Assessing the reliability and validity of Google Scholar indicators. The case of social sciences in Italy

Ferruccio Biolcati-Rinaldi, Francesco Molteni and Silvia Salini [1]

Abstract

Google Scholar is an appealing data source for the measurement of scientific production in the social sciences and humanities (SSH) fields. Its appeal derives from its extensive coverage of the literature. This contrasts with issues of data quality, which are still quite controversial. This chapter aims to evaluate the reliability and validity of bibliometric indicators taken from Scholar as well as their coverage of the scientific production in the social sciences. The analysis will be based on a comparison of Scholar with other bibliometric data sources (Web of Science and Scopus) plus an institutional dataset. The reliability of Scholar indicators will be investigated through correlational analysis, while their validity will be assessed using different external criteria (the results of national evaluation procedures based on a peer review approach). The analysis will be developed for the population of Italian university professors in a subset of SSH: political philosophy, history, political science and sociology. The final discussion of the results will take into account the various purposes that bibliometric exercises try to achieve.

[1] Ferruccio Biolcati Rinaldi
Francesco Molteni
Dipartimento di Scienze sociali e politiche
Università Statale di Milano
Via Conservatorio, 7
20122 Milano
e-mail: ferruccio.biolcati@unimi.it
e-mail: francesco.molteni@unimi.it

Silvia Salini
Dipartimento di Economia, management e metodi quantitativi
Università Statale di Milano
Via Celoria 2
20133 Milano
e-mail: silvia.salini@unimi.it

1. Introduction

One of the main debates in bibliometrics concerns its applicability to the fields of social sciences and humanities (SSH). The main problem stems from the structure of the scientific literature that characterises SSH and that reflects its peculiarities (Nederhof 2006). If we can assume that articles published in international journals are substantially comprehensive of the natural and life sciences literature, for SSH it is necessary to consider other communication channels of scientific literature (Hicks 2004).

Social sciences literature is not limited to what is published in international journals (Moed 2005, 42 and 122–131; Hicks 2004, 475–477), and it should be supplemented with the literature published through other scientific channels. One of these channels is certainly represented by books, which not only constitute a scientific research outcome more used in SSH than in other fields, but which also have more impact (in terms of citations) than articles. In addition, the relatively small values of the correlation coefficients between the citations from books and from articles support the hypothesis that these are somewhat two distinct forms of literature, although interdependent and with some overlaps (Hicks 2004, 481–482; Clemens et al. 1995).

Another important communication channel for SSH is represented by national journals. This is an obvious consequence of the fact that the subjects of interest are often nationally (if not locally) framed as well as the fact that the stakeholders (politics, media, public opinion) with whom researchers relate are national and local. While international journals are often included in Web of Science (WoS) or Scopus, national journals seldom are. It follows that national literature is not adequately represented in commercial bibliographies, with the exception of the US and UK. However, it should be considered that things are rapidly changing. The internationalisation of social sciences is increasing due to powerful forces such as globalisation processes, digital innovation, European funding mechanisms that require the establishment of international partnerships, and national assessment exercises that increasingly emphasise international publications (Hicks 2004, 484–489).

In recent years, both Web of Science and Scopus have opened themselves to these other channels of disseminating scientific literature, although in different ways. Nonetheless, the highest expectations for the resolution of these problems are placed by many researchers on Google Scholar. In particular, it is assumed that Scholar is a more extensive bibliography, able to cover most of the literature of SSH. It is also assumed that the data quality is still good enough to ensure the validity and reliability of bibliometric indicators derived from this source. This chapter aims to verify this hypothesis in the case of social sciences in Italy, combining the methodological practices for the assessment of indicators' reliability and validity with the appropriate data processing and data analysis procedures.

The chapter is divided into six sections, including the introduction and conclusions. The following section details the research design, introducing the description of the process of scientific production, the definitions of coverage, the reliability and validity and the field of application of the chapter, i.e., a subset of the social and political sciences in Italy. Then the data sources analysed in the chapter (Google Scholar, Scopus, Web of Science) are discussed in the third section. The fourth section will focus on the definition of the dataset used for the analysis and on the problem of Scholar coverage, while the fifth section presents the results of estimates of reliability and validity of the bibliometric indicators drawn from Scholar.

2. Research design

The aim of this section is to define the main coordinates behind the analysis, namely the various dimensions of the scientific production and the related bibliometric indicators, the desirable properties of such indicators and a description of the field of application of the analysis.

Our working hypothesis relies on three different dimensions of scientific production: output, recognition and relevance (of the kind and venue of the publication).

By *output* we mean the ability of the researcher to transform his or her scientific work into the main research product, that is, publications. Using this kind of definition, we aim to distinguish between the quantitative aspects of research work (the number of publications) and the qualitative ones (their scientific value). These represent different dimensions of scientific production, and the relationship between the two needs to be empirically assessed (Martin and Irvine 1983, 65–66; Research Evaluation and Policy Project 2005, 12). Already in the sixties, Cole and Cole (1967) pointed out this distinction, developing a typology of scientists based on the intersection between quantity and quality of scientific production. Starting from a sample of 120 university professors of physics in the United States, they identified four different typologies; two of them were identified by the convergence of the two dimensions ("the prolific": high quantity and quality; "the silent": low quantity and quality) and the other two by the divergence ("the mass producer": high quantity and low quality; "the perfectionist": low quantity and high quality). Indicators of scientific output are, for example, the number of publications in a given period in the different data sources, the number of books, chapters, articles etc.

With the second dimension, we adopt the interpretation of citations as a measure of scientific *recognition*. Relying on the normative structure of science described by Merton (1973), recognition represents the main incentive and reward mechanism for researchers. The main goal for scientific researchers is indeed to be recognised as competent members of a scientific community and possibly to have a central role in this community. There are several ways to achieve this scientific recognition – eponymy, awards, being part of an editorial committee, panel of referees and so on – but the primary means is surely being quoted in a scientific publication. Citations are also frequently used as impact indicators for scientific publications (Martin and Irvine 1983, 67–72), but here we intend to adopt the perspective of scientific recognition. This approach, being more general, permits consideration of the heterogeneity behind the reasons for citations, whereas the approach relying on impact implicitly presumes a rational approach that considers citations as being used strictly for scientific argumentation[2]. Examples of indicators of scientific recognition are the number of citations per publication and the author's H-index.

The third dimension (*relevance*) relies on the importance given by the scientific community to the various kinds of publications (books, articles, conference proceedings, maps, patents etc.) or to the various venues (journals, publishers, book series etc.). Regarding the latter, the underlying assumption is that as the relevance of the publication venue increases, the selection criteria become more compelling; it follows that the acceptance rate can somehow indicate the quality of publication. At the same time, publication in a relevant venue increases the odds of being read and quoted. With some simplification, we can state that the main objective of this kind of exercise is to provide SSH with an indicator similar to the *Impact Factor*, built upon peer assessment instead of journal citations. The publication of an article in a journal classified as Class A, according to the list published by ANVUR under the National Scientific Habilitation, is an example of an indicator of relevance.

[2] For an overview of the debate, see Moed (2005, 193–208) and the Research Evaluation and Policy Project (2005, 12–14).

For each of these three dimensions, it is possible to identify a series of indicators. These indicators should hold some desirable properties (reliability and validity), which we will try to assess. In addition, we will try to estimate the coverage of the scientific production provided by the different bibliometric datasets.

The analysis of the coverage, reliability and validity of bibliometric indicators will follow a two-level structure: a base level concerning the publications and an aggregate level concerning the authors. The base level about publications will permit an evaluation of the degree of coverage of the sources, using cross-comparison checks. The aggregate level will permit the assessment of the reliability and validity of the indicators.

According to Corbetta (2003, 81)

> Reliability has to do with the 'reproducibility' of the result, and marks the degree to which a given procedure for transforming a concept into a variable produces the same results in tests repeated with the same empirical tools (stability) or equivalent ones (equivalence).

Following this definition, bibliometric indicators, which are basically procedures for the translation of the different dimensions of scientific production into variables, can be considered reliable if they provide the same results with equivalent tools: in this case, with different data sources. The notion of reliability should not be confused with validity, which is defined as follows.

> Validity, on the other hand, refers to the degree to which a given procedure for transforming a concept into a variable actually operationalizes the concept that it is intended to (Corbetta 2003, 81).

Also in this case, there are different procedures with which to evaluate the validity of indicators. Herein we propose a validation criterion that tests "the correspondence between the indicator and an external criterion that, for some reason, is deemed to be correlated with the concept" (Corbetta 2003, 83). This test can be applied to all the dimensions linked to the quality of scientific production (recognition and relevance) using Research Quality Evaluation (*Valutazione della Qualità della Ricerca* - VQR 2004–2010) individual scores or National Scientific Habilitation (*Abilitazione Scientifica Nazionale* - ASN 2012) results as criteria.

The analysis presented here will focus on a selection of social and political sciences falling into the so-called Area 14 of the scientific sectors classification elaborated by MIUR (the Italian Ministry of Education, University and Research) upon the advice of the National University Council (CUN) (labelled *CUN areas* for this reason in the academic language). More specifically, Area 14 includes the following scientific disciplines:

- SPS/01 Political philosophy
- SPS/02 History of political thought
- SPS/03 History of political institutions
- SPS/04 Political science
- SPS/05 American history and institutions
- SPS/06 History of international relations
- SPS/07 General sociology
- SPS/08 Sociology of culture and communication
- SPS/09 Economic sociology and sociology of work and organisations
- SPS/10 Urban and environmental sociology
- SPS/11 Political sociology
- SPS/12 Sociology of law, deviance and social change
- SPS/13 African history and institutions
- SPS/14 Asian history and institutions

In administrative language, the disciplines are labelled *Settore concorsuale*, meaning that the recruitment process ("*concorso*", or competition) takes place within the disciplinary boundaries established in the classification. Each of the SCs may include one or more sub-disciplines, labelled *Settore scientifico disciplinare* (SSD). In the rest of the chapter we will deal with the SC aggregation level.

As will be noticed, Area 14 is a heterogeneous set of disciplines, with some disciplines close to the humanities. The VQR final report states that the share of articles among all publications subject to evaluation in the VQR ranges between 22.9% for SPS/06 History of international relations and 37.6% for SPS/04 Political science. The heterogeneity is even greater if we look at the share of publications in English, which ranges from 5.2% for SPS/03 History of political institutions to 38.6% for SPS/04 Political Science (ANVUR 2013).

3. Google Scholar and other bibliometric data sources

The data source upon which this research focuses is Google Scholar, the application for scientific research that is part of the well-known web search engine (http://scholar.google.com/). Using the *PageRank* algorithm (Franceschet 2011), Google Scholar indexes various kinds of research products (articles, working papers, reports, books, theses) found in the websites of publishers and publishing houses, academic and professional associations, universities, research institutes. For each research product, the algorithm also measures the number of times it is cited in other research products. Scholar is much more extensive than other data sources, but this advantage implies issues of data quality. The data quality significantly increases in the case of personal profiles; in this case, Scholar permits the removal of all those spurious publications or the unification of different records referring to the same publication. According to some authors, the main underlying problem is that "all the information about contributions selection, inclusion criteria, timing of updates and ways of indexing is essentially classified" (Baccini 2010, 75). Scholar can be consulted directly, or some interfaces can be used; among many of these (Scholarometer, Quadsearch, Scirus and others), the best known is Publish or Perish, developed by Anne-Wil Harzing (http://www.harzing.com/pop.htm). This interface permits the generation of many bibliometric indicators and the cleaning of the data obtained from the Scholar dataset. This is an important feature, considering the problems mentioned above.

Besides Scholar, this research focuses on other two data sources. The first one (Scopus) was released in 2004 by Elsevier (www.info.scopus.com) and contains a commercial bibliographical repertoire of abstracts and citations. Compared to Web of Science, Scopus is more extensive and includes more social sciences and humanities journals and national journals. According to 2014 data, Scopus indexes more than 21,000 journals (compared to the 12,000 of WoS) and 50,000 books (similar to WoS). Scopus and Web of Science share similar problems regarding journal admission (see below); despite being more extensive, the coverage problems for various fields in Scopus are lessened but not solved. Starting from Scopus data, some indicators similar to the Impact Factor are produced; examples include the *SCImago journal rank* (SJR) (www. scimagojr.com) and the *Source normalized impact per paper* (SNIP) (www. journalindicators.com).

Web of Science, managed by the Institute for Scientific Information (ISI) and supplied by Thomson Reuters (http://wokinfo.com), is organised into three different sections, the Science Citation Index Expanded, the Social Sciences Citation Index and the Arts & Humanities Citation Index. It aims at a selective coverage of influential journals as well as books and conference proceedings. As already discussed, WoS coverage for the social sciences and humanities is scarce when compared to that for the natural sciences or medicine (Moed 2005, 42). The admission criteria are demanding, especially for traditional national journals (regularity of publication, acceptance of international editorial conventions, number of citations per journal, English language, procedure of selection). The Journal Citation Reports, linked to WoS, develops some indicators of relevance for scientific journals, including the well-known Impact factor (IF).

4. The construction of the dataset and the coverage of Scholar

4.1. The population under investigation

The data analysed here refer to university professors who fall within the so-called Area 14 (Political and Social Sciences) and who were in service on 31 December 2014. By university professors we mean full professors, associate professors and assistant professors (researchers) with open-ended or temporary contracts. The list has been extracted through the website of the Ministry of Education, University and Research (http://cercauniversita.cineca.it/). Some information was added by drawing on the research data of *The Italian university system: a regional analysis*, a project developed for the Regional Observatory of Education and Training Systems of Regione Puglia with the scientific coordination of Daniele Checchi.

Table 4.1 shows the distribution of the main characteristics of the population: gender and age group, academic degree and discipline (*Settore concorsuale*), size of the university and area where the university is located. For many of these variables, the information is available even at a more disaggregated level. For the analyses performed here, the level of aggregation has been chosen in order to ensure a sufficient number of cases for group analysis.

Table 4.1. Frequency distribution of population characteristics (percentages and absolute values)

	%	(N)
Gender		
Male	60.9	(1034)
Female	39.1	(663)
Age group		
1940–1949	13.1	(218)
1950–1959	23.4	(391)
1960–1969	33.7	(562)
1970–1983	29.8	(498)
Academic degree		
Full professor	22.2	(376)
Associate professor	29.9	(507)
Assistant professor (fixed term contract)	38.5	(655)
Assistant professor (temporary contract)	9.4	(159)
Discipline		
14/A1 – Political philosophy	6.2	(105)
14/A2 – Political science	12.9	(218)
14/B1 – History of political thought and institutions	11.1	(189)
14/B2 – History of international relations, non-European history and institutions	8.8	(150)
14/C1 – General sociology	23.7	(401)
14/C2 – Sociology of culture and communication	18.3	(311)
14/C3 – Political sociology, sociology of law, deviance and social change	6.9	(117)
14/D1 – Economic sociology, urban and environmental sociology	12.1	(206)
University size		
Large	55.1	(922)
Medium	25.5	(427)
Small	16.0	(267)
Very small	3.4	(57)
University area		
North-West	23.8	(404)
North-East	19.2	(325)
Centre	24.3	(412)
South	19.7	(335)
Islands	10.4	(177)
Online	2.6	(44)

4.2. The Scholar dataset

The Scholar dataset was developed within the project *EVA – Extraction, validation and analysis of Google Scholar data in non-bibliometric scientific sectors,* developed by Alfio Ferrara, Stefano Montanelli and Stefano Verzillo, and described in a chapter in this volume.

Our work started from a dataset of 17,307 contributions; 1,641 of these cases (about 10%) were dropped due to the lack of information on the publication date or because the publication date proved patently implausible (prior to 1970), given the age structure of the population considered here. In this chapter, analyses are carried out mostly on contributions and not on publications, that is, the same publication recurs as many times as there are co-authors who belong to the target population described above.

Figure 1 describes the distribution of the publication years of the contributions. In interpreting this graph, as for the following graphs related to Scopus and Web of Science, keep in mind that the trend of scientific production, as estimated by the three bibliometric datasets, mixes different phenomena. On the one hand, it reflects the coverage capacity of the scientific production, which grows over time, and, on the other hand, the trend of scientific production. In turn, the latter suffers from two contrasting trends, the increase in individual productivity (perhaps following the introduction of evaluation) and the subsequent phases of expansion and contraction of academic staff (from 2000 to 2008, the number of professors rose from 1287 to 1801 and then subsequently dropped to 1679 in 2015). The difficulty of separating the two phenomena obviously has implications for any longitudinal comparison.

Figure 4.1 shows how the trend of scientific production has increased exponentially since the mid-nineties, so much so that in 2012 it reached a maximum of more than 1400 contributions. In 2013 and 2014 there is a decrease in the number of publications, but again this has nothing to do with the trend of scientific production but with the delay in some publications being indexed in Scholar.

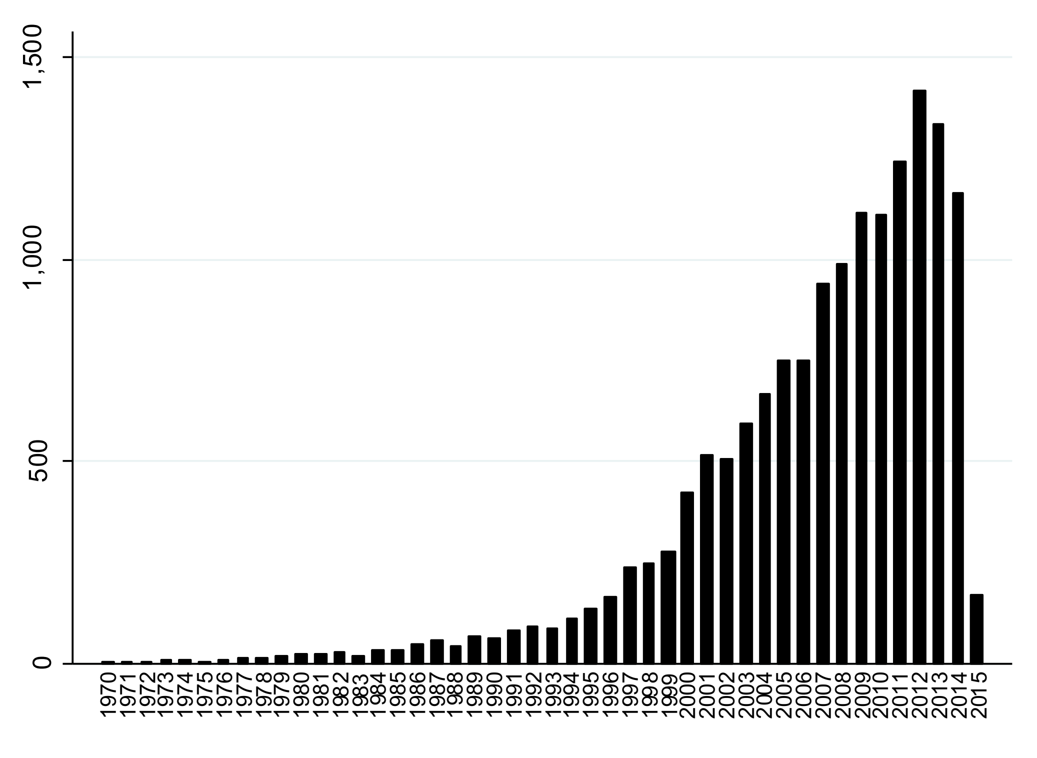

Figure 4.1. Publication year of contributions in Scholar (absolute values)

It is useful to underline how Scholar, compared with Scopus and WoS, is particularly poor in information. For example, neither the type of contribution (monograph, chapter, article etc.) nor the publication language or the name of the journal is immediately available. Information about the name of the journal would allow linking of the articles to the many bibliometric indicators available for journals. In some cases, this information can be retrieved but with some margin of error. In any case, the type of analysis that can be performed on the basis of the Scholar dataset is limited compared to the possibilities offered by the other datasets.

4.3. The Scopus dataset

While with Scholar data were retrieved in an automatic way using an algorithm, for Scopus (and WoS) data were retrieved in a controlled manner by an expert. This choice implies higher costs in identifying the contributions, which would consume much more human time and much less computer time, but this procedure presumably yields more accurate results. It should be emphasised that the difference between automatic and controlled retrieval in terms of the accuracy of the results must be seen in relative terms; even controlled retrieval is prone to errors (double surnames, composite names, abbreviated forms, problems in the identification of the author's membership). These errors derive from a variety of technical barriers of the selection procedure and from the knowledge of a large and complex field which, even if entrusted to an expert, cannot be complete.

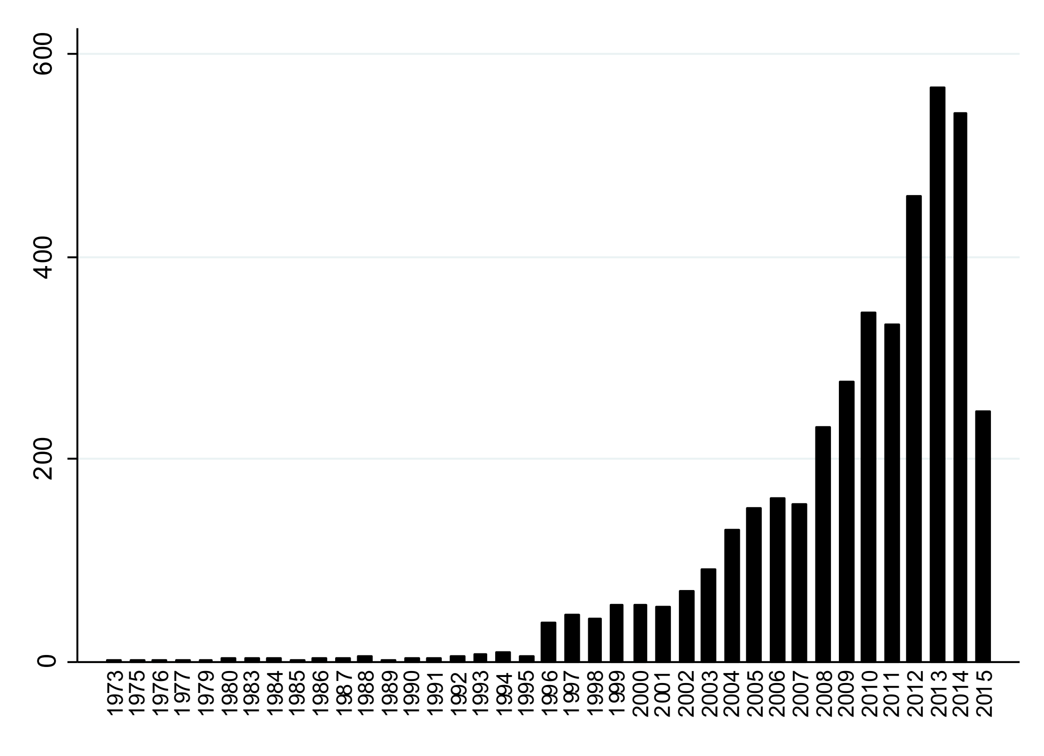

Figure 4.2. Publication year of contributions in Scopus (absolute values)

All publication years available in Scopus were considered here. The data collection was carried out between June and July 2015 and was completed for a subset of professors in November of that year. The procedure resulted in the identification of 4126 contributions. Figure 4.2 shows the trend of publication years in Scopus; this trend is similar to that of Scholar (exponential since the mid-nineties), even if the level is much lower. Here the peak is in 2013 with nearly 600 publications. Again, what is possible to appreciate is more the development of the capacity of the repertoire to include the scientific production of Italian political and social scientists rather than the trend of production.

4.4. The Web of Science (WoS) dataset

Similarly to what was done for Scopus, in WoS the identification of the contributions of Italian scholars was not carried out automatically by an algorithm but in a controlled manner by an expert.

We considered all available years and all bibliographic repertoire indexes of the Web of Science Core Collection: Science Citation Index Expanded (SCI-EXPANDED) from1985 to the present; the Social Sciences Citation Index (SSCI) from 1985 to the present; the Arts & Humanities Citation Index (A&HCI) from 1985 to the present; the Conference Proceedings Citation Index - Science (CPCI-S) from 1990 to the present; the Conference Proceedings Citation Index - Social Science & Humanities (CPCI-SSH) from 1990 to the present; the Emerging Sources Citation Index (ESCI) from 2015 to the present) together with the Book Citation Index (from 2005 to the present). Data collection from the Web of Science Core Collection was carried out in December 2015 and from the Book Citation Index in February 2016.

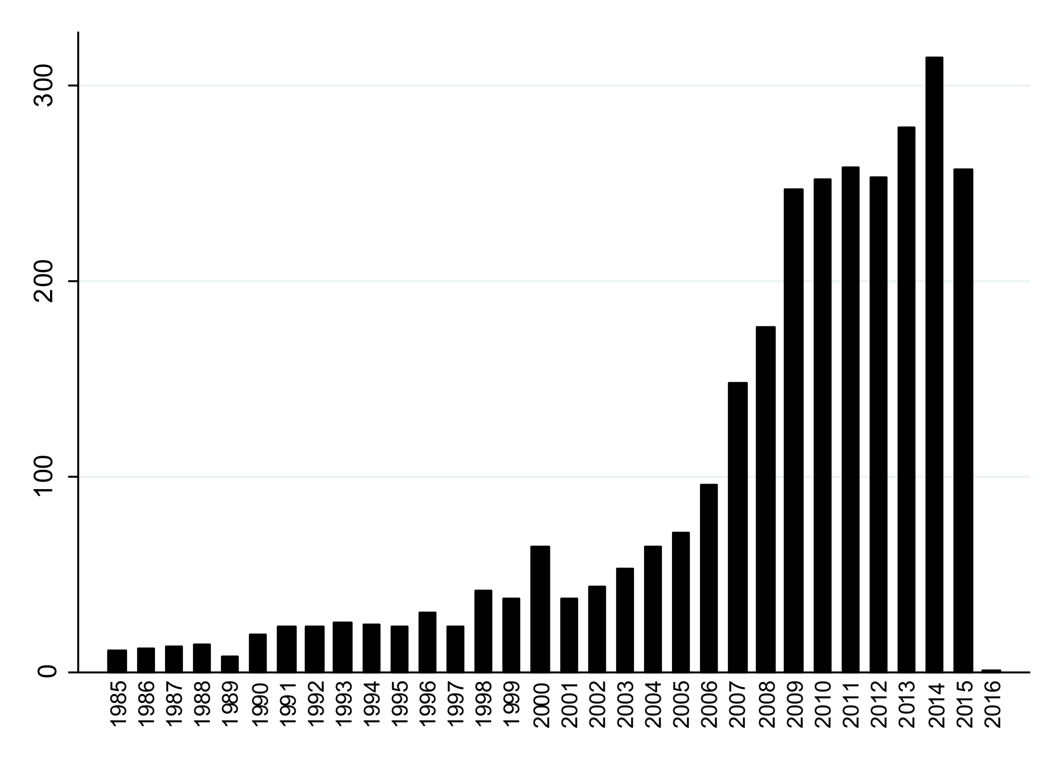

Figure 4.3. Publication year of contributions in WoS (absolute values)

The WoS dataset consists of 2941 contributions. Figure 4.3 shows the publication year trend, which is similar to the one already seen for Scholar and Scopus but with a period of stability between 2009 and 2012. The level here is even lower; the maximum is reached in 2014 with little more than 300 contributions.

4.5. The SUA-RD 2011–2013 dataset and the coverage algorithm

In addition to the citation databases, the SUA-RD dataset was considered. The dataset considers the publication activities of Area 14 inserted within the SUA-RD procedure in the 2011–2013 period, was made available by ANVUR[3]. The SUA-RD data is used to analyse the Scholar coverage.

The four datasets considered – Scholar, Scopus, WoS and SUA-RD 2011–2013 – were collected by different sources. In order to be able to perform a comparison between the publications they included, it was necessary to process some matching algorithms. This need was due to the possibility – not so remote – that the same record was indexed differently in different databases (i.e.

[3] We would like to thank Andrea Bonaccorsi and Marco Malgarini, who allowed us to use SUA-RD 2011–2013 data.

with slight differences in the title or even with different years of reference). The heuristic configuration adopted in the following project was particularly selective, to search with a high degree of reliability, even at the cost of discarding valid correspondences. In detail, for two records from two different databases to be considered as a match, the following conditions must be jointly verified:

- The publication year of the two must not differ by more than a year.
- The edit distance (minimum number of operations required to make two strings equal) between the two titles must be less than 2.

Before proceeding, it is necessary to make a clarification. As just mentioned, the data used for the following analysis were not the raw output of Scholar but the data disambiguated within the EVA project. It should be emphasised that there is a trade-off between correct disambiguation and proper coverage (see Figure 4). If you want to increase correct disambiguation, this inevitably increases the number of products that really are in SUA but that are not identified in Scholar – the under coverage. If you want to increase the coverage, you must be willing to tolerate a higher disambiguation error, that is, increase the number of instances that are not in SUA but that match with a record in Scholar.

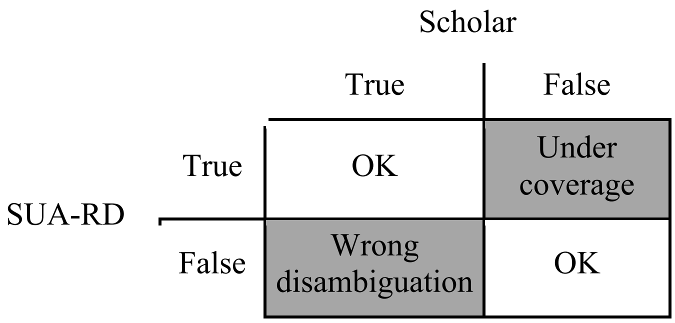

Figure 4.4. Trade-off between correct disambiguation and proper coverage

The parameters and thresholds have to be properly calibrated to favour one or the other objective. The following analyses, then, are conditioned by the choices that were made in the EVA project, and changes in these calibration parameters may have led to slightly different results.

Figure 4.5 shows the consistency of the SUA-RD publications in the period 2011–2013 and compares this with the raw output of Scholar, with the cleaned output of Scholar disambiguated by author (EVA1), with the cleaned output of Scholar disambiguated by author and sector (EVA2) and with Scopus and WoS. What is striking is certainly the different order of magnitude of the publications in the SUA-RD, not only compared to WoS and Scopus but also compared to EVA. SUA-RD collected 5728 contributions in 2011, 6112 in 2012 and 5593 in 2013; EVA2 never exceeds 1500 contributions, which is still fewer than those of Scopus and WoS. Even if the observation window is very narrow, it should be noted that for the SUA-RD there are no signs of the exponential curve that characterises the citation databases. This confirms that a similar trend does not capture the dynamics of scientific production but rather the growing ability of citation databases to cover such production.

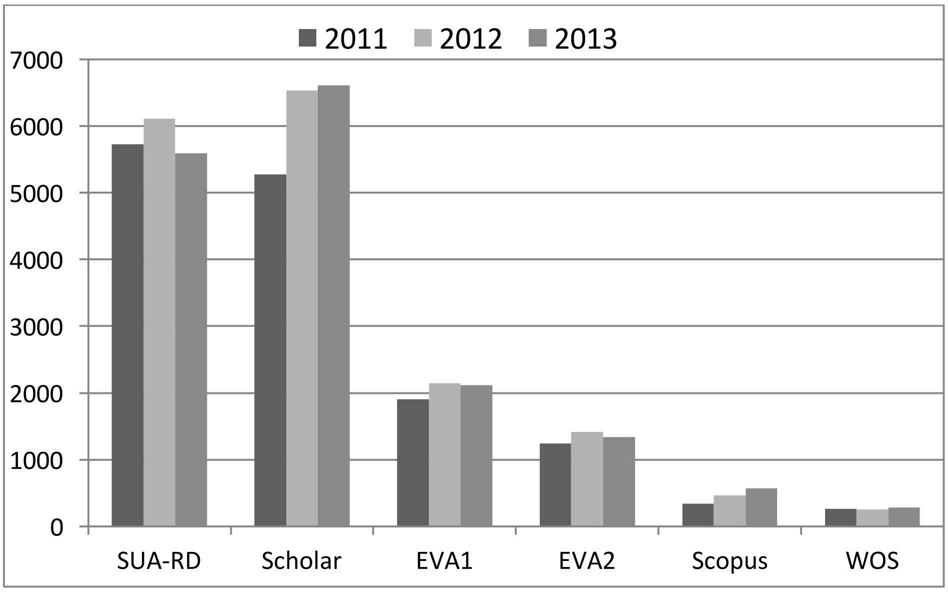

Figure 4.5. Year of publication of contributions in SUA-RD, Scholar, EVA, Scopus and WoS (absolute values)

SUA-RD is used as a benchmark. Therefore, it is important to evaluate how it covers Scopus and WoS in order to understand the appropriateness of this database.

Only 61.6% (total 1356) of Scopus records are covered by SUA-RD, and only 60.7% (total 780) of WoS records are covered by SUA-RD. These percentages respectively become 64.7% and 74.9% if only the type of article is considered, and they are obviously lower, at 46.1% and 26.2%, respectively, if other types of record are considered which are different from article, chapter and book chapter.

This result highlights the fact that there are inconsistencies in how the data are stored in citation databases and in the SUA-RD. Sometimes the titles are in Italian and sometimes in English; sometimes the years do not coincide; in one of the two sources, a subtitle is present; there are duplicates.

Note that this result refers to the period 2011–2013. In recent years, the institutional archives of universities have standardised and implemented mechanisms able to connect the publication to the corresponding records in the citation databases when they are present. If the exercise were to be repeated now, the result would surely be very different. Moreover, in the period considered only a few institutional archives of universities were Open Access. Recently, all institutional research archives have become Open Access. This has introduced transparency, reproducibility, accountability and visibility. In future exercises, the inconsistencies will probably be drastically reduced even if the problem is not completely solved.

Another issue is the types of publications represented by the databases. The distribution of product types, as seen from Figure 4.6, is very different in the three databases. Scopus and WoS primarily include articles, while the modal type for SUA-RD is instead the book chapter. Books comprise more than 15.2% of the production of scholars in Area 14, according to SUA-RD, but less than 3.2% in Scopus and WoS. Many books could be in Italian, and in fact, the percentages of book chapters are also very different, that is, 38.5% versus 16.6% for Scopus and 21.2% for WoS. It is important to note also that the Other category includes records that are very different: conference proceedings, translations, reviews, forewords and afterwords in SUA-RD and discussions and rejoints in Scopus and WoS.

Figure 4.6. Distribution of type of publication by database

After discussing the reliability of the database SUA-RD used as a benchmark, we shall now proceed with the analysis of the coverage of the three data sources Scholar, Scopus and WoS with respect to this benchmark database. As highlighted in Table 4.2, the Scholar coverage is much higher than that of Scopus and WoS, although the 14.8% coverage does not lead to the conclusion that Scholar is currently able to cover the scientific production of the scholars of Area 14.

Table 4.2. Coverage of SUA-RD by three sources: Scholar, Scopus and WoS

SUA	Scholar		Scopus		WoS	
	N	%	N	%	N	%
No	14,925	85.2	16,772	95.3	17,149	97.1
Yes	2,588	14.8	836	4.7	513	2.9
Total	17,513	100	17,608	100	17,662	100

We will now assess whether the coverage of Scholar differs by type of publication. We also consider the ASN classification for the publication, where present. The coverage may also vary according to the characteristics of the authors or universities the authors belong to. The same analyses were also carried out for Scopus and WoS, but they are not reported, since the percentage of coverage differs little.

Table 4.3: Coverage of SUA-RD in Scholar by the type of publication

Type		No	Yes	Total
Article	N	3,540	1,356	4,896
	%	72.3	27.7	
Book	N	2,187	468	2,655
	%	82.4	17.6	
Book Chapter	N	6,151	566	6,717
	%	91.6	8.4	
Other	N	3,099	194	3,203

	%	93.9	6.1	
Total	N	14,887	2,584	17,471
	%	85.2	14.8	

A first difference, as seen in Table 4.3, can be observed in the type of publication. Articles and books are covered by Scholar with respective percentages of 27.7% and 17.6%.

The coverage changes according to the level of the publications. In particular, 33% of the publications considered as Class A according to the ASN classification are covered by Scholar. Following the same ASN classification, 27% of scientific journals considered are covered by Scholar. Even the presence of a foreign co-author appears to increase the level of coverage: 27% of publications with a foreign co-author are covered by Scholar.

Regarding the characteristics of the authors, a different coverage does not seem to emerge, depending on the author's qualifications, while the disciplinary sector the publication belongs to seems to affect the SUA-RD coverage by Scholar. The sectors which are more covered are A2 (Political Science), C2 (Sociology of cultural and communication) and D1 (Economic sociology, urban and environmental sociology), with a coverage of 19%, 19% and 18%, respectively. The sectors which are less covered are A1 (Political Philosophy) at 10% and B1 (History of political thought and institutions) at 6%.

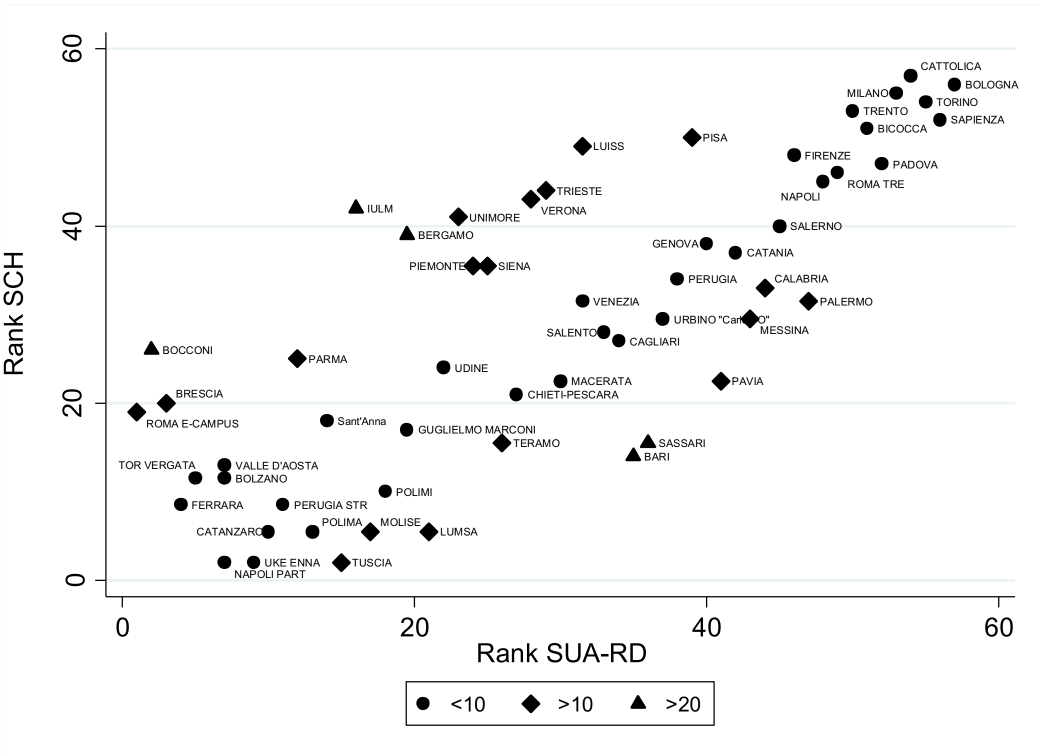

Figure 4.7. Ranks of universities for SUA-RD and Scholar

We now consider the characteristics of the university to which the authors belong. No difference is shown as a result of the size of the university. But the coverage is much higher for universities of the North-West (23%) and the North-East (18%) than for those in the Centre (10%) or the South and the Islands (8% for both).

As part of a further analysis, we wondered whether the ranking of universities or sectors greatly varied, considering Scholar instead of SUA-RD.

The following scatter plot shows the ranks of the universities (Figure 4.7) and the sectors (Figure 4.8) according to their rank on SUA-RD (horizontal axis) and their rank on Scholar (vertical axis). The plot shows that the sorting does not change substantially, even if, in the case of the universities, a large number change their ranking by 10 positions and five universities even change their ranking by 20 positions. The Spearman correlation indices are 0.77 for the universities and 0.89 for the sectors.

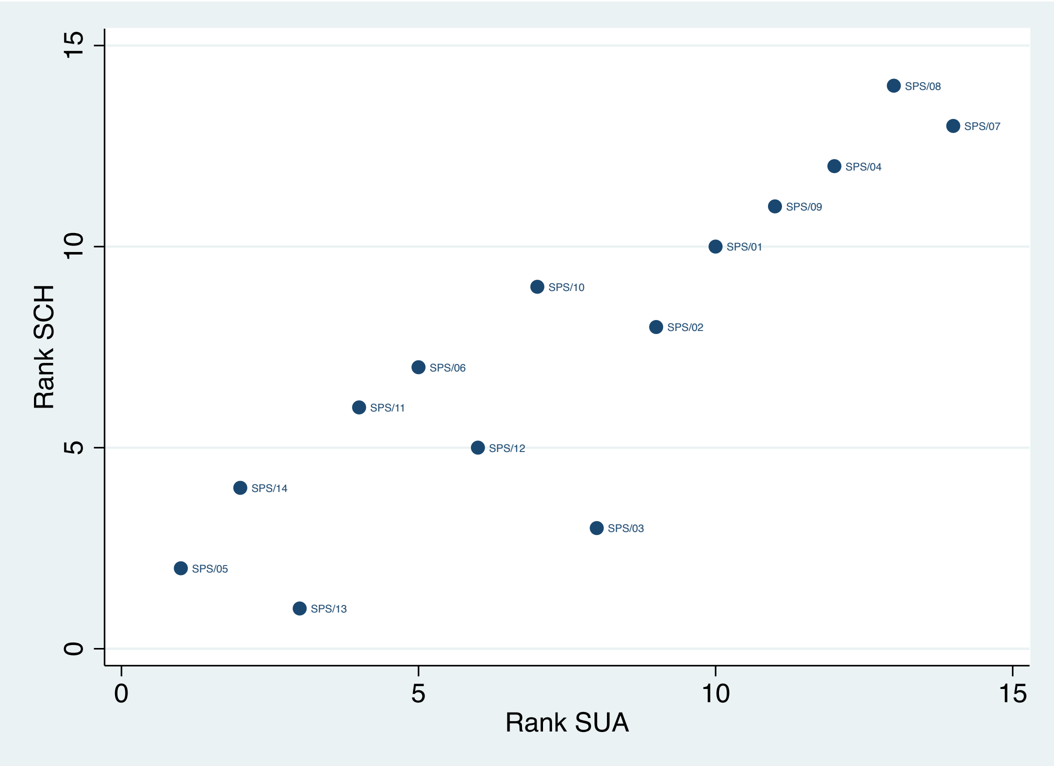

Figure 4.8. Ranks of sectors for SUA-RD and Scholar

5. Reliability and validity of bibliometric indicators

5.1. Descriptive statistics of output and recognition indicators

Starting from the large amount of data gathered, we decided to focus on three kinds of indicators.

- Number of contributions in Scholar, Scopus and WoS.
- Average number of citations per contribution in Scholar, Scopus and WoS.
- H-index in Scholar, Scopus and WoS.

The first indicator refers to scientific output, whereas the other two refer to scientific recognition. The average number of citations indicator was preferred to a simple arithmetical sum in order to normalise these values for the scientific output of the authors. Another option could have been to normalise for the activity period (computed as the difference between the publication years of the first and last contributions), but the number of contributions better grasped the volume of scientific output. Also, a preliminary analysis of distribution asymmetry confirmed the choice to normalise according to the number of contributions.

We also decided not to take into account other indicators related to the kind of publication or linked to the journal they were published in (Class A, IF, SCImago, SNIP etc.) because of the difficulty for Scholar in accurately identifying the kind of publication and the journal for the scientific articles. For further clarification, see the chapter by Alfio Ferrara, Stefano Montanelli and Stefano Verzillo in this volume.

All the indicators refer to the period 1998–2014, since all three data sources report reliable numbers for each of the years considered.

Table 5.1 presents the features of the statistical distribution for the nine indicators considered here. The results are not surprising. Regarding the number of contributions and the average citations per contribution, Scopus and WoS show similar values and they are very different from Scholar; the median of the distribution is significantly higher for Scholar, not to mention the range (here it coincides with the maximum value of the distribution) and the standard deviation. The situation changes slightly if we consider the variability in relative terms by means of the coefficient of variation, which shows similar distributions. The values for asymmetry and kurtosis are very high; to deal with them, we will use different techniques and different versions of the original variables through the entire analysis. Worthy of note is the median value for the number of citations in WoS (0), meaning that at least half of the authors from Area 14 have no publications at all in WoS.

Table 5.1. Descriptive statistics of bibliometric indicators

	Median	**Mean**	**Max**	**Std. dev.**	**Variation coefficient**	**Asym-metry**	**Kurtosis**	**(N)**
Number of contributions								
Scholar	4	8.2	147	11.7	1.4	3.7	26.3	(1697)
Scopus	1	2.1	41	3.9	1.8	3.3	19.0	(1697)
WoS	0	1.4	53	3.1	2.1	5.5	60.8	(1697)
Number of citations per contribution								
Scholar	2.0	5.5	216.6	13.5	2.4	8.5	102.5	(1372)
Scopus	0.6	2.5	72.6	5.8	2.2	5.9	52.4	(924)
WoS	1.0	2.5	87.0	6.0	2.3	7.0	74.9	(692)
H-index								
Scholar	2	2.6	22	2.9	1.1	2.3	10.1	(1372)
Scopus	1	1.1	12	1.5	1.3	2.4	11.3	(924)
WoS	1	1.0	10	1.2	1.2	2.0	9.3	(692)

The values for asymmetry and kurtosis are indeed lower for the H-index for all three data sources; this peculiarity makes the H-index very desirable from a statistical point of view if compared to the other citation indicators. The H-index is defined by its creator in this way: “A scientist has index h if h of his or her N_p papers have at least h citations each and the other (N_p - h) papers have $\leq h$ citations each” (Hirsch 2005, 16569). For instance, an H-index equal to 10 means that a researcher has 10 publications which are quoted at least 10 times. It is evident that this indicator aims to consider both scientific output and recognition.

5.2. Reliability of output and recognition indicators

As mentioned above, the reliability of the indicators of scientific output and recognition is assessed here by looking at the reproducibility of the results and is measured on the basis of a correlational analysis. Table 5 presents the main results of this analysis. First of all, all correlation coefficients are rather high, which means a general reliability for the three indicators in the three data sources considered. Secondly, for all three indicators, the correlation coefficients for Scopus and WoS are higher than the correlation coefficients for Scholar. Regarding this latter, its correlation with Scopus is always higher than the correlation with WoS.

This correlation structure seems to capture the specificities of the three data sources. If we hypothesise a continuum ranging from international journal articles to different types of publications even at the national level, we find a clear order of the three data sources on this continuum (WoS, Scopus, Scholar), with Scopus being much closer to WoS than to Scholar.

Moreover, it should be noted that the correlation coefficients between the citation indicators are higher than the correlations between the indicators counting the number of contributions. This might mean that the structure of scientific recognition is more coherent if compared to the structure of scientific output. Put another way, the various specificities of the three data sources are more evident for the indicators concerning scientific output, whereas the dimension of scientific recognition is captured in a more stable way from all three sources.

Table 5.2. Correlation coefficients for the indicators of output and recognition in Scholar, Scopus and WoS (Spearman correlations for the original versions) and Cronbach's Alpha

	Contributions	**Citations per contributions**	**H-Index**
Scholar-Scopus	0.48	0.64	0.65
Scholar-WoS	0.42	0.59	0.57
Scopus-WoS	0.65	0.74	0.72
Cronbach's Alpha	0.77	0.76	0.86
(N)	(1697)	(534)	(534)

To take into account the distribution asymmetry for the indicators of contributions and average citations per contribution, in addition to the rank correlations (Spearman) for the original variables, the standard correlation coefficients (Pearson) were also calculated on alternative versions of such indicators. In the truncated version, a threshold value – equal to three interquartile differences over the third quartile – was defined, and the cases with higher values were restricted to this threshold. The version using deciles recoded the cases according to the related decile, whereas the dichotomous version distinguished the values equal to "0" from all the other values coded as "1". The square root was computed for the average number of citations per contribution, as suggested by the analysis based on the "ladder of powers" to reduce the distributions' asymmetry (Hamilton 2012, 129–132). All the correlation coefficients between these different versions showed stable values, with the exception of the dichotomous one, which presented lower correlations, probably due to the loss of information. All these procedures were not necessary for the H-index, which presented a more symmetrical distribution. Table 5.2 shows the Spearman correlations similarly to what was done for the other indicators.

As further proof of the general reliability of the indicators considered, Table 5.2 shows the values of the Cronbach's Alpha computed for each indicator for the three data sources. In every case, the values are higher than the usual threshold of acceptability (0.7).

5.3. Validity of indicators of scientific recognition: VQR 2004–2010

In this section, we will deal with the issue of the validity of indicators of scientific recognition. As previously stated, we will apply the procedure of criterion validation which evaluates the indicator's validity on the basis of associations with an external criterion linked to the measured property. This evaluation is based on the assumption that the citation variables are indicators of scientific recognition, a property which is presumed to be linked with the quality of the scientific product.

This analysis comprises two different parts. In this section, we will use the average scores obtained in the Research Quality Evaluation (*Valutazione della Qualità della Ricerca* - VQR 2004–2010) as external criteria, whereas in the next section we will use the scores from the National Scientific Habilitation (*Abilitazione Scientifica Nazionale* - ASN 2012). Both these procedures are managed by ANVUR and are based on *peer review* processes aiming to ascertain the scientific quality of the research. It should be noted that in the first case only the scientific production is considered, whereas in the second other aspects of scientific work – research projects, teaching activities, research activity in international institutes, awards etc. – are also taken into account.

For what concerns the VQR 2004–2010, the researchers were asked to choose three scientific products. The variable considered here is the average score for these three products. One of the following scores was assigned to each product: "1" for "Excellent", "0.8" for "Good", "0.5" for "Acceptable" and "0" for "Limited". In the case of missing products, the score was "-0.5", in cases of products that were impossible to evaluate, it was "-1" and it was "-2" for verified situations of scam or plagiarism. Concerning Area 14, results from the VQR show that 36.0% of the products scored "Excellent-Good", 30.6% scored "Acceptable" and 27.3% "Limited". The remaining 4.0% consisted of penalised products, including missing products, that is, products not delivered (167 products) as well as explicitly penalised products, including non-eligible products, plagiarism, self-plagiarism or scams (13 products). All this information was taken from the final report of the Research Quality Evaluation (VQR 2004–2010) (ANVUR 2013).

In this case, we considered the H-Index and the number of citations per product as indicators of scientific recognition. For the latter, given its asymmetry problems, we used the rank correlation coefficient (Spearman's Rho), whereas for the H-Index we used the usual Pearson's correlation coefficient. For both indicators, we considered two different versions, the original one covering the whole observation window (1998–2014) and a second version covering only the period considered by the VQR (2004–2010)[4].

All the correlation coefficients are reported in Table 5.3. Starting from the H-Index computed for the extended time span (1998–2014), the table shows that all the correlation coefficients are statistically significant, ranging from 0.33 (Scholar-VQR) to 0.29 (Scopus-VQR) and 0.26 (WoS-VQR). Using the VQR score as the external criterion, the Scholar H-Index seems to be the most valid, followed by Scopus and WoS. If we use the 2004–2010 time span – the same period considered by the VQR – the correlation coefficients as well as their order are basically the same. Even if we had computed the Spearman's Rho, either the values would remain the same or the differences would increase for the 2004–2010 observation window.

If we look at the correlation coefficients between the VQR scores and the other indicator of scientific recognition – the average number of citations per contribution – we see lower values. These coefficients still remain statistically significant. In addition, the range stating a higher validity

[4] Given that the VQR scores are highly confidential, these analyses were conducted directly by ANVUR. In this regard, we thank Marco Malgarini, Tindaro Cicero and Marco De Santis Puzzonia from ANVUR for their precious collaboration. We also need to report that given some difficulties in combining the CAVIB dataset with that of the VQR 2004–2010, the analysis was conducted for 1394 over 1697 cases (82% of the total sample).

for Scholar (0.27) and a lower for Scopus (0.23) and WoS (0.21) is confirmed.[5] The results are highly consistent even if we look at the 2004–2010 time span.

Table 5.3. Correlation coefficients between the average score on VQR 2004–2010 and the original versions of the indicators of scientific recognition for Scholar, Scopus and WoS for the two different time spans

	H-Index/VQR *Pearson Correlation*			**Citations per contribution/VQR** *Spearman Correlation*		
1998–2014	*r*	*p*	*(N)*	*rho*	*p*	*(N)*
Scholar	0.33	0.000	(1127)	0.27	0.000	(1127)
Scopus	0.29	0.000	(750)	0.23	0.000	(750)
WoS	0.26	0.000	(559)	0.21	0.000	(559)
2004–2010	*r*	*p*	*(N)*	*rho*	*p*	*(N)*
Scholar	0.33	0.000	(970)	0.27	0.000	(970)
Scopus	0.30	0.000	(463)	0.26	0.000	(463)
WoS	0.25	0.000	(374)	0.19	0.000	(374)

5.4. Validity of indicators of scientific recognition: ASN 2012

The analysis of the indicators of scientific recognition based on the ASN 2012 uses a different version of the same criterion validation procedure. In the case of VQR 2004–2010, we used a *concomitant validation* process based on the assumption that the quality of the publications evaluated with a peer review process in the VQR 2004–2010 and the value of scientific recognition measured by the citation indicators were two contemporary features. With this being a symmetrical association the use of correlation coefficients is justified.

In the case of ASN 2012, the situation is different. Here, the underlying question is whether the use of citation indicators can predict the outcome of the habilitation procedure. This kind of procedure is called *predictive validation*. In this case, the association is asymmetrical. We assumed that the citation indicators impacted the habilitation, and therefore, we ran a logistic regression with the result of the habilitation procedure (yes/no) as dependent variable and the continuous bibliometric indicators as independent variables.

Although ASN considered other aspects of scientific work beyond the scientific quality of publication (research projects, teaching or research activities in international institutes, awards etc.), the results according to the two criteria (VQR and ASN) are essentially coherent.

Table 7 shows the results for the logistic regression model for the two citation indicators (average citations per contribution and H-Index) and for the three bibliometric indicators computed within ASN. These three indicators, hereby used as *benchmarks*, are the following:

1. Number of books
2. Number of journal articles or book chapters
3. Number of articles published in “Class A” journals

In order to allow a comparison between the various effects (odds ratio and predicted probabilities), we also used the base 10 normalised version of the indicators. This means that all the variables were

[5] If instead of the Spearman correlation coefficients we had calculated the Pearson’s coefficient, the results would have been even lower.

scaled to make the range of variation uniform (from 0 to 10) (Corbetta, Gasperoni and Pisati 2001, 80–82). For the abovementioned reasons, the normalisation of the variables concerning the average number of citations per contribution was computed starting from the square root of the original variables.

Table 7 shows some features of the nine models. Given the predictive validation perspective, the more relevant results regard the fit of the models rather than the size of the effects. The table displays three different indexes of model fit. The Pseudo (McFadden) R^2 is based on the maximum likelihood function, whereas the two different versions of Count R^2 define the model fit as a predictive power, which corresponds exactly to what we want to evaluate. The Count R^2 corresponds to the proportion of cases rightly predicted by the model, i.e. to the proportion of qualified researchers and non-qualified researchers rightly predicted by the model. However, this index does not consider the fact that even without the model it is possible to predict the dependent variable – the habilitation – using its mean, the proportion of habilitated researchers: if the value of the proportion of habilitated researchers exceeds 0.50 we can predict a successful habilitation for all the candidates, if the proportion is lower than 0.50 we can predict failure for all the candidates. Conversely, the Adjusted Count R^2 also considers the so-called Model 0, based on the mean of the dependent variable, and returns the proportion of rightly predicted cases among those not explicitly predicted by Model 0 (Menard 2002, 27–36). Put another way, beyond the cases rightly predicted by Model 0 (based on the mean of the dependent variable), how many cases can Model 1 – the one with the bibliometric indicator – predict?

Table 7. Logistic regressions for habilitation with different normalised bibliometric indicators (H-Indexes, average citations per contribution in Scholar, Scopus and WoS and ASN 2012 indicators)

	OR	**Std. Err.**	**z**	**P>|z|**	**Pseudo R^2**	**Count R^2**	**Adjusted Count R^2**	**(N)**
H-Index (normalised)								
Scholar	1.58	0.119	6.13	0.000	0.05	0.60	0.19	(696)
Scopus	1.60	0.156	4.83	0.000	0.04	0.57	-0.02	(509)
WoS	1.39	0.143	3.23	0.001	0.02	0.64	0.00	(397)
Average citations per contribution (normalised square root)								
Scholar	1.34	0.112	3.57	0.000	0.01	0.55	0.10	(696)
Scopus	1.22	0.088	2.88	0.004	0.01	0.58	0.00	(509)
WoS	1.22	0.111	2.20	0.028	0.01	0.64	0.00	(397)
ASN Indicator (normalised)								
1	0.98	0.071	-0.39	0.695	0.00	0.48	-0.04	(696)
2	2.09	0.235	6.61	0.000	0.04	0.61	0.21	(509)
3	1.93	0.163	7.84	0.000	0.07	0.64	0.28	(397)

The results shown in Table 7 are rather clear. The Scholar H-Index permits the prediction of 19% of cases, more than Model 0. Regarding the average citations per contribution indicator in Scholar, the Adjusted Count R^2 is instead 0.10. These values are lower than the two best performing indicators from ASN (the number of articles published in "Class A" journals and the number of published articles or book chapters), but they are decisively higher than the bibliometric indicators from Scopus and WoS, which are unable to increase the predictive ability of Model 0.

6. Conclusions

The set of results which we have discussed suggests to keep distinct the bibliometric level from the evaluative one. It must be stressed that bibliometrics and evaluation are not the same thing. Even if the connection between evaluation and bibliometrics is perceived by many scholars as very close, the two disciplines are relatively autonomous. Bibliometrics emerged from science studies and can be seen as a *tool* of research evaluation.

At the bibliometric level, the main results are as follows. First, the correlational analysis of the different indicators confirms the reliability of bibliometric indicators. The specificities of the different repertoires are also confirmed, with WoS and Scopus correlating highly with each other and Scholar correlating more with WoS than with Scopus. To assess the validity of the scientific recognition indicators, we used criterion validation. In particular, we used two criteria: the individual scores of VQR 2004–2010 and the results of ASN 2012. Both criteria confirm the validity of the scientific recognition indicators based on all three data sources, although Scholar is systematically more valid than the others.

Combining the two sets of results, a sort of scale (Scholar/Scopus/WoS) emerges, with the reliability increasing from Scholar to WoS while the validity instead increases from WoS to Scholar. Such a pattern is likely to be associated with the scientific perimeter covered by the different bibliometric datasets. While WoS and Scopus preferably cover articles in international journals, Scholar is more open to other types of publications and to national literatures. These different scientific perimeters become an advantage in terms of reliability for WoS and Scopus, which are more similar to each other, and an advantage in terms of validity for Scholar, which better discriminates among those researchers who do not publish articles in international journals.

Then there are some side results. The scientific recognition indicators are more reliable than output indicators. Furthermore, according to all the results – and not only in the validation analysis – the H-index seems to work better than the average number of citations for contribution. The bibliometric longitudinal comparisons are biased by a coverage capacity of bibliometric datasets that is growing over time.

Finally, we must not forget to mention a relevant weakness of Scholar that might go unnoticed. The difficulties in identifying for each contribution the type of publication (book, chapter, article etc.) and the journal of publication in the case of scientific articles drastically reduces the number of potential bibliometric indicators derivable from Scholar. It is therefore problematic to develop an analysis of the different types of publication and to define scientifically relevant indicators (IF, SCImago and the like) that are based on the characteristics of the publication venues.

With regard to evaluation, the main issues are those of Scholar coverage and data quality at the contribution level. Using as a benchmark SUA-RD data referring to the years 2011–2013, Scholar covers about 15% of the scientific production compared with 5% of Scopus and 3% of WoS. What is most striking is the Scholar absolute value. Although it certainly has a large degree of underestimation, the analysis shows that Scholar still processes a significant selection of the work produced by the scientific community. It also shows how Scholar is certainly more comprehensive than commercial repositories while far from being exhaustive. This is a first result which shortens the distance between Scholar, Scopus and WoS.

Second, it has been shown that the data collection process for the various sources – in particular the correct attribution of the contributions to the authors – is hampered by similar problems. It was also found that the quality of the data is certainly higher in commercial repertoires than in Scholar, but it is still not absolute. The three databases seem to be more similar than one might think. In other words, for all three sources, you have a trade-off between coverage and disambiguation, which, however, is magnified in the case of Scholar for the lower quality of the dataset.

If the validity and reliability of bibliometric indicators drawn from Scholar prove satisfactory at the aggregate level, there remains the problem of data quality at the level of single publications. Data quality is a critical point for any evaluation system, and the tolerable error threshold varies depending on the object of evaluation. If you evaluate a scientific structure (university, department, research lab etc.), you can appeal to the requirement that a mistake should not be such as to distort the results of the assessment. If you evaluate an individual, even this criterion crumbles because the error will induce a deep crisis of confidence in the system. The result is thus a suggestion for caution in the use of Scholar data, especially with respect to individual assessments.

Finally, it should be noted that these considerations are made in a changing environment, where great efforts have been undertaken to solve the thorny problem of correct attribution of publications to the authors. One example is the ORCID system adopted by ANVUR for the VQR 2011-2014. Commercial repertoires are also developing systems to improve disambiguation. It is therefore plausible that in just a short time the landscape may change dramatically by simplifying the combination of bibliometrics and evaluation. Another phenomenon to consider is the growing diffusion of Open Access in the Italian academic system. This will favour the presence on the web (and therefore on Scholar) of Italian scientific production.

References

ANVUR (2013) *Valutazione della Qualità della Ricerca 2004-2010 (VQR 2004-2010). Rapporto finale di area. Gruppo di Esperti della Valutazione dell'Area 14 (GEV14)*, Rome, 30 June.

Baccini, A. (2010) *Valutare la ricerca scientifica. Uso e abuso degli indicatori bibliometrici*, Bologna, Il Mulino.

Clemens, E.S., Powell, W.W., McIlwaine, K. and Okamoto, D. (1995) Careers in print: books, journals, and scholarly reputations, in *American Journal of Sociology*, vol. 101, no. 2, p. 433-494.

Cole, S. and Cole, J.R. (1967) Scientific output and recognition: a study in the operation of the reward system in science, in *American Sociological Review*, vol. 32, no. 3, p. 377-390.

Corbetta, P. (2003) *Social research. Theory, methods and techniques*, Thousand Oaks, CA, Sage.

Corbetta P., Gasperoni G. and Pisati M. (2001), *Statistica per la ricerca sociale*, il Mulino, Bologna.

Franceschet, M. (2011) PageRank: standing on the shoulders of giants, in *Communications of the ACM*, vol. 54, no. 6, p. 92-101.

Hamilton L.C. (2012) *Statistics with Stata: updated for version 12*, Brooks/Cole, Cengage Learning, Belmont, CA.

Hicks, D. (2004) The four literatures of social science, in Moed, H.F., Glaenzel, W. and Schmoch, U. (eds.) *Handbook of Quantitative Science and Technology Research*, Dordrecht, The Netherlands, Kluwer Academic Publishers, p. 473-496.

Hirsch, J.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, vol. 102, no. 46.

Martin, B.R. and Irvine, J. (1983) Assessing basic research. Some partial indicators of scientific progress in radio astronomy, in *Research policy*, vol. 12, p. 61-90; partial Italian translation in Viale, R. and Cerroni, A. (2003) (eds.) *Valutare la scienza*, Soveria Mannelli (CZ), Rubettino.

Menard S. (2002) *Applied logistic regression analysis*, second edition, Sage, Thousand Oaks, CA.

Merton, R.K. (1973) *The sociology of science*, Chicago, The University of Chicago Press; Italian translation: *Scienza, religione e politica*, Bologna, Il Mulino, 2011.

Moed, H.F. (2005) *Citation analysis in research evaluation*, Dordrecht, The Netherlands, Springer.

Nederhof, A.J. (2006) Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A review, in *Scientometrics*, vol. 66, no. 1, p. 81-100.

Research Evaluation and Policy Project (2005) *Quantitative indicators for research assessment – A literature review*, Discussion Paper 05/1, 30 March.