Bibliometric indicators for statisticians: critical assessment in the italian context

Francesca De Battisti, Silvia Salini

1 Introduction

The evaluation of the university and scientific research has become increasingly important in recent years. In particular, there is a growing interest in the evaluation of scientific publications and related bibliometric indicators (Marchant, 2009). The new criteria acquired in the university context, setting up the funding on the basis of assessments of the scientific productivity of universities and departments, as well as regulating the career advancement of individuals assessing their research products, require careful examination of databases available in different fields and kinds of information obtained from their query. It is important to notice that bibliometric indicators can not be self-sufficient instruments of assessment, but they must be integrated into more complex system of assessment; their oversimplified use, oriented to reduce the complexity of the evaluation, would have a severely negative impact on the resulting decision-making process. Despite that, the output of the databases is the image that the international reviewers (of journals, research projects, visiting demands and partnerships) have about the Italian statistics researchers and scientific community. Knowing of operational limitations about use, coverage and updating of databases (Falagas et al, 2008), the aim of this research is to gain awareness and knowledge of the image, true or false, obtained by them: the study analyses the scientific production of all italian statistics academic scholars (SECS/S01).

2 Main results

The databases that will be considered are:
1. Current Index to Statistics (CIS), created by the American Statistical Association and the Institute of Mathematical Statistics (http://www.statindex.org/).

Francesca De Battisti,
University of Milan e-mail: francesca.debattisti@unimi.it

Silvia Salini,
University of Milan e-mail: silvia.salini@unimi.it

2. Web of Science (WoS), edited by the Institute for Scientific Information and distributed by Thomson Reuters (http://isiwebofknowledge.com/).
3. Scopus, sponsored by Elsevier (www.info.scopus.com).
4. Google Scholar, with recommended interface Publish or Perish, developed by Anne-Wil Harzing (http://www.harzing.com/pop.htm).

By the database query, made in the period from February to April 2010, a dataset was built, in which there are the variables: number of publications for each database, corresponding time period and, excluding CIS, number of citations and h-index (Marchant, 2009). There are also descriptive variables such as title and affiliation, obtained by MIUR. Table 1 shows the joint distribution of the number of publications of italian researchers according to the CIS and WoS databases.

Table 1 Number of publications on CIS vs Number of publications WoS

		WoS						Total
		<= 5	6 - 10	11 - 15	16 - 20	21 - 25	26+	
CIS	<= 5	203	21	2	0	0	0	226
	6 - 10	71	23	5	1	1	0	101
	11 - 15	24	18	10	1	0	0	53
	16 - 20	2	8	5	5	2	1	23
	21 - 25	5	7	1	1	1	1	16
	26+	6	1	6	4	4	4	25
Total		311	78	29	12	8	6	444

First of all, the SECS/S01 scholars will be classified on the basis of 10 quantitative variables obtained from the databases, adding an additional dichotomous variable for each person that points out whether or not the subject has published on the top five journals resulting from the SIS Survey[1]. A preliminary classification shows that there is a group of ”better” researchers, that have high values on all variables, a group of scholars who publish much but have less citations, others have a lot of papers in other fields than statistics, etc. As a second step, using data reduction techniques, latent variables that give reason for the detected clusters, are identified: productivity, multi-disciplinarity and author impact. As final step, the possibility to build a composite index, based on all dimensinos and all databases, will be critically evaluated.

References

Falagas M.E., Pitsouni E. I., Malietzis G. A. and Pappas G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strenghts and weaknesses. *The FASEB Journal*, 22, 338-342.

Marchant T. (2009). An axiomatic characterization of the ranking based on the h-index and some other bibliometric rankings of authors *Scientometrics*, Vol. 80, No. 2 (2009) 327344

[1] http://www.stat.unibo.it/ScienzeStatistiche/Ricerca/Progetti+e+attivita/Materiali_Giornata_di_Studio_-_La_valutazione_della_ricerca_nelle_scienze_statistiche.htm

Joint Meeting

GfKl - CLADAG 2010

Florence, 8 - 10 September 2010

BIBLIOMETRIC INDICATORS FOR STATISTICIANS: CRITICAL ASSESSMENT IN THE ITALIAN CONTEXT

Francesca De Battisti and Silvia Salini

Department of Economics Business and Statistics

University of Milan

OUTLINE

INTRODUCTION

- Evaluation and bibliometric indicators: a very topical theme
- What happens to the statistics? Which databases and sources are used in the field?
- There are several sources with different characteristics. Are the information obtained from various sources consistent? Are the indicators obtained related to each other?
- Is it possible to synthesize information from different sources?

BIBLIOMETRIC DATABASES

1. **Current Index to Statistics**, created by the American Statistical Association and the Institute of Mathematical Statistics (http://www.statindex.org/) (**CIS**).
2. **Web of Science**, edited by the **Institute for Scientific Information** and distributed by Thomson Reuters (http://isiwebofknowledge.com/) (**ISI**).
3. **Scopus**, the mayor competitor of Web of Science, sponsored by Elsevier (www.info.scopus.com) (**SCO**).
4. **Google Scholar**, scientific research version of the famous search engine on the web; recommended interface for querying, which allows proper data cleaning, is **Publish or Perish**, developed by Anne-Wil Harzing (http://www.harzing.com/pop.htm) (**POP**).

BIBLIOMETRIC DATABASES: CIS

PLUS

- Only publications in statistics, probability and related topics
- Easy query
- Coverage time range: since 1974 and before

MINUS

- Not free
- Updating
- Inclusion criteria: all journals in which reported statistical papers are
- Operations: query only by surname
- Problems:

  - homonymy

  - some input errors in the database

BIBLIOMETRIC DATABASES: ISI

PLUS

- Selective coverage of most relevant journals (and other literature sources)
- Update
- Inclusion criteria: journals that meet particular technical criteria
- Operations: in the query it is possible to include only the surname and the initial, or to filter by category of work or affiliation. ISI also offers the possibility, by clicking on individual works, to identify sets of work automatically created by database; but not always

MINUS

- Not free
- Not easy query
- Coverage time range: University of Milan license since 1990
- With regard of affiliation, several problems arise:

  1 also the affiliations of the coauthors are reported

  2 it may be missing, in which case the paper is not detected

  3 it may have been some mobility, so you can lose all previous works

  4 it can be written in many different ways
- Problem: homonymy

GfKl - CLADAG 2010 DE BATTISTI, SALINI

BIBLIOMETRIC DATABASES: SCOPUS

PLUS

- More extensive than ISI initiative
- Easy query
- Coverage time range: papers since 1970
- Update
- Inclusion criteria: only journals cited monitored by Science Direct (Elsevier)
- Operations: query by surname and firts name, without affiliation. Then the database produces affiliation history of the author, matching name and history

MINUS

- Not free, but it is possible a free partial query
- Coverage time range: citations since 1996
- Operation problems:

  - homonymy

  - some errors in the mathing between author and affiliation

BIBLIOMETRIC DATABASES: POP

PLUS

- Free
- Inclusion criteria: anything on the web
- Coverage time range: unlimited
- It is more extended than the databases mentioned above

MINUS

- Not easy query
- It is not a database
- Coverage time range: unlimited
- Worse data quality

DATA SET: THE CASE STUDY

- Miur: SECS/S-01 (February 2010)
- 444 records
- Field:
  - affiliation (campus, faculty, department, title)
  - Npub (CIS, ISI, SCO, POP)
  - Ncit (ISI, SCO, POP)
  - H-index[1] (ISI, SCO, POP)
  - TOP5 Journals[2] (JASA, JRSSb, Annals, Biometrika, Biometrics)

[1]A scholar obtains a value h if he has h papers with at least h citations each and the remaining (N-h) papers have no more than h citations each.
[2] SIS Survey presented in Bologna on March 2010

DATA PREPARATION

- 444 total
- Missing values:
  - 29 are not applicable (NA)
  - 13 have 0 occurrences for each database (3 associate professors, 9 researchers)
- Outliers
  - no point in trying univariate outliers, scholars may simply be particularly productive or unproductive than other
  - a multivariate outlier, which is based on all available output, is represented by an unusual combination of the outputs of the 4 databases. It could be a great scholar or a data that needs a check.

DATA PREPARATION

- Outliers
  - Multivariate outliers detection is a way to detect anomalies and discrepancies between the databases.
- R Package 'mvoutlier'
  - Function **dd.plot**
    - Plots the classical Mahalanobis distance of the data against the robust Mahalanobis distance based on the mcd estimator.
    - P. Filzmoser, R.G. Garrett, C. Reimann. *Multivariate outlier detection in exploration geochemistry*. Computers & Geosciences, 31:579-587, 2005.
  - Function **p.cout**
    - Fast algorithm for identifying multivariate outliers in high-dimensional and/or large datasets.
    - P. Filzmoser, R. Maronna, M. Werner. *Outlier identification in high dimensions*, Computational Statistics and Data Analysis, 52, 1694-1711, 2008.

DATA PREPARATION

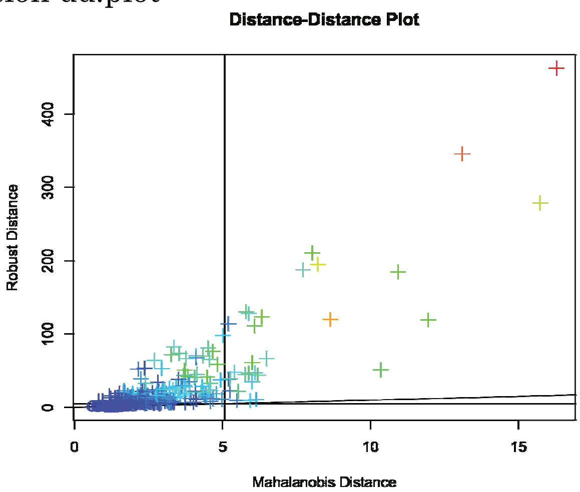
- Function dd.plot

23 outliers identified

```
$outliers
  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE
 [17] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
 [33] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
 [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [65] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [81] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[113] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[129] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[145] FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[161] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[177] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[193] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[209] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
[225] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
[241] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[257] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[273] FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
[289] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[305] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
[321] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
[337] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
[353] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[369] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
[385]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[401] FALSE FALSE
```

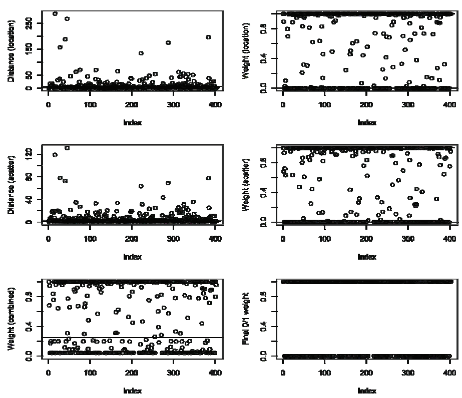
DATA PREPARATION

- Function dd.plot



DATA PREPARATION

- Function p.cout



•More than 23, for the presence of a lot of zero and for the skewness

•The 23 units identified before are the ones with the highest value of distance from the scatter

DATA PREPARATION

- 23 outliers
- detailed inspection of the individual records, using, if needed, also the curriculum
- 9 correct records: the unusual combination of the outputs is due to particular publication patterns [books (POP+) , National Statistical Journals (CIS+), disciplines with high impact (ISI+, SCOPUS+)]
- Errors:

5 POP
3 SCOPUS
2 ISI
1 CIS
1 POP & SCOPUS
1 POP & ISI
1 SCOPUS & POP & ISI

- special character in name
- homonymy
- change of affiliation
- wrong record in the database

DATA UNDERSTANDING

Mean (SD)

	Full Prof	Associate Prof	Assistant Prof
NpubCIS	15,37 (11,6)	6,91 (5,7)	2,88 (3,3)
NpubPOP	32,15 (29,1)	20,67 (17,7)	19,74 (17,5)
hindexPOP	5,31 (3,9)	3,72 (3,24)	3,18 (3,2)
NpubISI	7,03 (8,1)	4,61 (5,6)	3,22 (3,4)
hindexISI	2,47 (2,5)	1,54 (1,5)	1,12 (1,9)
NpubSCO	8,48 (11,5)	4,92 (6,0)	3,71 (5,3)
hindexSCO	2,16 (2,5)	1,39 (1,6)	1,17 (1,7)
N (MIUR)	143	111	148

DATA UNDERSTANDING

Median

	Full Prof	Associate Prof	Assistant Prof
NpubCIS	13	7	3
NpubPOP	24	18	15
hindexPOP	5	3	2
NpubISI	4	3	2
hindexISI	2	1	1
NpubSCO	5	3	2
hindexSCO	1	1	1
N (MIUR)	143	111	148

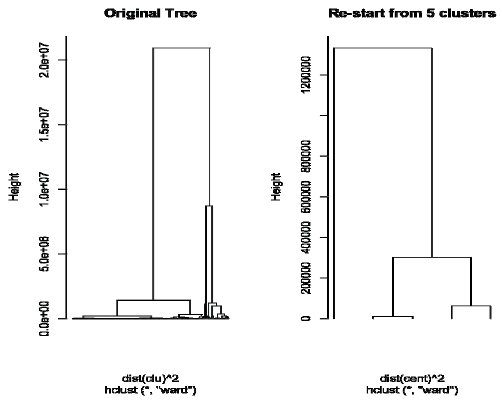
DATA UNDERSTANDING: TOP 5

Top 5	Authors	Mean of papers
Journal of the American Statistical Association	29	1.52 (0.8)
Journal of the Royal Statistical Society Series B	22	1.23 (0.4)
Biometrika	35	1.69 (1)
Annals of Statistics	19	1.53 (1)
Biometrics	16	1.19 (0.4)

	MIUR	TOP5 Author	Mean of papers (TOP5)
Full Prof	150	40 (26,6%)	2,78 (2)
Associate Prof	123	19 (16,4%)	2,21 (2,6)
Assistant Prof	171	20 (11,7%)	1,25 (0,5)

MODELLING: THE CLUSTERS

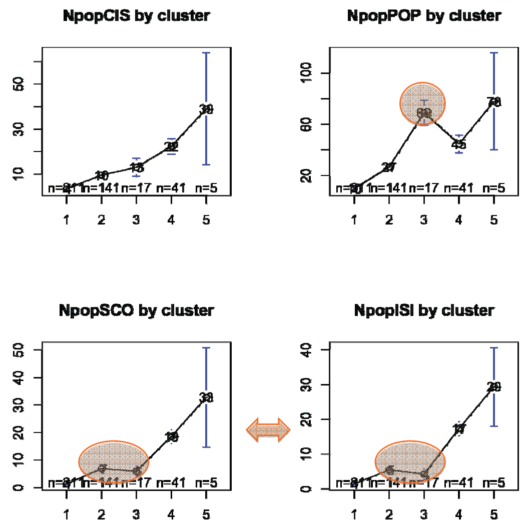
- Hierarchical Algorithm
  - Ward's method
  - Square Euclidean Distance



MODELLING: THE CLUSTER PROFILES

		Cluster					
		1	2	3	4	5	Total
N		211	141	17	41	5	415
NpubCIS	Mean	3,51	9,60	13,06	22,34	39,00	8,26
	Median	2,00	8,00	12,00	20,00	48,00	5,00
NpubSCO	Mean	1,28	6,87	5,94	18,54	32,80	5,46
	Median	,00	7,00	5,00	18,00	29,00	3,00
hindexSCO	Mean	,30	2,12	1,88	5,17	8,60	1,57
	Median	,00	2,00	2,00	5,00	9,00	1,00
NpubPOP	Mean	9,93	26,78	69,00	44,68	77,80	22,33
	Median	9,00	25,00	61,00	41,00	61,00	18,00
hindexPOP	Mean	1,62	4,72	10,18	7,88	14,60	3,80
	Median	2,00	5,00	10,00	8,00	13,00	3,00
NpubISI	Mean	1,45	5,46	4,24	17,34	29,40	4,83
	Median	1,00	6,00	3,00	17,00	26,00	3,00
hindexISI	Mean	,51	2,06	1,71	5,41	8,80	1,67
	Median	,00	2,00	2,00	5,00	9,00	1,00

MODELLING: THE CLUSTERS



MODELLING: THE CLUSTER PROFILES

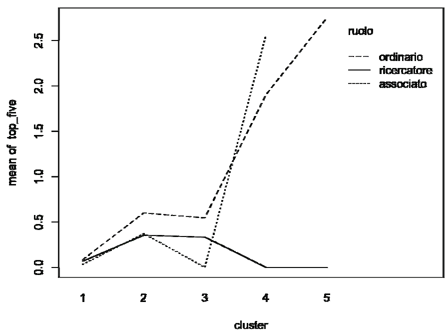
1) A very big group of scholars who have low values for all indices, half of them have at most one paper on ISI, but they have more than 2 statistical papers (CIS), they attend conferences and produce working papers (POP).
2) A big group of scholars that have good value for each indexes, half of them have more than 6 paper on ISI and SCOPUS and more than 8 statistical papers. Moreover they produce working paper and they attend conferences. Values for the dissemination are not very high.
3) A little group of scholars whose key feature is to have very high values for productivity and dissemination for POP. By analyzing in detail, they are people who have written important books, they often participate in conferences and events as organizers, they are editors of special issues and so on. The number of papers on ISI and Scopus is lower than in Cluster 2.
4) A group of scholars who have very high values for both production and dissemination on all databases, even if the amounts of POP are lower than in Cluster 3. Probably they invest more in the journals than in the other research activities.
5) Scholars with exceptional values on all databases for both productivity and dissemination.

MODELLING: THE CLUSTER PROFILES

Cluster	Full	Associate	Assistant
1	48	60	103
2	50	43	48
3	11	3	3
4	30	9	2
5	4	0	1

	Top Five										
	0	1	2	3	4	5	6	7	9	10	11
1	199	11	1	0	0	0	0	0	0	0	0
2	107	18	8	6	1	0	0	1	0	0	0
3	13	2	1	1	0	0	0	0	0	0	0
4	17	9	4	4	0	3	1	0	1	1	1
5	2	1	0	0	0	2	0	0	0	0	0

Interaction plot: Top Five Journals by Cluster and Title



MODELLING: DATA REDUCTION

- Item-item correlation matrix

Inter-Item Correlation Matrix

	NpubCIS	NpubSCO	NcitSCO	hindexSCO	NpubPOP	NcitPOP	hindexPOP	NpubISI	NcitISI	hindexISI
NpubCIS	1,000	,660	,506	,582	,574	,517	,573	,683	,452	,615
NpubSCO	,660	1,000	,715	,882	,622	,504	,648	,872	,508	,761
NcitSCO	,506	,715	1,000	,766	,431	,582	,515	,681	,779	,705
hindexSCO	,582	,882	,766	1,000	,582	,523	,643	,858	,613	,855
NpubPOP	,574	,622	,431	,582	1,000	,665	,837	,591	,326	,512
NcitPOP	,517	,504	,582	,523	,665	1,000	,799	,503	,598	,519
hindexPOP	,573	,648	,515	,643	,837	,799	1,000	,625	,430	,620
NpubISI	,683	,872	,681	,858	,591	,503	,625	1,000	,646	,888
NcitISI	,452	,508	,779	,613	,326	,598	,430	,646	1,000	,744
hindexISI	,615	,761	,705	,855	,512	,519	,620	,888	,744	1,000

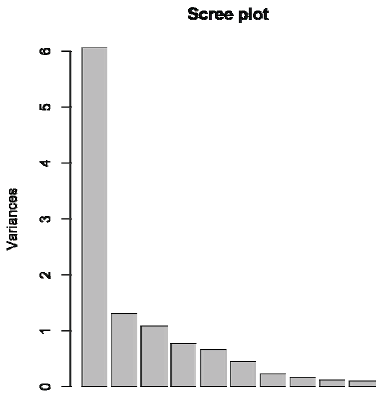
- Cronbach's alpha

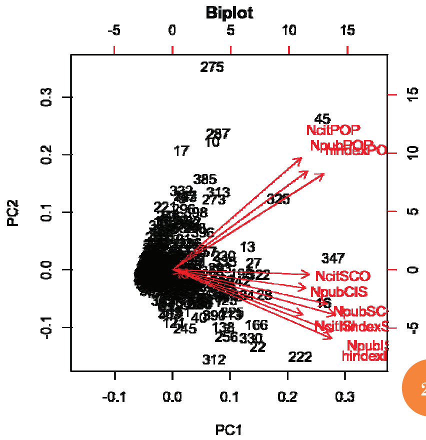
Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,594	,945	10

MODELLING: DATA REDUCTION

- Synthetic index?





CONCLUSION

- Using a single source is not recommended
- The combined use of multiple sources helps to control the results
- There is no single profile of a "good researcher"
- It is difficult to compare because everyone makes different research choices
- POP seems to measure a different dimension
- SCOPUS and ISI are very similar for statisticians
- CIS does not use selective criteria for inclusion
- Everyone should check his record and notify to the manager of the database what must be corrected, every database has a link / path to report errors

FUTURE TASKS

- Comparison between the journal coverage
- More information on researchers, links with outputs, co-authors
- MathSciNet instead of CIS
- Opportunity to use data from CINECA
- New scientific fields, comparisons

REFERENCES

- Abramo G. (2009), *Ci vuole metodo per valutare la ricerca*, www.lavoce.info.
- Bakkalbasi N., Bauer K., Glover J. And Wang L (2006), *Three options for citation tracking: Google Scholar, Scopus and Web of Science*, Biomedical Digital Libraries, 2006, 3:7.
- Bergstrom C.T., West J.D. and Wiseman M.A. (2008), The Eigenfactor metrics, *Journal of Neuroscience* **28** (45), pp. 11433–11434.
- Biolcati-Rinaldi F. (2010), *Quali indicatori bibliometrici per le scienze sociali?,* Working Paper 2, Dipartimento di Studi Sociali e Politici, UNIMI.
- Checchi, D. e Jappelli, T. (2008), *Ricerca per indice* h, www.lavoce.info.
- Falagas M.E., Pitsouni E. I., Malietzis G. A. and Pappas G. (2008), Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strenghts and weaknesses, *The FASEB Journal*, 22, 338-342.
- Franceschet M. (2010a), *Istruzioni per l'uso della bibliometrica*, www.lavoce.info.
- Franceschet M. (2010b), A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar, *Scientometrics,* 83(1), 243-258.
- Franceschet M. (2010c), The difference between popularity and prestige in the sciences and in the social sciences: a bibliometric analysis, *Journal of Informetrics,* 4(1), 55-63.
- Franceschet M (2009), A cluster analysis of scholar and journal bibliometric indicators, *Journal of the American Society for Information Science and Technology,*60(10), 1950-1964.
- Marchant T. (2009), An axiomatic characterization of the ranking based on the h-index and some other bibliometric rankings of authors, *Scientometrics*, Vol. 80, No. 2 (2009) 327344.
- Norris M. and Oppenheim C. (2007), Comparing alternatives to the Web of Science for coverage of the social sciences' literature, *Journal of Infometrics*, 1 (2007), 161-169.