DIPARTIMENTO DI INGEGNERIA INFORMATICA
AUTOMATICA E GESTIONALE ANTONIO RUBERTI

SAPIENZA
UNIVERSITÀ DI ROMA

An empirical approach to compare the performance of heterogeneous academic fields

Giancarlo Ruocco
Cinzia Daraio

Technical Report n. 3, 2012

An empirical approach to compare the performance of heterogeneous academic fields

Giancarlo Ruocco   Cinzia Daraio*

October 20, 2012

Abstract

In this paper we propose a "scaling-based" empirical approach to assess the scientific performance of heterogeneous academic disciplines. It relies on the idea that if we take into account for their two main sources of heterogeneity, the bibliometric distributions of different academic fields can be superimposed and collapse to a unique master curve by a single scaling parameter. By using data on the scientific production of around 2,500 scholars of the university of Rome "La Sapienza" from the Web of Science (WoS) over 2004–2008 we *i*) demonstrate the existence of a master curve; *ii*) determine the scaling factors which are the cornerstone to compare different academic fields; and *iii*) show that the master bibliometric distribution follows a Log-normal law.

Keywords: research assessment, normalization, scaling, universality, Italian universities

[*] **Ruocco**: Department of Physics, University of Rome "La Sapienza", Roma, Italy; email giancarlo.ruocco@roma1.infn.it. **Daraio**: Corresponding author. Department of Computer, Control and Management Engineering Antonio Ruberti (DIAG), University of Rome "La Sapienza", via Ariosto, 25, I00185 Roma, Italy; tel. +390677274068; fax +390677274074; email daraio@dis.uniroma1.it.

Acknowledgements: this paper has been presented during a MORE@DIAG Seminar, the 19 September 2012, Sapienza University of Rome. We thank seminar participants and in particular Henk F. Moed for useful comments and discussions. The usual disclaimers apply.

1 Introduction

The recent undergoing rapid changes in national systems of research and innovation, along with changes in economic conditions, are challenging European universities in acting a distinguishable role within the national economy. In this new context, universities are facing an important period of extraordinary change and transition, characterized by an increasing number of missions to accomplish while trying to have a more business-oriented behavior focused on competition.

Accordingly, in Europe, governments and national agencies for the evaluation of research activity are increasingly introducing elements of competition and research funds are more and more allocated according to some measures of "success" in the research activities. As a consequence, European universities[1] are learning how to compete (Deiaco et al. 2009; Deiaco et al. 2012).

This is particularly true for Italy where, historically, the evaluation of research outcome has never been accepted as a base for funds' allocation, and where the second national research evaluation exercise is under way and the results will be used to distribute accordingly a non negligible share of the governmental funds to universities. It is also planned that this share will be constantly increased in the following years.[2]

Beside the issue of funds allocation, the Italian Ministry of Education (MIUR) and the Italian National Agency for the Evaluation of Universities and Research Institutes (ANVUR) recently introduced explicitly bibliometric parameters, based on number of publications and citations, for the evaluation of candidates and evaluators for the national scientific qualification and identify some of the Italian Academic disciplines as "bibliometric" ones (ANVUR *Delibera* n. 50 of 21 June 2012), i.e. capable to be "measured" using bibliometric data.

As a matter of fact, Italian universities, pushed by the current macroeconomic situation and by the recent laws, are undergoing rapid changes in their governance and are starting to implement the new laws, developing internal systems of performance assessment.

As far as the assessment of performance is concerned, one of its main critical issue relies on the comparison of different academic research fields, each with its own "fertility ", publication practices and features. The existence of a different scientific production among different disciplines is self evident, and has been also recently stated by the League of European Research Universities (LERU, 2012) "Bibliometric outputs/outlets differ between disciplines. [...] These differences need to be taken into account in assessments in these areas".

Is the aim of the present paper to propose a bibliometric methodology for comparing the performances of different academic fields taking into account their own specificities. It

[1] For a comparative analysis on European universities microdata see Daraio et al. (2011).

[2] For a macro bibliometric analysis of Italian science with respect to the main European countries over the period 1980–2009, and its implications in terms of funding, see Daraio and Moed (2011).

is based on a "scaling approach", typical of statistical mechanics, but applied to assess university scientific performance.[3]

Each Italian academic staff member (scholar) belongs to an academic disciplinary sector (called in Italian "Settore Scientifico Disciplinare", SSD hereafter). In this paper we analyze the "bibliometric" academic disciplinary sectors reported in Appendix A.

The paper is organized as follows. In the next section we present our approach and previous literature. Section 3 describes the data and main bibliometric indicators used in the analysis, whilst Section 4 illustrates the method followed in the elaborations. In Section 5 the main results are reported. Section 6 analyzes the relationships between publications and impact indicators. Section 7 illustrates the distribution law calculated on the whole sample of scaled data, while Section 8 points to some potential limitations of our analysis. Finally Section 9 concludes the paper outlying further developments.

2 Previous literature and our approach

We apply a "physics" approach, based on scaling, in quantitative science and technology (S&T) studies as far as the investigation on the distribution of bibliometric indicators is concerned. S&T are conceived as a physical system of interacting sub-units the behaviour of which can be described by more general laws analogously to physical law.[4]

Quantitative studies of science have investigated the distribution of bibliometric indicators (publications and citations) since the seminal works of Lotka (1926), Naranan (1971) and Price (1976), finding power law characteristics of the science system. More recent empirical evidence can be found in Seglen (1992), Redner (1998), van Raan (2006), Radicchi et al. (2008), Glanzel (2010), Albarran et al. (2011), Evans et al. (2012). There is hence a wide empirical evidence that the distributions of bibliometric indicators are highly skewed.[5]

The presence of power laws might indicate that the underlying generating process is neither regular nor stochastic; power laws could point to the existence of "self organized" criticality (Bak et al. 1987), or to an "edge of chaos" dynamics (Langton, 1990). However, the exact mechanism behind the empirical laws found in the literature is still far from being reached, even if some attempts have been done in the literature (see e.g. van Raan, 2001).

In this paper we are not focusing on the ultimate mechanisms giving rise to these simple

[3] For a general presentation and a rich empirical evidence on universities as strategic making units and university performance in Europe, see Bonaccorsi and Daraio (2007a,b).

[4] For a review, see van Raan (2004).

[5] For a whole presentation of bibliometric and informetric distributions see Egghe and Rousseau (1990). Simon (1955) and Laherrére and Sornette (1998) are useful references for a general overview on skew distributions, see also Stock (2006). For a presentation of the mechanisms for generating power laws and the methods to detect them see Newman (2005).

distributions (nor to their specific mathematical expressions, although in the final section we provide a specific distribution law) but we show that such approach could be particularly useful for evaluation purposes. If an empirical general law is found, able to model different or heterogeneous disciplines (SSDs in the Italian system) by few specific discipline-dependent parameters, this would be of great value to derive useful information, to predict (estimate) e.g. the number of papers per year or the number of citations per year, and so on, which have to be produced by a scholar of a specific SSD to reach the median values; and/or the number of papers to be produced, or citations to receive, to be in the top 1% or 10% or 25% of their specific SSD distribution. The reference to median values is institutionally important because it is considered by the Italian law for being in the national scientific qualification committee and to apply for obtaining the national scientific qualification.

In this paper we consider one among the different activities of university, namely research, and show that modeling its evaluation analysing bibliometric distributions according to a scaling approach is a very useful and promising approach.

In the next section we describe the data and the main indicators used in the analysis, as well as the level of the analysis that best fits in our framework.

3 Data

We focus on three different indicators of the scientific production of a researcher, they are identified by the symbol ε ($\varepsilon \in \{P, C, IF\}$)[6] and are listed here below. Each indicator has been measured over the five years period 2004-2008 and the values considered are their yearly averages:

- PUB, number of publications authored by a scholar;
- IF, sum of the impact factor of the journals of all the author's publications; the impact factor of a journal has been divided by the median of all journals' impact factors in the same subject category.
- CIT, total number of citations (including self citations) of the scholar's publications; the citations of a publication have been divided by the median number of citations of all Italian publications, of the same type and year falling in the same subject category.

These indicators have been computed on the Web of Science database by Thomson Reuters[7] for a total of 2471 scholars, belonging to the bibliometric SSDs (around 125, see

[6] Where P stands for publications (PUB), C stands for citations (CIT) and IF for Impact factor.

[7] The indicators have been obtained by Sapienza university under a commercial agreement from Research Value Ltd which elaborated the data under license from WoS of Thomson Reuters. The authors do not have

Appendix), over 4200 scholars working at the university Sapienza in 2011, that have worked at Sapienza at least one year over the 5 years 2004–2008.

To the raw data of the Web of Science (WoS) has been applied an heuristic algorithm (D'Angelo et al. 2010) for reconciliation of the authors affiliation and disambiguation of the true identity of the authors, each publication (article, review and conference proceeding, according to WoS definition) is attributed to the university scholar that produced it. Further, a manual inspection and check was carried out over all scholars that in the period 2004–2008 did not have any publication in WoS or presented an average annual output in WoS higher or lower than the 20% of the data provided by departments. A total of 983 scholars where manually inspected and 1703 publications wrongly attributed were corrected[8].

To ensure the representativeness of publications in the WoS as proxy of the research output of the academic disciplinary sector (SSD) in the elaborations we considered only those SSDs where at least 50% of Italian scholars produced at least one publication (reported in the WoS) in the period 2004–2008.

The list of SSDs is reported in Appendix A.

In this paper we choose as the relevant unit of analysis the single scholar and her/his performance are measured against the SSD to which she/he belongs to. The SSD is a reasonable level of analysis able to deal with the heterogeneity of scientific production (indeed SSDs aggregate quite homogeneous disciplinary sectors) keeping the usefulness of the analysis for university strategic making; it can in fact be further aggregated at different levels, as for example Department level, *Settore Concorsuale* level, Area CUN level, and so on.

Of course, this level of analysis does not solve all problems. Universities themselves are collections of departments having considerable internal heterogeneity and also SSD have internal heterogeneity as well. In addition, from the point of view of research it is possible that a more relevant unit of analysis is the laboratory (Knorr-Cetina, 1995; Laredo and Mustar, 2001) or the research group level as observed by van Raan (2008, p. 566): "The research group is the most important working floor entity in science, as clearly shown by the internal structure of universities and research institutes, particularly in the natural sciences, in the medical research fields and increasingly in the social and behavioral sciences, but less in the humanities. However, obtaining data at the research group level is by far a trivial matter"; not the department or university.

We propose to use bibliometric indicators as a tool for assessing different SSDs that can be easily aggregated in Department, Schools and so on, providing useful support for the strategic decision making at the level of university. In particular for instance, resources for

access to the database on which the indicators have been calculated, so cannot calculate additional or different bibliometric indicators than those provided by Research Value.

[8] Also this analysis has been performed by Research Value Ltd under a commercial agreement with Sapienza University.

hiring new academic staff are centralized at university level but are allocated by SSD.

4 Method and analytical expression of the law

4.1 Rationale of the normalization

The modeling principle followed in the empirical fitting of the data is based on two general ideas about the possible sources of heterogeneity of academic fields' production that we detail below. In our framework, SSDs mainly differ for:

i) the percentage of researchers who do not have any product in WoS in the analysed period (2004-2008). We will refer hereafter to the researchers belonging to this group as "silent"[9];

ii) the skewness of their own SSD; i.e. the distribution of the top performers or outlying scholars is different across SSDs.

Hence, in order to obtain a general empirical law, which is able to model the general pattern of the distribution of the scientific performance of heterogeneous academic fields, we have to allow for a normalization which is able to take into account both components.

4.2 Empirical investigation on the distributions

In principle one is interested to study the distribution of the parameter ε ($\varepsilon \in \{P, C, IF\}$) for each SSD. However, the available sample includes several SSDs with small number of scholars (ranging from the most populated SSDs with a few tenths of observations to other SSDs with a minimum of 4 observations); this situation leads to a very noisy histogram representation. Therefore, we decided to work on the cumulative distributions - which are the object of the following analysis. By doing this choice we have the following advantages: *i*) the cumulative distributions, being the integral of the distributions itself are much less noising, and hence offer a more stable view of the pattern; *ii*) the data available -of the considered parameter- on the cumulative distributions are based on the national ranking of scholars, and hence come from a much larger population than the analysed sample (based on Sapienza).

[9] It is important to note that being "silent" does not mean being "inactive"; it could happen for instance that a "silent" researcher in our analysis has published many papers in journals not covered by WoS, or he/she has been a promotor of a big and challenging research project that will radically change a discipline, but the outputs of the project are not yet codified in scientific articles, or even that the researcher has played an important role in advising governmental body on strategic issues, and so on.

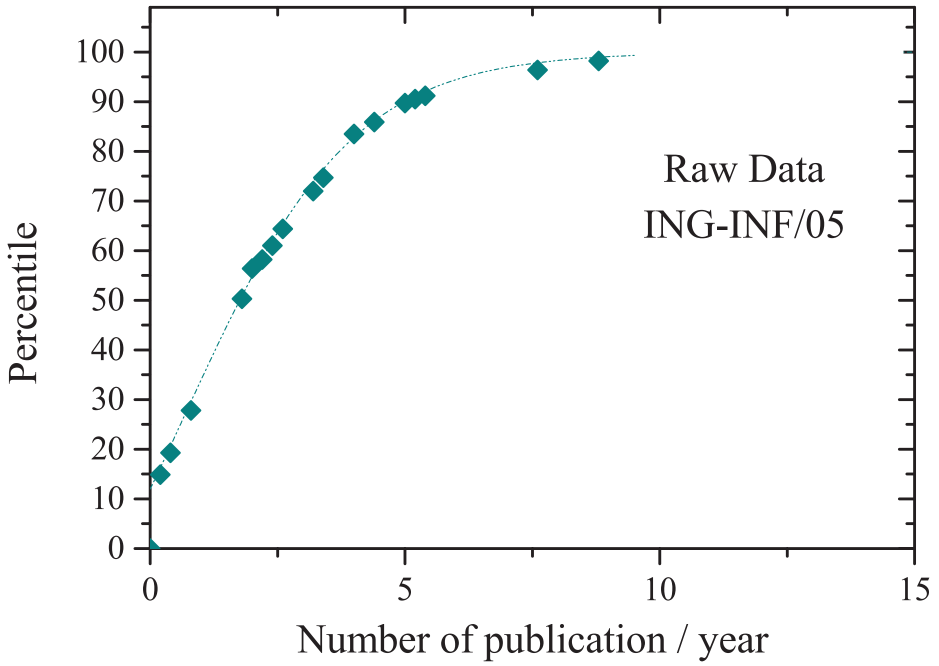

Figure 1: *Cumulative distribution of PUB for a specific SSD.*

In Fig. 1 we report as an example the cumulative distribution of the Sapienza's scholars in a specific SSD (ING-INF05), namely Information processing systems. In the vertical axis we report the "percentile" of a scholar (which indicates the percentage of the researchers of the indicated SSD that collected a number of "products" -in this case publications (PUB)- lower than that of the specific scholar) while the horizontal axis reports the (average yearly) number of publications. Each diamond in Fig. 1 represents a scholar and the dashed line is just a guide for the eye. As can be seen, the dots well cover the whole range of existence of the distribution, thus allowing us to be confident that the "Sapienza" sample can be profitably used to represent the real (national level) distribution.

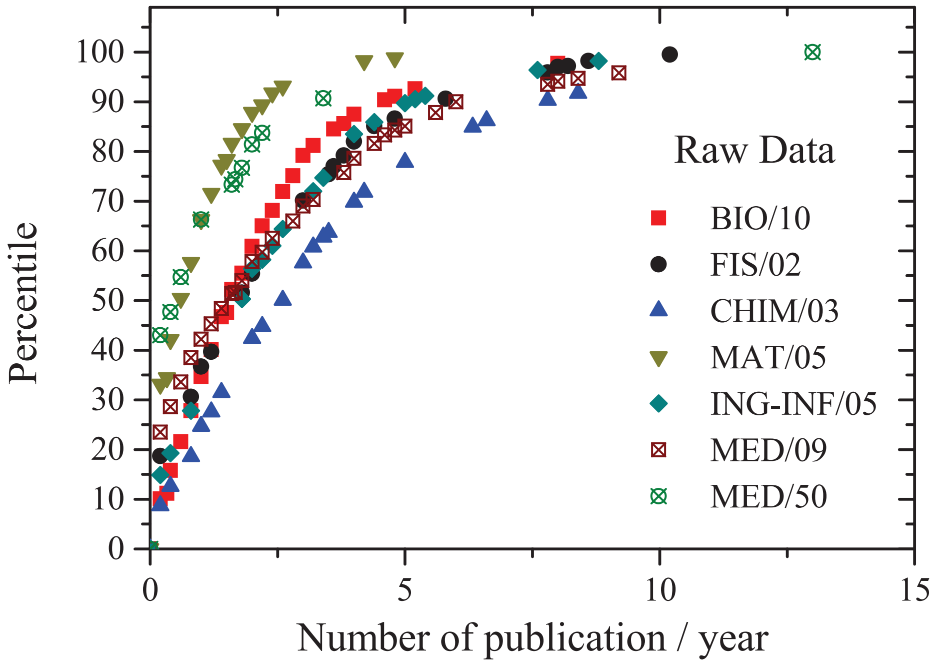

Figure 2: *Cumulative distributions of PUB for the indicated SSDs.*

Being interested in comparing different disciplines, in Fig. 2 we report the cumulative distribution of six different SSDs, chosen in order to cover different research areas. By looking at Fig. 2 one can observe the presence of the two sources of heterogeneity previously mentioned, i.e. the existence of a number of silent researchers, whose percentage is SSD-dependent, and the diverse SSDs' scientific fertility whose fingerprints can be found in the different slopes of the cumulative distributions.

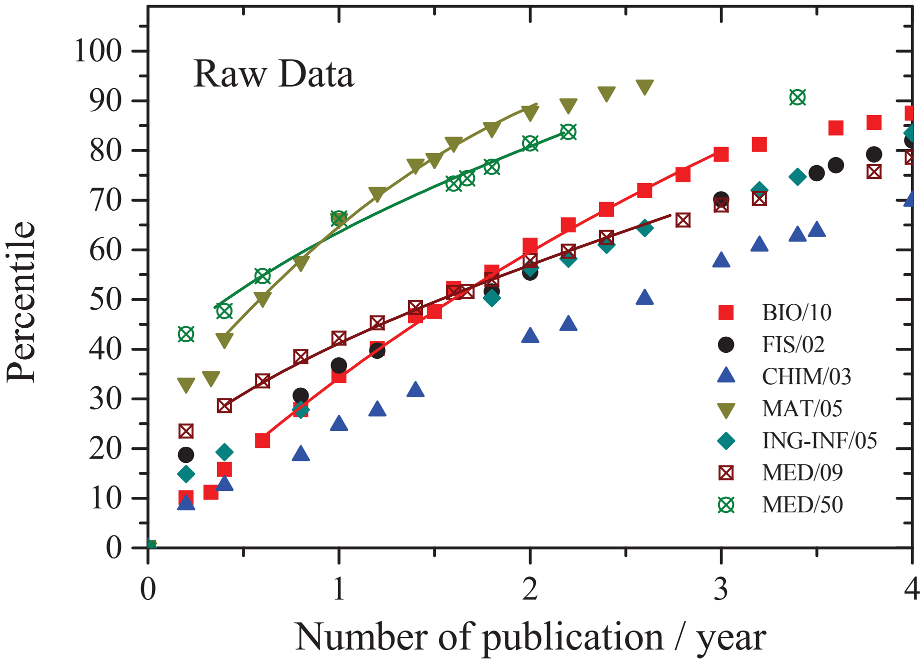

Figure 3: *Zoom of Figure 2 showing the intersection of some SSDs.*

In figure 3 we report a blow-up of Figure 2 which highlights the fact that different cumulative distributions cross each other, thus preventing a simple direct comparison between different disciplines. It is clear, in fact, that with the purpose of scaling the different SSDs' distributions into one single master curve we should avoid such intersections. Luckily we found that in order to solve this problem, it is sufficient to remove the silent researchers (scholars with no publications in WoS over the analyzed period) from the population, and repeat the analysis. The result is reported in Fig. 4, where one can still observe the diverse fertility of the various SSDs, but now their cumulative distributions do not cross each other any more. It is worth to emphasize that the observation of different SSDs' scholars crossing each other on the whole distribution (see Fig. 5) is an indication that the percentage of silent scholars and field fertility are not correlated factors.

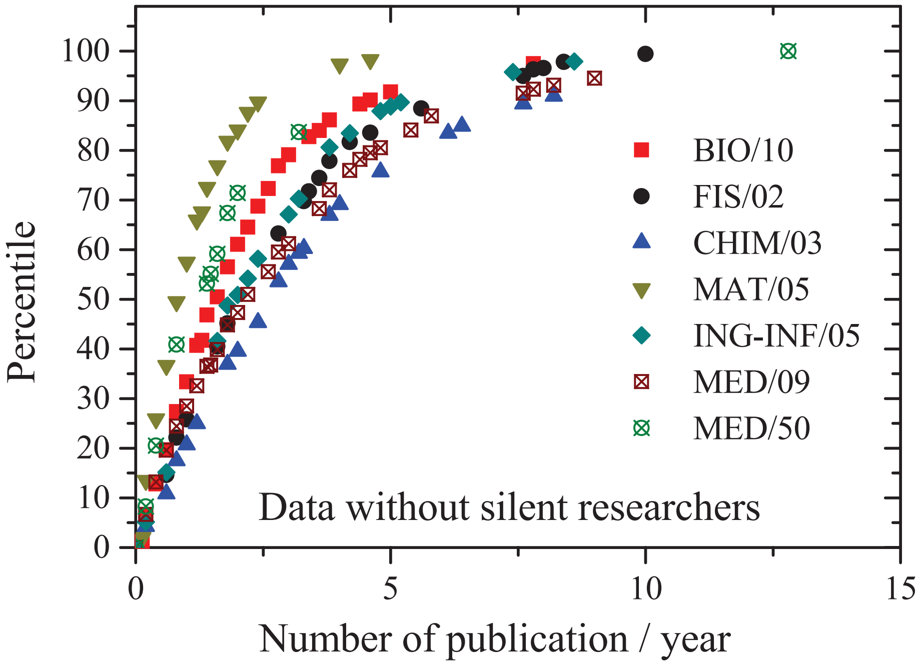

Figure 4: *Distributions of PUB for selected SSDs without silent researchers.*

Finally, figure 5 illustrates that dividing the number of publications (PUB) by its median value (determined graphically from Fig. 4 -at least for all the analyzed SSDs-) we obtain the collapse of all the cumulative distributions on a single master curve. This result is by far not trivial, and indicates that the distribution of the number of products has the same shape for all the SSDs, and that their only difference can be found in a single scaling parameter (beside the percentage of silent researchers).

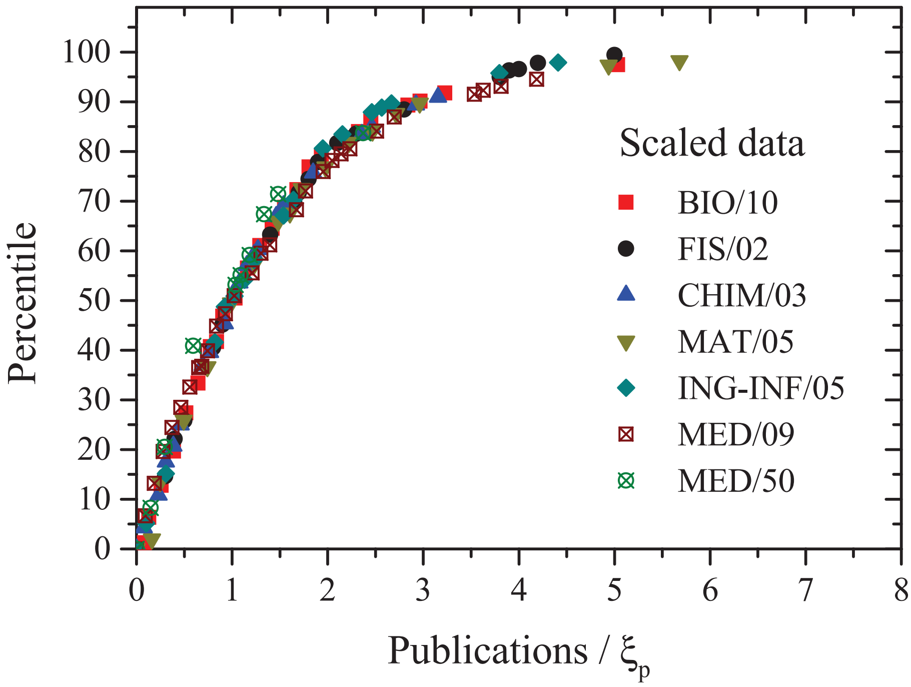

Figure 5: *Distributions of PUB for selected SSDs without silent researchers plotted as a function of the scaled variable PUB/ξ_p, being ξ_p the number of products that bring a researcher of the given SSD to the 50th percentile.*

In order to determine the scaling factors ξ_p for all the investigated SSDs, as well as the scaling factors ξ_{IF} and ξ_C for the distributions of respectively PUB, IF and CIT indicators, it is necessary to perform an automating procedure, thus to choose a fitting function and set-up a specific code for fitting this function to cumulative distributions.

4.3 Method applied to fit the empirical distribution

We are, at this stage, not interested in developing a theoretical model to explain the observed distributions, but we search for a simple relation, defined by two parameters, able to capture the two sources of heterogeneity, that we will use to fit the empirical data.

We are looking for a function that should verify the following conditions:

i) It should have value $B(s)$ when $x = 0$ (hence $B(s)$ is the percentage of silent researchers in the considered SSD s;

ii) it should have value A for $x \to \infty$ (and hence $A = 100$);

iii) it should grow almost linearly for small values of x and, *iv*) it should have a unique scale parameter (normalizing factor) $\gamma_{\varepsilon,s}$.

Being x the value of one of the bibliometric parameters analyzed ε ($\varepsilon \in \{P, IF, C\}$), the percentile of the analyzed population of a specific SSD (named s) that has the specific bibliometric parameter ε is found to be well represented by a *modified Boltzmann* function[10], $F(x;\varepsilon,s)$ $(0 \leq F \leq 100)$ as reported below:

$$F(x;\varepsilon,s) = A + \frac{3}{2}(B(s) - A)\frac{1}{1 + \frac{1}{2}\exp\left(x/\gamma_{\varepsilon,s}\right)}. \tag{4.1}$$

Let us define the function $\tilde{F}(x;\varepsilon,s)$ that we obtain eliminating the percentage of silent researchers in the considered SSD and setting $\tilde{F}(x;\varepsilon,s)$ to span the range between 0 and 1. We have then:

$$\tilde{F}(x;\varepsilon,s) = \frac{F(x;\varepsilon,s) - B(s)}{A - B(s)} = 1 - \frac{\frac{3}{2}}{1 + \frac{1}{2}\exp\left(x/\gamma_{\varepsilon,s}\right)}. \tag{4.2}$$

Being $\xi_{\varepsilon,s}$ the median of the distribution, i. e. the value of x such that the function $\tilde{F}(x;\varepsilon,s)$ reaches the 50-th percentile (i.e. $\tilde{F}(\xi_{\varepsilon,s};\varepsilon,s) = \frac{1}{2}$). The median is determined from the scale parameter $\gamma_{\varepsilon,s}$ as:

$$\xi_{\varepsilon,s} = \gamma_{\varepsilon,s}\, ln(4). \tag{4.3}$$

By using the empirical law found, we can derive all the other useful information on the considered bibliometric parameters such as, e.g., the value of x that reaches the P-th percentile of the specific SSD, for any value of P.

The whole set of available data (around 2,500 observations grouped into around 125 SSDs) is fitted to Eq. (4.1) using a Levemberg-Marquad least square fitting routine (Press et al., 2007).

In Fig. 6 we report some selected results of the fit to show the ability of the empirical relation (4.1) to represent the data.[11]

[10] A similar distribution was found in van Raan (2001).

[11] More detailed results are not showed to save space but are available from the authors upon request.

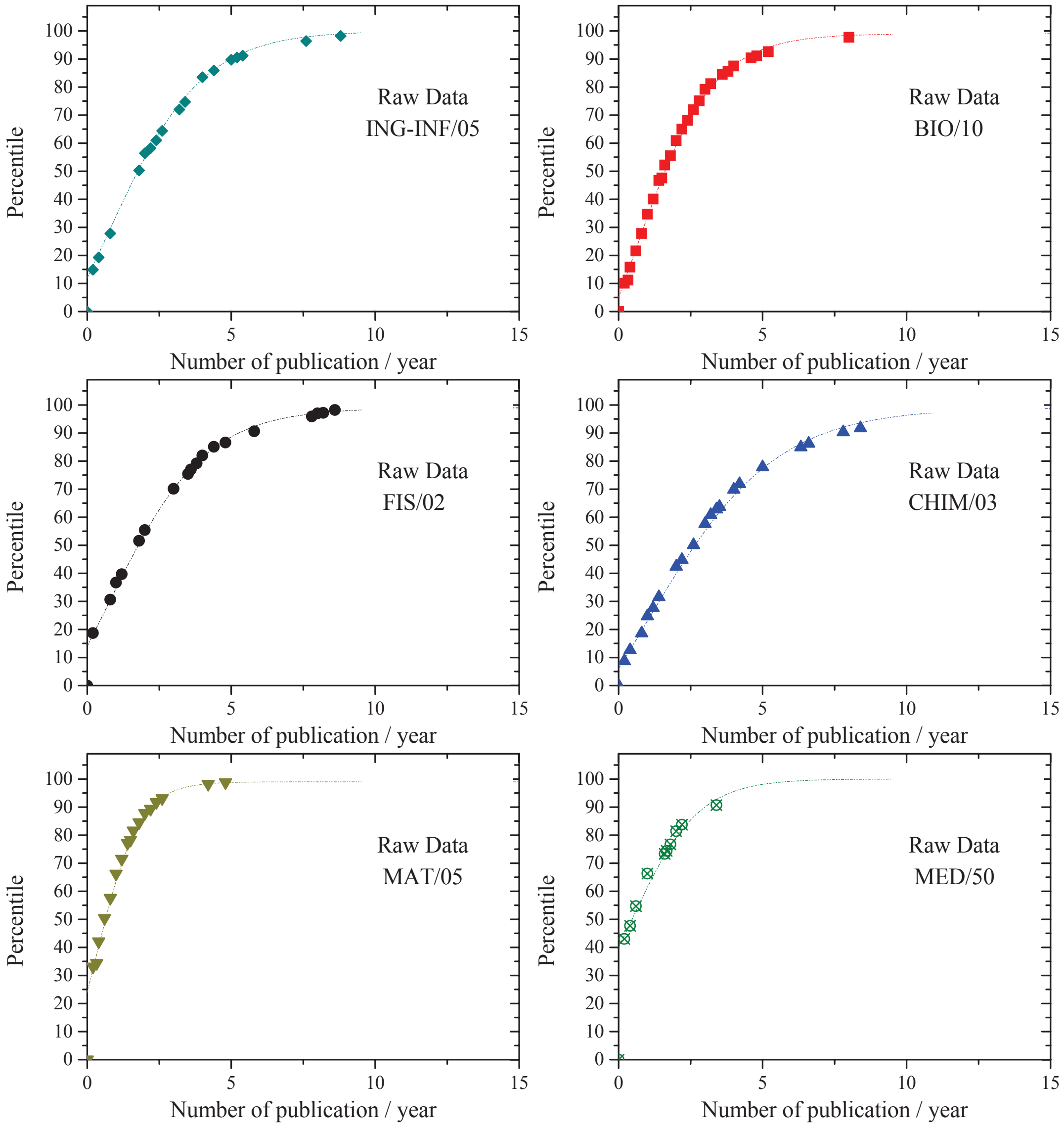

Figure 6: *Distributions of PUB for selected SSDs.*

In the next section we present the main empirical results.

5 Empirical results

Table 1 reports the median values of the bibliometric parameters (PUB, IF, CIT) for the selected SSDs illustrated in previous pictures. The table with all results is reported in Appendix A.

It is interesting to note that the ξ's parameters are useful to quantitatively compare different SSDs. As an example, from Table 1 one can see that a mathematician belonging to the SSD MAT05 need to publish 0.87 paper per year to reach the 50% of its distribution, while a physicist (FIS02) to have the same level of "productivity" should publish 2.16 papers per year. In other words, one can state that a mathematician with 5 publications per year is more productive (close to the 100^{th} percentile, indeed $\xi_{\varepsilon=P,s=\text{MAT05}} = 0.87$ hence $\gamma_{P,s=\text{MAT05}} = 0.63$, and $\tilde{F}(5, P, \text{MAT05}) \approx 1$) than a physicist with the same number of publications (92.5^{th} percentile, $\xi_{\varepsilon=P,s=\text{FIS02}} = 2.16$ hence $\gamma_{P,s=\text{FIS02}} = 1.56$, $\tilde{F}(5, P, \text{FIS02}) = 0.925$).

SSD	Definition	ξ_P	ξ_{IF}	ξ_C
BIO10	Biochemistry	1.57	2.35	1.62
FIS02	Theoretical Physics, Math. Mod. and Methods	2.16	3.98	2.73
CHIM03	General and inorganic chemistry	2.48	4.82	2.25
MAT05	Mathematical analysis	0.87	1.17	0.85
ING-INF05	Information processing systems	2.01	1.03	1.02
MED09	Internal medicine	1.97	2.85	1.95
MED50	Applied medical techniques	1.08	1.57	1.19

Table 1: *Selected results. ξ_P is the number of publications/year to be on the 50-th percentile of the specific SSD; ξ_{IF} is the average impact factor of the journals (normalized at the subject category level) to be on the 50-th percentile of the specific SSD; ξ_C is the average number of citations of a scholar's publications normalized on the Italian median.*

6 Correlations between publications and impact indicators

In the literature, the relationships between number of citations and number of publications across research fields, institutes and countries have been investigated. The production of the scientific community is characterized by cumulative advantages, known as the *Matthew effect* (Merton, 1968; Price, 1976). This specific feature of the scientific production implies that there is a non-linear increase of impact (no. of citations) with increasing size (no. of publications), demonstrated by the finding that the number of citations as a function of number of publications (assessed at sub-fields of science by Kats 1999, 2000) exhibits a power law dependence with an exponent larger than one. van Raan (2008) confirmed previous results at the level of research group, for which he found that citations increase in a power law relationship with the size (no. of publications) of the groups, and a Matthew (cumulative advantage) effect is also found at the group level. In particular, distinguishing in low-performance and top-performance groups it was found that mainly the lower performance groups have a size-dependent cumulative advantage for receiving citations, meaning that the number of citations "scales" in a disproportional non-linear way, according to a power law, with the size of the group in terms of number of publications.[12] Further, Costas et al. (2010) confirmed that these scaling rules apply also at the individual level. In particular they find that the number of citations received by scientists increases in a cumulatively advantageous way as a function of size (number of publications) for researchers in three areas: Natural resources, Biology and Biomedicine and Materials Science.

In this section we analyze the correlations between the bibliometric parameters ξ_{IF}, ξ_C and ξ_P estimated in Section 5. Each ball in the following figures represents an SSD.

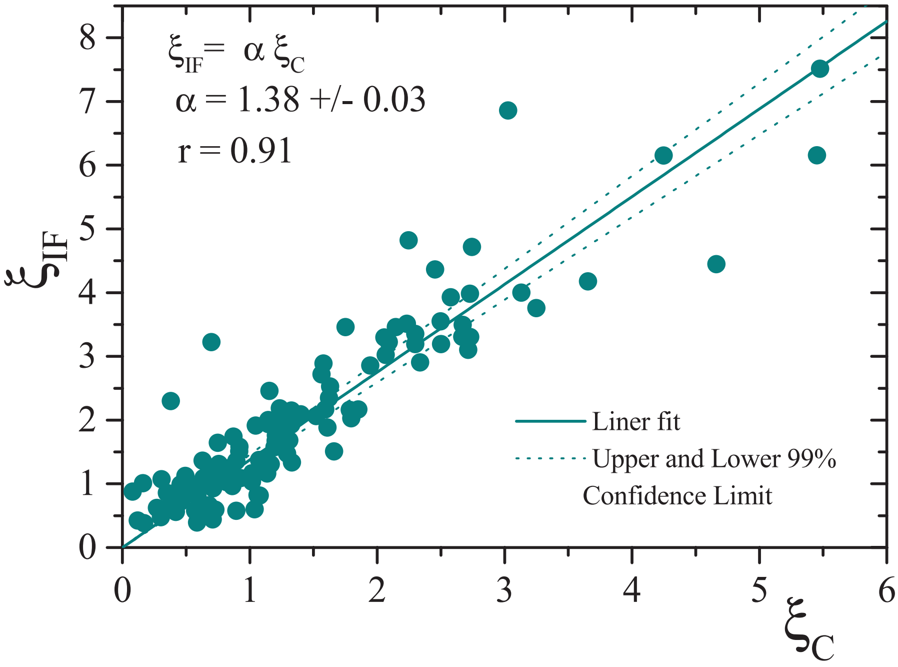

Figure 7: *Correlation between ξ_{IF} and ξ_C.*

Indeed, Figure 7 confirms that ξ_{IF} and ξ_C are highly correlated (linear correlation, with a slope $\alpha = \xi_{IF}/\xi_C$ =1.38 and a correlation r higher than 0.90). This strong correlation supports the choice of the Italian National Agency ANVUR to consider only one indicator between citations and IF among the relevant bibliometric parameters.

[12] For this reason the literature refers to "scaling" relationships to describe the correlations between number of citations and number of publications. Of course the meaning of "scaling" in this context is completely different from the scaling approach described in previous sections and used to search for a master curve (and its related scaling factors) to compare heterogeneous academic fields.

Figure 8 shows that the scatterplot of ξ_{IF} versus ξ_P is very similar to the one of ξ_C versus ξ_P.

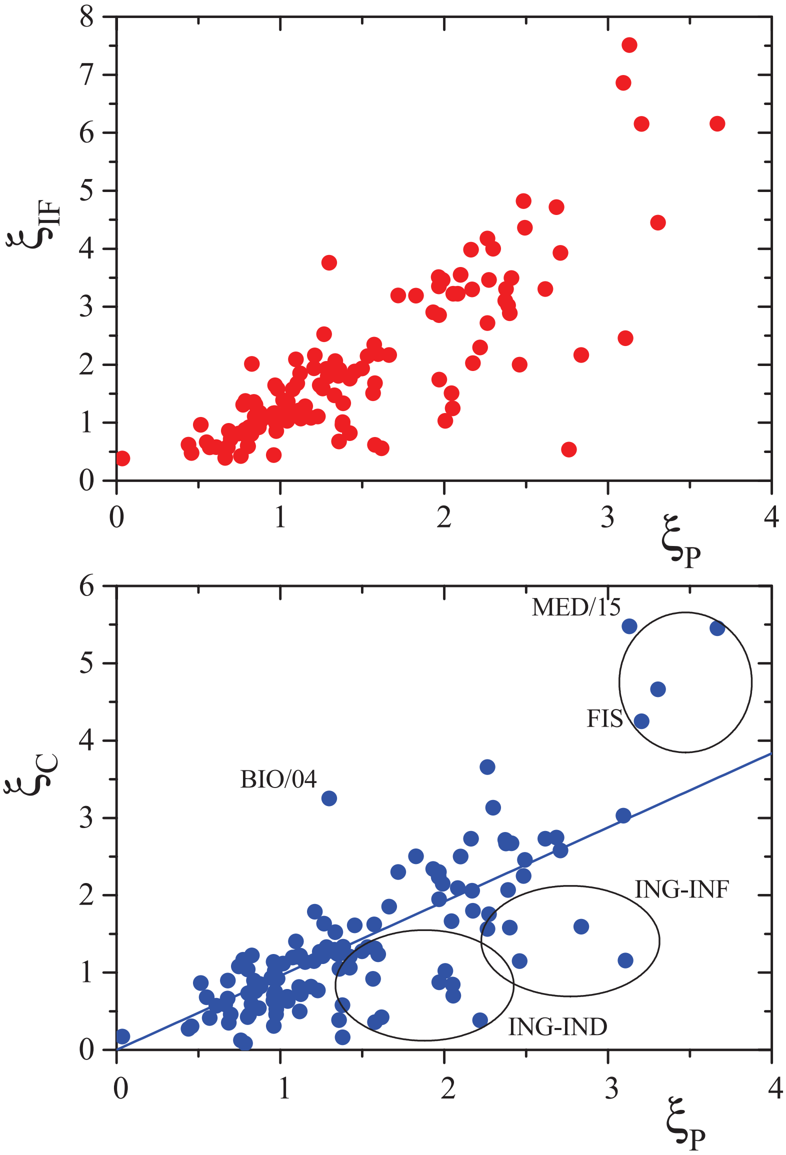

Figure 8: *Scatterplots of* ξ_{IF} *(top panel) and of* ξ_C *(bottom panel) versus* ξ_P*.*

Fig. 8, bottom panel illustrates clearly that publication and citation practices greatly differ among heterogeneous SSDs. While for a large part of Academic disciplines there is a linear relation between ξ_C and ξ_P, this is not the case for Physics' SSDs (indicated as FIS in Fig. 8, bottom panel) for instance, that have higher median values of CIT and PUB, compared to other SSDs such as the Computer Engineering and Industrial Engineering' SSDs (indicated respectively as ING-INF and ING-IND in Fig. 8 bottom panel) that show lower median values for CIT and PUB.

From the empirical evidence showed in this section, we might conclude that both number of publications and one between citations and impact factor could be considered as reasonable indicators for a bibliometric evaluation process.

7 Towards a general (non-empirical) distribution law

Once we have determined the scaling factors $\xi_{\varepsilon,s}$ for the different SSDs from the study of their cumulative distributions, we can in principle calculate for each scholar the "scaled indicators", that is (for example in the case of ε=P) the number of yearly publication divided by the appropriate scaling factor just determined for the researcher's SSD. Given the existence of a master curve, these scaled indicators should derive, for all Sapienza's scholars, from a single distribution. We gain, therefore, a large statistical basis for the study of the shape of the distribution.

In Fig. 9 we show the histogram of such data, the whole number of observations is equal to the number of scholars belonging to those SSDs with a sufficient number of observations, i.e. where it has been possible to determine the scaling parameters (≈2,400 observations). The histogram appears to be smooth enough to allow for a detailed shape analysis.

The solid curve reported in Fig. 9 (Top panel) represents the best fit obtained with a Log-normal distribution[13] with the following law:

$$g(x;\varepsilon) = \frac{1}{\sqrt{2\pi}\sigma_\varepsilon x} \ \exp\big\{-\frac{1}{2\sigma_\varepsilon^2}\big[ln(x/\xi_\varepsilon)\big]^2\big\} \tag{7.4}$$

A similar fit performed with the derivative of the Boltzmann distribution of Eq. (4.2) (dashed line in Fig. 9 -Top panel-) gives a worst agreement (normalized χ^2=2.2 in the case of the Boltzmann distribution and χ^2=1.1 for the Log Normal distribution although, in the latter case, there is one more parameter). Being the parameter ξ_ε in Eq. (7.4) equal to the median of the distribution, from the fit it turns out to be consistent with ξ_ε=1 ($\xi_P = 0.97 \pm 0.05$), while the parameter $\sigma_P = 0.43 \pm 0.03$.

To better emphasize the ability of the Log normal distribution to describe the empirical data, in Fig. 9 Bottom panel we report the same data as in Fig. 9 Top panel but plotted and binned as a function of $\log((x/\xi_\varepsilon))$, which -according to Eq. (7.4)- implies a gaussian shape for the histogram.

We notice that it is not our aim here to analyze the origin of the observed data distribution neither to validate previous empirical evidence (e.g. Radicchi et al. 2008; Evans et al. 2012; Waltman et al. 2012) on the existence of universality of bibliometric indicators.

[13] For a general overview on Log-normal distributions, see Limpert et al. (2001).

We only show that scaling phenomena exist and scaling factors may be estimated also to compare average yearly bibliometric indicators calculated at the academic disciplinary level. Moreover, we suggest to look for them validating empirically their existence. Finally, we emphasize that the empirically–based validation of the existence of a master curve (with its related scale factors) is a fundamental step, as far as a research assessment usage of the scaling factors is envisaged.

Nevertheless, the Log-normal distribution is ubiquitously observed in many human-decision driven phenomena (from stock price in economics to city sizes in sociology, and many others). It is not surprising therefore that it describes also the bibliometric indicators of scientific production analysed in this paper. Publication strategy is in fact a mix of individual decision and group attitude.

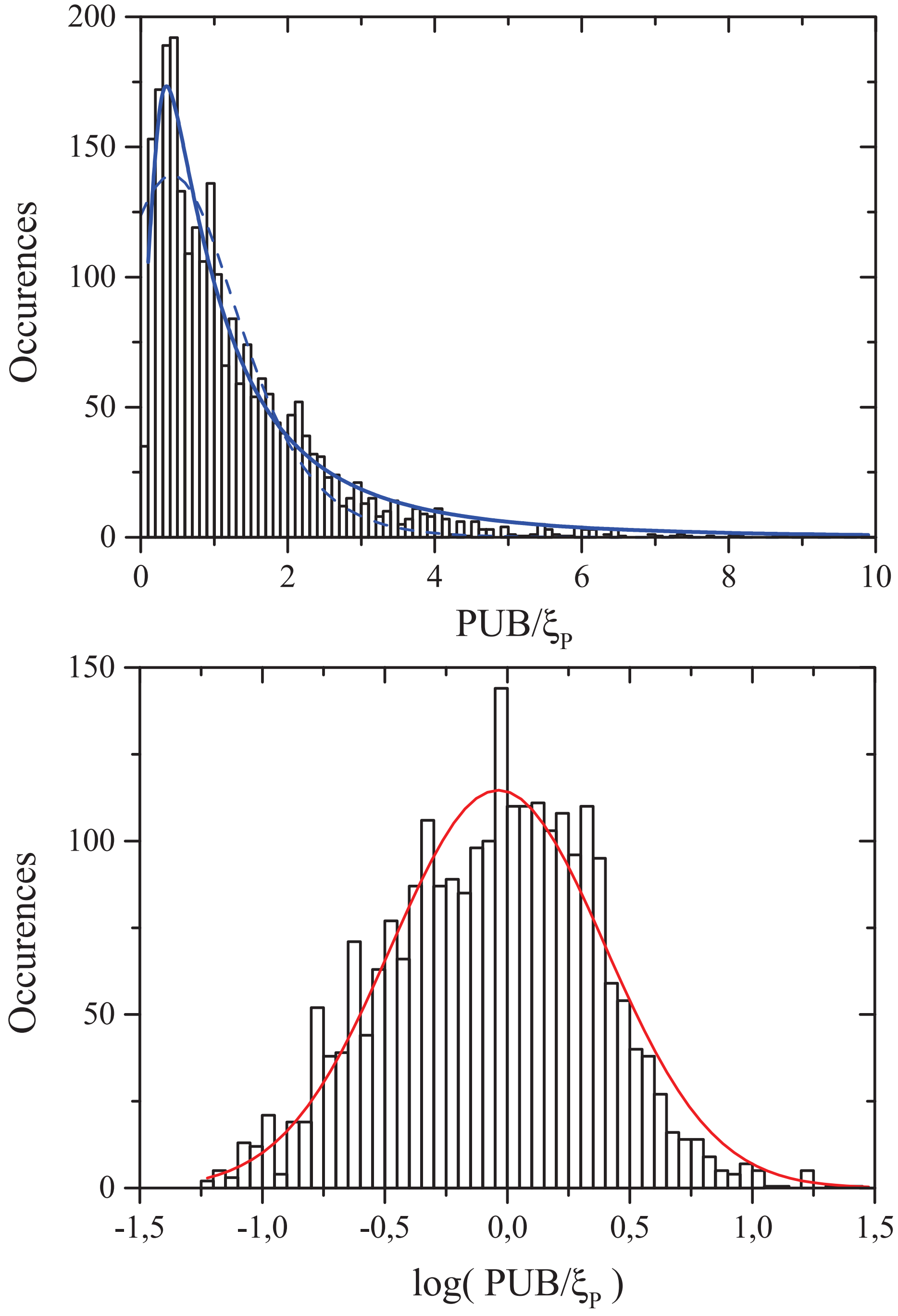

Figure 9: ***Top panel***: *Histogram of the distribution of the indicator PUB/ξ_P for all Sapienza's scholars. The solid curve is the fitted Log-normal distribution. Also reported as dashed line, for sake of comparison, is the fit to the derivative of the Boltzmann function.* ***Bottom panel***: *Histogram of the distribution of the indicator PUB/ξ_P for all Sapienza's scholars plotted and binned as a function of $log((x/\xi_\varepsilon))$. The solid curve is the fitted Log-normal distribution which, in this scale, appears as a gaussian distribution.*

8 Potential limitation of the analysis

A potential limitation of our analysis concerns the representativeness of Sapienza data for the estimation of the Italian SSDs analysed. The university of Rome La Sapienza is the biggest university in Europe (without considering long-distance learning universities) and is among the oldest ones. It accounts for around 7% of the total Italian academic staff. Given the large number of scholars considered in the analysis we consider that its representativeness is reasonable. In addition, in the elaborations we used the information related to the rank of Sapienza scholars in the Italian university system: it appears that Sapienza is well representative of the Italian distributions (see e.g. Figure 6 which shows the distributions of Sapienza scholars in the Italian national percentile, by SSD). Finally, IF, the impact factor indicator considered, is field normalized at international level (on all the median values of the journals by subject category). For all these reasons we consider the Sapienza sample as fully representative of the Italian university system.

9 Conclusions and further developments

We provide evidence that the distributions of the yearly average number of publications (PUB), citations (CIT) and impact factor (IF) of Italian bibliometric academic disciplines only differ by a scale factor, and after an appropriate normalization of the data - based on their main sources of heterogeneity, namely percentage of silent scholars and different skewness- could be rescaled, i.e. collapsed on one common empirical law that follows a Log-normal distribution. We show the usefulness of the obtained results in terms of research assessment. Interestingly, this approach is currently employed by the university of Rome "La Sapienza", within a complex system of performance evaluation, to allocate resources at departments and schools.

We estimate the empirical law by using data on 2471 scholars of the university of Rome "La Sapienza" which represents around 7% of the Italian academic staff. It could be worth to investigate if the findings of our analysis are confirmed by enlarging the sample. This further investigation would allow us to provide estimates of the scale factors for a larger number of academic disciplines.

It could be also interesting to investigate on the generative mechanism of the empirical law found for PUB, CIT and IF. We put forward a conjecture: it could be the convolution of two decreasing distributions: scientific productivity and age.

Another interesting extension of the analysis would be to move from national to international comparisons.

References

[1] Albarran P., Crespo J.A., Ignacio O., Ruiz-Castillo J., (2011), The skewness of science in 219 sub-fields and a number of aggregates, *Scientometrics*, 88, 385-397.

[2] Bak P., Tang C., and Wiesenfeld K. (1987), Self-Organized Criticality: An Explanation of $1/f$ Noise, *Physical Review Letters*, 59, 4, 381–384.

[3] Bonaccorsi A., Daraio C. (2007a), Theoretical perspectives on university strategy, in Bonaccorsi A. and Daraio C., edited by, *Universities and Strategic Knowledge Creation. Specialization and Performance in Europe*, Edward Elgar Publisher, Cheltenham (UK), pp. 3–30.

[4] Bonaccorsi A. Daraio C. (2007b), Universities as strategic knowledge creators. Some preliminary evidence, in Bonaccorsi A. and Daraio C., edited by, *Universities and Strategic Knowledge Creation. Specialization and Performance in Europe*, Edward Elgar Publisher, Cheltenham (UK), pp. 31–84.

[5] Costas, R., van Leeuwen, T., Bordons, M. (2010), A bibliometric classificatory approach for the study and assessment of research performance at the individual level: The effects of age on productivity and impact, *Journal of the American Society for Information Science and Technology*, 61(8), 1564-1581.

[6] D'Angelo, C. A., Giuffrida, C., Abramo, G. (2010), A heuristic approach to author name disambiguation in large-scale bibliometric databases, *Journal of the American Society for Information Science and Technology*, 62(2), 257-269.

[7] Daraio, C., et al., (2011), The European University landscape: A micro characterization based on evidence from the Aquameth project, *Research Policy*, 40, 148–164.

[8] Daraio, C., Moed, H.F., (2011), Is Italian science declining?, *Research Policy*, 40 (10), 1380–1392.

[9] Deiaco, E., Holmén, M. and McKelvey, M. (2009), What does it mean conceptually that universities compete? pp. 300-328, in McKelvey, M. and Holmén, M. (eds), *Learning to Compete in European Universities: From Social Institution to Knowledge Business*, Edward Elgar, Cheltenham (UK).

[10] Deiaco E., Hughes A., McKelvey M. (2012), Universities as strategic actors in the knowledge economy, *Cambridge Journal of Economics* 36, 525-541.

[11] Egghe L., Rousseau R. (1990) *Introduction to Informetrics. Quantitative Methods in Library, Documentation and Information Science*, Elsevier, Amsterdam.

[12] Evans T.S., Hopkins N., Kaube B.S. (2012), Universality of Performance Indicators based on Citation and Reference Counts, submitted to *Scientometrics*, arXiv:1110.3271v2 [physics.soc-ph].

[13] Glanzel, W. (2010), The role of the h-index and the characteristic scores and scales in testing the tail properties of scientometric distributions, *Scientometrics*, 83, 697-709.

[14] Katz, J.S. (1999), The Self-Similar Science System, *Research Policy*, 28, 501–517.

[15] Katz, J.S. (2000), Scale Independent Indicators and Research Assessment, *Science and Public Policy*, 27, 1, 23–36.

[16] Knorr-Cetina, K. (1995), Laboratory studies: the cultural approach to the study of science, in S. Jasanoff, G.E. Markle, J.C. Petersen and T. Pinch (eds), *Handbook of Science and Technology Studies*, London: Sage, pp. 140-66.

[17] Laherrre, J. and D. Sornette (1998), Stretched exponential distributions in nature and economy: "fat tails' with characteristic scales. *Eur. Phys. J. B*, 2, 525–539.

[18] Langton, C.G. (1990), Computation at the edge of chaos: Phase transitions and emergent computation *Physica D*, 42, 1-3, 12-37.

[19] Laredo, P. and P. Mustar (eds) (2001), *Research and Innovation Policies in the New Global Economy: An International Comparative Analysis*, Cheltenham, UK and Northampton, MA, Edward Elgar (USA).

[20] LERU (2012), *Research universities and research assessment*, Position paper, May 2012.

[21] Limpert E., Stahel W.A., and Abbt M. (2001), Log-normal distributions across the sciences: keys and clues, *BioScience*, 51 (5), 341–352.

[22] Lotka, A.J. (1926), The frequency distribution of scientific productivity. J. Washington Acad. Sci., 16, 317–323.

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

[24] Naranan, S. (1971), Power law relations in science bibliography a self consistent interpretation, *Journal of Documentation*, 27, 83-97.

[25] Newman, M.E.J., (2005), Power laws, Pareto distributions and Zipfs law, *Contemporary Physics*, 46, 323-351.

[26] Press, W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P.,(2007), *Numerical Recipes 3rd Edition: The Art of Scientific Computing*, Cambridge University Press, NY.

[27] Price, D. de Solla, (1976), Ageneral theory of bibliometric and other cumulative advantage processes, *Journal of the American Society for Information Science*, 27(5), 292-306.

[28] Radicchi, F., Fortunato, S., Castellano, C. (2008), Universality of citation distributions: Toward an objective measure of scientific impact, *Proceedings of the National Academy of Sciences of the United States of America*, 105, 17268-17272.

[29] Redner, S. (1998), How popular is your paper? An empirical study of the citation distribution, *European Physical Journal B*, 4, 131-134.

[30] Seglen, P. (1992), The skewness of science, *Journal of the American Society for Information Science*, 43, 628-638.

[31] Simon H. (1955), On a class of skew distribution functions, *Biometrika*, 42, 425–440.

[32] Stock W.G., (2006), On Relevance Distributions, *Journal of the American Society for Information Science and Technology*, 57, 8, 1126-1129.

[33] Van den Berghe, H., Houben, J. A., de Bruin, R. E., Moed, H. F., Kint, A., Luwel M., Spruyt, E. H. J., (1998), Bibliometric indicators of university research performance in Flanders, *Journal of the American Society for Information Science*, 49(1), 59-67.

[34] van Raan A.F.J. (2001) Competition amongst scientists for publication status: toward a model of scientific publication and citation distributions, *Scientometrics*, 51, 347-357.

[35] van Raan A.F.J. (2004), Measuring science, in H.F. Moed, W. Glanzel and U. Schmoch (edited by), *Handbook of Quantitative Science and Technology Research*, Kluwer Academic Publishers, pp. 19–50.

[36] van Raan, A.F.J. (2006), Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment of 147 chemistry research groups, *Scientometrics*, 67(3), 491–502.

[37] van Raan, A.F.J. (2008), Scaling rules in the science system: Influence of field-specific citation characteristics on the impact of research groups, *Journal of the American Society for Information Science and Technology*, 59(4), 565–576.

[38] Waltman L., van Eck N.J., van Raan A.F.J. (2012), Universality of Citation Distributions Revisited, *Journal of the American Society for Information Science and Technology*, 63(1), 72–77.

A Appendix: Detailed results

In this section we present the detailed results obtained for all the SSDs with at least 10 observations, grouped by disciplinary area. ξ_P is the number of publications/year to be on the 50-th percentile of the specific SSD; ξ_{IF} is the average impact factor of the journals (normalized at the subject category level) to be on the 50-th percentile of the specific SSD; ξ_C is the average number of citations of a scholar's publications normalized on the Italian median.

SSD	Definition	ξ_P	ξ_{IF}	ξ_C	No. obs.
MAT03	Geometry	0.55	0.66	0.68	22
MAT05	Mathematical analysis	0.87	1.17	0.85	57
MAT06	Probability and statistics	0.98	0.86	0.52	16
MAT07	Mathematical physics	1.19	1.08	0.82	27
MAT08	Numerical analysis	1.04	1.21	0.68	11
MAT09	Operational research	1.13	1.07	0.72	16
INF01	Informatics	1.42	0.82	1.06	46

Table 2: Results: Area 01 - Mathematics and informatics.

SSD	Definition	ξ_P	ξ_{IF}	ξ_C	No. obs.
FIS01	Experimental Physics	3.67	6.15	5.45	73
FIS02	Theor. phys. ,math. models and methods	2.16	3.98	2.73	30
FIS03	Physics of Matter	3.21	6.15	4.25	22
FIS05	Astronomy and astrophysics	3.09	6.86	3.03	18

Table 3: Results: Area 02 - Physics.

SSD	Definition	ξ_P	ξ_{IF}	ξ_C	No. obs.
CHIM01	Analytical chemistry	1.97	3.35	2.30	27
CHIM02	Physical chemistry	2.69	4.72	2.74	37
CHIM03	General and inorg. chemistry	2.48	4.82	2.25	37
CHIM06	Organic chemistry	2.49	4.36	2.45	37
CHIM08	Pharmaceutical chemistry	2.17	3.30	2.06	28
CHIM09	Pharm. and technol. applications of chem.	2.26	2.72	1.56	10

Table 4: Results: Area 03 - Chemistry.

SSD	Definition	ξ_P	ξ_{IF}	ξ_C	No. obs.
GEO01	Paleontology and Paleoecology	0.88	1.02	0.82	11
GEO02	Stratigraphic and sedim. geology	0.70	0.72	0.46	13
GEO04	Physical geogr. and geomorphology	0.61	0.58	0.57	11
GEO08	Geochemistry and volcanology	1.34	2.06	1.52	11

Table 5: Results: Area 04 - Earth sciences.

SSD	Definition	ξ_P	ξ_{IF}	ξ_C	No. obs.
BIO02	Systematic botany	0.66	0.39	0.59	12
BIO03	Environ. and applied botany	0.80	0.75	0.42	10
BIO05	Zoology	1.26	1.59	1.21	18
BIO06	Comparative anatomy and cytology	1.12	1.17	0.76	14
BIO07	Ecology	1.10	1.67	1.20	14
BIO09	Physiology	1.35	1.90	1.24	28
BIO10	Biochemistry	1.57	2.35	1.62	63
BIO11	Molecular biology	1.21	2.16	1.79	24
BIO12	Clinical bioch. and molecular bio.	0.83	2.01	1.22	12
BIO13	Experimental biology	1.21	1.94	1.15	20
BIO14	Pharmacology	1.93	2.90	2.34	31
BIO16	Human anatomy	1.50	1.93	1.27	25
BIO17	Histology	1.28	1.93	1.33	19
BIO18	Genetics	1.10	2.09	1.40	19

Table 6: Results: Area 05 - Biology.

SSD	Definition	ξ_P	ξ_{IF}	ξ_C	No. obs.
MED03	Medical genetics	2.26	4.18	3.66	12
MED04	Experimental medicine and pathophys.	1.72	3.19	2.30	63
MED05	Clinical pathology	1.12	1.85	1.22	26
MED06	Medical oncology	1.97	3.51	2.23	12
MED07	Microbiology and clinical microbiology	1.42	1.76	1.21	47
MED08	Pathology	2.38	3.30	2.67	33
MED09	Internal medicine	1.97	2.85	1.95	103
MED11	Cardiovascular diseases	1.99	3.46	2.15	42
MED12	Gastroenterology	2.10	3.55	2.50	27
MED13	Endocrinology	2.37	3.10	2.71	32
MED15	Blood diseases	3.13	7.51	5.48	19
MED17	Infectious diseases	1.53	2.15	1.33	17
MED18	General surgery	0.82	0.93	0.71	163
MED19	Plastic surgery	0.81	0.92	0.45	10
MED22	Vascular surgery	0.77	1.31	1.16	13
MED24	Urology	1.24	1.64	1.27	24
MED25	Psychiatry	1.58	1.68	1.31	16
MED26	Neurology	2.41	3.49	2.67	48
MED27	Neurosurgery	1.57	1.51	0.92	13
MED28	Oral diseases and dentistry	0.80	0.60	1.039	13
MED30	Eye diseases	0.75	0.81	1.08	21
MED31	Otorhinolaryngology	1.04	1.03	0.69	18
MED32	Audiology	0.96	0.44	0.71	10
MED33	Musculoskeletal system diseases	0.78	0.88	0.08	10
MED35	Dermatological and venerological diseases	1.05	1.36	0.63	11
MED36	Diagnostic imaging and radiotherapy	1.45	1.88	1.61	36
MED37	Neuroradiology	1.66	2.16	1.85	11
MED38	General and subspecialty paediatrics	1.29	1.79	1.27	56
MED39	Child neuropsychiatry	1.02	1.39	1.12	21
MED40	Obstetrics and gynaecology	1.15	1.28	1.14	49
MED41	Anaesthesiology	0.96	1.16	1.14	22
MED42	Hygiene and public health	1.11	1.21	0.81	15
MED46	Medical and biotechnology laboratory techniques	0.79	1.37	1.07	20
MED50	Applied medical techniques	1.08	1.57	1.19	17

Table 7: Results: Area 06 - Medicine.

SSD	Definition	ξ_P	ξ_{IF}	ξ_C	No. obs.
ING-IND04	Aerospace structures and design	0.87	0.92	0.54	11
ING-IND06	Fluid dynamics	0.85	1.31	0.76	14
ING-IND22	Materials science and technology	1.36	1.91	1.05	12
ING-IND25	Chemical plants	1.35	1.81	1.24	12
ING-IND31	Electrical engineering	2.05	1.25	0.84	16
ING-IND33	Electrical power systems	1.58	0.62	0.36	11
ING-INF01	Electronics	3.11	2.46	1.15	20
ING-INF02	Electromagnetic fields	2.40	2.89	1.58	14
ING-INF03	Telecommunications	2.84	2.17	1.60	19
ING-INF04	Systems and control engineering	2.46	2.00	1.15	17
ING-INF05	Information processing systems	2.01	1.03	1.02	27

Table 8: Results: Area 09 - Industrial and information engineering.

SSD	Definition	ξ_P	ξ_{IF}	ξ_C	No. obs.
M-PSI02	Psychobiology and physiological psy.	1.83	3.19	2.50	17
M-PSI03	Psychometrics	0.44	0.62	0.27	11

Table 9: Area 11 - History, philosophy, pedagogy and psychology.