Stat Methods Appl (2013) 22:269–283
DOI 10.1007/s10260-012-0217-0

Robust analysis of bibliometric data

Francesca De Battisti · Silvia Salini

Accepted: 1 October 2012 / Published online: 20 October 2012

Abstract This work stems from the idea of describing the scientific productivity of Italian statisticians. There are several problems that must be addressed in achieving this goal: What data should be used? Have the data been cleaned? What techniques can be used? We propose the use of multiple sources and multiple metrics to get a complete information base. We check the correctness of the data using multivariate outlier identification techniques. We appropriately transform the data. We apply robust clustering to verify the existence of homogeneous groups. We suggest the use of *forward search* to establish a ranking among scholars. The proposed methodology, which, in this case, allowed us to group scholars into four homogeneous groups and sort them according to multidimensional data, can be applied to other similar applications in bibliometrics.

Keywords Bibliometric indicators · Multivariate transformation · Cluster analysis · Forward search

1 Introduction

Evaluation of university and scientific research has increasingly become important in recent years. In particular, there is a growing interest in evaluation of scientific publications and related bibliometric indicators (Marchant 2009). Bibliometrics is concerned with measuring the output of science. Commonly, it is believed that bibliometrics owes its systematic development mainly to D.J.D Price and Eugene Garfield, its founders. But Godin (2006) documents how the systematic counting of publications originated

F. De Battisti · S. Salini (✉)
Department of Economics, Management and Quantitative Methods, University of Milan, Milan, Italy
e-mail: silvia.salini@unimi.it

F. De Battisti
e-mail: francesca.debattisti@unimi.it

with psychologists in the early 1900s. Despite this, when one speaks of bibliometric indicators, it is still unclear which indicators are being referred to and which the statistical unit of analysis is. The theme of bibliometric indicators is complex because there are different aspects to consider Moed (2005). The first is the source of data. There are, as we shall see, different databases from which one can obtain bibliometric indicators, some of which are based on open access and web searches, some not free that consider only articles actually published; some disciplinary; and some generalist. There are also different levels depending on statistical unit to which they relate: there are indexes referring to authors, indexes referring to journals and indexes related to individual research products (Ferrara and Salini 2012). In order to obtain a publication list for each person, one must query bibliometric databases, export the data author by author, clean the data and integrate them to obtain bibliometric indicators. This operation is definitely time consuming and necessarily incorporates a margin of error. This is one of the reasons because, at the moment, scientific societies, universities and institutions in general are not able to make quick direct and unambiguous bibliometric analyses. In this work, we consider measures aggregated by author of Italian statisticians. We identify the databases from which one can get information on scientific production of Italian statisticians, and we evaluate the characteristics of different sources of data, in terms of consistency and correlations between the indicators obtained. The most relevant databases are

1. Current Index to Statistics (CIS): Created by the American Statistical Association and the Institute of Mathematical Statistics (http://www.statindex.org/); it considers only publications in statistics, probability and related topics.
2. Web of Science (ISI): Edited by the Institute for Scientific Information and distributed by Thomson Reuters (http://isiwebofknowledge.com/); it has a selective coverage of most relevant journals (and other literature sources).
3. Scopus (SCO): The mayor competitor of ISI (www.info.scopus.com), sponsored by Elsevier; it is more extensive than the ISI initiative.
4. Google Scholar (GS): Scientific research version of the famous search engine on the web; it is more extensive than the databases mentioned above, but its data quality is worse. Publish or Perish (POP) (http://www.harzing.com/pop.htm) is a interface for querying, which allows proper data cleaning.

Much of the bibliometric literature discusses the characteristics of different databases and the relations among bibliometric measures. Among others, Falagas et al. (2008) compared the content coverage and practical utility of Scopus, Web of Science, Google Scholar and PubMed (they were interested in medical databases). Bakkalbasi et al. (2006) used citation analysis in an observational study examining Google Scholar, Scopus and Web of Science to test the hypothesis that the three search tools lead to different citation counts. They observed that the question of which tool provides the most complete set of citing literature may depend on the subject and publication year of a given article. Norris and Oppenheim (2007) compared Web of Science, Scopus, Google Scholar and CSA Illumina (Cambridge Scientific Abstracts) in order to analyse the social sciences literature. From their analysis, they found that Scopus offers the best coverage from amongst these databases, and that it could be used as an alternative to Web of Sciences as a tool to evaluate the research impact on the social sciences.

Archambault et al. (2009) used macro-level bibliometric indicators to compare results obtained from Web of Science and Scopus. They showed that the correlations between the measures obtained with both databases for the number of papers and the number of citations received by countries, as well as their ranks, are extremely high. Franceschet (2010) made a comparison of bibliometric indicators' scores and citation-based rankings computed on Web of Science and Google Scholar; he also provided some advice on their use.

The aim of the paper is to achieve two different goals: on the one hand, we provide a classification of the authors, in order to identify similar profiles, using a robust approach; on the other hand, we propose the use of *forward search* as a method applicable to obtain a generalised ranking. We want to illustrate that the shortcomings of the single databases can be overcome by handling the same bibliometric indicators derived from different databases contemporaneously. In Sect. 2, the data set is described. In Sect. 3, we identify multivariate outliers and clean the data. In Sect. 3.1, data transformation is presented. In Sect. 3.2, clusters and profiles are obtained through a robust approach. In Sect. 3.3, a ranking is suggested using the *forward search* method. Finally, we provide some conclusions.

2 The data

As already mentioned, the aim of this work is to produce a synthesis of the scientific productivity of Italian statisticians by querying four international databases: CIS, SCO, WOS and POP. In particular we consider all Researchers in Statistics, SECS/S01 (444 Scholars). Information about Italian statisticians were downloaded from the Cineca (MIUR).[1]

A significant limitation of bibliometric databases is that they are not self-compiled by researchers, and consequently, because of homonyms, affiliation changes and updates, the results obtained are approximations. The second limitation, at least for some disciplines, is that there is not a complete and multi-disciplinary comprehensive database that includes all types of products (articles, proceedings and monographs). CIS, that is the most popular international database of journals in which articles about statistics and probability appear with extensive coverage, has one of the major limitations in updating times: for some journals, the last four to five year are missing. The ISI database, regarded by many as being representative of the entire research output, is also used as a reference by the SIS Commission for the reform of the recruitment mechanisms of teaching (http://sis-statistica.it/); it does not include some journals with statistical contents in which statisticians are used to publish: the subject category Statistics & Probability includes 110 journals versus CIS which presents 2619 journals. Scopus follows less restrictive technical criteria for the inclusion of journals and it includes a larger number of them (De Moya-Anegón et al. 2007). Our decision to include Google Scholar, despite its data quality being worse than other databases, is due to the fact that research products of different types are catalogued in it (Jacsò 2005). With regard to Italian statisticians, many of their works are not present on ISI

[1] http://cercauniversita.cineca.it.

and Scopus. This is confirmed by the lists of journals produced by the Italian Statistical Society[2] and by the ANVUR (National Agency for the Evaluation of Universities and Research Institutes)[3] that are based on the real frequency distribution of publications. The different structure of various sources suggests that different situations for the same subject can be identified. The analysis aims to assess the coherence of the information obtained. The data collected (from February until April 2010) for each author are as follows: number of publications, corresponding time period and, where available, total number of citations and value of the h-index (Hirsch Index, Hirsch 2005). The database created and used for the analysis is arranged by author and is composed of 10 variables on bibliometric databases (number of publications for ISI; number of citations for ISI; h-index for ISI; number of publications for SCO; number of citations for SCO; h-index for SCO; number of publications for POP; number of citations for POP; h-index for POP and number of publications for CIS). It is important to note that in this study, authors belong to the same field but they have different academic roles. Descriptive variables, such as title, university, faculty and so on, are also available. An alternative method to query the databases is to download information on the single research product, so a better quality of the data can be achieved; this is the topic of another one of our current projects. With the availability of the product database, it will be possible to make more advanced analyses, for example, network analysis of the authors, groups (departments, faculties and universities) (Rivellini et al. 2006) or journals (Baccini et al. 2009; Baccini and Barabesi 2011); analysis of benchmarking between researchers or research groups, based on the journal ranking; and comparison of the median/mean individual Impact Factor (IF) versus the median/mean IF of the corresponding area.

3 Bibliometric data analysis

A bibliometric database produces data that are not clean, even if some cleaning filters are applied during data collection. Having four databases gives us the possibility to use multivariate outlier detection procedures. Univariate outliers could be simply scholars who are more productive or less productive than others. Otherwise, a multivariate outlier, which is based on all available output, is represented by an unusual combination of the outputs of the four databases. It could be a great scholar or data that need to be checked. Since there are strong anomalies in the data downloaded from different databases, this procedure was done, at this step, so as not to corrupt the subsequent transformations. In order to detect anomalies and discrepancies between databases, in Fig. 1, we plotted the classical Mahalanobis distance of the data against the robust Mahalanobis distance, based on the *mcd* estimator (Filzmoser et al. 2008).

The algorithm applied has identified 23 multivariate outliers that have been manually explored, using, if needed, the curriculum vitae of the scholars. In particular, 14 of them were assessed as incorrect records due to a special character in the name, homonymy, a change of affiliation or an incorrect record in the database, while nine

[2] http://www.sis-statistica.it/index.php?area=main\&module=contents\&contentid=520.

[3] http://www.anvur.org/sites/anvur-miur/files/gev_documenti/allegato_gev13_1.zip.

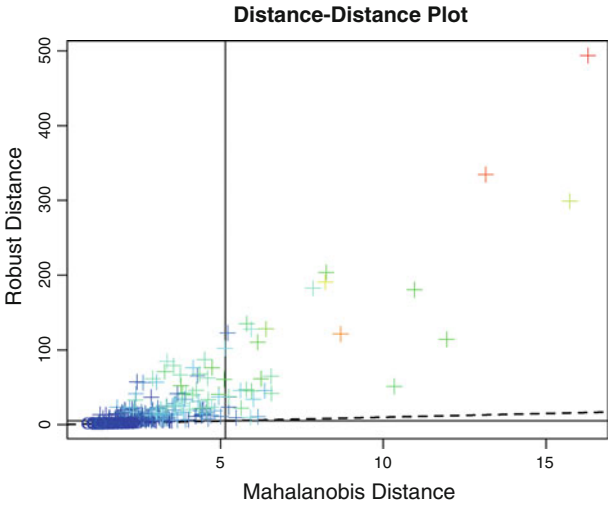

Fig. 1 Distance–distance plot: displays the robust distances versus the classical Mahalanobis distances. The dashed line is the set of points where the robust distance is equal to the classical distance. The horizontal and vertical lines are drawn at values equal to the cutoff which defaults to square root of the 97.5 distribution with p degrees of freedom. Points beyond these lines can be considered outliers

were due instead to the particular type of data analysed, because they are subjects with high values on the variables obtained, according to some sources over others.

Unproductive individuals,[4] who have nothing present in these databases, have been eliminated at this cleaning step. Records with only zeroes are not allowed in the following analysis.

At this point, data have been cleaned, and we can proceed with the next analysis.

3.1 Data transformation

A more detailed analysis of the data matrix allows us to make some considerations about the applicability of the traditional techniques of multivariate analysis. In the data matrix by author, there are a lot of zeros and the variable distributions are highly asymmetric with positive asymmetry; under these conditions, it is difficult to support the conjecture regarding the assumption of normality. Data in Fig. 2, which is the matrix of the scatterplots for all pairs of variables, do not seem to have the elliptical contours that would be expected from the pairwise bivariate normal distributions, and it is evident that there are many outliers.

It is necessary to identify a suitable transformation of the data (Emerson 1991). As in the Box–Cox transformation (1), (Box and Cox 1964), zeros in the data are not allowed; it is necessary to implement (2), as proposed by Yeo and Johnson (2000). In fact, the new transformation on the positive line is equivalent to the generalised Box–Cox transformation, $\{(x+1)^{\lambda}-1\}/\lambda$, for $x > -1$, where the shift constant 1 is included.

[4] 13 authors have zero occurrences for each database, 4 associate professors and 9 researchers.

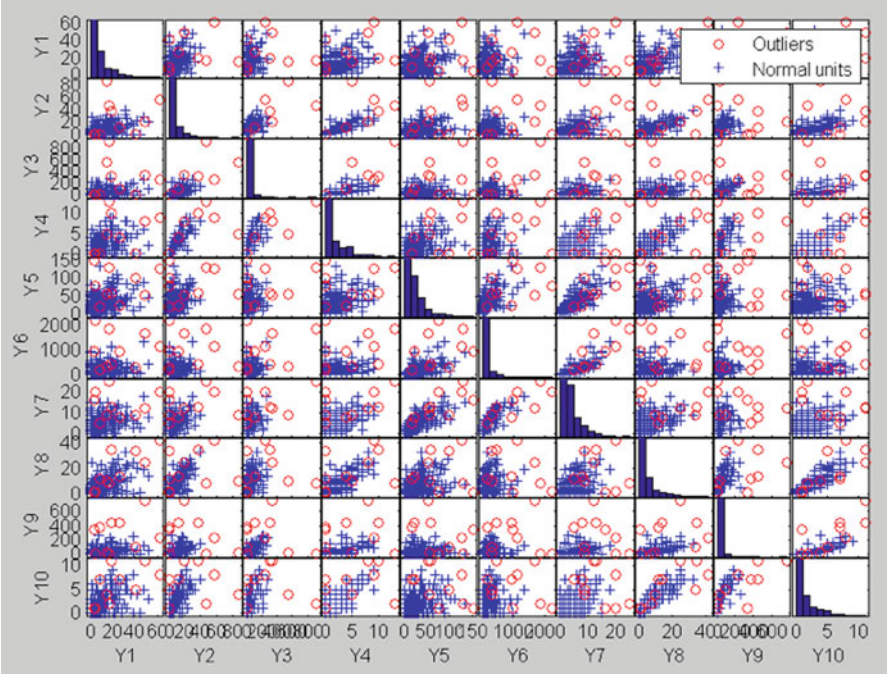

Fig. 2 Scatterplot matrix for the ten variables in the original scale

$$\psi^{BC}(\lambda, x) = \begin{cases} (x^{\lambda} - 1)/\lambda & (\lambda \neq 0) \\ \log(x) & (\lambda = 0) \end{cases} \tag{1}$$

$$\psi^{YJ}(\lambda, x) = \begin{cases} \{(x-1)^{\lambda} - 1\}/\lambda & (x \geqslant 0, \lambda \neq 0) \\ \log(x+1) & (x \geqslant 0, \lambda = 0) \\ \{(-x+1)^{2-\lambda} - 1\}/(2-\lambda) & (x < 0, \lambda \neq 2) \\ \log(-x+1) & (x < 0, \lambda = 2) \end{cases} \tag{2}$$

The transformation $Y1 = log(Y + 1)$ improves the closeness of the data to the normal distribution. However it may be that other transformations would give even better results. In order to test whether this occurs, we consider various transformations in the single parametric family; the aim is to obtain the best value for parameter λ, with respect to each variable considered.

It is not easy to immediately identify the optimal transformation for the data; a useful tool proposed for this purpose is the forward search procedure (Atkinson and Riani 2000). In particular we use the MATLAB toolbox *FSDA*.[5] This technique orders the observations from those most in agreement with a specified model to those least in agreement with it. The forward search estimators are effective in detecting masked multiple outliers, and more generally, in ordering data. Plots of diagnostic quantities

[5] http://www.riani.it/MATLAB.htm.

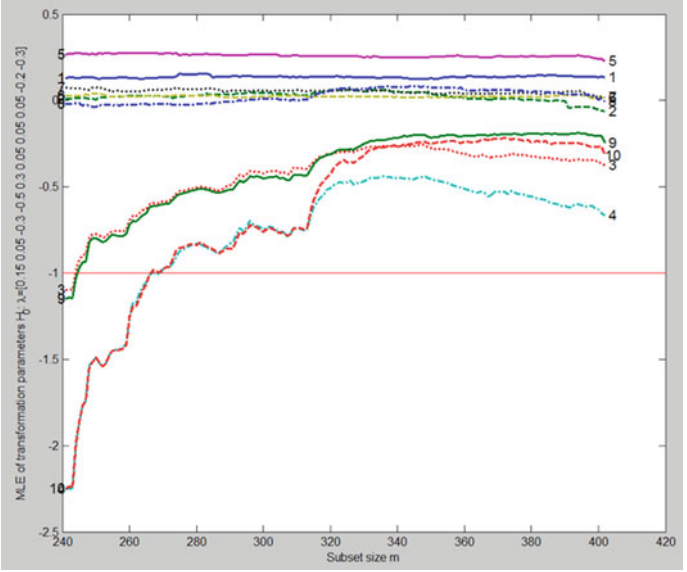

Fig. 3 Forward plot of the 10 elements of the maximum likelihood estimates $\hat{\lambda} = (0.15; 0.05; -0.4; -0.5; 0.3; 0.05; 0.05; 0.05; -0.2; -0.3)$

during the forward search clearly show the effect of individual observations on residuals and test statistics. This is a strength of the method. The transformation (2) has been used in the forward search (Atkinson et al. 2004, chapter 4) for the first time in this contribution.

We apply a forward search through the variables previously transformed ($Y2 = Y + 1$), estimating λ at each step.

With respect to each variable, the best value for the corresponding λ is obtained when the forward plot becomes stable. The forward plot of the maximum likelihood estimates of λ is in Fig. 3. The resulting values for the 10 elements of $\hat{\lambda}$ are as follows:

$$\hat{\lambda} = (0.15; 0.05; -0.4; -0.5; 0.3; 0.05; 0.05; 0.05; -0.2; -0.3).$$

How well defined these estimates of λ are can be determined from plots of the profile log-likelihood (Fig. 4). In Fig. 4, it is evident that the stability of the forward plot from a certain step of the procedure confirms the adequacy of the choice. In each panel, the values of the parameters are kept at their maximum likelihood estimates at step $m = 302$. The log-likelihoods are roughly parabolic and with maxima close to the proposed λ values rounded for each variable. All panels show a sharp definition of the estimates. The plot of likelihood ratio in Fig. 5 supports the transformation proposed—see, for example, Figures 4.10 and 4.11 of Atkinson and Riani (2000).

Finally, Fig. 6 shows the scatter plot matrix for the transformed variables; the outlier situation is improved; the univariate distributions are more symmetrical and the contours in the bivariate plots are more elliptical.

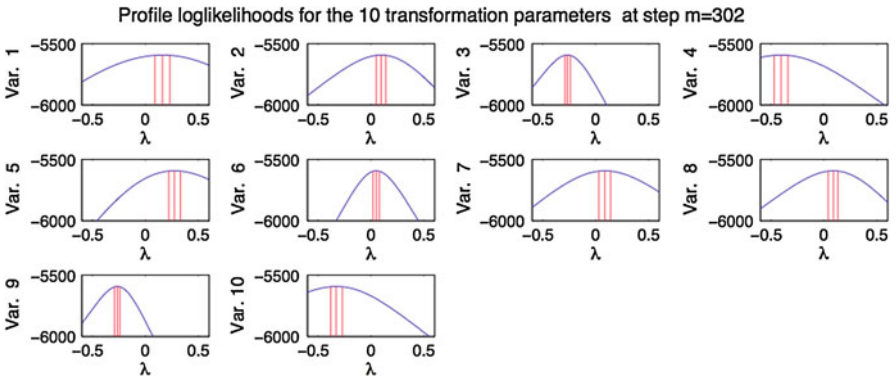

Fig. 4 Analysis of profile log-likelihood. The pairs of lines give asymptotic 95 % confidence intervals for each element of λ, based on asymptotic χ^2_1 distribution of twice the log-likelihood ratio

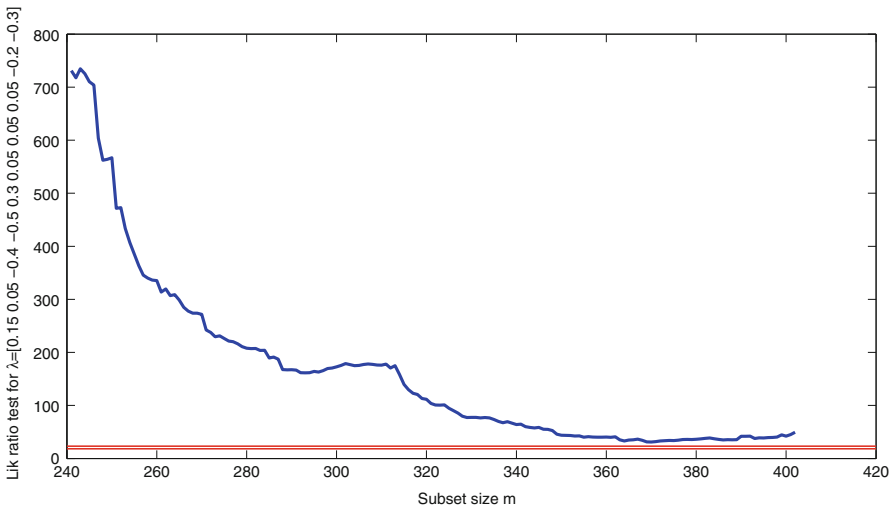

Fig. 5 Forward plot of the likelihood ratio test for the hypothesis of the transformation $\hat{\lambda} = (0.15; 0.05; -0.4; -0.5; 0.3; 0.05; 0.05; 0.05; -0.2; -0.3)$. The horizontal lines are the 95 and 99 % points of χ^2_{10}: this transformation is supported

3.2 Clusters and profiles

After the transformation of the data, it is interesting to see if groups of individuals with similar profiles exist.

To this end, we apply Model-based Methods of Classification (Fraley and Raftery 2002). This approach considers the problem of determining the structures of clustered data, without prior knowledge of the number of clusters or any other information about their compositions. Data are represented by a mixture model, in which each component corresponds to a different cluster. Models with varying geometric properties are obtained through Gaussian components with different parameterisations and cross-cluster constraints. Partitions are determined by the EM (expectation-maximisation)

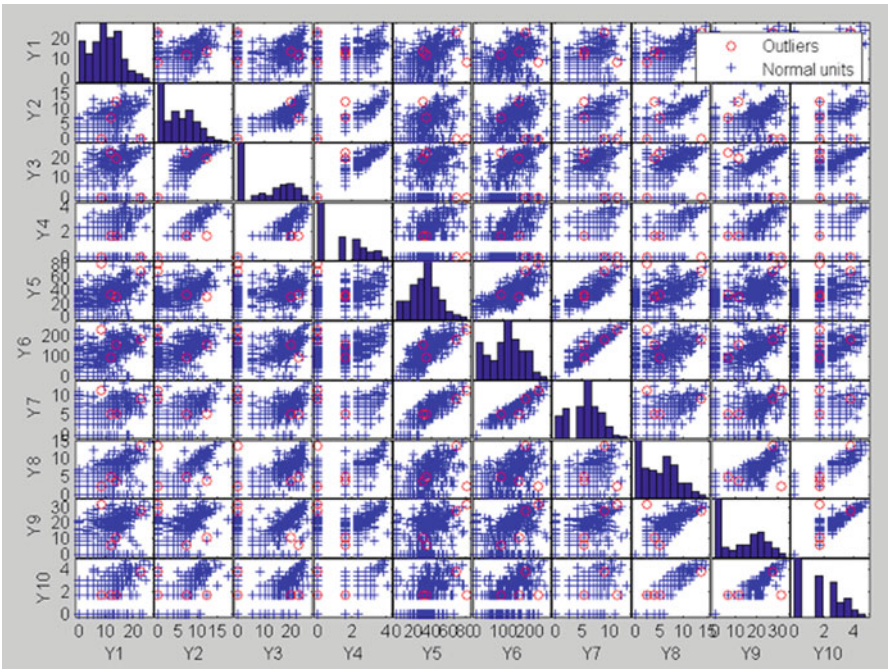

Fig. 6 Scatterplot matrix for the ten variables transformed $Y2 = Y + 1$ with $\hat{\lambda} = (0.15; 0.05; -0.4; -0.5; 0.3; 0.05; 0.05; 0.05; -0.2; -0.3)$

algorithm for maximum likelihood, with initial values from agglomerative hierarchical clustering. Models are compared using an approximation to the Bayes factor, based on the Bayesian Information Criterion (BIC). Unlike significance tests, this allows the comparison of more than two models at the same time, and removes the restriction that the models must be nested. The problems of determining the number of clusters and the clustering methods are simultaneously solved by choosing the best model.

This analysis is done using R library *mclust*.[6]

Figure 7 shows BIC from 10 different parameterisations of the covariance matrix in the Gaussian model and up to nine clusters. Different symbols and line types encode different model parameterisations. The *best* model is the one with the highest BIC among the fitted models. In this case, the best model is *VEV* with four clusters that correspond to ellipsoidal distributions with variable (V) volume, equal (E) shape and variable (V) orientation. For a description of the parametrisations of the covariance matrix in the Gaussian model and their geometric interpretation, see Banfield and Raftery (1993).

Following the best solution of the *mclust* algorithm, we explore it with four clusters. Table 1 gives the mean, the standard deviation (SD) and the median for the 10 variables by cluster and the number of units that composes each cluster. The quartiles are represented in Fig. 8. It is evident that the groups are ranked from the one with the lowest values for all the variables (cluster 1) to the one with the highest values (cluster 4).

[6] http://cran.r-project.org/web/packages/mclust/index.html.

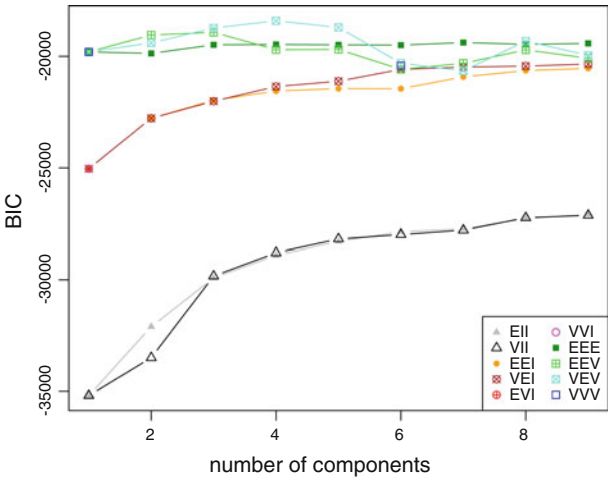

Fig. 7 BIC from *mclust* for the 10 available model parameterisations and up to nine clusters

Table 1 Mean, SD and median of the 10 variables by group

Groups	*N*	Statistics	Variables									
			pCIS	pSCO	cSCO	hSCO	pPOP	cPOP	hPOP	pISI	cISI	hISI
Clu 1	104	Mean	6.24	1.23	0.00	0.00	26.46	93.97	4.22	1.09	0.00	0.00
		SD	5.47	1.80	0.00	0.00	12.49	65.01	2.37	1.79	0.00	0.00
		Median	5.70	0.00	0.00	0.00	25.99	86.24	4.47	0.00	0.00	0.00
Clu 2	71	Mean	9.90	3.03	5.00	0.70	33.66	131.94	5.74	4.98	15.45	1.92
		SD	6.03	3.38	7.89	1.07	19.29	77.83	3.17	2.94	9.92	1.10
		Median	9.80	2.64	0.00	0.00	35.02	134.88	5.68	4.19	17.20	1.79
Clu 3	100	Mean	11.66	7.58	16.38	2.30	39.16	138.49	6.33	6.36	16.85	2.56
		SD	5.43	3.13	6.50	0.77	16.63	64.55	2.72	2.45	7.54	0.83
		Median	12.41	7.02	17.11	2.51	38.95	143.12	6.63	6.21	17.20	2.68
Clu 4	110	Mean	14.08	9.23	20.92	2.93	43.89	170.10	7.53	8.81	24.69	3.12
		SD	5.30	2.59	3.68	0.65	13.58	48.28	2.13	2.58	5.59	0.90
		Median	14.06	9.13	20.81	2.97	42.74	169.17	7.42	8.61	24.59	3.24

It must be noted that the considered indicators are not normalised. From the literature (Adler et al. 2009), we could expect that all indicators considered indeed tend to grow with the scientific age of researchers. If the aim is the comparative evaluation of scholars, then, obviously, the academic role has to be considered, as well as the scholar's scientific age. The latter, if it is not available, could be estimated by the year of the first publication, obtaining the average number of papers per year.[7] Choosing how to normalise the number of citations and the h-index can be rather complicated.

[7] The use of average instead of sum implicitly assumes a uniform distribution of academic publishing in the life cycle of individuals.

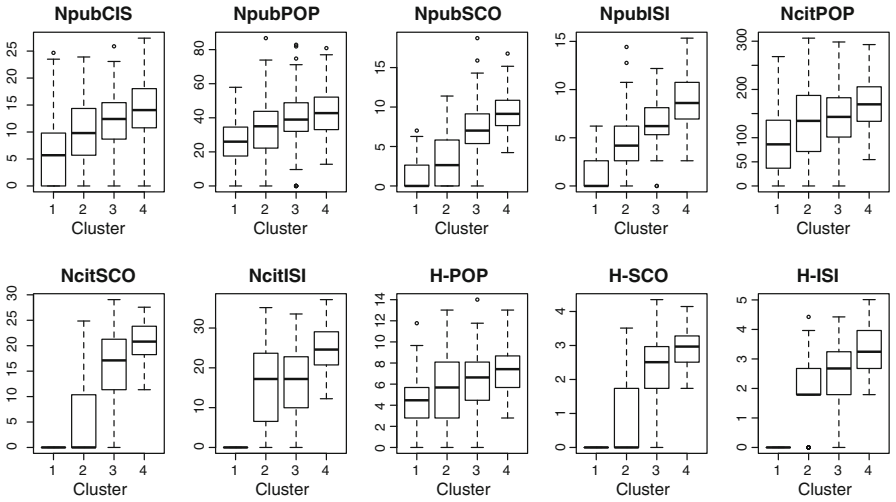

Fig. 8 The plots represent the boxplot of the 10 variables by cluster. The clusters are in the *horizontal axis*, the *vertical axis* shows the range of the relative variables. It can be noticed that the range changes according to measures (Npub, Ncit, H) and to databases (CIS, POP, SCO, ISI)

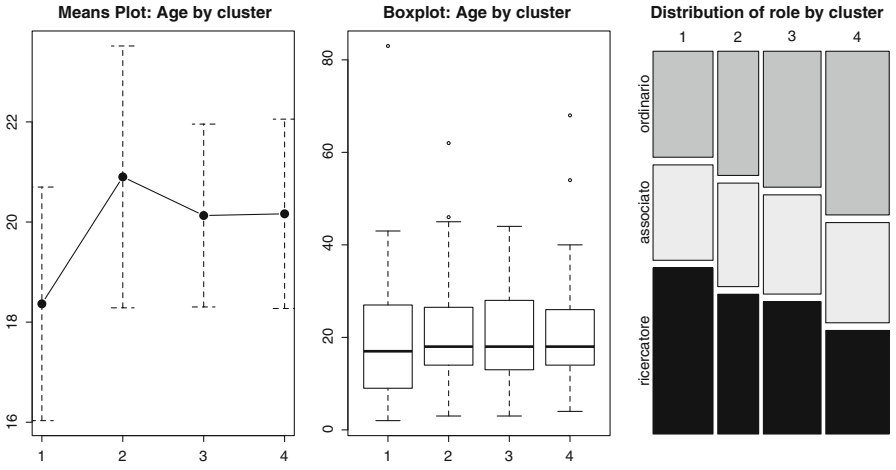

Fig. 9 The *first plot* contains the confidence intervals for the mean of age by cluster, the *second plot* shows the box plot of age by cluster and the *third* shows the distribution of roles by cluster

For the h-index variations are present, which take account of the number of authors (*individual h-index*, Batista et al. 2006), age of the individual (*m-quotient*, Hirsch 2005) and age of the papers (*contemporary h-index*, Katsaros et al. 2006).

In our exploratory exercise, in which the idea is to identify groups of scholars with similar profiles, the scientific age and academic role distribution could be considered as an outcome of the analysis. As shown in Fig. 9, the means of the age (based on the minimum first year of publication in the four databases) do not change significantly in the four clusters. The same is true for the median. The productivity does not increase

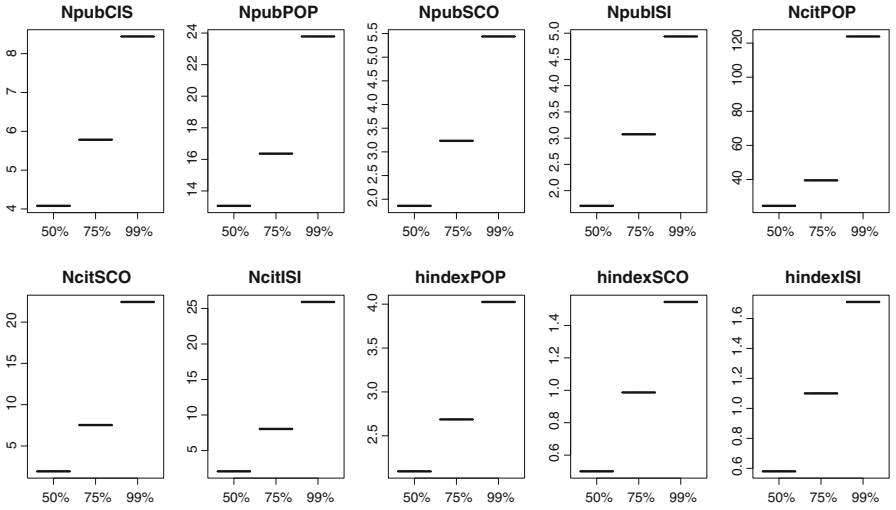

Fig. 10 Means of the variables for three steps of the forward search: 50, 75 and 99 %, the means are in the *vertical axis* and the forward search steps are in the *horizontal axis*

with increasing age. This may depend, in our opinion, on the changes happened during the last years in the methods of dissemination of research. Instead, the frequency distribution of the role changes in the four clusters; in particular, the percentages of full professor are 29, 34, 37 and 44 % respectively.

3.3 Ranking

When doing bibliometric analysis, one goal is to make a ranking of institutions or individuals. As previously mentioned in sub Sect. 3.1, the forward search in exploring multivariate data orders the observations from those closest to farthest from the bulk of the data (Atkinson et al. 2004, chapter 7). Proceeding with the analysis and applying the forward search, it is clear in Fig. 10 that the averages of the variables increase as the steps of the procedure increase (50, 75 and 99 %).

We propose to interpret the inclusion order of the units by forward search as a generalised ranking, where similar profiles (units entering in close steps) can be identified. In this case,[8] the bulk of the data, as shown in Fig. 2, is represented by unproductive units, with most indexes at or near zero. Looking at the averages for the various steps of the forward search for the ten variables shown in Fig. 9, we can conclude that individuals included later in the search have higher productivity levels. The last units to enter are outliers in the sense that they are individuals who have higher production levels than the others.

[8] The initial subset is computed by using robustly centered ellipses, by default. Only for non-transformed data, we expect that the bulk of the data is represented by unproductive scholars, for the positive asymmetry of the distributions. When the data are transformed to normality, the bulk becomes the center of the distribution and not the tails.

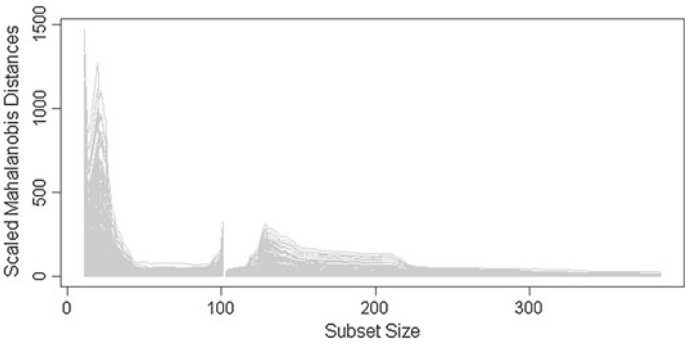

Fig. 11 Mahalanobis distance for each step of the search. Cluster 4 units are the initial subset

Table 2 Quantiles of inclusion order in the forward search by group, the initial subset consists of the units that belong to cluster 4

Group	N	Quantiles				
		Min	Q_1	Median	Q_3	Max
Clu 4	110	1	1	1	1	1
Clu 3	100	2	7	36	164	275
Clu 2	71	51	55	60	77	270
Clu 1	104	110	202	225	243	276

Another hypothesis we have checked is whether the inclusion order interpretation depends on the selection of the initial subset (Atkinson and Riani 2007). We now apply the forward search using the 110 units that belong to cluster 4 as an initial subset. We expect that it confirms the presence of groups; we also expect that the order of inclusion is consistent with the clusters identified above. Using the units of cluster 4 as the initial subset (i.e. the most productive), it should happen that the units belonging to cluster 3, closer to the units of cluster 4, enter in the search, for the most part, before those belonging to clusters 2 and 1. Figure 11 shows the Mahalanobis distance for each step of the search, and Table 2 shows the quartiles of the inclusion order in the search of the units.

Clear changes in the Mahalanobis distances indicate that a unit belonging to a new group enters in the search (Atkinson et al. 2006). The plot in Fig. 11 shows the presence of three groups over the initial subset, according to the Model-Based cluster in Fig. 7. It is important to note that although the groups are identified, they are quite dispersed. Most of the variables have high SDs (see Table 1). It would then not be reasonable to expect a clear separation. Table 2 shows, however, a consistent order of inclusion of units with respect to their cluster membership: 50 % of the units belonging to cluster 3 come before step 36, while 50 % of the units belonging to cluster 2 come before step 60; and 50 % of the units belonging to cluster 1 come before step 225. With respect to quartiles, 25 % of the units of cluster 1, the least productive scholars, come after step 243, while the first quartile for cluster 3 is 7, i.e. 25 % of the units of cluster 3 come

together before step 7—they are essentially very close to the initial subset. Even in this case, the order of inclusion produces a ranking from the most productive to the least.

4 Conclusion

In this work, we suggest handling the same bibliometric indicators derived from different databases in order to balance the shortcomings of a single database. Four international databases, CIS, ISI, SCO and POP, have been analysed in order to measure the scientific output of all Italian statisticians. Data are aggregated by author; thus, the only problem is to assess productivity and impact (through the citations and the h-index). We propose a procedure of multivariate outlier detection to identify errors due to an incorrect data download and to obtain a clean database; we implement, for the first time in literature, the generalised Box–Cox transformation in a forward search algorithm; we identify clusters/profiles of scholars with similar characteristics; finally, we propose the forward search procedure as an original method to establish a generalised ranking. In the near future, we want to carry out a simulation study in order to find empirical evidence that supports the use of forward search to produce a generalised ranking. We also want to study the distribution law of the Italian statistician productivity (Lotka 1926) and create a new database in which the units of analysis are the single research products or publications, instead of scholars. This database will allow more sophisticated analyses.

References

Adler R, Ewing J, Taylor P (2009) Citation statistics with discussion. Stat Sci 24:1–28

Archambault E, Campbell D, Gingras Y, Larivire V (2009) Comparing bibliometric statistics obtained from the web of science and Scopus. J Am Soc Inf Sci Technol 60(7):1320–1326

Atkinson AC, Riani M (2000) Robust diagnostic regression analysis. Springer, New York

Atkinson AC, Riani M (2007) Exploratory tools for clustering multivariate data. Comput Stat Data Anal 52:272–285

Atkinson AC, Riani M, Cerioli A (2004) Exploring multivariate data with the forward search. Springer, New York

Atkinson AC, Riani M, Cerioli A (2006) Random start forward searches with envelopes for detecting clusters in multivariate data. In: Zani S, Cerioli A, Riani M, Vichi M (eds) Data analysis classification and the forward search. Springer, Berlin

Baccini A, Barabesi L, Marcheselli M (2009) How are statistical journal linked? A network analysis. Chance 22(3):34–43

Baccini A, Barabesi L (2011) Seats at the table: the network of editorial boards in information and library sciences. J Infomet 5:382–391

Bakkalbasi N, Bauer K, Glover J, Wang L (2006) Three options for citation tracking: Google Scholar, Scopus and web of science. Biomed Digit Libr 3:7

Banfield JD, Raftery AE (1993) Model-based Gaussian and non-Gaussian clustering. Biometrics 49: 803–821

Batista PD, Campiteli MG, Konouchi O (2006) Is it possible to compare researchers with different scientific interests. Scientometrics 68(1):179–189

Box GEP, Cox DR (1964) An analysis of transformations. J R Stat Soc Ser B 26(2):211–252

De Moya-Anegón F, Chincilla-Rodriguez Z, Vargas-Qesada B, Corera-Álvarez E, JosèMunoz Fernandez FJ, Gonzáles-Molina A, Herrero-Solana V (2007) Coverage analysis of Scopus: a journal metric approach. Scientometrics 73(1):53–78

Emerson JD (1991) Introduction to transformation. In: Hoaglin DC, Mosteller F, Tukey JW (eds) Fundamentals of exploratory analysis of variance. Wiley, New York

Falagas ME, Pitsouni EI, Malietzis GA, Pappas G (2008) Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strenghts and weaknesses. FASEB J 22:338–342

Ferrara A, Salini S (2012) Ten challenges in modeling bibliographic data for bibliometric analysis. Scientometrics. doi:10.1007/S11192-012-0810-x

Filzmoser P, Maronna R, Werner M (2008) Outlier identification in high dimensions. Comput Stat Data Anal 52:1694–1711

Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis and density estimation. J Am Stat Assoc 97:611–631

Franceschet M (2010) A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar. Scientometrics 83(1):243–258

Godin B (2006) On the origins of bibliometrics. Scientometrics 68(I):109–133

Hirsch E (2005) An index to quantify an individual's scientific research output. In: PNAS. Proceedings of the National Academy of Sciences of the United States of America, Nov 15, vol 102, no 46

Jacsò P (2005) Google Scholar: the pros and the cons. Online Inf Rev 29(2):208–214

Katsaros C, Manolopoulos Y, Sidiropoulos A (2006) Generalized h-index for disclosing latent facts in citation networks. Retrieved 20 Dec 2008, from http://arxiv.org/abs/cs.DL/0607066

Lotka AJ (1926) The frequency distribution of scientific productivity. J Wash Acad Sci 16(12):317–324

Marchant T (2009) An axiomatic characterization of the ranking based on the h-index and some other bibliometric rankings of authors. Scientometrics 80(2):327–344

Moed HF (2005) Citation analysis in research evaluation. Springer, Berlin

Norris M, Oppenheim C (2007) Comparing alternatives to the Web of Science for coverage of the social sciences literature. J Infomet 1:161–169

Rivellini G, Rizzi E, Zaccarin S (2006) The science network in Italian population research: an analysis according to the social network perspective. Scientometrics 67:3

Yeo IK, Johnson RA (2000) A new family of power transformations to improve normality or symmetry. Biometrika 87(4):954–959