Journal of Informetrics 8 (2014) 390–405

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

How good is a model based on bibliometric indicators in predicting the final decisions made by peers?

Elizabeth S. Vieira[a,b], José A.S. Cabral[c], José A.N.F. Gomes[a,*]

[a] REQUIMTE/Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal

[b] Departamento Engenharia Industrial e Gestão, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal

[c] INESC-TEC, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal

ARTICLE INFO

Article history:
Received 15 October 2013
Received in revised form 16 January 2014
Accepted 23 January 2014
Available online 19 February 2014

Keywords:
Peer-review
Bibliometric indicators
Predictive power
Auxiliary instrument

ABSTRACT

This paper shows how bibliometric models can be used to assist peers in selecting candidates for academic openings.

Several studies have demonstrated that a relationship exists between results from peer-review evaluations and results obtained with certain bibliometric indicators. However, very little has been done to analyse the predictive power of models based on bibliometric indicators. Indicators with high predictive power will be seen as good instruments to support peer evaluations. The goal of this study is to assess the predictive power of a model based on bibliometric indicators for the results of academic openings at the level of *Associado* and *Catedrático* at Portuguese universities. Our results suggest that the model can predict the results of peer-review at this level with a reasonable degree of accuracy. This predictive power is better when only the scientific performance is assessed by peers.

1. Introduction

Peer-review is a process that uses a set of experts, considered qualified individuals for a given field, to perform a review. These experts formulate a set of qualitative judgments related to the object under assessment. Peer-review can be applied in several contexts: (1) it is frequently used by the academic community in internal evaluations; (2) it is systematically used by the editors of journals to evaluate submitted manuscripts; (3) applicants for academic or research positions are normally selected by a special committee of experts and (4) doctoral theses are submitted to a jury of experts. In the case of manuscripts submitted for publication, peer-review is used to improve the quality of the manuscripts by detecting weaknesses and errors. The feedback given by peers is used by the author to revise and improve the work. It is usually assumed that evaluation considers the originality and the contribution of the work for the advancement of knowledge in the scientific community. Similar evaluations occur in cases 3 and 4 mentioned above. In internal evaluations, peer-review is used for decisions regarding promotions. In this situation not only the aspects stated above are evaluated, but other parameters are also assessed depending on the final purpose. Peer-review has a long history and is well accepted by the scientific community, despite its limitations. In this sense, it seems correct to say that the final judgments made by peers are considered trustworthy.

* Corresponding author. Tel.: +351 220402507; fax: +351 220402659.
E-mail addresses: elizabeth.vieira@fc.up.pt (E.S. Vieira), jacabral@fe.up.pt (J.A.S. Cabral), jfgomes@fc.up.pt (J.A.N.F. Gomes).

1751-1577/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.joi.2014.01.012

Several criticisms have been raised against this methodology. Studies to assess the validity and to design strategies for improving the peer-review process are not very common. The absence of agreement among peers (or reliability) when they are asked to assess the same proposal is the main weakness of this methodology. This was analysed by Hodgson (1997) using the proposals submitted simultaneously to the Heart Stroke Foundation (HSF) and the Medical Research Council of Canada (MRC). Both agencies use peer-review to undertake evaluations and Pearson's correlation obtained between the scores was 0.592. The two agencies made the same decision for 72.5% of the proposals. The correlation between the ratings given by two independent peers to the same proposal out a set of proposals was also analysed using the Australian Research Council (ARC) database containing around 3000 proposals. These proposals were evaluated by more than 6000 external peers who were asked to rate the proposals on the quality of the project and the quality of the proponent researchers. The authors found reliabilities ranging between 0.15 and 0.53 (Jayasinghe, Marsh, & Bond, 2001; Jayasinghe, Marsh, & Bond, 2003; Jayasinghe, Marsh, & Bond, 2006). Reale, Barbara, and Costantini (2007) studied the reliability of the peer judgments in four disciplines (biology, chemistry, economics and humanities) at the Valutazione Triennale della Ricerca (VTR). The authors calculated Spearman's coefficient to evaluate the level of agreement between two peers when assessing the same research output. They found that Spearman's coefficient ranged between 0.25 in chemistry and 0.46 in economics.

A similar study was done by Cicchetti (1991), but in this case the reliability was studied using manuscript submissions to journals. The author studied the reliability considering the different characteristics of the disciplines. Reliability was analysed for general and diffuse disciplines and for specific and well-defined disciplines. He found that when disciplines are general and diffuse there is more agreement on rejection of documents than on acceptance. The opposite behaviour was observed when disciplines are specific and well-defined. Bornmann, Mutz, and Daniel (2010) catalog the results from more than 40 studies covering the reliability of journal peer-review.

Wood, Roberts, and Howell (2004) studied the reliability of the peer-review process at the UK Academy for Information System (UKAIS) conference. The authors found low levels of reliability.

Reliability was also studied for the selection of doctoral and post-doctoral applicants at the Boehringer Ingelheim Fonds (BIF). The author found that in 76% of the cases peers agreed on the decision to accept or reject an applicant (Bornmann & Daniel, 2005).

Up to this point, only works studying reliability have been discussed, but other studies have analysed potential biases in the peer-review process. In particular, features such as gender, institutional affiliation, academic title and nationality among others have been studied for both peers and research applicants. At the ARC, the fact that applicants are allowed to propose their own peers has been studied to determine whether this procedure introduces a potential biases. Marsh, Bonds, and Jayasinghe (2007) showed that the ratings given by peers nominated by the research applicants were higher than those given by the peers nominated by the funding panel.

Marsh, Jayasinghe, and Bond (2008) stated that peers from Australia tend to give lower ratings than peers from others nationalities, but part of this difference can be explained by the bias introduced by the applicants appointment. They verified that peers nominated by the applicants are more likely to come from other world regions (not Australia) than peers nominated by the panel. Even controlling for this aspect the authors observed that peers from Australia give lower ratings.

The studies looking at institutional affiliation as a source of bias suggest different findings. At the ARC the authors found that high prestige universities were more successful (Marsh et al., 2007). Reale et al. (2007) looking at the VTR, found there was no bias associated with the prestige of the institution.

Jayasinghe et al. (2003) showed that the academic title has a positive effect on the final rating attributed by peers at the ARC. They found that to be a Professor in the sciences has a significant and positive effect. In social sciences and humanities being a Professor is not significant, although they found Professorial status to interact significantly with university status. The study made by Reale et al. (2007) showed an opposite scenario at the VTR. The authors found no association between the academic level of the applicants and peer judgments. For all scientific fields analysed (chemistry, biology, humanities and economics) they found insignificant p-values.

Jayasinghe et al. (2003) observed that gender is not a potential bias in peer-review processes. In the same study they also showed that age was not significant in sciences, but was significant in social sciences and humanities.

Studies looking to see if peer judgments are influenced by the particular characteristics of the applicants and peers do not allow a general conclusion to be reached regarding a source of bias in the methodology. However, we can say that potential biases exist, but the way they influence the final rating depends on several factors, such as the scientific area and the scientific culture of the system that is being evaluated.

Another important issue in peer-review of grant proposals is the number of proposals assessed by peers. Jayasinghe et al. (2003) showed that when a peer is asked to evaluate three or more proposals the results are more reliable and valid. Indeed, when peers assess several proposals more references are available to draw judgments based on the originality, quality and contribution of the proposals to the advancement of knowledge.

In addition to these criticisms associated with peer-review there are other disadvantages of the methodology:

a) Time and implementation costs of peer-review methodology are very high. When applied on a large scale, e.g., an institutional or national level, it may be very expensive or impossible to implement. As a consequence, only the most significant research outputs are selected for evaluation. In the case of universities it may be difficult to select the best research outputs. In several cases, the selection could be based on the prestige and the position of the authors rather than the real

quality of the outputs. A study made by Abramo, D'Angelo, and Caprasecca (2009) showed that Italian universities did not select their best outputs for the VTR, distorting the final results due to an inefficient selection.

b) In some situations (as in the case of Portuguese academic openings) the parameters used for evaluation are pre-defined, but each panel member uses his own criteria to undertake the evaluation. One of the more frequent ways of assessing applicants for academic openings is by looking at where the applicants have been published, but again detailed criteria are not specified. In addition, if one of the applicants works in a very specialized field, some of the panel members may only have a general knowledge of that topic; in this situation it can be considered that these members do not have the knowledge needed to undertake the evaluation;

c) Peer-review has been criticized for taking a conservative view and not being receptive to new ideas. This may be prejudicial to emerging areas.

In fact, in the peer-review process the parameters that must be evaluated are stated (e.g. relevance of contribution; conception of the contribution, methodology (journal submissions); author's background; the originality of the proposed research project (grant submissions)) depending on the final purpose (Bornmann, 2011). But in all these situations peers will use their own knowledge and expertise to formulate a judgment. The way the evaluation is undertaken differs among peers because each peer applies his own methods.

Some of the limitations pointed out above can, in part, be counteracted if bibliometric indicators are used as an auxiliary instrument in peer-review. Several studies have tried to find a relationship between peer-review and bibliometric indicators at different levels (Abramo et al., 2009; Aksnes & Taxt, 2004; Bornmann & Daniel, 2006; Bornmann & Daniel, 2007; Bornmann, Wallon, & Ledin, 2008; Franceschet & Costantini, 2011; Nederhof & van Raan, 1987; Reale et al., 2007; Rinia, van Leeuwen, van Vuren, & van Raan, 1998; Taylor, 2011; van Raan, 2006). The authors of these studies used simple statistical methodologies to verify that the results of the peer-review process correlate with the results obtained through bibliometric indicators. For some of these models the design was primarily based on bibliometric indicators. Studies in this area should analyse the predictive power of the models defined, which in practice was often not the case.

The goal of the study presented here is to assess the predictive power of a model based on bibliometric indicators. The question is how far a model based on bibliometric indicators can go to predict the decisions of peer panels knowing that peer-review methodology is affected by the aspects mentioned above. If this was to be proven possible, then we could suggest the use of these bibliometric models as auxiliary instruments in peer assessment. In this analysis there is a set of covariates that influences the peer decision that cannot be observed by the analyst. Here, some of the limitations of the process (poor reliability, potential biases, among others) will be considered as if they introduce an "error" in the decision of each individual expert. These aspects will be treated as random errors. With this study we try to address the following research questions:

1) How well can a bibliometric model predict the final decision of an expert panel?
2) Will the proposed bibliometric model have significantly better predicting power than a purely random model?
3) Considering that the bibliometric model is being applied to decisions relating to academic openings, would the model perform better in a situation where research performance alone was considered?
4) Can a bibliometric model be used as an auxiliary instrument for peer judgment?

This paper is organized as follows: Section 2 describes the methodology later used in testing each model; Section 3 describes and interprets the results obtained; Section 4 draws together the main findings of our study.

2. Material and methods

2.1. Data set and statistical methodology

We consider a set of 27 academic job openings at the level of *Associado* and *Catedrático* at Portuguese universities, knowing the final decisions made by peers (the ranking of candidates for each opening). A total of 171 applicants applied to these 27 job openings. Table 1 shows how the numbers of applicants are distributed over the openings.

The scientific production of the 171 applicants indexed in the Web of Science (WoS) for the 10 years up to the date of their application was used in this study to produce a set of indicators for each candidate. The indicators were combined in a predicting model.

The process of defining the models started with a set of 12 indicators to characterize the scientific production of the applicants (Vieira, Cabral, & Gomes, 2013).

The indicators used were:

- *NDF* – the total number of documents published by the applicants. Each document was divided by the number of authors (N) associated with ($1/N$).
- *HCD* – percentage of highly cited documents. This indicator records the percentage of documents of a given researcher that are in the Top 10% most cited Portuguese documents.

- *CD* – percentage of citing documents without self-citations from the total number of citing documents.
- *PDC* – percentage of documents cited.
- h_{nf} index – This indicator is calculated in a way similar to the *h* index, but considers the different citation cultures of each subject category and the number of authors per publication (Vieira & Gomes, 2011).
- *NIR* – Normalized indicator for researchers. This indicator gives the normalized average number of citations per document taking into account the different citation culture of each subject category where the researcher has been published.
- $SNIP_m$ – the median of the *SNIP* indicator (Moed, 2010) for the set of journals where the applicants have been published.
- SJR_m – the median of the *SJR* indicator (Gonzalez-Pereira, Guerrero-Bote, & Moya-Anegon, 2010) for the set of journals where the applicants have been published.
- NA_m – Median number of authors per document. The NA_m is a normalized median number of authors per document, considering the variability of the number of authors in the subject categories.
- *DIC* – percentage of documents with international collaboration.

Table 1
Distribution of the numbers of applicants over the 27 openings.

Number of applicants	Number of openings
1 or more	27
2 or more	27
3 or more	21
4 or more	19
5 or more	18
6 or more	17
7 or more	14
8 or more	12
9 or more	10
10 or more	5
11 or more	1

Considering the features of our data set, a rank ordered logistic regression (ROLR) was used to adjust the model to the available data (Train, 2009). The basic concept of the ROLR is that the panels rank a set of applicants taking into account the so called utility associated with each applicant. Peers will first choose the applicant with the highest utility. The utility is defined by:

$$U_{nj} = V_{nj} + \varepsilon_{nj} \tag{1}$$

where $V_{nj} = \beta \cdot x_{nj}$ is a linear function of the observed explanatory indicators (x_{nj}) and β the vector of coefficients. The ε_{nj} represents those factors that influence utility, but are unknown to the analyst. The probability of an applicant, *i*, being placed first by the peer panel, *n*, within each job opening with *j* applicants is given as:

$$P_{nj} = \frac{e^{\beta \cdot x_{ni}}}{\sum_j e^{\beta \cdot x_{nj}}} \tag{2}$$

The numerator of the equation represents the exponential of the utility related to choice *i* (applicant) and the denominator the sum of the exponential of the utilities that represent the choice set available to the panels.

Initially, multicollinearity was detected, this being particularly strong between h_{nf} and *NDF*. As this is not desirable for our analysis, three strategies were considered:

1) To group the indicators using factor analysis (Strategy 1).
2) To eliminate the variable h_{nf} from our set of indicators and to define a model using the remaining indicators (Strategy 2).
3) To eliminate the variable *NDF* from our set of indicators and to define a model using the remaining indicators (Strategy 3).

Following these strategies, we constructed three models as shown below. Each model gives the probability of a given applicant being placed first (P_{ni}).

Model 1

$$P_{ni} = \frac{e^{1.06PC_{ni}}}{\sum_j e^{1.06PC_{nj}}} \tag{3}$$

where PC_{ni} is a composite indicator obtained through factor analysis and encompasses nine indicators (h_{nf}, *NIR, NDF, PDC, HCD, Q1, NI, CD* and SJR_m). The remaining indicators were not considered in the *PC* as they did not meet the requirements of the factor analysis.

Model 2

$$P_{ni} = \frac{e^{0.40h_{nf_{ni}}+0.032HCD_{ni}}}{\sum_j e^{0.40h_{nf_{ni}}+0.032HCD_{ni}}} \quad (4)$$

The indicators h_{nf} and HCD were found to have significant impact ($p < 0.05$) These represent the impact of the scientific production of the applicant. The h_{nf} indicator goes further, giving information about the dimension related to quantity.

Model 3

$$P_{ni} = \frac{e^{0.11NDF_{ni}+0.044HCD_{ni}+0.19NA_{mni}}}{\sum_j e^{0.11NDF_{ni}+0.044HCD_{ni}+0.19NA_{mni}}} \quad (5)$$

The indicators with significant impact ($p < 0.05$) in this model are related to quantity (NDF), impact (HCD) and collaboration (NA_m).

The rationale behind these models has already been presented and discussed (Vieira et al., 2013). Here we are interested in the performance of these models relating to the predictions of the final decisions made by the panel for this type of job openings.

2.2. The success of the models in predicting the final decisions

Train, in his book dedicated to discrete choice models states: "*Another goodness-of-fit statistic that is sometimes used, but should actually be avoided, is the "percent correctly predicted." This statistic is calculated by identifying for each sampled decision maker the alternative with the highest probability, based on the estimated model, and determining whether or not this was the alternative that the decision maker actually chose. The percentage of sampled decision makers for which the highest-probability alternative and the chosen alternative are the same is called the percent correctly predicted. This statistic incorporates a notion that is opposed to the meaning of probabilities and the purpose of specifying choice probabilities. The statistic is based on the idea that the decision maker is predicted by the researcher to choose the alternative for which the model gives the highest probability. However, as discussed in the derivation of choice probabilities in Chapter 2, the researcher does not have enough information to predict the decision maker's choice. The researcher has only enough information to state the probability that the decision maker will choose each alternative. In stating choice probabilities, the researcher is saying that if the choice situation were repeated numerous times (or faced by numerous people with the same attributes), each alternative would be chosen a certain proportion of the time. This is quite different from saying that the alternative with the highest probability will be chosen each time*" (Train, 2009). Here we adopted the following procedure with the aim of testing the success of the models. Initially, in a very simplistic way, we determined the percentage of openings where the applicant placed in first position by peers has a probability of being placed first that is higher than that expected in a situation where there is no information about the applicants and the selection is random. The information provided by the models allows us to go into more detail and to adopt the same procedure using pairs of applicants. In the set of openings available it is possible to find 426 pairs. For an opening with six applicants (ranked 1, 2, 3, 4, 5 and 6 by the peers) we have 14 pairs (1-2; 1-3; 1-4; 1-5; 1-6; 2-3; 2-4; 2-5; 2-6; 3-4; 3-5; 3-6; 4-5; 4-6).

The relative positions of the candidates ranked four and below (when more than four candidates apply) will not be considered in the evaluation of the model as it has been suggested that panels tend to pay less attention to the ranking given as they go down in a large set of applicants (Vieira et al., 2013). This is the reason why we considered all the pairs formed among candidates 1–4 and the pairs formed between applicants 1–4 and applicants below 4.

2.2.1. Probability distribution functions associated with the results provided by the models

While a simplistic method can be employed as described above, more information about the predictive power can be obtained using the probability distribution function associated with the data of each model for pairs of applicants. In these distributions the Y variable will be studied, representing the number of pairs predicted correctly using the models. The methodology adopted for defining the models allows the probabilities to be determined for different subsets. For each pair it is possible to use the probability of each applicant being placed first. An example is provided below.

Example

Consider an opening with three applicants (C1, C2, C3) where the probability of each one being placed first is 53.8%, 20.2% e 26.0%, respectively. The applicant C1 was placed in first position by the peers and C2 and C3 in the second and third position, respectively. For pair C1 and C2, the probability of being placed first is 72.7% for C1 and 27.3% for C2.

Two possible outcomes are possible for each pair; one is success (with probability p) and the opposite failure (with probability $1 - p$). Success means that the prediction given by the models coincides with the decision of the peers. The probability of success is the probability that the applicant would be placed in first position by the peers in the pair to be selected first using the models. As the success probability differs among pairs it was not possible to apply the binomial distribution. In this case the solution was to use the Monte Carlo method.

With information about the probability of success for each pair, we carried out a simulation to determine the probability distribution function of each model, i.e., the probability of correctly ordering 1, 2, . . ., 426 pairs of applicants.

2.2.2. Comparison of the results of the models with the results of a scenario where candidates are selected randomly

It is important to assess if the predictions given by our models are better than those that we would make in the absence of hard information on the applicants. In this case, the probability (p) of correctly guessing the winner of any pair is 50%. The probability of guessing correctly X pairs of the total of $n = 426$ follows a binomial distribution that can be approximated by a normal distribution with average equal to $n/2$ and a standard deviation equal to $\sqrt{n}/2$,

$$X \sim N(n \times p; \sqrt{n \times p(1-p)}) \tag{6}$$

$$X \sim N\left(\frac{n}{2}; \frac{\sqrt{n}}{2}\right) \tag{7}$$

Following the methodology described above, we can determine the actual distribution for the data of each model and approximate it by a normal distribution (this will be shown in the next section):

$$Y \sim N(\mu; \sigma) \tag{8}$$

Above, Y is the number of pairs predicted correctly by each model and μ and σ are the actual average and standard deviation of the determined distributions.

Finally, we calculate the following probability for each of the three models:

$$P(X > Y) = P(X - Y > 0) \tag{9}$$

2.2.3. Predictive power of the models when scientific performance alone is assessed

We defined our models using variables that describe some aspects of the scientific performance of the candidates. In the job openings other dimensions are considered in addition to scientific performance. For 74% of the total number of openings, the scientific performance has a weighting of between 60% and 70% on the final decision made by peers. For the remaining openings the weighting varies between 45% and 55%. The other dimensions are related to teaching, technology transfer and management activities. In this study these dimensions were not considered due to the impossible task of collecting information relating to these activities at the time of candidature. If we were to consider openings where the scientific performance is the unique aspect evaluated, we would expect to have better predictions. This will require a relatively long and very technical excursion into the realm of the statistical theory being used in this paper. This will be of hard reading for most of the readers, but we consider that this should be demonstrated. Those readers interested in a more detailed description of the methodology used here should consult Train (2009).

Starting from the expression that gives the utility and considering a new term representing the other dimensions (teaching, technology transfer and management activities), it is possible to infer the predictive power of each model if the scientific performance were the only aspect assessed.

$$U_{nj} = a_{nj1}\beta_{nj1}x_{nj1} + b_{nj2}\delta_{nj2}y_{nj2} + \varepsilon_{nj} \tag{10}$$

U_{nj} is the utility that the peer panels attribute to alternative j (applicants);

$\beta_{n1}x_{nj}$ represents the scientific performance dimension and was evaluated here using bibliometric indicators; a_{nj} is the weight attributed to the scientific performance in each opening;

$\delta_{n2}y_{nj}$ represents the other dimensions that cannot be assessed using bibliometric indicators; b_{nj} is the weight attributed in each opening to the unobservable dimension;

ε_{nj} is the error or stochastic term.

The terms ε_{nj} are those factors that influence the utility and are unknown to the analyst. It is assumed that the ε_{nj} follow an *iid* extreme value distribution. The probability density of each ε_{nj} is:

$$f(\varepsilon_{nj}) = e^{-\varepsilon_{nj}} e^{-e^{-\varepsilon_{nj}}} \tag{11}$$

The cumulative distribution is:

$$F(\varepsilon_{nj}) = e^{-e^{-\varepsilon_{nj}}} \tag{12}$$

We can now rewrite the expression (10):

$$\varepsilon_{nj} = U_{nj} - a_{nj}\beta_{nj1}x_{nj1} - b_{nj}\delta_{nj2}y_{nj2} \tag{13}$$

$$\varepsilon_{nj} = U_{nj} - a_{nj}V_{nj1} - b_{nj}V_{nj2} \tag{14}$$

The probability that a particular peer panel, n, chooses the alternative i is:

$$P_{ni} = P(U_{ni} > U_{nj}, \forall j \neq i) \tag{15}$$

$$P_{ni} = P((V_{ni1} + V_{ni2}) - (V_{nj1} + V_{nj2}) > \varepsilon_{nj} - \varepsilon_{ni}, \forall j \neq i) \tag{16}$$

Variables V_{nj1} and V_{nj2} were introduced to help simplify the expression. Using the density function $f(\varepsilon_{nj})$, this probability can be rewritten as:

$$P_{ni} = \int I(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni1} + V_{ni2}) - (V_{nj1} + V_{nj2}) \forall j \neq i)(\varepsilon_{nj}) d\varepsilon_{nj} \tag{17}$$

where $I(\ldots)$ is the indicator function, being equal to 1 if the argument is true and 0 if it is false.

As it is assumed that the ε_{nj}'s are independent, the cumulative distribution of all $j \neq i$ is the product of the individual cumulative distributions and the probability is given by:

$$P_{ni/\varepsilon_{ni}} = \prod_{j \neq i} e^{-e^{-(\varepsilon_{ni}(V_{ni1}+V_{ni2})-(V_{nj1}+V_{nj2}))}} \tag{18}$$

As we do not know ε_{ni} and V_{n2}, then we can calculate the probability as:

$$P_{ni/\varepsilon_{ni}} = \int f(V_{ni2}) dV_{ni2} \int f(V_{nj2}) dV_{nj2} \int \prod_{j \neq i} e^{-\varepsilon_{ni}} e^{-e^{-\varepsilon_{ni}}} e^{-e^{-(\varepsilon_{ni}+a_n*(V_{ni1}-V_{nj1})+b_n*(V_{ni2}-V_{nj2}))}} d\varepsilon_{ni} \tag{19}$$

$$P_{ni/\varepsilon_{ni}} = \int f(V_{ni2}) dV_{ni2} \int f(V_{nj2}) dV_{nj2} \frac{e^{a_n V_{ni1} + b_n V_{ni2}}}{e^{a_n V_{ni1} + b_n V_{ni2}} + e^{a_n V_{nj1} + b_{nj} V_{nj2}}} \tag{20}$$

$$P_{ni/\varepsilon_{ni}} = \int f(V_{ni2}) dV_{ni2} \int f(V_{nj2}) dV_{nj2} \frac{1}{1 + e^{a_n(V_{ni1} - V_{nj1}) + b_n(V_{ni2} - V_{nj2})}} \tag{21}$$

The calculation of expression (21) implies that the distribution of V_2 is known. Without information about V_2 the solution was to assume a few reasonable hypotheses. The working hypotheses and the procedure used are described in Appendix A.

The determination of $P_{ni/\varepsilon_{ni}}$ through expression (21) will allow an estimation of the improvement of the information given by the models in a situation where the scientific performance is the unique aspect assessed.

3. Results and discussion

3.1. The predictive power of the models

Having been placed in a given position by each panel, Table 2 presents the percentage of openings where the probability of the applicant being placed first is higher than that to be expected in a situation where we do not have information about the applicants and the selection is random.

Table 2
Comparison of the models' results with a random scenario.

Model	P_{R1} (%)	P_{Pairs} (%)
1	78	75
2	70	75
3	63	76

P_{R1} in Table 2 is the percentage of openings where the applicant having been placed in first position by the peer panel has a probability of being selected first higher than the probability expected in a random scenario. In the second column, P_{Pairs} gives the percentage of pairs (from the 426 counted as explained in Section 2) where the applicant placed in first position (by the peers) has a probability of being positioned first better than chance (50%).

It is considered that the percentage obtained for each model in both situations is higher as we only considered on the parameterization of the models those variables related with the scientific performance.

Fig. 1 presents histograms representing the frequency distribution of the probabilities determined using each model for the 426 pairs. The probabilities represented are those obtained for the applicant placed in the better position out of the pair.

In Fig. 1 we can see that there is a concentration of the pairs on the right side of the histogram, i.e. for probabilities above 50% showing, the superiority of the models in relation to the scenario where applicants are ordered randomly.

Using the data given by the models and simulation processes we can go further and formulate more detailed conclusions about the application of bibliometric analysis as an auxiliary instrument for the peer-review process. The assessment of the probability distribution function associated with the data of each model allows inferences to be made about the accuracy of the forecasts provided by each model. The distributions obtained, as explained in Section 2, are shown in Fig. 2 with the respective parameters.

The intention here is to use these distributions to determine the expected number of pairs predicted correctly (mean values) using the information given by each model. It is possible, in Fig. 2, to observe that the distributions obtained have a pattern close to the normal distribution. Here the distributions represent the probability of a given result occurring for the discrete variable 'number of pairs predicted correctly'. In the binomial distribution, the probability of success associated

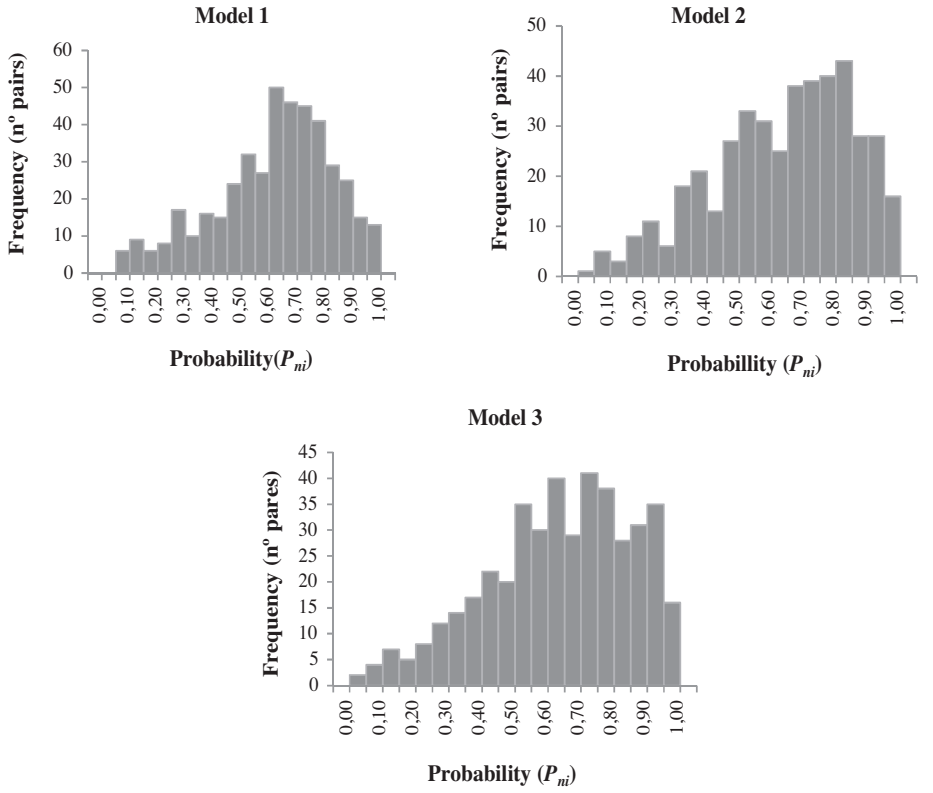

Fig. 1. Frequency distribution of P_{ni}.

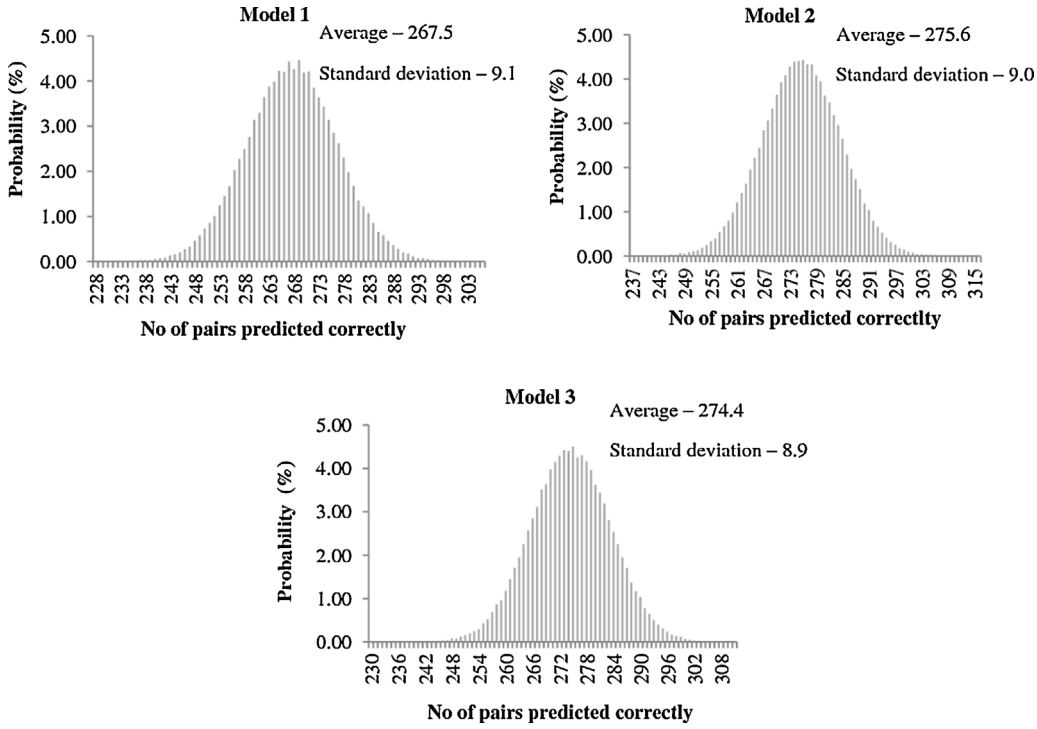

Fig. 2. Probability distribution function for the 426 pairs obtained using the Monte Carlo method.

with each event is the same for all attempts. In this situation, the probability of success differs among pairs, so we cannot consider that the binomial distribution describes the data's behaviour. However, the number of observations is large and this allows us to adjust the normal distribution to the data and to draw conclusions about the results obtained through the simulation.

The *Kolmogorov–Smirnov* test resulted in the rejection of the null hypothesis, i.e. there is no statistical evidence that the data comes from a normal distribution ($p < 0.001$). However, this test has some sensitivity to large samples as it has the capacity to detect minor deviations in relation to the assumed distribution. The option was taken to design the frequency histogram using the superimposed normal curve and information about skewness and kurtosis.

In Fig. 3 we can in fact see that the data follow a normal distribution well. The values for skewness, in Table 3, show that the distributions are slightly skewed to the left, with magnitudes close to zero suggesting small tails. The kurtosis values are also small and suggestive of a leptokurtic distribution for Models 1 and 2, and a mesokurtic distribution for Model 3.

Using Model 1 we can say that, on average, we can correctly predict 63% ± 2% of the pairs. This percentage is almost the same for Model 2 and 3 (65% ± 2% and 64% ± 2%). The values are well above that expected in a scenario where applicants are selected randomly (50%).

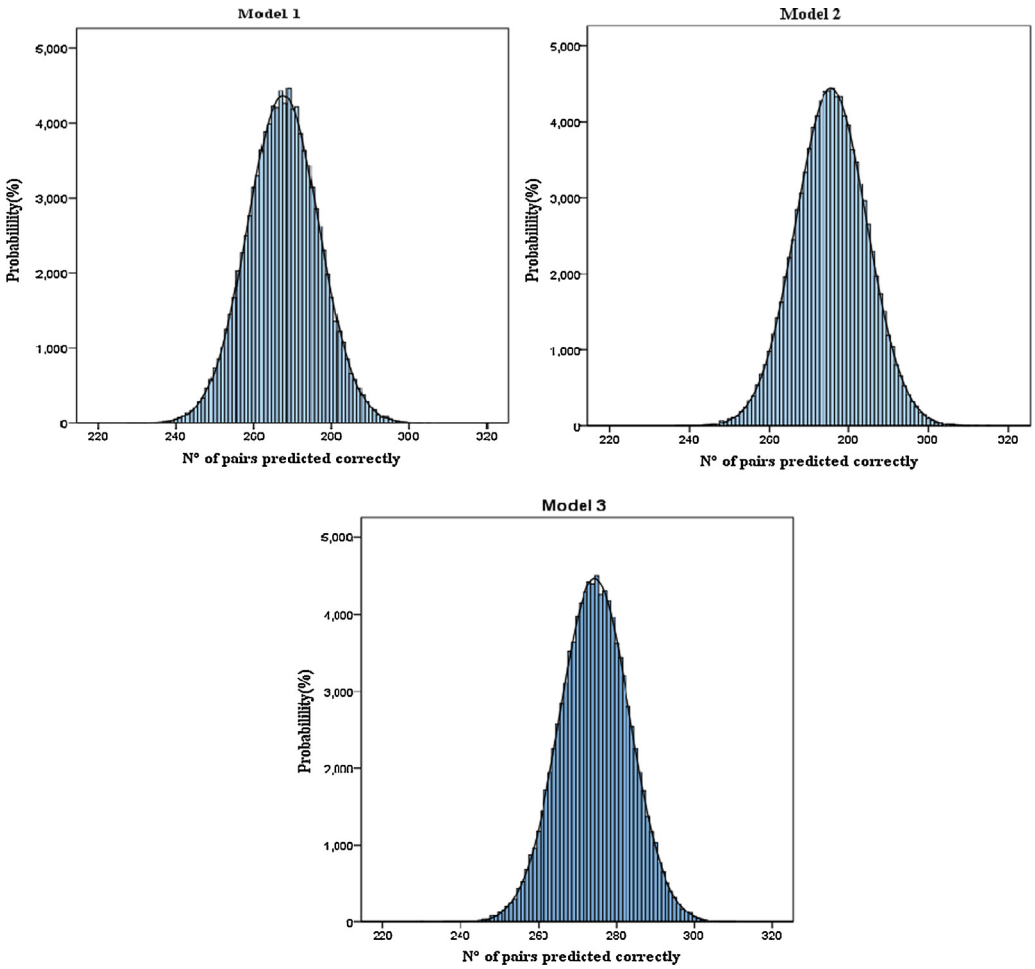

Fig. 3. Normal distribution adjusted to each model.

Table 3
Values obtained for skewness, kurtosis, average and standard deviation.

Models	Skewness	Kurtosis	Average	Standard deviation
Model 1	−0.014	0.012	267.5	9.1
Model 2	−0.028	0.003	275.6	9.0
Model 3	−0.028	0.000	274.4	8.9

Table 4
Average and standard deviation for each probability distribution function.

Variables	Model	Average	Standard deviation
Y	Model 1	267.5	9.1
	Model 2	275.6	9.0
	Model 3	274.4	8.9
X	Random selection	213	10.3
$(Y-X)$	Model 1	−54.5	13.5
	Model 2	−62.6	13.5
	Model 3	−61.4	13.4

3.2. Value added by the prediction models when compared to a random selection

If for a given opening we have two applicants and we do not have any information about them, we must assume that the probability of each candidate being placed first is 50%. With our models we are expecting to gain some information, i.e. between two candidates we will be able to say that one has a probability of being placed first higher than 50%. In this sense it is important to show how good our models are when compared with a situation where the choice among candidates is purely random.

As we saw before, a normal distribution can be adjusted to our results. In the scenario without information about the applicants a normal distribution can be used, as an approximation to the true binomial. Here we want to assess what the probability is that the predictions made in a random situation would be better than those that we have using the models. This is given by expression (9) as explained in detail in Section 2.

For the scenario where the applicants are selected randomly it was assumed that the number of pairs predicted correctly (X) follows a normal distribution characterized by the following parameters:

$$X \sim N\left(\frac{426}{2}; \frac{\sqrt{426}}{2}\right) \tag{22}$$

As our data have a normal distribution in both scenarios we know that:

$$(X \sim Y) \sim N\left(\frac{n}{2} - \mu; \sqrt{\left(\left(\frac{\sqrt{n}}{2}\right)^2 + (\sigma)^2\right)}\right) \tag{23}$$

Table 4 shows the average and the standard deviation for each situation.

For each model we determined that the probability of the forecasts available in a random scenario being higher than that given by the models is actually very low (lower than 0.01%). Indeed, the models supply more information than that accessible in a situation where peers do not have information about the applicants. This conclusion allows us to highlight the contribution of the bibliometric analysis as a decision support tool in peer-review evaluations.

3.3. Assessment of the method if scientific performance alone were considered

We are using bibliometric indicators as predictor variables but these indicators depend on just one of the dimensions evaluated. For the remaining dimensions (pedagogical, knowledge transfer and academic management) no metrics are readily available for use as indicators. At best, bibliometric indicators may be expected to predict one component of peer assessment. As we compare the global peer assessment with the predictions of our model, this affects our results negatively. If we were to apply our methodology to a situation where scientific performance is the single aspect evaluated we would expect to get much better predictions.

The weight attributed to scientific performance is not the same for all openings. We tried to relate the probability of a given candidate being selected first with the weight of the scientific performance in the final decisions, assuming total ignorance of the other components. In order to do this, several working hypotheses were considered as described in Appendix A. The results obtained are presented in Fig. 4. The y-axis represents the probability of each candidate being placed first if only the scientific performance were taken into account. The x-axis presents the probability of each applicant being chosen by first taking into account the weight of the dimensions not observed in the final decisions made by peers (mixed evaluation). Each curve represents the correlation for a given weight attributed to the dimensions not considered.

Two situations can be identified in Fig. 4. For values where the probability is higher than 50% in the mixed evaluation, the results show that the probability of an applicant being placed first can be improved substantially if the evaluation were based only on the scientific performance. Considering the two extreme cases (in the sample used here), i.e. the opening where the scientific performance represents 70% of the total evaluation and that where the scientific performance represents 45% it was possible to determine that the probabilities can be improved between [0.2%; 12.2%] and [0.2%; 29.9%] respectively. If the probability dips to values lower than 50% (mixed evaluation) then the probability of an applicant being placed first would

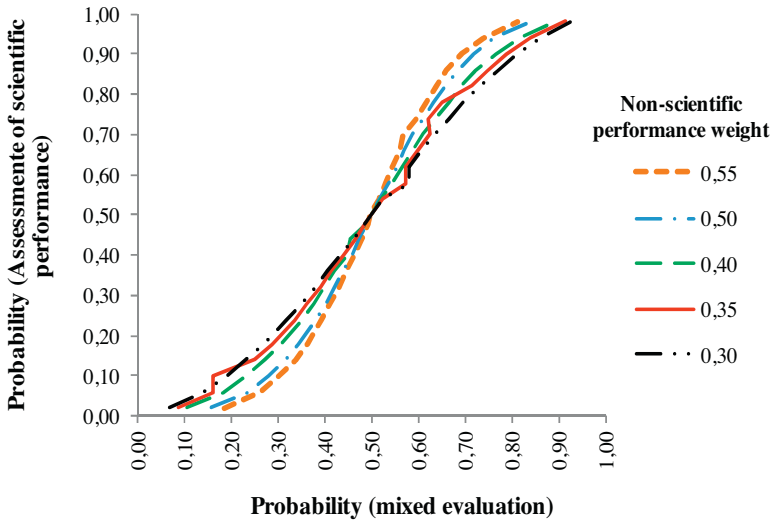

Fig. 4. Relation between the probability of a given candidate being placed first, when only the scientific performance is assessed, and the probability in the real case of mixed assessment, for different weights of the non-scientific components.

decrease under an evaluation based solely on academic performance. This behaviour was already to be anticipated; if the scientific performance of the applicant is poor the remaining parameters assessed by peers will be decisive and crucial for formulating a decision about the applicant.

With the probabilities obtained using our models and the weight given to the hidden dimensions in each opening, we can estimate the probability of success of each pair for the case when only scientific performance is being assessed. The probabilities estimated allow us to determine new probability distribution functions for each model. Once we have the distributions, we are in the position to quantify the improvement in our models when applied to situations where only the scientific performance is being assessed. Fig. 5 shows the obtained probability distribution functions.

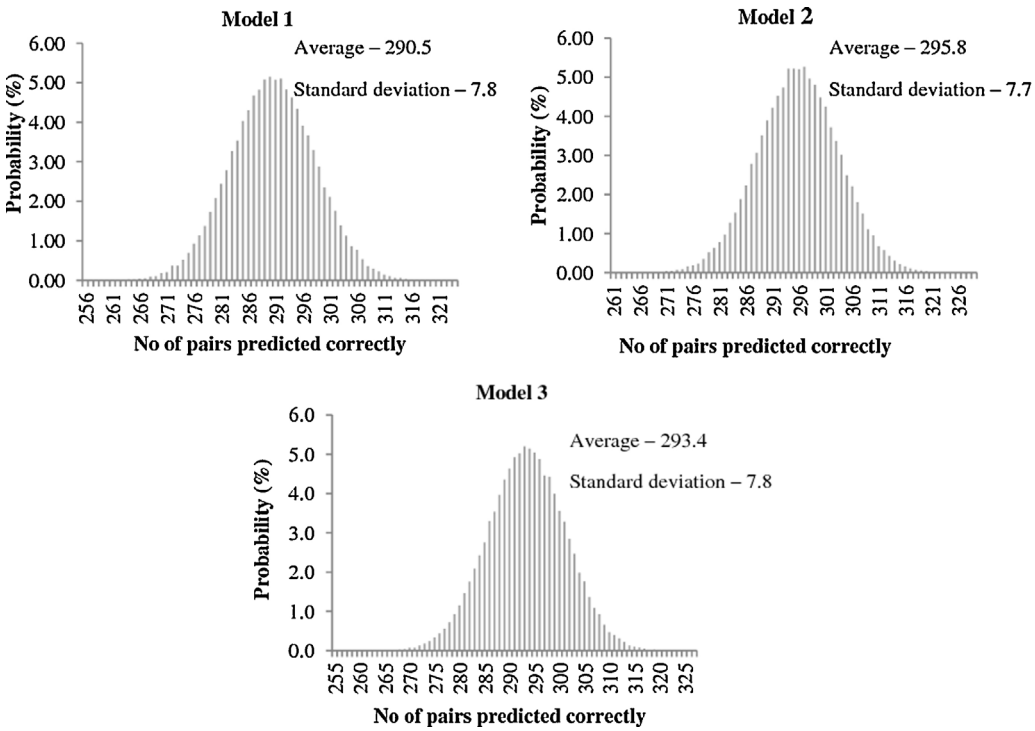

Fig. 5. Probability distribution functions for 426 pairs using the Monte Carlo method.

Table 5
Standard deviation estimated using the empirical data.

Model	Estimated σ_{V_2}
Model 1	1.09
Model 2	1.23
Model 3	1.72

The results confirm that the average number of successes increases if only the scientific performance is assessed. For these models the average percentage of pairs predicted correctly increases to $69\% \pm 2\%$. This represents an improvement of around 7% for Model 1, 6% for Model 2 and 8% for Model 3 in relation to the initial situation.

The results presented in Fig. 5 were obtained using a set of working hypotheses as described in Appendix A. However, with the empirical information available we can estimate the standard deviation associated with the unobservable parameters assessed by peers (V_2). The comparison of this value with that obtained using a theoretical approach allows us to infer the accuracy of the values estimated. A detailed description of the procedure used is available in Appendix A. Table 5 presents the results obtained for σ_{V_2}.

Table 5 shows that the values obtained for the σ_{V_2} are never far from the range $[\pi/\sqrt{6}; \pi/\sqrt{3}]$ as estimated before. The average number of correctly predicted pairs will not be far from those shown in Fig. 5, given the estimated values for σ_{V_2} in Table 5.

3.4. A model based on bibliometric indicators as an auxiliary instrument of peer evaluation

The results obtained here suggest that models based on bibliometric indicators may be a relevant auxiliary instrument for peer evaluation. However, this instrument may not be used abusively, as scientific performance is a very broad concept and some aspects are impossible to measure by applying these indicators. The example presented below shows a situation where the judgment cannot solely be based on the information given by these models.

Example

A researcher working in an emergent field might not have impact in the scientific community in the first years, but he/she might expect full acknowledgment from the scientific community after the emergent field becomes well accepted in the community. If this researcher is assessed using our models the results will suggest a weak scientific performance, but peers have competences that allow the identification of these special cases.

Scientific activities take many dimensions and the assessment of the performance depends on the policies of each scientific system. In a situation where one of the goals of the evaluation is to determine the socio-economic impact of the scientific activity, peer-review is crucial and the use of indicators may be limited.

In this study we explored the predictive power of the models comparing the results obtained with the decisions of peer-review. However, we have to be aware of the limitations of peer-review. The lack of reliability, potential bias and conflict of interests are some aspects that influence this methodology, as is suggested by the studies reviewed in the Introduction. This methodology is uniquely used in the scientific community when the assessment of research activities is the main goal. Without other options to validate our indicators the procedure adopted here is the best available.

4. Conclusions

The predictive power of three models based on bibliometric indicators was explored while accounting for several dimensions. The results obtained led us to the following findings:

- All models have a similar predictive power.
- Models 1, 2 and 3 found that in 78%, 70% and 63% of the openings, respectively, the applicant placed in first position by peers has a probability of being placed first that is better than chance. If the analysis is made considering pairs of applicants, the percentage of cases where this situation is observed is 75% for Models 1 and 2 and 76% for Model 3.
- The results given by the models overlap significantly with those from peer-review for a reasonably percentage of pairs, despite peers considering other dimensions besides scientific performance. Using the models it is expected, on average, to correctly predict between $63\% \pm 2\%$ and $65\% \pm 2\%$ of the pairs, i.e. the decision of the peers is coincident with that obtained using the models.
- The probability of the information existent in a random scenario being better than that available using our models is rather small (<0.01).
- The results given by the models improve when only the scientific performance is assessed by peers. An improvement between 6% and 8% was estimated for the average value.

The forecasts provided by the models are considered satisfactory, considering all the factors that affect the peer-review process.

The conclusions would be stronger if this study were applied to openings where the scientific performance was the unique parameter assessed.

From the three models available, Model 1 will probably be disregarded as this requires the calculation of nine indicators and the predictive power is similar to Model 2 and 3. This model is the best in providing forecasts in the situation where the analysis only takes into account the applicant placed in the first position. Model 2 and 3 have a similar predictive power, but the indicators represent different dimensions of the scientific performance. Here, peers have to state which dimensions are relevant.

In the selection of candidates for academic openings, information about the parameters that must be assessed and their weight on the final decisions have to be mentioned. The openings used here provided this information in the announcement. When bibliometric analyses are requested by the peer panels, the dimensions that they are intended to characterize have to be met. In this situation, bibliometricians have to inform the final users of the indicators about the sources used in the calculation, the limitations of the indicators, the set of assumptions considered and the difficulties that could arise in the interpretation of the results provided by the indicators. By supplying this information it will be possible to use the results obtained using bibliometric indicators in a responsible way.

It is also important to consider that individual performance cannot be reduced to a simple "number". The performance is influenced by several factors, such as researcher's age, position in the working group and scientific domain. Within the same academic environment and for researchers that have the same position, the activity profile may differ considerably. It is possible to find researchers that always work alone and others who collaborate occasionally. In this matter it is also important to consider if the working group is well established in the scientific community. In summary, when bibliometric indicators are used to compare the performance of researchers, the working context of each researcher has to be considered.

Our final suggestion is to use bibliometric indicators as an auxiliary instrument, offering objective information which can then be complemented with other relevant information that directly or indirectly impacts the scientific performance of researchers. A combination of a quantitative method (bibliometric techniques) and a qualitative method (peer-review) can improve the overall assessment for this kind of academic openings.

Acknowledgments

Elizabeth Vieira wish to acknowledge the financial support from FCT (Foundation of Science and Technology), Portugal, through grant No. SFRH/BD/75190/2010.

Appendix A.

This section describes the set of working hypothesis used to determine the performance of the models, for the case where scientific performance is the unique parameter used by the peers when assessing applicants. These hypotheses aim to determine the values of expression (21) in Section 2.

If b_n, in expression (21), is assumed to be equal to zero, then:

$$P_{ni/\varepsilon_{ni}} = \frac{1}{1 + e^{a_n(V_{ni1} - V_{nj1})}} \tag{24}$$

If b_n is different from zero we need to solve the integral in expression (21), but for this we need the distribution of V_2. As we do not have information about the variables that characterize V_2 some reasonable hypothesis are considered.

The term ε_{ni} essentially encompasses those factors of the dimensions V_1 and V_2 that influence the utility and are unknown to the analyst. In this sense, we can consider that the dispersion of the term ε_n is related to the dispersion of V_1 and V_2 and in order to determine the dispersion of the term ε_n we need to know the distributions of V_1 and V_2.

Hypothesis 1: V_1 and V_2 have a normal distribution.

$$V_1 \sim N(\mu_{V_1}; \sigma_{V_1}) \tag{25}$$

$$V_2 \sim N(\mu_{V_2}; \sigma_{V_2}) \tag{26}$$

Hypothesis 2: V_1 and V_2 have equal dispersion.

$$\sigma_{V_1} = \sigma_{V_2} = \sigma_V \tag{27}$$

If we consider that peers give equal attention to both dimensions, and we have no reason to suspect that this is not true, we can assume that the dispersion of V_1 and V_2 is the same.

With these two hypotheses we can say that:

$$\sigma_{\varepsilon_n} = \sqrt{((a_{n1}\sigma_{V_1})^2 + (b_{n2}\sigma_{V_2})^2)} \tag{28}$$

$$\sigma_{\varepsilon_n} = \sigma_V \sqrt{((a_{n1})^2 + (b_{n2})^2)} \tag{29}$$

We are using a normal distribution to draw inference about the standard deviation of the term ε_{ni}, when, in fact, this term follows an *iid* extreme value distribution. As we can see in expression (16), only the difference between utilities is important

for calculating the probabilities. If we have two alternatives, j and k, then $\varepsilon_n = \varepsilon_{nj} - \varepsilon_{nk}$ is the distributed logistic. Using the *idd* extreme value distribution for the error term or the logistic distribution for the difference between error terms is almost the same as using the normal distribution. In fact, the logistic distribution has heavier tails than the normal distribution allowing a better control of the atypical observations.

Now, with these hypotheses we can determine the range of variation of σ_{ε_n} for different values of b_n.

If $b_n = 1$ then:

$$\sigma_{\varepsilon_n} = \sigma_V \tag{30}$$

If $b_n = 1/2$ then:

$$\sigma_{\varepsilon_n} = \sigma_V \frac{1}{\sqrt{2}} \tag{31}$$

The final result is that:

$$\sigma_{\varepsilon_n} \in \left[\frac{1}{\sqrt{2}}; 1\right] \sigma_V \tag{32}$$

In the ROLR the scale of the utility is normalized using the variance of the error term. In this sense, the coefficients obtained are the impact of the observed variables relative to the standard deviation of the unobserved variables (Train, 2009). This is shown in the following expression. The error variance is fixed at a certain value for convenience. In the standard logistic distribution the variance of the error term is $\pi^2/6$.

If $\sigma^2 = \pi^2/6$ then:

$$U_{nj} = a_{nj} \frac{\beta^*_{nj1}}{\sigma_{\varepsilon_{nj}}} \frac{\pi}{\sqrt{6}} x_{nj1} + b_{nj} \frac{\delta^*_{nj2}}{\sigma_{\varepsilon_{nj}}} \frac{\pi}{\sqrt{6}} y_{nj2} + \varepsilon_{nj} \tag{33}$$

If

$$\sigma_{\varepsilon_n} \in \left[\frac{1}{\sqrt{2}}; 1\right] \sigma_V \tag{34}$$

Then:

$$\frac{\pi}{\sqrt{6}} = \frac{1}{\sqrt{2}} \sigma_{V_2} \tag{35}$$

$$\frac{\pi}{\sqrt{6}} = 1 \sigma_{V_2} \tag{36}$$

The final conclusion is that:

$$\sigma_{V_2} \in \left[\frac{\pi}{\sqrt{6}}; \frac{\pi}{\sqrt{3}}\right] \tag{37}$$

We stated above the hypothesis for the distribution of V_2. We have now the estimated standard deviation for this component and it is possible to solve the integral of the expression (21).

$$V_2 \sim N\left(0; \frac{\pi}{\sqrt{6}}\right) \tag{38}$$

We chose the lowest value for the standard deviation. As the standard deviation decreases, the influence of the error term also decreases and the results given by the models improve. However, the final results are not very sensitive to this choice.

The values in expression (21) were determined using the Monte Carlo method as there is no analytical solution. The following steps were carried out:

1) Consider that in expression (21):

$$K = e^{a_{n1}(V_{ni1} - V_{nj1})} \tag{39}$$

If b_n is equal to zero in expression (21), then the determination of P_{ni} is done using expression (24):

$$e^{(V_{ni1} - V_{nj1})} = \left(\frac{1}{P_{ni}} - 1\right) \tag{40}$$

and:

$$K = \left(\frac{1}{P_{ni}} - 1\right)^{a_{n1}} \tag{41}$$

2) Values for P_{ni} were fixed and determined the values of K.
3) With information about K, and the average and standard deviation of V_2, the expression (21) was resolved using the Monte Carlo method.

The probabilities estimated through the simulation process were used to build new probability distribution functions and to estimate the improvement of the information given by the models if only the scientific performance were assessed by peers.

The accuracy of the estimated σ_{V_2}

The empirical data available allows the reliability of the estimated values to be tested for the σ_{V_2} using the process described above.

Hypothesis 1: The dispersion of V_1 and V_2 are the same.

$$\sigma_{V_1} = \sigma_{V_2} \tag{42}$$

The utility is defined as:

$$U_{nj} = a_{nj1}\beta_{nj1}x_{nj1} + b_{nj2}\delta_{nj2}y_{nj2} + \varepsilon_{nj} \tag{43}$$

If the term V_1 ($\beta_{nj1}x_{nj1}$) has two variables (x_1 and x_2) we know that each variable has its own standard deviation. When we multiply the coefficient associated with each variable by the standard deviation (of the variable) we will obtain the dispersion of this term. The effect of dimension V_1 on the dispersion of the utility is given by:

$$a_{nj}(\beta_{nj1}\sigma_{x_{nj1}} + \beta_{nj2}\sigma_{x_{nj2}}) \tag{44}$$

The effect of dimension V_2 ($\delta_{nj2}y_{nj2}$) on the dispersion of the utility is given by:

$$b_{nj}\delta_{nj2}\sigma_{y_{nj2}} \tag{45}$$

As we already saw β and δ are normalized taking into account the variance of the unobservable factors of the error term:

$$\beta_{nj1} = \frac{\beta^*_{nj1}}{\sigma_{\varepsilon nj}} \times \frac{\pi}{\sqrt{6}} \tag{46}$$

$$\beta_{nj2} = \frac{\beta^*_{nj2}}{\sigma_{\varepsilon nj}} \times \frac{\pi}{\sqrt{6}} \tag{47}$$

$$\delta_{nj2} = \frac{\delta^*_{nj2}}{\sigma_{\varepsilon nj}} \times \frac{\pi}{\sqrt{6}} \tag{48}$$

If σ_{V_1} and σ_{V_2} are the same then σ_{V_2} can be estimated using the following expression:

$$(\beta_{nj1}\sigma_{x_{nj1}} + \beta_{nj2}\sigma_{x_{nj2}}) = \sigma_{V_2} \tag{49}$$

The value obtained for the σ_{V_2} by expression (49) can now be compared with its estimate in expression (35) and (36).

References

Abramo, G., D'Angelo, C. A., & Caprasecca, A. (2009). Allocative efficiency in public research funding: Can bibliometrics help? *Research Policy*, *38*(1), 206–215.

Aksnes, D. W., & Taxt, R. E. (2004). Peer reviews and bibliometric indicators: A comparative study at a norwegian university. *Research Evaluation*, *13*(1), 33–41.

Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology*, *45*, 199–245.

Bornmann, L., & Daniel, H. D. (2005). Selection of research fellowship recipients by committee peer review. Reliability, fairness and predictive validity of board of trustees' decisions. *Scientometrics*, *63*(2), 297–320.

Bornmann, L., & Daniel, H. D. (2006). Selecting scientific excellence through committee peer review – A citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. *Scientometrics*, *68*(3), 427–440.

Bornmann, L., & Daniel, H. D. (2007). Convergent validation of peer review decisions using the *h* index – Extent of and reasons for type i and type ii errors. *Journal of Informetrics*, *1*(3), 204–213.

Bornmann, L., Mutz, R., & Daniel, H. D. (2010). A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PLoS One*, *5*(12).

Bornmann, L., Wallon, G., & Ledin, A. (2008). Does the committee peer review select the best applicants for funding? An investigation of the selection process for two european molecular biology organization programmes. *PLoS One*, *3*(10).

Cicchetti, D. V. (1991). The reliability of peer-review for manuscript and grant submissions – A cross disciplinary investigation. *Behavioral and Brain Sciences*, *14*(1), 119–134.

Franceschet, M., & Costantini, A. (2011). The first italian research assessment exercise: A bibliometric perspective. *Journal of Informetrics*, *5*(2), 275–291.

Gonzalez-Pereira, B., Guerrero-Bote, V. P., & Moya-Anegon, F. (2010). A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of Informetrics*, *4*(3), 379–391.

Hodgson, C. (1997). How reliable is peer review? An examination of operating grant proposals simultaneously submitted to two similar peer review systems. *Journal of Clinical Epidemiology*, *50*(11), 1189–1195.

Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2001). Peer review in the funding of research in higher education: The Australian experience. *Educational Evaluation and Policy Analysis*, *23*(4), 343–364.

Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2003). A multilevel cross-classified modelling approach to peer review of grant proposals: The effects of assessor and researcher attributes on assessor ratings. *Journal of the Royal Statistical Society Series A-Statistics in Society*, *166*, 279–300.

Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2006). A new reader trial approach to peer review in funding research grants: An Australian experiment. *Scientometrics*, *69*(3), 591–606.

Marsh, H. W., Bonds, N. W., & Jayasinghe, U. W. (2007). Peer review process: Assessments by applicant-nominated referees are biased, inflated, unreliable and invalid. *Australian Psychologist*, *42*(1), 33–38.

Marsh, H. W., Jayasinghe, U. W., & Bond, N. W. (2008). Improving the peer-review process for grant applications – Reliability, validity, bias, and generalizability. *American Psychologist*, *63*(3), 160–168.
Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, *4*(3), 265–277.
Nederhof, A. J., & van Raan, A. F. J. (1987). Peer-review and bibliometric indicators of scientific performance – A comparison of *cum laude* doctorates with ordinary doctorates in physics. *Scientometrics*, *11*(5–6), 333–350.
Reale, E., Barbara, A., & Costantini, A. (2007). Peer review for the evaluation of academic research: Lessons from the italian experience. *Research Evaluation*, *16*(3), 216–228.
Rinia, E. J., van Leeuwen, T. N., van Vuren, H. G., & van Raan, A. F. J. (1998). Comparative analysis of a set of bibliometric indicators and central peer review criteria – Evaluation of condensed matter physics in the netherlands. *Research Policy*, *27*(1), 95–107.
Taylor, J. (2011). The assessment of research quality in UK universities: Peer review or metrics? *British Journal of Management*, *22*(2), 202–217.
Train, K. (2009). (2nd ed.). *Discrete choice methods with simulation* (Vol. 2009) Cambridge University Press.
van Raan, A. F. J. (2006). Comparison of the hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, *67*(3), 491–502.
Vieira, E. S., Cabral, J. A. S., & Gomes, J. A. N. F. (2013). Definition of a model based on bibliometric indicators for assessing applicants to academic positions. *Journal of the American Society for Information Science and Technology*, http://dx.doi.org/10.1002/asi.22981/abstract (in press)
Vieira, E. S., & Gomes, J. A. N. F. (2011). An impact indicator for researchers. *Scientometrics*, *89*(2), 607–629.
Wood, M., Roberts, M., & Howell, B. (2004). The reliability of peer reviews of papers on information systems. *Journal of Information Science*, *30*(1), 2–11.