POLITECNICO DI TORINO
Repository ISTITUZIONALE

Regularity in the research output of individual scientists: An empirical analysis by recent bibliometric tools

Original
Regularity in the research output of individual scientists: An empirical analysis by recent bibliometric tools / Franceschini, Fiorenzo; Maisano, DOMENICO AUGUSTO FRANCESCO. - In: JOURNAL OF INFORMETRICS. - ISSN 1751-1577. - 5, n.3:(2011), pp. 458-468. [10.1016/j.joi.2011.04.004]

Availability:
This version is available at: 11583/2424000 since:

Publisher:
Elsevier Ltd.

Published
DOI:10.1016/j.joi.2011.04.004

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

06 May 2026

Regularity in the research output of individual scientists: An empirical analysis by recent bibliometric tools

Fiorenzo Franceschini*, Domenico Maisano

Politecnico di Torino, DISPEA (Department of Production Systems and Business Economics), Corso Duca degli Abruzzi 24, 10129 Torino, Italy

Keywords:
Research evaluation
Individual scientist
Publication regularity
Citation regularity
Ferrers diagram
h-Index
Citation/publication distribution
Borda's method

A B S T R A C T

This paper proposes an empirical analysis of several scientists based on their time regularity, defined as the ability of generating an active and stable research output over time, in terms of both quantity/publications and impact/citations. In particular, we empirically analyse three recent bibliometric tools to perform qualitative/quantitative evaluations under the new perspective of regularity. These tools are respectively (1) the *PY*/*CY* diagram, (2) the publication/citation Ferrers diagram and triad indicators, and (3) a year-by-year comparison of the scientists' output (Borda's ranking). Results of the regularity analysis are then compared with those obtained under the classical perspective of overall production.

The proposed evaluation tools can be applied to competitive examinations for research position/promotion, as complementary instruments to the commonly adopted bibliometric techniques.

1. Introduction

Evaluating the scientific output of individual scientists is an important issue, with significant consequences for their promotion, tenure, faculty positions, research grants, etc. (Van Raan, 2000). Two of the most important aspects usually taken into account are: overall production – generally measured in terms of publications – and overall impact – generally measured in terms of citations received by each publication (Bornmann, 2011; Cronin, 1984; Franceschini, Galetto, & Maisano, 2007). These two quantities can be aggregated into other indicators, such as the *h*-index, *g*-index or others (Hirsch, 2005). For more on the *h*-index, *g*-index and the large number of variants and improvements, we refer the reader to the vast literature and extensive reviews (Alonso, Cabrerizo, Herrera-Viedma, & Herrera, 2009; Egghe, 2010a; Rousseau, 2008; Franceschini & Maisano, 2010a).

Besides the traditional evaluation of research output in its entirety, an aspect that is sometimes invoked in the rules of competitive examinations for research position/promotion is time regularity of one scientist's output (ASPHER, 2010; Collegio dei presidenti di corso di studi in Matematica, 2008; IPEA, 2009). The basic idea is that a scientist who not only performs well compared to his/her peers, but also is able to "spread" his/her scientific output over time, should be preferred to another scientist, with equivalent overall scientific output, but not homogeneously distributed and with significant fluctuations. In this sense, regularity is likely to denote a more persistent commitment to research. In a recent "conceptual" paper, regularity has been defined as the ability of generating an active and stable research output over time (Franceschini &

* Corresponding author. Tel.: +39 011 5674 7225; fax: +39 011 5674 7299.
E-mail addresses: fiorenzo.franceschini@polito.it (F. Franceschini), domenico.maisano@polito.it (D. Maisano).

Maisano, in press-a). Active means that the output has to be substantial and stable means that the scientist should "spread" his/her research output over time. We are fully aware that this point can be controversial.

Ideally, researchers should publish when they have something of interest or new to say, no matter when. It is also true that it would seem better that publications occurred as frequently as possible, avoiding, however, manipulative behavior, like publishing the research results "in installments". Also, one can think of many cases that could produce significant variations of outputs over time, although not necessarily linked to the quality of research or researchers (e.g., fluctuations in external funding, fluctuations of teaching loads, interruptions of careers for women having children or other reasons). Besides, regularity is quite relevant, given the fact that public funds of many university departments and research institutions are allocated annually, depending on the published output of the year(s) ahead.

Having said that, the debate on the validity of regularity analysis remains open and is somehow related to that on the validity of bibliometric analysis in general for evaluating individual scientists (Haeffner-Cavaillon & Graillot-Gak, 2009; Snizek, 1995). Nevertheless, it does not seems absurd to reassert that the ability to do significant research with some persistence over time is something desirable from the point of view of the research institution. Thus, we think that an analysis based on regularity can be useful, at least as a complement to the classical bibliometric techniques for evaluating research performance (Moed, 2005).

The purpose of this paper is a structured comparison of 24 Italian researchers in the same discipline and with similar career lengths (i.e., around 15 years), on the basis of the time regularity of their scientific output. Input data for this evaluation are the temporal distributions of the publications and/or corresponding citations. Analysis is carried out by three major bibliometric tools introduced and described in detail in Franceschini and Maisano (in press-a): (1) the *PY*/*CY* diagram; (2) the Ferrers diagrams and triad indicators; (3) a procedure based on a year-by-year comparison of scientists, according to their publication and citation temporal distributions (Borda's ranking).

The remaining of this paper is organised into four sections. Section 2 illustrates the methodology used for data collection. Section 3 comments results obtained by applying each of the three above mentioned tools, underlining their advantages and limitations. Section 4 compares the results obtained from the evaluation of scientists' output in terms of regularity, with those obtained under the classical perspective of overall production. Finally, the conclusions are given, summarising the original contribution of the paper.

2. Data collection

This study concerns the analysis of time regularity relating to the publications/citations of some Italian academic scientists involved in the scientific sector of Production Technology and Manufacturing Systems. This specific sector is the authors' primary research field; this fact makes it easier to select a homogeneous sample of researchers, as well as verify the accuracy of the scientists' publication statistics. In particular, we have selected 24 scientists with the same academic position – i.e., associate professor – who, realistically, can be seen as potential "competitors" for promotion to a full professor position. As expectable, these scientists have similar career lengths (defined as the number of years from the first publication to date) of about 12–15 years.

For each scientist, input data for the analysis are given by publications and the corresponding citations, using a 15 year time window (from 1996 to 2010), which embraces the whole production of all the 24 scientists. These data are used to construct the *PY* distribution – i.e., the yearly distribution of total publications according to the age – and the *CY* distribution – i.e., the yearly distribution of the citations accumulated up to the moment of the analysis, by the publications issued in one year. The choice of using yearly time-buckets derives from the need for a reasonable analysis resolution and simplicity in the subdivision of publications depending on their publication date.

Data have been collected using the Google Scholar search engine for two main reasons: (1) despite the lower accuracy, Google Scholar's coverage is superior to that of Web of Science and Scopus databases in many fields such as Social Sciences, Computer Science or Engineering Science; (2) it can be automatically queried through dedicated software applications, such as Publish or Perish or other *ad hoc* applications (Harzing & van der Wal, 2008). The point on the Google Scholar's coverage and (low) accuracy is very subtle and deserves more attention. We are aware that, as a rule for a generic bibliometric study, it would be better to limit document types to journal publications – specifically, articles, notes, letters and reviews. However, regarding the scientists of interest, a significant portion of their publication contributions consists of conference papers, book chapters, monographs, and even articles published in Open-Access journals, national journals or other journals not indexed by Web of Science and Scopus (Franceschini & Maisano, 2011). The fact that these two databases constantly widen their portfolio of indexed journals and conference proceedings is emblematic. Google Scholar database probably provides a more comprehensive picture of recent impact, even if it should improve significantly before it becomes fully operational. In the recent literature, there are several studies on the reasons for the low quality of Google Scholar, as well as comparisons with other bibliometric databases; for example (Bar-Ilan, 2010; Bornmann et al., 2009; Franceschini & Maisano, in press-b; Labbé, 2010; Neuhaus & Daniel, 2008). To limit the effect of database inaccuracies, data have been manually checked and cleaned. This way, we eliminated the most common database mistakes – e.g., false references, duplicate records, author ambiguities, etc. The resulting dataset is reported in Table 1, where scientists are anonymous for reasons of confidentiality.

It can be noticed that the duration of the publishing activity is not exactly the same for the 24 scientists. In particular, in the years before the first publication, P_y and C_y values of a scientist – i.e., respectively the total number of publications for each year and the corresponding total number of citations, accumulated up to the moment of the analysis – are marked by the

Table 1
Data relating to 24 Italian scientists in the scientific sector of Production Technology and Manufacturing Systems, considering 15 consecutive years. For each scientist the following indicators are reported: h, PY and CY distribution with the corresponding sum (P and C respectively), mean value (μ), median (Med), standard deviation (s), interquartile range (IQR) and an indicator of continuity ($Cont$) given by the percentage of years with $P_y > 0$ and the percentage of years with $C_y > 0$. We precise that C_y values are the citations received by the papers published in one year, accumulated up to the date of the analysis. Scientists are sorted in descending order according to their P values.

Scient.	h		1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	Sum	μ	Med	s	IQR	$Cont$
S1	13	P_y	4	3	4	3	6	1	5	7	6	22	27	45	43	21	22	$P=219$	14.6	6	14.7	18	15/15
		C_y	9	19	23	31	15	0	23	72	15	95	34	140	30	9	2	$C=517$	34.5	23	38.8	20.5	14/15
S2	9	P_y	–	1	0	0	5	1	5	4	3	13	9	8	32	16	6	$P=103$	6.9	5	8.5	7.5	12/14
		C_y	–	0	0	0	11	1	30	56	15	78	56	28	83	21	1	$C=380$	25.3	15	29.4	42.5	11/14
S3	8	P_y	1	1	6	4	5	7	4	6	3	4	6	15	16	14	7	$P=99$	6.6	6	4.7	3	15/15
		C_y	0	0	1	0	13	9	17	11	3	59	25	31	19	13	0	$C=201$	13.4	11	16.0	17.5	11/15
S4	8	P_y	5	3	0	1	6	1	6	8	4	7	5	9	13	9	7	$P=84$	5.6	6	3.5	4	14/15
		C_y	30	93	0	7	83	36	22	12	2	32	7	14	16	14	2	$C=370$	24.7	14	28.0	24	14/15
S5	4	P_y	2	5	2	5	3	0	4	1	7	7	5	6	10	12	7	$P=76$	5.1	5	3.3	4.5	14/15
		C_y	2	14	0	10	0	0	0	0	0	2	0	3	3	5	0	$C=39$	2.6	0	4.2	3	7/15
S6	9	P_y	–	1	0	2	3	4	1	13	11	6	4	3	4	1	1	$P=54$	3.6	3	3.8	3	13/14
		C_y	–	1	0	10	7	3	4	78	52	41	12	2	8	0	0	$C=218$	14.5	4	23.4	10.5	11/14
S7	6	P_y	–	–	–	2	0	2	6	9	6	4	1	2	6	4	6	$P=48$	3.2	2	2.9	5.5	11/12
		C_y	–	–	–	4	0	0	53	18	0	11	2	3	26	14	4	$C=135$	9.0	3	14.5	12.5	9/12
S8	5	P_y	2	6	1	1	3	1	1	1	1	6	9	6	4	3	1	$P=46$	3.1	2	2.6	4	15/15
		C_y	0	23	0	0	10	14	1	0	0	6	0	3	0	0	0	$C=57$	3.8	0	6.8	4.5	6/15
S9	7	P_y	1	3	1	1	3	2	5	4	4	4	1	6	1	3	6	$P=45$	3.0	3	1.8	3	15/15
		C_y	8	27	6	1	28	8	19	26	25	10	1	12	1	12	1	$C=185$	12.3	10	10.2	18.5	15/15
S10	8	P_y	4	8	8	4	0	1	2	2	0	0	0	6	3	3	1	$P=42$	2.8	2	2.8	3.5	11/15
		C_y	44	37	23	62	0	5	0	4	0	0	0	12	6	1	0	$C=194$	12.9	4	19.6	17.5	9/15
S11	8	P_y	4	0	3	3	1	4	3	1	5	0	1	3	5	3	3	$P=39$	2.6	3	1.6	2.5	13/15
		C_y	34	0	10	7	0	0	26	46	34	0	0	8	11	8	0	$C=184$	12.3	8	15.2	18.5	9/15
S12	6	P_y	–	1	2	0	0	3	1	1	3	5	3	1	8	6	4	$P=38$	2.5	2	2.4	2.5	12/14
		C_y	–	1	3	0	0	25	4	3	16	28	10	6	23	6	0	$C=125$	8.3	4	9.9	12.5	11/14
S13	7	P_y	–	–	1	1	4	1	2	5	2	4	1	6	3	2	2	$P=34$	2.3	2	1.8	2.5	13/13
		C_y	–	–	2	1	36	2	3	9	16	26	0	20	19	1	0	$C=135$	9.0	2	11.6	17	11/13
S14	6	P_y	–	–	3	0	3	2	0	2	4	3	2	6	6	1	1	$P=33$	2.2	2	2.0	2.5	11/13
		C_y	–	–	3	0	5	15	0	1	33	10	3	24	17	8	0	$C=119$	7.9	3	10.2	12.5	10/13
S15	5	P_y	–	1	4	0	3	0	6	1	2	0	1	3	2	8	1	$P=32$	2.1	1	2.4	2.5	11/14
		C_y	–	0	5	0	7	0	31	0	7	0	0	51	3	8	0	$C=112$	7.5	0	14.4	7	7/14
S16	6	P_y	–	–	–	–	3	1	1	4	0	4	3	4	2	6	1	$P=29$	1.9	1	1.9	3.5	10/11
		C_y	–	–	–	–	36	4	0	41	0	24	4	25	4	8	0	$C=146$	9.7	4	14.3	16	8/11
S17	4	P_y	1	1	0	1	0	0	0	1	2	2	5	4	2	5	2	$P=26$	1.7	1	1.7	1.5	11/15
		C_y	0	1	0	1	0	0	0	0	0	9	8	12	6	1	0	$C=38$	2.5	0	4.1	3.5	7/15
S18	10	P_y	–	1	4	3	5	4	1	0	2	0	2	2	1	0	1	$P=26$	1.7	1	1.6	2	11/14
		C_y	–	8	60	151	40	110	2	0	32	0	2	9	6	0	0	$C=420$	28.0	6	45.9	36	10/14
S19	4	P_y	1	2	2	0	0	2	1	0	2	0	5	4	3	1	2	$P=25$	1.7	2	1.5	1.5	11/15
		C_y	10	2	0	0	0	2	2	0	9	0	12	12	9	1	3	$C=62$	4.1	2	4.7	9	10/15
S20	2	P_y	1	1	0	0	3	0	0	3	1	3	0	1	6	2	3	$P=24$	1.6	1	1.7	3	10/15
		C_y	2	2	0	0	1	0	0	2	0	8	0	1	4	0	0	$C=20$	1.3	0	2.2	2	7/15
S21	8	P_y	1	6	2	3	0	0	0	2	2	1	2	0	2	2	1	$P=24$	1.6	2	1.5	1.5	11/15
		C_y	2	22	15	8	0	0	0	25	22	3	31	0	2	2	0	$C=132$	8.8	2	11.0	18.5	10/15
S22	4	P_y	3	0	0	0	0	3	1	2	0	0	3	3	2	3	3	$P=23$	1.5	2	1.4	3	9/15
		C_y	9	0	0	0	0	4	7	17	0	0	22	2	2	3	0	$C=66$	4.4	2	6.8	5.5	8/15
S23	5	P_y	2	2	3	1	4	1	1	2	1	2	0	0	1	1	0	$P=21$	1.4	1	1.1	1	12/15
		C_y	10	17	13	3	11	2	0	0	8	4	0	0	2	3	0	$C=73$	4.9	3	5.5	9	10/15
S24	4	P_y	–	–	2	4	1	1	0	1	1	3	0	0	2	1	1	$P=17$	1.1	1	1.2	1.5	10/13
		C_y	–	–	4	12	9	1	0	2	4	29	0	0	4	4	1	$C=70$	4.7	2	7.6	4	10/13

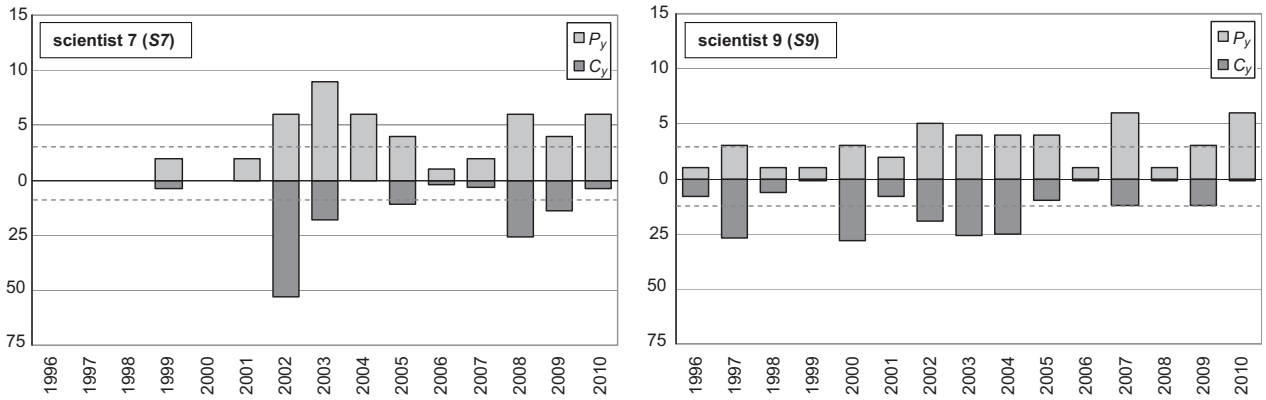

Fig. 1. *PY*/*CY* diagrams of two of the scientists in Table 1, plotting the yearly distribution of publications (P_y) at the upper-hand side of the horizontal axis, that of the citations (C_y) at the lower-hand side. Distribution mean values (μ) are graphically represented by horizontal dotted lines.

sign "−". For each scientist the following indicators are reported: h-index, P_y and C_y values (namely *PY* and *CY* distribution) with the corresponding sums (P and C respectively) mean value (μ), median (*Med*), standard deviation (s), interquartile range (*IQR*, calculated as the difference between the 3rd quartile and the 1st quartile of the distribution of interest) and a rough indicator of continuity (*Cont*), defined as the percentage of years with P_y and C_y: ($P_y > 0$, $C_y > 0$) over the years of activity (thus neglecting those years before the one containing the first publication). In particular, the indicators of dispersion (i.e., s and *IQR*) can be used to provide a rough indication on the distribution regularity, under the (questionable) assumption that an ideally regular pattern is uniform over time.

3. Analysis results

3.1. PY/CY diagram

This double diagram provides a qualitative picture of the temporal evolution of one researcher's scientific output, representing the relative *PY* and *CY* distributions. A similar graphical representation was used by Glänzel and Zhang (2010). Fig. 1 plots the *PY*/*CY* diagrams relating to two of the examined scientists, selected on a random basis (i.e., *S7* and *S9* in Table 1). In order to facilitate visualisation, citations are rescaled by factor 5.

This diagram is easy to construct and different information can be deduced from it:

- the shape of the *PY* and *CY* distribution;
- first year of publication activity (for instance 1999 for scientist 7 and 1996 for scientist 9);
- duration of publication activity (15 years in the example);
- amount of publications for each year (P_y) and corresponding impact in terms of total citations (C_y) accumulated up to the moment of the analysis;
- presence of discontinuities/interruptions in the scientific output, represented by null P_y or C_y values. In the example, publication activity of scientist 7 is null in 1996, 1997, 1998 and 2000, while publications of 2001 and 2004 have not yet received any citation. To quantify this aspect, the two above defined indicators of continuity, *Cont*(P_y) and *Cont*(C_y), are used. Of course, in case of absence of discontinuities, the indicator values are 1. For the two scientists in Fig. 1, publication continuity is respectively $11/12 \approx 0.92$ for scientist 7 and 1 for scientist 9, while citation continuity is $9/12 = 0.75$ for scientist 7 and 1 for scientist 9. Continuity values (*Cont*) relating to the other examined scientists are reported in the last column of Table 1.

By these diagrams, the two bibliometric aspects of production and impact (represented by the *PY* and *CY* distribution respectively) are analysed only separately. One might think of merging them by means of an aggregated indicator, such as the *annual* h-index (h_y), defined as the number such that, for a group of papers issued in the same year, h_y papers received at least h_y citations while the other papers received no more than h_y citations. We remark that the original aggregation criterion of h makes sense when the publications (elements of interest) and the corresponding citations (countable characteristic) are represented by numbers with the same order of magnitude (Franceschini & Maisano, 2010a). When considering the yearly production of an individual scientist, the typical number of citations per paper tend to be generally larger than the total number of papers, then h_y would tend to "degenerate" into P_y, losing its synthesis effectiveness. To clarify this fact, let consider the example in Fig. 2, reporting the complete production output of scientist 9 and the corresponding P_y and h_y values.

	year	'96	'97	'98	'99	'00	'01	'02	'03	'04	'05	'06	'07	'08	'09	'10	
	P_y	1	3	1	1	3	2	5	4	4	4	1	6	1	3	6	$P = \Sigma P_y = 45$
scientist 9 (*S9*)	citations received by each paper	8	16 7 4	6	1	12 9 7	6 2	7 4 4 4 0	12 5 5 4	11 9 3 2	6 2 1 1	1	5 4 3 0	1	5 4 3	1 0 0 0 0 0	$h = 7$
	C_y	8	27	6	1	28	8	19	26	25	10	1	12	1	12	1	$C = \Sigma C_y = 185$
	h_y	1	3	1	1	3	2	4	4	3	2	1	3	1	3	1	

Fig. 2. Published contributions of scientist 9 (*S9*), considering 15 consecutive years. It can be noted that the values of the annual h-index (h_y) are very similar to those of P_y. The fact that the number of annual papers of a scientist cannot be very large, compared to the typical number of citations that they received, entails that h_y values are generally limited by the corresponding P_y values.

3.2. Ferrers diagrams and triad indicators

A complementary representation of the P_y and C_y statistics can be obtained by Ferrers diagrams. Considering the *PY* distribution, each row of the diagram represents a partition of the publications among years (see the example in Fig. 3, relating to scientists 7 and 9). Years are sorted in descending order according to P_y. If there are several years with exactly the same publications, priority is given to the most recent ones. The largest completed (filled in) square of points in the upper left hand corner of a Ferrers diagram is called the Durfee square (Andrews, 1998; Egghe, 2010b). The Durfee square side is h_{PY} (in the example, it is 5 and 4 for *S7* and *S9* respectively). Precisely, a scientist has index h_{PY} if h_{PY} of his or her career years have at least h_{PY} publications each and the other years have $\leq h_{PY}$ publications each.

Clearly, h_{PY} is an extension of the classical Hirsch's h-index. In this case career years are the elements of interest and the corresponding yearly publications are the countable characteristic. h_{PY} may be not highly discerning when the career years and yearly number of publications do not have the same order of magnitude. For this reason, we think that analysis time window should embrace 10–15 years at least.

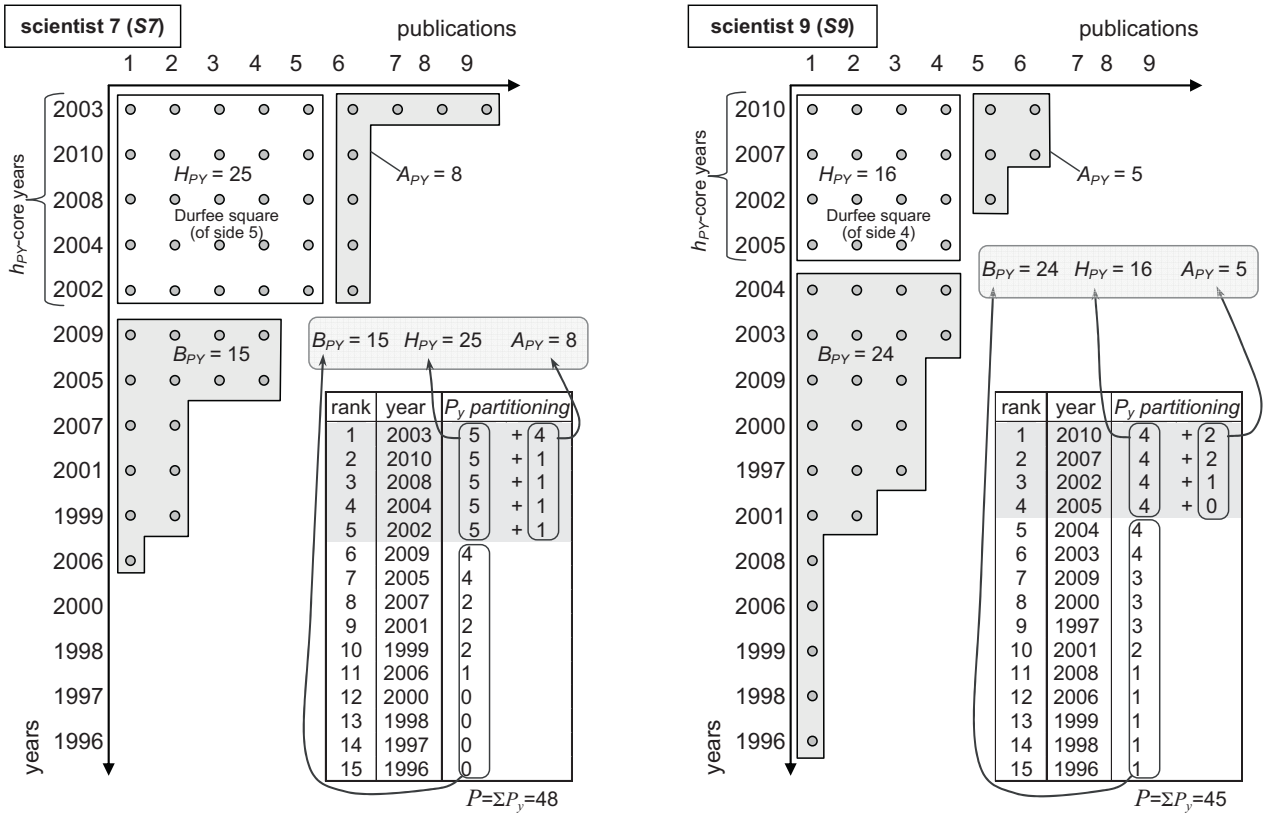

Fig. 3. Ferrers diagrams and calculation of the *triad* indicators (i.e., H_{PY}: publications in the Durfee square, A_{PY}: publications to the right of the Durfee square and B_{PY}: publications below the Durfee square) relating to the *PY* distribution (i.e., the yearly distribution of publications) of scientists 7 and 9 (see Table 1). P_y values are ranked in descending order and reported in the tables below. The largest completed (filled in) square of points in the upper left hand corner of a Ferrers diagram is called the Durfee square and it corresponds to h_{PY} (Egghe, 2010b).

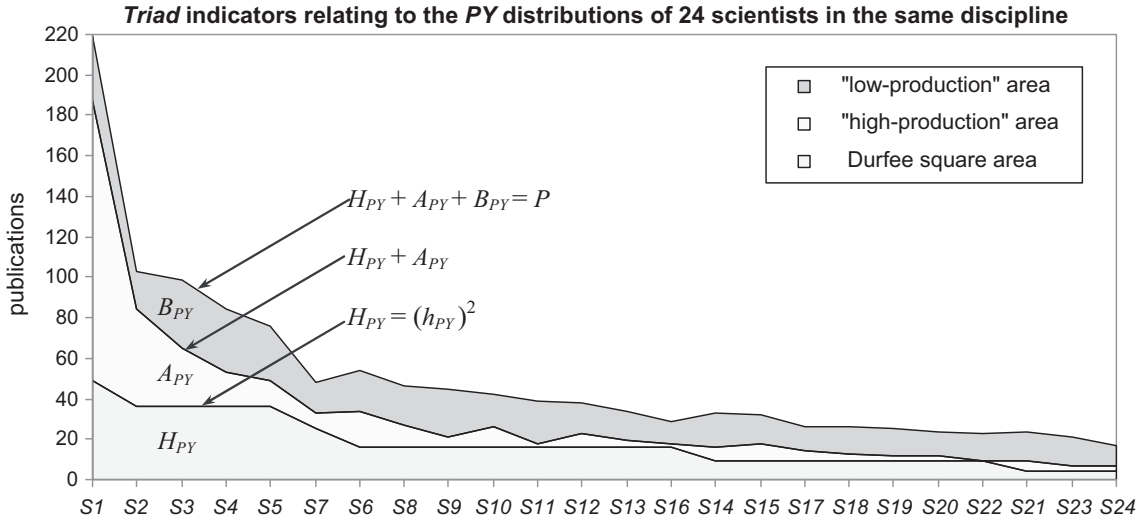

Fig. 4. Chart representing the triad indicators relating to the *PY* distributions of the 24 scientists in Table 1. Scientists are reported on the horizontal axis, while H_{PY}, A_{PY} and B_{PY} values on the vertical axis. Conventionally, scientists are ranked in lexicographic order $H_{PY} \rightarrow P$ (sort according to H_{PY} and, in case of equality, according to $P = H_{PY} + A_{PY} + B_{PY}$). Fig. 5(a) reports the scientists' complete *PY* distribution and the relevant H_{PY}, A_{PY}, B_{PY} and P values.

By the Ferrers diagram in Fig. 3, publications can be immediately subdivided into two categories:

1. the series of the h_{PY} most productive years, forming the h_{PY}-core. They can be classified as "high-production years", i.e., those years with a high number of publications, many of which to the right of the Durfee square;
2. years with relatively few publications (below the Durfee square). They can be classified as "low-production years", i.e., those years with not enough publications to be included within the h_{PY}-core.

The most productive and regular scientists are reasonably those with high h_{PY} values, since they are able to produce a conspicuous quantity of publications that are spread over time. Thus h_{PY} provides a rough quantification of the scientific output of a scientist from the regularity of production viewpoint.

Using the related Ferrers diagram, the complete time distribution of one scientist's publications can be subdivided into three main contributions:

(H_{PY}) publications in the Durfee square. H_{PY} coincides with h_{PY}^2
(A_{PY}) publications to the right of the Durfee square ("high-production" years)
(B_{PY}) publications below the Durfee square ("low-production" years)

This triple of indicators (H_{PY}, A_{PY} and B_{PY}), denominated as *triad*, was introduced in Franceschini and Maisano (2010b), but it was associated with another type of Ferrers diagram: that of the overall publications and citations of a scientist, ignoring their temporal distributions. In the present case, triad indicators provide a snapshot of a scientist's publication contributions. We highlight that $P = H_{PY} + A_{PY} + B_{PY}$. An example of calculation is shown in Fig. 3.

Triad's information content is certainly superior than that one given by a single indicator, such as h_{PY} or P. Our proposal is to associate these three indicators to each scientist, giving an instant overview of his/her publication output over time and facilitating comparisons among several scientists.

An effective way to compare different scientists on the basis of the triad indicators is given by the graph in Fig. 4. The corresponding numeric values are reported in Fig. 5(a).

Despite the fact that some authors are superior to others in terms of P values, sometimes they can be inferior in terms of h_{PY} values. For example, let consider scientist 6 in comparison to scientist 7 and scientist 14 in comparison to scientist 16. The scientific production of the former ones is relatively more concentrated in a limited number of years, as evidenced by their relatively high A_{PY} values.

Ferrers diagrams and triad indicators can also be constructed for evaluating the regularity relating to the impact of the scientific output, by using one scientist's *CY* distribution, instead of *PY* distribution. Likewise h_{PY}, h_{CY} is the Durfee square side of this new Ferrers diagram and a scientist has index h_{CY} if h_{CY} of his or her career years have publications with at least h_{CY} total citations (accumulated up to the moment of the analysis) and the other years have publications with $\leq h_{CY}$ total citations each. We remind that h_{CY} makes sense only if C_y values (countable characteristics) are not much larger than the number of years analysed (elements of interest). In this sense, the fact that most of the C_y values are likely to be larger than the number of analysed years makes h_{CY} potentially less discerning than h_{PY}.

Then, analogous triad indicators (H_{CY}, A_{CY} and B_{CY}) and a chart similar to the one in Fig. 4, but based on *CY* distributions, could be used for complementing h_{CY} and easing comparison among different scientists. Fig. 5(a) reports these other triad

a

h_{PY}, h_{PY}, P, C and *triad* indicators

	relating to *PY* distribution					relating to *CY* distribution				
Scient.	h_{PY}	H_{PY}	A_{PY}	B_{PY}	P	h_{CY}	H_{CY}	A_{CY}	B_{CY}	C
S1	7	49	138	32	219	11	121	376	20	517
S2	6	36	48	19	103	9	81	297	2	380
S3	6	36	29	34	99	9	81	116	4	201
S4	6	36	17	31	84	10	100	252	18	370
S5	6	36	13	27	76	3	9	20	10	39
S6	4	16	18	20	54	7	49	159	10	218
S7	5	25	8	15	48	5	25	97	13	135
S8	4	16	11	19	46	4	16	37	4	57
S9	4	16	5	24	45	8	64	95	26	185
S10	4	16	10	16	42	6	36	148	10	194
S11	4	16	2	21	39	8	64	113	7	184
S12	4	16	7	15	38	6	36	72	17	125
S13	4	16	3	15	34	6	36	90	9	135
S14	3	9	7	17	33	6	36	71	12	119
S15	3	9	9	14	32	5	25	79	8	112
S16	4	16	2	11	29	5	25	109	12	146
S17	3	9	5	12	26	4	16	19	3	38
S18	3	9	4	13	26	7	49	361	10	420
S19	3	9	3	13	25	5	25	27	10	62
S20	3	9	3	12	24	2	4	8	8	20
S21	2	4	5	15	24	6	36	87	9	132
S22	3	9	0	14	23	4	16	39	11	66
S23	2	4	3	14	21	5	25	34	14	73
S24	2	4	3	10	17	4	16	38	16	70

b

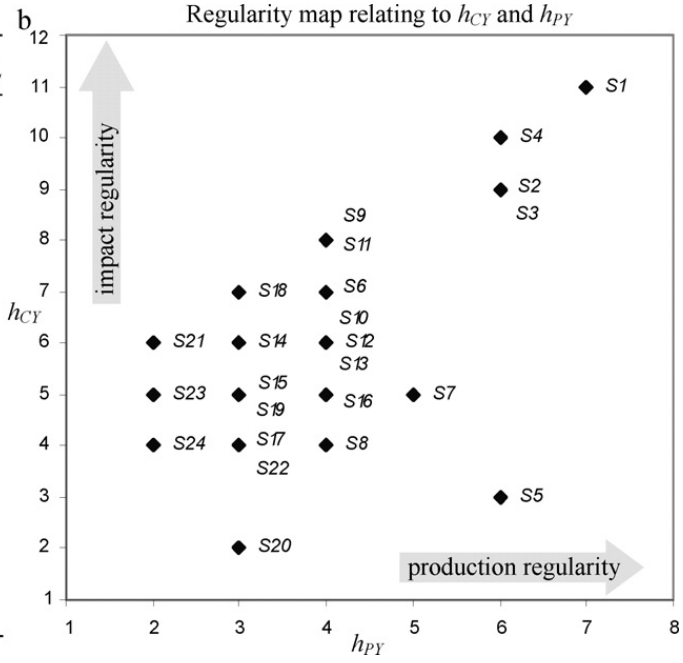

Fig. 5. (a) h_{PY}, h_{CY}, P, C and triad indicators, relating to the *PY* and *CY* distribution of the 24 scientists in Table 1. It can be noticed that the ranking based on H_{PY} is different from the one based on H_{CY}. (b) Regularity map representing the bibliometric positioning of the examined scientists, according to their h_{PY} and h_{CY} values.

indicators, associated to the 24 scientists of interest. It is worth remarking that B_{CY} values are generally small, while A_{CY} values are much higher, confirming the fact that C_y values are likely to be larger than the number of analysed years.

It can be noticed that the ranking based on H_{CY} is different from the one based on H_{PY}, confirming the fact that regularity in the production and regularity in the impact are two not necessarily related aspects (Glänzel, 2006). These two aspects, which have been analysed separately, can be represented together through the regularity map in Fig. 5(b).

3.3. Year-by-year comparison of several scientists (Borda's ranking)

We now propose a structured comparison among the *PY* and *CY* distributions of different scientists based on the assumption that the most regular scientists are those who are able to overcome their competitors for most years. This logic is pretty alike to the one of Formula One races, where the world championship is generally the driver who is regularly in the top positions for as many competitions as possible during the season, not the one who alternates outstanding with poor performances.

A very simple method for comparing the *PY* and *CY* distributions of different scientists is the Borda's method (Borda, 1781; Saari, 1995). Referring to each year, an xth scientist has a rank $r_y(x)$: 1 for the first position, 2 for the second, ... and n for the last. The Borda score (B) for the xth scientist is the sum of his yearly ranks:

$$B(x) = \sum_{i=1}^{n} r_i(x) \tag{1}$$

The winner is the scientist (x^*) with the lowest Borda score:

$$B(x^*) = \min_x B(x) \tag{2}$$

Borda's algorithm is applied to the *PY* and *CY* distributions of the 24 examined scientists. Borda yearly ranks and Borda scores (B) relating to the previous *PY* and *CY* distributions are reported in Fig. 6(a) and (b). It is worth noting that final rankings based on regularity may be different from those based on the overall scientific output. For instance, scientist 4 – despite having the 4th highest P and C values – is in 3rd and 2nd position respectively for the relevant Borda ranks. This is a consequence of the fact that most of publications and citations of scientist 6 are relatively well distributed over the years.

Apart from being simple, this comparison on an annual basis makes it possible to "filter out" other generalized trends, which are not necessarily related to the performance of scientists; in particular:

- the increasing tendency towards publishing and citing, favored by recently introduced rewards and incentives (Bornmann, 2011; Stephan, 2008);

Borda yearly ranks relating to PY distributions

Scient.	P	rank	r'96	r'97	r'98	r'99	r'00	r'01	r'02	r'03	r'04	r'05	r'06	r'07	r'08	r'09	r'10	B	$r_B(P_y)$
S1	219	1st	2	5	3	5	1	11	4	4	3	1	1	1	1	1	1	44	1st
S2	103	2nd	15	10	17	17	3	11	4	7	9	2	2	4	2	2	5	110	5th
S3	99	3rd	9	10	2	2	3	1	7	5	9	8	4	2	3	3	2	70	2nd
S4	84	4th	1	5	17	11	1	11	1	3	6	3	5	3	4	5	2	78	3rd
S5	76	5th	6	4	9	1	8	20	7	16	2	3	5	5	5	4	2	97	4th
S6	54	6th	15	10	17	9	8	2	12	1	1	5	9	14	11	19	15	148	8th
S7	48	7th	15	18	17	9	18	7	1	2	3	8	16	18	7	10	5	154	9th
S8	46	8th	6	2	14	11	8	11	12	16	18	5	2	5	11	11	15	147	7th
S9	45	9th	9	5	14	11	8	7	4	7	6	8	16	5	22	11	5	138	6th
S10	42	10th	2	1	1	2	18	11	10	11	22	19	21	5	13	11	15	162	11th
S11	39	11th	2	18	6	5	16	2	9	16	5	19	16	14	10	11	9	158	10th
S12	38	12th	15	10	9	17	18	5	12	16	9	7	10	20	6	7	8	169	12th
S13	34	13th	15	18	14	11	6	11	10	6	12	8	16	5	13	16	12	173	13th
S14	33	14th	15	18	6	17	8	7	20	11	6	13	13	5	7	19	15	180	14th
S15	32	15th	15	10	3	17	8	20	1	16	12	19	16	14	16	6	15	188	15th
S16	29	16th	15	18	17	17	8	11	12	7	22	8	10	11	16	7	15	194	16th
S17	26	17th	9	10	17	11	18	20	20	16	12	16	5	11	16	9	12	202	19th
S18	26	17th	15	10	3	5	3	2	12	23	12	19	13	18	22	24	15	196	18th
S19	25	19th	9	8	9	17	18	7	12	23	12	19	5	11	13	19	12	194	16th
S20	24	20th	9	10	17	17	8	20	20	10	18	13	21	20	7	16	9	215	23rd
S21	24	20th	9	2	9	5	18	20	20	11	12	18	13	22	16	16	15	206	21st
S22	23	22nd	5	18	17	17	18	5	12	11	22	19	10	14	16	11	9	204	20th
S23	21	23rd	6	8	6	11	6	11	12	11	18	16	21	22	22	19	24	213	22nd
S24	17	24th	15	18	9	2	16	11	20	16	18	13	21	22	16	19	15	231	24th

a

Borda yearly ranks relating to CY distributions

Scient.	C	rank	r'96	r'97	r'98	r'99	r'00	r'01	r'02	r'03	r'04	r'05	r'06	r'07	r'08	r'09	r'10	B	$r_B(C_y)$
S1	517	1st	6	6	2	3	6	17	5	2	9	1	2	1	2	6	3	71	1st
S2	380	3rd	12	15	14	15	8	15	3	3	9	2	1	4	1	1	5	108	3rd
S3	201	6th	12	15	13	15	7	6	8	11	15	3	4	3	5	4	8	129	5th
S4	370	4th	3	1	14	8	1	2	6	10	16	5	10	8	8	2	3	97	2nd
S5	39	22nd	9	8	14	5	16	17	16	18	17	18	16	16	18	12	8	208	22nd
S6	218	5th	12	12	14	5	12	11	10	1	1	4	6	19	11	21	8	147	6th
S7	135	11th	12	15	14	10	16	17	1	8	17	10	13	16	3	2	1	155	10th
S8	57	21st	12	4	14	15	10	5	15	18	17	15	16	16	24	21	8	210	23rd
S9	185	8th	8	3	7	12	5	7	7	6	5	11	15	9	23	5	5	128	4th
S10	194	7th	1	2	2	2	16	8	16	13	17	19	16	9	12	17	8	158	12th
S11	184	9th	2	15	6	8	16	17	4	4	2	19	16	14	9	7	8	147	6th
S12	125	14th	12	12	10	15	16	3	10	14	7	7	8	15	4	11	8	152	9th
S13	135	11th	12	15	12	12	3	12	12	12	7	8	16	7	5	17	8	158	12th
S14	119	15th	12	15	10	15	14	4	16	17	3	11	12	6	7	7	8	157	11th
S15	112	16th	12	15	8	15	12	17	2	18	13	19	16	2	18	7	8	182	17th
S16	146	10th	12	15	14	15	3	9	16	5	17	9	11	5	15	7	8	161	14th
S17	38	23rd	12	12	14	12	16	17	16	18	17	13	9	9	12	17	8	202	21st
S18	420	2nd	12	9	1	1	2	1	13	18	4	19	13	13	12	21	8	147	6th
S19	62	20th	4	10	14	15	16	12	13	18	11	19	6	9	10	17	2	176	16th
S20	20	24th	9	10	14	15	15	17	16	15	17	14	16	21	15	21	8	223	24th
S21	132	13th	9	5	4	7	16	17	16	7	6	17	3	22	20	16	8	173	15th
S22	66	19th	6	15	14	15	16	9	9	9	17	19	5	19	20	14	8	195	20th
S23	73	17th	4	7	5	11	8	12	16	18	12	16	16	22	20	14	8	189	19th
S24	70	18th	12	15	9	4	11	15	16	15	14	6	16	22	15	13	5	188	18th

b

c

Scient.	overall B	final r_B
S1	2	1st
S2	8	4th
S3	7	3rd
S4	5	2nd
S5	26	14th
S6	14	6th
S7	19	8th
S8	30	15th
S9	10	5th
S10	23	10th
S11	16	7th
S12	21	9th
S13	25	12th
S14	25	12th
S15	32	17th
S16	30	15th
S17	40	20th
S18	24	11th
S19	32	17th
S20	47	24th
S21	36	19th
S22	40	20th
S23	41	22nd
S24	42	23rd

Fig. 6. Application of Borda method to the *PY* and *CY* distributions in Table 1. Borda scores (B) and the relevant ranks – respectively $r_B(P_y)$ and $r_B(P_y)$ – are reported in the last two columns of tables (a) and (b). The overall Borda score (*overall B*) relating to the previous rankings and the (unique) final ranking (*final* r_B) are reported in table (c).

- the physiological decrease in the C_y values in the most recent years, due to the citation accumulation process; e.g., according to some authors, the amount of time to collect most of the citations is about 3–5 years for papers in the engineering field (Amin & Mabe, 2000).

It is important to highlight that this method does not adequately consider the year-by-year "gap" among scientists. For example, considering a specific year, the gap between two scientists with rank positions 4th and 6th is not necessarily coincident to the gap between two groups with rank positions 1st and 3rd. This is one of the typical problems of the indicators based on rankings (Billaut, Bouyssou, & Vincke, 2010). Growing in complexity, we could introduce other more refined methods that take into account also the magnitude of gaps between rank positions.

Also, the method penalizes scientists who started to publish later on (*S2*, *S6*, *S7*, *S12*, *S13*, *S14*, *S15*, *S16*, *S18* and *S24* in Table 1). To overcome this problem, one could replace the B value of a scientist with his/her mean yearly rank (mean r_y), obtained by excluding the years of inactivity.

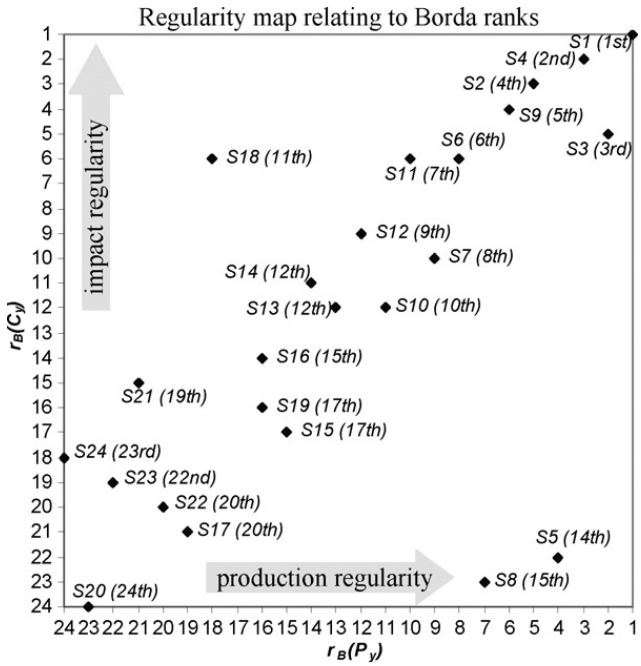

Fig. 7. *Regularity map* representing the bibliometric positioning of the examined scientists, according to their Borda rankings – respectively $r_B(P_y)$ and $r_B(C_y)$, see Fig. 6(a) and (b). The unique ranks, obtained by reapplying Borda algorithm to $r_B(P_y)$ and $r_B(C_y)$ (see Fig. 6(c), are reported in the data labels (in brackets).

The output of the comparison carried out so far consists of two independent rankings, respectively according to the *PY* and the *CY* distribution. Synthesising these rankings into a single one is a delicate operation. For example, one could do this by reapplying Borda's algorithm (see the *final* r_B ranking in Fig. 6(c)). However, this entails that regularity in production and in impact have the same importance. This rank brings out interesting cases. For example, the "massive producer" *S5* (with many publications) is "outranked" by *S18*, who is a most selective but "efficient" researcher (with only 26 publications he/she gets almost the same number of citations as *S1*).

An alternative way to use the two rankings together, without merging them, is to draw a regularity map (see Fig. 7). Such a map illustrates the bibliometric positioning of different scientists, from the point of view of regularity. The most regular scientists are those with low Borda scores (relating to both *PY* and *CY* distributions). They are located near the top-right corner. Comparing this regularity map with that based on h_{PY} and h_{CY} (Fig. 5(b)), some differences emerge, due to different logic with which regularity is evaluated.

4. Regularity versus overall output

The proposed tools make it possible to evaluate and compare scientists according to their regularity, which is a "novel" bibliometric property. A first correlation analysis between the scientists' regularity and their overall output can be performed by considering the two bibliometric aspects of production and impact separately. The diagram in Fig. 8(a) shows the Borda ranks, i.e., $r_B(P_y)$, of the 24 examined scientists relating to their *PY* distributions against the ranks in terms of *P* (both reported in Fig. 6(a)). The diagram in Fig. 8(b) instead shows the Borda ranks, i.e., $r_B(C_y)$, of the scientists relating to their *CY* distributions against the ranks in terms of *C* (both reported in Fig. 6(b)).

Results relating to the two approaches (i.e., regularity and overall output) are quite correlated (high R^2 values). This is a sign that scientists have more or less the same tendency to spread the scientific production over the years. Therefore, the "initial advantage" of scientists with large overall output is generally maintained in terms of comparisons on an annual basis. In any case, this does not mean that a massive production must always imply production regularity.

Regarding impact, we remark that this correlation is relatively weaker, probably because impact is not controlled by scientists. To be precise, we are aware that scientists do not always have control over the publication time of their papers (e.g., because of frequent delays in the editorial and publication process); however, the fact remains that this type of control is undoubtedly greater than that over the temporal distribution of the papers' impact.

Despite the general correlation, the approach based on regularity enriches the one based on overall output, making it possible to identify "lone voices" i.e., scientists with remarkable overall output but spread irregularly over time or vice versa. For example, let consider scientist 18, who has the second largest *C* value but is ranked only 6th according to the Borda score, since most of his/her citations are concentrated in relatively few years. Scientist 9, on the other hand, has the 9th largest *P* value but is ranked 6th in terms of publication regularity. This means that there are 3 scientists (i.e., *S6*, *S7* and *S8*, see Fig. 8(a)) who are more prolific but irregular as well.

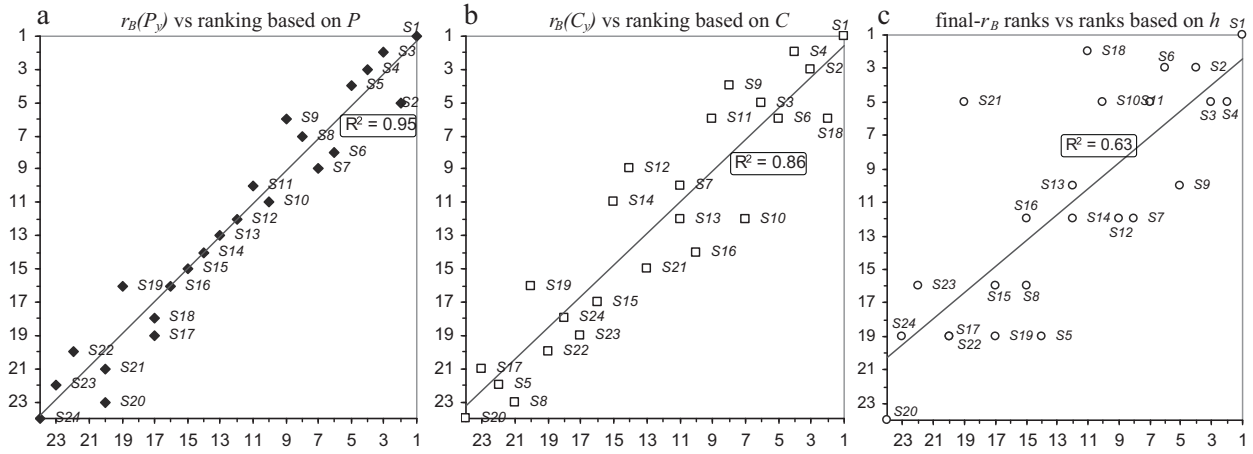

Fig. 8. (a) Borda ranks of the 24 examined scientists relating to their *PY* distributions against the ranks in terms of *P* (numeric data are reported in Fig. 6(a). (b) Borda ranks of the 24 examined scientists relating to their *CY* distributions against the ranks in terms of *C* (numeric data are reported in Fig. 6(b)). (c) *Final* r_B ranks against ranks based on *h*. Numeric data are reported in Fig. 6(c) and Table 1. These graphs show that results of the approach based on regularity and that one based on the overall output are quite correlated (high R^2 values).

A second comparison of the two approaches can be performed by considering indicators that aggregate the two bibliometric aspects of production and impact. Specifically, regarding regularity, we consider the unique ranking obtained by reapplying Borda method (*final* r_B in Fig. 6(c), while, regarding the overall production, we consider the ranking obtained by the very well-known *h*-index (see the corresponding data in the second column of Table 1). As Fig. 8(c) shows, the correlation between these two rankings is less pronounced ($R^2 = 0.63$) probably due to different aggregation logics.

5. Concluding remarks

This paper analyses the scientific output of 24 Italian academic scientists involved in the same scientific discipline, from the regularity viewpoint. This analysis represents a useful complement to the classical bibliometric techniques. In particular, three recent tools are presented and described in detail in the article. They are respectively: (1) the *PY*/*CY* diagram, (2) the publication/citation Ferrers diagram and triad indicators, and (3) a year-by-year procedure for comparison of scientists according to their *PY* and *CY* distributions (Borda's ranking).

Being based on indicators (basically, the number of annual publications and citations of a scientist) that are commonly used in bibliometric studies concerning any scientific discipline, regularity analysis can be reasonably generalized to other scientists and scientific fields. A first limitation is that comparison should be restricted to scientists in the same discipline – owing to the different citation rates (Amin & Mabe, 2000) – and with similar career lengths, otherwise those with longer careers are favored.

Another limitation is that the two bibliometric aspects of production and impact are analysed only separately and their aggregation remains an open issue. This problem is only partially overcome by introducing some regularity maps, which represent the bibliometric positioning of the scientists according to their regularity in production and impact.

Despite these limitations, the suggested regularity analysis enriches the traditional analysis approach based on indicators of overall output, like *P*, *C*, *h*, etc., and these two approaches can be combined together to identify "abnormal" situations represented by scientists for which overall output and regularity "do not go hand in hand". In conclusion, since in case of competitive examinations for research position/promotion/tenure acquisition scientists are generally within the same field and have similar career lengths, we believe that the proposed tools may be useful as they are.

Regarding the future, the analysis tools will be tested on the basis of a larger amount on empirical data on specific research fields. Moreover, regularity analysis may be extended to other dimensions of research performance in addition to total production and total impact; first of all "efficiency" in terms of citations – roughly estimated by the citation rate (*CPP*, citations per paper) – which can be studied both in overall terms and on an annual basis.

References

Alonso, S., Cabrerizo, F., Herrera-Viedma, E., & Herrera, F. (2009). *h*-Index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*, *3*(4), 273–289.

Amin, M, & Mabe, M. (2000). *Impact factors: Use and abuse*. *Perspectives in Publishing* Elsevier Science.

Andrews, G. E. (1998). *The theory of partitions*. Cambridge, UK: Cambridge University Press.

ASPHER – The Association of Schools of Public Health in the European Region (2010). *Doctoral Education and Research Capacities in Public Health – Background and guidelines*. Retrieved from http://www.aspher.org/pliki/pdf/ASPHER_Report_DoctoralStudiesPH.pdf

Bar-Ilan, J. (2010). Citations to the 'introduction to informetrics' indexed by WOS, Scopus and Google Scholar. *Scientometrics*, *82*(3), 495–506.

Billaut, J. C., Bouyssou, D., & Vincke, P. (2010). Should you believe in the Shanghai ranking? An MCDM view. *Scientometrics*, *84*(1), 237–263.

Borda, J. C. (1781). Mémoire sur les élections au scrutin, Comptes Rendus de l'Académie des Sciences, translated by Alfred de Grazia as Mathematical derivation of an election system. *Isis*, *44*, 42–51.

Bornmann, L., Marx, W., Schier, H., Rahm, E., Thor, A., & Daniel, H. D. (2009). Convergent validity of bibliometric Google Scholar data in the field of chemistry – Citation counts for papers that were accepted by Angewandte Chemie International Edition or rejected but published elsewhere, using Google Scholar, Science Citation Index, Scopus, and Chemical Abstracts. *Journal of Informetrics*, *3*(1), 27–35.
Bornmann, L. (2011). Mimicry in science? *Scientometrics*, *86*(1), 173–177.
Collegio dei presidenti di corso di studi in Matematica (2008). *Considerazioni e proposte relative agli indicatori di qualità di attività scientifica e di ricerca, e ai parametri per le valutazioni comparative*. Retrieved from http://users.unimi.it/barbieri/indicatoriMAT_29nov08.pdf
Cronin, B. (1984). *The citation process: The role and significance of citations in scientific communication*. London: Taylor Graham.
Egghe, L. (2010a). The Hirsch-index and related impact measures. *Annual Review of Information Science and Technology*, *44*, 64–114.
Egghe, L. (2010b). Conjugate partitions in informetrics: Lorenz curves, h-type indices, Ferrers graphs and Durfee squares in a discrete and continuous setting. *Journal of Informetrics*, *4*(3), 320–330.
Franceschini, F., Galetto, M., & Maisano, D. (2007). *Management by measurement: Designing key indicators and performance measurements*. Berlin: Springer., ISBN 9783540732112.
Franceschini, F., & Maisano, D. (2010a). Analysis of the Hirsch index's operational properties. *European Journal of Operational Research*, *203*(2), 494–504. doi:10.1016/j.ejor.2009.08.001
Franceschini, F., & Maisano, D. (2010b). The citation triad: An overview of a scientist's publication output based on Ferrers diagrams. *Journal of Informetrics*, *4*(4), 503–511.
Franceschini, F., & Maisano, D. Proposals for evaluating the regularity of a scientist's research output. *Scientometrics*, in press-a, doi:10.1007/s11192-011-0371-4.
Franceschini, F., & Maisano, D. Influence of database mistakes on journal citation analysis: Remarks on the paper by Franceschini and Maisano, QREI (2010), *Quality and Reliability Engineering International*, in press-b, doi:10.1002/qre.1174.
Franceschini, F., & Maisano, D. (2011). Structured evaluation of the scientific output of academic research groups by recent h-based indicators. *Journal of Informetrics*, *5*(1), 64–74.
Glänzel, W. (2006). On the h-index: A mathematical approach to a new measure of publication activity and citation impact. *Scientometrics*, *67*(2), 315–321.
Glänzel, W., & Zhang, L. (2010). A demographic look at scientometric characteristics of a scientist's career. *ISSI Newsletter*, *6*(3), 66–84.
Haeffner-Cavaillon, N., & Graillot-Gak, C. (2009). The use of bibliometric indicators to help peer-review assessment. *Archivum Immunologiae et Therapiae Experimentalis*, *57*(1), 33–38.
Harzing, & van der Wal. (2008). Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics*, *8*(11), 61–73.
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 16569–16572.
IPEA – Institute of Applied Economic Research. (2009) *Simplified public call for proposals IPEA/PVE Nr 001/2009 – Selection of candidates for research funding.* Retrieved from http://www.ipea.gov.br/sites/000/2/pdf/PublicCall_001_09PVE.pdf
Labbé, C. (2010). Ike Antkare one of the great stars in the scientific firmament. *ISSI Newsletter*, *6*(2), 48–52.
Moed, H. F. (2005). *Citation analysis in research evaluation*. Dordrecht: Springer.
Neuhaus, C., & Daniel, H. D. (2008). Data sources for performing citation analysis: An overview. *Journal of Documentation*, *64*(2), 193–210.
Rousseau, R. (2008). Reflections on recent developments of the h-index and h-type indices. In *Proceedings of WIS 2008, Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting Berlin* Germany.
Saari, D. G. (1995). *Basic geometry of voting*. Berlin: Springer., ISBN 3-540-60064-7.
Snizek, W. E. (1995). Some observations on the use of bibliometric indicators in the assignment of university chairs. *Scientometrics*, *32*(2), 117–120.
Stephan, P. E. (2008). Science and the university: Challenges for future research. *CESifo Economic Studies*, *54*(2), 313–324. doi:10.1093/cesifo/ifn014
Van Raan, A. F. J. (2000). The Pandora's box of citation analysis: Measuring scientific excellence, the last evil? In B. Cronin, & H. B. Atkins (Eds.), *The web of knowledge: A festschrift in honor of Eugene Garfield* (pp. 301–319). Medford, NJ: ASIS Monograph Series, Information Today Inc.