UNIVERSITÀ
DI SIENA
1240

QUADERNI DEL DIPARTIMENTO
DI ECONOMIA POLITICA E STATISTICA

Alberto Baccini, Lucio Barabesi
Martina Cioni, Caterina Pisani

Crossing the hurdle: the determinants of individual scientific performance

n. 691 – Dicembre 2013

Abstract - An original dataset referring to a medium-sized Italian university is implemented for analyzing the determinants of scientific research production at individual level. Three different indicators, based on the number of publications and/or citations, are considered. Their distributions are highly skewed, displaying an excess of zero-valued observations, thus zero-inflated and hurdle regression models are introduced. Among them, the Hurdle Negative Binomial model exhibits a good fitting and appears to be reasonably coherent with the underlying generating data process. Indeed, the performance of active researchers is described by the count component of the model, while the odds to be in a non-active status is modelled by the zero component. Individual characteristics, teaching and administrative activities, as well as the features of the department the researcher belongs to, are considered as explanatory variables. The analysis of the results highlights that scientific productivity is lower for oldest active researchers, and that there is a significant effect of academic position on research production. Evidence of clear-cut and significant substitution or complementarity effect between teaching and research activities is not found. Indeed, a major teaching load does not apparently affect the odds to be a non-active researcher, while it has mixed and very weak effects on publication performance of active researchers. A negative relationship among productivity and administrative tasks is highlighted. Finally, the analysis points out the effects of department composition on both the odds to be non-active and the value of the production indicators of active researchers.

Keywords: Academic research productivity, Scientist productivity, Hurdle models, Zero-Inflated models, Negative Binomial distribution, Poisson distribution

Acknowledgments: We would like to thank Sonia Boldrini from the Evaluation Committee of the University of Siena for her valuable assistance in data harvesting and database building.

Alberto Baccini, Lucio Barabesi, Martina Cioni, Caterina Pisani
Department of Economics and Statistics - University of Siena

alberto.baccini@unisi.it, lucio.barabesi@unisi.it, martina.cioni@unisi.it, caterina.pisani@unisi.it

1. Introduction

In the last thirty years, there has been a growing interest in the role of academic research activity and in its contribution to economic growth and social development. One of the least studied and most puzzling feature of this debate is the question of individual scientific productivity. Researcher’s activity is basically a multi-output activity, producing outcomes as research, teaching and others products (newspapers articles, medical protocols, etc.) with relevant impact on society. This idea is so widespread that in the national research assessment exercises (for example in the British Research Excellence Framework 2014) information is collected about all these different activities to evaluate not only the quality of research produced by universities, but also their multifaceted societal impact. At the best of our knowledge, while no study addresses the question of the determinants of researchers’ overall production, many papers focus solely on one dimension of their multi-output activities.

However, the idea that scientific publications represent the essence of the research activity is widely accepted (Wooton, 2013). In this respect, scholars have addressed two different issues. The first is about the development of models able to accurately describe the law governing the frequency distribution of researcher’s publications; the second is about the determinants of individual productivity. In this literature, the number of publications of a researcher is the most used proxy for quantifying her/his productivity. The main drawback of this indicator is that each publication counts for one: a short paper addressing a limited issue counts as a seminal paper. So, the holy grail of scientometric research is the construction of indicators addressing at the same time the issue of individual productivity and of “quality” of scientific work, also at a researcher level (Van Leeuwen et al., 2003).

Indeed, the most used strategy in empirical research consists in the substitution of the notion of “research quality” with the one of “scientific impact”, as defined in Martin and Irvine (1983), that can be more easily handled using data on citations. One possible line of inquiry consists in counting a subset of the publications of a researcher, such as the highly cited papers or those published in top journals - defined in reference to the impact factor or to others similar indicators. A different approach aims to obtain composite indicators of productivity and impact of a researcher considering her/his published articles together with the citations they received. Among them, the most widespread indicator is the h-index (Hirsch, 2005).

The present paper - based on an original dataset - contributes to the debate by focusing on the determinants of individual scientific production as proxied by three different indicators based on the number of publications and/or their citations. These indicators are integer-valued and show distributions which are highly skewed with an excess of zeros. Since most of the scientific contributions are produced by few prolific researchers - as also stylized by David (1994) - the distribution of the indicator adopted to quantify research performance tends to display heavy tails and high frequency of zero values, irrespective of the proxy considered and the discipline investigated. Nevertheless, many existing papers dealing with modelling the determinants of scientific production do not specifically address the skewed and zero-inflated nature of the data (Carayol and Matt, 2004 & 2006; Lissoni et al., 2011; Rivera-Huerta et al., 2011). Moreover, some authors impose artificial smoothness to the data, although recognizing the discrete nature of the production indicators they use (Kelchtermans and Veugelers, 2011; Lissoni et al., 2011).

In order to model skewed count data with an excess of zero-valued observations, we introduce zero-inflated and hurdle regression models, which, even if well-established in other disciplines, have been quite neglected in the framework of the analysis of production process of academic research.

The remainder of the paper is divided into four sections. Section 2 surveys the main literature addressing the issues of researcher's production. Section 3 illustrates the data and the methodology used. Section 4 presents and discusses the main results of our empirical analysis. Finally, in Section 5 conclusions are drawn and avenues for further research are considered.

2. Literature review

Two alternative approaches have been considered in the analysis of scientific production: the first focuses on the laws underlying the frequency distribution of the number of publications (or citations), while the second aims to identify the determinants of scientific performance. The first approach dates back to the 1920s, particularly to the publication of Lotka's seminal article (1926). Lotka investigated the frequency distribution of scientific productivity of chemists and physicists showing that "...the number (of authors) making *n* contributions is about l/n^2 of those making one; and the proportion of all contributors, that make a single contribution, is about 60%...". This kind of approach has survived until the recent attempts to create theoretical models able to foresee the future pattern of production of a scholar given her/his past performance (e.g. Wang et al., 2013 and the bibliography cited thereon).

Alternative explanations of Lotka findings and, more in general, of the highly skewed nature of scientific production of scholars have been proposed. The simplest one - highly criticized for example by Allison and Stewart (1974) and by David (1994) - is the so-called “sacred-spark hypothesis”, i.e. the differences in productivity reflect unequal and predetermined capabilities of researchers. In the late 1960s, a so called “Matthew-effect hypothesis” was advanced by Merton (1968). Merton highlighted that well-known researchers receive more recognition for their work than less well-known researchers. This hypothesis was then generalized by Cole and Cole (1973) to be valid not only for recognition, but, more generally, for scientific productivity. In this form, this hypothesis was called “cumulative advantage hypothesis”. The idea is that recognition received early in researchers’ career may be reinforced over time as it would enable easier access to research resources; this issue means that any advantage will be cumulative (Defazio et al., 2009). This kind of explanation exclusively focuses on the social structure in which scholars are embedded and deploys their works. A second approach addresses the academic research production aiming to identify both individual and collective determinants of scientific productivity: the former refer to personal characteristics of a researcher and the latter to the features of the organization in which she/he is active. In these works the sacred-spark and the Mattew-effect hypotheses are considered - so to speak - as residual or unexplained components of a roughly defined production function addressing all relevant explanatory variables. This explanatory approach has been applied both to the individual level of analysis - where the survey unit is the researcher - and to the aggregate level - where the survey unit is the research unit (e.g. department, laboratory, university). A complete review of all the elements that a broad and growing applied literature considers as possible explanatory variables of individual and aggregate scientific productivity is out of the aim of our work. We focus our attention solely on the issues used in our models.

The set of individual determinants considered can be grouped into the following categories: personal characteristics of a researcher, such as gender and age; individual career features, such as seniority, career progress and scientific field in which a scholar is working; activities in which the researcher is engaged, as teaching activities and administrative tasks; and, finally, access to research funding.

As to the personal characteristics, we are interested in evaluating if age affects the publication activity, as suggested by life-cycle models. Indeed, some studies predict a decrease of research production as age increases (Diamond, 1984, 1986); others find that publication activity tends to increase in the early career, reaches a peak, and then decreases (Levin and Stephan, 1991; Weiss and Lillard, 1982; Zuckerman and Merton, 1972); while others find a productivity curve with two peaks (Bayer and Dutton, 1977). These relationships have to be taken with caution, since it is difficult to distinguish between age effect and cohort effect. Indeed, the latter can be associated, for example, with a progression of knowledge or a different availability of resources, as discussed by Stephan (1996, 2012). At the aggregate level of analysis, a study on the Italian National Research Council highlights a negative relationship between age and research productivity indicators (Bonaccorsi and Daraio, 2003), while the results by Carayol and Matt (2004) suggest an “inversed-U shape” relationship between laboratories productivity and age. In contrast, a lack of significant relationship between age and publication rate within the faculties of the University of Vienna is claimed by Wallner et al. (2003).

Regarding the role of gender in scientific publication performance, since 1980s - when Cole and Zuckerman (1984) defined the gender differences in productivity among academic scientists as “a puzzle” - several studies tried to address this issue (Fox, 2005; Fox et al., 2011; Leahey, 2006; Levine and Stephan, 1998; Xie and Shauman, 2003). Mairesse and Pezzoni (2013) revisit the gender gap in scientific production, offering a critical review of the empirical evidence throughout the analysis of what influences women scientific productivity (family engagements, marital status and policies in favour of women, institutional specificities, discipline specificities, etc.). Abramo et al. (2009) document differences in productivity between men and women, but highlight a progressive reduction of the performance gap over time for Italian scientists (hard sciences and life sciences). Similar conclusions are achieved by van Arensbergen et al. (2012) who suggest that, even if men outperform women in terms of scientific production in the older generations, the gendered differences are disappearing in the younger generations.

As to the effects of seniority and career progress, the role of tenure and position are ambiguous. On one hand, an improvement in the professional status of the researcher can positively affect research performance, since - for example - she/he can have easier access to funding or attract talented young students/scientists in her/his research team (the so-called “status effect”). On the other hand, once a career progress has been obtained, the incentives to production can be reduced. The relationship between seniority or career progress and research performance is nearly universally addressed in this literature. Among others, Fabel et al. (2008) find a negative effect of career age on publications for full professors, while Rivera-Huerta et al. (2011) consider career years as a control variable in modelling individual research output.

It is well known that publishing activities and citation patterns vary among scientific disciplines. These differences - as emphasized by Anania and Caruso (2013) - are particularly relevant in many areas of Social Sciences and Humanities, where scientific productivity and citation practices typically yield fewer citations per paper. Two strategies are used for tackling this problem. The first strategy tends to limit the analysis on researchers working in homogeneous scientific fields (see e.g. Lissoni et al., 2011; Pezzoni at al., 2012). The second strategy aims to model the research performance of scientists from heterogeneous areas including control variables for the researcher scientific discipline - defined according to some available classification (see e.g. Carayol and Matt, 2006; Rivera-Huerta et al., 2011).

As to the multidimensional nature of scientists' activity, the question to be addressed is if the engagement in activities - such as teaching and administrative tasks - may crowd out research. According to some authors (Fox, 1992; Taylor et al., 2006), these activities conflict as the more productive researchers spend less time for teaching and students in general. This substitutive relationship is also documented by a recent paper on French professors in Economics (Kossi et al., 2013). In contrast, an analysis on business economists in Germany and Switzerland shows that higher teaching loads in terms of class sizes do not deteriorate research productivity (Fabel et al., 2008) and an investigation performed by Kelchtermans and Veugelers (2011) on scientists employed at the KU Leuven highlights that alternative activities have very small and mostly insignificant effects on research output.

However, as suggested by Stephan (1996), given the collaborative nature of science, individual determinants solely represent a part of the drivers of scientific production. Thus, determinants relating to the organization in which the researcher operates have to be considered. Allison and Long (1990) highlight the role of prestigious departments in encouraging individual scientific productivity. The composition of laboratories or departments in terms of type of researchers (full professors, assistant professors, PhD students, etc.) and their average age are also considered in literature (Carayol and Matt, 2006), as well as the quality of colleagues' production (Lissoni et al., 2011; Mairesse and Turner 2006), and the fundraising ability (Carayol and Matt, 2006). At the aggregate level of analysis, some scholars concentrate on the organization size, showing a positive size effect on laboratory productivity (Cainelli et al., 2006; Fabel et al., 2008); others find that small-sized departments are more productive (Carayol and Matt, 2004 & 2006) while others focus on the effects of the composition and average age of the research unit (Bonaccorsi and Daraio, 2003). On one side senior researchers may enhance the productivity of the younger due to co-author works or informal contacts; on the other side, the younger can act as incentives to stimulate the research activities of the older (Carayol and Matt, 2004 & 2006).

3. Material and methods

3.1. Data

In order to address the issues raised in the previous section, our analysis is based on a large original dataset implemented by using either internal administrative sources or external sources. The data refer to the University of Siena (Italy), one of the oldest (established in 1240) publicly-funded universities in Italy. It is a medium-sized university with about 17,000 students covering 8 scientific areas: Arts and Humanities, Economics, Engineering, Law, Mathematical, Physical and Natural Sciences, Medicine and Surgery, Pharmacy, Political Sciences. The dataset is composed by 942 individual records referred to permanent researchers of the University of Siena. For each researcher, we collected information dealing with personal characteristics, research activity, teaching activity, administrative tasks and departmental affiliation. The composition of such a database allows us to assess the determinants of researchers' publication productivity.

3.1.1 Response variables

In order to quantify publication production, three different response variables are considered:

i. number of publications in Anagrafe.UNISI, denoted by *repository*. The Anagrafe.UNISI is the research repository of the University of Siena. It is used by researchers in order to record all their outlet and therefore it is filled with a broad range of outputs: scientific publications, teaching materials and informative articles - such as newspapers and magazine articles. The variable *repository* is constructed by counting - for each scholar - the number of scientific outputs recorded in the repository during the period 2008-2010 and classified as articles, books, chapters in books, conference proceedings;

ii. number of publications in Scopus, the bibliographic database developed by Elsevier (http://www.scopus.com), denoted by *scopus*. This variable is the number of publications in the database referred to the period 2008-2010. In order to get this information for each researcher, the last and first names were queried in the Scopus Author search. If the author's name was not unique, the results were refined to ensure that the correct publications were attributed to the researcher. The database was accessed in July 2011;

iii. h-index score, described by the variable *h_index*. This index - introduced by Hirsh (2005) - combines the author's article count and citation count into a single value. The value of the h-index was extracted from the Scopus database and refers to December 31st, 2010. The database was accessed in July 2011 using the same procedure adopted for the variable *scopus*.

It must be pointed out that the variable *scopus* supplies a measure of the outcome of research production more rigorous than the variable *repository*, since in the Scopus database solely peer-reviewed literature appears. On the contrary, the h-index is an indicator mixing a proxy of scientific production (i.e. the number of publications) and research impact (i.e. the number of received citations). Moreover, as stressed by Hirsch (2005, p.16572), the h-index "... gives an estimate of the importance, significance, and broad impact of a scientist's cumulative research contributions ...". Figure 1 displays the box-plots of the three variables.

Finally, it is worth noting that, since the h-index completely ignores the issue of co-authorship, for the variables *repository* and *scopus* we collected in turn the number of publications for each researcher without considering the co-author number.

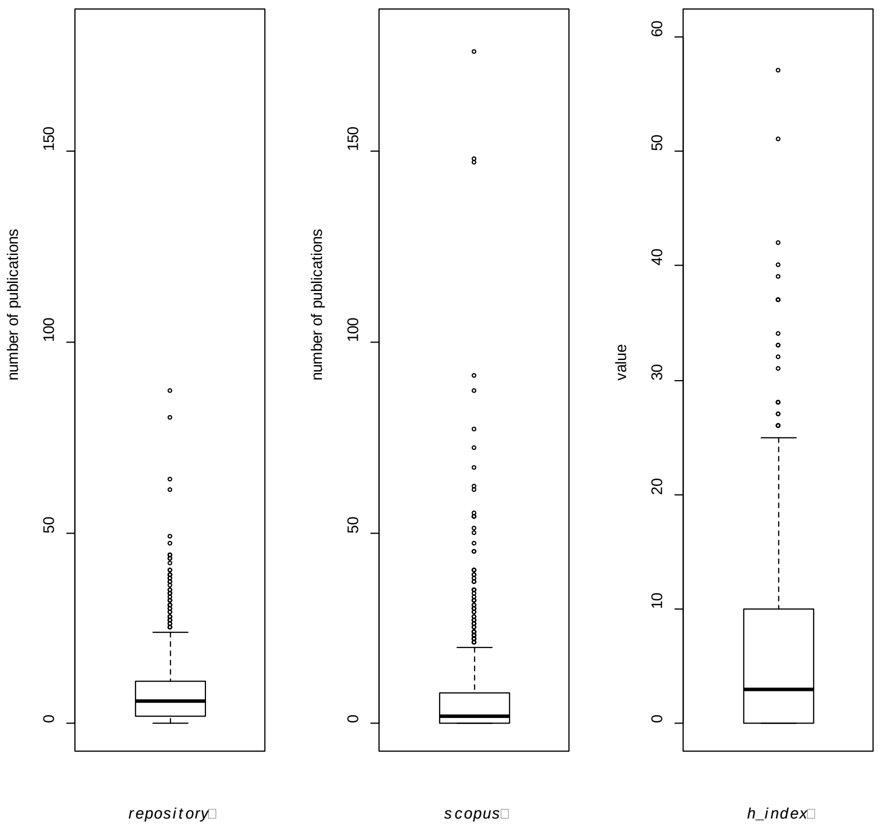

Figure 1. Box-plots of the number of publications in the research repository of the University of Siena (*repository*), of the number of publications in Scopus database (*scopus*) and of the h-index value (*h_index*).

3.1.2 Explanatory variables

The explanatory variables are collected from internal administrative sources and they are aimed to highlight the individual determinants of academic productivity in terms of the features of the researchers, i.e. their personal characteristics, their teaching activities, their participation to the university governance and the research working context. Among the variables describing researcher personal characteristics, gender (indicated as the variable *gender*) is the first considered. The database also includes the age of the researcher at December 31st, 2010. In order to achieve a better understanding of the age effect on production, we considered age quartiles which turned out to be 45.93, 53.77 and 60.80 years. Subsequently, each researcher was assigned to one of the four groups defined by the quartiles according to her/his age (say, the variable *age*). Moreover, the number of years since she/he got a permanent position (indicated as the variable *tenure*) and her/his position (indicated as the variable *position*) are considered. The latter variable gives the professional role in the University of Siena at December 31st, 2010, as contemplated by the Italian academic system: “ricercatore universitario” (assistant professor), “professore associato” (associate professor) and “professore ordinario” (full professor). Research and teaching activities, as well as wages, are defined by national legislation. Also the recruitment and career promotion system, from assistant professor to associate or full professor, and from associate to full professor, are ruled by national laws and based - at least on principle - on scientific productivity. Moreover, to take into account that Italian legislation allows for choosing between full-time or part-time academic positions, the variable *full-time* (providing this information at December 31st, 2010) is considered. Part-time researchers have a teaching load of about one-half with respect to full-time researchers. It is mandatory that researchers involved in private practice (such as lawyers, engineers or doctors, as an example) have a part-time contract.

Since publishing activities and citation practices are deeply different across disciplines, we consider the sector of activity of the researcher. To this aim, we re-classified the scientific disciplines, as defined in the Italian academic system, into the European Research Council Sectors in such a way to give rise to the variable *erc*, which categorizes the disciplines in Life Sciences (LS), Physical and Engineering Sciences (PE), and Social Sciences and Humanities (SH).

Prin projects - acronym of “Progetti di Ricerca di Interesse Nazionale” - are three-years research projects granted by the Italian Ministry of University and Research on a competitive basis, open to all disciplines. Each project is evaluated by peer reviewers. Projects receiving positive evaluation are grouped on scientific fields and ranked and solely a small minority of them are funded. The dummy variable *prin* indicates if a researcher has been involved at least in one positive-evaluated Prin project during the period 2008-2010. Thus, the variable *prin* is a proper proxy of the ability of the researcher to be involved in well-evaluated projects, rather than a proxy of financial resources available for her/his research.

In order to tackling the question if teaching and research are complementary or conflicting activities, we gathered information on researchers teaching activity through three variables. The variable *teaching* is the average number of teaching hours per month during the period 2008-2010, computed excluding months on leave. The variable *thesis* is the number of bachelor and master dissertations supervised by each researcher in the period 2008-2010, while the variable *students* is the average number of students attending lectures in the same period. Similarly, we would assess if the time devoted to the governance could badly affect the individual productivity. To this purpose, we introduce the variable *presence_faculty_meeting*, i.e. the proportion of faculty meetings attended by a researcher during the period 2008-2010. This variable can be considered as a rough proxy of the institutional duties of a researcher. Finally, we wonder if research productivity is affected by the research context and, particularly, by the characteristics of the department staff, i.e. the so-called "departmental effect". In order to contemplate the department composition, we include two variables, denoted by *young_ratio* and *taw_ratio*. The former variable represents the ratio of the number of PhD students and research fellows to the number of permanent researchers (assistant, associate and full professors). The latter variable is the ratio of the number of non-research staff units (i.e. technical and administrative workers) to the number of permanent researchers. Table 1 reports the definition of the explanatory and response variables adopted in the analysis, as well as some corresponding descriptive statistics.

3.2 Method

From Figure 1, it is at once apparent that the distributions of the three response variables show the presence of a remarkable number of zeros and a high level of skewness. When dealing with modelling skewed count data with an excess of zeroes, it is well known that the usual Poisson (P) regression can be inappropriate. Indeed, the data tend to exhibit over-dispersion, i.e. a larger variance than that predicted by the mean, and a large number of zero counts. In order to address over-dispersion, the Negative Binomial (NB) regression can be alternatively used. However, when the major source of over-dispersion is related to a preponderance of zero counts, more flexible count data models - such as zero-inflated or hurdle models - have to be adopted.

Table 1. Definition of variables. Mean, median, minimum and maximum values, Standard Deviation (S.D.) are reported for quantitative variables; relative frequencies are reported for categorical variables.

Response variables	Description	Mean	Median	Min.	Max.	S.D.
repository	number of publications in Anagrafe.UNISI	8.47	6.00	0	87.00	1.11
scopus	number of publications in Scopus database	6.45	2.00	0	176.00	2.11
h_index	h- index	6.12	3.00	0	57.00	1.28
Quantitative explanatory variables	**Description**	**Mean**	**Median**	**Min.**	**Max.**	**S.D.**
tenure	years in permanent position	18.11	18.00	2.00	42.00	0.58
teaching	average number of hours of teaching per month worked	7.39	7.22	0	52.00	0.53
thesis	number of thesis as supervisor	2.60	1.30	0	44.00	1.51
students	number of students attending lectures	35.16	26.50	0	415.00	0.98
presence_faculty_meeting	proportion of presence to the Faculty meeting	0.50	0.55	0	1.00	0.55
young_ratio	ratio of PhD students and research fellows and permanent researchers	1.09	1.11	0	2.21	0.47
taw_ratio	ratio of technical and administrative workers and permanent researchers	0.51	0.43	0.08	1.41	0.71
Categorical explanatory variables	**Description**	**Frequency**				
age	age of the scholar					
1	age ≤45.93	0.25				
2	45.93<age ≤53.77	0.25				
3	53.77<age ≤60.80	0.25				
4	age>60.80	0.25				
gender	gender					
M	male	0.66				
F	female	0.34				
position	role at the university					
AR	assistant professor	0.41				
AP	associate professor	0.29				
FP	full professor	0.30				
erc	European Research Council Sectors					
LS	Life Sciences	0.29				
PE	Physical and Engineering Sciences	0.28				
SH	Social Sciences and Humanities	0.43				
prin	positive evaluation in Prin project					
0	no positive evaluation	0.26				
1	yes positive evaluation	0.74				
full-time	full-time service to the university					
0	no full-time	0.06				
1	yes full-time	0.94				

3.2.1 Poisson and Negative Binomial models

The Poisson regression is the simplest regression model for count data, assuming a log-linear dependence between the mean of a Poisson distribution and a set of covariates (see e.g. McCullagh and Nelder, 1989). More precisely, let us assume that Y be the random variable (r.v.) representing the response variable and that $Y_1,\ldots,Y_n$ be a sample of *n* stochastically independent counts. In the Poisson regression, the r.v. Y has a Poisson distribution with parameter $\lambda(z)$, where z denotes a suitable covariate vector. If z_i denotes the value of the covariate vector for the *i*-th individual, the r.v. Y_i has probability function (p.f.) given by

$$P(Y_i = k \,|\, z_i) = \exp(-\lambda(z_i))\frac{\lambda(z_i)^k}{k!}, \quad k = 0,1,\ldots \ .$$

Hence, by adopting log-linear regression, the model assumes that

$$\log\lambda(z_i) = z_i{}^T\gamma ,$$

where γ denotes an unknown parameter vector. It must be pointed out that assuming a Poisson distribution has several implications which cannot be neglected: first, the expectation and variance are equal, even if count data often exhibit over-dispersion. Alternatives - overcoming this violation - include the Poisson regression with over-dispersion parameter or the use of the Negative Binomial regression (Hilbe, 2011). Under the Negative Binomial model, the r.v. Y has a Negative Binomial distribution with parameter $\lambda(z)$, i.e. the p.f. of the r.v. Y_i is given by

$$P(Y_i = k \,|\, z_i) = \frac{\Gamma(k+\theta)}{\Gamma(\theta)\Gamma(k+1)}\left(\frac{\lambda(z_i)}{\lambda(z_i)+\theta}\right)^k\left(\frac{\theta}{\lambda(z_i)+\theta}\right)^\theta, \quad k = 0,1,\ldots,$$

where θ represents the dispersion parameter, while $\Gamma(\cdot)$ represents the usual Euler's Gamma function. In this case, similarly to the Poisson model, the mean is $\lambda(z_i)$, while the variance is given by $\lambda(z_i)+\theta^{-1}\lambda(z_i)^2$, thus allowing the modelling of over-dispersed data. In turn, log-linear regression is commonly adopted to describe the dependence with the covariate vector.

A second issue deals with the observed number of zero counts, which highly exceeds the predicted number. This drawback should be addressed by means of more refined count data models, such as zero-inflated or hurdle models.

3.2.2 Zero-inflated models

In the zero-inflated models, the response variable is modelled as a mixture of a Dirac mass at zero and an integer-valued distribution - usually referred to as the count component. Among them, the most widely applied model is arguably the Zero-Inflated Poisson model (ZIP) (see e.g. Bohning et al., 1999; Dalrymple et al., 2003; Hall, 2000; Lambert, 1992; Rathbun and Fei, 2006). In the ZIP regression, the response variable Y is assumed to be zero with probability $p(x)$ - where x denote a further covariate vector - while it displays a Poisson distribution of parameter $\lambda(z)$ with probability $1-p(x)$. Thus, If x_i denotes the value of the covariate vector for the *i*-th individual, the p.f. of the r.v. Y_i is given by

$$P(Y_i=0|x_i,z_i)=p(x_i)+(1-p(x_i))\exp(-\lambda(z_i))$$

and

$$P(Y_i=k|x_i,z_i)=(1-p(x_i))\exp(-\lambda(z_i))\frac{\lambda(z_i)^k}{k!},\quad k=1,2,\dots.$$

From the previous expressions, it is at once apparent that - as is typical in zero-inflated setting - zero counts arise from the Dirac distribution describing the excess of zero counts as well as from the Poisson distribution modelling the remaining zero and non-zero counts. Thus, two regression expressions are actually considered: a logistic regression predicting whether an “inflated” zero occurs and a log-linear regression predicting the remaining zero and non-zero counts, i.e.

$$\log\frac{p(x_i)}{1-p(x_i)}=x_i^T\beta$$

and

$$\log\lambda(z_i)=z_i^T\gamma\ ,$$

where β and γ denote parameter vectors to be estimated. However, count data often exhibit a high variability, which precludes the use of a Poisson distribution - even after accounting for inflated zeros. Subsequently, an alternative way of modelling such a data is to assume a Zero-Inflated Negative Binomial model (ZINB) (see e.g. Minami et al., 2007; Rose et al., 2006; Zhang et al., 2012) for the r.v. Y. In this case, the probability of obtaining an inflated zero is in turn given by

$p(x)$, even if the r.v. Y shows a Negative Binomial distribution of parameters $\lambda(z)$ and θ with probability $1-p(x)$. Hence, the model reduces to

$$P(Y_i = 0|x_i, z_i) = p(x_i) + (1-p(x_i))\left(\frac{\theta}{\lambda(z_i)+\theta}\right)^{\theta}$$

and

$$P(Y_i = k|x_i, z_i) = (1-p(x_i))\frac{\Gamma(k+\theta)}{\Gamma(\theta)\Gamma(k+1)}\left(\frac{\lambda(z_i)}{\lambda(z_i)+\theta}\right)^{k}\left(\frac{\theta}{\lambda(z_i)+\theta}\right)^{\theta}, \quad k = 1,2,\dots,$$

where the parameter $\lambda(z_i)$ and θ respectively denote the mean number of counts and the dispersion parameter. Similarly to the case of ZIP, both the logistic and the log-linear regressions are considered for predicting “inflated” zeros and the remaining counts.

3.2.3 Hurdle models

The hurdle models, originally introduced by Mullahy (1986), are two-component models: the first component is constituted by a Dirac distribution at zero, while the second component - i.e. the count component - is a truncated integer-valued distribution modelling strictly positive counts. In the framework of the Poisson distribution, the Hurdle Poisson model (HP) postulates that the r.v. Y is equal to zero with probability $p(x)$, while with probability $1-p(x)$ has a truncated Poisson distribution of parameter $\lambda(z)$, i.e. the p.f. of the r.v. Y_i is given by

$$P(Y_i = 0|x_i) = p(x_i)$$

and

$$P(Y_i = k|x_i, z_i) = (1-p(x_i))\frac{\exp(-\lambda(z_i))}{1-\exp(-\lambda(z_i))}\frac{\lambda(z_i)^k}{k!}, \quad k = 1,2,\dots.$$

Alternatively, when dealing with a marked variability, the count component can be modelled by means of a truncated Negative Binomial distribution of parameters $\lambda(z)$ and θ, in such a way to obtain the Hurdle Negative Binomial model (HNB) (Dalrymple at al., 2003; Zhang et al., 2012), i.e.

$$P(Y_i = 0|x_i) = p(x_i)$$

and

$$P(Y_i = k \left| x_i, z_i \right.) = (1 - p(x_i)) \frac{\Gamma(k+\theta)}{\Gamma(\theta)\Gamma(k+1)} \left(\frac{\lambda(z_i)}{\lambda(z_i)+\theta} \right)^k \left[\left(1 + \frac{\lambda(z_i)}{\theta} \right)^\theta - 1 \right]^{-1}, \quad k = 1,2,\dots .$$

Similarly to the framework of zero-inflated models, $p(x_i)$ and $\lambda(z_i)$ are generally modelled by means of logit and log-linear regression, respectively. It is worth noting that - even if the hurdle model may apparently resembles the zero-inflated model - since they essentially mixture a Dirac mass at zero with a count distribution - their interpretation is rather different. Indeed, hurdle models assume that zero counts can solely arise with probability $p(x_i)$, while under zero-inflated models $p(x_i)$ represents the probability of getting “excess zeros”. More precisely, in the last case zero counts may be obtained from the Dirac distribution as well as from the count component.

3.2.4 Parameters estimation and model selection

In the present case study, the P, NB, ZIP, ZINB, HP, HNB models were considered. The vectors of parameters γ and/or β - and the shape parameter θ when dealing with Negative Binomial distribution - were estimated by means of the maximum-likelihood method. The computational procedures needed for the estimation were carried out by means of the R software (R Development Core Team, 2012). The `glm( )` function (Chambers and Hastie, 1992) in the `stats` package and the `glm.nb( )` function in the `MASS` package (Venables and Ripley, 2002) were adopted with Poisson and Negative Binomial regression respectively, while the functions `zeroinfl( )` and `hurdle( )` in the `pscl` package (Zeileis et al., 2008) were used for dealing with zero-inflated and hurdle regression models. The models were subsequently compared on the basis of the log-likelihood values, as well as on the Akaike Information Criterion (AIC).

4. Results

We estimated the parameters of the zero-inflated and hurdle models for the three response variables - i.e. *repository*, *scopus* and *h_index* - by adopting the same explanatory covariate vectors *x* and *z* for the two components. The log-likelihood and AIC values of the estimated models are reported in Table 2. The analysis of this table highlights that, on one hand, the AIC values of the zero-inflated and hurdle models are rather similar and, on the other hand, the superiority of Negative-Binomial-based models over Poisson-based models is apparent for fitting purposes. Particularly, the lowest AIC values are observed for the ZINB and HNB models.

Table 2. Log-likelihood (log-L) and Akaike's information criterion (AIC) values of the fitted models for the three response variables.

Models	*repository*		*scopus*		*h_index*	
	log-L	AIC	log-L	AIC	log-L	AIC
P	-4404.08	8844.10	-4068.33	8172.70	-2490.77	5017.50
NB	-2861.43	5760.90	-2103.64	4244.10	-2090.73	4219.50
ZIP	-4070.00	8211.71	-3623.00	7317.44	-2265.00	4602.92
ZINB	-2829.00	5731.99	-2037.00	4148.35	-1979.00	4031.34
HP	-4070.00	8212.99	-3623.00	7317.81	-2265.00	4601.11
HNB	-2831.00	5735.59	-2036.00	4145.84	-1981.00	4036.65

These findings are further supported by the predicted frequencies of the models, as is shown in Figures 2, 3 and 4. Indeed, the simple Poisson model cannot account for the large proportion of zero counts, while the Zero-Inflated and Hurdle Poisson models can address this lack of fitting, even if they are not able to predict the nonzero frequencies correctly. The ZINB and HNB models exhibit the best fittings. Moreover, these models not solely produce rather similar AIC values, but - as is common with zero-inflated skewed data (see e.g. Rose et al., 2006; Zhang et al. 2012) - also give rise to similar parameter estimates.

Notwithstanding that zero-inflated and hurdle models are sometimes indistinguishable on the basis of goodness-of-fit statistics, one model type may be more appropriate in order to describe the underlying generating data process. In our framework, zero-inflated models allow for zeros to arise either from "potentially productive" or "unproductive" populations: unproductive researchers can never produce a research outcome, thus giving rise to structural zeros, while those potentially productive can either produce or not produce a research outcome. In contrast, under hurdle models, all the researchers are considered potentially productive so that no structural zeros are assumed, but solely some researchers "pass the hurdle". If a researcher has passed the hurdle producing at least one research outcome, she/he becomes an "active" researcher, and her/his performance is described by the count component of the hurdle model. Since the Italian recruitment system is based on scientific productivity as previously remarked, there is no reason to assume the existence of an unproductive population and hence - in presence of similar goodness-of-fit statistics - the use of hurdle models might be preferred. Thus, in Table 3 we present the parameter estimates, their standard errors and the corresponding significance values when the HNB model is adopted (the same quantities for the other models are available from the authors upon request).

Table 3. Estimated parameters, Standard Errors (S.E.) and Significance (Sig.), log-likelihood (log-L) and Akaike's information criterion (AIC) values for HNB model.

Variables	*repository*			*scopus*			*h_index*		
	Estimate	S.E.	Sig.	Estimate	S.E.	Sig.	Estimate	S.E.	Sig.
Count component of the model									
Intercept	2.672	0.232	***	3.352	0.331	***	2.625	0.202	***
*age*_1	Ref.								
*age*_2	-0.168	0.093	.	0.102	0.132		0.114	0.086	
*age*_3	-0.504	0.123	***	-0.484	0.189	*	-0.213	0.114	.
*age*_4	-0.501	0.153	**	-0.542	0.235	*	-0.468	0.144	**
*gender*_F	Ref.								
*gender*_M	0.141	0.067	*	0.223	0.098	*	0.031	0.061	
*position*_FP	Ref.								
*position*_AP	-0.478	0.084	***	-0.709	0.123	***	-0.493	0.076	***
*position*_AR	-0.991	0.100	***	-1.281	0.146	***	-0.900	0.089	***
*erc*_LS	Ref.								
*erc*_PE	-0.043	0.088		-0.192	0.102	.	-0.285	0.061	***
*erc*_SH	-0.305	0.099	**	-2.089	0.177	***	-1.795	0.124	***
tenure	-0.009	0.005	.	-0.020	0.008	**	-0.003	0.005	
teaching_per_month	-0.015	0.010		-0.001	0.012		-0.002	0.007	
thesis	0.033	0.008	***	0.015	0.016		-0.010	0.011	
students	0.002	0.001	*	-0.003	0.001	*	-0.002	0.001	*
*prin*_0	Ref.								
*prin*_1	0.262	0.080	***	0.753	0.120	***	0.396	0.073	***
*full-time*_0	Ref.								
*full-time*_1	-0.048	0.149		0.079	0.208		0.001	0.127	
presence_faculty_meeting	0.048	0.125		-0.605	0.174	***	-0.013	0.107	
young_ratio	-0.023	0.067		-0.114	0.094		0.128	0.055	*
taw_ratio	0.432	0.120	***	0.132	0.156		0.264	0.093	**
Log(theta)	0.543	0.076	***	0.308	0.094	***	1.394	0.099	***
Zero component of the model									
Intercept	-2.088	0.781	**	-3.578	0.740	***	-4.088	0.847	***
*age*_1	Ref.								
*age*_2	-0.248	0.429		0.591	0.317	.	1.024	0.352	**
*age*_3	0.111	0.511		0.946	0.403	*	1.533	0.453	***
*age*_4	0.556	0.619		1.216	0.501	*	1.827	0.561	**
*gender*_F	Ref.								
*gender*_M	-0.052	0.265		-0.331	0.223		-0.513	0.256	*
*position*_FP	Ref.								
*position*_AP	0.568	0.350		1.077	0.287	***	1.061	0.320	***
*position*_AR	1.597	0.373	***	1.985	0.340	***	2.160	0.386	***
*erc*_LS	Ref.								
*erc*_PE	1.304	0.392	***	1.157	0.351	***	1.158	0.494	*
*erc*_SH	1.063	0.417	*	4.742	0.428	***	5.078	0.528	***
tenure	0.046	0.020	*	0.042	0.017	*	0.011	0.019	
teaching_per_month	0.040	0.028		0.015	0.028		0.020	0.040	
thesis	-0.030	0.039		0.007	0.025		0.017	0.027	
students	-0.001	0.005		-0.003	0.004		-0.002	0.004	
*prin*_0	Ref.								
*prin*_1	-1.291	0.261	***	-1.082	0.267	***	-1.009	0.302	***
*full-time*_0	Ref.								
*full-time*_1	-0.760	0.422	.	-0.768	0.400	.	-1.311	0.476	**
presence_faculty_meeting	-1.370	0.438	**	-0.119	0.385		1.219	0.428	**
young_ratio	-0.433	0.239	.	0.108	0.207		0.224	0.238	
taw_ratio	-0.954	0.478	*	-0.768	0.413	.	-1.824	0.556	**
log-L	-2831.000			-2036.000			-1981.000		
AIC	5735.591			4145.838			4036.646		

. Significant at 10%
* Significant at 5%
** Significant at 1%
*** Significant at 0.1%

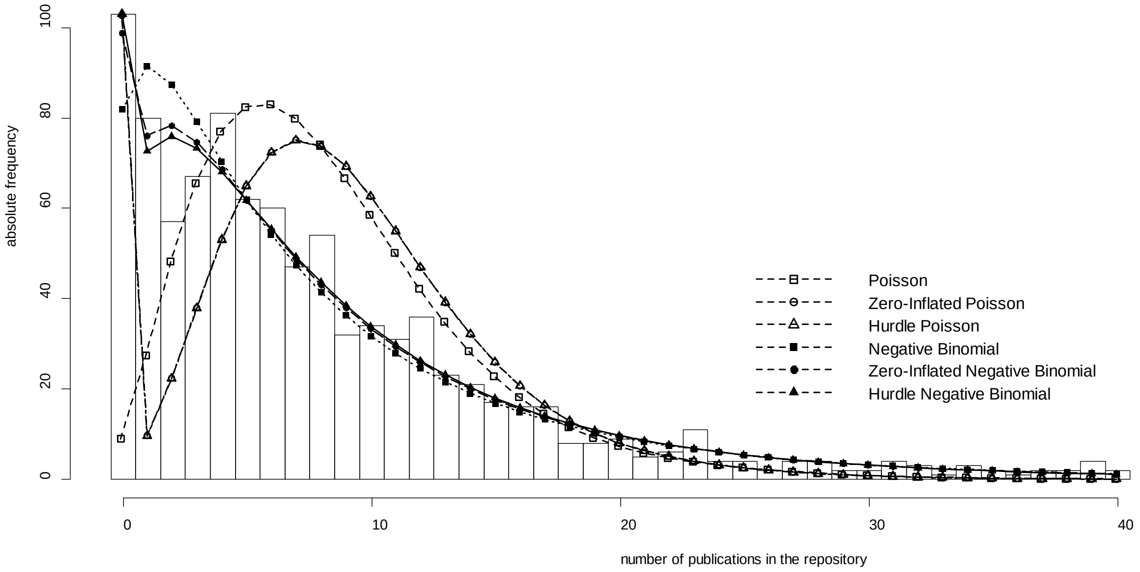

Figure 2. The observed distribution of the number of publications in the research repository of the University of Siena along with the predicted distributions obtained using the P, NB, ZIP, ZINB, HP and HNB models. Values greater than 40 are omitted for improving readability.

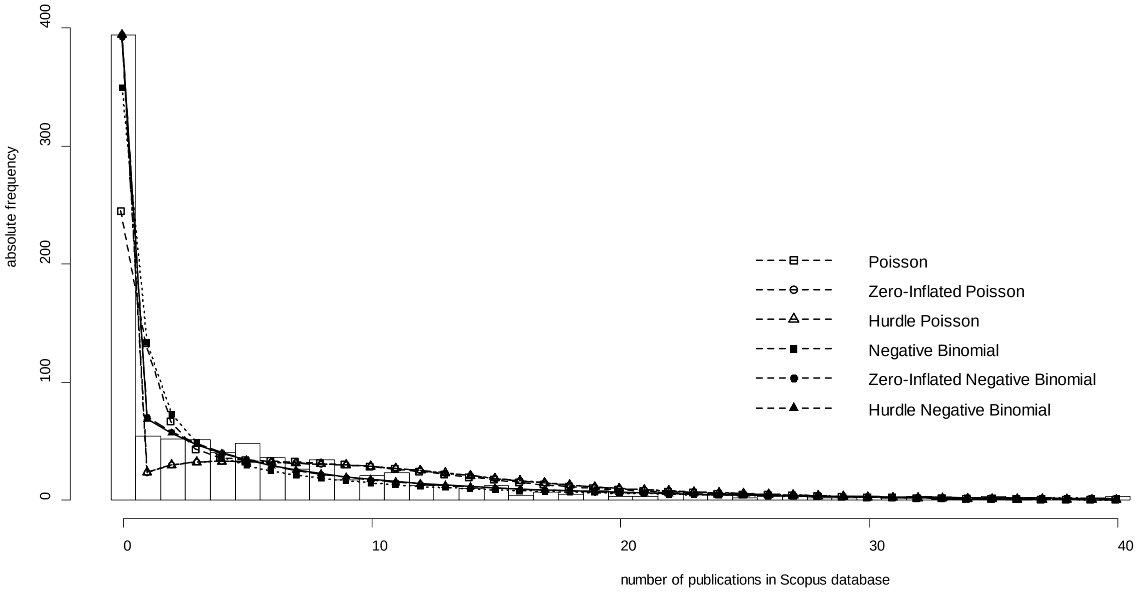

Figure 3. The observed distribution of the number of publications in Scopus along with the predicted distributions obtained using the P, NB, ZIP, ZINB, HP and HNB models. Values greater than 40 are omitted for improving readability.

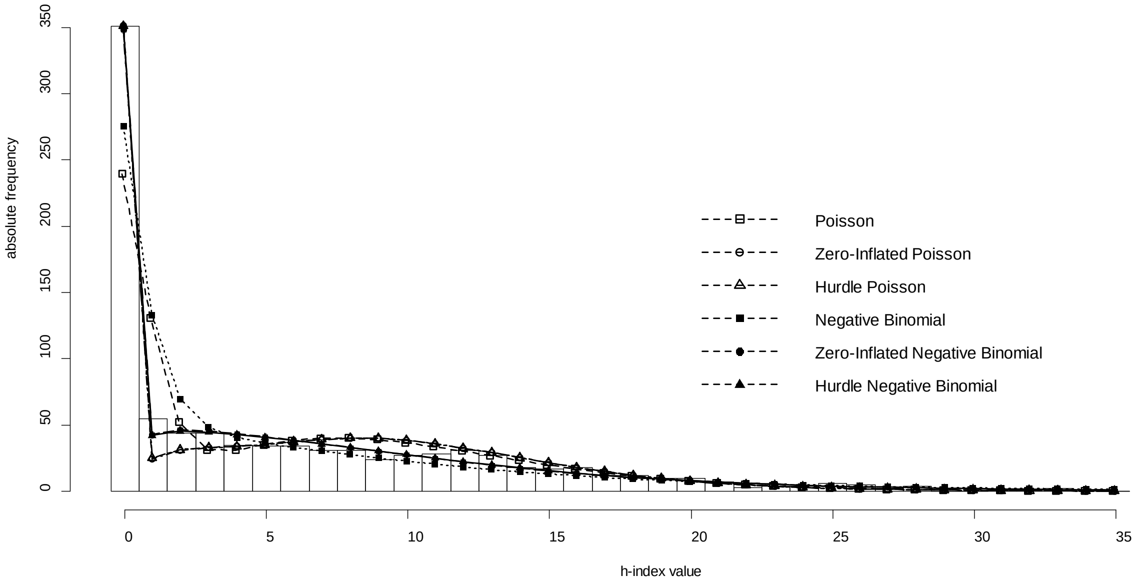

Figure 4. The observed distribution of the h-index value along with the predicted distributions obtained using the P, NB, ZIP, ZINB, HP and HNB models. Values greater than 35 are omitted for improving readability.

The analysis of Table 3 motivates the following conclusions. *Age* has a different effect on the three measures of research output. By considering the number of publications in the repository, this variable is significantly and negatively affected by age for the active researchers, i.e. those passing the hurdle. This result especially holds for the researchers whose age is greater than the median. Similar - even if less significant - conclusions also hold when the number of publications in Scopus is considered. Such findings are not in contrast with the hypothesis of an inversed U-shape productivity-age relation. As to the h-index value, the negative effect of age on the active researchers turns out to be significant solely for the oldest scholars. Moreover, the variable *age* has a strong impact on the probability of having a zero h-index value, i.e. the estimated odds become larger as the researcher age increases. Considering the cumulative nature of the h-index as a production measure, this finding is slightly puzzling. The issue might be interpreted as the result of two connected processes. The first refers to the different publication strategies (probably) adopted by older and younger scholars, respectively. The younger researchers are more sensible to the "publish-or-perish" pressure and they usually pay more attention to the outlet in which their works are published. In such a case, the choice of journals - where articles are submitted - is driven by the Impact Factor or by other similar indicators. Journals indexed in bibliographic databases, such as Scopus, are usually preferred with respect to non-indexed journals. The second process refers to the coverage of the Scopus database. Older researchers - particularly the oldest - have publications dating back up to 35-45 years and which could have been appeared in journals not indexed in the Scopus database. Moreover, citations to oldest articles could not be recorded in Scopus.

The gender effect is moderately significant, concerning all the research production measures. In fact, *gender* is slightly significant in the case of *repository* and *scopus* with respect to the count component and for the variable *h_index* when considering the zero component - suggesting that women face with more difficulties than men in publishing.

As to the academic position, its effect is highly significant for all the response variables. In particular, referring to the model count component, the number of publications and h-index value of the active researchers decrease for associate professors and assistant professors with respect to full professors. Moreover, as is evidenced by the zero component of the model, the academic position has a marked effect even on the probability to be a non-active researcher. Indeed, when the variable *h_index* is considered, the odds of having a zero value for associate professors and assistant professors are approximately 3 and 9 times the odds of the full professors, respectively. In turn, similar results hold for the variables *scopus*, while for the variable *repository* the effect is significant on the odds only for assistant professors.

As expected, the scientific sector of activity significantly impact on research performance: the researchers who have passed the hurdle and belong to the LS sector show a higher level of production than those belonging to the SH sector. This result also holds for the PE sector for *h_index* and - with a lower significance level - *scopus* variables. Analogously, in the zero component part of the models, the odds of the non-active status is significantly smaller for the researcher of LF sector. However, such results should be cautiously interpreted owing to the different coverage of the Scopus database, as well as the different publication and citation patterns in the sectors (Iglesias and Pecharroman, 2007). As is well known, the Scopus database has a weaker coverage for SH, especially for non-English language countries, such as Italy. In the SH sector, research results are communicated to the academic community mainly by means of books and chapters in books. Subsequently, for this sector, the bibliographic databases, such as Scopus, are largely incomplete in terms of publications and citations (Hicks, 2004). It is also worth remembering that co-authorship patterns are very different across scientific sectors. It can be argued that for sectors where articles have usually dozens of authors, the probability to be non-active is lower than in other sectors - such as SH - where groups of co-authors are very small and single authorship often prevails.

As to the connection between teaching and research activities, when the zero component of the model is considered, teaching activities do not seem to affect the probability to be a non-active researcher. However, if active researchers are considered, it is worth noting that, for the three explanatory variables adopted to proxy teaching tasks (i.e. *teaching_per_month*, *thesis* and *students*), only *students* weakly significantly affects the research performance in terms of a weak decrease in the number of publication in the Scopus and h-index value. On the contrary, when the number of publications in the repository is considered, there is a weak evidence that the variable *students* positively influences the research output and a strong evidence that the number of thesis increases the number of publications. It may be argued that the results achieved in the thesis can be used by supervisors to produce a research outcome suitable to appear in the repository, but not in peer-reviewed journals indexed by the Scopus database.

As to the connection between the administrative duties of a researcher and her/his research activities, we found that the participation to Faculty meetings has a significant negative effect on the number of publications in the Scopus database for active researchers. Moreover, it increases the odds of having zero h-index value, while significantly reduces the odds of having zero publications on the research repository. The mixed evidence of our analysis does not allow to draw a clear-cut conclusion about the relation between teaching and research activities, while administrative tasks appear to negatively affect research productivity, especially when research outcomes filtered by a reviewing process are considered.

A positive evaluation received for the Prin projects significantly increases both the expected output of active researchers (as highlighted by the positive coefficient estimates in the count component of the model) and the probability of passing the hurdle (as shown by the negative sign of the estimates in the zero part of the model).

A part-time contract significantly increases the probability to be non-active, with the most marked effect when the variable *h_index* is considered. However, it is worth noting that, among the active researchers, a part-time contract does not significantly affect the research performance.

By considering the collective level of department organization, the ratio between young and permanent researchers (*young_ratio*) seems to have a moderately significant positive effect solely for the *h_index* variable in the count part of the model, while an increase in the ratio between administrative workers and researchers (*taw_ratio*) has a significant positive effect on the production of active researchers when *repository* or *h_index* variables are considered and a significant effect on the reduction of the odds to be a non-active researcher.

5. Discussion and conclusions

The analysis of research production has attracted a growing interest in the recent decades. We are contributing to the stream of literature investigating the determinants of researcher performance. We analyze original data referring to a medium-sized Italian university which employs 942 researchers covering many scientific fields in Life Sciences, Physical and Engineering Sciences and Social Sciences and Humanities. Our analysis contributes with respect to previous literature on scientific production by adopting models suitable to manage count data with a zero excess, which are often the main features of the response variables adopted to quantify research production. The currently-used models impose artificial smoothness to data. In contrast, we handle data using zero-inflated and hurdle models. Among them, the Hurdle Negative Binomial exhibits a good fitting and appears to be reasonably coherent with the underlying generating data process. In this case, all the researchers are considered as potentially productive and when a researcher passes the hurdle - writing a paper and becoming active - her/his performance is described by the count component of the model. Moreover, the odds that a researcher be non-active is modelled by the zero component of the model.

In order to highlight the determinants of a researcher production, we consider her/his individual characteristics, her/his research, teaching and administrative activities and the features of the environment which she/he belongs to.

Our models found that scientific productivity is lower for oldest active researchers. This result is in agreement with the life-cycle hypothesis, and with the prevailing evidence reported in the previous literature. We also found that the probability to be non-active in terms of h-index significantly increases with age. This result is slightly puzzling, given that h-index is a cumulative indicator of research production and impact. We suggest to interpret this evidence in terms of a lower coverage of the Scopus database for older papers and of a change in the publication/editorial strategies across generations of scholars. Younger researchers are probably more attentive and sensible than older scholars to the selection of the journals where their outlet is going to be published.

Moreover, there is a significant effect of academic position on research production: researchers with higher position are more productive. At a first sight, this finding is coherent with organizational characteristics of the Italian university system, in which promotions are mainly based on research activities. Moreover, the still-surviving hierarchical organization of the university allows for full professors to act as both coordinators of national and international projects - which generally give rise to many co-authored publications - and supervisors for PhD students and junior researchers - who can stimulate their production.

As to the effects of teaching activities on research productivity, we do not find evidence of clear-cut and significant substitution or complementarity effects. In particular a major teaching load in terms of teaching hours, number of students and thesis supervised does not affect the odds to be a non-active researcher. Indeed, our finding tends to reject the hypothesis that the odds to be non-active is affected by the crowding out effect between teaching and publications. Analogous conclusions have been achieved - using a completely different approach - by Kelchtermans and Veugelers (2011). The same variables have mixed and very weak effects on publication performance of active researchers. Indeed, the number of students attending lessons tends to positively affect the number of publications recorded in the repository - which, it is worth to remember, includes any type of publication - and to reduce the number of publications recorded in Scopus database and the h-index value. The number of theses supervised has a significant positive effect solely on the number of publication in the repository. Our results suggest that neither substitution nor complementarity simple hypotheses seem to adequately represent the multifaceted relation among research and teaching activities.

As to the features of the department a researcher belongs to, it is worth noting that the relative number of PhD students and research fellows does not seem to impact the odds to be non-active. In contrast, it can have a positive effect on the h-index when the active researchers are considered. This effect can be explained by two, probably interrelated, processes. On one hand, the more productive researchers act as supervisors for the young non-permanent researchers, thus enhancing the probability to publish co-authored papers. On the other hand, the presence of productive PhD students widen the citation network of permanent staff improving the number of citations received by the latter.

The composition of the department, in terms of the ratio of the number of non-research staff units to the number of permanent researchers, affects both the odds to be non-active and the value of the production indicators, thus evidencing a positive effects of administrative resources availability on scientific productivity. In this case, we can suppose that administrative staff conveniently help a researcher in administrative and bureaucratic tasks connected with teaching and with the management of projects and research activities. Moreover, we found a negative relation among productivity and administrative tasks, as proxied by the presence to Faculty meetings. These results seem to suggest that researchers productivity is negatively affected by bureaucratic and administrative tasks - a topic not covered in previous literature and deserving more scrutiny.

Finally, we caution against generalizations of our results since the adopted database include scientists working in a single university and we are aware that more general country-level investigation on the main determinants of scientific production should be undertaken.

We believe that further research is needed in order to create improved measures of the effort devoted to institutional duties and university governance to better understand if and how these activities could affect scientific production. Similarly, additional proxies for the teaching load have to be exploited. Indeed, the evaluation of the effects of administrative and teaching tasks on scientific output is mandatory to verify if current incentive policies for stimulating research production are effective. To this purpose, future research could surely benefit from the availability of information - at national level - concerning researcher's final outputs (publications, patents, products, etc.), but also projects, scientific areas, teaching and administrative and institutional activities.

Our future research could address different issues emerging from this paper, such as the relationship of research production with career progress and academic collaborations or the effects of co-authorship when evaluating scientific research performance.

References

Abramo, G., D’Angelo, C.A., Caprasecca, A., 2009. Gender differences in research productivity: a bibliometric analysis of the Italian academic system. Scientometrics 79, 517-539.

Allison, P.D., Long J.S., 1990. Departmental effects on scientific productivity. American Sociological Review 55, 469-478.

Allison, P.D., Stewart, J.A., 1974. Productivity differences among scientists: evidence for accumulative advantage. American Sociological Review 39, 596-606.

Anania, G., Caruso, A., 2013. Two simple new bibliometric indexes to better evaluate research in disciplines where publications typically receive less citations. Scientometrics 96, 617-631.

Bayer, A.E., Dutton, J.E., 1977. Career age and research professional activities of academic scientists. Journal of Higher education 48, 259-282.

Böhning, D., Dietz, E., Schlattmann, P., Mendonça, L., Kirchner, U., 1999. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. Journal of the Royal Statistical Society Series A 162, 195-209.

Bonaccorsi, A., Daraio, C., 2003. Age effects in scientific productivity. The case of the Italian National Research Council (CNR). Scientometrics 58, 49-90.

Cainelli, G., de Felice, A., Lamonarca, M., Zoboli, R., 2006. The publications of Italian economists in ECONLIT: quantitative assessment and implications for research evaluation. Economia Politica 23, 385-423.

Carayol, N., Matt, M., 2004. Does research organization influence academic production? Laboratory level evidence from a large European university. Research Policy 33, 1081-1102.

Carayol, N., Matt, M., 2006. Individual and collective determinants of academic scientists’ productivity. Information Economics and Policy 18, 55-72.

Chambers, J.M., Hastie, T.J. (Eds), 1992. Statistical Models in S. Chapman & Hall, London.

Cole, J.R., Cole, S., 1973. Social Stratification in Science. University of Chicago Press, Chicago.

Cole, J.R., Zuckerman, H. (1984). The productivity puzzle: persistence and change in patterns of publication of men and women scientists, in: Steinkempt, M.W., Maehr, M.L. (Eds), Advances in Motivation and Achievement. JAI Press, Greenwich, Conn., pp. 217-258.

Dalrymple, M.L., Hudson, I.L., Ford, R.P.K., 2003. Finite mixture, zero-inflated Poisson and hurdle models with application to SIDS. Computational Statistics & Data Analysis 41, 491-504.

David, P. (1994). Positive feedbacks and research productivity in science: reopening another black box, in: Grandstrand, O. (Ed), Economics and Technology. Elsevier, Amsterdam, pp. 65-85.

Defazio, D., Lockett, A., Wright, M., 2009. Funding incentives, collaborative dynamics and scientific productivity: evidence from the EU framework program. Research Policy 38, 293-305.

Diamond, A.M., 1984. An economic-model of the life-cycle research productivity of scientists. Scientometrics 6, 189-196.

Diamond, A.M., 1986. The life-cycle research productivity of mathematicians and scientists. The Journal of Gerontology 41, 520-525.

Fabel, O., Hein, M., Hofmeister, R., 2008. Research productivity in business economics: an investigation of Austrian, German and Swiss universities. German Economic Review 9, 506-531.

Fox, M.F., 1992. Research, teaching and publication productivity: mutuality versus competition in academia. Sociology of Education 65, 293-305.

Fox, M.F., 2005. Gender, family characteristics, and publication productivity among scientists. Social Studies of Science 35, 131-150.

Fox, M.F., Fonseca, C., Bao, J., 2011. Work and family conflict in academic science: Patterns and predictors among women and men in research universities. Social Studies of Science 41, 715-735.

Hall, D.B., 2000. Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics 56, 1030-1039.

Hicks, D. (2004). The four literatures of social science, in: Moed, F.H., Glaenzel, W., Schmoch, U. (Eds), Handbook of Quantitative Science and Technology Research. Kluwer Academic Publishers, Dordrecht, Boston and London, pp. 473-496.

Hilbe, J. 2011. Negative Binomial Regression. Cambridge University Press, Cambridge.

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.

Iglesias, J.E., Pecharroman, C., 2007. Scaling the h-index for different scientific ISI fields. Scientometrics 73, 303-320.

Kelchtermans, K., Veugelers, R., 2011. The great divide in scientific productivity: why the average scientist does not exist. Industrial and Corporate Change 20, 295-336.

Kossi, Y., Lesueur, J.Y., Sabatier, M., 2013. Publish or teach? The role of the scientific environment on academics multitasking. Groupe d’Analyse et de Théorie Économique Lyon-St Étienne, GATE WP 1315.

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34, 1-17.

Leahey, E., 2006. Gender differences in productivity: research specialization as a missing link. Gender & Society 20, 754-780.

Levin, S., Stephan, P.E., 1991. Research productivity over the life cycle: evidence for academic scientists. American Economic Review 81, 114-132.

Levin, S., Stephan, P.E., 1998. Gender differences in the rewards to publishing in academe: science in the 1970s. Sex Roles 38, 1041-1064.

Lissoni, F., Mairesse, J., Montobbio, F., Pezzoni, M., 2011. Scientific productivity and academic promotion: a study on French and Italian physicists. Industrial and Corporate Change 20, 253-294.

Lotka, A.J., 1926. The frequency distribution of scientific productivity. Journal of the Washington Academy of Science 16, 317-323.

Mairesse J., Pezzoni M., 2013. Does gender affect scientific productivity? A critical review of the empirical evidence and a panel data econometric analysis for French physicists. Presented to AFSE Meeting, Aix en Provence, 26 June 2013.

Mairesse, J., Turner, L. (2006). Measurement and explanation of the intensity of co-publication in scientific research: an analysis at the laboratory level, in: Antonelli, C., Foray, D., Hall, B.H., Steinmueller, W.E. (Eds), New Frontiers in the Economics of Innovation and New Technology: Essays in Honour of Paul A. David. Edward Elgar, Cheltenham and Northampton, pp. 255-295.

Martin, B.R., Irvine, J., 1983. Assessing basic research: some partial indicators of scientific progress in radio astronomy. Research Policy 12, 61-90.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Chapman and Hall, London.

Merton, R., 1968. The Matthew effect in science. Science 159, 56-63.

Minami, M., Lennert-Cody, C.E., Gao, W., Román-Verdesoto, M., 2007. Modeling shark bycatch: the zero-inflated negative binomial regression model with smoothing. Fisheries Research 84, 210-221.

Mullahy, J., 1986. Specification and testing of some modified count data models. Journal of Econometrics 33, 341-365.

Pezzoni, M., Sterzi, V., Lissoni, F., 2012. Career progress in centralized academic systems: social capital and institutions in France and Italy. Research Policy 41, 704-719.

R Core Team, 2012. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, from http://www.R-project.org/.

Rathbun, S.L., Fei, S., 2006. A spatial zero-inflated Poisson regression model for oak regeneration. Environmental and Ecological Statistics 13, 409-426.

Rivera-Huerta, R., Dutrénit, G., Ekboir, J.M., Sampedro, J.L., Vera-Cruz, A.O., 2011. Do linkages between farmers and academic researchers influence researcher productivity? The Mexican case. Research Policy 40, 932-942.

Rose, C.E., Martin, S.W., Wannemuehler, K.A., Plikaytis, B.D., 2006. On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. Journal of Biopharmaceutical Statistics 16, 463-481.

Stephan, P.E., 1996. The economics of science. Journal of Economic Literature 34, 1199-1235.

Stephan, P.E., 2012. How Economics Shapes Science. Harvard University Press, Cambridge, Mass.

Taylor, S.W., Fender, B.F., Burke, K.G., 2006. Unraveling the academic productivity of economists: the opportunity costs of teaching and service. Southern Economic Journal 72, 846-859.

van Arensbergen, P., van der Weijden, I., van den Besselaar, P., 2012. Gender differences in scientific productivity: a persisting phenomenon? Scientometrics 93, 857-868.

van Leeuwen, T.N., Visser, M.S., Moed, H.F., Nederhof, T.J., van Raan, A.F.J., 2003. The Holy Grail of science policy: exploring and combining bibliometric tools in search of scientific excellence. Scientometrics 57, 257-280.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S. Springer-Verlag, New York.

Wallner, B., Fieder, M., Iber, K., 2003. Age profile, personnel costs and scientific productivity at the University of Vienna. Scientometrics 58, 143-153.

Wang, D., Song, C., Barabási, A.L., 2013. Quantifying long-term scientific impact. Science 342, 127-132.

Weiss, Y., Lillard, L.A., 1982. Output variability, academic labor contracts, and waiting times for promotion. Research in Labor Economics 5, 157-188.

Wooton, R., 2013. A simple, generalizable method for measuring individual research productivity and its use in the long-term analysis of departmental performance, including between-country comparisons. Health Research Policy and Systems 11, 1-14.

Xie, Y., Shauman, K.A., 2003. Women in Science: Career Processes and Outcomes. Harvard University Press, Cambridge, Mass.

Zeileis, A., Kleiber, C., Jackman, S., 2008. Regression models for count data in R. Journal of Statistical Software 27, 1-25.

Zhang, X., Lei, Y., Cai, D., Liu, F., 2012. Predicting tree recruitment with negative binomial mixture models. Forest Ecology and Management 270, 209-215.

Zuckerman, H.A., Merton, R.K. (1972). Age, aging, and age structure in science, in: Riley, M.R., Johnson, M., Foner, A. (Eds), A Sociology of Age Stratification: Aging and Society. Russell Sage Foundation, New York, pp. 292-356.