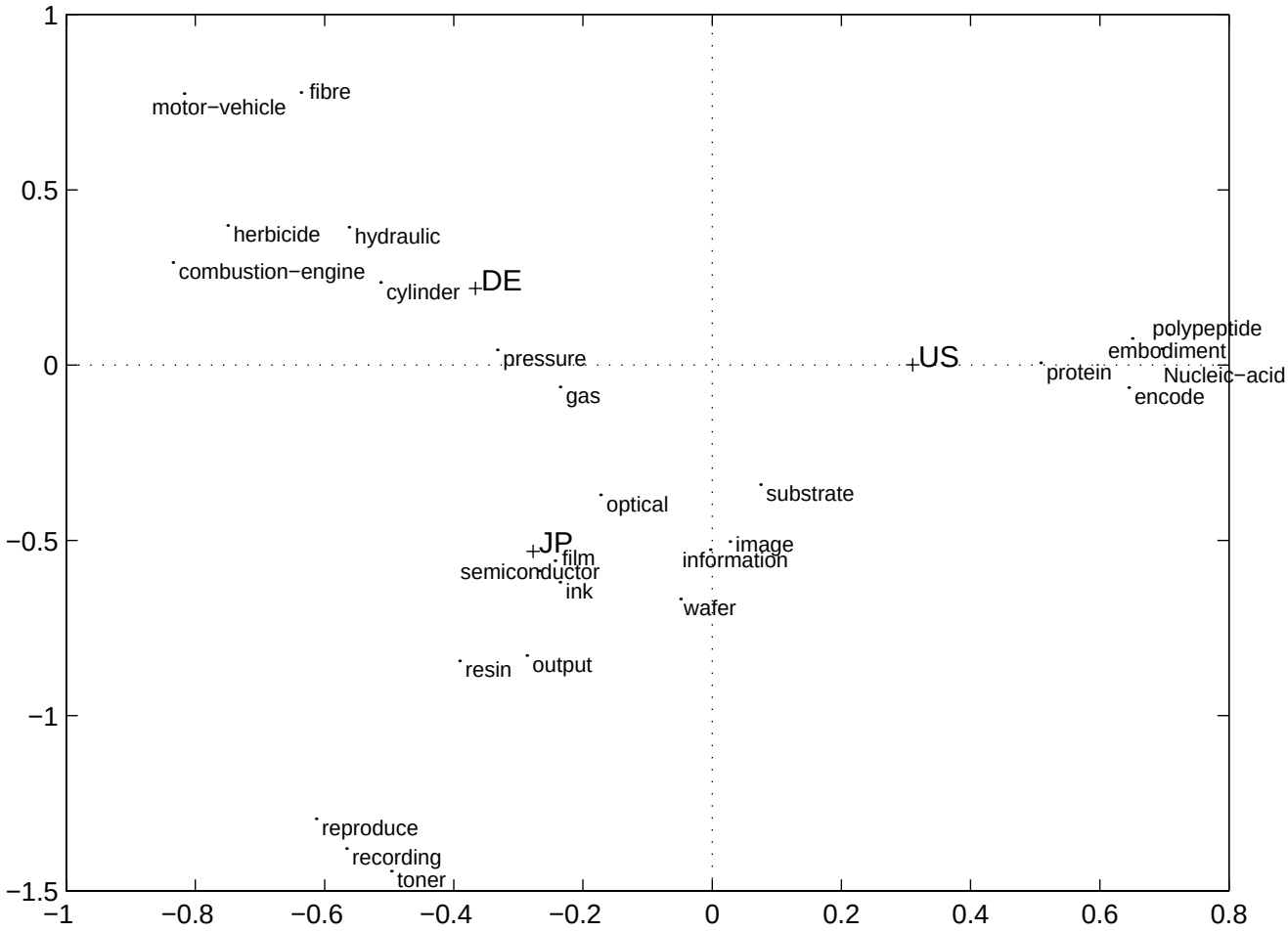


Figure 4: *Projection of words in the first two factorial axis*

of the countries.

Figure 4 illustrates the nature and the position of the lemmas selected for the interpretation of the first two factors. It is striking to notice how efficiently the heuristically selected lemmas seem to contribute to a possible interpretation of the relative positions of the US, JP and DE patents. Indeed, the lemmas "*Nucleic-acid*", "*protein*", "*encode*", ... characteristic for factor 1 that opposes US to both JP and DE, strongly suggest a clear positioning of the US patents in the domain of biotechnologies, while the JP and DE patents remain in more traditional engineering fields. In addition, the opposition on the axis linked with factor 2 between words such as "*reproduce*", "*recording*", "*toner*", "*ink*", "*semiconductor*", "*wafer*", ... on the JP side and words such as "*cylinder*", "*combustion-engine*", "*motor-vehicle*", "*pressure*", ... on the DE side also strongly suggest the relative positions might be interpreted as the indication of the strong involvement of Japan (JP) in innovation within the electronics and printing industry, Germany (DE) being positioned in rather more traditional engineering.

In the specific case of patent data, one additional confirmation technique can be implemented: the use of the IPC codes that are assigned to each of the patents and provide a very synthetic description of its content.

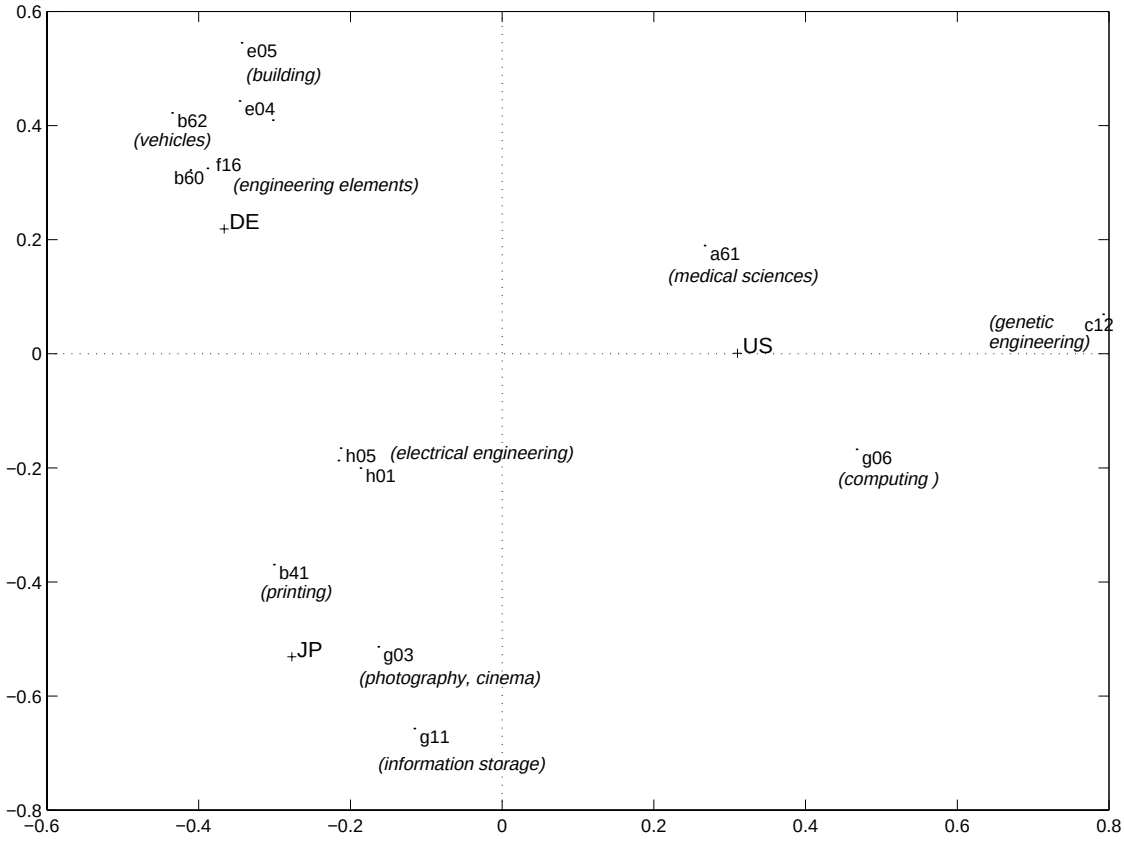


Figure 5: *Projection of IPC codes in the first two factorial axis*

The idea underlying the confirmation method is then the following:

- each of the patents is associated with the (unique) 3 character IPC code derived from the complete IPC assigned in its main IPC code field;
- a new contingency table is computed: in this matrix, the rows are associated with the different IPC codes, the columns with the lemmas used in the previous contingency table and the cell records the number occurrences of a given lemma in patents associated with a given IPC code;
- due to the fact that the columns of the new contingency matrix are the same as the ones that were used to perform the correspondence analysis with the countries, standard transfer formulae can be used to project the rows of the new contingency matrix (i.e. the different IPC codes) as supplementary elements in the factor space produced by the correspondence analysis.

Figure 5 displays the positioning of some of the IPC codes that are best represented in the plane corresponding to the first two factors (as measured by the "$\cos^2$" values), and, as interpretation is concerned, it is interesting to notice that the codes (`a61`, `g06`, `c12`) that mostly explain the position of the US patents in opposition to the JP and DE patents are respectively associated with the domains "*medical sciences*", "*computing*" and "*micro-organisms*", while, on the other axis, codes associated with DE are essentially in the traditional engineering domain (with codes such as the `b` codes representing "*various industrial techniques*") whereas the codes associated with JP indeed concern the fields of "*printing*" and "*information storage*".

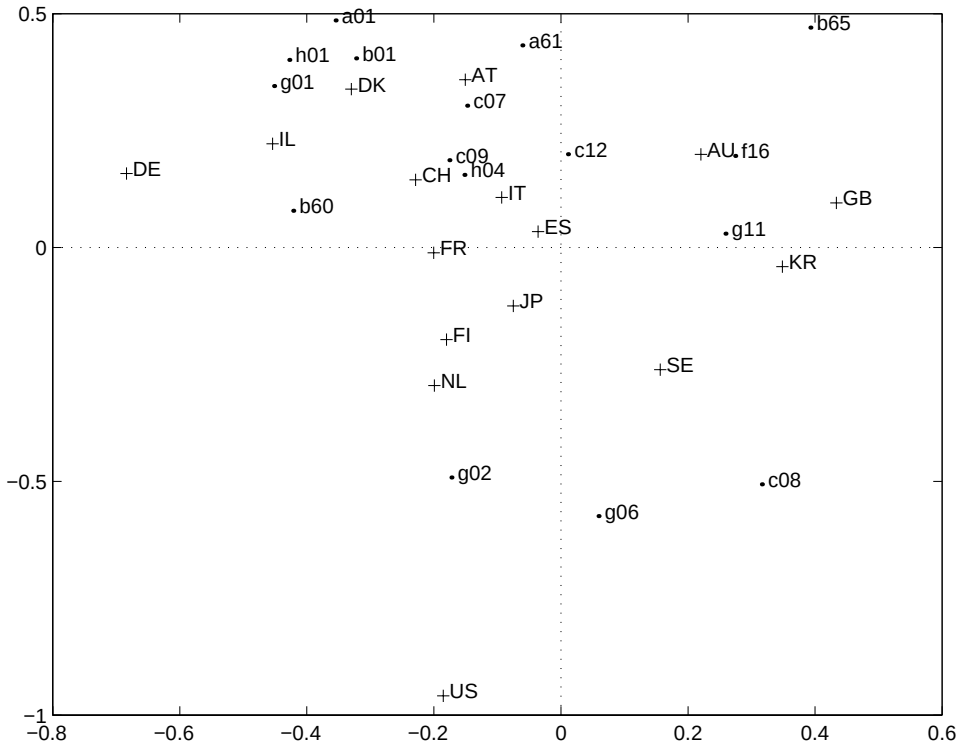


Figure 6: *Projection in the first two factorial axis for Country ⊗ IPC correspondence analysis*

Notice that such results can not be obtained for the direct correspondence analysis of the countries against IPC codes (i.e. simply using these 2 closed categories, not using words as support) as illustrated in figure 6. This shows that performing textual analysis of the content of the patent can bring more information about the technological trends.

**Stability** It is an important aspect of the proposed methodology to provide the user with efficient means to test the stability of the results produced by correspondence analysis. To this end, bootstrap techniques have are used within the STING project. More specifically, to study the stability of the results presented in this contribution, we performed 50 times bootstrap replicates of the 23,750 patents (Davison and Hinkley, 1997). For a few selected countries, an example of their projection in the first two factorial axis for each of the 50 replicates is given in figure 7. It can be seen that for the most contributing countries, the results are very stable (small dispersion of the replicates). One less stable country (IL) has also been displayed for comparison.

## 4. Cluster analysis for patent data

**Description of the method** Since the scope of this paper is to present the general idea of the statistical methodology, we do not provide here an extended description of the methods from a theoretical point of view but rather insist on the main aspects that illustrate the way they work.

Cluster analysis could directly take place after the linguistic preprocessing step. However, due to the specific format of the contingency tables used and in order to increase the robustness of the resulting clusters, a correspondence analysis is first performed. This correspondence analysis is not presented to the user and is performed in a purely automated way since otherwise it would

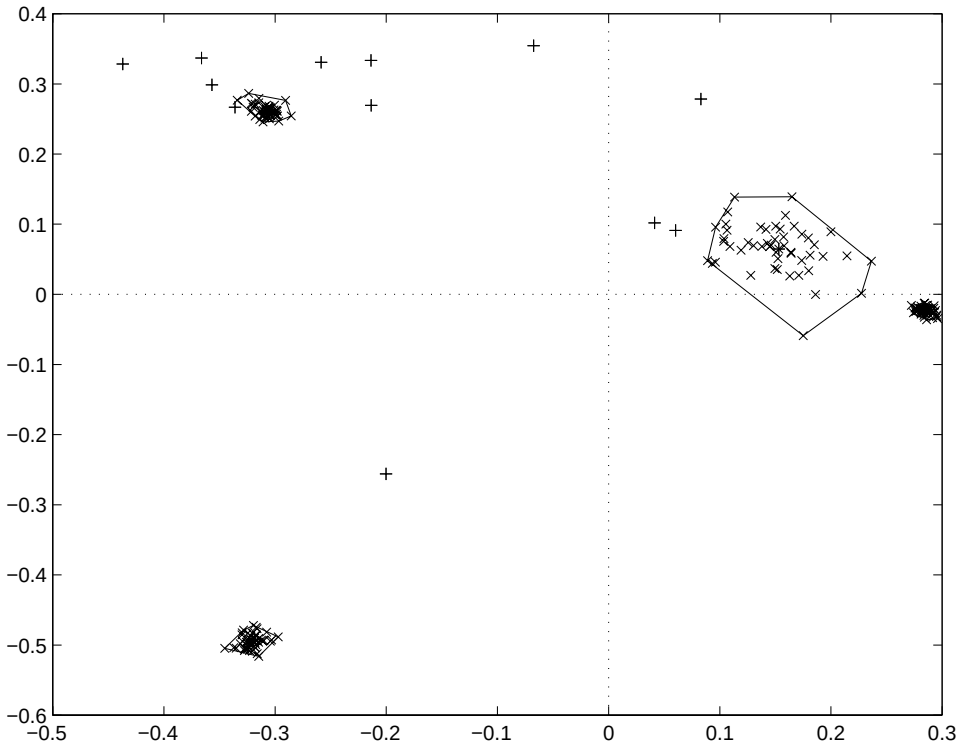


Figure 7: *50 bootstrap replicates for the three most contributing countries (US, DE and JP) in the first two factorial axis*

require from the user the ability to appropriately define its parameters.

As far as cluster analysis itself is concerned, hierarchical clustering is usually performed, apart from the case of too large contingency tables for which the system first performs a k-means cluster analysis followed by a hierarchical clustering applied to the groups created by the k-means procedure.

Hierarchical clustering produces a hierarchy of groups partially nested in one another, starting with a set of elements that are characterized by variables. Grouping is done on the basis of similarities or distances, in our case the $\chi^2$-distance.

As already mentioned previously, two different contingency tables are used.

The first contingency table cross-tabulates words (lemmas) extracted from the titles and abstracts with the IPC codes that describe the patents. The goal is to identify words with similar profile and the resulting word clusters can then be used to describe the different IPC codes.

In addition, as each IPC code is most often related to only one specific patent, all the information associated with the patent can be used for the created clusters. Thus a representation of clusters in terms of codes that belong to each cluster is obtained while additional information such as assignees, inventors can be taken into account.

The second approach is complementary to the first one as the associated contingency table cross-tabulates IPC codes with words (lemmas). The goal is then to identify groups of IPC codes that share common technologies as characterized by the words that describe each patent. again,

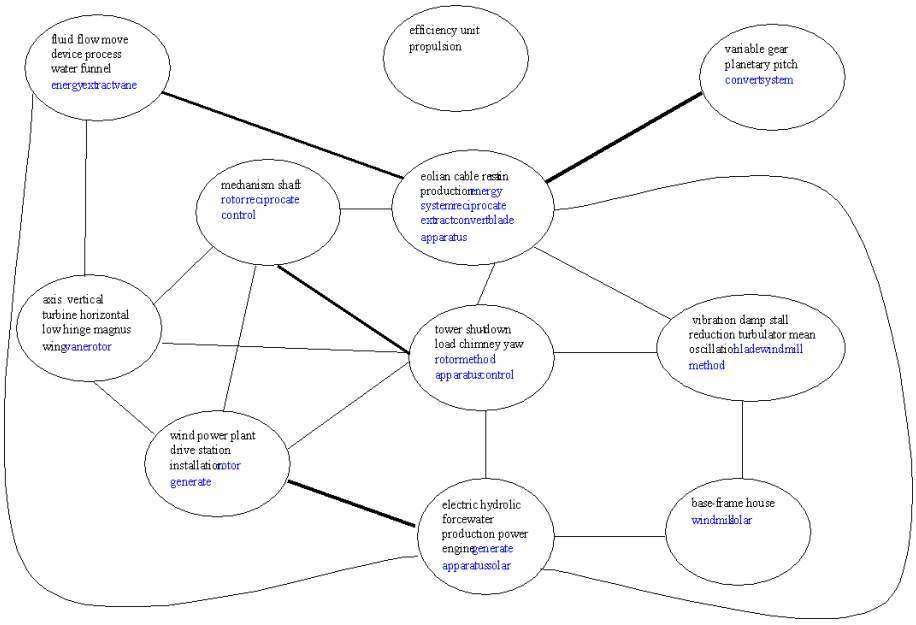


Figure 8: *Relationships between word clusters*

supplementary information associated with the patents referenced by the IPC codes can be used to capture information related not only to specific patents but also to areas or technology.

Furthermore both approaches allow to draw relationships between clusters based on common codes or common words respectively.

**Case study** A test on real patent data was carried out to provide a fist validation of the accuracy of the proposed methodology. The sample data used, again derived from the ESPACE/EPA database (Vol. 2000/006), consisted of all the patents related to wind motors for the years 1995 to 2000. The hierarchical structure adopted in the International Classification System for the IPC codes is used to describe the patents made the initial selection of the required patents quite easy as the 4 character IPC code F03D identifies the subsector corresponding to wind motors (the F descriptor corresponding to the general sector related to motors).

Figure 8 and 9 present the clusters obtained analysis for the two considered contingency tables, as well as the associated relationship maps.

Figure 8 shows the word clusters and their relationships obtained with a contingency table cross-tabulating words and IPC codes, while Figure 9 shows the code clusters and their relationships obtained with a contingency table cross-tabulating IPC codes and words.

The two figures clearly illustrate that the identified relationships are quite complex and that additional information such as inventors, assignees, etc. needs to be taken into account in order to draw synthetic conclusions about area of technology.

## 5. Conclusions

The developed statistical methodology permits to analyze patent data based on multidimensional analysis techniques that make use of all the information describing a patent. This approach

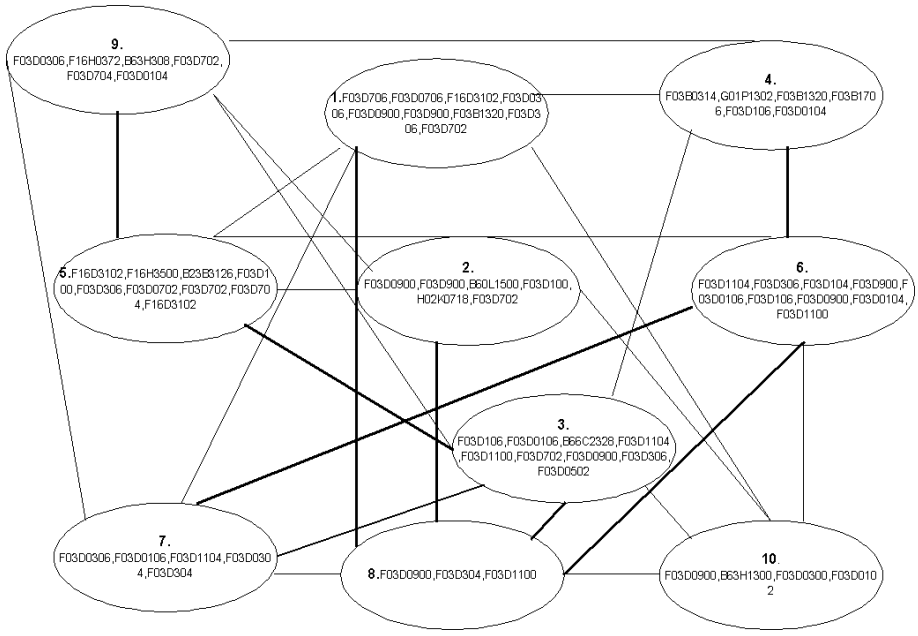


Figure 9: *Relationships between IPC code clusters*

enables to capture information not only at the level of a sector but also perform comparisons at the level of a country or at the level of set of sectors. In addition it permits to identify technological trends and innovation.

Additional research is ongoing to study the impact of the specific distance used between the words and alternative (dis)similarity criteria are currently under investigation to allow applications of cluster analysis techniques that do not require preliminary stabilizing correspondence analysis.

## References

Benzécri J.-P. et al. (1973). *L'Analyse des Données*, volume II : L'Analyse des Correspondances. Dunod.

Comanor W. S. and Scherer F. M. (1969). Patent statistics as a measure of technical change. *Journal of Political Economy*, 77(3):392–398.

Constant P. (1995). *Manuel de développement SYLEX-BASE*. INGÉNIA-LN, Paris, France.

Davison A. C. and Hinkley D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press.

Dou H. (1995). *Veille technologique et compétitivité – L'intelligence économique au service du developpement industriel*. Dunod, Paris.

Guellec D. and van Pottelsberghe B. (1998). New indicators from patent data. In *Proc. of Joint NEST/TIP/GSS Workshop*.

Johnson R. and Wichern D. (1998). *Applied multivariate statistical analysis*. Prentice-Hall, Inc.

Lebart L., Morineau A., and Piron M. (1997). *Statistique exploratoire multidimentionnelle*. Dunod, 2e edition.

Lebart L. Salem A. B. L. (1998). *Exploring Textual Data*, volume 4. Kluwer Academic Publishers.

Narin F. (1995). Patents as indicators for the evaluation of industrial research output. *Scientometrics*, 34(4):489–496.