Predicting the results of evaluation procedures of academics

Francesco Poggi Corresp., 1, **Paolo Ciancarini** 1, **Aldo Gangemi** 2, **Andrea Giovanni Nuzzolese** 3, **Silvio Peroni** 2, **Valentina Presutti** 3

1 Computer Science and Engineering - DISI, Università di Bologna

2 Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

3 STLab, Institute of Cognitive Science and Technologies, National Research Council, Roma, Italy

Corresponding Author: Francesco Poggi
Email address: fpoggi@cs.unibo.it

Background. The 2010 reform of the Italian university system introduced the National Scientific Habilitation (ASN) as a requirement for applying to permanent professor positions. Since the CVs of the 59149 candidates and the results of their assessments have been made publicly available, the ASN constitutes an opportunity to perform analyses about a nation-wide evaluation process.

Objective. The main goals of this paper are: (i) predicting the ASN results using the information contained in the candidates' CVs; (ii) identifying a small set of quantitative indicators that can be used to perform accurate predictions.

Approach. Semantic technologies are used to extract, systematize and enrich the information contained in the applicants' CVs, and machine learning methods are used to predict the ASN results and to identify a subset of relevant predictors.

Results. For predicting the success in the role of associate professor, our best models using all and the top 15 predictors make accurate predictions (F-measure values higher than 0.6) in 88% and 88.6% of the cases, respectively. Similar results have been achieved for the role of full professor.

Evaluation. The proposed approach outperforms the other models developed to predict the results of researchers' evaluation procedures.

Conclusions. Such results allow the development of an automated system for supporting both candidates and committees in the future ASN sessions and other scholars' evaluation procedures.

Predicting the Results of Evaluation Procedures of Academics

Francesco Poggi[1], Paolo Ciancarini[1], Aldo Gangemi[2], Andrea Giovanni Nuzzolese[3], Silvio Peroni[2], and Valentina Presutti[3]

[1]**Department of Computer Science and Engineering (DISI), University of Bologna, Italy**
[2]**Department of Classical Philology and Italian Studies, University of Bologna, Italy**
[3]**STLab, Institute of Cognitive Science and Technologies - National Research Council, Italy**

Corresponding author:
Francesco Poggi[1]

Email address: francesco.poggi5@unibo.it

ABSTRACT

Background. The 2010 reform of the Italian university system introduced the National Scientific Habilitation (ASN) as a requirement for applying to permanent professor positions. Since the CVs of the 59149 candidates and the results of their assessments have been made publicly available, the ASN constitutes an opportunity to perform analyses about a nation-wide evaluation process.
Objective. The main goals of this paper are: (i) predicting the ASN results using the information contained in the candidates' CVs; (ii) identifying a small set of quantitative indicators that can be used to perform accurate predictions.
Approach. Semantic technologies are used to extract, systematize and enrich the information contained in the applicants' CVs, and machine learning methods are used to predict the ASN results and to identify a subset of relevant predictors.
Results. For predicting the success in the role of associate professor, our best models using all and the top 15 predictors make accurate predictions (F-measure values higher than 0.6) in 88% and 88.6% of the cases, respectively. Similar results have been achieved for the role of full professor.
Evaluation. The proposed approach outperforms the other models developed to predict the results of researchers' evaluation procedures.
Conclusions. Such results allow the development of an automated system for supporting both candidates and committees in the future ASN sessions and other scholars' evaluation procedures.

INTRODUCTION

Quantitative indicators have been extensively used for evaluating scientific performances of a given research body. International institutions, national authorities, research and funding bodies have an increasing interest on indicators, mainly based on bibliometric data, which can be used to algorithmically assess the performance of their institutions. SCImago[1] (for journals), the Performance Ranking of Scientific Papers for World Universities[2] and the Academic Ranking of World Universities[3] (for universities) are popular examples of rankings that use bibliometric indicators to rate scientific performances.

Peer review is still the Holy Grail for research evaluation, but the pressure for more frequent and extensive assessments of the performance of researchers, research groups and institutions makes bibliometry attractive. Currently, several countries use a combination of peer review and bibliometric indicators to allocate funding and evaluate the performance of higher education institutions. Examples of this mixed strategy are the Excellence in Research for Australia (ERA) and the Valutazione della Qualità della Ricerca (VQR) in Italy. The British Research Excellence Framework (REF), successor of the Research Assessment Exercise (RAE), is another example, in which experts can make use of citation data as an

[1] https://www.scimagojr.com/
[2] http://nturanking.lis.ntu.edu.tw/
[3] http://www.shanghairanking.com/

additional input of their reviews. In many countries bibliometric indicators are one of the factors that can be used for assessing individuals or institutions to allocate funding at national level. For instance, in Germany the impact factor of the publications is used in performance-based funding systems, in Finland the reallocation system uses the number of publications as one of the considered measures, in Norway a two-level bibliometric indicator is used for similar purposes, etc. (Vieira et al., 2014a).

The growing importance of quantitative indicators may be mainly explained by their advantages compared to peer review processes: objectivity, low time and implementation costs, possibility of quick and cheap updates, ability to cover a large number of individuals, etc. However, in many cases peer review is still the only method available in practice, and is hence intensively used in many situations. We know that bibliometric indicators are more accepted in the assessment of large research bodies, but they are still used frequently for individuals. It is therefore very important to benchmark bibliometric indicators against traditional peer assessments in real situations.

Some studies have been carried out in recent years with the main goal of finding a relation between the two methods at several levels. At national level, the relation between bibliometric indicators and the results of the Research Assessment Exercise (RAE) in Britain (Norris and Oppenheim, 2003; Taylor, 2011) or the Italian Triennial Assessment Exercise (VTR) (Abramo et al., 2009; Franceschet and Costantini, 2011) have been investigated. Other studies focused on the assessments of departments (Aksnes, 2003) and research groups (Van Raan, 2006). Just a few works have been made at the individual level (Nederhof and Van Raan, 1987; Bornmann and Daniel, 2006; Bornmann et al., 2008), while many analyzed the correlation between indicators and research performances (Leydesdorff, 2009; Franceschet, 2009). Recent works analyzed the correlation between traditional bibliometric indicators and altmetrics by also taking into account quality assessment procedures performed by peers (Nuzzolese et al., 2018; Wouters et al., 2015; Bornmann and Haunschild, 2018). All these works share the general finding that a positive and significant correlation exists between peer review and bibliometric indicators, and suggest that indicators can be useful tools to support peer reviews.

In this work we investigate the relation between quantitative indicators and peer review processes from a different perspective. The focus of the study is to analyze if and to what extent **quantitative indicators can be used to predict the results of peer reviews**. This problem is interesting for many different reasons. First of all, since an high number of factors are involved in peer review processes (e.g. cultural, social, contextual, scientific, etc.), the feasibility of reproducing such a complex human process through computational and automatic methods is a relevant topic per se. Moreover, the possibility of predicting human assessments has many practical applications. Having an idea of the results of an evaluation procedure may be very useful for candidates (e.g. to understand if they are competitive for a given position, to decide if to apply or not, etc.). Also reviewers can benefit of such information (e.g. for supporting a first screening of the candidates, for spotting possible errors to investigate, etc.). In other words, the final goal of our work is not substituting peer committees by automatic agents, but **providing tools for supporting both candidates and reviewers in their tasks**.

This study analyzes the Italian National Scientific Habilitation (ASN)[4], a nation-wide research assessment procedure involving a large number of applicants from all academic areas. The ASN is one of the main novelties in the national university system introduced by Law 240/2010 (Law dec. 30, n. 240, 2011), and it is similar to other habilitation procedures already in place in other countries (e.g., France and Germany) in that it is a prerequisite for becoming a university professor. The ASN is meant to attest that an individual has reached the scientific maturity required for applying for a specific role (associate or full professor) in a given scientific discipline; however, the qualification does not guarantee that a professorship position will eventually be granted. The assessments of the candidates of each discipline are performed by committees composed of four full professors from Italian universities and one professor from a foreign research institution. The evaluation is performed considering the CVs submitted by the applicants and three quantitative indicators computed for each candidate.

The first session of the ASN started on November 2012 and received 59149 applications spanning 184 Recruitment Fields (RFs), which correspond to scientific fields of study in which Scientific Areas (SAs) are organized. The curricula of all applicants, the values of their bibliometric indicators and the final reports of examination committees have been made publicly available. This work focuses on the

[4]The acronym ASN stands for *Abilitazione Scientifica Nazionale*. For the rest of the paper, all acronyms (e.g. ASN, MIUR, ANVUR, etc.) are based on the original Italian names, since they are well established in the Italian scientific community. The English translations are also provided for the benefit of the international readers.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27582v1 | CC BY 4.0 Open Access | rec: 12 Mar 2019, publ: 12 Mar 2019

analysis of applicants' curricula. For this purpose, we processed this vast text corpus, extracted the contained information and used it to populate a Knowledge Graph by exploiting semantic technologies. This Knowledge Graph contains a collections of relevant data for each applicant and it has then been used to perform different kinds of analyses at the level of category of discipline (i.e. *bibliometric* and *non-bibliometric*), Scientific Area, and RF.

An approach based on machine learning techniques has been used to answer the following research questions:

- *RQ1:* Is it possible to predict the results of the ASN using only the information contained in the candidates' CVs?
- *RQ2:* Is it possible to identify a small set of predictors that can be used to predict the ASN results?

The rest of the work is organized as follows. Section 'Related Work' presents an overview of the related work. In Section 'The Italian Scientific Habilitation', we introduce our case study by providing the necessary background on the ASN. Section 'Methods and Material' gives an overview of the ASN dataset, describes the techniques used to analyze and enrich this text corpus, and introduces the ontology developed for storing the information in a semantic format. In Section 'Results' we describe the results of the analyses performed to answer the two aforementioned research questions, and in Section 'Evaluation' we evaluate our work by comparing the predictive power of our approach with others at the state of the art. Finally, in the last section we discuss the results and draw some conclusions.

RELATED WORK

Quantitative indicators have been extensively used for evaluating the scientific performance of a given research body. Many recent studies have focused on the predictive power of such indicators for different purposes. These works can be divided in two main groups: those that use bibliometric indicators to predict other indicators and those that use bibliometric indicators to predict the results of evaluation procedures performed through a peer review process or a mixed strategy (i.e. a combination of peer review and bibliometric indicators). The next subsections discuss the main recent works on this topic. To facilitate the readers, Table 1 summarizes the main information about them and our study.

Prediction of Bibliometric Indicators

A first challenge concerns the problem of identifying a subset of bibliometric indicators for predicting other bibliometric indices. Ibáñez et al. (2016) introduced an approach based on Gaussian Bayesian networks to discover multivariate relationships among bibliometric indices and identify the best subset of predictive variables. The approach has been tested on the data of 280 Spanish full professors of Computer Science. For each professor, 12 bibliometric indicators have been computed (i.e. number of documents and citations, h-index, g-index, hg-index, a-index, m-index, q^2-index, h_r-index, h_i-index, h_c-index, c-index) by querying Thomson Reuters Web of Knowledge. The network that best performs uses a bibliometric core composed of four indicators (i.e. citations, the g-index, the q^2-index, and the h_r-index) A second analysis is performed on these core indicators using Gaussian Bayesian networks learned by a genetic algorithm to look for the optimal model that best predicts the other eight bibliometric indices. The main drawback of the work is that no evaluation is presented: only a test on a small sample composed of three cases is discussed in the paper.

Other works focused on the prediction of papers citations. Danell (2011) used previous publication volume and citation rate of authors to predict the impact of their articles. The aim of this work is to investigate whether evaluations systems based on researchers' track records actually reward excellence. For this purpose the work focused on the authors of two disciplines (i.e. episodic memory research and Bose-Einstein condensate) and developed a model based on quantile regression to predict their relative citation rate. The results indicate that previous publication volume has no significant effect on the citation rate of articles. A better predictor of the impact of the articles was achieved using previous citation rate. In particular, this measure has a very high accuracy in predicting who will write an highly cited article, while it is not very accurate in predicting who will write a median-cited or an uncited article. The author concludes that selecting researchers based on how much they have previously written does not seem to be a good strategy for a selective research policy, while it would be better to consider citation rate to identify future excellence.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27582v1 | CC BY 4.0 Open Access | rec: 12 Mar 2019, publ: 12 Mar 2019

Table 1. Comparison of the related work with our study. Missing data are labeled with "n.a.". PoC stands for "Prediction of Citations", PoPJ for "Prediction of Peer Judgements, and AoH for "Analysis of H-index for peer judgements".

Work	Papers	Authors	Discipline	Predictors	Task	Method
Ibáñez et al. (JASIST, 2016)	n.a.	280	Computer Science	12	PoC	Gaussian Bayesian networks
Danell (JASIST, 2011)	6030	8149	Neuroscience and Physics	2	PoC	Quantile regression
Fu and Aliferis (Scientometrics, 2010)	3788	n.a.	Medicine	12 (+ textual features)	PoC	Support vector machines
Lindahl (J.of Informetrics, 2018)	n.a.	406	Mathematics	4	PoC	Logistic regression
Bornmann and Daniel (J.of Informetrics, 2007)	n.a.	414	Biomedicine	1	AoH	Correlation analysis
Van Raan (Scientometrics, 2006)	n.a.	700	Chemistry	1	AoH	Correlation and error analysis
Cronin and Meho (JASIST, 2006)	n.a.	31	Information Science	1	AoH	Correlation analysis
Vieira et al. (JASIST, 2014a)	7654	174	Hard sciences	3 (based on 12 bibl. indices)	PoPJ	Rank ordered regression logic
Jensen et al. (Scientometrics, 2009)	n.a.	3659	All	8	PoPJ	Binomial regression
Tregellas et al. (PeerJ, 2018)	n.a.	363	Biomedicine	10 (3 for the best model)	PoPJ	Logistic regression, Support vector machines
This work	1910873	59149	All	326	PoPJ	Support vector machines (CFS for feature selection)

Another work (Fu and Aliferis, 2010) faces the problem of predicting the number of citations that a paper will receive within an horizon of ten years using only the information available at publication time. The authors investigate the predictive power of three different approaches (i.e. support vector machines, logistic regression and decision trees) testing them on a dataset composed of 3788 biomedical articles. The experiments show that it is feasible to predict future citation counts with a mixture of content-based features (approximately 20000 features have been extracted from the text of the papers) and bibliometric features (e.g. journal impact factor, number of authors and institutions, number of articles and citations for the first and last author) using machine learning methods.

A recent work (Lindahl, 2018) examines the relation between bibliometric indicators and research excellence. In particular the authors investigated the ability of four indices (i.e. publication rate, top journal publications, average number of coauthors and highly cited publications) computed on the first four years of the career to predict whether an author will attain excellence (operationalized by the percentile-based indicator defined in (Bornmann, 2013) as a measure for highly cited papers) in the following four years. The dataset of the study consisted of track records collected from the MathSciNet database of 406 early career mathematicians in the field of number theory with at least one paper between 2000 and 2003. Logistic regression and dominance analysis was conducted on the data, producing the following rank of the predictors based on their relative importance (i.e. contribution to model fit): 1. top journal publications (which contributed 44.51% to model fit); 2. top 10% publications (which contributed 43.58%); 3. publication rate (which contributed 11.91%). The major conclusions were that publishing many articles in top journals is the most important factor in the process of achieving excellence in the early career, followed by having an high publication rate (which is an implicit requirement of the previous factor) and publishing many papers.

Prediction of the Results of Evaluation Procedures

Only a few works focused on the problem of using bibliometric indicators to predict the results of evaluation procedures performed through peer-review processes. Vieira et al. (2014a) compare three models for predicting the success of applicants to academic positions. The test dataset is composed of

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27582v1 | CC BY 4.0 Open Access | rec: 12 Mar 2019, publ: 12 Mar 2019

the track records of 174 candidates to 27 selection processes for associate and full professor in hard sciences that took place in Portugal between 2007 and 2011. The areas of Chemistry, Physics, Biology, Mathematics, Mechanics, Geology, and Computer Science were considered. In all cases, candidates have been assessed by a panel of peers, producing a ranking of the applicants. Starting from 12 bibliometric indicators (i.e. number of documents, percentage of cited, highly cited and citing documents, average number of authors, h_{nf}-index, NIR, SNIP, SJR, percentage of international collaborations, normalized impact and the number of Scimago's Q1 journals) a few composite indices have been derived through a factor analysis. Following a discrete choice model, three predictive models based on Rank Ordered Logistic Regression (ROLR) have been defined. The best model is able to predict the applicants placed in the first position by peers in 56% of the cases. By considering the problem of predicting the relative position of two candidates (i.e. who will be ranked in the higher position), the best model is able to predict 76% of the orderings. In another work (Vieira et al., 2014b), the performances of these models have been compared with a random model, observing that in 78% of the cases the applicant placed in first position by peers has a probability of being placed first that is better than chance. The authors conclude that the predictions provided by the models are satisfactory, and suggest that they can be used as an auxiliary instrument to support peer judgments.

Another work tested the predictive power of eight bibliometric indicators for predicting scientists promotions (Jensen et al., 2009). The dataset used in the study is composed of the track records of 3659 CNRS researchers from all disciplines that have filled the CNRS report between 2005 and 2008, whose data has been obtained by querying the Web of Science database. In the same timespan, the promotions of about 600 CNRS researchers at all the five CNRS levels have been considered. A binomial regression model (logit) has been used to assess the overall relevance of eight quantitative indicators (h-index, normalized h-index, number of publications and citations, mean citations per paper, h-index per paper, age, gender) and to study their dependence. The results showed that the h-index is the best index for predicting the promotions, followed by the number of publications. Differences exist between disciplines: in Engineering, for instance, the number of publications is the best predictor. A logit model based on the best overall predictor (i.e. h-index) has been tested for each subdiscipline, leading to correct predictions in 48% of the cases. The authors conclude that bibliometric indicators do much better than randomness, which would achieve 30% of guessed promotions.

A recent study (Tregellas et al., 2018) focused on the problem of predicting career outcomes of academics using the information in their publication records. The objective of the work is to identify the main factors that may predict the success of young researchers in obtaining tenure-track faculty research positions. The dataset used in this study is composed of the track records of 363 PhD graduates from biomedical sciences programs at the University of Colorado from 2000 to 2015. The ratio of faculty/non-faculty members (i.e. individuals employed/not employed in faculty positions) is 12%. For each PhD graduate, 10 indicators has been computed (i.e. sex, date of graduation, number of first-author and non-first-author publications , average impact factor of first-author and non-first-author publications, highest impact factor of first-author and non-first-author publications, weighted first-author and non-first-author publication count). Logistic regression models and support vector machines has been used to investigate and compare the ability of the aforementioned indicators to predict career outcomes. The best prediction has been performed by the logistic regression model using three predictors (i.e. sex, date of graduation and weighted first-author publication count), showing 73% accuracy . A similar result (i.e. 71% accuracy) has been obtained by the best model based on support vector machines using the same predictors. The results suggest that, while sex and months since graduation also predict career outcomes, a strong predoctoral first-author publication record may increase likelihood of obtaining an academic faculty research position. The analysis of the results also showed for all models high negative predictive values (i.e. high accuracy in predicting those who will not obtain a faculty position), while low positive predictive values. This suggest that first-author publications are necessary but not sufficient for obtaining a faculty position. The main limitation of the study concerns the dataset size, since it was conducted on a small set of individuals at only one institution, focusing on a single discipline. The authors observe that it is then necessary to determine how generalizable the current findings are. Finally, the fact that all the best models are less than 75% accurate suggests that variables other than those considered here are also likely to be important factors in predicting future faculty status.

Other empirical studies focused on a single indicator (i.e. the h-index) to assess how it correlates with peer judgements. These works have the main limitation of being carried out on small samples for

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27582v1 | CC BY 4.0 Open Access | rec: 12 Mar 2019, publ: 12 Mar 2019

Table 2. The 14 Italian scientific areas. For each we report the numeric ID, a three-letter code, the name of the area and the number of RFs it contains.

ID	Code	Area Name	N. of Recr. Fields
01	MCS	Mathematics and Computer Sciences	7
02	PHY	Physics	6
03	CHE	Chemistry	8
04	EAS	Earth Sciences	4
05	BIO	Biology	13
06	MED	Medical Sciences	26
07	AVM	Agricultural Sciences and Veterinary Medicine	14
08	CEA	Civil Engineering and Architecture	12
09	IIE	Industrial and Information Engineering	20
10	APL	Antiquities, Philology, Literary Studies, Art History	19
11	HPP	History, Philosophy, Pedagogy and Psychology	17
12	LAW	Law	16
13	ECS	Economics and Statistics	15
14	PSS	Political and Social Sciences	7
	Total		**184**

technical reasons (i.e. the difficulty of obtaining large sets of robust bibliometric data). In practice, they were generally limited to a single discipline: Bornmann and Daniel (2007) studied 414 applications to long-term fellowships in biomedicine, Van Raan (2006) analyzed the evaluation of about 700 researchers in chemistry, Cronin and Meho (2006) studied 31 influential information scientists from the US.

To the best of our knowledge, no other work analyzed the predictive power of quantitative indicators for predicting the results of peer judgments of researchers.

THE ITALIAN SCIENTIFIC HABILITATION

The Italian Law 240/2010 (2011) introduced substantial changes in the national university system. Before 2010, in the Italian universities there were three types of tenured positions: assistant professor, associate professor and full professor. The reform suppressed the position of assistant professor and replaced it with two types of fixed term positions called type A and type B researcher. Type A positions last for three years and can be extended for other two years. Type B positions last for three years and have been conceived as a step for becoming tenured associate professor, since at the time of recruitment universities must allocate resources and funding for the promotion. Each academic is bound to a specific Recruitment Field (RF), which corresponds to a scientific field of study. RFs are organized in groups, which are in turn sorted in 14 Scientific Areas (SAs). In this taxonomy defined by Decree 159 (Ministerial Decree 159, 2012), each of the 184 RFs is identified by an alphanumeric code in the form AA/GF, where AA is the ID of the SA (in the range 01-14), G is a single letter identifying the group of RFs, and F is a digit denoting the RF. For example, the code of the RF ”Neurology” is 06/D5, which belongs to the group ”Specialized Clinical Medicine” (06/D), which is part of the SA ”Medicine” (06). The 14 SAs are listed in Table 2, and the 184 RFs are listed in Appendix A (Poggi et al., 2018b).

Under the new law, only people that attained the National Scientific Habilitation (ASN) can apply for tenured positions in the Italian university system. It is important to note that an habilitation does not guarantee any position by itself. The ASN has indeed been conceived to attest the scientific maturity of researchers and is a requirement for accessing to a professorship in a given RF. Each university is responsible for creating new positions for a given RF and professional level provided that financial and administrative requirements are met, and handles the hiring process following local regulations and guidelines.

The first two sessions of the ASN took place in 2012 and 2013. Although the Law 240/2010 prescribes that the ASN must be held at least once a year, the next sessions took place in 2016 (1 session), 2017 (2 sessions) and 2018 (2 sessions). At the time of the writing of this article the last session of the 2018 ASN was still in progress, and the dates of the next sessions have not yet been set. For each of the 184 RFs, the Ministry of University and Research (MIUR) appoints an examination committee for the evaluation of the candidates. The committees are composed of five full professors who are responsible for the evaluation of the applicants for associate and full professor. Committee members are randomly selected from a list of eligible professors, for a total of 920 professors. Different committees have been appointed for 2012, 2013 and 2016-18 sessions, respectively.

In order to apply to a session of the ASN, candidates have to submit a curriculum vitae with detailed information about their research activities. Although the ASN is bound to a specific RF and professional level, it is possible to apply in different RFs and roles. In 2012, for example, 136/260 (52.3%) applicants for full professor in the RF 09/H1 (Information Processing Systems) also applied to 01/B1 (Informatics). Those who fail to get an habilitation cannot apply again to the same RF and level in the next session. Once acquired, an habilitation lasts for six years.

The ASN introduced two types of parameters called *bibliometric* and *non-bibliometric* indicators, respectively. Bibliometric indicators apply to scientific disciplines for which reliable citation databases exist. The three bibliometric indicators are:

- Normalized number of journal papers
- Total number of citations received
- Normalized h-index

Since citations and paper count increase over time, normalization based on the scientific age (the number of years since the first publication) is used to compute most of the indicators. The aforementioned indicators are used for all RFs belonging to the first nine SAs (01-09), with the exception of the RFs 08/C1, 08/D1, 08/E1, 08/E2, 08/F1 and the four RFs belonging to the group Psichology (11/E). These RFs are collectively denoted as *bibliometric disciplines*.

Non-bibliometric indicators apply for the RFs for which MIUR assessed that citation databases are not ”sufficiently complete”, and hence bibliometric indices can not be reliably computed. The three non-bibliometric indicators are:

- Normalized number of published books
- Normalized number of book chapters and journal papers
- Normalized number of paper published on ”top” journals

These are used for all RFs belonging to the last five SAs (10-14) with the exceptions described above. These RFs are denoted as *non-bibliometric* disciplines. It is important to remark that this terminology (i.e. ”bibliometric” and ”non-bibliometric”) is used in the official MIUR documents but it is not consistent with that used by the scientometric community. Non-bibliometric indicators, for instance, are indeed bibliometric being based on paper counts. Given that these terms became standard within the Italian research community, we will follow the MIUR ”newspeak” according to the definitions above.

The values of the indicators for each candidate were computed by the National Agency for the Assessment of Universities and Research (ANVUR), a public agency established with the objective of assessing Italian academic research. Data from Scopus[2] and Web of Science[3] were used for this computation, and only publications in a time window of ten years before the ASN session were considered. The computed indicators and the candidates’ CVs are the only information provided to the evaluation committees for their assessments. The 2012 session of the ASN has been analyzed by a quantitative point of view in (Marzolla, 2015).

METHODS AND MATERIAL

This section describes the dataset of the applicants’ CVs submitted to the 2012 session of the ASN, the Academic Career (AC) ontology, which is the model we developed for representing the information contained in the applicants’ CVs, and the PDF to Academic Career Ontology (PACO) converter, the software tool we developed to extract such information and produce a Knowledge Graph which conforms to the AC ontology. This Knowledge Graph is the starting point of the analyses presented in the following section.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27582v1 | CC BY 4.0 Open Access | rec: 12 Mar 2019, publ: 12 Mar 2019

Table 3. The number of applications for associate and full professor for each session of the ASN.

Session	Associate Professor	Full Professor	Total
2012	41088	18061	59149
2013	11405	5013	16418
2016	13119	7211	20330
2017a	3254	1515	4769
2017b	2501	1322	3823
2018a	5176	2445	7261
Total	**76543**	**35567**	**112110**

ASN Data

The number of applications submitted to the six sessions of the ASN are reported in Table 3. We decided to focus our analysis on the 2012 session of the ASN because: (i) it is a representative and exhaustive sample of the whole population; (ii) since in the following sessions different committees have been appointed, in this way we exclude biases and other problems introduced by changes in the evaluation committees.

Overall, the 2012 session of the ASN received 59149 applications spanning 184 RFs. For each applications, we collected three different documents: the CV, the official document with the values of the three quantitative indicators described in the previous section and the final reports of the examination committee. These documents are in PDF, and have been made publicly available on the ANVUR site for a short period of time. Some basic information and statistics about the 2012 ASN session are summarized in Appendix B (Poggi et al., 2018b).

Since ANVUR did not provide a template for the habilitation, the CVs are very heterogeneous, varying in terms of formatting, internal structure and organization. This heterogeneity and the massive amount of information contained in the 59149 PDFs are two of the main challenges faced in this project. In order to manage this problem we developed an ontology which provides an uniform representation of the information and a reference conceptual model. It is the basis of both the data processing and subsequent analyses, as described in the following sections.

Ontology Description

The objective of the Academic Career (AC) ontology is to model the academic career of scholars. AC is an OWL2 (W3C, 2012) ontology composed of fifteen modules, each of which is responsible for representing a particular aspect of the scientific career of a scholar. The first two modules of the AC ontology concern personal information and publications. The next modules pertain to ten categories suggested by ANVUR:

1. Participation to scientific events with specific roles (eg. speaker, organizer, attendee, etc.)
2. Involvement and roles in research groups (management, membership, etc.)
3. Responsibility for studies and researches granted by qualified institutions
4. Scientific responsibility for research projects
5. Direction or participation to editorial committees
6. Academic and professional roles
7. Teaching or research assignments (fellowships) at qualified institutes
8. Prizes and awards for scientific activities
9. Results of technological transfer activities (e.g. spin-offs, patents, etc.)
10. Other working and research experiences

The last three modules concern scholars' education, scientific qualifications, and personal skills and expertises.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27582v1 | CC BY 4.0 Open Access | rec: 12 Mar 2019, publ: 12 Mar 2019

Data Processing

The processing of a vast set of documents such as the corpus of the ASN curricula is not a trivial task. The main issue to face in this process is the management and harmonization of its heterogeneity in terms of kinds of information, structures (eg. tables, lists, free text), styles, languages, just to cite a few. Nonetheless, the automatic extraction of information from CVs and its systematization in a machine processable format is a fundamental step for this work, since all the analyses described in Section "Results" are based on these data.

For this purpose, we developed PDF to Academic Career Ontology (PACO), a software tool that is able to process the researchers' CVs, extract the most relevant information, and produce a Knowledge Graph that conforms to the AC ontology. The processing performed by PACO is composed of four consecutive steps, that correspond to the software modules constituting PACO's architecture, as shown in Figure 1. The processing of an applicant's CV can be summarized as follows:

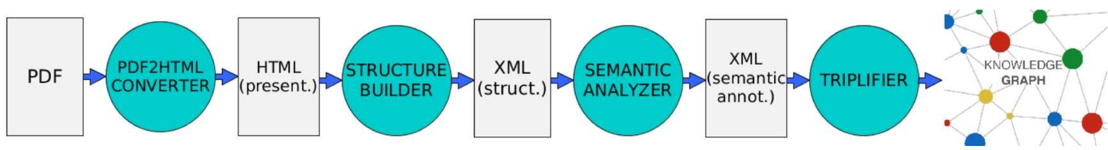


Figure 1. An overview of the architecture of the PACO converter composed of four sub-modules (blue circles).

- **HTML conversion:** The *PDF2HTML converter* takes as input a PDF and produces as output an HTML version of the CV composed of inline elements and presentational elements. The structure of the document is not reconstructed in this phase. In particular, the containment relations between elements (e.g. cells in a table, items in a list, etc.) are missing. For instance, a table is converted into a series of rectangles with borders (the cells) followed by a series of inline elements (the text). All the elements are at same level in the output document hierarchy, and no explicit relation between them is maintained.

- **Structure re-construction:** the *Structure Builder* uses the presentational information computed in the previous phase to infer the structure of the document. Different strategies have been developed to recognize meaningful patterns in the presentation and reconstruct the document hierarchy. For example, a mark positioned near an inline element containing text is interpreted as a list item, a sequence of consecutive list items is interpreted as a list. The output is an XML document, in which the original textual content is organized in meaningful structural elements.

- **Semantic analysis:** the objective of the *Semantic Analyzer* is to annotate the output of the previous phase with information about its content. For example, it has to infer if a list is a list of publications, awards, projects, etc. A series of analyses is performed for each element, from simple ones (e.g. to test if an element contains a name, surname, birth date, etc.) implemented through basic techniques such as the use of heuristics or pattern matching, to more complex ones (e.g. to identify publications, roles, etc.) implemented using external tools and libraries. Another important technique is to leverage the homogeneity of structured elements (e.g. of all the items in a list or of all the cells of a column) to infer meaningful information about their content, using the approach described in (Poggi et al., 2016). The basic idea is that, for instance, if the majority of the elements of a list have been recognized as publications, it is then reasonable to conclude that also the others are publications. The output of this phase is an XML document annotated with the results of the semantic analysis.

- **Triplification:** the *Triplifier* is responsible of populating a Knowledge Graph with the information inferred in the previous phase. The marked XML document is the input of this stage, and the output is a Knowledge Graph that conforms to the AC ontology.

The data extracted from the applicants' CVs by PACO have also been semantically enriched with information from the following external sources:

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27582v1 | CC BY 4.0 Open Access | rec: 12 Mar 2019, publ: 12 Mar 2019

- Cercauniversita[5]: for information about the candidates' careers within the Italian university system;
- TASTE database[6]: for data about reserchers' entrepreneurship and industrial activities from the TASTE database;
- Semantic Scout[7]: for information about researchers of the Italian National Council of Research (CNR).

The final outcome of this process is the Knowledge Graph from which we computed the predictors used in the analyses discussed in the next section.

RESULTS

The aim of the analyses presented in this section is to answer the two Research Questions (RQs) dicussed in Section "Introduction". Given the huge amount of data provided by the curricula of the applicants, we want to understand if machine learning techniques can be used to effectively distinguish between candidates who got the habilitation and those who did not (RQ1). We are also interested in identifying a small set of predictors that can be used to perform accurate predictions for the different RFs and scientific levels of the ASN (RQ2).

Analysis of the Recruitment Fields and Areas

In order to implement a supervised learning approach, we needed to create a training set in which the ground truth is obtained from the final reports of the examination committees. The instances of our dataset correspond to the 59149 applications submitted to the 2012 ASN. For each instance, we collected 326 predictors, 309 of which are numeric and 17 are nominal. The only source of data used to build our dataset is the Knowledge Graph containing the data extracted from the applicants' curricula and enriched with external information.

The predictors that have been computed belong to one of the following two categories:

- numeric and nominal values extracted from the CVs (e.g. the number of publications) or derived from the CVs using external sources (e.g. the number of journal papers has been computed using the publication list in the CVs and querying online databases like Scopus);
- quantitative values calculated using the values from the previous point. For example, we computed statistical indicators such as the variance of the number of journal papers of each applicant in the last N years.

The aforementioned 326 predictors and the habilitation class feature are our starting point to investigate the performances of different machine learning approaches. We decided not to explicitly split the dataset in training and test sets, and systematically rely on cross-fold validation instead. In particular, the data reported in this work are related to the 10-fold validation, but we have also performed a 3-fold one with very similar results.

The following supervised machine learning algorithms have been tested:

- **NB:** Naïve Bayes (John and Langley, 1995)
- **KN:** K-nearest neighbours classifier (K chosen using cross validation) (Aha et al., 1991)
- **C45:** C4.5 decision tree (unpruned) (Quinlan, 2014)
- **RF:** Random Forest (Breiman, 2001)
- **SVM:** Support Vector Machine trained with sequential minimal optimization (Keerthi et al., 2001)

The rationale behind this choice is to have representatives for the main classification methods that

[5] http://cercauniversita.cineca.it/ is a MIUR service that provides information and statics about Italian professors, universities, degree programs, students, fundings, etc.

[6] TAking STock: External engagement by academics (TASTE) is an European project founded under the FP7 program that developed a database with data about the relation between universities and enterprises in Italy - see https://eventi.unibo.it/taste

[7] Semantic Scout is a service that provides CNR scientific and administrative data in a semantic format - see http://stlab.istc.cnr.it/stlab/project/semantic-scout/

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27582v1 | CC BY 4.0 Open Access | rec: 12 Mar 2019, publ: 12 Mar 2019

Table 4. Performance of the machine learning algorithms investigated for the classification of the applicants to the RF 11/E4 (level II). For each algorithm we report Precision, Recall and F-Measure values.

	Precision	Recall	F-measure
NB	0.856	0.850	0.853
KN	0.867	0.906	0.886
C45	0.865	0.914	0.888
RF	0.844	1.000	0.916
SVM	0.894	0.951	0.922

have shown effectiveness in past research. DFE has been introduced because it is known to provide a good Bayesian approach for feature-rich datasets like the one we are dealing with.

All learners have been tuned using common best practices. SVM has been tested with various kernels (in order to account for complex non-linear separating hyperplanes). However, the best results were obtained with a relatively simple polynomial kernel. The parameters for the resulting model have been tuned using the grid method (He and Garcia, 2009). We tested the learners on different data samples obtaining similar results for both bibliometric and non-bibliometric RFs. For example, Table 4 shows the results we obtained with these machine learning algorithms for the applicants to the RF 11/E4 (level II).

Notice that we tested the performances of the learners only with respect to the not qualified class. We do that because we are mainly interested in understanding if we can use machine learning techniques to identify unsuccessful applicants who got not qualified. We are also reporting a limited amount of analysis data, specifically in this work we focus on precision and recall (and the related F-measure). Other aspects of the learners (such as the ROC curve) have been analyzed in our tests but they were always aligned with the results expressed by the three measure we are providing here. The results show that the best classifiers are those known to perform better on feature-rich datesets. In particular, SVM outperforms the others classification methods, and for this reason has been used in the rest of our analyses.

The objective of the first experiment is to predict the results of the ASN (RQ1). To this end, we classified our dataset with respect to the class of candidates who got the habilitation using the SVM learner. We first split the dataset in two partitions containing the data about candidates for level I and level II, respectively. For each partition, we classified separately the applicants of each RF. The results of our analysis are published in (Poggi et al., 2018a), and are summarized by the boxplots in Figure 2. The boxplot is a method for graphically depicting the distribution of data through their quartiles. The central rectangle spans the first quartile to the third quartile. The segment inside the rectangle shows the median, and ”whiskers” above and below the box show the locations of the minimum and maximum.

From these results we observe that the performance of the learners for bibliometric and non-bibliometric RFs are very similar, and that they are distributed evenly (i.e. there is not a polarization of bibliometric and non-bibliometric RFs). Moreover, we note that 154/184 (83.7%) and 162/184 (88%) RFs have F-measure scores higher than 0.6 for professional level I and II, respectively.

We also investigated the performance of the SVM learner on the data partitioned in the scientific areas in which RFs are organized. To do so, we split the dataset in 16 partitions: nine for bibliometric SAs (01-09), one for the macro sector 11/E (Psicology) which is bibliometric, five for non-bibliometric SAs (10-14), and one for the RFs 08/C1, 08/D1, 08/E1, 08/E2 and 08/F1 which are non-bibliometric.

The results for both professional levels are summarized in Figure 3, and the whole data are reported in (Poggi et al., 2018a). Also in this case, results are very accurate for both bibliometric and non-bibliometric disciplines, with F-measure scores spanning from a minimum of 0.622 (07-AVM) and 0.640 (02-PHY) for professionals level I and II, and a maximum of 0.820 (11-HPP) and 0.838 (14-PSS) for professional levels I and II. We observe that, at the associate professor level, the performance for non-bibliometric SAs (Figure 3d) are significantly better than for bibliometric SAs (Figure 3c). Moreover, the variance of the values is much lower for non-bibliometric SAs, as showed by the boxplots which are significantly more compressed.

Analysis of the Quantitative Indicators of Applicants

The objective of the next experiment is to identify a small set of predictors that allows to perform accurate predictions of the ASN results (RQ2). To this end, we analyzed the relevance of the various predictors for

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27582v1 | CC BY 4.0 Open Access | rec: 12 Mar 2019, publ: 12 Mar 2019

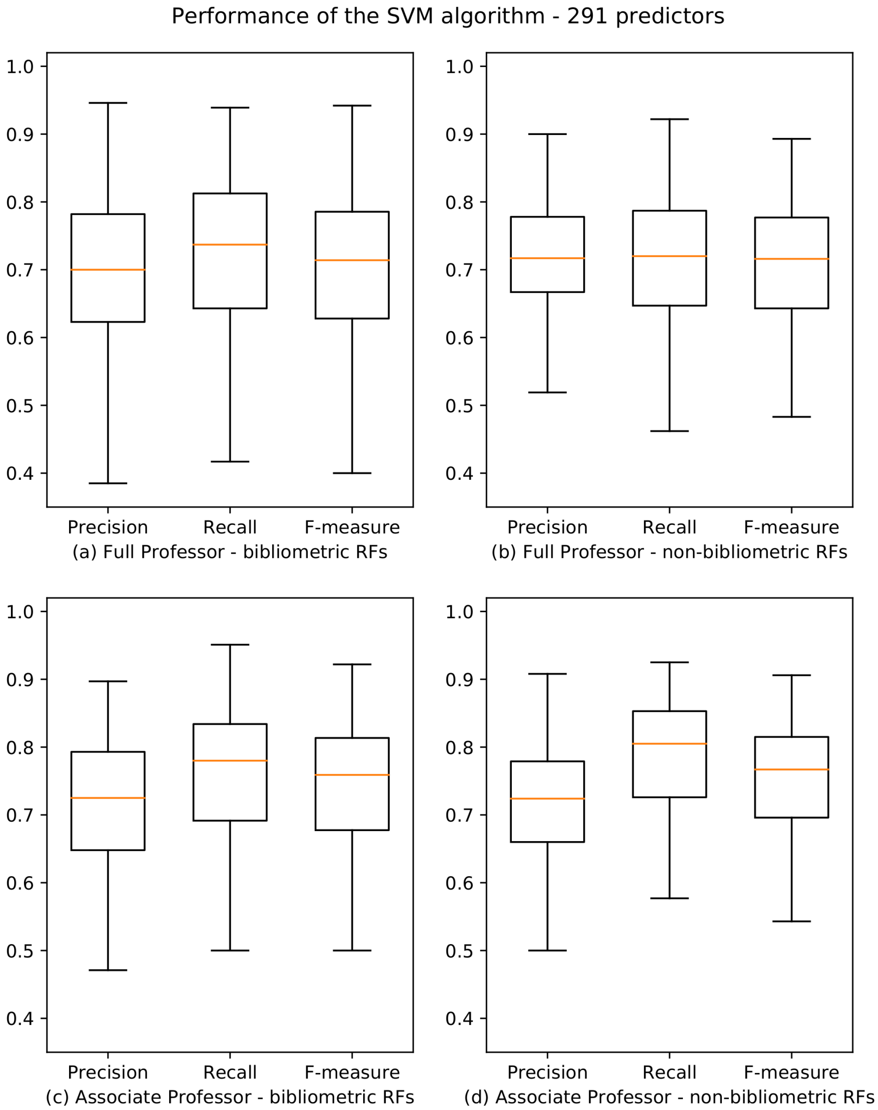


Figure 2. Boxplots depicting the performance of the SVM algorithm for academic level I and II. Precision, Recall and F-measure values are reported for bibliometric (a,c) and non-bibliometric (b,d) RFs.

classification purposes. In case of large number of predictors, several attribute engineering methods can be applied. The most widely adopted is attribute selection, whose objective is identifying a representative set of attributes from which to construct a classification model for a particular task. The reduction of the number of attributes can help learners that do not perform well with a large number attributes. This helps also in reducing the computation time needed to create the predictive model.

There are two main classes of attribute selection algorithms: those who analyze the performance of the learner in the selection process (i.e. wrappers) and those who do not use the learner (i.e. filters). The first class is usually computationally expensive since the learner runs continuously to check how it performs when changing the attributes in the dataset. That leads to computation times that are two or more orders of magnitude larger compared to the learner itself. For this reason, we did only some limited experiments with learner-aware attribute selection. In our test cases the results obtained were marginally better than those obtained with processes not using the learner. Consequently, we used a filter-based approach in our in-depth analysis.

We used Correlation-based Feature Selection (CFS) (Hall and Holmes, 2003), which is the first method that evaluates (and hence ranks) subsets of attributes rather than individual attributes. The central

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27582v1 | CC BY 4.0 Open Access | rec: 12 Mar 2019, publ: 12 Mar 2019

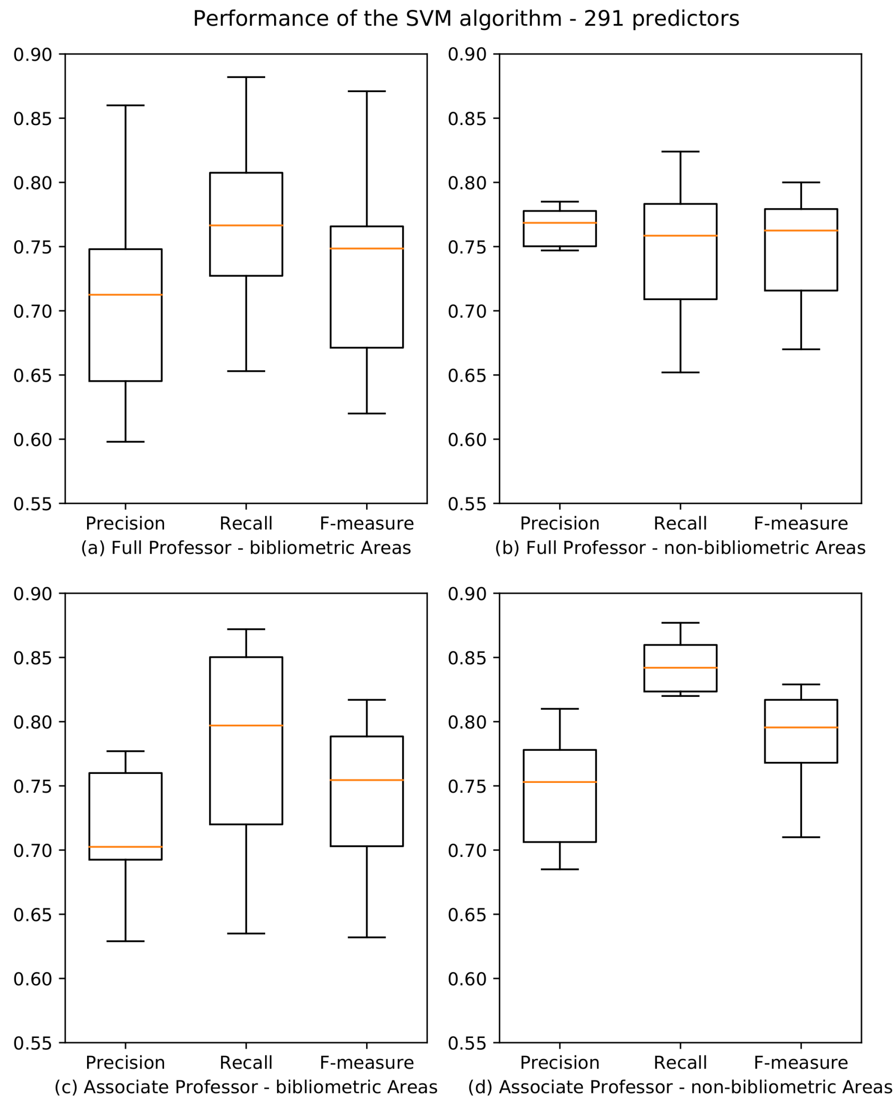


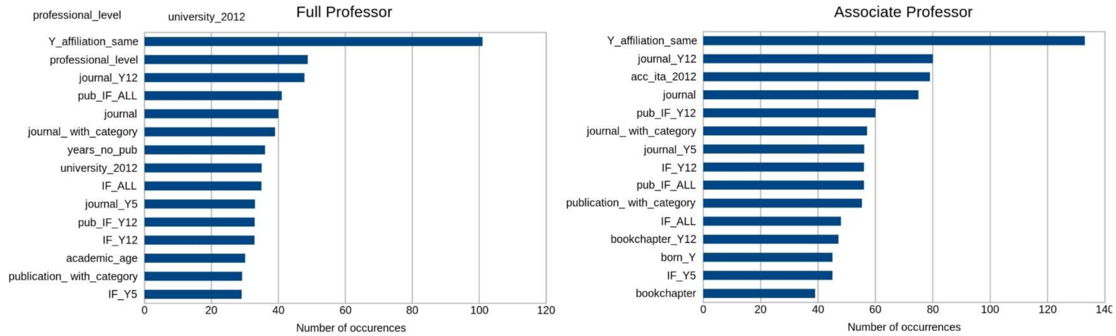
Figure 3. Boxplots depicting the performance of the SVM algorithm for academic level I and II. Precision, Recall and F-measure values are reported for bibliometric (a,c) and non-bibliometric (b,d) SAs.

hypothesis of this approach is that good attribute sets contain attributes that are highly correlated with the class, yet uncorrelated with each other. At the heart of the algorithm is a subset evaluation heuristics that takes into account the usefulness of individual attributes for predicting the class along with the level of intercorrelation among them.

The first step of our investigation consists on splitting our training set in partitions corresponding to the two professional levels of the ASN, and running the CFS filters on the data of each RF. We then produced a ranking of the selected predictors by counting the occurrences of each of them in the results of the previous computation. Figure 4 reports the top 15 predictors for the two professional levels considered.

We used the best overall learner emerged from the aforementioned tests (i.e. SVM) and applied it, for each academic level and RF, considering the top 15 predictors. The results of our analysis on the 184 RFs are summarized in Figure 5, and the whole data are reported in (Poggi et al., 2018a). We observe that there has been a slight improvement in performances if compared to those obtained using all the predictors: 162/184 (88%) and 163/184 (88.6%) RFs have F-measure scores higher than 0.6 for professional level I and II, respectively. Moreover also in this case the results for bibliometric and non-bibliometric RFs are similar. An analysis of the indicators selected as top 15 predictors is presented in Section 'Discussion'.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27582v1 | CC BY 4.0 Open Access | rec: 12 Mar 2019, publ: 12 Mar 2019



LEGENDA: PREDICTORS CODEBOOK

PREDICTOR	DESCRIPTION	PREDICTOR	DESCRIPTION
academic_age	Academic age (i.e. years since first publication)	journal_with_category	Number of journal papers with a category
acc_ita_2012	Faculty MIUR in 2012	journal_Y12	Journal paper since 2000
bookchapter	Total number of book chapters	journal_Y5	Journal papers in 2008-2012
bookchapter_Y12	Book chapters since 2000	professional_level	Professional level in 2012
born_Y	Year of birth	pub_IF_ALL	Publications with Impact Factor (IF) since ever
IF_ALL	Cumulative IF since ever	pub_IF_Y12	Publications with impact factor since ever
IF_Y12	Cumulative IF since 2000	publication_with_category	Number of publications with category
IF_Y5	Cumulative IF in 2008-2012	university_2012	Italian university in 2012 (if any)
journal	Total number of journal papers chapters	Y_affiliation_same	Maximum number of years with affiliation to the same university
years_no_pub	Number of years with no journal papers published		

journal_Y5
journal_ with_category

Figure 4. Top 15 predictors selected by the CFS filter for professional level I and II. The x-axis shows how many times the predictors have been chosen by the CFS algorithm.

EVALUATION

In order to assess the predictive power of our approach, in this section we compare our best models with those that have been proposed in literature to solve similar problems. As discussed in Section ”Related Work”, three works are particularly relevant for this task: Viera’s model (2014a) based on rank ordered regression, Jensen’s binomial regression model (2009), and the models developed by Tregellas et al. (2018).

A first analysis can be performed comparing the information summarized in table 1 about the sizes of the datasets and the scopes of these works with our investigation. By considering the number of authors and papers, we observe that our dataset is some orders of magnitude greater than the others: i.e. 59149 authors (our work) vs 174 (Vieira), 3659 (Jensen) and 363 (Tregellas) authors; 1910873 papers (our work) vs 7654 papers (Vieira). We also remark that Viera’s and Tregellas’s work are limited to very small samples of researchers from Portugal and the United States, while our and Jensen’s work analyze a nationwide population. Moreover, while the other works focused on a limited set of indicators (Vieira’s model is based on three indicators, Jensen’s on eight and Tregellas’s on ten), we extracted a richer set of indicators from candidates’ CVs (326 predictors). We also observe that, while our work and Jensen’s cover all the disciplines, Vieira limits the analysis to seven disciplines in hard sciences, and Tregellas to biomedical sciences. Overall, our dataset is very wide and rich, and less exposed to issues (e.g. biases) than those used in the other three works.

In order to evaluate the predictive power of our approach, we have to compare its performances with those of the aforementioned works. For this purpose, all the proposed predictive models must be tested on the same data. Since none of the datasets used in the considered works are freely available, we decided to test the models on representative samples extracted from our dataset, and compare the results with our approach.

The first model proposed by Vieira is based on a composite predictor that encompasses 12 standard bibliometric indicators and that is obtained through factor analysis. Unfortunately, the authors don’t provide a definition of such composite predictor, nor they discuss the details on how it has been computed. Given the lack of such information, we observed that is impossible to replicate the model and decided to

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27582v1 | CC BY 4.0 Open Access | rec: 12 Mar 2019, publ: 12 Mar 2019

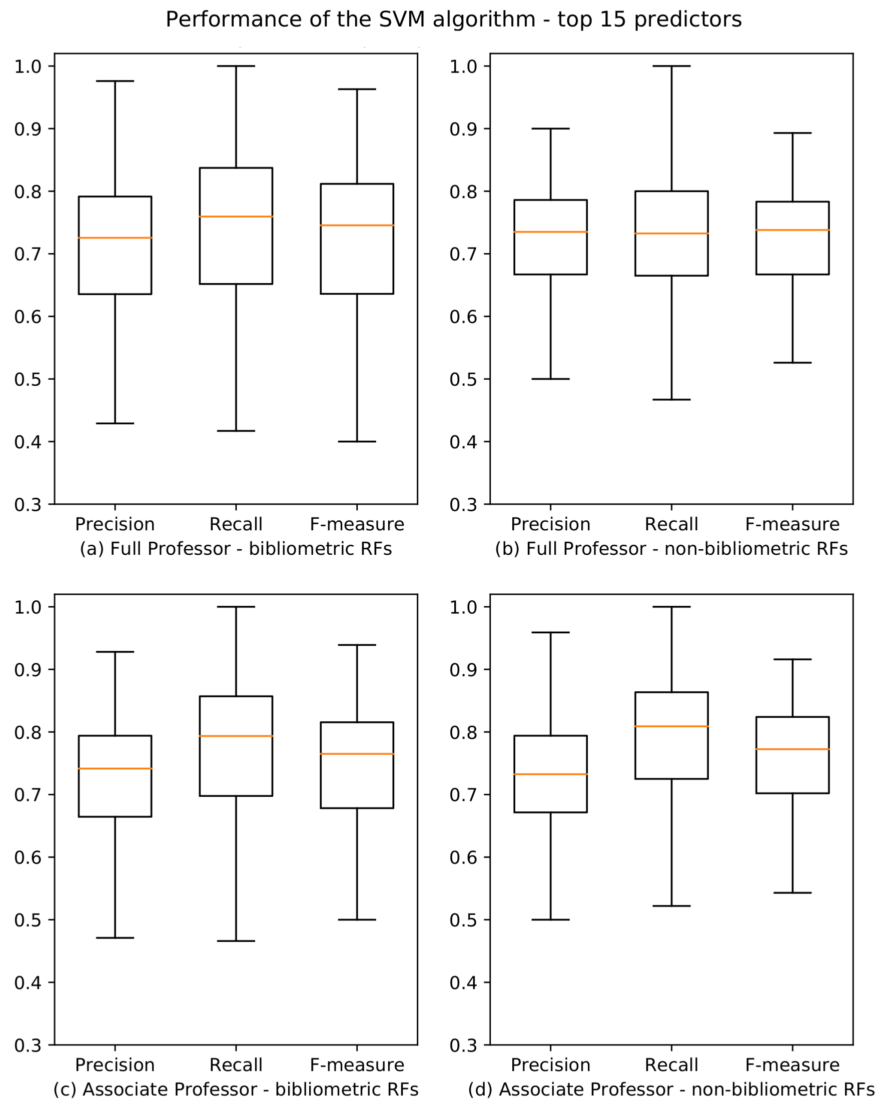


Figure 5. Boxplots depicting the performance of the SVM algorithm for academic level I and II using the top 15 predictors. Precision, Recall and F-measure values are reported for bibliometric (a,c) and non-bibliometric (b,d) RFs.

exclude Vieira's model from this experiment.

Jensen's model is a binomial regression model based on eight indicators: h, h_y, number of publications and citations, mean citations/paper, h/number of papers, age and gender. We decided to focus this analysis on the applicants to the associate professor level for two RFs: Informatics (01/B1) and Economics (13/A1). These two RFs have been chosen as representatives of bibliometric and non-bibliometric recruitments fields because they best meet two important criteria: i) they received a very high number of applications; ii) the two populations (i.e. those who attained the habilitation and those who did not attained it) are well balanced. For the same reason we also considered the SAs "Mathematics and Computer Science" (MCS-01, bibliometric) and "Economics and Statistics" (ECS-13, non-bibliometric). In this way we are able to assess the predictive power of the models at different levels of granularity, both for bibliometric and non-bibliometric RFs and SAs. Since the indicators used by Jensen's models that were not present in our dataset (i.e. mean citations/paper, h/number of papers) could be derived from our data, we computed and added them to the test dataset. We then built the regression models using the aforementioned eight indicators and, as suggested by the authors, we also repeated the experiment using only the h-index, which has been identified as the one with the highest relevance. The results obtained by Jensen's models and our

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27582v1 | CC BY 4.0 Open Access | rec: 12 Mar 2019, publ: 12 Mar 2019

Table 5. Comparison of the performances of our models (OUR-SVM) with Jensen's models using eight predictors (J-LOG8) and one predictor (J-LOG1). Best Precision, Recall and F-measure values are in bold.

Field/	Precision			Recall			F-measure		
Area	J-LOG8	J-LOG1	OUR-SVM	J-LOG8	J-LOG1	OUR-SVM	J-LOG8	J-LOG1	OUR-SVM
01/B1	0.592	0.611	**0.718**	0.588	0.578	**0.773**	0.590	0.594	**0.744**
13/A1	0.611	0.635	**0.724**	0.683	0.579	**0.787**	0.672	0.606	**0.754**
MCS-01	0.677	0.638	**0.692**	0.719	**0.782**	0.753	0.697	0.703	**0.721**
ECS-13	0.676	0.633	**0.685**	0.705	0.658	**0.736**	0.690	0.645	**0.710**

models are reported in Table 5.

The results show that our approach outperforms Jensen's regression models in all the considered RFs and SAs. The only exception is the recall value of the regression model based on the only h-index (LOG1) for the MCS-01 area. However, we report that the relative F-measure, which is a measure of the overall model accuracy, is much lower than our model. This can be explained by considering the low model precision, which is probably caused by an high number of false positives.

By comparing the F-measure values of the models we also observe that the regression models have the worst performances in non-bibliometric fields and areas (i.e. RF 13/A1 and SA ECS-13). The main reason is that the quantitative indicators used by the Jensen's models, which are mostly bibliometric, do not provide enough information for performing accurate predictions for non-bibliometric disciplines. In contrast, our approach is more stable, and leads to similar results in all RFs and SAs. The ability of our model to manage the variability of the different disciplines can be explained by the richness of the dataset on which the model is based.

We also compared the performance of our approach with Tregellas's two best models based on three indicators: sex, date of graduation, and number of first-author papers. As in the previous experiment, we decided to perform the test on two RFs, one bibliometric and one non-bibliometric, following the aforementioned criteria. As representative of bibliometric RFs we chose "Molecular biology" (05/E2) since Tregellas's work focused on the biomedical domain, and "Economics" (13/A1) as representative of non-bibliometric RFs (as in the previous experiment). Two out of the three indicators used by Tregella's models were not present in our dataset: number of first-author papers and date of graduation. While the first indicator can be easily computed using the publication list in the candidates' CVs, the latter (i.e. date of graduation) has to be gathered from external sources. Unfortunately, no freely-available database contains this information. We then had to search the web for authoritative sources (such as professional CVs, personal web pages, etc.) and manually process them to find information about the candidates' education. For this reason, we decided to focus our analysis on a sample of 50 randomly selected candidates for each of the considered RF. The output test dataset has been used for our experiment. The results of our model and Tregellas's models based on linear regression and SVM classifiers are reported in Table 6.

The results show that overall our approach outperforms Tregella's models. Also in this case there is an exception: the recall value of Tregella's model based on SVMs in RF 05/E2. However, by analyzing the relative F-measure, we note that Tregella's overall model accuracy is lower than our model: 0.720 for Tregella's SVM-based model, and 0.738 for our model. This is caused by the high number of false positives produced by Tregella's predictive model, which consequently results in lower precision and F-measure values compared to our model.

By comparing the F-measure values of the models we observe that Tregella's models have very low performances in the non-bibliometric RF (13/A1). We also note that, even considering the specific discipline for which Tregella's models have been designed for (i.e. RF 05/E2 - "Molecular biology", which is a discipline in the the biomedical domain), our model has better performances than two Tregella's regression models. This confirms that our approach is more stable and general, being able to perform accurate predictions in very different RFs and disciplines. As discussed in the previous experiment, the ability of our models to manage the variability and specificity of different disciplines can be explained by the richness of the features in our datasets, which have been automatically extracted from candidates' CVs, and that are fundamental to accurately predict the result of complex human processes (such as evaluation procedures).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27582v1 | CC BY 4.0 Open Access | rec: 12 Mar 2019, publ: 12 Mar 2019

Table 6. Comparison of the performances of our model (OUR-SVM) with Tregellas's two best models based on linear regression (T-LR) and support vector machines (T-SVM). Best Precision, Recall and F-measure values are in bold.

Field	Precision			Recall			F-measure		
	T-LR	T-SVM	OUR-SVM	T-LR	T-SVM	OUR-SVM	T-LR	T-SVM	OUR-SVM
05/E2	0.649	0.628	**0.750**	0.750	**0.844**	0.750	0.696	0.720	**0.750**
13/A1	0.440	0.550	**0.690**	0.393	0.393	**0.645**	0.415	0.458	**0.667**

DISCUSSION AND CONCLUSIONS

This research has been driven by the two research questions described in the introduction, and that can be summarized as follows:

- *RQ1:* Is it possible to predict the results of the ASN using only the information contained in the candidates' CVs?
- *RQ2:* Is it possible to identify a small set of predictors that can be used to predict the ASN results?

The analyses presented in Section 'Results' show that machine learning techniques can successfully resolve the binary classification problem of discerning between candidates that attained the habilitation and those who did not on the base of the huge amount of quantitative data extracted from applicants' CVs with a good accuracy. In fact, the results of the experiments for RQ1 have F-measure values higher 0.6 in 154/184 (83.7%) RFs and in 162/184 (88%) RFs for academic levels I and II, respectively. Moreover, the performances are very similar and uniform for both bibliometric and non-bibliometric disciplines, and do not show a polarization of the results for the two classes of disciplines.

Through an attribute selection process we identified 15 top predictors, and the prediction models based on such predictors resulted to have F-measure values higher than 0.6 in 162/184 (88%) RFs and 163/184 (88.6%) RFs for academic levels I and II, respectively (RQ2). Also in this case, the results are uniform and equally distributed among bibliometric and non-bibliometric disciplines.

Some interesting considerations can be made by analyzing and comparing the top 15 predictors for the two academic levels (i.e. associate and full professor). First of all we remark that, as is obvious, many standard bibliometric indicators have been identified as relevant. In particular, seven of them are shared by both associate and full professor levels: `pub_IF_ALL`, `pub_IF_Y12`, `publication_with_category`, `journal_Y12`, `IF_ALL`, `IF_Y5` and `journal` (see Figure 4). However we note that the first predictor (i.e. the one selected by the feature selection algorithm for most of the RFs) for both levels is `Y_affiliation_same` (i.e. the maximum number of years with affiliation to the same university). This is a non-bibliometric indicator which has not been considered by any of the papers reviewed in the 'Related Work' Section. We plan to investigate whether working for the same institutions correlates positively or negatively with the success to the ASN as future work, and to analyze if there are differences among disciplines.

We also remark that there are interesting observations that concern each of the two levels and highlight peculiar aspects of each of them. For instance, we note that `born_Y` is among the top 15 predictors for associate professors (and not for full professor), suggesting that the age may be a relevant feature for the success at the beginning of an Italian scholar's career. This result is analogous to the one presented in Tregellas et al. (2018), in which a similar indicator (i.e. the date of graduation) is used for predicting career outcomes of young researchers. Conversely, `years_no_pub` (i.e. the number of years in which no papers written by the candidate has been published) is a relevant predictor for full professor (and not for associate professor). An explanation of this fact is that evaluation committees may have considered continuity in publications as a relevant factor in the evaluation of candidates to the full professor level (e.g. for discerning between candidates who have been active throughout their careers, and those who have not always been productive). Also in this case we plan to perform a deeper analysis of this point as future work.

An evaluation of the predictive power of our approach has been performed by comparing the results of our models with the best models that have been proposed in literature to predict academic promotions. The comparison shows that our model outperforms Jensens' binomial regression models and Tregella's models on both bibliometric and non-bibliometric disciplines. This outcome proves that it is possible to predict with a good accuracy the results of complex human processes such peer-review assessments through computational methods. Moreover, the performance difference between the two approaches is more evident for non-bibliometric disciplines. We observe that the outperformances of our results (overall and for non-bibliometric disciplines) are a straight consequence of the richness and quality of the predictors extracted form candidates' CVs. An explanation is that models which are mostly based on bibliometric indicators are not able to fully catch and explain all the different factors (e.g. cultural, social, contextual, scientific, etc.) that play a key role in peer-review evaluation processes.

The results of this work are encouraging and suggest that it is possible to develop automatic systems that support both committees and applicants in complex assessment processes such as the ASN. Future directions of this research line consists in a deeper analysis of the results. In particular, it would be interesting to consider the applicants that have not been correctly classified by the learner in order to improve the approach and also have a more precise understanding of the factors that have been more relevant for assessments of academics performed by humans such as the ASN.

ACKNOWLEDGMENTS

This research has been supported by ANVUR within the Uniform Representation of Curricular Attributes (URCA) project. We would also like to thank Riccardo Fini (University of Bologna) and Andrea Bonaccorsi (University of Pisa), who provided important considerations and insights on this work.

REFERENCES

Abramo, G., D'Angelo, C. A., and Caprasecca, A. (2009). Allocative efficiency in public research funding: Can bibliometrics help? *Research Policy*, 38(1):206–215. DOI:10.1016/j.respol.2008.11.001.

Aha, D. W., Kibler, D., and Albert, M. K. (1991). Instance-based learning algorithms. *Machine learning*, 6(1):37–66. DOI:10.1023/A:1022689900470.

Aksnes, D. (2003). A macro study of self-citation. *Scientometrics*, 56(2):235–246. DOI:10.1023/A:102191922.

Bornmann, L. (2013). How to analyze percentile citation impact data meaningfully in bibliometrics: The statistical analysis of distributions, percentile rank classes, and top-cited papers. *Journal of the Association for Information Science and Technology*, 64(3):587–595. DOI:10.1002/asi.22792.

Bornmann, L. and 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. DOI:10.1007/s11192-006-0121-1.

Bornmann, L. and 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. DOI:10.1016/j.joi.2007.01.002.

Bornmann, L. and Haunschild, R. (2018). Do altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000Prime data. *PloS one*, 13(5):e0197133. DOI:10.1371/journal.pone.0197133.

Bornmann, L., Wallon, G., and 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):e3480. DOI:10.1371/journal.pone.0003480.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32. DOI:10.1023/A:1010933404324.

Cronin, B. and Meho, L. (2006). Using the h-index to rank influential information scientists. *Journal of the Association for Information Science and Technology*, 57(9):1275–1278. DOI:10.1002/asi.20354.

Danell, R. (2011). Can the quality of scientific work be predicted using information on the author's track record? *Journal of the Association for Information Science and Technology*, 62(1):50–60. DOI:10.1002/asi.21454.

Franceschet, M. (2009). A cluster analysis of scholar and journal bibliometric indicators. *Journal of the Association for Information Science and Technology*, 60(10):1950–1964. DOI:10.1002/asi.21152.

Franceschet, M. and Costantini, A. (2011). The first Italian research assessment exercise: A bibliometric perspective. *Journal of informetrics*, 5(2):275–291. DOI:10.1016/j.joi.2010.12.002.

Fu, L. D. and Aliferis, C. F. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics*, 85(1):257–270. DOI:10.1007/s11192-010-0160-5.

Hall, M. A. and Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering*, 15(6):1437–1447. DOI:10.1109/TKDE.2003.1245283.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284. DOI:10.1109/TKDE.2008.239.

Ibáñez, A., Armañanzas, R., Bielza, C., and Larrañaga, P. (2016). Genetic algorithms and Gaussian Bayesian networks to uncover the predictive core set of bibliometric indices. *Journal of the Association for Information Science and Technology*, 67(7):1703–1721. DOI:10.1002/asi.23467.

Jensen, P., Rouquier, J.-B., and Croissant, Y. (2009). Testing bibliometric indicators by their prediction of scientists promotions. *Scientometrics*, 78(3):467–479. DOI:10.1007/s11192-007-2014-3.

John, G. H. and Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. In *Proc. 11th Conference on Uncertainty in Artificial Intelligence*, pages 338–345. Morgan Kaufmann.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., and Murthy, K. R. K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*, 13(3):637–649. doi:10.1162/089976601300014493.

Law dec. 30, n. 240 (2011). Rules concerning the organization of the universities, academic employees and recruitment procedures, empowering the government to foster the quality and efficiency of the university system (Norme in materia di organizzazione delle università, di personale accademico e reclutamento, nonche' delega al Governo per incentivare la qualità e l'efficienza del sistema universitario), Gazzetta Ufficiale n. 10 del 14 gennaio 2011 - Suppl. Ordinario n. 11. Available at http://www.gazzettaufficiale.it/eli/id/2011/01/14/011G0009/sg. (Accessed 9 October 2018).

Leydesdorff, L. (2009). How are new citation-based journal indicators adding to the bibliometric toolbox? *Journal of the Association for Information Science and Technology*, 60(7):1327–1336. DOI:10.1002/asi.21024.

Lindahl, J. (2018). Predicting research excellence at the individual level: The importance of publication rate, top journal publications, and top 10% publications in the case of early career mathematicians. *Journal of Informetrics*, 12(2):518–533. DOI:10.1016/j.joi.2018.04.002.

Marzolla, M. (2015). Quantitative analysis of the Italian national scientific qualification. *Journal of Informetrics*, 9(2):285–316. DOI:10.1016/j.joi.2015.02.006.

Ministerial Decree 159 (2012). Redefinition of scientific disciplines (Rideterminazione dei settori concorsuali), Gazzetta Ufficiale Serie Generale n. 137 del 14-06-2012 - Suppl. Ordinario n. 119). Available at www.gazzettaufficiale.it/eli/id/2012/06/14/12A06786/sg. Accessed 9 October 2018.

Nederhof, A. J. and Van Raan, A. F. (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. DOI:10.1007/BF02279353.

Norris, M. and Oppenheim, C. (2003). Citation counts and the Research Assessment Exercise V: Archaeology and the 2001 RAE. *Journal of Documentation*, 59(6):709–730. DOI:10.1108/00220410310698734.

Nuzzolese, A. G., Ciancarini, P., Gangemi, A., Peroni, S., Poggi, F., and Presutti, V. (2018). Do altmetrics work for assessing research quality? *Scientometrics*, pages 1–24. DOI:10.1007/s11192-018-2988-z.

Poggi, F., Ciancarini, P., Gangemi, A., Nuzzolese, A. G., Peroni, S., and Presutti, V. (2018a). Predicting the Results of Evaluation Procedures of Academics: Additional Materials. Available at https://doi.org/10.6084/m9.figshare.6814550. DOI:10.6084/m9.figshare.6814550.

Poggi, F., Ciancarini, P., Gangemi, A., Nuzzolese, A. G., Peroni, S., and Presutti, V. (2018b). Predicting the Results of Evaluation Procedures of Academics: Appendices. Available at https://doi.org/10.6084/m9.figshare.6814502. DOI:10.6084/m9.figshare.6814502.

Poggi, F., Cigna, G., and Nuzzolese, A. G. (2016). Enhancing Open Data to Linked Open Data with ODMiner. In *LD4IE@ISWC*, pages 44–50. Available at http://ceur-ws.org/Vol-1699/paper-06.pdf.

Quinlan, J. R. (2014). *C4.5: programs for machine learning*. Elsevier.

Taylor, J. (2011). The assessment of research quality in UK universities: peer review or metrics? *British Journal of Management*, 22(2):202–217. DOI:10.1111/j.1467-8551.2010.00722.x.

Tregellas, J. R., Smucny, J., Rojas, D. C., and Legget, K. T. (2018). Predicting academic career outcomes by predoctoral publication record. *PeerJ*, 6:e5707. DOI:10.7717/peerj.5707.

Van Raan, A. F. (2006). Comparison of the hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, 67(3):491–502. DOI:10.1556/Scient.67.2006.3.10.

Vieira, E. S., Cabral, J. A., and Gomes, J. A. (2014a). Definition of a model based on bibliometric indicators for assessing applicants to academic positions. *Journal of the Association for Information Science and Technology*, 65(3):560–577. DOI:10.1002/asi.22981.

Vieira, E. S., Cabral, J. A., and Gomes, J. A. (2014b). How good is a model based on bibliometric indicators in predicting the final decisions made by peers? *Journal of Informetrics*, 8(2):390–405. DOI:10.1016/j.joi.2014.01.012.

W3C OWL Working Group (2012). OWL 2 Web Ontology Language. Available at https://www.w3.org/TR/owl2-overview/.

Wouters, P., Thelwall, M., Kousha, K., Waltman, L., de Rijcke, S., Rushforth, A., and Franssen, T. (2015). The metric tide: Correlation analysis of REF2014 scores and metrics (Supplementary Report II to the Independent Review of the Role of Metrics in Research Assessment and Management). *London: Higher Education Funding Council for England (HEFCE)*. DOI:10.13140/RG.2.1.3362.4162.