Can we automate expert-based journal rankings? Analysis of the Finnish publication indicator

Mirka Saarela[a,*], Tommi Kärkkäinen[a]

[a] *University of Jyvaskyla, Faculty of Information Technology, P.O. Box 35, FI-40014 University of Jyvaskyla, Finland*

Abstract

The publication indicator of the Finnish research funding system is based on a manual ranking of scholarly publication channels. These ranks, which represent the evaluated quality of the channels, are continuously kept up to date and thoroughly reevaluated every four years by groups of nominated scholars belonging to different disciplinary panels. This expert-based decision-making process is informed by available citation-based metrics and other relevant metadata characterizing the publication channels. The purpose of this paper is to introduce various approaches that can explain the basis and evolution of the quality of publication channels, i.e., ranks. This is important for the academic community, whose research work is being governed using the system. Data-based models that, with sufficient accuracy, explain the level of or changes in ranks provide assistance to the panels in their multi-objective decision making, thus suggesting and supporting the need to use more cost-effective, automated ranking mechanisms. The analysis relies on novel advances in machine learning systems for classification and predictive analysis, with special emphasis on local and global feature importance techniques.

Keywords: Performance-based research funding system, Machine learning, Automation, Feature importance

1. Introduction

Various countries have created their own performance-based research funding system (PRFS) for motivating and rewarding national faculty and research institutes to actively add to the knowledge base in their respective disciplines. As summarized by Zacharewicz et al. (2019, Figure 1), a PRFS should be based on assessment, which can result from the individual application or a combination of peer review, metrics-informed peer review, or a formula based on quantitative metrics. Hence, all assessment procedures are based on sufficient data to gather evidence for evaluation and decision making. When focusing on the scholarly output, it is the quality and quantity of publications that matter (Sandström & van den Besselaar, 2018).

Creating and analyzing a system that attempts to translate scholarly activity fairly into monetary incentives is not straightforward. One might argue that if a funding allocation formula based on quantitative metrics is too simple, it leads to adverse effects. For instance, Butler (2003) showed that the allocation of resources in Australia based on the number of publications resulted in more publications but mainly in lower-level publication channels. However, this finding and the corresponding analysis were later invalidated by van den Besselaar et al. (2017), who found a positive correlation between the national publication quantity and quality. Although not conclusive, the academic peer review tradition is based on the presumption that the research output should be assessed based on quality and not solely on quantity. Moreover, quality can be quantified (i.e., systematically measured and evaluated) in diverse ways (e.g., Kulczycki et al., 2017; Rodríguez-Navarro & Brito, 2019; Subochev et al., 2018; Wallace & Perri, 2018; Walters & Markgren, 2019).

The quality of an individual publication channel can be determined by citation-based indicators or by experts. Examples of citation-based indicators are metrics based on the impact factor, whereas an expert ranking is usually

*Corresponding author.

Email address: `mirka.saarela@jyu.fi` (Mirka Saarela)

established through peer review of a subset of the publication output. The main problem with citation-based indicators is their inequality between disciplines. For example, research in the social sciences and humanities—which is usually more locally oriented (Hicks et al., 2015) and often published in national languages—cannot compete in popularity with, for example, research in the natural sciences, which is usually published in the English language only (Kulczycki et al., 2018; Sivertsen, 2016a, 2019; Verleysen & Engels, 2014). Another well-known example is that mathematics papers accumulate citations at a much lower rate than, for example, biology papers (Bensman et al., 2010; Vieira & Gomes, 2010). The main problems with expert-based ranking, on the other hand, are the associated costs and the fact that they are not objective. The first problem arises because expert-based evaluations are expensive and not straightforward (Thelwall, 2017). The latter issue might be caused by the tendency of reviewers to deliberately or unintentionally evaluate according to their own (research) interests, a biased selection of reviewers, or other conflicts of interest (Dondio et al., 2019; Haddawy et al., 2016; Saarela et al., 2016; Serenko & Bontis, 2018; Walters, 2017; Zacharewicz et al., 2019). Especially if a publication is published in a relatively unusual national language (such as Finnish, where experts who actually speak this language might compose only a very small group), it can be challenging to find knowledgeable reviewers who do not have any conflict of interest (Letto-Vanamo, 2019). Thus, all evaluations of publication channel quality come with certain costs.

Recently, the idea of automatically classifying publication channels by using machine learning models has become more pronounced (Halim & Khan, 2019; Saarela et al., 2016; Tüselmann et al., 2015). These models can work independent of human intervention, so they promise to be more objective. They are also cheaper to implement than expert-based classification procedures. In addition, they can (in comparison to citation-based indicators) take all available quality indicators into account, which means the utilization of all possible explanatory features. This seems desirable because publication channel ranking is essentially a multicriteria decision problem (Subochev et al., 2018).

In bibliometric studies, machine learning has been used, for example, to predict citation counts (Abrishami & Aliakbary, 2019), knowledge flow (Hassan et al., 2018), and rising scholars (Daud et al., 2015). In particular, the tree-based classifiers have been increasingly utilized (e.g., Bai et al., 2019; Kim & Kim, 2018; Heinisch & Buenstorf, 2018; Treeratpituk & Giles, 2009; Tüselmann et al., 2015). Heinisch & Buenstorf (2018), for example, used different machine learning methods (regularized logistic regression, support vector machine, random forests, and adaptive boosting) to predict dissertation supervisors from the characteristics of co-authors. Random forests outperformed the other algorithms in terms of precision and recall. Kim & Kim (2018) and Treeratpituk & Giles (2009) reported the success of random forest classifiers in author name disambiguation research. Bai et al. (2019) applied gradient boosting to predict the citations of research papers. Compared with currently popular deep learning methods, which, with many kinds of processing layers, are black-box models, tree-based approaches allow the determination of feature importances to explain the predictions. The importance of explainable models in bibliometrics has been emphasized, for example, by Hicks et al. (2015).

With the goal to automate publication channel quality assessment, we analyze in this paper the publication indicator of the Finnish PRFS by using explainable machine learning techniques. This means that we are examining the PRFS here only as a system of publication channel rankings. The publication indicator, the rank of a publication channel, is expert based, and the expert panels are renewed every four years. The current panels were nominated in the beginning of 2018. The main activity of the panels, in addition to continuously ranking new publication channels, is the re-evaluation of all publication channel ranks of the Finnish PRFS. These new classifications have been in effect since the beginning of 2019. Our research questions, linked to the diverse analysis of all existing ranks and the new expert-based evaluations, are as follows:

- What features characterize publication channels of different ranks?
- Which publication channel rankings have changed during the last reevaluation, and what features characterize the publication channels with changed ratings? In other words, can we predict and explain the latest rank changes?
- Can we automate and explain the rankings?

2. The Finnish PRFS

The Finnish database underlying the national PRFS, *JuFo* (Publication Forum in English, "JulkaisuFoorumi" or "JuFo" in Finnish), was introduced in 2010, in conjunction with a renewed university legislation. The national PRFS is based on the Norwegian model (i.e., the PRFS developed for Norway in 2004) and the Danish PRFS that adopted the Norwegian model in 2009 (Aagaard, 2019; Pölönen, 2018).[1] The goal of the Norwegian model is to treat all scholarly disciplines equally and accurately. It is composed of three components: (i) an exhaustive national database of publications, (ii) a publication indicator that weights/ranks publications based on the publication channel they were published in, and (iii) a funding model that allocates resources to national research institutions based on their shares in the total publication points resulting from the ranks (Sivertsen, 2018, 2016b).

The publication indicator, i.e., the second component of the Norwegian model in Finland, is called a *JuFo rank*, or simply rank. The rank of a publication channel is determined by a group of nominated scholars belonging to a disciplinary panel, who negotiate and decide the ranks of all identified publication channels in their field. It can be 1 (basic), 2 (leading), or 3 (top). Other identified publication channels that do not qualify for the basic rank 1 are ranked 0. Essentially, this means that such a channel does not apply reliable operating principles in the peer review process.[2]

The aggregated rankings of the publication indicator currently constitute 13% of the public funding of Finnish universities[3] and are computed using the weights shown in Table 1. This means that as a part of the expert evaluation of the ranks done by the panel members, the members try to ensure that all relevant subfields of a discipline (especially those that each member represents) have highly ranked publication channels. However, at most, 20% of publications can have a rank higher than 1 and only 5% can be assigned to the top rank 3. These percentages are calculated based on the total publication volume of the publication channels allocated to a panel. Hence, the number of papers published determines the number of high-ranking publication channels there can be (i.e., if a publication channel has a large publication volume, it takes up the higher level's quota).

Publication type	**rank 3**	**rank 2**	**rank 1**	**rank 0**
Monograph	16	12	4	0.4
Journal article, book article, conference proceedings article, or edited work	4	3	1	0.1

Table 1: Weights of peer-reviewed publications in the *JuFo* funding model for the time frame 2017–2020 (Letto-Vanamo, 2019; Pölönen, 2018). A monograph refers to a peer-reviewed scientific book that was written entirely by its author(s).

The decision-making process of the ranks is led by the chairperson of the panel and supported by two coordinators (Pölönen, 2018). As a part of this assessment, different research fields and traditions related to the panel's discipline are attempted to be balanced in the highest ranks. The contentious cases are solved by the chairperson and through panel discussions. These discussions indeed reflect personal opinions on the quality of a channel.[4] Thus, the final decision on a publication channel's rank can be subjective.

Besides the volume of the publication channel and equality between research fields, the panels must also take other criteria into account. For example, when a new publication channel is evaluated, because of a scholarly suggestion or a published article by a national scholar in the previous year, it can obtain rank 1 through the manual checking of the peer review process, members of the editorial board, and the actual contents of the published papers. For ranks 2 and 3, the assessment takes into account the limits of the publication volume on the rank level and all available quantitative information on the channels, especially Danish and Norwegian rankings ranked on a scale from 0 to 2 and the citation indicators *SCImago journal rank* (SJR), *source normalized impact* (SNIP), and *impact per publication* (IPP), which are all stored in the *JuFo* database. Moreover, certain panels might have additional evaluation criteria. For example, a special form of publication channel with rank 2 in panel 2 on *computer and information sciences* is the most distinguished conferences (Li et al., 2018; Meho, 2019; Vrettas & Sanderson, 2015). Ranking activity, therefore, is characterized as a multicriteria decision-making process done by experts (Subochev et al., 2018). According to

[1] Meanwhile, the Norwegian model has been adopted also by Poland and Flanders in Belgium, as well as by some individual research institutes in Sweden and Ireland (Sivertsen, 2019).

[2] http://www.julkaisufoorumi.fi/en/evaluations/classification-criteria

[3] https://minedu.fi/documents/1410845/4392480/Universities_funding_2017.pdf/

[4] The second author of this paper has been a member of the panel on *computer and information sciences* since January 2018.

Zacharewicz et al. (2019) (see Section 1), the quality of a publication channel is assessed using a metrics-informed peer review, resulting in a formula based on quantitative metrics for the actual resource allocation.

Table 2: Panels of the different disciplines in Finland. Each panels is responsible for the evaluation and classification of those publication channels in their discipline.

panel	discipline
Panel 1	Mathematics and statistics
Panel 2	Computer and information sciences
Panel 3	Physical sciences, space science and astronomy
Panel 4	Chemical sciences
Panel 5	Geosciences and environmental sciences
Panel 6	Biosciences I
Panel 7	Biosciences II
Panel 8	Civil and mechanical engineering
Panel 9	Electrical and Electronic engineering, information engineering
Panel 10	Chemical engineering, materials engineering and environmental engineering
Panel 11	Medical engineering, biotechnology and basic medicine
Panel 12	Clinical medicine I
Panel 13	Clinical medicine II and dentistry
Panel 14	Health sciences and other medical sciences
Panel 15	Agricultural sciences
Panel 16	Economics and business
Panel 17	Social sciences, media and communications, interdisciplinary social sciences
Panel 18	Psychology and educational sciences
Panel 19	Political science, public administration and law
Panel 20	Philosophy and theology
Panel 21	Languages
Panel 22	Literature, arts and architecture
Panel 23	History, archaeology and cultural studies
Panel 24	Multidisciplinary

Another interesting (but ambivalent) indicator that is available (according to the Norwegian model) in another national database[5] and might be considered indirectly for the national rank is the number of publications by national scholars in the evaluated channels. In fact, previous research indicates that PRFSs can create changes in publication practices (Aagaard, 2019; Aagaard et al., 2015; Ahlgren & Waltman, 2014; Butler, 2003; Haddawy et al., 2016; Saarela et al., 2016; Serenko & Bontis, 2018). This means that if a PRFS advocates certain publication channels it is likely that national scholars will adapt their publishing behavior accordingly (Sīle & Vanderstraeten, 2019). Moreover, a study by Sandström & van den Besselaar (2016) showed a strong correlation between the number of papers (i.e., productivity of individual researchers) and number of citations (i.e., impact). Another reason why the number of publications could serve as an indicator is that the experts who decide the ranks (and who are usually active scholars themselves) might be biased toward their own discipline and research. Serenko & Bontis (2018), for example, found that experts had a higher opinion about publication channels in which they have published or those that they are considering as a possible channel for publication. Thus, the number of publications by national scholars (although not an official ranking criterion) seems to be an interesting indicator to be studied when analyzing ranks in a PRFS.

3. Data retrieval and preprocessing

The datasets for this study were collected in the beginning of 2019 mainly from two publicly available collections; the data of the Finnish publication channel evaluation system (*JuFo*[6]) and the data consisting of all publications by Finnish researchers (*Juuli*[7]). These datasets can be easily joined by utilizing the unique *Juuli ID* that each publication channel in the Finnish database has. In the latest reevaluation, the volume of a publication channel also had an impact on the rankings (as described in Section 2), so we added this information to our final dataset (retrieving it from the *JuFo Portal*[8] by leveraging the *JuFo ID* to match the different datasets).

[5]The *Juuli* database described in Section 3.
[6]See `https://www.tsv.fi/julkaisufoorumi/haku.php?lang=en`
[7]See `http://www.juuli.fi/`
[8]See `http://www.julkaisufoorumi.fi/en`

The data collection and analysis of the *JuFo* and *Juuli* data are follow-ups of the studies of these datasets before the latest reevaluation (Saarela et al., 2016; Akusok et al., 2019). The reasons for collecting and analyzing the new data were mainly twofold: the panel members changed during this time, and all ranks were completely reevaluated by the new panels. This means that we are analyzing here the dynamics and the results of the last thorough reevaluation by the new panels that have been in operation since the beginning of 2019.

The *JuFo* data contain 31,172 publication channels. For the sake of completeness, we removed all channels from our list that do not have a rank, a panel, an ID, a name, or a history of previous evaluation in the system (this information is needed for the second part of the analysis). This left us with 26,346 publication channels. Moreover, ambiguous or duplicate data values and different letter cases were cleaned. For example, for the *language* column, the values "FINNISH" and "Finnish" were both mapped to "Finnish" and the values "Multiple languages," "Miscellaneous languages," and "Undetermined" were all mapped to "Other language."

Table 3: List of variables, meaning, and transformation into features

Feature	**Meaning**	**Numerical representation**
Rank	Current (2019) *JuFo* ranking	Integer values in range {0, 1, 2, 3}
Title	Title of the publication channel	String values, only used for explaining performed analysis
Publications	Number of publications of Finnish scholars in publication channel ⟨title⟩ found in *Juuli*	Integer value of the sum of all articles by Finnish scholars, zero if no publications (`NrOfPublications`)
Volume	Volume how many articles the publication channel publishes in a year as stored in the *JuFo Portal*	Integer value of the sum of all articles the publication channel publishes in a year (`VOL`) plus a binary variable representing missing values (`VOL_NaN`)
Type	Type of the publication channel type (i.e., whether it is a journal or conference)	Represented in one-hot encoding with 2 binary variables; this feature has no missing values
StartYear	Year when the publication channel ⟨title⟩ was founded	Start year of the channel (`startYear`), plus a binary variable representing missing values (`startYear_NaN`)
Language	Language of the publication channel only separating English, Finnish, Swedish and other	One-hot encoding of the four categories {English, Finnish, Swedish, other Language}
Norway	Publication channel ranking in the Norwegian analog of *JuFo*	The Norwegian rank (0, 1, or 2) or `Norway_NaN` if not evaluated in Norway
Denmark	Publication channel ranking in the Danish analog of *JuFo*	The Danish rank (0, 1, or 2) or `Denmark_NaN` if not evaluated in Denmark
SJR	The *SCImago Journal Rank* indicator (González-Pereira et al., 2010)	Real-valued variable for the impact indicator plus a binary variable representing missing values
IPP	The *Impact per Publication* indicator (see `https://www.journalindicators.com/methodology`)	Real-valued variable for the impact indicator plus a binary variable representing missing values
SNIP	The *Source Normalized Impact* indicator (Moed, 2010)	Real-valued variable for the impact indicator plus a binary variable representing missing values
Panel	The scientific panel that is responsible for assigning a rank to the publication channel; all panels numbers and their disciplines are listed in Table 2	One-hot encoding of the panel number with 24 binary variables, one for each panel (`panel_Nr1`, . . ., `panel_Nr24`); this feature has no missing values
Sherpa/ Romeo Code	Colour codes[9] to depict policies regarding the self-archiving of publication channel articles on the web and in open access repositories	One-hot encoding of the four different sherpa/romeo color codes (green, blue, yellow, white) plus one for not available (`SherpaRomeoCode_NaN`)
Publisher	Number of publication channels with the same publisher in *JuFo*	Frequency of the publisher in *JuFo* (`freqpublisher`) or `freqpublisher_NaN` if the publisher was not available in *JuFo*

All important variables and our transformations and preprocessings to obtain the individual features of the machine learning models are listed in Table 3. Hence, our analysis is based on the three citation-based indicators (i.e., SJR, IPP, and SNIP), the expert-based rank in the related ranking systems (i.e., the Norwegian and Danish rank), the start year of the publication channel, the language the publication channel publishes in, the type of the publication channel (i.e., whether it is a journal or conference), the evaluating panel (which essentially indicates the discipline of the publication channel), the volume (i.e., the number of articles this publication channel publishes in a year), and the number of publications of Finnish researchers within the publication channel. Another feature that can be associated with the quality of a publication channel is the sherpa/romeo-code, i.e., the policies regarding open-access publishing (van Vlokhoven, 2019). Thus, we also transformed this information into features (see Table 3).

[9]The color codes are explained, for example, at `http://www.sherpa.ac.uk/romeoinfo.html`.

As the publisher behind a publication channel is often an indicator of its quality, we also wanted to harness this information. However, even after cleaning (letters were lower-cased, and all whitespaces and symbols were removed), the publisher variable in the dataset still had 11,799 unique values. Thus, we did not use all of these unique publishers as features (which would have resulted in 11,799 separate features with one hot encoding). Instead, we just leveraged the frequency of the publisher, i.e., the number of publication channels with the same publisher in the *JuFo* database, as a single feature (see Table 3). As can be seen from Fig. 1, the majority of publishers have only a few publication channels in *JuFo*.

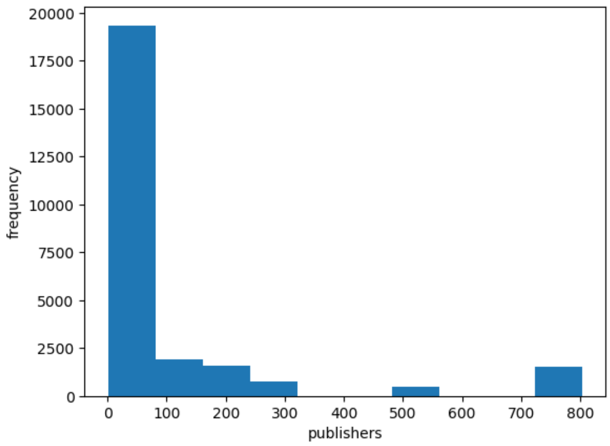

Fig. 1. Histogram visualizing the frequencies of publishers in the publication channel data. Only three publishers have more than 400 publication channels (Springer with 803, Elsevier with 733, and Taylor & Francis with 494). A total of 9,151 publishers in the data have only one publication channel, and more than two thirds of the publishers (i.e., 18,306) have less than 50 publication channels.

The dataset has a considerable amount of missing data that are not missing at random (Little & Rubin, 2014). According to Tüselmann et al. (2015), missing data problems are often not addressed sufficiently in publication channel ranking studies. Moreover, the usual strategies to deal with missing data are to either discard or impute them. However, both of these strategies are unsuitable for our data. Discarding data rows or features with missing values would result in a huge data loss. Imputation should only be performed if the data are missing at random (Little & Rubin, 2014). Another reason why both deletion and imputation are inappropriate here is that the missingness of a value actually contributes valuable information. For example, if the SJR of a publication channel is missing, it is often an indicator that this channel might be of lower quality (Saarela et al., 2016). Thus, we encoded for each feature with missing values a new binary feature containing information on whether this value was available for the individual publication channel.

All data retrieval, preprocessing, and analysis were performed with `Python 3.7.1` using `scikit-learn`, `pandas`, `imbalanced-learn`, `seaborn`, and `LIME` libraries.

4. Methods

To answer our research questions, we diversely analyzed the data by leveraging state-of-the-art machine learning methods. This analysis was divided into several parts: (i) analysis (and prediction) of overrated publication channels, (ii) prediction of rank changes, and (iii) description (and prediction) of publication channel ranks. The prediction of changed or overrated ranks are binary classification tasks, whereas the prediction of publication channel ranks is a multiclass classification problem. In all parts of the analysis, we especially focused on feature importance or relevance (Guyon & Elisseeff, 2003; Molnar, 2019), i.e., the characteristics that discriminate the publication channels with (i) overrated, (ii) changed, or (iii) different ranks. These feature importances were computed for both the models and the individual predictions. Since all our classification tasks have to deal with imbalanced data, we first describe our strategies to address this issue (Section 4.1) and then, the used classifiers and measured feature importances (Section 4.2 and Section 4.3).

4.1. *Class imbalance*

All classification tasks in our analysis need to deal with imbalanced data. Rank prediction is a multiclass imbalanced data classification problem. 2,392 (9.1%) publication channels in *JuFo* are rank 0, 20,574 (78.1%) are rank 1, 2,587 (9.8%) are rank 2, and 790 (3.0%) are rank 3. This means that the default predictor that would always predict a channel as rank 1 would readily achieve an accuracy of 78%. The imbalance issue is even more pronounced in both of the binary classification tasks addressed. Only 4,817 (18.3%) of the 26,343 publication channels are overrated, and only 2,069 (7.9%) have a changed rank. This means that the default predictor predicting always that a channel is not overrated would achieve an accuracy of 82%, and predicting that its rank did not change would achieve an accuracy of 92%. As a whole, the most interesting class is often severely underrepresented in the analysis.

An imbalance in class frequencies is a common problem in many machine learning studies (see, e.g., Kim & Kim, 2018; Saarela et al., 2019), and different techniques and methods have been introduced to address it (Vanhoeyveld & Martens, 2018). These procedures are mostly based on either cost-sensitive or sampling methods. In cost-sensitive methods, the classifier is penalized more when classifying an observation from the minority class as the majority class, which prevents systematic classification to the majority class. Sampling methods reshape the training data by sampling a smaller set of the observations from the majority class in the training set (undersampling) or by repeating observations from the minority training set (oversampling). In this paper, we use the synthetic minority oversampling (`SMOTE`) introduced by Chawla et al. (2002). `SMOTE` was found to be the best sampling technique in previous comparisons (Bach et al., 2017; Saarela et al., 2019). Instead of randomly duplicating observations from the minority class, it creates clusters around each observation from the minority class. Hence, new observations that are synthetic interpolations of the minority class are created.

4.2. *Classifiers*

Our goal was to employ classifiers that are well-performing without sacrificing the explainability of the models (Hicks et al., 2015). To accomplish this, we used a linear (logistic regression) and two nonlinear tree-based (random forest and gradient boosting) classifiers for the classification tasks. In comparison to black-box models (such as currently famous deep learning methods), traditional linear methods as well as tree-based approaches are often preferred choices in sensitive domains because they are explainable. Classification performance, on the other hand, depends normally on the dataset at hand, and no single classifier performs best or worst for all datasets. Nevertheless, tree-based methods are frequently the winner in general comparisons of classification algorithms (e.g., Olson et al., 2018) and have been reported to outperform even extreme learning machines (Boemer et al., 2018), which were used in the previous analysis of the old *JuFo* dataset instance (Akusok et al., 2019). Another reason why we used tree-based methods in this study is that they are explainable despite their nonlinearity (see, e.g., Saarela et al., 2019).

Our first benchmark prediction method was logistic regression. Logistic regression is probably the most traditional and well-established technique to predict a categorical label and similar to other linear methods, known for its straightforward interpretability. In bibliometrics, logistic regression has been applied, for example, to predict and characterize highly cited method papers (Small, 2018), to test whether papers with continuously rising citations (so-called evergreens) can be identified (Zhang et al., 2017), and to examine how the characteristics of a paper's references (most importantly, their country assignments) affect the citedness of the paper (Bornmann et al., 2018). Logistic regression measures the relationship between the categorical label (i.e., the class variable) and the explanatory features by estimating probabilities using a logistic function. More precisely, given M training observations $(\boldsymbol{x_i}, y_i), i = 1, \ldots, M$, where $\boldsymbol{x}_i$ is an n-dimensional observation (i.e., the vector containing the n feature values) and y_i is its class label, logistic regression estimates the probability of y given $\boldsymbol{x}$ by using the sigmoidal transformation $\frac{1}{1+\exp(-\theta^T \boldsymbol{x})}$, where θ are the linear coefficients.

We trained the `LogisticRegression` models by running a fivefold `GridSearchCV` over the penalty to determine whether to use lasso or ridge (i.e., l_1 or l_2 regularized) logistic regression, the strength of regularization `C` (using the values 0.0001, 0.001, 0.01, 1, 10, and 100), and the `class_weight` (`None` or `balanced`). One reason why logistic regression remains to be a popular classification method for many applications is that it can be easily explained and interpreted. The magnitude of the absolute values of feature coefficients θ can be interpreted as the importance of that feature, with a larger value indicating that the feature has more relevance in the classification. Moreover, the sign of the coefficient indicates whether the feature increases or decreases the probability of belonging to a certain class. If the logistic regression model is penalized with the l_1 norm, some of the feature coefficients shrink to exactly zero, which makes the model simpler and more interpretable (Tibshirani, 1996).

Random forests are ensemble learners based on decision trees. They overcome the disadvantages of decision trees, while at the same time keeping their main advantages (Breiman, 2001). The main advantages of decision trees are their low bias. More precisely, if they are grown fully (i.e., without any pruning), they can often find rules to separate the classes of training data perfectly. Their main disadvantage is that they suffer from high variance. This means that they may not generalize well, and a slight change in the training data can change the structure of the tree completely. Random forests aggregate a large number of uncorrelated decision trees, each of which has low bias and high variance. To reduce the variance in a forest, each tree is grown on a bootstrap sample from the original data, and for each node in each tree randomly, only a subset of the features (controlled by the `max_features` parameter) is used; this random variable selection is repeated at each node. Thus, random forests achieve both low bias and low variance. Their low bias is acquired because each tree in the forest separates the classes in the training almost perfectly. The low variance results from bagging many random (uncorrelated) trees.

Another perk of random forests is that they do not need much parameter tuning. We grid-searched the maximum number of features to sample at each node (`max_features`) and the minimum number of samples required to be at a leaf node (`min_samples_leaf`) by using fivefold `GridSearchCV`. Moreover, as all our classification problems were defined with imbalanced class frequencies, we also grid-searched the best `class_weight` (`None` or `balanced`). Finally, we set our `n_estimators` to a high value (`1000`) because more trees in a forest are always better; they reduce overfitting and improve the reliability of the results. The random forest implementation in Python provides feature importance measures by calculating the Gini importance (Breiman, 2017). After each node split, the total decrease of the Gini index (node impurity) is calculated and then averaged over all trees. The more impurity decreases, the more important the input feature is.

Gradient boosting (Friedman, 2001) is like random forest; it is an ensemble method utilizing weak learners (as most commonly done, we use here decision trees as the weak learners). However, while the trees in a random forest are built in parallel, the idea of boosting is to train the decision trees sequentially, each trying to correct its predecessor. This correction is accomplished by fitting as many trees as there are mutually exclusive classes on the negative gradient of the multinomial (binomial in the binary case) deviance loss function. This means that gradient boosting can be illuminated as an optimization algorithm on this deviance loss function (Breiman, 1997).

Gradient boosting classifiers are generally more sensitive to parameter settings than random forests are. We ran a fivefold cross-validation grid search over the shrinkage factor of each tree's contribution (`learning_rate`), the maximum depth of the individual trees (`max_depth`), the minimum number of samples required to be at a leaf node (`min_samples_leaf`), and the number of features to consider when looking for the best split (`max_features`). While for the random forest, a higher number of estimators is always better, for the gradient boosting, there is a trade-off between the `learning_rate` and the `n_estimators`. Thus, we also grid-searched for the best number of estimators. Similar to the random forest, the gradient boosting implementation in Python computes feature importances by utilizing the Gini index (node impurity) for a single decision tree by the amount that each feature split point improves the performance measure. To obtain the modular feature importance for the entire model, these importances are then weighted by the number of observations the node is accountable for, and they are subsequently averaged across all of the decision trees within the model.

For all three classifiers (logistic regression, random forest, and gradient boosting), we divided our data into the training (75%) and validation (25%) sets. For the logistic regression classifier,[10] we additionally normalized the training data by using standard scaling. Then, we ran a stratified fivefold cross-validation grid search over the (normalized) training data to determine the best parameters for the models. Finally, we fitted our models by using the best parameters from the grid search, and applied the prediction on the (normalized[11]) validation data that were untouched the entire time during model training and parameter optimization.

4.3. *Global and local explanations*

Feature importance measures provide a way to explain machine learning models to domain experts. The more important a feature is in the model, the more important it is for the decision-making process in which the model is applied. Generally, the explainability of a model can be divided into global and local interpretability (Molnar,

[10]Tree-based classifiers do not require feature scaling and gave better results without normalization.

[11]The mean and standard deviation determined from normalizing the training data were used to normalize the validation data.

2019). Global interpretability means that the entire trained model is used in the explanation. Some models, such as the above explained logistic regression, random forest, and gradient boosting classifiers, have inbuilt modular global interpretability methods that provide information between the input data and the prediction. Local interpretability, on the other hand, refers to the results of a trained model on a specific input. For a feedforward transformation of the features, the local and global interpretability can be considered with a single trained model when the global feature sensitivity is aggregated over the observations (Saarela & Kärkkäinen, 2015). A general framework for this purpose is the local interpretable model-agnostic explanations (LIME) tool developed by Ribeiro et al. (2016). LIME provides the rules and features that are important for classifying a specific observation. It accomplishes this by weighting neighboring observations by their proximity to the observation being explained. The explanation is obtained by training a local linear model based on the weighted neighbouring observations. Thus, in order to explain specific predictions by our models, we also trained explainers with the `TabulaLimeExplainer` function (Ribeiro et al., 2016).

5. Results

Next, we present the experimental results with the datasets and methods depicted in Sections 3 and 4, separately analyzing the overrated (Section 5.1), changed (Section 5.2), and different ranks (Section 5.3).

5.1. Analysis of deviating ranks

Different countries have implemented diverse PRFSs. As described in Section 2, the Norwegian and Danish PRFSs (Sivertsen, 2018, 2016b) are the foundation of the Finnish PRFS. This means that normally, the publication channel ranks in *JuFo* are similar to those in the Norwegian and Danish PRFS, which were used as the reference values during the very first overall evaluation in 2012. Thus, in the first part of our triangulated analysis, we considered a publication channel rank as overrated if the difference to the Norwegian and Danish ranks is at least 1. As rank 2 is the highest in Norway and Denmark, a channel is not labeled overrated if it is rank 3 in Finland and rank 2 in Norway or Denmark. Moreover, it is enough if at least one of the Nordic counterpart ranks is similar to the Finnish rank. In summary, the rules for overrated publication channels are (FIN $>$ 1 & NOR $\leq$ 1 & DK $\leq$ 1) $\vee$ (FIN $=$ 1 & NOR $<$ 1 & DK $<$ 1). This means that by definition only rank 1, rank 2, and rank 3 publication channels can be overrated.

Table 4: Mean number of publication in overrated channels versus the non-overrated channels.

type	**all channels**	**overrated channels**	**non-overrated channels**
journal	5.23	5.87	5.08
conference	5.04	8.8	2.78
all	5.22	5.99	5.05

To investigate if the publication activity of Finnish scholars affect the *JuFo* rank (or vice versa), we compared the average publication activity for publication channels with and without overrated ranks. Table 4 shows the average number of publications for all channels in *JuFo*, for the overrated channels, and for the non-overrated channels. The overrated channels clearly have the most publications. This is especially true for conferences. Overrated conference publication channels have, on average, 6.02 publications more than the non-overrated ones (see Table 4). This means that publication activity could serve as a feature when predicting if a publication channel is overrated. The higher evaluation of publication channels with more Finnish publication activity was already recognized by Saarela et al. (2016).

Another characteristic of overrated ranks might be the publication language of the channel. Fig. 2 shows the difference in publication activity of Finnish researchers in overrated and not overrated publication channels of different languages. As noted previously, the average Finnish publication activity is generally larger in overrated than in non-overrated channels. However, there are considerable differences between the languages of the publication channels. In fact, the publication activity difference between overrated and not overrated publication channels is largest for those channels that publish articles in Finnish. The mean number of publications of Finnish scholars in Finnish publication channels is 63 for non-overrated channels and 85 for overrated ones, +35% (while it is generally 5 in non-overrated and 6 in overrated ones, +20%, see Table 4). Thus, the language of a publication channel appears to be another important indicator for the unexpected ranks.

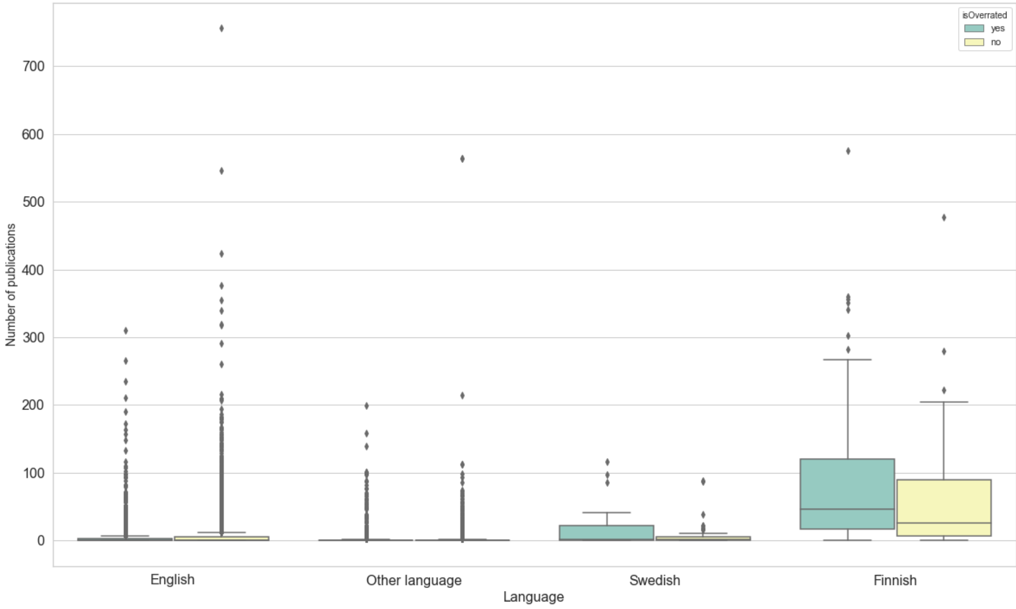

Fig. 2. Boxplot showing the publication activity of Finnish researchers in overrated and not overrated publication channels of different languages. The plot shows that there are considerably more publications of Finnish researchers in overrated channels that publish articles in Finnish and Swedish language (i.e., the two national languages).

On the other hand, there are also 148 underrated publication channels in *JuFo*. Similar to overrated publication channels, the underrated ones are defined as the subset of publication channels for which both reference indicators are higher than the Finnish rank (NOR = 2 & DAN = 2 & FIN < 2) | (NOR ≥ 1 & DAN ≥ 1 & FIN = 0). Again, this subset has a clear profile. Most of the publication channels are evaluated by panels 17–23 (i.e., the panels with a social sciences and humanities orientation, see Table 2), they have zero or very few publications by national scholars,[12] and 40% of these are published in a language other than English, Finnish, or Swedish. The remaining underrated publication channels are mainly published in English language, but also have either zero or very few publications of Finnish scholars or many publications but with a very high volume,[13] which probably caused the underrating.

To predict whether a channel is overrated, we leveraged all features depicted in Table 3 as input for the classifiers, except the Norwegian and Danish levels, which were used to define the output variable ("overrated"). As outlined in Section 4, we used one linear (logistic regression) and two nonlinear tree-based classifiers for the prediction. The division into the training and validation sets and the parameter grid search on the training data were performed as described in Section 4.

Table 5 shows the validation set performance results for the different classifiers predicting overrated publication channels. It also shows the performance differences when `SMOTE` oversampling was applied on the minority class observations of the training data (i.e., those publication channels in the training set that had the label "overrated") and whether the number of publications of Finnish scholars was included as a feature in the models. Clearly, `SMOTE` oversampling improved the f1 performance of all classifiers, and removing the number of publications decreased the performance of such classifiers. The overall best-performing classifier predicting if a publication channel was

[12]There are two exceptions to this rule, the *European Early Childhood Education Research Journal* and the *Temanord*. The latter can be explained by the fact that this is a Danish publication channel.

[13]Examples of the case of the latter include the journals *Astrophysical Journal*, and *Electrochimica Acta*, which have a volume of 2,900 and 2,250 respectively.

Table 5: Validation set performance of classifiers predicting if a publication channel is overrated (the best values in each column are emphasized). Since the number of publications feature is ambivalent (see our discussion in the last paragraph in Section 2), the table shows the performance for both the models including this feature (central three columns) and excluding it (last three columns).

	all features			number of publications excluded		
model	**precision**	**recall**	**f1-score**	**precision**	**recall**	**f1-score**
logistic regression	*0.829*	0.728	0.756	*0.829*	0.726	0.755
logistic regression with SMOTE	0.825	0.735	0.761	0.825	0.733	0.759
random forest	0.826	0.844	0.828	0.820	0.839	0.823
random forest with SMOTE	0.827	0.843	0.830	0.826	*0.841*	*0.830*
gradient boosting	*0.829*	*0.847*	0.831	0.817	0.831	0.822
gradient boosting with SMOTE	0.828	0.839	*0.832*	0.820	0.832	0.825

overrated (with an f1 score of 0.832) was gradient boosting with SMOTE oversampling; it includes the number of publications of Finnish scholars as a feature in the model. Besides the number of publications, the most important features predicting with the gradient boosting classifier whether a channel was overrated were the rank of the publication channel, the binary feature indicating if the language of the publication channel is English, the three citation-based indicators (SJR, IPP, and SNIP), the binary indicator if the IPP is available, the start year of the publication channel, the frequency of the publisher in *Jufo*, and the volume of the publication channel.

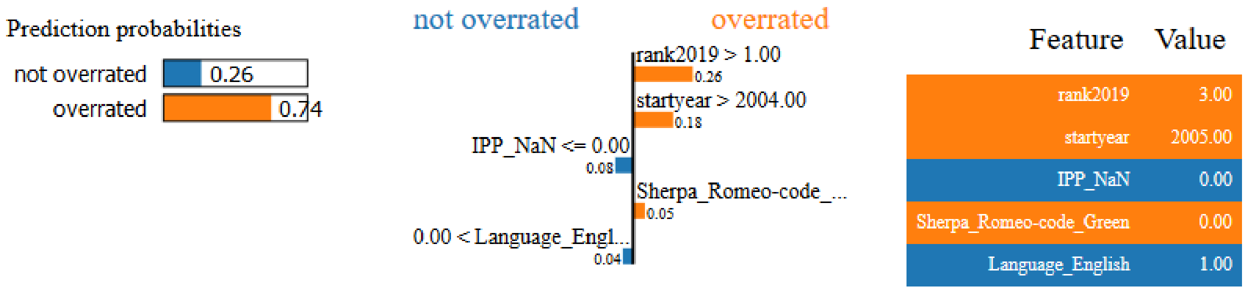

Fig. 3. Five most important LIME explanations of the gradient boosting classifiers predicting if the rank of a publication channel is overrated for the *International Journal of Education through Art*. The high current rank, the young age of the journal (start year 2005), and the green sherpa code were the most important local explanations classifying the channel as overrated (which it is). However, the language of the publication channel (English) and the fact that the IPP is available (not missing) were reasons for the model to classify this publication channel with some probability as not overrated.

With the LIME explainer, we can obtain the most important local rules indicating on what basis our models predict why individual publication channels are overrated. Fig. 3 shows the most important local importances when the LIME explainer was trained on top of the gradient boosting model for the example of the *International Journal of Education through Art*, i.e., the overrated publication channel in the validation set that had the highest predicted probability to be overrated and was actually overrated (true positive). This journal is rank 3 in *JuFo* but only rank 1 in Denmark and Norway. As can be seen from Fig. 3, the most important local importances explaining why this publication channel rank is overrated were the high current rank (rank 3), the young age of the journal (start year 2005), and the sherpa/romeo code (green). Such importances can be of interest when explaining to interested parties (such as scholars or research resource allocation decision makers) why the publication channel has a higher rank than expected.

5.2. *Analysis of changed ranks*

In the second part of the analysis, we aim at predicting publication channels for which the rank has changed. The *JuFo* database contains a column *JuFo-history* that depicts the rank for the previous years 2012–2018. Using this column, we compared the 2015 rank (presenting the previous major update of all *JuFo* ranks) of each publication channel with the current rank. We then defined a rank as "changed" if the *rank 2015* was different from the current rank (*rank 2019*). As input for the classification, we used all the features and their transformation explained in Table 3, except the current *JuFo* rank (*rank 2019*), which was used to define the output variable (*changed*).

Table 6: Validation set performance of classifiers predicting if the rank of a publication channel was recently changed. Since the number of publications feature is ambivalent (see our discussion in the last paragraph in Section 2), the table shows the performance for both the models including this feature (central three columns) and excluding it (last three columns).

	all features			number of publications excluded		
model	**precision**	**recall**	**f1-score**	**precision**	**recall**	**f1-score**
logistic regression	*0.906*	0.687	0.759	0.905	0.685	0.758
logistic regression with `SMOTE`	0.905	0.689	0.760	*0.906*	0.688	0.760
random forest	0.904	*0.924*	0.906	0.900	0.920	0.905
random forest with `SMOTE`	0.904	0.923	*0.908*	0.900	0.919	*0.906*
gradient boosting	0.896	0.922	0.889	0.893	*0.922*	0.888
gradient boosting with `SMOTE`	0.900	0.899	0.899	0.898	0.886	0.892

Table 6 shows the performance results for the different classifiers on the validation set. As can be seen from the table, random forest predicted the changed ranks the best with an f1 score of 0.906 without and 0.908 with `SMOTE` oversampling of the minority class in the training data (i.e., those publication channels in the training data with changed ranks). For both the random forest model and the slightly worse-performing gradient boosting model, the most important feature predicting the rank change is the start year of the publication channel (see Table 7). This makes sense, as publication channels with increased rank are often newer journals. The mean of the start year of all publication channels in *JuFo* for which the rank did not change is 1986, while the mean starting date of the publication channels with increased rank is 1997.

Table 7: Ten most important features of the two tree-based classifiers predicting if a rank was recently changed, arranged by importance.

random forest		gradient boosting	
feature	**importance**	**feature**	**importance**
startyear	0.21796	startyear	0.55994
NrOfPublications	0.08215	SJR	0.08695
SNIP	0.08103	SNIP	0.08695
SJR	0.07805	IPP	0.06326
IPP	0.07379	Denmark	0.04884
VOL	0.05638	NrOfPublications	0.04435
Denmark	0.03337	Denmark_NaN	0.03767
Norway	0.02455	VOL	0.02021
Denmark_NaN	0.02004	SNIP_NaN	0.01914
Sherpa_Romeo-code_Green	0.01721	Sherpa_Romeo-code_Green	0.01167

Fig. 4 shows the most important local importances for one publication channel in the validation set, the journal *Acta Dermato-Venereologica*, when the LIME explainer was trained on the best-performing random forest model. Similarly as before (see Section 5.1), this publication channel was chosen because out of all true positives (publication channels that were predicted to have a changed rank when their rank was actually changed) in the validation set, it had the highest probability. The rank of the *Acta Dermato-Venereologica* journal was changed from 1 to 2. As can be seen from the figure, the relatively large number of publications of Finnish researchers in this journal (69) and the values of the citation-based indicators were the most important local explanations why the random forest model predicted that the rank for this publication channel was recently changed.

5.3. Analysis of ranks

The last part of our triangulated analysis focuses on examining and predicting the current rank of an individual publication channel.

5.3.1. Fractionalization

To understand the difference between citation- and expert-based rankings of the publication channels in *JuFo*, we first fractionalized the data by using the three available citation-based quality indicators SJR, IPP, and SNIP. This means that we divided the available citation-based values into categories (0–3) such that the same frequencies of *JuFo* ranks were present also in the SJR, SNIP, and IPP categories. If we fractionalize all data as in previous bibliometric

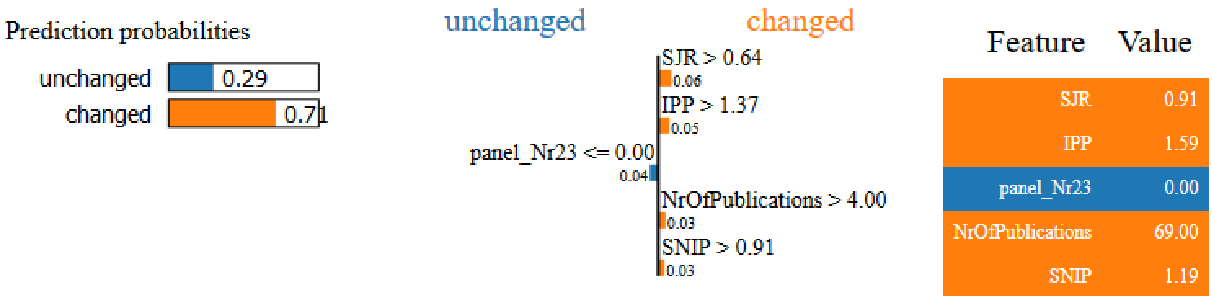

Fig. 4. Top five LIME explanations of the random forest classifiers predicting if the rank of the *Acta Dermato-Venereologica* publication channel was recently changed. The model was 71% sure that this publication channel was changed (correct), mainly because of the large number of publication of Finnish researchers in this channel and the values of the citation-based indicators.

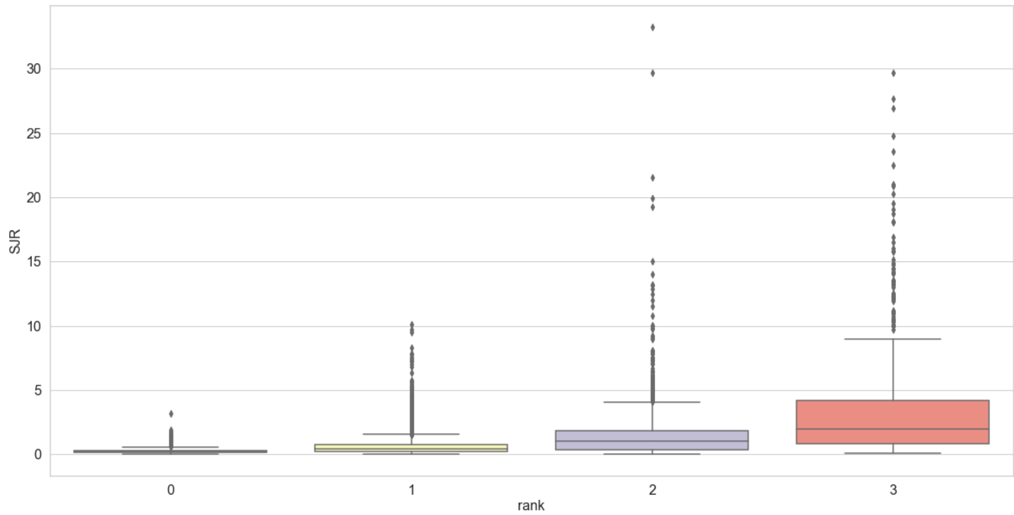

Fig. 5. Boxplot showing the distribution of the SJR per rank. On average, publication channels that have a higher SJR also have a higher rank.

articles (e.g., Ahlgren & Waltman, 2014),[14] a publication channel should generally be rank 3 if it has an IPP in (4.68, 89.23], rank 2 if it has an IPP in (2.48, 4.68], and rank 1 if it has an IPP in (0.0, 2.47]. The values for SJR are for rank 3 if in (2.559, 39.285], rank 2 if in (1.171, 2.558], rank 1 if in (0.1, 1.17], and rank 0 if smaller than 0.1. The categories for SNIP are for rank 3 if in (2.308, 50.569], rank 2 if in (1.323, 2.308], and rank 1 if smaller than 1.323.

However, it seems that a general division into categories is not meaningful, as it must also take the discipline into account. The differences of the citation-based indicators between the different panels are considerable. Fig. 5 shows the distribution of the SJR for each rank (0–3). The figure demonstrates that on average, publication channels with a higher SJR are also ranked higher. Fig. 6 shows the same information but panelwise. While it displays again that on average, higher-ranked publication channels have a higher SJR, it also demonstrates the large differences in the actual SJR values between the panels. For example, a publication channel that is in the range of rank 3 for panel 23 (i.e., the panel for *history, archaeology and cultural studies*) would only be in the range of rank 1 for panel 7 (i.e., the panel for *biosciences II*). This is true although the SJR tries to be field-normalized by its algorithm. The same

[14]Note that fractionalization in bibliometrics has also been used to compute for how much of a publication a co-author can take credit for (e.g., Mutz & Daniel, 2019) and to normalize citation-based indicators by taking citation behavior among citing authors into account (e.g., Leydesdorff & Opthof, 2010).

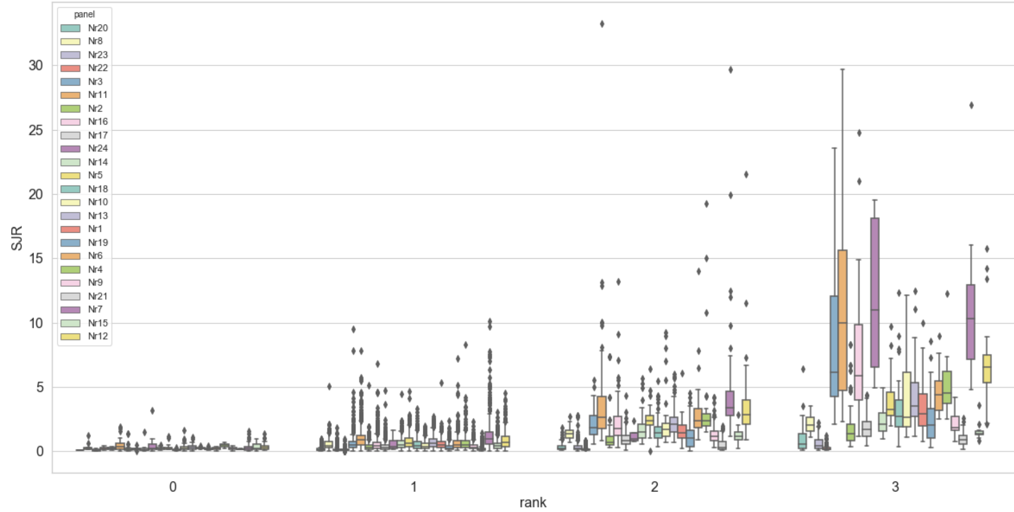

Fig. 6. Boxplot showing the distribution of the SJR per rank for the different panels. In general, publication channels with a higher SJR have a higher rank. However, there are considerable differences between the different panels (Note: The panels are listed in Table 2).

applies to the IPP in an even more distinctive manner since the IPP is not field-normalized.[15] In summary, these depictions clearly illustrate the known differences in citation practices between disciplines (Kulczycki et al., 2018; Sīle & Vanderstraeten, 2019; Sivertsen & Larsen, 2012; Sivertsen, 2016a; Verleysen & Engels, 2014).

The SNIP measures the contextual citation impact by weighting citations based on the total number of citations in a discipline (Moed, 2010). Thus, it should be like the SJR more balanced between the different disciplines. However, even for the SNIP, there are still considerable differences between the panels.[16] In summary, the citation-based indicators provide a good suggestion for the rank of a publication channel, but the rank cannot be evaluated or estimated by solely using this information.

5.3.2. *Prediction of ranks*

To predict the rank and determine the features that influence it the most, we utilized once again the logistic regression and the two tree-based classifiers with a division into the training and validation sets, as well as the parameter grid search on the training data, as described in Section 4. All features depicted in Table 3 were leveraged as input for the classifiers, except the rank itself, which readily defined the multinomial output variable.

Table 8 summarizes the validation set performance of the different classifiers by using the best parameters from the grid search and different feature sets (the number of publications included or not) with and without SMOTE oversampling. As can be seen from the table, the nonlinear classifiers (gradient boosting and random forest) again outperformed the linear one. In fact, the best f1 score on the validation set was 0.816 for logistic regression with l_1 penalization, whereas it was 0.879 for random forest and even 0.88 for gradient boosting. Gradient boosting was the best method throughout (see the emphasized values in Table 8). As the classification task seems to be a nonlinear problem, we focused on the better-performing tree-based classifiers (i.e., random forest and gradient boosting) to determine the important features. The figure showing the feature importances for random forest and gradient boosting can be found online.[17] As can be seen from the figure, the models emphasized different features, but the 10 most important ones were the same, even though their order was different. These 10 most important features were the three

[15]This is illustrated in the figure at `https://doi.org/10.6084/m9.figshare.9928775`.
[16]This is illustrated in the figure at `https://doi.org/10.6084/m9.figshare.9928799`.
[17]See `https://doi.org/10.6084/m9.figshare.9928997`

Table 8: Validation set performance of classifiers predicting the rank of a publication channel by using the best parameters from the grid search with fivefold cross-validation on the training set. Since the number of publications feature is ambivalent (see our discussion in the last paragraph in Section 2), the table shows the performance for both the models including this feature (central three columns) and excluding it (last three columns).

	all features			number of publications excluded		
model	**precision**	**recall**	**f1-score**	**precision**	**recall**	**f1-score**
logistic regression	0.830	0.730	0.758	0.830	0.730	0.757
logistic regression with `SMOTE`	0.824	0.841	0.816	0.822	0.840	0.815
random forest	0.876	0.882	0.877	0.867	0.874	0.869
random forest with `SMOTE`	*0.878*	0.881	0.879	0.873	0.877	*0.875*
gradient boosting	0.877	0.883	0.878	0.869	0.870	0.869
gradient boosting with `SMOTE`	*0.878*	*0.884*	*0.880*	*0.873*	*0.880*	*0.875*

citation-based indicators (i.e., SJR, SNIP, and IPP), the Norwegian and Danish ranks, the binary indicator specifying whether the Danish rank is available, the start year of publication channel, the number of publications of Finnish scholars, the frequency of the publication channel's publisher in *JuFo*, and its volume.

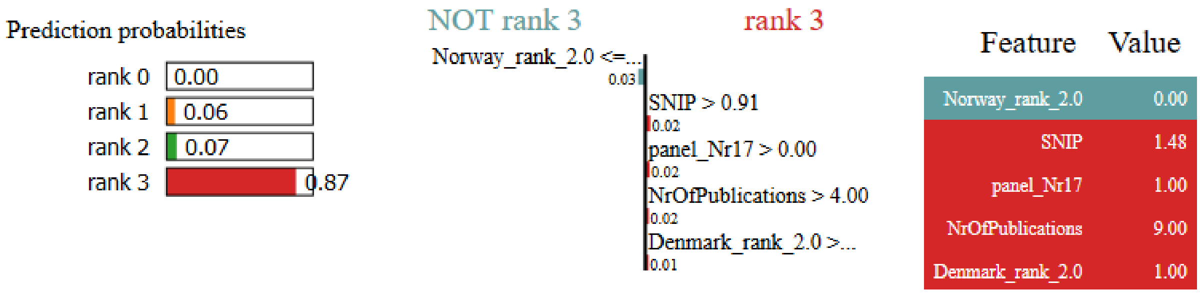

Fig. 7. Top five LIME explanations of the gradient boosting classifiers predicting the rank of the *Journal of Informetrics* publication channel. The model is 87% sure that this publication channel has the highest rank, rank 3 (correct rank) because the rank in Denmark is the highest, and the citation-based indicator SJR and the number of publications of Finnish researchers are higher than the thresholds for panel 17 (*social sciences, media, and communications; interdisciplinary social sciences*). The model predicts tiny probabilities that this publication channel is rank 2 or rank 1 but is 100% certain that it cannot be rank 0. The most important feature indicating that this channel might not be rank 3 is the rank in Norway, which is only 1.

With the LIME explainer, we can identify again the most important rules determining the rank of an individual publication channel. Fig. 7 shows the most important local importances when LIME is used on top of the gradient boosting model for the example of the *Journal of Informetrics* publication channel. As can be seen from the figure, one of the most important features for the prediction was the Danish rank, which is the highest. Other local explanations classifying this publication channel into the highest rank category were the citation-based indicator SJR and the number of publications of Finnish scholars, which were both higher than the thresholds for the panel responsible for evaluating this publication channel (i.e., the panel for *social sciences, media, and communications; interdisciplinary social sciences*).

6. Conclusions

Quality of scholarly research activities can be quantified in various ways. In the Finnish PRFS system, according to the Norwegian model (Aagaard, 2019; Pölönen, 2018; Sivertsen, 2019), the quality of a publication channel is being ranked by expert panels. The purpose of this paper was to test if such an expert-based rank assignment of the publication indicator can be explained and automated. The automation was studied from three different perspectives utilizing several techniques, always highlighting the features with the greatest effect on the ranking. For researchers using the system and leveraging it to determine where they should submit their publications or which publications to read, it is of interest why certain publication channels have a certain rank. Thus, it is not enough if an automated system accurately classifies a publication channel into a certain category. It is the explanations of why this channel belongs to a category that provides helpful guidelines. Hence, we analyzed not only the classification models but also the feature importances.

The analysis showed that the three citation-based indicators (i.e., SJR, SNIP, and IPP), the Norwegian and Danish ranks, the binary indicator specifying whether the Danish rank is available, the start year of the publication channel, its volume, the number of publications of Finnish scholars in this publication channel, and the frequency of the publication channel's publisher are the 10 most important features classifying a publication channel into a rank. These features were identified to be significant by both the nonlinear methods used in this work (random forest and gradient boosting). The number of publications of Finnish scholars was found to be the most important feature predicting if a publication channel rank in the Finnish system is overrated compared with the rank of its Nordic counterparts, or the tendency of a rank to be improved during reevaluations. These findings emphasize the high correlation between citation- and expert-based rankings in such cases in which matured citation-based indicators are available, also confirming the conclusions of earlier studies (Akusok et al., 2019; Saarela et al., 2016), despite the use of different techniques, another study design, and the old dataset instance in these.[18]

We found many cases illustrating the known differences in citation practices between different disciplines (Kulczycki et al., 2018; Sīle & Vanderstraeten, 2019; Sivertsen & Larsen, 2012; Sivertsen, 2016a; Verleysen & Engels, 2014). The special role of language (the native language or that other than English) or the type of publication channel (highly rated conferences in *computer and information sciences*) when a PRFS covers all disciplines was highlighted. The language issue is particularly difficult to model appropriately: even if citation-based indicators can be normalized by the averages of citation metrics in the discipline, the number of national scholars within a discipline is not available for such a normalization.

The start year of a publication channel was the most important feature predicting whether a rank was changed. Hence, an expert-based rank that was higher than expected was linked to a newer publication channel. This finding shows that the expert-based system can adapt to the possible improvements of the quality of a publication channel faster than citation-based indicators can, which, by construction, are slower in improving their value.

The relation between national publication activity (i.e., the feature specifying the number of publications of Finnish scholars in respective publication channels) and the rank of a publication channel is a complicated issue. On one hand, our analysis showed that the performance of all prediction models improved when the number of publications was leveraged as a feature. On the other hand, it is obvious that this feature should not be used when building a system that automatically classifies publication channels into ranks. If it will be used in automatic models, national scholars can easily realize adversarial attacks (Yuan et al., 2019) by intensively publishing in a channel whose rank is to be increased artificially.

The purpose of ranking is to provide simple access to better publication channels where national visibility should be increased. However, when monetary incentives come into play, there will be a push from the scholarly community to increase the ranks of already *conquered* channels. As we have shown, there is a clear dependency between expert rank and publication activity, but this is not necessarily a causal effect (Waltman, 2017). The fact that those channels that recently improved their ranking are largely characterized by high national publication activity is an alarming finding. Nevertheless, our analysis also showed that the expert-based system adapts to the prospective changes of the publication channel quality faster than do citation-based indicators. Thus, the relation between quantity and quality of publications (i.e., productivity and impact) remains to be an interesting study topic (Kolesnikov et al., 2018; Sandström & van den Besselaar, 2016).

Naturally, the results presented here are limited to the used techniques and data. A further study could repeat the presented analysis with data from another country. Several countries (including Denmark, Flanders in Belgium, and Poland) have adopted a national PRFS inspired by the Norwegian model, and others, such as Sweden, have implemented the model locally at several national research institutes (Hammarfelt, 2018). It would be interesting to analyze how closely machine learning models could predict the other national quality indicators and to determine what (if any) differences exist between the countries. As shown by Ahlgren & Waltman (2014) through univariate analysis, single citation-based indicators (such as the SNIP) had some predictive power in classifying publication channels into ranks in Norway. Enlarging the input data space with more features and utilizing multivariate techniques, such as the machine learning methods leveraged in this study, would certainly improve prediction performance. Again, it would

[18] Both studies used the *JuFo* data before the reevaluations, Akusok et al. (2019) used extreme learning machines to study the overrated higher-level publication channels, whereas Saarela et al. (2016) used confusion matrices, association rule mining, and decision trees to identify characteristics of publication channel ranks.

be interesting to determine the most important features in comparison to the results obtained from the data of the Finnish publication indicator presented here.

Finally, having an automated predictor of publication channel quality could save costs and efforts in all countries with similar PRFSs. Introducing, managing, and maintaining such systems create costs for governments and essentially taxpayers. If the publication indicator in PRFSs would use automatic systems instead of expert-based classifications, these costs would be reduced or even vanish completely. Moreover, one part of the Norwegian model emphasizes the importance of open data. Making not only the underlying data but also the feature importances (of both the models and the individual classifications) openly available and visible in such systems would support individual scholars and other interest groups using the system, such as research policy decision makers, in understanding certain classifications better. This understanding, in turn, can help drive more informed decisions about what to read and where to publish, both nationally and internationally.

Acknowledgement

This research was supported by the Academy of Finland (grants no. 315550 and 311877) and is related to the thematic research area DEMO (Decision Analytics Utilizing Causal Models and Multiobjective Optimization, jyu.fi/demo) of the University of Jyväskylä.

References

Aagaard, K. (2019). Performance-based Research Funding in Denmark: The Adoption and Translation of the Norwegian Model. *Journal of Data and Information Science*, *3*, 20–30.

Aagaard, K., Bloch, C., & Schneider, J. W. (2015). Impacts of performance-based research funding systems: The case of the Norwegian Publication Indicator. *Research Evaluation*, *24*, 106–117.

Abrishami, A., & Aliakbary, S. (2019). Predicting citation counts based on deep neural network learning techniques. *Journal of Informetrics*, *13*, 485 – 499. URL: http://www.sciencedirect.com/science/article/pii/S1751157718303821. doi:https://doi.org/10.1016/j.joi.2019.02.011.

Ahlgren, P., & Waltman, L. (2014). The correlation between citation-based and expert-based assessments of publication channels: SNIP and SJR vs. Norwegian quality assessments. *Journal of Informetrics*, *8*, 985 – 996. URL: http://www.sciencedirect.com/science/article/pii/S1751157714000911. doi:https://doi.org/10.1016/j.joi.2014.09.010.

Akusok, A., Saarela, M., Kärkkäinen, T., Björk, K.-M., & Lendasse, A. (2019). Mislabel Detection of Finnish Publication Ranks. In J. Cao, C. M. Vong, Y. Miche, & A. Lendasse (Eds.), *Proceedings of ELM-2017* (pp. 240–248). Cham: Springer International Publishing.

Bach, M., Werner, A., Żywiec, J., & Pluskiewicz, W. (2017). The study of under-and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis. *Information Sciences*, *384*, 174–190.

Bai, X., Zhang, F., & Lee, I. (2019). Predicting the citations of scholarly paper. *Journal of Informetrics*, *13*, 407 – 418. URL: http://www.sciencedirect.com/science/article/pii/S1751157718301767. doi:https://doi.org/10.1016/j.joi.2019.01.010.

Bensman, S. J., Smolinsky, L. J., & Pudovkin, A. I. (2010). Mean citation rate per article in mathematics journals: Differences from the scientific model. *Journal of the American Society for Information Science and Technology*, *61*, 1440–1463.

Boemer, F., Ratner, E., & Lendasse, A. (2018). Parameter-free image segmentation with SLIC. *Neurocomputing*, *277*, 228–236.

Bornmann, L., Adams, J., & Leydesdorff, L. (2018). The negative effects of citing with a national orientation in terms of recognition: National and international citations in natural-sciences papers from Germany, the Netherlands, and the UK. *Journal of Informetrics*, *12*, 931 – 949. URL: http://www.sciencedirect.com/science/article/pii/S1751157718300336. doi:https://doi.org/10.1016/j.joi.2018.07.009.

Breiman, L. (1997). *Arcing the edge*. Technical Report Statistics Department, University of California, Berkeley.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32. doi:10.1023/A:1010933404324.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Butler, L. (2003). Explaining Australia's increased share of ISI publications—the effects of a funding formula based on publication counts. *Research policy*, *32*, 143–155.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Daud, A., Ahmad, M., Malik, M., & Che, D. (2015). Using machine learning techniques for rising star prediction in co-author network. *Scientometrics*, *102*, 1687–1711.

Dondio, P., Casnici, N., Grimaldo, F., Gilbert, N., & Squazzoni, F. (2019). The "invisible hand" of peer review: The implications of author-referee networks on peer review in a scholarly journal. *Journal of Informetrics*, *13*, 708 – 716. URL: http://www.sciencedirect.com/science/article/pii/S1751157718304206. doi:https://doi.org/10.1016/j.joi.2019.03.018.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, (pp. 1189–1232).

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

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Haddawy, P., Hassan, S.-U., Asghar, A., & Amin, S. (2016). A comprehensive examination of the relation of three citation-based journal metrics to expert judgment of journal quality. *Journal of Informetrics*, *10*, 162–173.

Halim, Z., & Khan, S. (2019). A data science-based framework to categorize academic journals. *Scientometrics*, *119*, 393–423. URL: https://doi.org/10.1007/s11192-019-03035-w. doi:10.1007/s11192-019-03035-w.

Hammarfelt, B. (2018). Taking Comfort in Points: The Appeal of the Norwegian Model in Sweden. *Journal of Data and Information Science*, *3*, 84–94.

Hassan, S.-U., Safder, I., Akram, A., & Kamiran, F. (2018). A novel machine-learning approach to measuring scientific knowledge flows using citation context analysis. *Scientometrics*, *116*, 973–996.

Heinisch, D. P., & Buenstorf, G. (2018). The next generation (plus one): an analysis of doctoral students' academic fecundity based on a novel approach to advisor identification. *Scientometrics*, *117*, 351–380. URL: https://doi.org/10.1007/s11192-018-2840-5. doi:10.1007/s11192-018-2840-5.

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., & Rafols, I. (2015). The Leiden Manifesto for research metrics. *Nature News*, *520*, 429. URL: https://doi.org/10.1038/520429a. doi:10.1038/520429a.

Kim, J., & Kim, J. (2018). The impact of imbalanced training data on machine learning for author name disambiguation. *Scientometrics*, *117*, 511–526. URL: https://doi.org/10.1007/s11192-018-2865-9. doi:10.1007/s11192-018-2865-9.

Kolesnikov, S., Fukumoto, E., & Bozeman, B. (2018). Researchers' risk-smoothing publication strategies: Is productivity the enemy of impact? *Scientometrics*, *116*, 1995–2017. URL: https://doi.org/10.1007/s11192-018-2793-8. doi:10.1007/s11192-018-2793-8.

Kulczycki, E., Engels, T. C., Pölönen, J., Bruun, K., Dušková, M., Guns, R., Nowotniak, R., Petr, M., Sivertsen, G., Starčič, A. I. et al. (2018). Publication patterns in the social sciences and humanities: evidence from eight European countries. *Scientometrics*, (pp. 1–24).

Kulczycki, E., Korzeń, M., & Korytkowski, P. (2017). Toward an excellence-based research funding system: Evidence from Poland. *Journal of Informetrics*, *11*, 282 – 298. URL: http://www.sciencedirect.com/science/article/pii/S1751157716302231. doi:https://doi.org/10.1016/j.joi.2017.01.001.

Letto-Vanamo, P. (2019). Evaluation of academic legal publications in Finland. In R. van Gestel, & A. Lienhard (Eds.), *Evaluating Academic Legal Research in Europe: The Advantage of Lagging Behind* chapter 8. (p. 218–237). Edward Elgar. doi:https://doi.org/10.4337/9781788115506.00013.

Leydesdorff, L., & Opthof, T. (2010). Scopus's source normalized impact per paper (SNIP) versus a journal impact factor based on fractional counting of citations. *Journal of the American Society for Information Science and Technology*, *61*, 2365–2369. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21371. doi:10.1002/asi.21371. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.21371.

Li, X., Rong, W., Shi, H., Tang, J., & Xiong, Z. (2018). The impact of conference ranking systems in computer science: a comparative regression analysis. *Scientometrics*, *116*, 1–29.

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data* volume 793. Wiley.

Meho, L. I. (2019). Using Scopus's CiteScore for assessing the quality of computer science conferences. *Journal of Informetrics*, *13*, 419 – 433. URL: http://www.sciencedirect.com/science/article/pii/S1751157718304176. doi:https://doi.org/10.1016/j.joi.2019.02.006.

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of informetrics*, *4*, 265–277.

Molnar, C. (2019). *Interpretable Machine Learning*. Lean Publishing.

Mutz, R., & Daniel, H.-D. (2019). How to consider fractional counting and field normalization in the statistical modeling of bibliometric data: A multilevel Poisson regression approach. *Journal of Informetrics*, *13*, 643 – 657. URL: http://www.sciencedirect.com/science/article/pii/S1751157718302852. doi:https://doi.org/10.1016/j.joi.2019.03.007.

Olson, R., Cava, W., Mustahsan, Z., Varik, A., & Moore, J. (2018). Data-driven advice for applying machine learning to bioinformatics problems. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (pp. 192–203). volume 23.

Pölönen, J. (2018). Applications of, and Experiences with, the Norwegian Model in Finland. *Journal of Data and Information Science*, *3*, 31–44.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You? Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). ACM.

Rodríguez-Navarro, A., & Brito, R. (2019). Probability and expected frequency of breakthroughs: basis and use of a robust method of research assessment. *Scientometrics*, *119*, 213–235. URL: https://doi.org/10.1007/s11192-019-03022-1. doi:10.1007/s11192-019-03022-1.

Saarela, M., & Kärkkäinen, T. (2015). Analysing Student Performance using Sparse Data of Core Bachelor Courses. *Journal of Educational Data Mining*, *7*, 3–32.

Saarela, M., Kärkkäinen, T., Lahtonen, T., & Rossi, T. (2016). Expert-based versus citation-based ranking of scholarly and scientific publication channels. *Journal of Informetrics*, *10*, 693–718.

Saarela, M., Ryynänen, O.-P., & Äyrämö, S. (2019). Predicting hospital associated disability from imbalanced data using supervised learning. *Artificial Intelligence in Medicine*, *95*, 88 – 95. URL: http://www.sciencedirect.com/science/article/pii/S0933365718303063. doi:https://doi.org/10.1016/j.artmed.2018.09.004.

Sandström, U., & van den Besselaar, P. (2018). Funding, evaluation, and the performance of national research systems. *Journal of Informetrics*, *12*, 365 – 384. URL: http://www.sciencedirect.com/science/article/pii/S1751157717302882. doi:https://doi.org/10.1016/j.joi.2018.01.007.

Sandström, U., & van den Besselaar, P. (2016). Quantity and/or Quality? The Importance of Publishing Many Papers. *PloS one*, *11*, e0166149.

Serenko, A., & Bontis, N. (2018). A critical evaluation of expert survey-based journal rankings: The role of personal research interests. *Journal of the Association for Information Science and Technology*, *69*, 749–752.

Sīle, L., & Vanderstraeten, R. (2019). Measuring changes in publication patterns in a context of performance-based research funding systems: the case of educational research in the University of Gothenburg (2005–2014). *Scientometrics*, *118*, 71–91. URL: https://doi.org/10.1007/s11192-018-2963-8. doi:10.1007/s11192-018-2963-8.

Sivertsen, G. (2016a). Patterns of internationalization and criteria for research assessment in the social sciences and humanities. *Scientometrics*, *107*, 357–368.

Sivertsen, G. (2016b). Publication-Based Funding: The Norwegian Model. In M. Ochsner, S. E. Hug, & H.-D. Daniel (Eds.), *Research Assessment in the Humanities: Towards Criteria and Procedures* (pp. 79–90). Cham: Springer International Publishing. URL: https://doi.org/10.1007/978-3-319-29016-4_7. doi:10.1007/978-3-319-29016-4_7.

Sivertsen, G. (2018). The Norwegian Model in Norway. *Journal of Data and Information Science*, *3*, 3 – 19. URL: https://content.sciendo.com/view/journals/jdis/3/4/article-p3.xml.

Sivertsen, G. (2019). Understanding and Evaluating Research and Scholarly Publishing in the Social Sciences and Humanities (SSH). *Data and Information Management*, *3*, 61–71. URL: https://content.sciendo.com/view/journals/dim/ahead-of-print/article-10.2478-dim-2019-0008.xml.

Sivertsen, G., & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: An empirical analysis of the potential. *Scientometrics*, *91*, 567–575.

Small, H. (2018). Characterizing highly cited method and non-method papers using citation contexts: The role of uncertainty. *Journal of Informetrics*, *12*, 461 – 480. URL: http://www.sciencedirect.com/science/article/pii/S1751157718300440. doi:https://doi.org/10.1016/j.joi.2018.03.007.

Subochev, A., Aleskerov, F., & Pislyakov, V. (2018). Ranking journals using social choice theory methods: A novel approach in bibliometrics. *Journal of Informetrics*, *12*, 416 – 429. URL: http://www.sciencedirect.com/science/article/pii/S1751157718300919. doi:https://doi.org/10.1016/j.joi.2018.03.001.

Thelwall, M. (2017). Three practical field normalised alternative indicator formulae for research evaluation. *Journal of Informetrics*, *11*, 128 – 151. URL: http://www.sciencedirect.com/science/article/pii/S175115771630205X. doi:https://doi.org/10.1016/j.joi.2016.12.002.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*, 267–288.

Treeratpituk, P., & Giles, C. L. (2009). Disambiguating Authors in Academic Publications using Random Forests. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries* (pp. 39–48). ACM.

Tüselmann, H., Sinkovics, R. R., & Pishchulov, G. (2015). Towards a consolidation of worldwide journal rankings – a classification using random forests and aggregate rating via data envelopment analysis. *Omega*, *51*, 11 – 23. URL: http://www.sciencedirect.com/science/article/pii/S0305048314000966. doi:https://doi.org/10.1016/j.omega.2014.08.002.

van den Besselaar, P., Heyman, U., & Sandström, U. (2017). Perverse effects of output-based research funding? Butler's Australian case revisited. *Journal of Informetrics*, *11*, 905 – 918. URL: http://www.sciencedirect.com/science/article/pii/S1751157717301943. doi:https://doi.org/10.1016/j.joi.2017.05.016.

Vanhoeyveld, J., & Martens, D. (2018). Imbalanced classification in sparse and large behaviour datasets. *Data Mining and Knowledge Discovery*, *32*, 25–82. URL: https://doi.org/10.1007/s10618-017-0517-y. doi:10.1007/s10618-017-0517-y.

Verleysen, F. T., & Engels, T. C. (2014). Internationalization of peer reviewed and non-peer reviewed book publications in the social sciences and humanities. *Scientometrics*, *101*, 1431–1444.

Vieira, E., & Gomes, J. (2010). Citations to scientific articles: Its distribution and dependence on the article features. *Journal of Informetrics*, *4*, 1 – 13. URL: http://www.sciencedirect.com/science/article/pii/S1751157709000534. doi:https://doi.org/10.1016/j.joi.2009.06.002.

van Vlokhoven, H. (2019). The effect of open access on research quality. *Journal of Informetrics*, *13*, 751 – 756. URL: http://www.sciencedirect.com/science/article/pii/S1751157718304814. doi:https://doi.org/10.1016/j.joi.2019.04.001.

Vrettas, G., & Sanderson, M. (2015). Conferences versus journals in computer science. *Journal of the Association for Information Science and Technology*, *66*, 2674–2684. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.23349. doi:10.1002/asi.23349. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23349.

Wallace, F. H., & Perri, T. J. (2018). Economists behaving badly: publications in predatory journals. *Scientometrics*, *115*, 749–766. URL: https://doi.org/10.1007/s11192-018-2690-1. doi:10.1007/s11192-018-2690-1.

Walters, W. H. (2017). Do subjective journal ratings represent whole journals or typical articles? Unweighted or weighted citation impact? *Journal of Informetrics*, *11*, 730 – 744. URL: http://www.sciencedirect.com/science/article/pii/S1751157717300925. doi:https://doi.org/10.1016/j.joi.2017.05.001.

Walters, W. H., & Markgren, S. (2019). Do faculty journal selections correspond to objective indicators of citation impact? Results for 20 academic departments at Manhattan College. *Scientometrics*, *118*, 321–337. URL: https://doi.org/10.1007/s11192-018-2972-7. doi:10.1007/s11192-018-2972-7.

Waltman, L. (2017). Special section on performance-based research funding systems. *Journal of Informetrics*, *11*, 904. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85020481507&doi=10.1016%2fj.joi.2017.05.015&partnerID=40&md5=3a12f3d62248a5d33437ace91247871a. doi:10.1016/j.joi.2017.05.015. Cited By 2.

Yuan, X., He, P., Zhu, Q., & Li, X. (2019). Adversarial Examples: Attacks and Defenses for Deep Learning. *IEEE Transactions on Neural Networks and Learning Systems*, (pp. 1–20). doi:10.1109/TNNLS.2018.2886017. Early Access, published online on January 14, 2019.

Zacharewicz, T., Jonkers, K., Lepori, B., & Reale, E. (2019). Performance-based research funding in EU Member States—a comparative assessment. *Science and Public Policy*, *46*, 105–115. URL: https://doi.org/10.1093/scipol/scy041. doi:10.1093/scipol/scy041. arXiv:http://oup.prod.sis.lan/spp/article-pdf/46/1/105/27775768/scy041.pdf.

Zhang, R., Wang, J., & Mei, Y. (2017). Search for evergreens in science: A functional data analysis. *Journal of Informetrics*, *11*, 629 – 644. URL: http://www.sciencedirect.com/science/article/pii/S1751157716303583. doi:https://doi.org/10.1016/j.joi.2017.05.007.