ROBUST ANALYSIS OF BIBLIOMETRIC DATA

FRANCESCA DE BATTISTI **SILVIA SALINI**

Working Paper n. 2011-36

OTTOBRE 2011

DIPARTIMENTO DI SCIENZE ECONOMICHE AZIENDALI E STATISTICHE

Via Conservatorio 7
20122 Milano
tel. ++39 02 503 21501 (21522) - fax ++39 02 503 21450 (21505)
http://www.economia.unimi.it
E Mail: dipeco@unimi.it

Noname manuscript No.
(will be inserted by the editor)

Robust Analysis of Bibliometric Data

Francesca De Battisti · **Silvia Salini**

Received: date / Accepted: date

Abstract The aim of the work is to reproduce the image of the research profile of the Italian statisticians derived from the querying of bibliometric databases. We highlighted the need for multiple sources in order to convey a truer picture and how data could be combined in order to have a classification or an index of the overall productivity, which took into account all sources and metrics. The data matrix contains a set of metrics from a variety of databases for each author and it is a sparse matrix (there are many zeros). Furthermore, the variables are leptokurtic and characterized by positive asymmetry. In order to apply the classical techniques of multivariate analysis, data must first be transformed or, alternatively, robust analysis techniques have to be used. In the paper we will focus our attention on this type of bibliometric data, describing their main characteristics and problems. In addition, a robust approach to the analysis of these data will be presented.

Keywords Bibliometric Indicators · Multivariate Transformation · Cluster Analysis · Forward Search

1 Introduction

In the recent years, in Italy, the interest in the evaluation of the research has been widespread. The first national exercise for the evaluation of the research was conducted by the CIVR (Committee for the evaluation of research) between 2001-2003

F. De Battisti
Department of Economics, Business and Statistics - University of Milan
Tel.: ++39 (0)2.50321464
Fax: ++39 (0)2.50321505
E-mail: francesca.debattisti@unimi.it

S. Salini
Department of Economics, Business and Statistics - University of Milan
Tel.: ++39 (0)2.50321538
Fax: ++39 (0)2.50321505
E-mail: silvia.salini@unimi.it

period; the Law of 9 January 2009 No. 1 states that some of the funds of the universities are allocated according to indicators related to the outcomes of the VTR (three year evaluation of research). Minister Gelmini stated that in the future an increasing share of the funds will be allocated on merit. To evaluate this research becomes therefore indispensable. An interesting paper considers the rankings of Italian university based on bibliometric indicators [21]. Two methods are typically used to assess the quality of academic publications: the peer review and the use of bibliometric indicators: despite the low prevalence of "bibliometric culture" in some countries, particularly for some disciplines, it should be noted that bibliometrics is an ancient discipline, as evidenced by [18]. As described in the article cited, both approaches have their advantages and limitations. The combined use of these tools (informed peer review) would probably be more apt and correct. Anglo-Saxon future evaluation exercises will move in this direction and it seems likely the new VQR (five year evaluation of research) will take into account not only the peer review but also bibliometric indicators. When one speaks of bibliometric indicators it is not clear to what indicators a reference is made and what is the statistical unit of the analysis. The theme of bibliometric indicators is complex because there are different aspects to consider [9]. First of all the source of data: there are, as we may see, different databases from which you can obtain bibliometric indicators, some of which easily accessible on the web, some others not free, that consider only articles actually published; some disciplinary, some generalists. There are also different levels of analysis depending on the statistical unit they are related to. There are indexes referring to the authors, indexes referring to journals and indices related to individual research products. Then, there are also different types of metrics, some indexes of quantity, some indexes of quality and other indicators of impact / dissemination / awareness. In our opinion a reliable and robust bibliometric analysis should begin with an appropriately structured database that takes into account all the available information, at all levels. A clear indication of how bibliometric data should be structured is CERIF (The Common European Research Information Format), when identifies a European standard for building research database [4] (http://www.eurocris.org). The idea is that there should be a single repository of research, complete, clean and public. In Italy there are two main systems for institutional repositories of universities: SURPLUS (CILEA) and U-Gov (CINECA). The former is open to access, the latter is not. At the moment neither system is CERIF compatible[1]. Now, in order to obtain a publication list for each person one should query bibliometric databases, export data author by authors, cleaning them and integrating them to obtain bibliometric indicators. This operation is definitely time consuming and necessarily incorporates a margin of error. This is one of the reasons why at the moment scientific societies, universities, institutions in general are not able to make quick, direct and advanced bibliometric analysis. In this work we consider measures aggregated by author (Italian Statisticians) and we aim 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 gen-

[1] The National Agency for the Evaluation of Universities and Research Institutes (ANVUR) plans to build, in the coming years, an Institutional Research Database.

eralized ranking. In section 2 the data set is described. In Section 3.1 data cleaning and transformation are presented. In Section 3.2 clusters and profiles are obtained through a robust approach. In Section 3.3 a ranking is suggested using the forward search. Some conclusions are given.

2 The data

As already mentioned, the aim of the work is to produce a synthesis of the scientific productivity of Italian Statisticians. Unfortunately, at the moment a single public database of all the research products it is not available. For this reason, the only way to have a comprehensive database about authors belonging to different structures is to use bibliometric databases. The first limitation of these databases is that they are not self-compiled by researchers and consequently - because of homonyms, affiliations change and updates - the results obtained are approximations. The second limit, 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, monographs). A way to get a good approximation is to use more than one database. As first exercise, we queried four international databases, in order to know the scientific output of all Researchers in Statistics, SECS/S01 (444 Subjects). In particular:

1. Current Index to Statistics (CIS), created by the American Statistical Association and the Institute of Mathematical Statistics (http://www.statindex.org/); only publications in statistics get considered, 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 Web of Science, sponsored by Elsevier (www.info.scopus.com); it is more extensive than ISI initiative.
4. Google Scholar (POP), scientific research version of the famous search engine on the web; recommended interface for querying, which allows proper data cleaning, is Publish or Perish (http://www.harzing.com/pop.htm), developed by Anne-Wil Harzing; it is more extensive than the databases mentioned above, but with a worse quality of data [2].

CIS is the most popular, extended in terms of coverage and shared international databases of journals in which articles about statistics and probability appear; one of the major limitations of this database is the update times. For some magazines the last 4/5 years are still missing. The ISI database, regarded by many as representative of the entire research output - with an error margin inferior to 5 percent [1] - also used as a reference in the SIS Commission for the reform of the recruitments mechanisms of Teaching (http://sis-statistica.it/), does not include major Italian statistical journals, like Metron. Sankhya, not Italian, is not included as well its editors were also Mahalanobis and Rao. Scopus follows less restrictive technical criteria for the inclusion of journals and it includes a larger number of them; Metron and Sankhya are present

[2] Now it is available a new Scholar h-index calculator (https://addons.mozilla.org/en-US/firefox/addon/scholar-h-index-calculator/) that improves the data quality.

in this case. The decision to include Google Scholar, despite a data quality that is worse than in other databases, is due to the fact that research products of different types (articles, working papers, reports, books, theses), part of our scientific history, are catalogued in it. There are a lot of famous Italian statisticians who have published articles of high scientific value in collections like Vita e Pensiero, Quaderni di Statistica, Statistica, and so on. The different structure of the 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 [3] for each author are: the number of publications, the corresponding time period and, where available, the total number of citations and the value of h-index (Hirsch Index, [22]). So, the database created and used for the analysis 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 notice that in this exercise authors belong to the same field and they have different roles (unfortunately the seniority is not known, this can be a normalization factor). The main limitations of the bibliometric indicators have to be highlighted: they are measures of impact and not of quality ([3]; [2]); a conscious and responsible use is then recommended. Descriptive variables such as title, university, faculty and so on are also available. For 29 authors it is not possible to obtain the corresponding record (due to POP), while for 13 subjects a value of 0 for all the variables considered in the databases is obtained. The data matrix created is by author and not by product. In this way the information available are aggregated. 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 than in the previous case; this is the topic of another our current project. With the availability of the product database it will be possible to make more advanced analysis. For example, network analysis of the authors or groups (departments, faculties, universities) [24] or magazines ([10], [11]); analysis of benchmarking between researchers or research groups based on the classification ratings (AA, A, B, C); comparison of the median/mean individual Impact Factor (IF)[4] versus the median/mean IF of the corresponding area. Again, it is important to advise that the assessments, whether shared or not, can only be done from a clean and complete database of individual research products and not from the aggregate measures for each author.

3 Bibliometric Data Analysis

A first study on the above mentioned data was presented in [15], along with various analysis: a synthesis of the situation about Italian Statistician publications; the detection of clusters that highlight different research profiles; the identification, using data reduction techniques, of the latent variables that give reason for the detected clus-

[3] The data collection period is from January 2010 until April 2010.

[4] The IF is a measure of the frequency with which the "average article" in a journal has been cited in a given period of time and it is from Journal Citation Report (JCR), a product of Thomson ISI (Institute for Scientific Information), that provides quantitative tools for evaluating journals.

ter: productivity, multi-disciplinarity and author impact. In particular, as the output of the bibliometric database produces data that are not clean, even if some cleaning filters are applied. In [15] a multivariate outliers detection was applied in order to detect anomalies and discrepancies between the databases. The identification of univariate outliers shows scholars who are particularly 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 4 databases. One is either a great scholar or a data need some serious check. It is important to notice the difference between univariate and multivariate outliers, because a subject can be an outlier for one variable, but not for many variables taken together. We plotted the classical Mahalanobis distance of the data against the robust Mahalanobis distance based on the *mcd* estimator [20] and we applied the PCOut algorithm [19], a fast algorithm for identifying multivariate outliers in high-dimensional and/or large datasets. The algorithm applied has identified 23 multivariate outliers; for later analysis 14 of them were assessed as incorrect records, while 9 were due instead to the particular type of data analyzed, because they are subjects with particularly high values on the variables obtained according to some sources over others. These subjects were identified as the best; therefore, with a view to ranking, will be at the top of the list. However, the previous work presents a weakness that lies in the type of data. The matrix that contains a set of metrics from a variety of databases for each author is a sparse matrix (there are many zeros). Furthermore, the variables are leptokurtic and characterized by positive asymmetry. In order to apply the classical techniques of multivariate analysis, the data must be transformed; in this paper more attention will be devoted to data cleaning and initial transformation.

3.1 Data cleaning and transformation

As previously introduced, 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 conjecture the assumption of normality. See in Figure 1 the scatterplot matrix for the ten variables in original scale, that is the matrix of scatterplots for all pairs of variables. The data do not seem to have the elliptical contours which would be expected from the pairwise bivariate normal distributions and it is evident that there are many outliers.

So it is necessary to identify a suitable transformation of the data, except the logarithmic one [16]. It is not easy to immediately identify the optimal trasformation for the data; a useful tool proposed for this purpose is the forward search procedure [5]. In particular we use the MATLAB toolbox *FSDA*[5]. This technique orders the observations from those most in agreement with a specified generalized linear 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

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

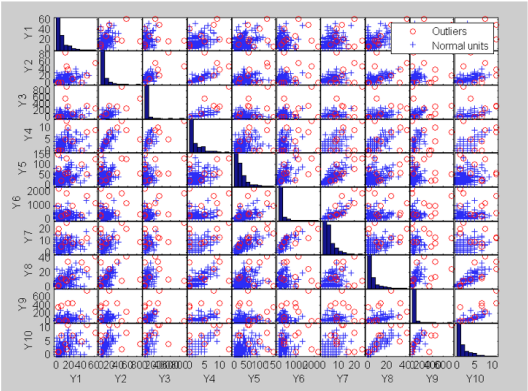

Fig. 1 Scatterplot matrix for the ten variables in original scale

diagnostic quantities during the forward search clearly show the effect of individual observations on residuals and test statistics. This is a strength of the method. As in the Box-Cox transformation (1), [13], zeros in the data are usually not allowed, it is necessary to implement (2) proposed by [26]. 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.

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

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

The use of this transformation through the forward search ([7], chapter 4) represents an innovative methodological contribution. The proposed transformation improves the closeness of the data to the normal distribution. It may however be that other transformations would give even better results. In order to test whether it is so, we embed the various transformations in the single parametric family; the aim is to obtain the best value for parameter λ, with respect to each variable considered. The first step is to apply a forward search through the variables previously transformed $(Y1=log(Y+1))$, 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. So the chosen values for the 10 elements of $\hat{\lambda}$ are:

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

In the second step we repeat the analysis using the transformation proposed from step 1; the forward plot of the maximum likelihood estimates of λ is in Figure 2.

How well defined these estimates of λ are can be determined from plots of the profile loglikelihood (Figure 3). In Figure 4 it is evident the stability of the forward

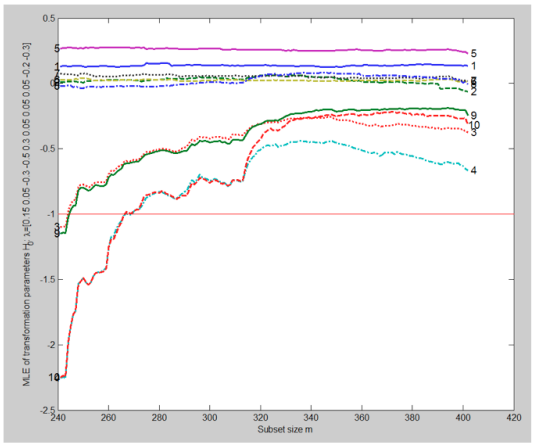

Fig. 2 Forward plot of the ten elements of the maximum likelihood estimates $\hat{\lambda}$ (Analysis of transformations $\hat{\lambda} = (0.15; 0.05; -0.3; -0.5; 0.3; 0.05; 0.05; 0.05; -0.2; -0.3)$).

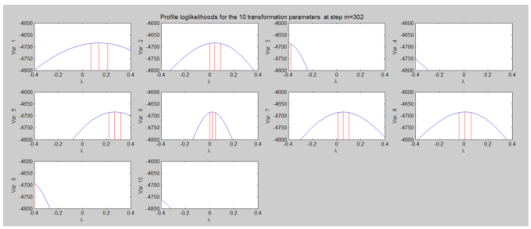

Fig. 3 Analysis of profile loglikelihood

plot from a certain step of the procedure, to confirm 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 loglikelihoods are roughly parabolic close the λ values proposed for each variable; the pairs of lines give asymptotic 95% confidence intervals for each element of λ, based on asymptotic χ^2_1 distribution of twice the loglikelihood ratio. All panels show a sharp definition of the estimates. The plot of the likelihood ratio in Figure 4 shows support for the transformation proposed.

Finally, Figure 5 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.

3.2 Clusters and profiles

After the transformation of the data, it is interesting to see if groups of individuals with similar profiles exist. We may verify the existence of clusters applying the method of Calinsky [14]. Starting from a classical hierarchical cluster analysis we assign the initial centers and then we apply the k-means algorithms for the solution

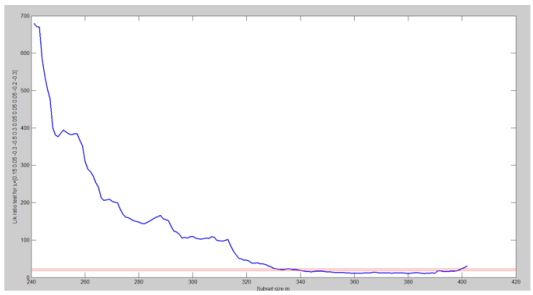

Fig. 4 Forward plot of the likelihood ratio test for the hypothesis of the transformation $\hat{\lambda} = (0.15; 0.05; -0.3; -0.5; 0.3; 0.05; 0.05; 0.05; -0.2; -0.3)$). The horizontal lines are the 95% and 99% points of χ^2_{10}: this transformation is supported

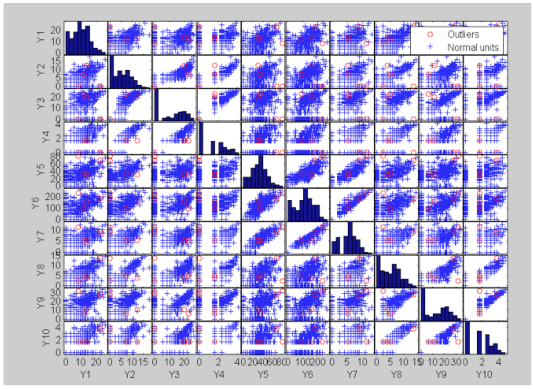

Fig. 5 Scatterplot matrix for the ten variables transformed $Y1 = log(Y + 1)$ with $\hat{\lambda} = (0.15; 0.05; -0.3; -0.5; 0.3; 0.05; 0.05; 0.05; -0.2; -0.3)$).

with 2,3,4, ...15 clusters. This method shows that an optimal number of clusters is 3 or 4, as shown in Figure 6 in which the Calinsky index is maximum. This analysis is done using R library *cclust*.[6]

As known, the *K-mean* algorithm is affected by a lot of problems of non robustness, with respect to initial center and to the sorting of the unit in the data set. In order to obtain a more robust result we apply Model-based Methods of Classification [17]. This approach considers the problem of determining the structure of clustered data, without prior knowledge of the number of clusters or any other information about their composition. 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 parameterizations and cross-cluster constraints. Partitions are determined by the EM (expectation-maximization) algorithm for maximum likelihood, with initial values from agglomerative hierarchical clustering. Models are compared using an approximation to the Bayes factor based

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

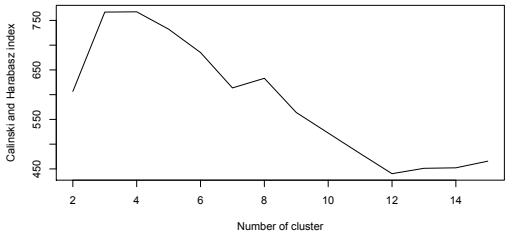

Fig. 6 Calinsky and Harabasz Index for different number of clusters

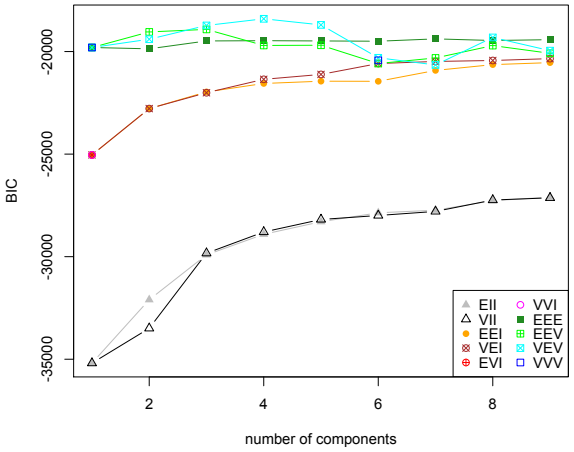

Fig. 7 BIC from *mclust* for the 10 available model parameterizations and up to 9 clusters.

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 compared 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*.[7]

Figure 7 shows BIC from 10 different parameterizations of the covariance matrix in the Gaussian model and up to 9 clusters. Different symbols and line types encode different model parameterizations. The *best* model is the one with the highest BIC among the fitted models. In this case the best model is *VEV* with 4 clusters that correspond to ellipsoidal distributions with variable (V) volume, equal (E) shape and variable (V) orientation. For a description of the parameterizations of the covariance matrix in the Gaussian model and their geometric interpretation see [12].

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

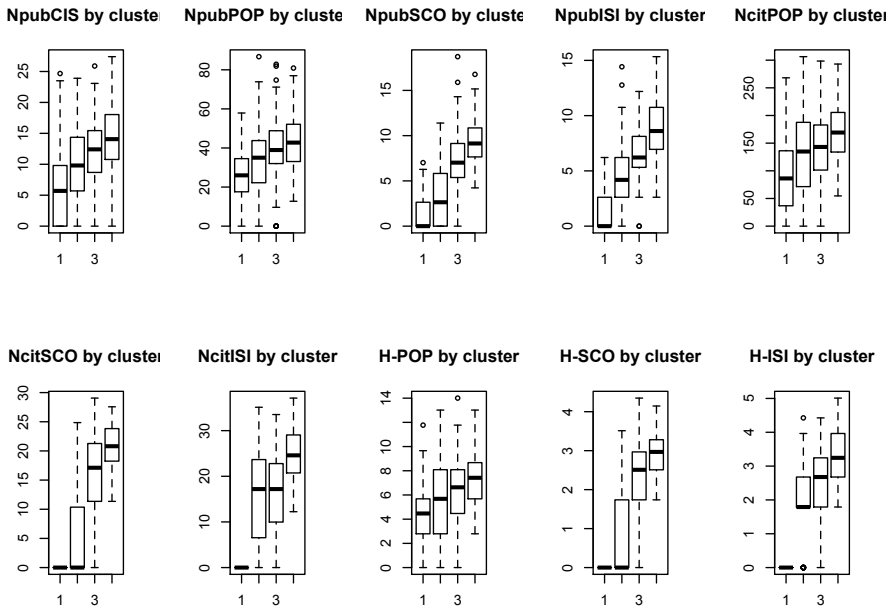

Fig. 8 Boxplot of the variables by groups

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

Table 1 Mean, Standard Deviation and Median of the ten variables by groups

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

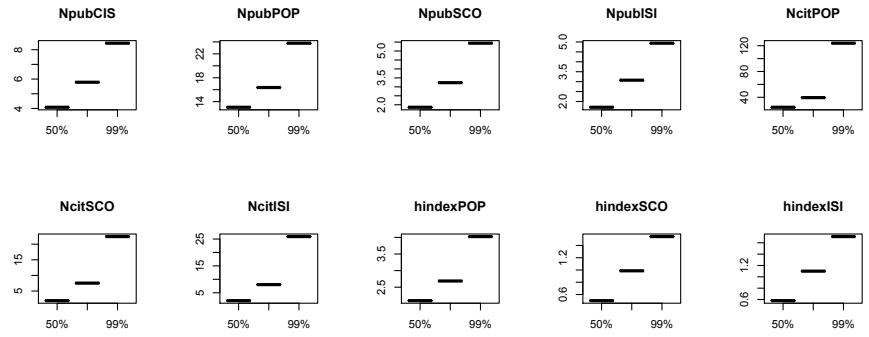

Fig. 9 Means of the variables for three steps of the forward search: 50%, 75% , 99%

3.3 Ranking

A goal to acheve when you do bibliometric analysis is to make a ranking of the institutions or the individuals. As just mentioned in the sub section 3.2 the forward search in exploring multivariate data orders the observations from those closest to farthest from the initial subset, the bulk of data ([7], chapter 7). Proceeding with the analysis and applying the forward search on the non transformed data, it is clear in Figure 9 that the averages of the variables increase, with increasing steps of the procedure (50%, 75%, 99%).

So we propose to interpret the inclusion order of the units by forward search as a generalized ranking, where similar profiles (units entering in close steps) can be identified. In this case, the bulk of the data, as shown in Figure 1, is represented by unproductive and unpopular units, with most indexes at or near zero. Looking at the averages for the various steps of the forward of the ten variables, we can consider that if the order of inclusion increases then the level of productivity / diffusion of the individuals increases too. The last units to enter are outliers in the sense that they are individuals who have higher production and popularity than the others.

Another hypothesis we have been checking is that, by monitoring units in transformed data, the inclusion order interpretation depends on the selection of the initial subset [6]. We apply the forward search using the 81 units that belong to cluster 4 as initial subset. We expect that it is confirmed the presence of groups, we also expect that the order of inclusion is consistent with the clusters identified above. Using the units of the cluster 4 as initial subset, i.e. the most productive / popular, 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, first of those belonging to clusters 2 and 1. Figure 10 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.

A clear change in the Mahalanobis distances indicates that a unit belonging to a new group enters in the search [8]. The plot in Figure 9 shows the presence of 3 groups over the initial subset, according with the Calinsky monitoring on Figure 6

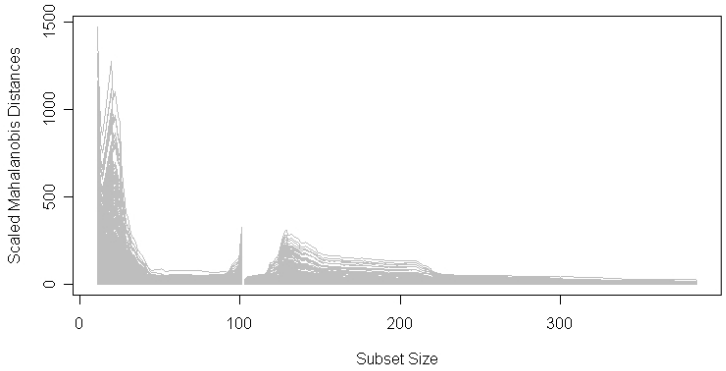

Fig. 10 Mahalanobis distance for each step of the search

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

		Quantiles				
Groups	*N*	*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

and Model-Based cluster on Figure 7. It is important to note that although the groups are identified they are quite dispersed. Most of the variables have high standard deviations (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 the cluster 3 comes before step 36, while 50% of the units belonging to cluster 2 comes before step 60 and before step 225 the 50% of units of cluster 1. Looking to quartiles, 25% of the units of the cluster 1, the less productive/popular scholars, comes after step 243, while the first quartile for the cluster 3 is 7, i.e. 25% of the units of the cluster 3 comes together before step 7, i.e. they are essentially very close to the initial subset. Even in this case, the order of inclusion produces a ranking but from the most productive / popular to the least.

4 Conclusion

In this work four international databases have been analysed, in order to know the scientific output of all Italian Statisticians: CIS, ISI, SCO, POP. The data are considered aggregated by author. The only problem is to assess productivity and impact (through the citations and the h-index). We propose the use of multiple sources to reduce and

correct the errors and we apply robust methods of analysis of bibliometric data. After suitable transformations of data, on one hand we identify clusters/profiles of scholars with similar characteristics, then, secondly, we establish a generalized ranking using the forward search.

What we want to face in the future is to study the italian statistician productivity distribution law [23] and to realize a simulation study in order to find an empirical evidence that supports the use of forward search to produce a generalized ranking. Moreover, a new database related to the single product of research is under construction. The idea is to organize the data in a suitable way in order to reach new goals using new models. Possible dimensions involved in the analysis will be: the topics extracted from abstracts and keywords using topic models [25], the time (year of publication), the co-autorship, the affiliation and countries of authors, the journal bibliometric indexes (i.e Impact Factor, Scimago Journal Rank, ect.). We also wish to identify a measure of multidisciplinarity for each author and study its expected effect on the impact of the research.

References

1. Abramo G. (2009), Ci vuole metodo per valutare la ricerca, www.lavoce.info.
2. Arnold D.N. (2009). Integrity under attack: The state of scholarly publishing. SIAM News 42-10.
3. Adler R., Ewing J., Taylor P. (2009). Citation Statistics with discussion. Statistical Science, 24:1-28.
4. Asserson A., Jeffery K., and Lopatenko A. (2002). CERIF: Past, present and future: An overview. In Proceedings of the 6th International Conference on Current Research Information Systems. University of Kassel, 33-40.
5. Atkinson A.C., Riani M. (2000). Robust Diagnostic Regression Analysis. Springer, New York.
6. Atkinson A.C. and Riani M. (2007). Exploratory tools for clustering multivariate data, Computational Statistics and Data Analysis. 52, 272-285.
7. Atkinson A.C., Riani M. and Cerioli A. (2004). Exploring Multivariate Data with the Forward Search, Springer, New York.
8. Atkinson A.C., Riani M. and Cerioli A. (2006). Random start forward searches with envelopes for detecting clusters in multivariate data, in: S. Zani, A. Cerioli, M. Riani and M. Vichi (Eds.), Data Analysis, Classification and the Forward Search, Springer-Verlag, Berlin.
9. Baccini A. (2010). Valutare la ricerca scientifica. Uso e Abuso degli indicatori bibliometrici, Il Mulino, Bologna.
10. Baccini A., Barabesi L., Marcheselli M. (2009). How are Statistical Journal Linked? A Network Analysis. Chance.
11. Baccini A., Barabesi L. (2011). Seats at the table: the network of editorial boards in information and library sciences, Journal of Infometrics.
12. Banfield J. D., Raftery A. E. (1993). Model-based Gaussian and non-Gaussian clustering. Biometrics, 49:803-821.
13. Box G. E. P. and Cox D. R. (1964). An analysis of transformations. Journal of the Royal Statistical Society, Series B 26 (2): 211-252.
14. Calinski R.B., Harabasz J. (1974). A dendrite method for cluster analysis. In Communication in statistics, 3, 1-27.
15. De Battisti F., Salini S. (2010). Bibliometric indicators for statisticians: critical assessment in the Italian context. Joint meeting GfKl-CLADAG Firenze. http://air.unimi.it/bitstream/2434/152106/2/deba_sal_cladag2010_u.pdf
16. Emerson J.D. (1991). Introduction to Transformation. In Hoaglin D.C., Mosteller F., Tukey J.W., Fundamentals of Exploratory Analysis of Variance, Wiley & Sons, New York.
17. Fraley C., Raftery A.E. (2002). Model-based Clustering, Discriminant Analysis and Density Estimation. Journal of the American Statistical Association, 97, 611631.
18. Franceschet M. (2010). Istruzioni per l'uso della bibliometria, www.lavoce.info.

19. Filzmoser P., Garrett R.G., Reimann C. (2005). Multivariate outlier detection in exploration geochemistry. Computers & Geosciences, 31,579-587.
20. Filzmoser P., Maronna R., Werner M. (2008). Outlier identification in high dimensions, Computational Statistics and Data Analysis, 52, 1694-1711.
21. Geraci M., Degli Espositi M. (2011). Where do Italian universities stand? An in-depth statistical analysis of national and international rankings. Scientometrics, 87-3, 667-681.
22. 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, November 15, a. 102, n. 46.
23. Lotka A. J. (1926). The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences, 16 (12): 317-324.
24. 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.
25. Steyvers M., Griffiths T. (2007). Probabilistic topic models. Handbook of latent semantic analysis, 427(7):424-440.
26. Yeo I. K., Johnson R.A. (2000). A new family of power transformations to improve normality or symmetry. Biometrika, 87, 4, 954-959.