

Cluster methods for assessing research performance: exploring Spanish computer science

Alfonso Ibáñez · Pedro Larrañaga · Concha Bielza

The objective of this paper is to propose a cluster analysis methodology for measuring the performance of research activities in terms of productivity, visibility, quality, prestige and international collaboration. The proposed methodology is based on bibliometric techniques and permits a robust multi-dimensional cluster analysis at different levels. The main goal is to form different clusters, maximizing within-cluster homogeneity and between-cluster heterogeneity. The cluster analysis methodology has been applied to the Spanish public universities and their academic staff in the computer science area. Results show that Spanish public universities fall into four different clusters, whereas academic staff belong into six different clusters. Each cluster is interpreted as providing a characterization of research activity by universities and academic staff, identifying both their strengths and weaknesses. The resulting clusters could have potential implications on research policy, proposing collaborations and alliances among universities, supporting institutions in the processes of strategic planning, and verifying the effectiveness of research policies, among others.

Keywords Cluster analysis methodology · Bibliometric techniques · Universities · Academic staff · Computer Science · Spain

Introduction

The important role played by higher education institutions in national research and the increasing competition among them for limited resources have led to the appearance of

analyses to establish the most influential institutions in recent years. Many countries have introduced national exercises for the evaluation of research activity, responding to demands for greater accountability and efficiency in funding for institutions. Governments and their national agencies are gradually imposing elements of competition in the allocation of public funds.

Research activity assessment is a fairly new phenomenon in Spain. Nowadays, three organizations are in charge of assessing Spanish scientific activity: National evaluation and foresight agency, National Evaluation Committee of Research Activity (CNEAI) and National Agency for Quality Assessment and Accreditation (ANECA).

Research has grown exponentially in Spain over last fifty years. It now accounts for 3.3 % of global output stored in the Thomson Reuters database, compared with a share of only 0.2 % in 1963. But, as in other areas, quality is more important than quantity in science. And this is where Spain falls down. According to Thomson Reuters' Essential Science Indicators, Spain ranks 9th among the top-performing countries for papers, 11th for citations, and 34th for citations per paper in all fields. Spain fares no better in the field of computer science, which is the focus of this paper. In this case, Spain ranks 10th for papers, 9th for citations, and 35th for citations per paper. On the other hand, computer science is an active research area in Spain that has recently advanced substantially. Also, importantly, it has a huge potential for bridging the gap between higher education and the business world.

Spanish higher education has expanded remarkably over the last half century. There are now 50 public universities, compared with only 15 in 1968. Most of these institutions sprang up between the 1970s and the 1990s. Also, there are currently 28 private universities. These universities tend to focus more on teaching than public universities, which also pursue research interests. For this reason, this analysis is confined to public universities, and also circumscribed to 48 out of the 50 public universities, because two of them (Universidad Internacional de Andalucía and Universidad Internacional Menéndez Pelayo) have no academic staff specialized in the computer science field.

Let us now turn to the academic staff of the selected universities. We analyze the 2004 tenured academic staff at the above universities (December 2009). All the academic staff are associated with one of the following three specific areas: Computer architecture and technology, computer science and artificial intelligence (CSAI), and computer languages and systems (CLS). Members of the academic staff specialize in one specific area, in which they lecture, regularly publish and are assessed by national organizations. There are four types of permanent (civil servant) positions in the Spanish higher education system. All four positions are associated with tenure obligations. These academic staff work full-time and engage in teaching, mentoring and research at the university. We translate these four academic positions (from the highest to the lowest level) as: full professor (FP), associate professor-type1 (AP1), associate professor-type2 (AP2) and associate professor-type3 (AP3).

Nowadays, the process of evaluation of scientific research has become a central element in the management and governance policies of national research systems. The most widespread evaluation methodologies can be classified into two general types: peer-review and bibliometric techniques. Although if used properly peer review is assumed to be the most reliable methodology, it is slow, expensive and unwieldy (Mulligan 2005; Cobo et al. 2007; Scarpa 2006). Other authors contest this appraisal (Hanks 2005; Horrobin 2001). This difference of opinion among authors has led to the development of methodologies based on bibliometric techniques. Both types of methodologies have pros and cons, extensively discussed in the literature (Horrobin 1990; MacRoberts and MacRoberts 1996;

Moxham and Anderson 1992; VanRaen 2005), in terms of costs, execution times, limitations and objectiveness of measurement.

The objective of this paper is to develop a cluster analysis methodology for measuring the performance of research activities in terms of productivity, visibility, quality, prestige and internationalization, while overcoming some of the limitations related to methodologies that have been proposed in the literature. Our cluster analysis methodology is based on bibliometric techniques and, therefore, has many advantages (objectivity, rapidity, and low costs, among others) over a peer-review methodology. The proposed methodology does not depend on the quality judgment of experts, so it does not suffer severe limitations related to subjectivity. It also overcomes the traditional limits of bibliometric analyses based on simple rankings and permits a robust multi-dimensional cluster analysis at the level of universities and academic staff. The cluster analysis methodology has been applied to the Spanish public universities and their academic staff in the computer science area. The results can be used to characterize the research activity of universities and academic staff, identifying both their strengths and weaknesses. These analyses afford a comprehensive overview of the current situation in the area of computer sciences in Spain.

Using our methodology, policy-makers could discover knowledge related to universities and their staff. The goal of the cluster analysis methodology is to form different clusters, maximizing within-cluster homogeneity and between-cluster heterogeneity. In this way, universities/academics that belong to the same cluster are very similar to each other, whereas universities/academics belonging to different clusters are very different in terms of bibliometric data. Each cluster is interpreted as providing a characterization of research activity by universities and academic staff, identifying both their strengths and weaknesses. These value-added clusters could have potential implications on research policy. Finally, this methodology supports institutions in the processes of strategic planning, in verifying the effectiveness of policies and initiatives for continuous improvement.

The remainder of the paper is organized as follows. The next section reviews some related work. “Cluster analysis methodology” section describes the procedures on which our cluster analysis methodology is based. “Exploring Spanish computer science research” section presents how both Spanish universities and their academic staff are grouped into different clusters. Finally, “Discussion and conclusions” section contains some discussions and conclusions about the results and future research on the topic.

Related work

Some methodologies have been published in the literature for assessing the research performance at different levels. First, Abramo et al. (2008) developed a bibliometric-non-parametric methodology for measuring the performance of research activities in the Italian university system. They used the data envelopment analysis technique to compare scientific production performances of Italian universities during the period 2001–2003. They defined an efficient production frontier based on three outputs (number of publications, contributions to publications and scientific strength) to assess the inefficiency production of universities, based on minimum distance from the frontier. Second, Costas et al. (2010) proposed a general bibliometric methodology for informing the assessment of research performance of individual scientists. The authors set up a bibliometric profile for every researcher composed of nine performance variables (number of publications, number of citations, h-index, percentage of highly cited papers, internationally normalized impact, citations per publication, median impact factor, normalized journal position, and subfield

normalized impact). Through factor analysis, the nine variables were then reduced to three dimensions (impact, journal quality, and production). The proposed methodology analyzed publications from 1994 to 2004 belonging to researchers working at the Spanish National Research Council. Third, Abramo and D'Angelo (2011) also suggested a bibliometric methodology for large-scale comparative evaluation of research performance by individual researchers, research groups and departments within research institutions. The proposed methodology provided performance ratings for bibliometric indicators (productivity, fractional productivity, scientific strength, fractional scientific strength and quality indexes) and relative rankings with respect to other researchers in the same area. Their methodology was also applied to the Italian university research system and the analyzed period was 2004–2006. Finally, Torres-Salinas et al. (2011a) proposed a methodology for comparing academic institutions. Their methodology is based on a bidimensional quantitative-qualitative index, the *IFQ²A*. The quantitative dimension shows the net production of an institution by using raw indicators (number of documents, number of citations and *h*-index), whereas the qualitative dimension focuses on the ratio of high-quality production on each institution (ratio of papers published in journals in the top JCR quartile, average number of citations received by all citable papers and ratio of papers that belong to the top 10 % most cited). Their methodology was applied to the Spanish university research system during the period 2000–2009.

We noted that none of the previous methodologies performed a cluster analysis, they just presented absolute values for bibliometric indicators achieved for researchers and institutions. Unlike these methodologies, Palomares-Montero and García-Aracil (2010) performed a fuzzy clustering algorithm to analyze universities according to three aspects (teacher mission, research mission and knowledge transfer mission). On the one hand, the teaching indicators were performance rate (rate between the number of credits passed with respect to the number of credits enrolled), student-teacher ratio, and running expenses per student. On the other hand, the research indicators were thesis awarded by professor holding a PhD degree, ISI publication-tenured professor ratio, and sexenio-tenured professor ratio. Finally, the knowledge transfer indicators were patent-teacher ratio, contracts-teacher ratio, and grants income by fulltime teacher. Their methodology was applied to the Spanish universities in the academic year 2006–2007.

We have also found some studies that did not develop new methodologies, they just performed simple descriptive exercises. Some bibliometric studies, like He and Guan (2008), Rojas-Sola and Jorda-Albinana (2009), Wainer et al. (2009), among others, have also analyzed computer science research. Again, these studies performed no cluster analysis. They just show simple descriptive information related to a specific set of the publications. First, He and Guan (2008) analyzed the productivity and visibility of Chinese researchers from 1997 to 2005 indexed by the ISI Web of Knowledge. They focused on proceeding papers published in Lecture Notes in Computer Science. A total of 5916 proceeding papers are analyzed in the study. Results showed that Chinese researchers seldom work alone, and they are more likely to collaborate with domestic fellows than overseas scientists. Also, in spite of the increasing numbers of Chinese papers on computer science in the last few years, they receive only a few citations. Second, Rojas-Sola and Jorda-Albinana (2009) studied the journal articles published by researchers from Venezuelan universities from 1997 to 2007 through the ISI Web of Knowledge. They found 181 journal articles for computer science journals. They discovered that four universities account for 93.9 % of all university scientific output. Universidad Simón Bolívar stands out on scientific output, whereas Universidad de los Andes excels in terms of impact factor. Finally, Wainer et al. (2009) explored Brazilian research published in journal articles and proceeding papers

indexed by ISI Web of Knowledge and Scopus. They compared Brazilian production from 2001 to 2005 with some Latin American, European and other countries with sizeable research output. A total of 2,357 and 3,343 publications were extracted from ISI Web of Knowledge and Scopus, respectively. Results showed that Brazil has by far the largest scientific output among the Latin American countries, produces about a third as much as Spain and a quarter as much as Italy, and about the same as India and Russia.

Other studies, like Ibáñez et al. (2013), Torres-Salinas et al. (2011b) and Rojo and Gómez (2006), analyzed Spanish research in computer science. These studies analyzed productivity, visibility and collaboration patterns without cluster analysis. First, Ibáñez et al. (2013) focused on the relationship among research collaboration, number of documents and number of citations of computer science research activity. They analyzed the number of documents and citations and how they vary by number of authors. These measures were also analyzed under different circumstances, that is, when documents are written in different types of collaboration, when documents are published in different document types, when documents are published in different computer science subdisciplines, and, finally, when documents are published by journals with different impact factor quartiles. To investigate the above relationships, they analyzed the publications listed in the Web of Science and produced by active Spanish university professors between 2000 and 2009, working in the computer science field. Second, Torres-Salinas et al. (2011b) analyzed Spanish universities according to quantitative and qualitative measures related to production, impact and journal quality. They studied scientific output published in journals from 2001 to 2010 indexed by ISI Web of Knowledge. Torres-Salinas et al. showed that Universidad Politècnica de Catalunya, Universidad de Granada (UGR) and Universidad Politècnica de València stand out in terms of quantitative measures (number of papers, number of citations, and *h*-index), whereas Universidad de Barcelona, Universidad de Santiago de Compostela and Universidad de Girona (UDG) excel with respect to qualitative measures (percentage of papers published in first-quartile journals, average citations per paper, and percentage of highly cited papers). Finally, Rojo and Gómez (2006) provided an overview of scientific (publications) and technological (patents) production during the 1990–2002 period, ranking Spanish universities according to some measures. Scientific output data were obtained from the Information Service for the Physics and Engineering database, whereas technological output was obtained from the European and Spanish Patent Office patent databases. Rojo and Gómez observed that the most active universities were: Universidad Politècnica de Catalunya, Universidad Politècnica de Madrid (UPM) and Universidad Politècnica de València, whose output is specially oriented towards conference papers. UGR, with a more basic character, has a high production of journal papers. Results also showed that the universities holding most patents were: Universidad Complutense de Madrid, Universidad Politècnica de Catalunya and UPM.

Cluster analysis methodology

Our cluster analysis is divided into several procedures. These procedures are in charge of defining and describing the bibliometric variables, collecting bibliometric records from different databases, ensuring the reliability of data, calculating bibliometric indices, presenting statistical description of bibliometric indices, performing partitional, hierarchical and probabilistic cluster analysis at different levels, visualizing clustering results, identifying the achieved clusters and, finally, supporting institutions on research policy decisions. We explain them in the following sections.

Definition of bibliometric variables

The scientific literature is the main channel for disseminating and validating research results. Two basic measures (number of documents and number of citations) are responsible for quantifying, respectively, the productivity and the visibility of any researcher or institution. Therefore, these measures have become crucial bibliometric indices for assessing research activity.

Bibliometric indices are quantitative metrics for evaluating and comparing the research activity of individual researchers and institutions according to their output. These indices essentially involve counting the number of times scientific papers are cited. They are based on the assumption that influential studies will be cited more frequently than others. Nowadays, bibliometric indices are an increasingly important topic for the scientific community. They represent an objective and verifiable method whose results are reproducible. This method is applicable to a large volume of data and is able to test the statistical significance of different hypotheses. In this way, Ibáñez et al. (2011) learned a Bayesian network model from bibliometric data to analyze how bibliometric indices relate to each other. The induced Bayesian network was then used to discover probabilistic conditional (in)dependencies among the bibliometric indices and, also for probabilistic reasoning.

The eight bibliometric indices analyzed in this study are:

- *Normalized documents* This measure indicates the ability of each university to produce scientific knowledge. *Normalized documents* is defined as the ratio between the number of documents published by each university and the number of academics affiliated with that university. It is calculated allowing for the influence of university size in order to obtain a fair measure of production.
- *Normalized citations* This measure shows the scientific impact that each university has on the scientific community. It again allows for the influence of university size. *Normalized citations* is the ratio between the number of citations received by each university and the number of academics affiliated with that university.
- *Journal publication* This indicator analyzes the penchant towards either of the two most important types of research output (journals or conferences). *Journal publication* represents the ratio between the number of documents published in journals and the total number of documents published both in journals and in conferences. This indicator establishes each university's main dissemination channel.
- *First-quartile documents* This indicator shows the percentage of publications that a university publishes in the world's most influential scholarly journals. Journals considered for this indicator are ranked in the first quartile of their categories as ordered by Journal Citation Reports. *First-quartile documents* is the percentage of documents published in first-quartile journals with respect to the sum published in all other quartiles.
- *Fourth-quartile documents* This is a similar indicator to *First-quartile documents* but for the least influential scholarly journals according to Journal Citation Reports (fourth-quartile).
- *Citations per journal article* This measure is associated with the impact of journal articles. *Citations per journal article* represents the mean number of citations received by documents published in journals. This indicator reflects the quality of journal articles published by each university.
- *Citations per proceeding paper* This measure is associated with the impact of proceeding papers. *Citations per proceeding paper* represents the mean number of

citations received by documents published in conference proceedings. This indicator reflects the quality of proceeding papers published by each university.

- *International collaboration* This indicator shows the ability of each university to create international research links through publications. *International collaboration* represents the percentage of publications that a university publishes in collaboration with overseas institutions.

Another two bibliometric indices (*total documents* and *total citations*) are analyzed in this study. These indices replace, respectively, *normalized documents* and *normalized citations* for clusterings of academic staff

- *Total documents* This measure indicates the ability of each academic to produce scientific knowledge. *Total documents* is defined as the number of documents published by each academic. It represents the academic's productivity.
- *Total citations* This measure shows the scientific impact that each academic has on the scientific community. *Total citations* is defined as the number of citations received by each academic. It represents the academic's visibility.

Data collection

Our work is based on the analysis of the research activity of Spanish public universities and their academic staff in the computer science area using a cluster analysis methodology. Two datasets are built for this purpose. The first dataset includes the values of the bibliometric indices of Spanish universities from the date of the first publication by an active member of their academic staff (January 1, 1973) to December 31, 2009. We use this dataset to group universities into different clusters. The second dataset includes the values of the bibliometric indices for each academic from his/her first publication until December 31, 2009. We use this dataset to group academics into different clusters. In the following, we illustrate the different phases for building these datasets.

The first step was to apply to the Spanish Ministry of Education for a list of active academics associated with three specific areas (CAT, CSAI and CLS) as of December 31, 2009. This list includes the full name of each academic (out of 2004 academics), and their associated university, position and research area. The next step was to retrieve a list of publications and citation data (as of December 31, 2009) for each academic. This information was carefully downloaded from the Web of Science (ISI Web of Knowledge).

ISI Web of Knowledge contains databases specialized in journals (Science Citation Index and Journal Citation Reports) and conferences (Conference Proceedings Citation Index) covering more than 400 computer science journals and more than 7,000 of the most significant computer science conferences, respectively. While this platform does not store all the scientific literature, it does store what really matters (Garfield 1996). Finally, it is one of the most important tools used by CNEAI and ANECA in order to assess Spanish scientific activity (Ruiz-Pérez et al. 2010).

Regarding data extraction, we recorded all publications (journal articles and proceeding papers) associated with the 2004 Spanish computer science academics. Publications by each academic were then filtered by his/her affiliated university. This step took into account different Spanish university name variations. Also, only publications related to the computer science area (according to JCR categories) were selected. In order to ensure the reliability of results, we checked our final list of publications against other databases like DBLP Computer Science Bibliography, personal webpages and institutional websites,

among others. This step was carefully carried out due to problems related to Spanish personal name variations in international databases (Ruiz-Pérez et al. 2002).

The last phase was to develop software using all this information to calculate the bibliometric indices (variables in the cluster analysis) associated with the selected universities and academics. We selected eight specific bibliometric indices which are widely accepted among the scientific community, measure different aspects of scientific activities and are easily interpretable.

Statistical description of bibliometric indices

Before performing any clustering approach, a statistical summary of bibliometric indices values are presented. The objective is to provide an overview of the performance of Spanish computer science research in terms of productivity, visibility, quality, prestige and internationalization. After computing all the eight bibliometric indices for all 48 universities and 2004 academics, we show box plots that represent the smallest observation (extreme of the lower whisker), lower quartile (Q1), median (Q2), upper quartile (Q3), and largest observation (extreme of the upper whisker) for each bibliometric index. Box plots may also indicate which observations, if any, might be considered outliers. This statistical description also shows the top five universities ranked according to our bibliometric variables. These rankings are useful to compare different universities in a one-dimensional basis.

Cluster analysis at different levels

Clustering can be considered as the most important unsupervised learning problem. It is concerned with finding a structure in a collection of unlabeled elements that are characterized by several variables. The goal is to group elements in this collection so that elements that belong to a cluster are very similar to each other, whereas different clusters are highly heterogeneous.

Different starting points and criteria usually lead to different taxonomies of clustering algorithms (Jain et al. 1999; Everitt et al. 2001; Xu and Wunsch 2005). A simple agreed frame is to classify clustering techniques as partitional clustering, hierarchical clustering and probabilistic clustering, based on the properties of clusters generated. Partitional clustering groups elements exclusively, so that any element belonging to one specific cluster cannot be a member of another cluster. On the other hand, hierarchical clustering produces a hierarchical structure of clusters. Hierarchical clustering proceeds successively by either merging smaller clusters into larger ones (agglomerative clustering) or by splitting larger clusters (divisive clustering). Finally, probabilistic clustering provides a cluster membership probability for each element, where elements have a specific probability of being members of several clusters. We have used R implementations of the above clustering approaches for obtaining the results (R Development Core Team 2011).

One of the most important issues in cluster analysis is the evaluation of clustering results (Halkidi et al. 2001). Clustering validation is concerned with determining the optimal number of clusters (the best for the input dataset) and checking the quality of clustering results. We have used both internal and external validity indices in order to evaluate the clustering results. Internal validity indices do not require a priori information from dataset, they are based on the information intrinsic to the dataset alone, whereas external validity indices require previous knowledge about dataset.

Partitional clustering

Partitional clustering algorithms assign a set of objects into k clusters with no hierarchical structure. In principle, the optimal partition, based on some specific criterion, can be found by enumerating all possibilities. But this brute force method is infeasible in practice, due to the expensive computation (Liu 1968). Therefore, heuristic algorithms have been developed in order to seek approximate solutions.

The k -means algorithm (McQueen 1967) is a well-known partitional clustering algorithm. In this algorithm a cluster is represented by its centroid, which is a mean of points within a cluster. This algorithm works conveniently only with numerical attributes (for mean calculation) and can be negatively affected by a single outlier. Some outliers, which are quite far away from the cluster centroid, are still forced into a cluster and, thus, distort the cluster shapes. In this way, new algorithms, like PAM (Kaufman and Rousseeuw 1990), have appeared in order to overcome these obstacles.

Partitioning around medoids (PAM) has several advantages with regard to k -means. First, this algorithm presents no limitations on attributes types because it utilizes real data points (medoids) as the cluster prototypes (medoids do not need any computation and always exist). Second, the choice of medoids is dictated by the location of a predominant fraction of points inside a cluster and, therefore, it is lesser sensitive to the presence of outliers. Finally, unlike k -means the resulting clustering is independent of the initial choice of medoids. For these reasons, we used PAM as a representative algorithm of partitional clustering. The objective of this algorithm is to determine a representative element (medoid) among the elements of the dataset for each cluster. For k clusters, the goal is to find k representative elements which minimize an objective function

$$E = \sum_{i=1}^k \sum_{x \in C_i} d(x, m_i), \quad (1)$$

where k is the number of clusters, x is an element belonging to the cluster C_i , m_i is the medoid of cluster C_i , and $d(x, m_i)$ is the Euclidean distance between x and m_i .

The algorithm begins by selecting an element as a medoid for each cluster C_i . After selecting a set of k medoids, k clusters are constructed by assigning each element to its nearest medoid. If the objective function can be reduced by switching a selected medoid for an unselected (non-medoid) element, then they are switched. This continues until the objective function can be decreased no further.

Hierarchical clustering

Hierarchical clustering algorithms build a tree of clusters called dendrogram. This dendrogram allows exploring data on different levels of granularity. Hierarchical clustering algorithms are categorized into agglomerative and divisive (Jain and Dubes 1988). Agglomerative clustering starts with clusters and each of them includes exactly one object. A series of merge operations are then followed out that finally lead all objects to the same group. Divisive clustering proceeds in an opposite way. It starts with one cluster of all objects and recursively splits the most appropriate objects. For a cluster with n objects, there are $2^{n-1} - 1$ possible two-subset divisions, which is very expensive in computation (Everitt et al. 2001). Therefore, divisive clustering is not commonly used in practice. Despite this, some divisive clustering algorithms, like MONA and DIANA (Kaufman and Rousseeuw 1990), are also developed in the literature.

We focus on the agglomerative clustering. Based on the different definitions for distance between two clusters, there are many agglomerative clustering algorithms. The simplest methods include single linkage (Sneath 1957) and complete linkage technique (Sorensen 1948). Other linkage metrics, such as average linkage, median linkage and centroid linkage, are also developed. Unlike methods based on linkage metrics, a more complicated agglomerative clustering algorithm called the Ward's method (Ward 1963) uses an analysis of variance approach to evaluate the distances between clusters. It is also known as Ward's minimum variance method.

Ward's algorithm (Ward 1963) was used as an advanced hierarchical clustering procedure. Given k clusters, this algorithm reduces them to $k - 1$ mutually exclusive clusters by considering the union of all possible $k(k - 1)/2$ pairs. It selects the union of clusters which minimizes the heterogeneity among cluster elements

$$E = \sum_{i=1}^k \sum_{x \in C_i} d(x, ct_i), \quad (2)$$

where k is the number of clusters, x is an element belonging to the cluster C_i , ct_i is the centroid of cluster C_i , and $d(x, ct_i)$ is the squared Euclidean distance between x and ct_i .

Thus, homogeneous clusters are linked to each other. The complete hierarchical structure can be obtained by repeating this process until only one cluster remains.

Probabilistic clustering

From a probabilistic perspective, we find the most likely set of clusters given the data. This type of clustering is based on a statistical model called finite mixture model that is assumed to generate the data and whose parameters are estimated using methods such as the expectation–maximization (EM) algorithm (Dempster et al. 1977), algorithms SNOB (Wallace and Dowe 1994), AutoClass (Cheeseman and Stutz 1996), Mclust (Fraley and Raftery 1999).

Given a number of clusters k , probabilistic clustering models data as a finite mixture of k probability density functions. Each cluster i is represented by one component $f_i(\mathbf{x})$ of the mixture. We have modeled the n variables as conditionally independent Gaussian distributions given the cluster value. Each distribution (i) is characterized by two parameters for each variable (j): the mean (μ_{ij}) and the standard deviation (σ_{ij}).

The EM algorithm, which is the most popular algorithm of this type of clustering (McLachlan and Krishnan 1997), is used to find the maximum likelihood estimates of the mixing coefficients (π_i) and the parameters of the conditional Gaussian distributions (μ_{ij} and σ_{ij}). Thus,

$$\begin{aligned} f(\mathbf{x}) &= \sum_{i=1}^k \pi_i f_i(\mathbf{x}) \\ &= \sum_{i=1}^k \pi_i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \\ &= \sum_{i=1}^k \pi_i \prod_{j=1}^n \left[\frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{1}{2} \left(\frac{x_j - \mu_{ij}}{\sigma_{ij}} \right)^2} \right]. \end{aligned} \quad (3)$$

The algorithm converges to a locally optimal solution by iteratively updating values for π_i , μ_{ij} and σ_{ij} . This whole process is embedded in a cross-validation procedure that is capable of estimating the number of clusters k without this having to be set a priori.

Visualization of clustering results

Several figures are presented to represent different aspects of the clustering results. After performing the partitional clustering, several tables show all universities/academics grouped into disjoint clusters and the medoid bibliometric values within each cluster. Even if universities/academics belong to the same cluster, they may behave differently depending on the bibliometric indices. In this way, we present several figures showing cluster projection for some specific bibliometric indices. Then, we represent the hierarchical structure of clusters (dendrogram) obtained by merging smaller clusters into larger ones. This dendrogram shows how the clusters are related. By cutting the dendrogram at a target level, we obtain all universities/academics grouped into disjoint clusters. Regarding probabilistic clustering, we present the mean and standard deviation values for each bibliometric variable within the resulting clusters. After that, we list each university/academic's probability of being a member of each cluster.

The resulting clusters are also visually inspected using a representation in a lower dimensional space. The goal is to obtain a three-dimensional representation that approximates our eight-dimensional bibliometric variables and check whether or not the clusters were visually distinguishable. We use a principal component analysis for this purpose. Finally, we also plot the distribution of academics grouped in each cluster for analyzing each cluster by areas and positions associated with each academic.

Identification of final clusters

Each cluster can be defined according to different research activity aspects e.g. productivity (documents per academic), visibility (citations per academic), quality (citations per journal articles and proceeding papers), prestige (first-quartile journals), and internationalization (international collaboration). We set global labels (high, medium-high, medium-low and low) for the values of each bibliometric index in the different clustering algorithms, so each cluster can be represented as a set of global labels associated with research activity aspects. Using the resulting clusters and the above labels, we could conclude that some universities/academics produce more scientific knowledge and have a bigger scientific impact than other universities/academics, whereas other universities/academics usually publish in the most influential journals, and thus they have a selective strategy, and finally, we could also conclude that specific universities/academics have an excellent ability to create international research collaborations.

Implications on research policy

Methodologies for the evaluation of research activities have been raising an increasing amount of interest in the last few years. The conclusions of these methodologies have great relevance for the design of policies to promote research and development.

Thanks to our cluster analysis methodology a comprehensive overview of the current situation in a specific discipline and region is achieved. This overview could help policy-makers for making decisions. The proposed methodology could be considered as a tool,

which could also help university presidents and heads of departments and research groups in the processes of strategic planning, in verifying the effectiveness of policies and initiatives for continuous improvement, in the optimization of limited economic resources, and in the promotion of academic staff, among others. The resulting clusters are interpreted as providing characterizations of research activity by universities and academic staff, identifying both their strengths and weaknesses. Using our methodology, policy-makers could propose collaborations and alliances among universities. These universities could perhaps merge strategically in order to exploit their resources, enhance their reputation and visibility, and compete with the most active international universities.

Exploring Spanish computer science research

Spanish public universities

We first calculate all the bibliometric indices for all 48 universities. Figure 1 shows the box plots of the distribution of each bibliometric index. Taking *Normalized documents* as an example, we found that 1.4 was the value of the lower whisker, whereas 11.9 was the value of the upper whisker. The 25th percentile (Q1), 50th percentile (Q2) and 75th percentile (Q3) were 3.6, 5.8, and 7.5 documents per academic, respectively. We also found an outlier (17.3) which corresponded with Universidad Pablo de Olavide (UPO) de Sevilla. Taking another example (*Journal publication*), we noted three outliers corresponding with Universidad Pompeu Fabra (UPF) (94.4), Universidad de León (70.0) and Universidad de Córdoba (UCO) (66.3). In this case, the five-number summaries were: lower whisker (19.6), 25th percentile (30.8), 50th percentile (36.5), 75th percentile (44.1), and upper whisker (64.0). Finally, we also report the minimum and maximum value of the analyzed indices: *Normalized documents* [1.4, 17.3], *Normalized citations* [1.3, 106.3], *Journal publication* [19.6, 94.4], *First-quartile documents* [0.0, 67.3], *Fourth-quartile documents* [0.0, 57.1], *Citations per journal article* [0.7, 18.3], *Citations per proceeding paper* [0.0, 2.0], and *International collaboration* [0.0, 83.3].

Table 1 shows the top five universities ranked according to our eight variables. Analyzing the values for all universities, we discovered that UGR had the highest value of *Normalized citations*. This means that the mean citations received by each academic affiliated with UGR was 106.3 citations. We also noted that the best university according to *First-quartile documents* was UCO, that is, 67.3 % of its journal articles were published in first-quartile journals. Similarly, UPF was the best university regarding *International collaboration* because 83.3 % of its collaborative documents were co-authored by researchers with overseas affiliations.

Before running any algorithm, the number of clusters should be fixed using clustering validation. The optimal number of clusters is usually determined based on internal validity indices like the silhouette coefficient (Rousseeuw 1987). This index is used to measure the goodness of a clustering structure without external information. Its value ranges from -1 to 1 . A larger average silhouette coefficient indicates a better overall quality of the clustering result, so the optimal number of clusters is the one that gives the largest average silhouette value. After running clustering validation, we found that the partitions with two clusters and four clusters had the highest silhouette coefficients. Although four-cluster partition had a little lower silhouette coefficient (0.65) than two-cluster partition (0.67), we selected four-cluster partition ($k = 4$) because it more realistically explained our dataset.

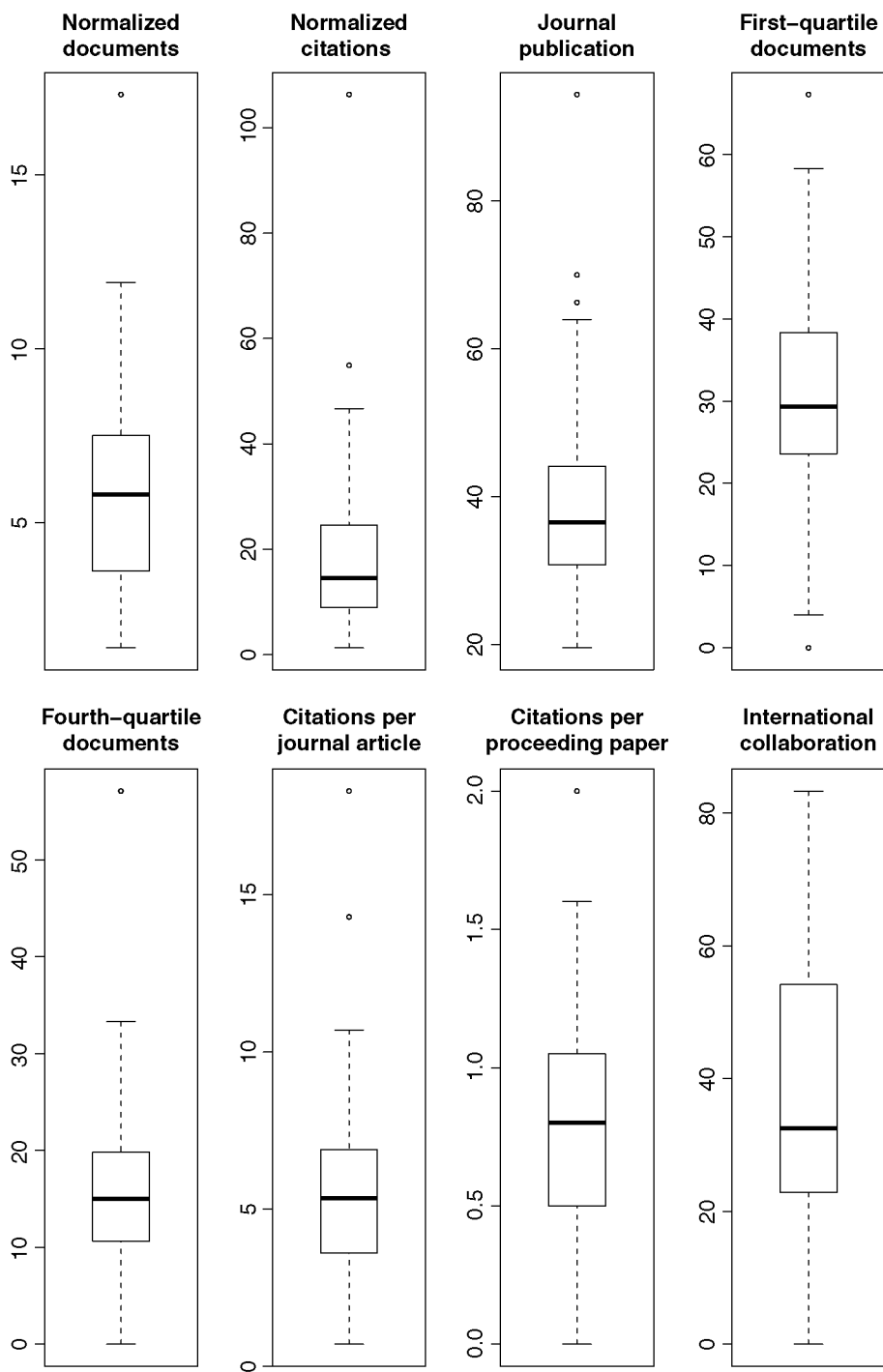


Fig. 1 Box plot of each bibliometric index

Table 1 Top five universities ranked according to our eight variables

Variables	Univ 1	Univ 2	Univ 3	Univ 4	Univ 5
<i>Normalized documents</i>	UPO (17.3)	UGR (11.9)	UC3M (10.7)	UCM (9.4)	UAM (9.4)
<i>Normalized citations</i>	UGR (106.3)	UJA (54.9)	UPNA (46.7)	UPO (45.0)	UPF (34.3)
<i>Journal publication</i>	UPF (94.4)	ULE (70.0)	UCO (66.3)	UPNA (64.0)	UGR (59.5)
<i>First-quartile documents</i>	UCO (67.3)	ULPGC (58.3)	UBU (57.1)	UB (52.6)	UV (45.5)
<i>Fourth-quartile documents</i>	UCA (57.1)	UNEX (33.3)	ULE (33.3)	UPO (29.4)	UAB (26.2)
<i>Citations per journal article</i>	UJA (18.3)	UGR (14.3)	UDG (10.7)	UBU (9.6)	UPNA (9.5)
<i>Citations per proceeding paper</i>	UAM (2.0)	UCM (1.6)	US (1.4)	URV (1.4)	UPC (1.3)
<i>International collaboration</i>	UPF (83.3)	UDG (79.3)	UC (64.3)	UII (61.1)	UPC (60.5)

Table 2 Partitional clustering: four clusters of universities

Clusters	Universities
Cluster A	A Coruña, Almería, Cádiz, Carlos III de Madrid, Extremadura, Huelva, La Laguna, Las Palmas de Gran Canaria, León, Lleida, Miguel Hernández de Elche, Salamanca Nacional de Educación a Distancia, Politécnica de Cartagena, Rey Juan Carlos, Vigo
Cluster B	Alcalá de Henares, Alicante, Autónoma de Barcelona, Autónoma de Madrid, Cantabria, Castilla-La Mancha, Complutense de Madrid, Girona, Jaume I de Castello, La Rioja, Málaga, Politécnica de Catalunya, Politècnica de València, Pompeu Fabra, Sevilla
Cluster C	Barcelona, Burgos, Córdoba, Illes Balears, Murcia, Oviedo, País Vasco, Valencia, Politécnica de Madrid, Rovira i Virgili, Santiago de Compostela, Valladolid, Zaragoza
Cluster D	Granada, Jaén, Pablo de Olavide, Pública de Navarra

After choosing the number of clusters, we performed the partitioning around medoids (partitional clustering). Table 2 shows all universities grouped into four disjoint clusters. We observed that the number of universities belonging to each cluster were: cluster A (16 universities), cluster B (15 universities), cluster C (13 universities) and cluster D (4 universities).

Table 3 shows the medoid values within the four clusters (A, B, C and D). Analyzing the variable values, we noted that there were some differences among clusters. For example, we found that universities belonging to cluster D had the highest value for *Normalized citations* (54.9 citations per academic). They also excelled in terms of *Journal publication* and *Citations per journal article*. Universities associated with the other clusters (A, B and C) excelled with respect to the other variables: cluster A (*Fourth-quartile documents*), cluster B (*Normalized documents*, *Citations per proceeding paper* and *International collaboration*), and cluster C (*First-quartile documents*). Finally, it shows the medoid university within each cluster. In this way, we noted that Universidad de A Coruña (UDC) was representative of cluster A, Universidad de Málaga (UMA) was representative of cluster B, UPM was representative of cluster C, and Universidad de Jaén (UJA) was representative of cluster D.

Even if universities belong to the same cluster, they may behave differently depending on the bibliometric indices. Figure 2 shows cluster analysis projection for some specific bibliometric indices. Universities belonging to clusters A, B, C and D are represented by point-down triangles, squares, circles and point-up triangles, respectively.

Figure 2 (top) shows the projection on the *Normalized documents* and *Normalized citations* axes. Taking cluster D (point-up triangles) as an example, we found that there were important differences among the four universities. UPO belonged to cluster D and ranked 1st for *Normalized documents*, whereas UJA, which also belonged to cluster D, ranked 24th for *Normalized documents*. We also observed big differences between UGR and the other three universities regarding *Normalized citations*. Despite these differences, the four universities were the top scorers for *Normalized citations*.

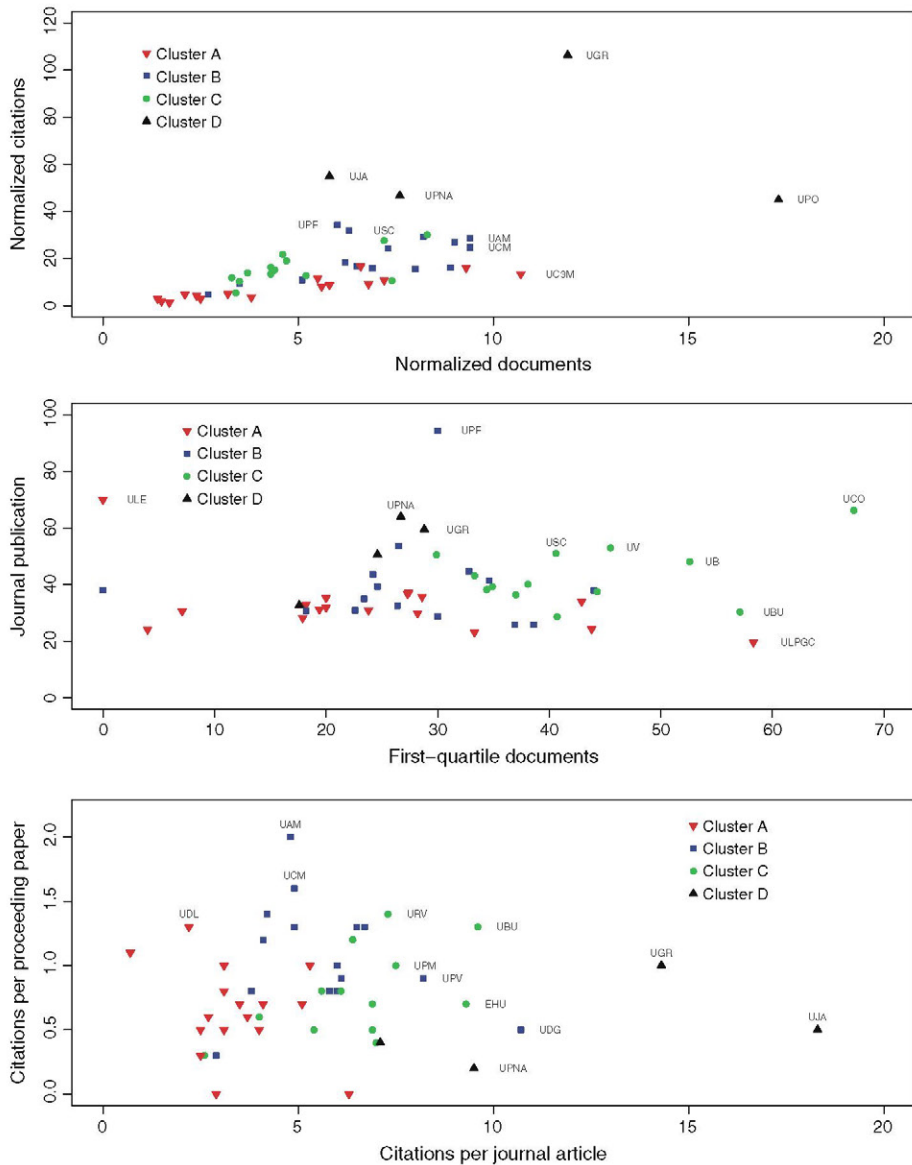
Figure 2 (middle) shows the projection on the *First-quartile documents* and *Journal publication* axes. Note that these rankings are very different to the previous ones (Fig. 2, top). According to Table 3, we observed that cluster A had the lowest value for *Journal publication*. Despite this, Universidad de León (ULE), which belonged to cluster A, ranked 2nd for *Journal publication*, outperforming universities belonging to better clusters. UPF ranked 1st for *Journal publication* and was a member of cluster B, which was not the highest scoring group for the analyzed variable. On the other hand, universities belonging to cluster C usually ranked top for *First-quartile documents*. UCO, which ranked 1st for *First-quartile documents*, also ranked 3rd for *Journal publication*.

Figure 2 (bottom) shows the projection on the *Citations per journal article* and *Citations per proceeding paper* axes. In this case, two universities belonging to cluster D (UJA and UGR) were among highest scorers for *Citations per journal article*. Universities belonging to cluster B (squares) did not score high for *Citations per journal article*. Even so, UDG, which belonged to cluster B, ranked 3rd for the above measure. On the other hand, universities belonging to cluster A (point-down triangles) did not score high on *Citations per proceeding paper*. Even so, Universidad de Lleida (UDL), which belonged to cluster A, ranked 6th for the above measure. Universities belonging to cluster B, like Universidad Autónoma de Madrid (UAM) and Universidad Complutense de Madrid (UCM), ranked top for *Citations per proceeding paper*.

Figure 3 represents the hierarchical structure of clusters (dendrogram) obtained by merging smaller clusters into larger ones (Ward's algorithm). This dendrogram shows how

Table 3 Partitional clustering: Medoid values within the four clusters (A, B, C and D) and the number of universities (in parentheses) associated with each cluster

Variables	Four resulting clusters			
	A (16 univ)	B (15 univ)	C (13 univ)	D (4 univ)
<i>Normalized documents</i>	7.2	7.3	4.4	5.8
<i>Normalized citations</i>	10.8	24.4	15.1	54.9
<i>Journal publication</i>	29.8	39.3	38.2	50.6
<i>First-quartile documents</i>	28.2	24.6	34.3	24.6
<i>Fourth-quartile documents</i>	17.9	17.7	10.8	12.3
<i>Citations per journal article</i>	3.1	6.5	7.5	18.3
<i>Citations per proceeding paper</i>	0.8	1.3	1.0	0.5
<i>International collaboration</i>	25.2	53.8	39.3	21.8
<i>Medoid university within each cluster</i>	UDC	UMA	UPM	UJA



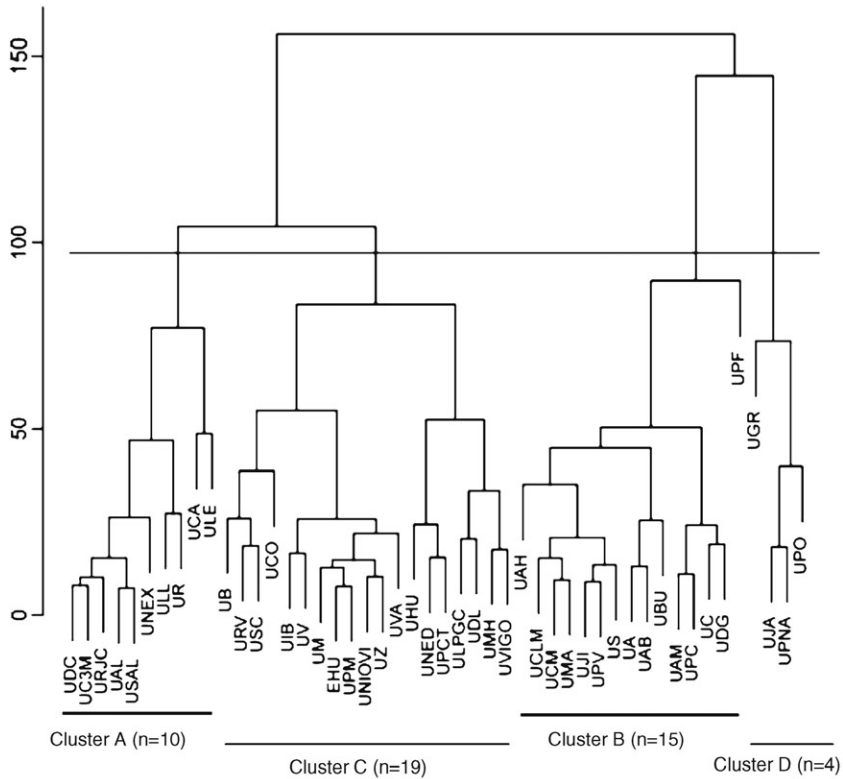


Fig. 3 Hierarchical clustering: Hierarchical structure of clusters (dendrogram) obtained by merging smaller clusters into larger ones using Ward's algorithm

universities). These results are very similar to the outcomes for partitional clustering. Note that the four universities belonging to cluster D, and 14 out of 15 universities belonging to cluster B are the same as before. Regarding cluster A, note that it contains fewer universities (down from 16 to 10), six universities having moved to cluster C.

Regarding probabilistic clustering, we ran the EM algorithm and formed four different clusters. Table 4 shows the mean and standard deviation values for each variable within the four resulting clusters. Taking *Normalized citations* as an example, we found that universities belonging to cluster D received an average number of citations equal to 62.4 ± 25.7 citations per academic. Universities belonging to cluster A, B, and C received on average fewer citations: 7.2 ± 4.1 , 21.9 ± 7.4 and 16.9 ± 5.1 , respectively.

After obtaining the four resulting clusters, we turned to the question of which universities belong to each cluster. Table 5 shows the answer. It lists each university's probability of being a member of each cluster. The highest membership probability for almost all universities was close to 1.00. This means that there was no doubt about which cluster they belong to. For example, the members of cluster D were: UGR, UJA, UPO de Sevilla (their membership probability of cluster D was 1.00000) and Universidad Pública de Navarra (its membership probability of cluster D was 0.99999). On the other hand, we noted that all universities belonging to cluster A had a probability >0.82 , whereas universities belonging to cluster B and C, had a probability >0.80 and 0.88 , respectively.

Table 4 Probabilistic clustering: Mean \pm standard deviation values for each variable within the four clusters (A, B, C and D) and the number of universities (in parenthesis) associated with each cluster

Variables	Four resulting clusters			
	A (19 univ)	B (16 univ)	C (9 univ)	D (4 univ)
<i>Normalized documents</i>	4.6 \pm 2.6	7.1 \pm 1.7	4.9 \pm 1.2	10.6 \pm 4.4
<i>Normalized citations</i>	7.2 \pm 4.1	21.9 \pm 7.4	16.9 \pm 5.1	62.4 \pm 25.7
<i>Journal publication</i>	34.8 \pm 11.4	38.5 \pm 15.5	46.5 \pm 9.3	51.4 \pm 12.1
<i>First-quartile documents</i>	26.1 \pm 16.1	32.7 \pm 9.9	39.4 \pm 11.2	24.5 \pm 4.2
<i>Fourth-quartile documents</i>	19.2 \pm 12.4	15.9 \pm 4.6	9.6 \pm 3.6	15.8 \pm 7.9
<i>Citations per journal article</i>	3.4 \pm 1.2	6.3 \pm 1.8	6.7 \pm 1.2	12.2 \pm 4.4
<i>Citations per proceeding paper</i>	0.6 \pm 0.4	1.1 \pm 0.4	0.7 \pm 0.2	0.5 \pm 0.3
<i>International collaboration</i>	24.0 \pm 15.9	56.5 \pm 12.9	31.9 \pm 7.9	23.9 \pm 4.8

Numbers in boldface represent the highest value for each variable

Clustering validation is also concerned with checking the quality of clustering results using external validity indices like the Rand index (Rand 1971). The Rand index is used to measure the agreement between two different clusterings. The Rand index lies between 0 and 1. It takes the value of 1 when the two clusterings are identical. After running clustering validation, we found important similarities among the clustering algorithms: partitional vs hierarchical (0.8262), partitional vs probabilistic (0.8245), hierarchical vs probabilistic (0.7819). Note that the agreement between the hierarchical and probabilistic clustering pair had the lowest Rand index value, whereas the other clusterings had similar agreements. Table 6 compares the results of the three cluster algorithms (partitional, hierarchical and probabilistic) to see the robustness of the results. Universities listed in grey shaded rows were not grouped in the same cluster by all the cluster algorithms.

Each cluster can be defined according to different research activity aspects [e.g. productivity (documents per academic), visibility (citations per academic), quality (citations per journal articles and proceeding papers), prestige (first-quartile journals), and internationalization (international collaboration)]. We set global labels (high, medium-high, medium-low and low) for the values of each bibliometric index in the different clustering algorithms. Table 7 represents each cluster according to research activity aspects. Taking cluster B as an example, we found that the productivity of universities belonging to this cluster was medium-high but visibility was medium-low. Also, their values for quality, prestige and internationalization were medium-high, medium-high and high, respectively.

In order to summarize all results, we conclude that universities belonging to cluster D produce more scientific knowledge and have a bigger scientific impact than other universities. Universities belonging to cluster C usually publish in the most influential journals, and thus they have a selective strategy. In contrast, universities belonging to cluster B have an excellent ability to create international research publications, whereas universities belonging to cluster A do not stand out on any research activity aspect.

Finally, other variables like the *number of computer science theses* published during the 2005–2009 period, which was not used for the clustering, is also used as a external variable to describe the four resulting clusters. We found that cluster B had the highest value, followed by cluster D, cluster C and cluster A. Analyzing all universities, we also observed that the three top ranked universities for *number of computer science theses* were Universidad Politècnica de Catalunya (cluster B), Universidad Politècnica de València (cluster B) and UGR (cluster D).

Table 5 Cluster membership probability of each university

University	Cluster A	Cluster B	Cluster C	Cluster D
A Coruña (UDC)	0.98983	0.00950	0.00026	0.00040
Alcalá (UAH)	0.99902	0.00098	0.00000	0.00000
Alicante (UA)	0.02357	0.97587	0.00056	0.00000
Almería (UAL)	0.99987	0.00003	0.00002	0.00008
Autónoma de Barcelona (UAB)	0.05224	0.94776	0.00000	0.00000
Autónoma de Madrid (UAM)	0.00000	1.00000	0.00000	0.00000
Barcelona (UB)	0.99745	0.00020	0.00235	0.00000
Burgos (UBU)	0.00000	0.99723	0.00277	0.00000
Cádiz (UCA)	1.00000	0.00000	0.00000	0.00000
Cantabria (UC)	0.00000	0.99999	0.00001	0.00000
Carlos III de Madrid (UC3M)	0.97364	0.01380	0.00000	0.01257
Castilla-La Mancha (UCLM)	0.19839	0.80157	0.00004	0.00001
Complutense de Madrid (UCM)	0.00000	1.00000	0.00000	0.00000
Córdoba (UCO)	0.00000	0.00003	0.99997	0.00000
Extremadura (UNEX)	1.00000	0.00000	0.00000	0.00000
Girona (UDG)	0.00000	1.00000	0.00000	0.00000
Granada (UGR)	0.00000	0.00000	0.00000	1.00000
Huelva (UHU)	1.00000	0.00000	0.00000	0.00000
Illes Balears (UIB)	0.00008	0.00246	0.99746	0.00000
Jaén (UJA)	0.00000	0.00000	0.00000	1.00000
Jaume I de Castellón (UJI)	0.00035	0.99862	0.00103	0.00000
La Laguna (ULL)	0.99931	0.00044	0.00025	0.00000
La Rioja (UR)	0.94613	0.05385	0.00002	0.00000
Las Palmas de Gran Canaria (ULPGC)	1.00000	0.00000	0.00000	0.00000
León (ULE)	1.00000	0.00000	0.00000	0.00000
Lleida (UDL)	0.99801	0.00199	0.00000	0.00000
Málaga (UMA)	0.00000	0.99999	0.00001	0.00000
Miguel Hernández de Elche (UMH)	0.99999	0.00001	0.00000	0.00000
Murcia (UM)	0.07037	0.00822	0.92141	0.00001
Nacional de Educación a Distancia (UNED)	0.09010	0.02112	0.88724	0.00154
Oviedo (UNIOVI)	0.08176	0.02303	0.89521	0.00000
Pablo Olavide de Sevilla (UPO)	0.00000	0.00000	0.00000	1.00000
País Vasco (EHU)	0.00000	0.00886	0.99110	0.00003
Politécnica de Cartagena (UPCT)	0.99994	0.00000	0.00000	0.00006
Politécnica de Catalunya (UPC)	0.00000	1.00000	0.00000	0.00000
Politécnica de Madrid (UPM)	0.00008	0.03569	0.96423	0.00000
Politécnica de València (UPV)	0.00000	0.99949	0.00051	0.00000
Pompeu Fabra (UPF)	0.00000	1.00000	0.00000	0.00000
Pública de Navarra (UPNA)	0.00000	0.00001	0.00000	0.99999
Rey Juan Carlos (URJC)	0.82268	0.12601	0.00047	0.05084
Rovira i Virgili (URV)	0.00000	0.99930	0.00070	0.00000
Salamanca (USAL)	0.96119	0.01620	0.02137	0.00125
Santiago de Compostela (USC)	0.00000	0.01375	0.98619	0.00006

Table 5 continued

University	Cluster A	Cluster B	Cluster C	Cluster D
Sevilla (US)	0.00394	0.99606	0.00000	0.00000
València (UV)	0.00117	0.00068	0.99815	0.00000
Valladolid (UVA)	0.97670	0.00070	0.02261	0.00000
Vigo (UVIGO)	1.00000	0.00000	0.00000	0.00000
Zaragoza (UZ)	0.01693	0.95124	0.03183	0.00000

Numbers in boldface represent the highest probability for each university

Spanish public university academic staff

We calculated all bibliometric indices for all 2004 academics and report the minimum and maximum value of the distribution of each selected bibliometric index: *Total documents* [0, 178], *Total citations* [0, 4570], *Journal publication* [0, 100], *First-quartile documents* [0, 100], *Fourth-quartile documents* [0, 100], *Citations per journal article* [0, 82.5], *Citations per proceeding paper* [0, 16.0], and *International collaboration* [0, 100]. Taking *Total citations* as an example, we found that 0 was the lowest number of citations received by a specific academic, whereas 4,570 was the highest value.

We also performed an internal clustering validation to find the optimal number of clusters for our academic staff dataset. After running clustering validation, we found that the partition with six clusters ($k = 6$) had the highest silhouette coefficient. In this way, we run a partitioning clustering algorithm (partitioning around medoids) setting the number of clusters to six. We did not perform hierarchical and probabilistic clusterings for space reasons. Figures associated with these cluster analyses were very big for representing 2004 academics.

Table 8 shows the number of academics (in parentheses) associated with each cluster and the mean and standard deviation values for each variable within the six resulting clusters. We observed that the number of academics belonging to each cluster were: cluster A (321 academics), cluster B (839 academics), cluster C (416 academics), cluster D (166 academics), cluster E (248 academics), and cluster F (14 academics). Analyzing the variable values, we found that there were some differences among clusters. Taking *Total documents* as an example, we observed that academics belonging to cluster F had the highest mean value (74.5 ± 39.8). They also stood out on *Total citations*, *Journal publication* and *Citations per journal article*. Academics associated with cluster E excelled in terms of *First-quartile documents*, whereas academics associated with cluster C excelled with respect to *Fourth-quartile documents*. Finally, academics in cluster D had the highest value of *Citations per proceeding paper* and academics belonging to cluster A stood out on *International collaboration*.

The clusters obtained with partitioning around medoids were visually inspected using a representation in a lower dimensional space (see Fig. 4). The goal was to obtain a three-dimensional representation that approximates our eight-dimensional variables and check whether or not the clusters were visually distinguishable. A principal component analysis (Pearson 1901) was performed, and we studied the three principal components which account for the highest proportion of variance (95.0 %).

Figure 4 plots the values of the bibliometric indices for each academic in the transformed three-dimensional space. Different symbols and colors were used to show the cluster assigned by the clustering algorithm to each academic. Two-dimensional projections were also included for ease of interpretation. The first principal component (1st PC),

Table 6 Comparisons among three different clustering results

University	Partitional	Hierarchical	Probabilistic
A Coruña (UDC)	A	A	A
Alcalá (UAH)	B	B	A
Alicante (UA)	B	B	B
Almería (UAL)	A	A	A
Autónoma de Barcelona (UAB)	B	B	B
Autónoma de Madrid (UAM)	B	B	B
Barcelona (UB)	C	C	A
Burgos (UBU)	C	B	B
Cádiz (UCA)	A	A	A
Cantabria (UC)	B	B	B
Carlos III de Madrid (UC3M)	A	A	A
Castilla-La Mancha (UCLM)	B	B	B
Complutense de Madrid (UCM)	B	B	B
Córdoba (UCO)	C	C	C
Extremadura (UNEX)	A	A	A
Girona (UDG)	B	B	B
Granada (UGR)	D	D	D
Huelva (UHU)	A	C	A
Illes Balears (UIB)	C	C	C
Jaén (UJA)	D	D	D
Jaume I de Castellón (UJI)	B	B	B
La Laguna (ULL)	A	A	A
La Rioja (UR)	B	A	A
Las Palmas de Gran Canaria (ULPGC)	A	C	A
León (ULE)	A	A	A
Lleida (UDL)	A	C	A
Málaga (UMA)	B	B	B
Miguel Hernández de Elche (UMH)	A	C	A
Murcia (UM)	C	C	C
Nacional de Educación a Distancia (UNED)	A	C	C
Oviedo (UNIOVI)	C	C	C
Pablo Olavide de Sevilla (UPO)	D	D	D
País Vasco (EHU)	C	C	C
Politécnica de Cartagena (UPCT)	A	C	A
Politécnica de Catalunya (UPC)	B	B	B
Politécnica de Madrid (UPM)	C	C	C
Politécnica de València (UPV)	B	B	B
Pompeu Fabra (UPF)	B	B	B
Pública de Navarra (UPNA)	D	D	D
Rey Juan Carlos (URJC)	A	A	A
Rovira i Virgili (URV)	C	C	B
Salamanca (USAL)	A	A	A
Santiago de Compostela (USC)	C	C	C

Table 6 continued

University	Partitional	Hierarchical	Probabilistic
Sevilla (US)	B	B	B
València (UV)	C	C	C
Valladolid (UVA)	C	C	A
Vigo (UVIGO)	A	C	A
Zaragoza (UZ)	C	C	B

Table 7 Definition of clusters regarding different research activity aspects

	Cluster A	Cluster B	Cluster C	Cluster D
Productivity	Medium-low	Medium-high	Medium-low	High
Visibility	Low	Medium-low	Medium-low	High
Quality	Medium-low	Medium-high	Medium-high	High
Prestige	Medium-low	Medium-high	High	Medium-low
Internationalization	Medium-low	High	Medium-high	Medium-low

which accounted for 85.9 % of the variance, distinguished academics in cluster D and cluster F from the other clusters. The second principal component (2nd PC) distinguished academics belonging to cluster A from cluster B, and accounted for 5.9 % of the variance. Finally, the third principal component (3rd PC), which accounted for 3.2 % of the variance, distinguished between academics belonging to cluster E and cluster A. It also distinguished between academics belonging to cluster E and cluster B.

Table 9 shows the number of academics at each university belonging to each of the six clusters. Taking cluster F as an example, we found that its 14 members were: 1 academic from UDG, 8 academics from UGR, 1 academic from UJA, 1 academic from UPM, 2 academics from Universidad Politècnica de València and 1 academic from Universidad Pública de Navarra. Moreover, the biggest cluster B was composed mainly of a group of academics from the UPM (106), Universidad Politècnica de València (70) and Universidad de Las Palmas de Gran Canaria (45).

In order to summarize all results, we found that academics belonging to cluster F usually produced more scientific knowledge and had more impact than other academics. These academics had the highest impact in terms of journal articles, whereas academics belonging to cluster D excelled with respect to proceeding papers. Academics belonging to cluster E usually published in the most influential journals, whereas academics belonging to cluster C usually published in journals with lower impact factors. Academics belonging to cluster A stood out for their ability to author international research publications. Finally, academics belonging to cluster B did not stand out on any research activity aspect.

By areas (CAT, CSAI, CLS) and positions (FP, AP1, AP2, AP3) associated with each academic, Figure 5 shows the distribution of academics grouped in each cluster. For example, we found that cluster F had 14 members, 4 of whom (28.6 %) work on CAT, 8 (57.1 %) on CSAI, and 2 (14.3 %) on CLS. Also, we noted that cluster F was composed of 4 FP working on CAT, 6 FP and 2 AP1 working on CSAI, and 1 FP and 1 AP1 working on CLS. Figure 5 also shows that cluster A, cluster B, cluster C and cluster D were mainly

Table 8 Mean \pm standard deviation values for each variable within the six clusters (A, B, C, D, E and F) and the number of academics (in parentheses) associated with each cluster

Variables	Six resulting clusters					
	A (321)	B (839)	C (416)	D (166)	E (248)	F (14)
<i>TD</i>	14.0 \pm 9.4	2.5 \pm 4.8	9.2 \pm 8.9	33.1 \pm 20.4	11.0 \pm 8.1	74.5 \pm 39.8
<i>TC</i>	26.4 \pm 23.2	2.3 \pm 6.5	19.8 \pm 21.5	175.5 \pm 89.0	21.0 \pm 20.4	1249.4 \pm 1071.7
<i>JP</i>	37.7 \pm 24.9	2.8 \pm 7.6	55.9 \pm 26.5	54.7 \pm 22.1	42.1 \pm 24.5	69.6 \pm 14.7
<i>Q1</i>	23.6 \pm 28.6	0.0 \pm 0.9	5.6 \pm 10.8	31.7 \pm 19.9	70.6 \pm 25.6	37.4 \pm 15.0
<i>Q4</i>	14.2 \pm 23.5	0.0 \pm 0.0	28.4 \pm 36.1	11.8 \pm 13.0	3.9 \pm 11.1	8.9 \pm 7.7
<i>CJ</i>	4.3 \pm 6.4	0.4 \pm 1.7	4.1 \pm 5.0	11.0 \pm 8.6	4.5 \pm 6.3	23.5 \pm 8.7
<i>CP</i>	0.8 \pm 0.9	0.3 \pm 0.9	0.6 \pm 0.8	1.4 \pm 1.6	0.6 \pm 0.7	0.9 \pm 0.7
<i>IC</i>	83.5 \pm 20.0	0.6 \pm 4.3	5.5 \pm 12.3	48.0 \pm 31.7	8.7 \pm 16.5	39.8 \pm 19.5

Numbers in boldface represent the highest value for each variable

TD total documents, *TC* total citations, *JP* journal publication, *Q1* first-quartile documents, *Q4* fourth-quartile documents, *CJ* citations per journal article, *CP* citations per proceeding paper, *IC* international collaboration

composed of CLS academics, whereas members of clusters D and F were mainly CSAI academics. Taking into account academic positions, cluster A, cluster C, and cluster E were mainly composed of AP1, cluster B was mainly composed of AP1 and AP3, cluster D was mainly composed of FP and AP1, and finally, cluster F was mainly composed of FP.

Discussion and conclusions

Our work proposes a cluster analysis methodology to evaluate the research activity (in terms of bibliometric indices) of institutions and their academic staff. This paper focuses on the study of Spanish public universities and academics working in the computer science field, but we believe that this methodology can also be applied in other academic settings as well as in other research areas and countries.

The proposed methodology offers a series of advantages when it is compared to the classic peer review methodologies. Specially, our methodology does not suffer limitations related to subjectivity since it does not depend on the quality judgment of experts, it is an objective technique for assessing research performance. It also overcomes the traditional limits of bibliometric analyses based on simple rankings and permits a multi-dimensional cluster analysis at different levels.

This cluster analysis methodology groups similar universities or academics in the same cluster, maximizing within-cluster homogeneity and between-cluster heterogeneity. These results are useful for characterizing the research activity of universities and their academic staff. Three well-known clustering approaches (partitional, hierarchical and probabilistic) are used to give a comprehensive overview of the current situation by means of their useful different outputs (cluster medoids, dendrograms and cluster probabilities, among others). Other clustering approaches, such as combinatorial search-based techniques, kernel-based techniques, graph theory-based techniques, neural networks-based techniques and fuzzy techniques, have not been used in this paper. Further analysis, including the above approaches, could give a more sophisticated overview. Regarding clustering validation, the silhouette coefficient and Rand index were used to determine the optimal number of

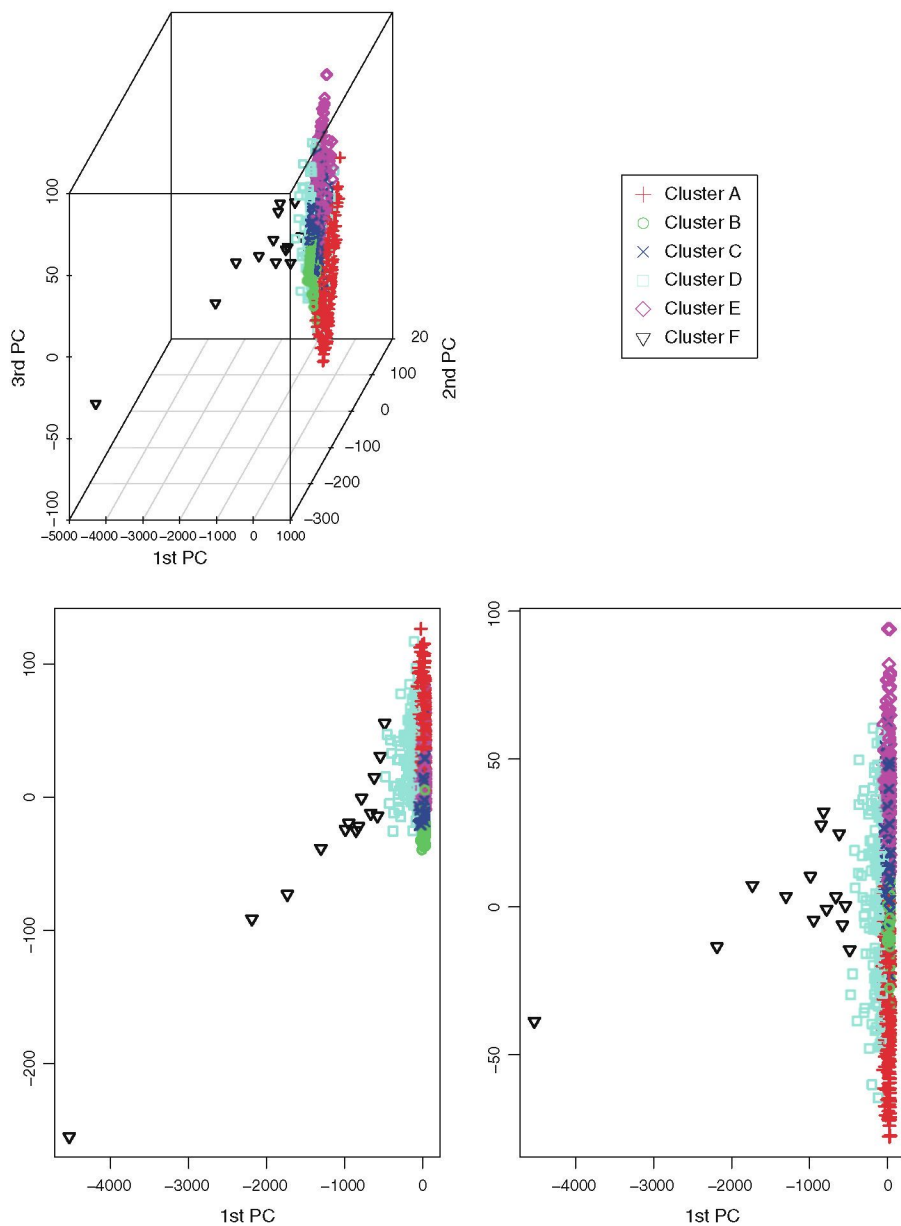


Fig. 4 Visualization of the academic clusters in three and two-dimensional spaces obtained with principal component analysis

clusters and the agreement between two different partitions, respectively. We have also used other internal indices (e.g. Dunn index) and external indices (e.g. Adjusted Rand index), but the results (not shown) did not vary so much.

Spanish public universities were grouped into four different clusters. Universities that belong to cluster D (UGR, UJA, UPO de Sevilla and Universidad Pública de Navarra)

Table 9 Number of academics within the six clusters by universities

Academics from university	Six resulting clusters					
	A (321)	B (839)	C (416)	D (166)	E (248)	F (14)
A Coruña	6	16	15	0	8	0
Alcalá	6	23	9	1	2	0
Alicante	13	30	14	4	16	0
Almería	3	21	8	0	5	0
Autónoma de Barcelona	12	12	11	7	4	0
Autónoma de Madrid	10	6	5	3	3	0
Barcelona	0	3	1	0	3	0
Burgos	1	7	0	1	0	0
Cádiz	1	22	9	0	0	0
Cantabria	5	2	4	1	1	0
Carlos III de Madrid	4	2	13	1	7	0
Castilla-La Mancha	12	18	15	7	15	0
Complutense de Madrid	17	19	11	9	9	0
Córdoba	0	6	3	5	6	0
Extremadura	6	29	8	0	3	0
Girona	13	8	2	2	1	1
Granada	8	11	27	33	6	8
Huelva	0	6	2	0	0	0
Illes Balears	5	23	6	5	8	0
Jaén	2	12	7	1	4	1
Jaume I de Castellón	18	24	10	3	3	0
La Laguna	6	4	7	0	0	0
La Rioja	1	3	2	0	0	0
Las Palmas de Gran Canaria	0	45	6	0	7	0
León	0	3	4	0	0	0
Lleida	4	3	3	0	2	0
Málaga	12	30	22	13	15	0
Miguel Hernández de Elche	0	5	1	0	1	0
Murcia	7	22	15	3	6	0
Nacional de Educación a Distancia	3	9	9	2	3	0
Oviedo	7	32	7	1	4	0
Pablo Olavide de Sevilla	0	0	0	1	2	0
País Vasco	5	37	13	6	8	0
Politécnica de Cartagena	0	5	1	0	2	0
Politécnica de Catalunya	45	27	29	20	22	0
Politécnica de Madrid	19	106	30	10	11	1
Politécnica de València	41	70	32	13	17	2
Pompeu Fabra	2	0	0	0	1	0
Pública de Navarra	3	1	3	1	6	1
Rey Juan Carlos	3	3	10	0	4	0
Rovira i Virgili	0	15	5	2	1	0

Table 9 continued

Academics from university	Six resulting clusters					
	A (321)	B (839)	C (416)	D (166)	E (248)	F (14)
Salamanca	1	20	1	1	2	0
Santiago de Compostela	1	12	8	4	4	0
Sevilla	6	18	10	3	4	0
València	3	21	6	1	4	0
Valladolid	6	19	3	0	4	0
Vigo	0	14	4	0	2	0
Zaragoza	4	15	5	2	12	0

score highest for the following research activity aspects: productivity, visibility and quality. Universities belonging to cluster C (UCO, Universidad del País Vasco, and UPM, among others) excel in terms of prestige, whereas universities belonging to cluster B (UDG, Universidad Politécnica de València, and UPF, among others) stand out on international collaboration. Finally, universities belonging to cluster A have worse scores for research activity aspects than the other universities.

Unlike Bornmann and Leydesdorff (2012) who showed that northern cities perform better than southern cities in some countries like Italy, we found that most of universities belonging to cluster D, which score highest for productivity, visibility and quality, are southern universities. In contrast, we found that some northern universities like Universidad Autónoma de Barcelona, Universidad de Cantabria, UDG, Universidad Politécnica de Catalunya stand out on international collaboration, whereas Universidad de Oviedo, Universidad del País Vasco, and Universidad Santiago de Compostela excel in terms of prestige.

Spanish computer science output originates mainly in higher education institutions. Analyzing Spanish university results, we find that they do not stand out for their quality. Citations per document is used as an indicator of Spanish computer science research quality in order to compare Spanish universities with other international universities. According to Essential Science Indicators, we found that ten Spanish universities rank in the top 350 positions, but only two (Universidad de Barcelona and Universidad de Vigo) are among the top 100 for citations per document. A possible reason for this situation is the constant cuts in the Spanish science budget (Pain 2012).

The cluster analysis methodology grouped Spanish academics into six different clusters: cluster A (321 academics), cluster B (839 academics), cluster C (416 academics), cluster D (166 academics), cluster E (248 academics), and cluster F (14 academics). Each cluster can be summarized with respect to different research activity aspects. Academics belonging to cluster F excel in terms of productivity, visibility and quality, whereas academics belonging to cluster E and cluster A stand out for their prestige and internationalization, respectively. Other academics that belong to clusters B, C, and D score worse in terms of research activity aspects. Focusing on cluster F (the best in terms of productivity, visibility and quality), we find that academics from UDG, UGR, UJA, UPM, Universidad Politécnica de València and Universidad Pública de Navarra are members of cluster F. Also, this cluster is composed mainly by FPs of the CSAI area.

Agrait and Poves (2009) state that not all Spanish academics publish research. Even so, they have paid time for researching. Results show that 43.7 % of Spanish academics regularly publish documents or patents, etc. By positions (FP, AP1, AP2, AP3) associated

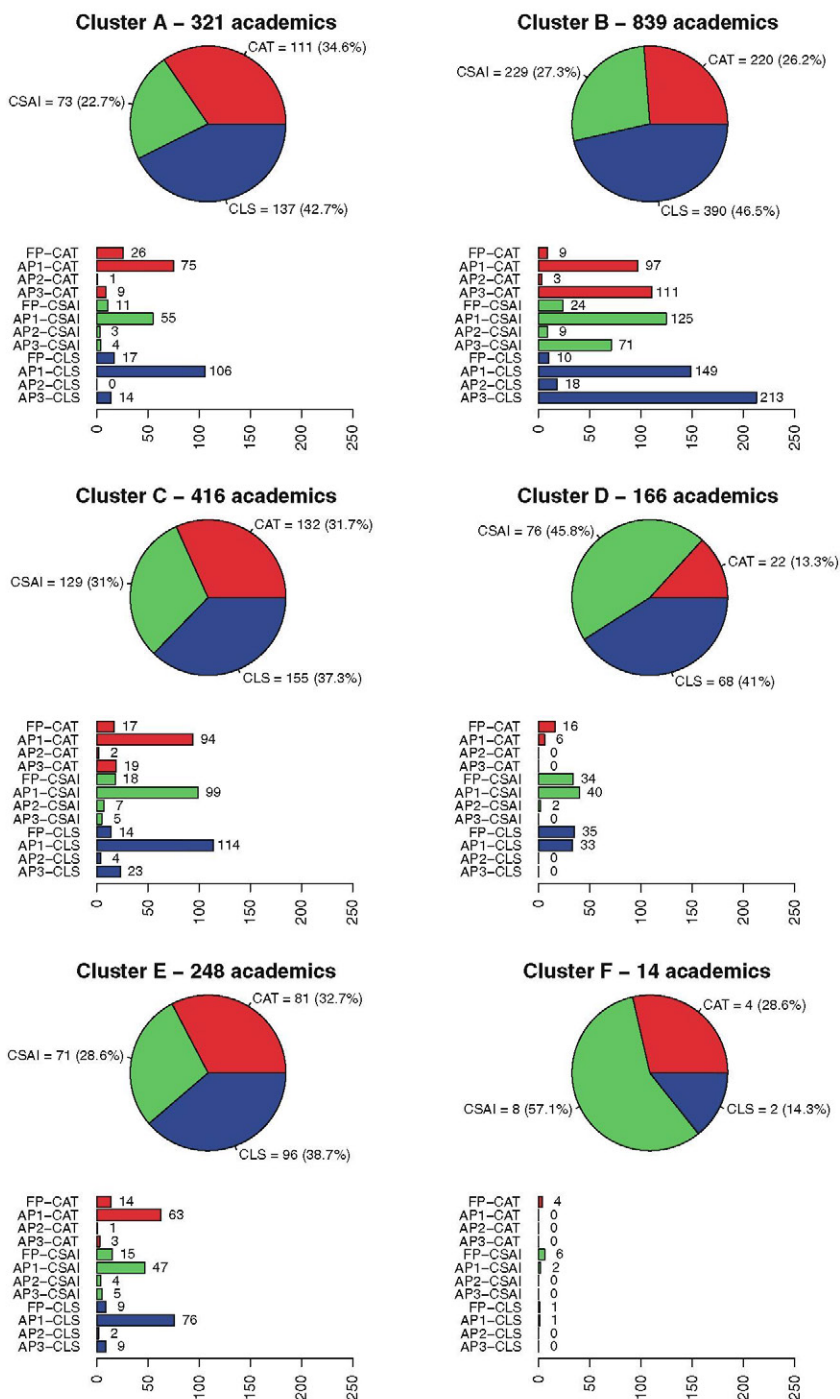


Fig. 5 Distribution of academics belonging to each cluster by areas and positions

with each academic, results show that 69.5 % of FP, 40.6 % of AP1, 21.5 % of AP2 and 4.9 % of AP3 usually do research. These results corroborate our findings in computer science. Cluster B includes the highest number of academics (839 out of 2004). Table 8 shows that the academics in cluster B have a low score for publications and citations. Also, this cluster is mainly composed of AP3 academics (see Fig. 5).

Our cluster analysis methodology can help institutions to compare themselves to each other and motivate them to improve their outcomes, since the proposed methodology characterizes research activity, identifying both their strengths and weaknesses. According to the results, academic researchers should improve the quality (number of citations per paper) in universities belonging to cluster A, the visibility (number of citations per academic) in universities belonging to cluster B, the productivity (number of publications per academic) in universities belonging to cluster C, and prestige (the percentage of documents published in first-quartile journals) in universities belonging to cluster D. On the other hand, academics belonging to AP2 and AP3 positions should increase their productivity and visibility, AP1 academics working in CAT and CLS should improve their quality, AP1 academics working in CSAI should publish in journals with higher impact factor, and finally, FP academics should collaborate with foreign institutions.

Using the cluster analysis methodology, policy-makers could propose collaborations and alliances among universities belonging to the same cluster. Several universities should perhaps merge strategically in order to compete with the most active international universities. In this way, Spanish universities could exploit their resources, enhance their reputation and visibility, and rise in the international rankings.

In the future, our target will be to incorporate private universities and non-tenured academics. Also, we will use other aspects (number of patents, number of projects, number of spin-offs, etc.) as variables in the cluster analysis. Finally, bibliometric analysis results are dependent on the selected source (ISI Web of Knowledge, Scopus, Google Scholar, etc). This is another point to be taken into account.

Acknowledgements This work was partially supported by the Spanish Ministry of Science and Innovation, grants TIN2010-20900-C04-04, Cajal Blue Brain and Consolider Ingenio 2010-CSD2007-00018.

- Abramo, G., & D'Angelo, C. A. (2011). National-scale research performance assessment at the individual level. *Scientometrics*, 86(2), 347–364.
- Abramo, G., D'Angelo, C. A., & Pugini, F. (2008). The measurement of Italian universities' research productivity by a non parametric-bibliometric methodology. *Scientometrics*, 76(2), 225–244.
- Agrait, N., Poves, A. (2009). Report on CNEAI assessment results. Technical report, National Evaluation Committee of Research Activity (in Spanish).
- Bornmann, L., & Leydesdorff, L. (2012). Which are the best performing regions in information science in terms of highly cited papers? Some improvements of our previous mapping approaches. *Journal of Informetrics*, 6(2), 336–345.
- Cheeseman, P., & Stutz, J. (1996). *Bayesian classification (autoclass): Theory and results*. Menlo Park: AAAI Press.
- Cobo, E., Selva O'Callaghan, A., Ribera, J., Cardellach, F., Dominguez, R., & Vilardell, M. (2007). Statistical reviewers improve reporting in biomedical articles: A randomized trial. *PLoS ONE*, 2(3), 332.
- Costas, R., VanLeeuwen, T. N., & Bordons, M. (2010). A bibliometric classificatory approach for the study and assessment of research performance at the individual level: The effects of age on productivity and impact. *Journal of the American Society for Information Science and Technology*, 61(8), 1564–1581.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the

- Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster analysis*. London.: Arnold.
- Fraley, C., Raftery, A. (1999). *Mclust: Software for model-based cluster and discriminant analysis*. Technical report, Department of Statistics, University of Washington.
- Garfield, E. (1996). The significant scientific literature appears in a small core of journals. *The Scientist*, 10(17), 13.
- Halkitis, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2/3), 107–145.
- Hanks, G. (2005). Peer review in action: the contribution of referees to advancing reliable knowledge. *Palliative Medicine*, 19(5), 359–370.
- He, Y., & Guan, J. C. (2008). Contribution of Chinese publications in computer science: A case study on LNCS. *Scientometrics*, 75(3), 519–534.
- Horrobin, D. (2001). Something rotten at the core of science. *Trends in Pharmacological Sciences*, 22(2), 51–52.
- Horrobin, D. L. (1990). The philosophical basis of peer review and the suppression of innovation. *Journal of the American Medical Association*, 263, 1438–1441.
- Ibáñez, A., Larrañaga, P., & Bielza, C. (2011). Using Bayesian networks to discover relationships between bibliometric indices. A case study of computer science and artificial intelligence journals. *Scientometrics*, 89(2), 523–551.
- Ibáñez, A., Bielza, C., Larrañaga, P. (2013). Relationship among research collaboration, number of documents and number of citations: A case study in Spanish computer science production in 2000–2009. *Scientometrics*. doi:10.1007/s11192-012-0883-6.
- Jain, A., & Dubes, R. (1988). *Algorithms for clustering data*. Englewood Cliffs: Prentice-Hall.
- Jain, A., Murty, M., & Flynn, P. (1999). Data clustering: A review. *ACM Computing Surveys*, 31(3), 264–323.
- Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: Wiley.
- Liu, G. (1968). *Introduction to combinatorial mathematics*. New York: McGraw-Hill.
- Maarek, Y. S., & BenShaul, I. Z. (1996). Automatically organizing bookmarks per contents. *Computer Networks and ISDN Systems*, 28(7–11), 1321–1333.
- MacRoberts, M. H., & MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics*, 36, 435–444.
- McLachlan, G., & Krishnan, T. (1997). *The EM algorithm and extensions*. New York: Wiley.
- McQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceeding of the Fifth Berkeley symposium on mathematical statistics and probability* (pp. 281–297).
- Moxham, H., Anderson, J. (1992). Peer review. A view from the inside. *Science and Technology Policy* 5(1), 7–15.
- Mulligan, A. (2005). Is peer review in crisis?. *Oral Oncology*, 41, 135–141.
- Pain, E. (2012). Research cuts will cause “exodus” from Spain. *Science*, 336(6078), 139–140.
- Palomares-Montero, D., García-Aracil, A. (2010). Fuzzy cluster analysis on Spanish public universities. In: *Investigaciones de Economía de la Educación, Asociación de Economía de la Educación* (Vol. 5, Chapt. 49, pp. 976–994).
- Pearson, K. (1901). On lines and planes of closest fit to systems of point in space. *Philosophical Magazine*, 2(6), 559–572.
- R Development Core Team. (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. <http://www.R-project.org>. Accessed 14 Nov 2011.
- Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336), 846–850.
- Rojas-Sola, J. I., Jorda-Albinana, B. (2009). Bibliometric analysis of Venezuelan publications in the computer sciences category of the JCR data base (1997–2007). *Interciencia*, 34(10), 689–695 (in Spanish).
- Rojas, R., & Gómez, I. (2006). Analysis of the Spanish scientific and technological output in the ICT sector. *Scientometrics*, 66(1), 101–121.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1), 53–65.
- Ruiz Pérez, R., Delgado-López-Cózar, E., & Jiménez-Contreras, E. (2002). Spanish personal name variations in national and international biomedical databases: Implications for information retrieval and bibliometric studies. *Journal of the Medical Library Association*, 90(4), 411–430.
- Ruiz Pérez, R., Delgado-López-Cózar, E., & Jiménez Contreras, E. (2010). Principles and criteria used by the National Evaluation Committee of Research Activity (CNEAI-Spain) for the assessment of scientific publications: 1989–2009. *Psicothema*, 22(4), 898–908.

- Scarpa, T. (2006). Peer review at NIH. *Science*, 311(5757), 41.
- Sneath, P. (1957). The application of computers to taxonomy. *Journal of General Microbiology*, 17(1), 201–226.
- Sorensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyzes of the vegetation on Danish commons. *Biologiske Skrifter*, 5(1), 1–34.
- Torres Salinas, D., Moreno Torres, J. G., Delgado-López-Cózar, E., & Herrera, F. (2011). A methodology for institution-field ranking based on a bidimensional analysis: the IFQ^2A index. *Scientometrics*, 88(3), 771–786.
- Torres-Salinas, D., Moreno-Torres, J. G., Robinson-García, N., Delgado-López-Cózar, E., Herrera, F. (2011). Rankings ISI of Spanish universities according to fields and scientific disciplines (2nd ed. 2011). *El Profesional de la Información*, 20(6), 701–709 (in Spanish).
- VanRaen, A. F. J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, 62(1), 133–143.
- Wainer, J., Xavier, E. C., & Bezerra, F. (2009). Scientific production in computer science: A comparative study of Brazil and other countries. *Scientometrics*, 81(2), 535–547.
- Wallace, C., Dowe, D. (1994). Intrinsic classification by MML-The SNOB program. In *Proceeding of the 7th Australian Joint Conference on artificial intelligence* (pp. 37–44).
- Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of American Statistical Association*, 58(301), 236–244.
- Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3), 645–678.