Multivariate approach to classify research institutes according to their outputs: the case of the CSIC's institutes

José Luis Ortega[1], Elena López-Romero and Inés Fernández

R&D Unit, CSIC, Serrano, 113, 28006 Madrid. Spain

e-mail: {jortega, elena.lopez, ines.fernandez}@orgc.csic.es

Abstract

This paper attempt*s* to build a classification model according to the research products created by those institutes and hence to design specific evaluation processes. Several scientific input/output indicators belonging to 109 research institutes from the Spanish National Research Council (CSIC) were selected. A multidimensional approach was proposed to resume these indicators in various components. A clustering analysis was used to classify the institutes according to their scores with those components (principal component analysis). Moreover, the validity of the *a priori* classification was tested and the most discriminant variables were detected (linear discriminant analysis). Results show that there are three types of institutes according to their research outputs: Humanistic, Scientific and Technological. It is argue that these differences oblige to design more precise assessment exercises which focus on the particular results of each type of institute. We conclude that this method permits to build more precise research assessment exercises which consider the varied nature of the scientific activity.

[1] Corresponding author

Article accepted to be published in *Journal of Informetrics*

Keywords: Scientometrics, Principal Component Analysis, Linear Discriminant Analysis, research centres classification

Introduction

Research activity is affected by multiple variables that influence the success of their results. The disposal of human and economic resources determines the quantity and quality of the research products. Moreover, the different types of results (publication, patents, thesis, etc.) are examples of different type of researches. Thus, applied sciences create patents which contain detailed description of inventions, while the humanities need books which permit the expression of textual criticism and speculative and discursive reasoning. Disciplines with a rapid obsolescence tend to use fast communication media such as the proceeding papers (Line, 1970). This involves assessing the scientific research according to multiple output indicators which express the different activities of each research discipline (Martin, 1996).

The assessment exercises are based on a reward system in which the results obtained by a research unit are valuated for a research institution through a qualitative (peer review), quantitative (scientometric assessment) or mixed approach (Shapira and Kuhlmann, 2003). However, one of the most important challenges of a research evaluation system is to value or quantify the importance of each scientific result in the context of the multidisciplinary organizations. In European research councils, such as CNRS, CNR or CSIC there are specialized institutes in all spheres of knowledge whose research outputs considerably differ one from another.

We think that before measuring the value of each output, it is necessary to identify the research institutes by their common research results, and then to classify them in different profiles which allow the application of specific assessment exercises. This

paper attempts to build a classification model according to the research products created by those institutes and in this way to design specific evaluation processes.

Related Research

Several papers have addressed the statistical and automatic classification of research units (scholars, institutes, universities, etc.) using R&D indicators. One of the first works was developed by Giese (1990). He used principal component and discriminant analysis to rank German universities according to input indicators (staff, funding). Discriminant analysis was also used by Coccia (2004; 2005) to classify research institutes in high and low performance. He analysed a set of different indicators such as training courses, publications, staff, etc. Ramani (2002) classified the Indian biotech firms according to their expenditure, publications, staff and other variables, using principal component analysis. Tagarelli et al. (2004) proposed data mining techniques for classifying research centres according to their publications, European projects and patents. However, their results are more focused on the reliability of the model rather than on the validity of the obtained classes. Tikoria et al. (2009) used analytics hierarchy process to measure the performance of the R&D organizations in India.

Other studies have used partial approaches based mainly on bibliometric data in order to classify journals (Schubert and Braun, 1996), articles in thematic categories (Glänzel and Schubert, 2003), authors by their publications and citations (Zhou et al., 2007; Harris and Kaine, 1994), article-related indicators according to the features they measure (Bollen et al., 2009), research institutes by their publications (Chen and Liu, 2006; Thijs and Glänzel, 2008) and to build thematic maps from bibliographic data (Polanco et al., 1998).

Objectives

This paper attempts to build a classification model according to the research outputs produced by the 109 research institutes of the CSIC and in this way to design specific evaluation processes for each group. Methodologically, we intend to answer the following questions:

- Is it possible to implement a statistical method in order to classify research centres according to R&D indicators?
- How many classes would be found and according to what?
- What indicators would characterize those classes?

Methods

Data

Several indicators were used to characterize each research institute. These data were obtained from an internal assessment exercise, which quantitatively measures the achievement of research objectives previously defined each year. This study contains the accumulated results from 2005 to 2008. We have selected this time period because there are fluctuations in the research activity of the institutes each year due to the variability of their resources (economic, human, etc.). This process measured the research activity of 109 research institutes through 13 indicators grouped in four thematic blocks (see Table 1).

Blocks	*Indicator*	*Definition*
Funding	Projects	Amount of funds obtained by their participation in competitive research projects
Scientific Production	ISI publications	Number of papers published in journal indexed in the ISI-Thomson database
	International non-ISI publications	Number of papers published in international journals which are not indexed in the ISI-Thomson database. It also includes international proceeding papers and book chapters
	National non-ISI publications	Number of papers published in national journals which are not indexed in the ISI-Thomson database. It also includes national proceeding papers and book chapters
	Books	Number of books edited or written
Technological Production	Spin-offs	Number of new firms established
	Licensed patents	Number of licensed patents
	Private R&D contracts	Amount of funds obtained by private research contract with a company or private foundation
	Public R&D contracts	Amount of funds obtained by public research contract with an administration or public foundation
	International patent application	Number of Patent Applications to the European Patent Office (EPO). World Intellectually Property Office (WIPO)and to offices of foreign countries
	National patent application	Number of Patent Application to the Spanish Patent Office (OEPM)
Training	Thesis	Number of PhD Thesis directed or co-directed by a CSIC researcher
	Curses	Number of teaching hours in graduate, doctoral and specialized courses

Table 1. Classification and definition of the 13 indicators used in the assessment exercise.

The number of institutes which participate in the exercise changes each year because there are institutes that disappear, merge or start. So, we made a snapshot of the CSIC's institutes in 2005. The Appendix I lists the acronyms of the institutes beside their *a priori* and *a posteriori* classification, and their membership probability. The name in brackets is the former name of the institute in 2005, while split institutes after to 2005 were merged in their original institute. Due to internal restrictions, we cannot publish the full name of the institutes. However, we put in bracket the research issues of those institutes.

The distribution of the research resources (incomes by projects and contracts) and results (publications, patents, books, etc.) do not follow a Gaussian distribution but a power law (Katz, 1999). Thus, few institutes produce the large majority of the CSIC results. This is due to a size effect, in which the largest institutes produce more results than de other ones because they also gather more resources. Then we divided their outputs by the total number of scientific and technical personnel –it does not include administrative staff. These new variables are now Gaussian and they do not show a size effect.

Statistics

Three statistical techniques were used to create a classification method of research centres. The Principal Component Analysis was firstly used to extract the principal components that resume the information of the 13 indicators. Next, a cluster analysis leads to *a priori* classify the institutes according to their scores in the obtained components. Finally, the Linear Discriminant Analysis was used to test the *a priori* classification done by the clustering technique and to identify which variables are the most important to differentiate between groups. This procedure was previously used by

Giese (1990) and Coccia (2004; 2005), although they do not use the clustering technique to previously classify the observations.

Principal Component Analysis

The Principal Component Analysis (PCA) (Pearson, 1901; Hotelling, 1933) is a multivariable technique related with the factor analysis. The aim of the PCA is to reduce the dimension of p variables to a set of new variables (principal components) which contain the highest amount of information from the previous variables. It is desirable that all variables are well correlated between them, because this is symptomatic of redundant information and therefore a lower number of new variables (components) will be necessary to explain the model. These components are uncorrelated between them, because the fist one has the highest amount of information, the second one has the information that the previous does not contain and so on.

These components are interpreted according to their correlation with the previous variables, because they contain part of the information of the original variables. Thus, these components allow us to plot the observations in a new reduced space and to see how the variables are related with the institutes. To simplify the component structure and therefore to make its interpretation easier and more reliable, it is usual to apply rotations to the components, Varimax, which was developed by Kaiser (1958), is the most popular rotation method; because it makes that each component represents only a small number of variables.

Agglomerative Hierarchical Clustering

The Agglomerative Hierarchical Clustering (AHC) (Sneath and Sokal, 1973) is a method of cluster analysis which intents to build a hierarchy of groups. The

agglomerative one starts with a similarity matrix in which each element shows a similarity degree regarding the other ones. Then a linkage method is successively jointing elements, creating a bunch of clusters called Dendrogram. A cut-off point is selected to identify the most important clusters from the tree plot.

Linear Discriminant Analysis

The Linear Discriminant Analysis (LDA) (Fisher, 1936) is a statistical method that comes from the multiple regression analysis and its principal objective is to obtain several classification functions that classify an observation in a set. The LDA starts from a set of continuous variables, which are grouped according to a categorical variable (the classification criterion). It selects the variables that more clearly separate those groups and creates several linear classification functions that reclassify the observations. The method compares the new classification with the previous one as a way to test the model. Finally, the LDA enable us to assign new observations to the groups through the classification functions.

Software

We used several statistical software packages to develop different statistical methods and to obtain several graphical outputs. SPSS 17 was used to calculate the LDA, while XLStat 2008 was used to calculate the PCA. XLStat was mainly used due to its graphical outputs, much better than the SPSS ones. SPSS was used to develop calculi not included in the XLStat 2008 software, i. e. the insertion method in the LDA. XLStat was also used to calculate the clustering analysis and to visualize the tree plot.

Results

Principal Component Analysis

As we said before, PCA was firstly used to extract the principal components and to *a priori* classify the institutes according to their scores in those components. Three components were obtained with a variance of 30.67% to the first one, 27.08% to the second one and 13.24% to the third one, being a cumulate variance of 70.99%. *Public R&D Contracts* and *Spin-offs* variables were rejected from the model due to their low correlation with the three first components. They highly correlated with the F4 and F5, respectively. However, to also consider these factors would produce an excessively complex model.

Variables	F1 (Technological)	F2 (Humanistic)	F3 (Scientific)
Licensed patents	**.757**	.030	-.369
Private R&D contracts	**.653**	.234	-.173
Inter. patent applications	**.880**	.081	-.239
Nat. patent applications	**.877**	.111	-.188
Inter. Non-ISI publications	.306	**.718**	-.078
Nat. Non-ISI publications	-.181	**.908**	.124
Books	-.328	**.848**	-.079
Thesis	.231	**.590**	.509
Courses	-.247	**.691**	-.143
Projects	.539	-.139	**.562**
ISI publications	.438	-.026	**.758**

Table 2. Correlation between variables and components (in bold r>.55)

Table 2 shows the Pearson's correlation between the variables and the obtained principal components. These correlations allow the interpretation of the meaning of each component according to their relationship with the original variables. Thus, the first component (F1) correlates with *Licensed patents, Private R&D contracts, International patent applications* and *National patent applications*. These variables are

related with the technology transfer, so we tagged this component as "Technological". The second one (F2) is correlated with *International no-ISI publications, National no-ISI publications, Book, Thesis* and *Courses*. This component was defined as "Humanistic" because the no-ISI publications (book chapters, conference papers) and books are common outputs of the Humanities. Finally, the third component (F3) correlates with *Projects* and *ISI publications,* so we decided to name this component "Scientific" because both indicators are characteristic of the hard sciences.

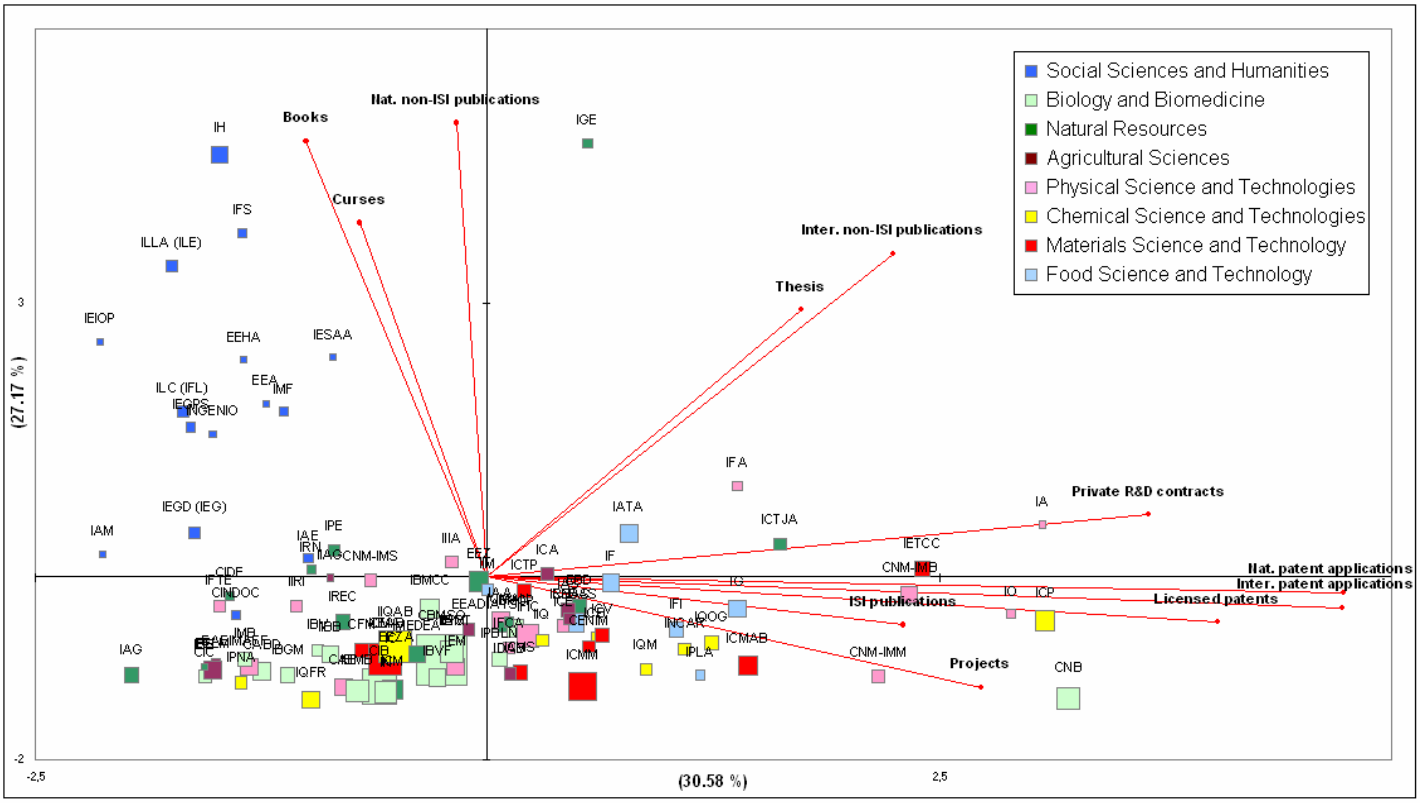

Figure 1. PCA map with the two main components, variables and observations.

Variance (57.75%) with Varimax rotation

Figure 1 shows the bi-dimensional projection of the variables (lines) and observations (dots) according to the two principal components. Unfortunately, XLStat does not generate 3D plots, so we can only represent the first two components, the

"Technological" and the "Humanistic" ones. The colours represent the Scientific and Technical Area of each institute –the CSIC's institutes are organized in eight Scientific and Technical Areas–, while the size is proportional to their scientific staff. Two institutes were removed from the graph because their scores were very high and they were located too far away. These institutes are the IAI (Industrial Automation) with respect to the horizontal axe and the IHCD (History of Science and Information Sciences) with respect to the vertical one.

Figure 1 lets us to appreciate how the research institutes from the Social Sciences and Humanities Area (blue) are displayed vertically along the "Humanistic" component, showing positive scores with that axis. These "Humanistic" institutes focus their activity on publishing books and non-ISI papers and teaching courses. The institutes with the largest scores in that component are the IHCD and the IH (History). The horizontal axis represents the "Technological" component. Institutes with positive scores in that axis are institutes with a strong technological activity such as obtaining private research contracts and patenting new inventions. For instance, the abovementioned IAI and the CNB (Biotechnology) are the institutes with the highest scores in that axis.

The picture also shows how the different S&T Areas are related with the components. We have already seen that all the institutes of the Social Sciences and Humanities Area have positive scores with the second component, with the exception of the IEDCYT (Information Science). Notice the high "humanistic" profile of the IGE (Geology) despite the fact that it belongs to the Natural Resources Area. We also observe that the great majority of the Biology and Biomedicine institutes have negative scores in both components, whereas they show high positive scores in the "Scientific" component. The only exception is the CNB which has an important technological profile. Finally, it is

interesting to note that all the institutes of the Food Science and Technology Area are positively located on the “Technological” axe.

Agglomerative Hierarchical Clustering

Once obtained the principal components and named them according to their correlations, each institute was *a priori* classified through the AHC. This mix procedure is widely used in different disciplines such as Chemistry (Michel and Jeandenans, 1993), Computing (Lin et al., 2007) and Medicine (Modlin et al., 2009). A similarity matrix was built from the scores of each institute in the PCA. This lets us to group institutes that are closer to a component or other one. Cosine similarity measure was used because it is sounder to non-parametric variables. The linkage method used was the Average Linkage. This method is computed as the average distance between objects from the first cluster and objects from the second cluster. The resulting dendrogram (Figure 2) shows three defined groups that fit with the three PCA components. The truncation was automatically set-up by the statistical software which shows a high cut-off (similarity>0.25) and the identified clusters show solid differences. Although the dendogram shows more sets at a higher level, these are not differentiated by the LDA because the proportion of misclassified increases and it does not found discriminant variables (Martinez and Kak, 2001). This is because some of these groups are set up by multiples fuzzy characteristics that the LDA does not achieve to differentiate. Thus, the institutes in the green cluster are observations that have high scores with the “Technological” component, while the institutes of the pink cluster are related to the “Scientific” one and the brown set with the “Humanistic” one. This allows us to class *a priori* each institute in one of the three groups (Technological, Scientific and Humanistic).

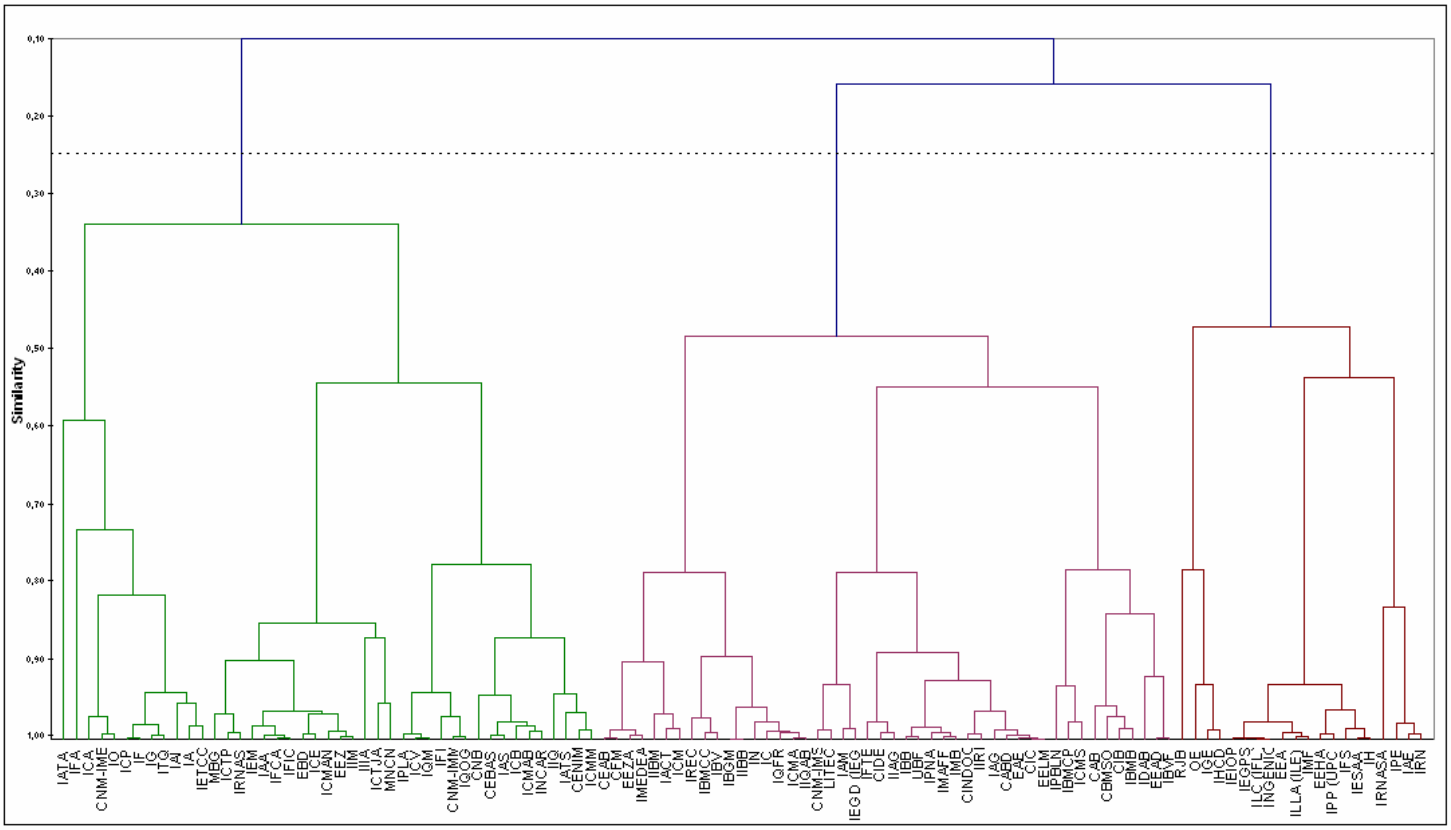

Figure 2. Dendrogram from the Hierarchical Clustering Analysis (HCA). Similarity: cosine; Method: Average link

Linear Discriminant Analysis

Next, we used the LDA to test the *a priori* classification, to observe the most discriminant variables and to obtain the discriminate functions as well. Since there are three groups to test, the model found two discriminant functions to separate the groups. We have used a stepwise discriminant method to select only those variables that have discriminant power, rejecting the redundant variables or with a low discriminant factor. The Wilks' Lambda method was used because is the most extended and it minimizes the Wilks' Lambda value. This is because the lower is this value the higher is the discriminant power of the functions. We have used a restrictive model in order to only

select the most discriminant variables. Thus, we have reduced the significance at the *p-value*=.01.

	Function	
	1	2
ISI publications	.631	.622
Inter. non-ISI publications	-.474	.418
Books	-.483	.502
Nat. patent applications	.693	.228

Table 3. Coefficients of the discriminant functions (standardized)

Table 3 shows the two functions and their coefficients. The stepwise method detected only four variables with the highest discriminant ability. Those variables are *ISI publications, International non-ISI publications, Books* and *National patent applications*. The first function distinguishes between Humanistic and non-Humanistic institutes because the humanistic variables (*International non-ISI publications* and *Books)* have negative coefficients. While, the second one separates the Scientific and Technological institutes because the *National patent applications* coefficient is the lowest and the *ISI publications* coefficient is the largest positive one. The Wilks' Lambda value of the first function (λ=.199) is rather low, so the differences between the Humanistic and non-Humanistic institutes are strong. However the second one (λ=.569) is larger, causing less clear differences between the Scientific and Technological groups.

	Predicted Group Membership			
Class code	Humanistic	Scientific	Technological	Total
Humanistic	16	4	0	20
Scientific	0	41	5	46
Technological	0	4	39	43
Total	16	49	44	109

Table 4. Reclassification of the institutes according to the discriminant functions

Table 4 shows the predicted classification from the *a priori* one. The LDA found that the 86.2% of cross-validated grouped cases are correctly classified. This lets us to validate the model because the ratio of classification is rather high. "Technological" was the group with most cases correctly classified (90.7%), while the "Humanistic" group has the lowest percentage of correctly classified cases (70%).

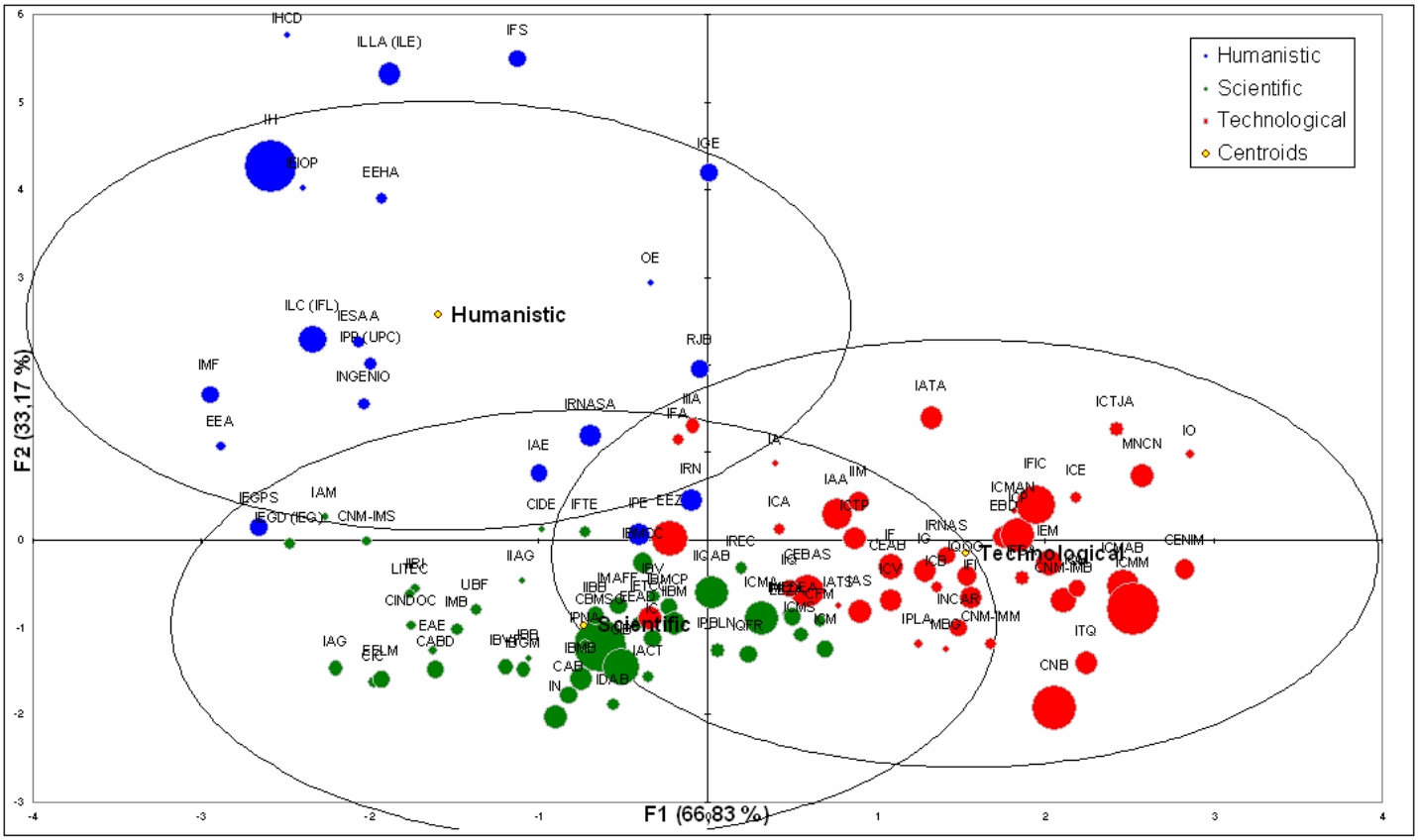

Figure 3. Plot of the research institutes according to the LDA classification

	Humanities		Scientific		Technological		Total
	institutes	%	institutes	%	institutes	%	
Social Sciences and Humanities	12	70.6	5	29.4	0		17
Biology and Biomedicine	0		17	94.4	1	5.6	18
Natural Resources	1	5	11	55.0	8	40	20
Agricultural Sciences	0		5	55.6	4	44.4	9
Physical Science and Technologies	1	5.3	7	36.8	11	57.9	19
Chemical Science and Technologies	0		4	40	6	60.0	10
Materials Science and Technology	0		2	22.2	7	77.8	9
Food Science and Technology	0		0		7	100	7
Total	14	12.8	51	46.8	44	40.4	109

Table 5. Percentage of classified institutes by Scientific and Technical Areas

Figure 3 shows the graphical representation of the three groups detected by the LDA and the *a posteriori* classification of the research institutes. Table 5 shows the percentage of research institutes in each class by S&T Areas. A large percentage of Social Sciences and Humanities institutes are located in the "Humanistic" class (70.6%), while the remaining 29.4% are located in the "Scientific" group. These five institutes work in Economics and Information Science which are research areas closer to the "Scientific" pattern than to the "Humanistic" one because they produce a large proportion of ISI publications. The Areas with the highest proportion of "Scientific" institutes are Biology and Biomedicine (94.4%) and Agricultural Sciences (55.6%), while the largest number of "Technological" institutes come from Food Science (100%) and Materials Science and Technology (77.8%). We also observe that there are Areas that share out their institutes between "Scientific" and "Technological" groups. This is the case of Agricultural Sciences (55.6%; 44.4%), Chemical Science and Technologies (40%; 60%) and Natural Resources (55%; 40%). It is interesting to notice that the 5.3% of Physical Science and Technologies institutes (1 institute) is classified as "Humanistic". This is an astrophysical observatory which has an above average of non-ISI publications, mainly articles published in their own journal.

Discussion

The main objective of this work is to present a statistical classification method which allows to describe the principal features of the research institutes and to group them in solid sets. The obtained results reinforce the suitability of the method because it detects three differentiated classes: Humanistic, Scientific and Technological. These groups were corroborated by the PCA, founding three components; by the AHC, showing three sets from those components; and by the LDA, testing those groups with an 86% of

correct classified and founding their respective discriminant functions. It is interesting to note that these institutes are classified in those classes according to their scientific outputs, while their inputs do not make possible to differentiate them. This allows us to argue that these institutes may share the same type of funding (projects, public or private contracts, etc.), but they produce differentiated products. Although the PCA and LDA have been profusely used in Scientometrics (Giese, 1990; Coccia, 2004; 2005), they were used to rank research units detecting groups of high or low performance. However, this work does not expect to present a new ranking method but to evidence that there are research institutes which produce different outputs and then they cannot be compared or ranked together. They have to be assessed in a separated way, according to their research outputs and their particular research performance. Moreover, we consider that some research rankings make mistake comparing research units without distinguish between technological institutions oriented to patent production or humanistic institutions centred in book edition. Furthermore, these rankings are just based on bibliographical databases (Thomson-ISI, Scopus) which use indicators related to a unique scientific result: research articles. We think that to base the research assessment on only published papers could show an unrealistic view of such complex activity (Van Raan, 2005).

This argument fits with Martin (1996) who defends that the research evaluation must be done from a multidimensional scope which assesses the multiple results that the research activity is able to produce. Thus, if we just consider ISI publications as only indicator, the best valued institutes are those that only publish ISI papers, while institutes focused on patenting or publishing books are underestimated. For instance, this is the case of the ITQ, a prestigious chemical institute which licenses the 18.5% of the CSIC's patents, but it only contributes the .9% of the ISI papers; or the case of the

IH with the 10.1% of books but a .3% of ISI publications. These examples show that the bibliometric indicators are good if they describe research areas where the scientific publication is the principal output of their activity (Giese, 1990), but in areas where the publications are in no way the only product of research, bibliometric indicators have to be carefully used (Skoie, 1999). Our results allow us to claim that those indicators can only be used in evaluation processes if they come with other non Thomson-ISI or Scopus based indicators such as published books and applied, licensed or granted patents.

Results allow us to improve our evaluation exercise as well:

- Defining three types of evaluation models for each group of institutes.
- Distinguishing between proceeding papers and other non-ISI publications, avoiding the misclassification of the Computing institutes in "Humanistic" institutes
- Putting more attention on the output indicators than the input ones.

Conclusions

The obtained results enable us to claim that the principal component analysis is a suitable tool to reduce R&D activity indicators to different components and the agglomerative hierarchical clustering a reliable classification method of observations according to the principal components. We can also state that the discriminant analysis has been a proper method to validate the *a priori* classification and to identify the most discriminant variables that make possible to classify the research institutes. These statistics allow us to build a robust methodology to characterize research units according to their research inputs/outputs. Regardless of the results, the principal advantage of this method is that does not allow to classify research institutes but also

other research elements such as researchers, organizations and countries, being an important tool to the improvement of research units assessment.

Three classes of research institutes have been found: Humanistic, Scientific and Technological. These classes are defined from the characteristic research products of each institute. Thus, a "Scientific" institute is one which mainly publishes ISI papers, a "Humanistic" is one that mainly publishes books and non-ISI publications and the "Technological" ones are those which produce patent applications. Obviously, the importance of these results are not that certain institutes produce particular outputs but that this method has identified three classes and hence makes possible to design different evaluation models focused on their principal research outputs. We conclude that this method permits to build more precise research assessment exercises which consider the varied nature of the scientific activity.

Acknowledgments

We would like to thank to the anonymous referee for helful comments on methodology and the friendly discussion about the results.

References

Shapira, P., & Kuhlmann, S. (2003). Learning for science and technology policy evaluation. In: P. Shapira & S. Kuhlmann (eds.). Learning from science and technology policy evaluation: experiences from the United States and Europe. Edward Elgar: Cheltenham

Bollen, J., Van De Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures, PLoS One, 4(6), e6022. http://www.plosone.org/article/info:doi/10.1371/journal.pone.0006022.

Chen, Y., & Liu, N. C. (2006). A first approach to the classification of the top 500 world universities by their disciplinary characteristics using scientometrics, Scientometrics, 68(1): 135-150

Coccia. M. (2004). Models for measuring the research performance and identifying the productivity of public research institutes, R&D Management, **34**(3): 267–280

Coccia. M. (2005). Scientometric model for the assessment of the scientific research performance within the public institutes, Scientometrics, **65**(3): 97–311.

Cunningham, S. J., & Bocock. D. (1995). Obsolescence of computing literature, Scientometrics, 34(2): 255-262

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems, Annals of Eugenics, 7: 179–188.

Giese, E. (1990). Ranking of Universities in the FRG, Scientometrics, 19(5-6): 363-375

Glänzel, W., & Schubert. A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes, Scientometrics, 56(3):357-367

Harris, G., & Kaine, G. (1994). The determinants of research performance: a study of Australian university economists, Higher Education, 27: 191-201.

Hotelling, H. (1933). Analysis of a complex of statistical variables into Principal Components, Journal of Educational Psychology, 24: 417-520.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis, Psychometrika, 23: 187-200

Katz, J. S. (1999). The Self-similar science system, Research Policy, 28(5): 501-517

Lin, J. S., Tien, S. W., Chen, T. S., Kao, Y. H., Lin, C. C., & Chiu, Y. H. (2007). Referential hierarchical clustering algorithm based upon principal component analysis and genetic algorithm. In: A. Xu, H. Zhu, S. Y. Chen, et al. (Eds.) *Proceedings of the 6th Conference on WSEAS International Conference on Applied Computer Science* Stevens Point, Wisconsin, USA, 138-142.

Line, M. B. (1970). The 'half-life' of periodical literature: apparent and real obsolescence, Journal of Documentation, 26(1): 46-54

Martin, B. R. (1996). The use of multiple indicators in the assessment of basic research, Scientometrics, 36(3): 343-362

Martinez, A. M., & Kak, A. C. (2001). *PCA versus LDA,* IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(2): 228–233

Michel, A. G., Jeandenans, C. (1993). Multiconformational investigations of polypeptidic structures, using clustering methods and principal components analysis, Computers & Chemistry, 17(1): 49-59

Modlin, I.M., Gustafsson, B.I., Drozdov, I., Nadler, B., Pfranger, R., & Kidd, M. (2009). Principal Component Analysis, Hierarchical Clustering, and Decision Tree Assessment of Plasma mRNA and Hormone Levels as an Early Detection Strategy for Small Intestinal Neuroendocrine (Carcinoid) Tumors. Annals of Surgical Oncology, 16(2):487-498

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space, Philosophical Magazine, Series 6, 2 (11): 559-572.

Polanco, X., François, C., & Keim, J. P. (1998). Artificial neural network technology for the classification and cartography of scientific and technical information, Scientometrics, 41(1–2): 69–82.

Schubert, A., & Braun, T. (1996). Cross-field normalization of scientometric indicators, Scientometrics, 36(3): 311-324

Shapira, P., & Kuhlmann, S. (2003). Learning for science and technology policy evaluation. In: P. Shapira & S. Kuhlmann (eds.). Learning from science and technology policy evaluation: experiences from the United States and Europe. Edward Elgar: Cheltenham

Sneath, P. H. A., & Sokal, R. R. (1973). Numerical Taxonomy. Freeman: San Francisco

Skoie, H. (1999). Bibliometrics—some warnings from the north, Scientometrics, 45(3):433-437

Tagarelli, A., Trubitsyna, I., & Greco, S. (2004). Combining linear programming and clustering techniques for the classification of research centers, The European Journal on Artificial Intelligence, AI Communications, 17(3):111-122

Thijs, B., & Glänzel, W. (2008). A structural analysis of publication profiles for the classification of European research institutes, Scientometrics, 74(2): 223-236

Tikoria, J., Banwet, D. K., & Deshmukh, S. G. (2009). Performance measurement of national R&D organisations using analytic hierarchy process: a case of India, International Journal of Innovation and Regional Development, 1(3): 276-300

Zhou, F., Guo, H. C., Ho, Y. H., & Wu, C. Z. (2007). Scientometric analysis of geostatistics using multivariate methods, Scientometrics, 73(3): 265-279

Van Raan, A. F. J., (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods, Scientometrics, 62(1): 133-143

Appendix 1. Institutes name and acronym, *a priori* and *a posteriori* classification and membership probabilities.

Observation	A priori	A posteriori	*Pr(Human.)*	*Pr(Scient.)*	*Pr(Tech.)*
CAB	Scientific	Scientific	*0.009*	*0.884*	*0.106*
CABD	Scientific	Scientific	*0.003*	*0.987*	*0.011*
CBMSO	Scientific	Scientific	*0.001*	*0.946*	*0.053*
CEAB	Scientific	Technological	*0.000*	*0.156*	*0.844*
CEBAS	Technological	Technological	*0.001*	*0.455*	*0.544*
CENIM	Technological	Technological	*0.000*	*0.005*	*0.995*
CFM	Scientific	Technological	*0.001*	*0.459*	*0.540*
CIB	Scientific	Scientific	*0.002*	*0.913*	*0.085*
CIC	Scientific	Scientific	*0.006*	*0.986*	*0.008*
CIDE	Scientific	Scientific	*0.014*	*0.909*	*0.077*
IEDCYT (CINDOC)	Scientific	Scientific	*0.045*	*0.937*	*0.018*
CNB	Technological	Technological	*0.000*	*0.202*	*0.798*
CNM-IMB	Technological	Technological	*0.000*	*0.045*	*0.955*
CNM-IMM	Technological	Technological	*0.000*	*0.108*	*0.892*
CNM-IMS	Scientific	Scientific	*0.391*	*0.602*	*0.007*
EAE	Scientific	Scientific	*0.004*	*0.979*	*0.017*
EBD	Technological	Technological	*0.000*	*0.079*	*0.920*
EEA	Humanistic	Humanistic	*0.970*	*0.030*	*0.000*
EEAD	Scientific	Scientific	*0.004*	*0.818*	*0.178*
EEHA	Humanistic	Humanistic	*0.996*	*0.004*	*0.000*
EELM	Scientific	Scientific	*0.006*	*0.985*	*0.009*
EEZ	Technological	Scientific	*0.015*	*0.814*	*0.172*
EEZA	Scientific	Scientific	*0.003*	*0.560*	*0.437*
IA	Technological	Technological	*0.013*	*0.289*	*0.698*
IAA	Technological	Technological	*0.030*	*0.206*	*0.764*
IAE	Humanistic	Scientific	*0.011*	*0.951*	*0.037*
IAG	Scientific	Scientific	*0.011*	*0.983*	*0.006*
IAI	Technological	Technological	*0.000*	*0.000*	*1.000*
IAM	Scientific	Scientific	*0.285*	*0.712*	*0.003*
IAS	Technological	Technological	*0.001*	*0.235*	*0.765*
IATA	Technological	Technological	*0.001*	*0.097*	*0.902*
IACT	Scientific	Scientific	*0.014*	*0.862*	*0.124*
IATS	Technological	Technological	*0.000*	*0.273*	*0.727*
IBB	Scientific	Scientific	*0.004*	*0.919*	*0.077*
IBGM	Scientific	Scientific	*0.002*	*0.978*	*0.020*
IBMB	Scientific	Scientific	*0.001*	*0.945*	*0.053*
IBMCC	Scientific	Scientific	*0.002*	*0.912*	*0.086*
IBMCP	Scientific	Scientific	*0.001*	*0.876*	*0.124*
IBV	Scientific	Scientific	*0.002*	*0.855*	*0.143*
IBVF	Scientific	Scientific	*0.003*	*0.975*	*0.022*
IC	Scientific	Scientific	*0.002*	*0.881*	*0.116*
ICA	Technological	Technological	*0.003*	*0.347*	*0.651*
ICB	Technological	Technological	*0.000*	*0.193*	*0.807*

ICE	Technological	Technological	*0.001*	*0.011*	*0.988*
ICM	Scientific	Technological	*0.001*	*0.445*	*0.554*
ICMA	Scientific	Scientific	*0.001*	*0.558*	*0.442*
ICMAB	Technological	Technological	*0.000*	*0.017*	*0.983*
ICMAN	Technological	Technological	*0.000*	*0.107*	*0.892*
ICMM	Technological	Technological	*0.000*	*0.014*	*0.986*
ICMS	Scientific	Technological	*0.000*	*0.492*	*0.507*
ICP	Technological	Technological	*0.000*	*0.035*	*0.965*
ICTJA	Technological	Technological	*0.001*	*0.029*	*0.970*
ICTP	Technological	Technological	*0.001*	*0.217*	*0.782*
ICV	Technological	Technological	*0.000*	*0.171*	*0.829*
IDAB	Scientific	Scientific	*0.000*	*0.883*	*0.117*
IEGD (IEG)	Scientific	Scientific	*0.391*	*0.605*	*0.004*
IEGPS	Humanistic	Humanistic	*0.801*	*0.198*	*0.001*
IEIOP	Humanistic	Humanistic	*0.997*	*0.003*	*0.000*
IEM	Technological	Technological	*0.000*	*0.034*	*0.966*
IESAA	Humanistic	Humanistic	*0.991*	*0.008*	*0.000*
IETCC	Technological	Scientific	*0.032*	*0.832*	*0.136*
IF	Technological	Technological	*0.000*	*0.111*	*0.889*
IFA	Technological	Technological	*0.346*	*0.196*	*0.457*
IFCA	Technological	Technological	*0.000*	*0.067*	*0.932*
IFI	Technological	Technological	*0.000*	*0.061*	*0.939*
IFIC	Technological	Technological	*0.001*	*0.037*	*0.962*
ILC (IFL)	Humanistic	Humanistic	*0.985*	*0.015*	*0.000*
IFS	Humanistic	Humanistic	*1.000*	*0.000*	*0.000*
IFTE	Scientific	Scientific	*0.002*	*0.890*	*0.108*
IG	Technological	Technological	*0.000*	*0.115*	*0.885*
IGE	Humanistic	Humanistic	*0.944*	*0.006*	*0.050*
IH	Humanistic	Humanistic	*0.973*	*0.027*	*0.001*
IHCD	Humanistic	Humanistic	*1.000*	*0.000*	*0.000*
IIAG	Scientific	Scientific	*0.023*	*0.897*	*0.081*
IIBB	Scientific	Scientific	*0.003*	*0.935*	*0.062*
IIBM	Scientific	Scientific	*0.001*	*0.887*	*0.112*
IIIA	Technological	Scientific	*0.367*	*0.587*	*0.046*
IIM	Technological	Technological	*0.003*	*0.371*	*0.626*
IIQ	Technological	Scientific	*0.000*	*0.604*	*0.396*
IIQAB	Scientific	Scientific	*0.001*	*0.710*	*0.289*
IIRI	Scientific	Scientific	*0.073*	*0.916*	*0.011*
ILLA (ILE)	Humanistic	Humanistic	*1.000*	*0.000*	*0.000*
IMAFF	Scientific	Scientific	*0.005*	*0.839*	*0.156*
IMB	Scientific	Scientific	*0.003*	*0.986*	*0.011*
IMEDEA	Scientific	Scientific	*0.003*	*0.504*	*0.494*
IMF	Humanistic	Humanistic	*0.985*	*0.015*	*0.000*
IN	Scientific	Scientific	*0.002*	*0.976*	*0.022*
INCAR	Technological	Technological	*0.001*	*0.132*	*0.866*
INGENIO	Humanistic	Scientific	*0.059*	*0.923*	*0.018*
IO	Technological	Technological	*0.000*	*0.004*	*0.996*
IPBLN	Scientific	Scientific	*0.001*	*0.766*	*0.233*
IPE	Humanistic	Scientific	*0.056*	*0.606*	*0.337*
IPLA	Technological	Technological	*0.000*	*0.215*	*0.784*
IPNA	Scientific	Scientific	*0.001*	*0.908*	*0.091*
IQFR	Scientific	Scientific	*0.001*	*0.559*	*0.441*
IQM	Technological	Technological	*0.000*	*0.020*	*0.980*
IQOG	Technological	Technological	*0.000*	*0.095*	*0.905*

IREC	Scientific	Technological	*0.002*	*0.464*	*0.534*
IRN	Humanistic	Scientific	*0.023*	*0.561*	*0.415*
IRNAS	Technological	Technological	*0.000*	*0.190*	*0.809*
IRNASA	Humanistic	Scientific	*0.238*	*0.557*	*0.205*
ITQ	Technological	Technological	*0.000*	*0.026*	*0.974*
LITEC	Scientific	Scientific	*0.003*	*0.987*	*0.010*
MBG	Technological	Technological	*0.000*	*0.069*	*0.931*
MNCN	Technological	Technological	*0.000*	*0.003*	*0.997*
OE	Humanistic	Humanistic	*0.993*	*0.005*	*0.002*
RJB	Humanistic	Scientific	*0.366*	*0.385*	*0.249*
UBF	Scientific	Scientific	*0.002*	*0.978*	*0.020*
IPP (UPC)	Humanistic	Humanistic	*0.886*	*0.112*	*0.001*