Available online at www.sciencedirect.com

ScienceDirect

Procedia Economics and Finance 17 (2014) 47 – 54

Procedia Economics and Finance

www.elsevier.com/locate/procedia

ELSEVIER

Innovation and Society 2013 Conference, IES 2013

A support for classifying scientific papers in a University Department

Daniela Cocchi*, Giuseppe Cavaliere, Marzia Freo, Simone Giannerini, Mario Mazzocchi, Carlo Trivisano, Cinzia Viroli [†]

Department of Statistical Sciences- University of Bologna, Via delle Belle Arti 41, Bologna 40126, Italy

Abstract

Measuring the productivity of the research community is a challenging and relevant issue at the national and institutional levels. To this aim several lists which classify scientific journals have been provided by both public and private companies accordin g to specific motivations.

The existence of a multiplicity of lists from one hand, the always increasing number of journals and the variegate publishing strategies of the single researchers from the other one, pose the problem of the definition and assessment of the set of scientific journals that maximally cover the potential heterogeneous research domains.

This work proposes a procedure for merging the classifications provided by competing lists of journal from different institutions, solving indeterminacies and missing attributions.

Selection and peer-review under responsibility of the Organizing Committee of IES 2013.

Keywords: journal classification, research evaluation, statistics research.

* Corresponding author. Tel.: +39-051-2098234; fax: +39-051-232153.
E-mail address: daniela.cocchi@unibo.it

[†] The Authors are the members of the Research Committee of the Department of Statistical Sciences, University of Bologna.

Selection and peer-review under responsibility of the Organizing Committee of IES 2013.

doi:10.1016/S2212-5671(14)00877-6

1. Introduction

In the recent years, increased emphasis has been placed on the evaluation of scientific research of academic institutions and its members, on the basis of their publications in peer reviewed journals and volumes. As a result of the increasing interest on publishing in refereed journals, we also witness a rapid increment of the number of the scientific journals available for the scientific communities. To give an example, the number of ISI journals in the subject category "statistics & probability" was 69 in 2000, 81 in 2005 and 117 in 2012. As a consequence, the identification of the set of scientific journals of potential interest and their assessment are crucial issues for the research community.

Different lists which classify scientific journals are available for helping research institutions to address and orientate their research lines. Mostly, such lists are provided by both public and private companies according to specific motivations. At the opposite side, the members of university departments or research institutions follow their own publishing strategies; after publication, individual scientific performances are assessed, while departments are evaluated as a whole. The existence of a multiplicity of lists could be seen an odd complication, on the contrary it is a resource for both individual and institution choices.

In this work, we develop a proposal for merging alternative classifications that can support the publication policies of research institutions, taking a Department of Statistics as a working example. It is well known that each department has its own history; in particular, Statistics Departments may be very heterogeneous, since their discipline of interest spreads over a multiplicity of research domains. This poses the problem of the definition of the lists of scientific journals that maximally cover the potential heterogeneous research domains.

The proposal we develop takes into account official classifications of journals and offers a strategy for solving:
a) the classification differences and indeterminacies among lists
b) the cases of indeterminacy due to the non-inclusion of relevant journals in official lists.
The procedure works as follows:

1) we define the lists of scientific journals which are relevant for the department/institution involved;
2) we classify such journals according to available rankings exogenously defined;
3) we develop a tool to solve indeterminacies which occur when a) the joint use of the selected lists is not conclusive or b) the researchers published in journals that do not belong to any of the selected lists.

2. The definition of a list of journals at the local level

It is very common that the members of a research institution (denoted as department from now on) publish in a set of journals that usually do not fit in just one of the available lists. Hence, the department needs to resort to multiple lists as to maximize their coverage, although a number of journals might still be excluded from the resulting set; also, this induces discrepancies that have to be solved so that a suitable classification is needed.

2.1. The example of a Department of Statistics

The example of an Italian department of Statistics is typical (Table 1), since statistical disciplines expand in several cultural domains and some of these are not included in the Italian "CUN Area 13: Economics and Statistics" (which roughly covers Economics, Statistics, Mathematics for Economics and Management) to which academic researchers in Statistics have been associated[1]. In any case, Area 13 is important since, besides Economics, it includes the typical statistical sectors going from SECS-S01 to SECS-S05 (Statistics, Statistics for experimental and technological research, Economic Statistics, Demography, Social Statistics), together with SECS-S06 (Mathematical

[1] In the Italian University System, the disciplines taught in higher education are classified into 14 CUN (Consiglio Universitario Nazionale, Italian National University Council) main fields (CUN, 2013).

methods of economy, finance and actuarial sciences), SECS-P02 (Economic policy) and SECS-P05 (Econometrics)[2].

Moreover, research assessment and resource allocation depend on the lists that are officially released within such area. These lists are denoted as GEV13 (GEV, 2012) and ANVUR13 (ANVUR, 2013a); GEV13 has been used as a benchmark in the recent national VQR exercise (ANVUR, 2013b), while ANVUR13 has been the support for the recent ASN (Abilitazione Scientifica Nazionale, i.e. National Scientific Qualification; ASN, 2011). Besides these, the lists proposed by scientific societies are a further tool for the construction of a classification system. Here, we utilize the list of the Italian Statistical Society SIS (sectors SECS-S01 - SECS-S05) (SIS, 2012) and the AMASES (Associazione per la Matematica Applicata alle Scienze Economiche e Sociali; AMASES, 2009) list for SECS-S06.

The union of the 4 lists consists of 2786 titles and covers 79% of the publications of the members of the Department of Statistical Sciences, University of Bologna (Department from now on). In the time span considered (January 2004-December 2012), 64 journals do not belong to any of the 4 lists because

a) the journal does not belong to the union of the 4 lists considered or

b) the journal has not yet been included in an official list.

In Table 1 we report a brief description of the 4 lists considered together with i) the number of departmental journals that do not fit into the lists; ii) the coverage of the lists.

Note that a union of lists does not directly correspond to a database. When a list of scientific journals is prepared, each record is generally identified by a ISSN code. When an author submits a paper to a journal for publication, she/he is attracted mainly by the “title”, irrespective of possible changes in the name, in the ISSN code and of the fact that the journal may have a paper and an electronic version. Record linkage is an issue in this context (Christen, 2012), basically since the same title can be differently recorded in each separate list, due to the separation of words with spaces or commas, to tiny typos, to the consideration of cancellation of the determinative article in the precise writing. Linkage has to be exact and not probabilistic; a lot of editing has been performed, essentially in order to avoid duplication and performing all matching.

Table 1. The four official lists considered by the Department of Statistical Sciences, University of Bologna. The last 2 columns refer to the (January 2004-December 2012) Department publications in 312 journals.

Source	Classification	N. of journals	Excluded dept. journ.	Coverage
GEV 13 (MIUR)	4 classes (A, B, C, D)	1902	154	51%
ANVUR A	A$^{(*)}$:(D1–D4)	686	230	26%
SIS	4 classes (A, B, C, D)	1265	85	73%
AMASES	None	182	287	8%

(*) Only class A has been considered by ANVUR. D1-D4 is another official way to denote the higher education topics relevant to statistics

3. A unique classification method from a multiplicity of lists

We propose a hierarchical scheme for combining the classification of the lists as follows: first, we merge the classifications of the lists proposed by governmental institutions (List 1, i.e. GEV13 and ANVUR/A). Then, we combine the lists of scientific societies (List 2, i.e. SIS and AMASES). Third, we perform a final combination of the previous two separate aggregations.

A peculiar feature of our proposal is that the rankings produced by external institutions are maintained to the maximum possible extent. In particular, we consider the classification by GEV13 as a minimum threshold that cannot be lowered further. Moreover, a journal classified in D by GEV13 or in D by SIS but eventually included as

[2] The official denominations of the disciplines can be retrieved in CUN (2011).

relevant journal in the list of scientific journals of ANVUR or AMASES is upgraded to the class C of the aggregated classification.

Any remaining indeterminacy is solved by means of appropriate indices.

List 1 is constructed from combining the classifications of GEV13 and ANVUR/A according to the scheme of Table 2; the last 3 columns contain the frequency distribution of the journals.

Table 2. Cross classification according to GEV13 and ANVUR/A (left panel: classes; right panel: number of cases)

	ANVUR/A	Not ANVUR/A	ANVUR/A	Not ANVUR/A	TOTAL
GEV A	A	A	430	19	449
GEV B	B	B	18	333	351
GEV C	C	C	3	175	178
GEV D	**C**	D	0	924	924
Not GEV	**Indeterminacy**	nc	235	649	884
TOTAL			686	2100	2786

List 2 is constructed starting from the classifications of SIS and AMASES, according to the scheme of Table 3. Since SIS has proposed a different classification for each sector SECS-S01/SECS-S05, we have chosen the most favorable class over the sectors.

Table 3. Cross classification according to scientific societies (left panel: classes; right panel: number of cases)

	AMASES	Not AMASES	AMASES	Not AMASES	TOTAL
SIS A	A	A	17	76	93
SIS B	B	B	46	198	244
SIS C	C	C	11	188	199
SIS D	**C**	D	7	722	729
Not SIS	**Indeterminacy**	nc	101	1420	1521
TOTAL			182	2604	2786

In Table 4 we summarize the classification scheme that integrates the two lists and Table 5 contains the frequency distribution of the 2786 journals according to the indications of Table 4, which illustrates our decision rule for integrating external classifications and cases that deserve further investigation.

Table 4. A decision rule for integrating external classifications and an internal index

	List 2: Scientific Societies					
List 1: Ministry	**A**	**B**	**C**	**D**	**Indeterminacy**	**Not classified**
A	A	A	A	A	A	A
B	INDEX	B	B	B	B	B
C	INDEX	INDEX	C	C	C	C
D	INDEX	INDEX	INDEX	D	D	D
Indeterminacy	INDEX	INDEX	INDEX	INDEX	INDEX	INDEX
Not classified	INDEX	INDEX	INDEX	INDEX	INDEX	DEPT

Table 4 reports the classification scheme that has been adopted: for a subset of cells, the classification of List 1 prevails, while for another subset, the index developed in the Department will be computed. The last cell (last row and last column, here denoted as DEPT) counts the journals where the members of the Department have published but that do not belong to the lists.

Three different principles are synthetized in Table 4: I) the adoption of an exogenous classification, II) the proposal of using an internally developed tool for solving uncertainties and indeterminacies and III) the identification of the publication peculiarities of the Department.

Table **5.** A decision rule for integrating external classifications and an internal index: a case study

	List 2: Scientific Societies						
List 1: Ministry	**A**	**B**	**C**	**D**	**Indeterminacy**	**Not classified**	**TOT**
A	62	87	12	16	32	240	**449**
B	14	36	18	18	17	248	**351**
C	5	18	14	14	3	125	**179**
D	5	22	58	115	14	711	**925**
Indeterminacy	4	36	19	75	5	96	**235**
Not classified	3	45	85	484	30	DEPT	**647**
TOT	**93**	**244**	**206**	**722**	**101**	**1420**	**2786**

The integration scheme proposed in Table 4 and performed in Table 5 is the following:

a) Attribution which accepts an exogenous classification

- the journals of List 1 which are not contained in List 2 maintain the exogenous classification (A, B, C, D) (first 4 values of the "Not classified" column, i.e. 1324 titles).
- the journals of List 1 which are undetermined in List 2 maintain the exogenous classification (A, B, C, D) (first 4 values of the "Indeterminacy" column, i.e. 66 titles).
- the journals that in List 2 are classified in a class lower or equal to the class of List 1 (upper triangle of the sub-table that classifies according to A, B, C, D in both lists) receive the class of List 1, i.e. 165 titles.

b) Attribution via an index developed in the department

- for journals that in List 2 are classified in a higher class than in List 1: they cannot be assigned to a class lower than the GEV class. This occurs for the cells that are in the lower triangle of the sub-table that classifies according to A, B, C, D in both lists (122 cases)
- for the cases of indeterminacy of List 1, i.e. the total of the line Indeterminacy (235 cases)
- for the journals not contained in List 1 but contained in List 2 (total of nc line: 647 cases)

The 64 journals that do not belong to the union of Lists 1 and 2, but where the members of the Department published, are also assigned to a class by means of the method proposed by the Department.

3.1. Dealing with indeterminacies and exclusion from lists

In order to deal with the indeterminacies highlighted under case b) above, we have built a normalized index in the interval (0, 1] associated to each journal. The index can be computed if the journal belongs either to the ISI or to the Scopus databases.

The main features of the index are: 1) it allows to compare journals across different scientific areas; 2) it does not depend on the choice of the scientific categories chosen by a department. First, we derive the ISI index for the whole database (10,743 journals in the 2011 version of Journal of Citation Reports). If the information for this index is unavailable, we compute a Scopus index instead, based on the 19,124 journals listed in Scopus for 2011. Second, we normalize the index within each subject category as to take values in (0, 1]. If a journal belongs to more than one category, the median of the normalized indices for each category is assumed. Third, we derive a classification in 4 categories by choosing the following thresholds that maximize the agreement with the GEV list:

if $P > .80$ → A
if $0.30 < P <= 0.80$ → B
if $P <= 0.30$ → C
no index → D.

3.2. The normalized index proposed by the Department

Several bibliometric indices are available in the ISI-Thomson Journal Citation Reports, one of the two main commercial bibliometric data bases. Among these, the Impact Factor (*IF*), the 5-year Impact Factor (*IF5*), the

Article Influence Score (*AIS*), and the Eigenfactor Score (*ES*) explain various elements of the multidimensional citation outcome. We consider, as a starting point, the bibliometric indices of all 10,743 journals available from the 2011 ISI Journal Citation Reports. These are classified in 176 categories under the Science Edition section, and 56 under the Social Edition section, for a total of 232 categories.

The procedure for computing the synthetic index, which follows a framework designed by a Committee nominated by the Head of the Department in 2011, is the following:
for each category *s* (*s=1,...,* 232) to which the journal belongs, the two indices *AIS* and *ES* are projected into a 0-1 interval (hereinafter indices *AIS** and *ES**) by means of the following function:

$$\begin{cases} I_s^* = 1 & \text{if } I_s > I_{s,perc} \\ I_s^* = \dfrac{I_s - I_{s,\min}}{I_{s,perc} - I_{s,\min}} & \text{if } I_s \le I_{s,perc} \end{cases}$$

where

I_s is the bibliometric index (*AIS* or *ES*) of the journal as computed from the Journal Citation Reports data base
$I_{s,perc}$ is the index (*AIS* or *ES*) which corresponds to a pre-defined percentile for the *s*-th subject category
$I_{s,min}$ is the index (*AIS* or *ES*) which corresponds to the minimum value for the *s*-th subject category
I_s^* is the (*AIS* or *ES*) index normalized within the 0-1 interval for the *s*-th subject category.

For journals which belong to more than one subject category, the median value of the index *AIS** and *ES** was computed.
The final score for each journal (P_{ISI}) was computed as follows:

$$\begin{cases} P_{ISI} = 1 & \text{if } AIS^* = 1 \\ P_{ISI} = \left(AIS^* + ES^*\right) & \text{if } AIS^* < 1 \end{cases}$$

The *ES* indicator refers to the number of citations for the articles of a given journal over the last 5 years found in journals which also belong to the JCR, with a weighting factor which assigns a weight on citing journals which is proportional to their own citation levels. The *ES* excludes self-citations. The *AIS* index can be interpreted as a measure of the average influence of articles published on a given journal which belongs to the JCR over the 5 years which follow their publications. It is computed by dividing the *ES* of the journal by the number of articles published by the journal itself, and is normalised in a way that an *AIS* above (below) 1 indicates that the articles in that journal have an impact on the scientific community above (below) the average. The *AIS* is the indicator which is most similar to the 5-year Impact Factor, but it overcomes its limitations. By construction, the *AIS* does not count self-citations.

Scopus, another major commercial bibliometric database, provides two indicators (*SJR* and *SNIP*) and has an almost double coverage relative to the JCR (19,124 journals). For this reason, Scopus was chosen as the official data-base for the computation of author-level bibliometric indices by ANVUR. The SJR indicator for a journal is obtained as the average number of citations received in a given year for all the articles published by the same journal over the previous three years. The *SNIP* index normalizes the average number of citations relative to the “potential” citations for the reference discipline for each journal, so that it represents a comparable indicator across disciplines.

As a starting point, we considered the bibliometric indices of all 19,124 journals available in the 2011 Scopus data base, subdivided in 306 categories. The procedure to compute the derived index is as follows:
For each category *s* (*s=1,...,*306) to which a journal belongs, the *SJR* index is projected into the 0-1 interval based on the following function:

$$\begin{cases} I_s^* = 1 & \text{if } I_s > I_{s,perc} \\ I_s^* = \dfrac{I_s - I_{s,\min}}{I_{s,perc} - I_{s,\min}} & \text{if } I_s \leq I_{s,perc} \end{cases}$$

where:

I_s is the index (SJR) of the journal obtained from the Scopus data base
$I_{s,perc}$ is the index (SJR) which corresponds to a pre-defined percentile for the s-th subject category
$I_{s,min}$ is the index (SJR) which corresponds to the minimum value for the s-th subject category
I_s^* is the (SJR) index normalized within the 0-1 interval for the s-th subject category

For journals which belong to more than one subject category, the median value of the SJR^* indices is taken, which provides the final P_{SCOPUS} score.

As a general conclusion, a journal for which ambiguity has been detected due to different classifications performed by different institutions, may remain in the class attributed as worst, but, on the contrary, can be upgraded. A class can be attributed also to journals that do not fall in any of the lists considered, provided that they enter in ISI or Scopus classifications. With our proposal we however do not claim any alternative local classification for the journals that can receive a class from exogenous work.

The computer codes for assigning a value of the index to any journal are available on request.

3.3. A solution for journals not appearing in the selected lists

The members of any scientific organization may publish in journals that do not belong to the official lists that can be selected (64 out of 312 in the Department of Statistics of the University of Bologna between 2004 and 2012). One part of these publications is classified by means of the normalized indices illustrated before. The remaining ones may still be associated to a class by exploiting the relationship between the ISI/SCOPUS index above and the Google Scholar H-Index (25 out of 64 journals). The procedure is the following:

1) the Scholar H-Index (last 3 years) is computed for each non classified-journal;
2) for each class, the journal with ISI/SCOPUS index nearest to the lowest class threshold is identified and its Google Scholar H-Index (last 3 years) is obtained;
3) each nc-journal is assigned to the best class (among A, B, C) for which:

nc-journal Scholar H-index > threshold-journal Scholar H-index.

3.4. The final classification

Under the proposed procedure, a class from A to D is attributed to the complete set of 2786+ 64 = 2850 journals as reported in Tab. 6. In this way, an exogenous classifying label is attributed to much more than the half of the titles belonging in the four identified lists: in these cases, the contrast with decisions taken by influential external institutions are annulled. In addition, all titles where the department members have published in the time interval considered are associated to a class.

Table 6. The final classification of the titles of interest in our Department

	From lists	**Attributed**	**Dept. specific**	TOTAL
A	449	248	6	703
B	337	294	11	642
C	156	208	24	388
D	840	254	23	1117
TOTAL	1782	1004	64	2850

4. Conclusions

In this work we have illustrated a way for classifying the scientific publications of any institution, suggesting to identify the most suitable set of lists prepared by official institutions and to decide ex ante the rules according to which the exogenous classification can be used. The adoption of a multiplicity of lists carries contradictions and ambiguities. For such cases and for the journals that do not belong to any list, we suggest to employ internal classifications tools. With our criterion, a new bibliometric index, proposed when official lists of journals were not yet available, is suggested to this aim. We arrive at classifying all the scientific publications of a specific department, but the method we propose can be easily transferred to any other situation.

The work performed for constructing the indices is a complete exercise of standardisation. The coverage of the proposed index is very extended, since all ISI or Scopus categories are considered for normalizing the results. Any scientific department could compute the index for the set of journals of interest.

Acknowledgements

We thank the Heads of the Department of Statistical Sciences of the University of Bologna Ignazio Drudi and Rosella Rettaroli for promoting the work that suggest this paper. We specially thank ARIC (Area della RIerca: Research and Knowledge Transfer Division), University of Bologna for providing the data base on the research outputs of the Department of Statistical Sciences in a computationally tractable form. We thank SIS (the Italian Statistical Society) for giving us the opportunity to illustrate our proposal.

References

AMASES (Associazione per la Matematica Applicata alle Scienze Economiche e Sociali), 2009. http://www.amases.it/Documenti/PDF/Lista%20riviste_29Apr09.pdf

ANVUR, 2013a. ANVUR13/CLASSEA http://www.anvur.org/attachments/article/254/area13_classe_a.pdf

ANVUR, 2013b. http://www.anvur.org/index.php?option=com_content&view=article&id=28&Itemid=119&lang=it

ASN (Abilitazione Scientifica Nazionale), 2011. http://abilitazione.miur.it/public/index.php?lang=eng

Christen P., (2012). Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, Springer.

CUN (Consiglio Universitario Nazionale), 2013 http://www.cun.it/about-us.aspx

CUN (Consiglio Universitario Nazionale), 2011 http://www.cun.it/media/116411/settori_scientifico_disciplinari_english.pdf

GEV, 2012. http://anvur-miur.cineca.it/?q=it/content/lista-riviste-gev-13

SIS (Società Italiana di Statistica), 2012. http://old.sis-statistica.org//files/pdf/2012/documento_presentazione_liste_riviste_sis_02042012_con_integrazione_20062012.pdf