Austrian Journal of Statistics
June 2020, Volume 49, 35–58.
http://www.ajs.or.at/
doi:10.17713/ajs.v49i5.1186

Modeling and Prediction of the Impact Factor of Journals Using Open-Access Databases

Matthias Templ
Zurich University of Applied Sciences

Abstract

This article is motivated by the work as editor-in-chief of the Austrian Journal of Statistics and contains detailed analyses about the impact of the Austrian Journal of Statistics.

The impact of a journal is typically expressed by journal metrics indicators. One of the important ones, the journal impact factor is calculated from the Web of Science (WoS) database by Clarivate Analytics.

It is known that newly established journals or journals without membership in big publishers often face difficulties to be included, e.g., in the Science Citation Index (SCI) and thus they do not receive a WoS journal impact factor, as it is the case for example, for the Austrian Journal of Statistics.

In this study, a novel approach is pursued modeling and predicting the WoS impact factor of journals using open access or partly open-access databases, like Google Scholar, ResearchGate, and Scopus. I hypothesize a functional linear dependency between citation counts in these databases and the journal impact factor. These functional relationships enable the development of a model that may allow estimating the impact factor for new, small, and independent journals not listed in SCI. However, only good results could be achieved with robust linear regression and well-chosen models.

In addition, this study demonstrates that the WoS impact factor of SCI listed journals can be successfully estimated without using the Web of Science database and therefore the dependency of researchers and institutions to this popular database can be minimized. These results suggest that the statistical model developed here can be well applied to predict the WoS impact factor using alternative open-access databases.

Keywords: bibliometrics, journal impact factor, open-access, statistical modelling.

1. Introduction

The journal impact factor (hereinafter also referred to as JIF) is one of the most well-known indicators calculated from science citation indexed (SCI) journals listed in the journal citation reports (JCR) and calculated from the WoS database. It is a proxy of the relevance of a scientific journal. The higher the JIF, the more often the journal has been cited within a certain time period (typically 2 or 5 years). About 3750 journals are included in the WoS SCI. In addition to the SCI, there is the expanded SCI (SCIE), which includes around 8800 journals such as the Austrian Journal of Statistics. Both SCI and SCIE are part of the Web

of Science database, but the calculation of journal impact factors is done for SCI (and SSCI - Social Sciences Citation Index) listed journals only.

The Web of Science database is currently maintained by Clarivate Analytics and this commercial company calculates the WoS SCI impact factors from it. If a journal is included in WoS SCI is based on decisions of Clarivate Analytics.

There are other databases than the Web of Science available and also appropriate to calculate journal impact factors or similar metrics. The journal impact factor can be for instance calculated from the database of Scopus and Google Scholar. Let's take the example of the Austrian Journal of Statistics, which is listed in SCIE, Scopus, DOAJ, and many other databases. It has currently (accessed 17.01.2019) an SJR SCImago journal metric of 0.422 (for 2018). The SCImago journal metric (SJR indicator) accounts for both the number of citations received by a journal and the importance or prestige of the journals where these citations come from. SCImago (ranking service) is is found and run by research groups from the University of Granada, Extremadura, Carlos III (Madrid), and Alcalá de Henare and retrieve data from Scopus. However, the JIF of the Austrian Journal of Statistics is unknown even the journal is included in the Web of Science database, because the journal is not listed SCI, only in SCIE.

None of the alternatives (such as SJR) has an approximately high degree of awareness like the JIF. Even if - to the best of my knowledge - universities in the Asian and South-East-Asian region focus more and more on the Scopus database and the SJR indicator.

Even the JIF is one of the most established bibliometric indicators of the prestige and influence of a scientific journal, the JIF is controversial, as it is also often used to qualitatively assess the scientific performance of a scientist, for example in appointment procedures. However, the JIF is not a suitable measure of the quality of the research results set out in an article or even for the evaluation of a scientist and his/her scientific performance. The San Francisco Declaration on Research Assessment (DORA) initiative is committed to evaluating and amending the criteria for evaluating research results. It calls on organizations and scientists not to use journal-based metrics - such as the Journal Impact Factor - as a substitute for the quality of individual research articles when making recruitment, promotion, or funding decisions.

The common view in the scientific literature is that the JIF cannot be calculated from alternative sources. Recently, a team of academics spent months (because of the non-existence of an API of Google Scholar) on collecting data about 2.3 million papers from the academic search engine Google Scholar, see Martín-Martín, Costas, van Leeuwen, and López-Cózar (2018a). They found that as long as the data stored in Google Scholar is not made available to the scientific community in an automated process with mass export features, Google Scholar and other alternatives to the current commercial providers cannot be considered a viable option Martín-Martín *et al.* (2018a).

Mongeon and Paul-Hus (2016) compared the coverage of Scopus and Web of Science using descriptive comparisons as well as Meho and Yang (2007) describes differences between JIF and Google scholar in a descriptive manner. Interestingly, Harzing and van der Wal (2009) compared Spearman correlation coefficients from Google Scholar h-indices of journals to WoS impact factors, and López-Cózar and Cabezas-Clavijo (2013) also looked at correlations between SJR, Google Scholar metrics, and the JIF. Martín-Martín, Orduna-Malea, Thelwall, and Delgado López-Cózar (2018b) found out that the Spearman correlation coefficient between citation counts in Google Scholar and Web of Science or Scopus are high (0.78-0.99) and many other articles compare the coverage and correlations between literature databases and/or differences in metrics calculated from different literature databases. However, to the best of my knowledge, nobody thought of using a statistical model to estimate the journal impact factor from alternative databases.

The aim of this article is to introduce a method, which estimates the JIF using Google Scholar, the database from Scopus, and the RG Score from ResearchGate. Using these databases, the functional dependency between these outcomes and the SCI impact factor of SCI-indexed journals calculated by Clarivate Analytics are modelled. In addition, feature engineering

was applied to improve the prediction of the model. I hypothesised that in case a strong relationship is found, open-access databases can be used for estimating the WoS journal impact factor without the need of a restricted and closed-access database. It also allows us to estimate the WoS impact factor of such journals that are not listed in SCI.

Overview In Section 2, the journal impact factor is explained, while the differences between the bibliographic databases and their features are described in Section 3. Section 4 reports the data collection and feature engineering as well as the evaluation metrics. Section 5 include the analysis and model results obtained from citation data that was gathered from Google Scholar, Scopus and ResearchGate. All results are compared with those from WoS impact factors. Especial attention is given to the results obtained about the Austrian Journal of Statistics, which is a non-SCI journal (but indexed in SCIE) and corresponding predictions are discussed in more detail. The models were additionally tested using scientific journals in the field of statistics, food science and sport science. Section 6 concludes and discusses the application of the proposed models in practice. Possibilities for future research are also receives attention.

2. The journal impact factor

The JIF evaluates the frequency of citations, namely, how often articles are cited in other scientific journals within a certain time range. It is thus an indicator of how well the articles perceived and cited by scientists in a specific journal.

The WoS journal impact factor is calculated and published annually in the Journal Citation Reports (Seeger, Kohlen, and Strauch 2004) for SCI indexed journals based using the Web of Science database (maintained by Clarivate Analytics). Basically, it was built on two index databases, the Science Citation Index (it consists of literature from 1900 to the present) and the Social Science Citation Index (it consists of literature from 1956 to the present). These cover the source of scientific literature in the field of natural sciences, medicine and social sciences (Gorraiz 1992). However, the coverage of the different subjects is very different. The sciences and medicine are covered very well, while the coverage of the social sciences and humanities is rather low (Stock and Stock 2003).

2.1. Estimating the WoS journal impact factor

For a journal, the number of citations given for all articles published in the journal in question within the last two or five years is important. The WoS journal impact factor (JIF) is calculated on the basis of the articles published in the past two or five years (Andrade, Gonzaelez-Jonte, and Campanario 2009). For example, the five-year JIF in year 2019 of a journal in year 2019 is calculated on the basis of the published articles in 2014, 2015, 2016, 2017, and 2018.

More precisely, the JIF is calculated based on Equation 1, where the nominator of the two years JIF in a given year is the number of citations received in that year, but only for articles published in that journal during the two preceding years. This number is divided by the total number of *citeable items* published in that journal during the two preceding years:

$$\mathrm{IF}_{tj} = \frac{\mathrm{Citations}_{t-1,j} + \mathrm{Citations}_{t-2,j}}{\mathrm{Publications}_{t-1,j} + \mathrm{Publications}_{t-2,j}} \tag{1}$$

A journal impact factor of, say 1.12 of a journal j in t =2017 means that, on average, its papers published in 2015 and 2016 received roughly 1.12 citations each in 2017.

2.2. Potential bias of the JIF

There are different kinds of potential bias included in the estimation of the JIF from Web of

Science.

The JIF is a measure with a skewed distribution. A few articles generate the most citations, while many other articles are often rarely or never cited. In Section 4.4 we will see that the high impact factor from the Journal of Statistical Software around the year 2017 was only due to thousands of citations of a single article.

Non-English journals and articles are underrepresented in Web of Science (Holmberg 2015).

The JIF is often regarded as a decisive factor and accordingly, journals with a higher JIF are more often cited as a reference in scientific publications.

It is difficult to get indexed in the SCI, especially for new or smaller journals and for journals independent from big publishers. This has at least two consequences. On the one hand, journals that are not listed in SCI, do not receive a WoS/SCI journal impact factor and therefore have less chance of being cited in other scientific journals. Furthermore, these journals are not as attractive for researchers to submit there their work. It happens because the researchers are often evaluated indirectly by the impact factor of their papers, thus by the impact factor of the journal where they have published work. Thus publishing in a journal with a high impact factor improves their own scientific reputation.

Another potential bias arises from the fact that scientific journals often publish reviews and accompanying materials such as editorials, meeting abstracts, technical information and letters (Kaltenborn and Kuhn 2003). Review articles are usually cited more often than the original articles and therefore review articles increases the JIF, although original research papers might often have a higher scientific value.

The JIF underestimates the true impact factor, because it's based only on SCI journals. Citations to books, book chapters, theses, conference papers, and journal articles published in non-ISI journals are not included (Harzing and van der Wal 2009; Meho and Yang 2007). This obstacle related also to different document types such as commentary-type contributions that are counted in the numerator of Formula 1), but they excluded to count as source articles (Simons 2008) in the official SCI journal impact factor. Only original articles, meeting abstracts, technical information and reviews are serving as source (thus counted in the denominator of Formula 1). As a result, the JIF of journals with commentary-type contributions are in advantage over journals without publishing this kind of contributions.

However, these biases should be taken into account when estimating the JIF. Thus, when the estimation of JIF was made using Google Scholar, for example, the commentary-type articles were not excluded as source.

Also note that some scientific journals have a longer reviewing, copy-editing and proof-reading process than others. This can also have an impact on the JIF.

Another source of potential bias comes from the various publishing frequency of an issue per journal. There are journals, which publish issues regularly over the whole year. While other journals may not publish any issue at the end of the year, say from October to December thus those issues are published only in January of the next year. This may also increase the JIF, especially the two-year JIF which has a short observation period.

An article which is cited very often in the upcoming two years after publication raise up the two-years impact factor dramatically, while other measures of importance, such as the h-index is not influenced by such an outlier (cf. Table 2 and Section 4.4)

Note that also errors in titles of articles leads to a bias (see Vanclay 2012, for further details on this).

Many of these potential biases are only specific for Web of Science data, but not for citation data from Google Scholar. A model to estimate the impact factor using robust methods naturally respect this kind of biases in some way by fitting the functional relationship of the impact factor to a bunch of explanatory variables, see Sections 4.4 and 5.

3. Bibliographic databases and features

At least in the field of statistics, the Web of Science, Scopus, Google Scholar (GS) and ResearchGate (RG) are the most relevant bibliographic databases, which contain information about bibliometric values for individuals, journals and institutions. An overview about these databases are given in Table 1. Not all databases make the citation statistics freely available. Typically, access to WoS is subject to a fee, as is Scopus, the latter having 100 free queries. It is not possible to connect to the ResearchGate database with an interface, nor to download the bibliographic information of articles in an automatised manner without extensive web scrapping. Also software like Publish or Perish does not offer the possibility to download the citation information of articles on ResearchGate.

Table 1: Bibliographic databases and their coverage and access. Note that the WoS platform contains even more as WoS core, but also patents, etc. are included.

	n. of journals	from	access
Web of Science (core)	$21{,}177^{(*)}$	1900	paid service
Google Scholar	$\text{largest}^{(+)}$	2004	free, but no API
Scopus	$21{,}950^{(x)}$	2004	partially paid service
ResearchGate	unknown	2008	free, but no API and limited access

$^{(*)}$ figure from July 12, 2019. $^{(+)}$ coverage not known. $^{(x)}$ figure from August, 2017.

In principle, the impact factor can be calculated from all of them, but there are three major differences. The first is related to the coverage of articles, the second is the distinguishment of the kind of articles, and the third is the labeling of the year of citing an arctile.

Note that another big source of bibliographic information is CrossRef. They are a non-profit organization of several publishers, founded in 2000. It facilitates finding, citing, linking, and evaluating research results, but millions of references are still missing (Pentz 2001).

3.1. Web of Science

The Web of Science is a commercial platform that offers a common search language, navigation environment and data structure of all journal articles in the SCI and SCIE (and more). The Web of Science Core Collection includes the Science Citation Index Expanded (SCIE), Emerging Sources Citation Index (ESCI), Social Sciences Citation Index (SSCI) and Arts & Humanities Citation Index (AHCI). The number of items in the Web of Science is about 342,000 pieces of magazines, books, proceedings, patents and data sets, and more than 155 million articles from journals, books and proceedings. The area of bio-medical sciences, natural sciences, engineering, computer science, materials science, social sciences, art and humanities are covered by the journals. The journal literature covers a period from 1900 to the present day, while patents cover a period from 1963 to the present. In the citation analysis, the author's Web of Science citation tracking, quotation counts and h-index (Hirsch 2005) are included.

3.2. Scopus

Scopus is one of the world's largest and most frequently cited databases for scientific journals, literature, books and conference proceedings. It covers research topics from all scientific and technical disciplines. From arts and humanities to medicine and the social sciences. With the tools provided by Scopus, research can be recorded, analysed and visualized. When Elsevier introduced Scopus in 2004, it attracted a lot of attention. On the one hand, this was due to an enormous marketing effort, on the other hand, it was because Scopus is the first major competitor to the Web of Science, especially for citation tracking. Citation tracking

also provides information about other institutions and authors who are doing similar work (Goodman 2005). Scopus generally includes more articles as Web of Science, see e.g. Li, Burnham, Lemley, and Britton (2010); Mongeon and Paul-Hus (2016), but as WoS it is selective. For further readings on this we refer to Aksnes and Sivertsen (2019).

SCImago is a platform containing journals and country-specific scientific indicators, using information from the Scopus database. The platform covers over 34100 titles and more than 5000 international publishers. With this information, journals and country rankings can be compared or analysed. Journals can be grouped in 27 different subject areas, 313 subject categories or in 239 countries. In this contribution, we use the scientific journal ranking (SCImago Journal Rank - SJR). The SJR expresses the average number of weighted citations published per document in the selected year, out of the three previous years in the selected journal. The algorithm to calculate the SJR metric is based on Google PageRank^{TM}. Thus, citations are weighted depending on the citations and prestige of the source where they originate.

3.3. Google Scholar

Google Scholar (GS) is a major academic search engine that provides an easy way to search for articles, reports, books, theses, abstracts and court opinions from academic publishers, professional societies, online repositories, universities and other websites. From the authors' point of view, Google Scholar offers the possibility to graphically display the automatically updated citation metrics for their articles, as well as the calculation of multiple citation metrics. Moreover, everything is kept very simple, no matter how many articles have been written or whether the name is shared by a large number of scientists. The articles of a journal or of an author is ranked by the Google Scholar Rank. The Google Scholar Rank (partially based on Google's PageRank^{TM}) basically ranks articles according to their number of citations, but also other effects like the clicks and discussion of an article are taken into account. Most notable is that these numbers do not well incooperate for the calculation of the two- or five year impact factor, because these time restriction cannot be selected, i.e. only the total number of citations for an article is known. There exist several tools to access Google Scholar by web scrapping, but - as mentioned in the introduction - there is no API available. All in all, this makes it complicated to extract information from Google Scholar in an automatized and large-scale manner (Martín-Martín *et al.* 2018a) and it is nearly impossible to restrict the citations of an article due to a certain time span as needed for the calculation of the JIF. In addition, not only research articles from peer-reviewed and indexed journals contributes to the citation statistics, also many research articles that are not published in a peer-reviewed journal are accounted for in the citation analysis.

Google Scholar has been shown as a rich source of information and it has good coverage of disciplines and languages, also in the Humanities and Social Sciences, where WoS is known to be weak (Chavarro, Ráfols, and Tang 2018; Prins, Costas, van Leeuwen, and Wouters 2016). One the one hand Google Scholar has consistently returned higher numbers of publications and citations as WoS or Scopus (Harzing and Alakangas 2016), but on the other hand citation counts from a range of different sources have been shown to correlate positively with GS citation counts (Martín-Martín *et al.* 2018b). For further discussions of the weaknesses and strength of Google Scholar, see e.g. (Delgado López-Cózar, Orduña-Malea, and Martín-Martín 2018).

3.4. ResearchGate

ResearchGate (RG) is a social networking site for researchers where they can create their own profiles, list their publications, and interact with each other. It was founded in 2008 by the physicians Dr. Ijad Madisch and Dr. Sören Hofmayer and the computer scientist Horst Fickenscher and has today more than 15 million members. It offers specialist circles a new opportunity to disseminate their work and thus change the dynamics of informal science communication. The service thus takes into account the fact that scientists are building

up a personal network. ResearchGate has articles from various disciplines and over several years. However, documents from older years and from certain disciplines, such as the arts and humanities have the potential for expansion. Researcher are motivated to upload the most recent articles to the website to attract a wider audience (Thelwall and Kousha 2016). It is unknown, how exactly the algorithm works to calculate the RG Scores for authors (Kraker and Lex 2015) or the RG journal impact. Several authors analyzed the RG Score that is assigned to authors (Orduna-Malea, Martín-Martín, Thelwall, and Delgado López-Cózar 2017; Copiello and Bonifaci 2018). The RG journal metric is based on average citation counts from work published in this journal. It is calculated using the ResearchGate database.

The coverage of ResearchGate is 100M+ of publications according to the information on the website of ResearchGate. The coverage of journals with RG Scores is questionable. For example, the Austrian Journal of Statistics has no RG Score assigned.

Table 2: h-index and journal metrics (or journal indicators) of selected journals using different data sources in 2018.

	index	Journal of statistical software	Annals of Mathematics	Econometrica
Web of Science	JIF	22.737	4.768	3.750
Scopus/SCImago	JIF	17.569	9.257	17.653
ResearchGate	RG Score	5.18	6.49	4.08

As can be seen in Table 2, the journal metrics fluctuate when using different sources of citation data. One reason for this fluctuation is the different coverages and types of information in the related databases, but also different time spans can play a role.

The Journal of Statistical Software has one specialty related to the impact factor for 2018 caused by a large number of citations of exactly one article. We come back to this issue in Section 4.4, and point out that the impact factor is sensible to outliers. For example, an h-index is not highly influenced by such an outlier, and also the ResearchGate Score seems to be more robust to such an outlier (see Table 2).

4. Data collection, models and methods

Google Scholar works neither with an API, JSON nor a similar programming interface. It is therefore not or only partially possible, to access key figures via the individual journals, except if time-consuming web scraping is applied. This takes months of team-work, see e.g Meho and Yang (2007) who needed about 100 hours of processing time, extraction of data from Scopus consumed 200 hours, and Google Scholar a grueling 3,000 hours, or (Else 2018) (Nature News from April 11, 2018) reports months of data collection and web scrapping related work.

4.1. Data collection

Another solution is to use the software Publish or Perish, which makes it possible to retrieve citation statistics for articles from Google Scholar (restricted by 1000 articles per year and journal), CrossRef and Scopus and extract it to a file. Note that there is also a possiblity to extract citation statistics for articles in Web of Science, but this is not open-access and needs a (paid) licence.

In this work, individual data from articles were downloaded via Publish or Perish (Harzing and Van der Wal 2008) and pre-processed for the needs of this study. We refer to the Publish or Persish website, their manuals and training programs, how to install this software, how to collect information by point-and-click. All articles from 32 journals has been investigated to calculate a journal impact factor for each journal from all articles citing each other.

The focus was the linear relationship of the two-year WoS journal impact factor of statistical journals with the impact factors estimated from other sources (Google Scholar, Scopus) and from other sources (than Publish or Perish) like ResearchGate Scores. There is no interface and option to automatically scrap ResearchGate Scores per journal and automated data extraction from ResearchGate is not allowed under official terms of service. The RG Scores were manually extracted from the ResearchGate website for all selected journals.

Note that from the data extracted with Google Scholar and Scopus through Publish or Perish, it is only known how often an article has been cited from the publication date to the present day, but the year of the citation is not known. For this reason the total number of citations was divided by the number of years since the paper was published. In the case of the journal impact factor using the Web of Science database, outliers always only influence the values for the respective years. However, when using Google Scholar, all subsequent years are affected when calculating the journal impact factor. We come back to the problem of outliers in Section 4.4 and 5 and consequently propose to use robust methods to control the influence of outliers.

4.2. Feature engineering

To model and predict the journal impact factor from Clarivate Analytics, not only the above mentioned databases were used, but also new features were generated as described below.

It was hypothised that if additional features are included in the model, those improve the model estimations. The following features were investigated and simple descriptive statistics on them are visualized.

Size of the journal: It was assumed that the absolute number of published articles in a journal influence the model. In order to distinguish between the size of the journal, the median number of articles published were calculated from the statistical journals that are listed in SCI. Afterwards, it was possible to distinguish between small to median-sized and median to large-sized journals. The descriptive analysis (Figure 1) highlights that the size of a journal may not have significant influence on the JIF, however, this is the outcome of a bivariate analysis and it give just a first expression when comparing the boxes between different databases. It does not consider possible interactions with other explanatory variables in a larger model.

Physical address of the journal: Another feature to consider when calculating the JIF is the origin of the journal. The following question was tested through descriptive analysis: is the relationship of the JIF calculated from the Web of Science differ from the JIF that was calculated from alternative databases when the origin of the journal is in Europe or in the USA? In other words, is the coverage different for different databases for Europe and USA? The descriptive analysis (Figure 2) shows that the journal impact factor calculated by Clarivate Analytics using the Web of Science database, may be higher for US-based journals compared to European journals. This pattern is also recognizable when the impact factors are from other sources (eg. Scopus and ResearchGate) estimated. On the other hand, this pattern is not as high as related to the first (Web of Science) or the first two (Web of Science and Google Scholar) groups. This is a surprising result whereby we can only guess about the reasons.

Kind of the journal: The journals were categorized into three groups based on the kind of research that is published in the journal. The first group contained those journals that publish theoretical and mathematical oriented articles, while the second group contained journals that are focused on applied research articles. A third group was also created that publish both applied and theoretical research. The partition of journals into these three

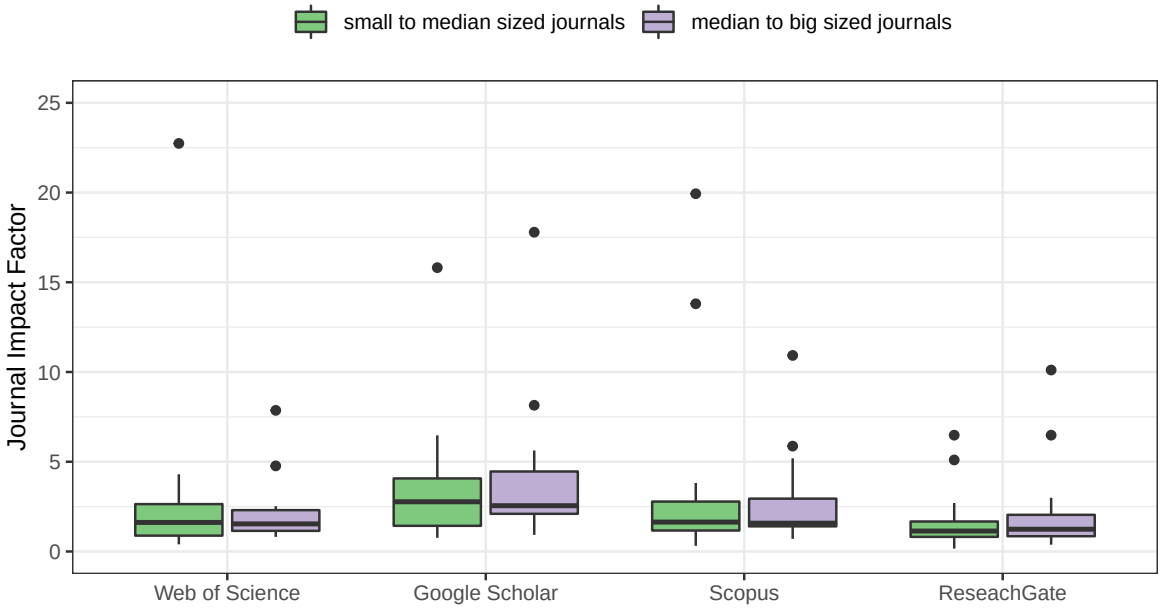

Figure 1: Journal impact factor of statistical journals calculated from alternative databases (Web of Science, Google Scholar, Scopus und ResearchGate) and their relation to the size of the journal. The colour of the boxplot corresponds to the size (green: small to median; lilac: median to big).

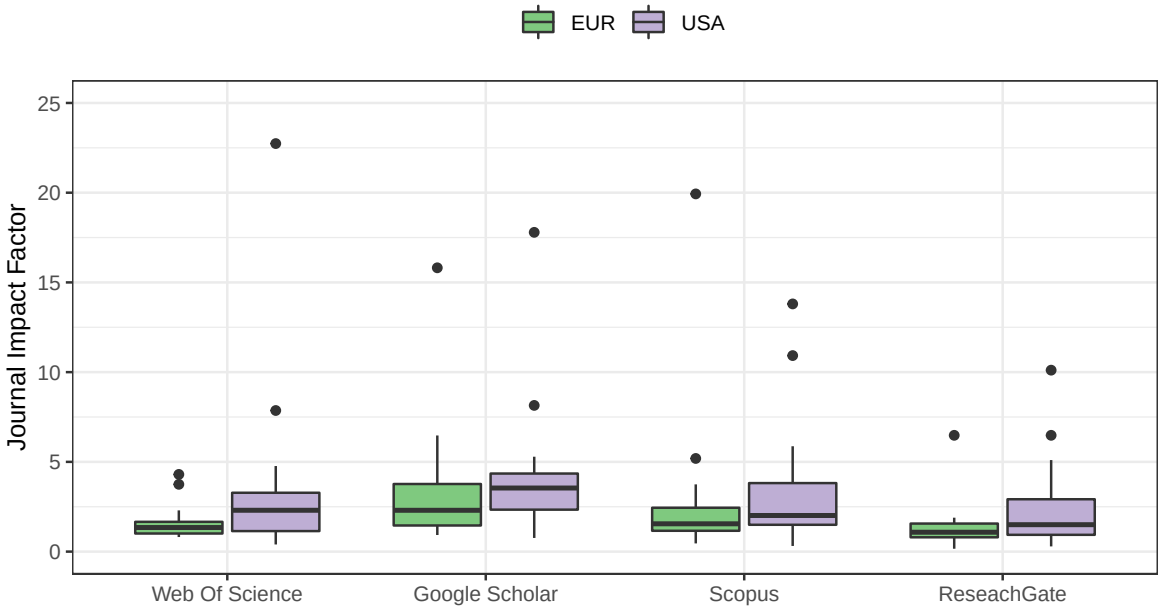

Figure 2: Journal impact factor of statistical journals calculated from alternative databases (Web of Science, Google Scholar, Scimago und ResearchGate) and their relation to the origin of the journal. The colour of the boxplot corresponds to the origin (green: Europe; lilac: USA).

groups was based on expert rating. The following question was tested through descriptive analysis: is the relationship of the JIF calculated from the Web of Science differ from the JIF that was calculated from alternative databases when the kind of the journal is theoretical, applied, or mixed type?

The descriptive analysis of this feature resulted in a non-significant difference between the kind of research published in the journal (boxplot not shown). However, the JIF was slightly lower for applied journals compared to the theoretical and mixed journals, independently from the source of the database that was used for the calculation of the journal impact factor.

Further features: When further features are taking into account as explanatory variable in the model, these have to be carefully selected due to potential multicollinearity. Such features are for example, h-indices or similar indices calculated from Google Scholar or other alternative sources.

Several other kind of engineered features were experimentally tested using text mining to differentiate between the amount of potential keywords and the style of the articles. No useful features could be extracted by using this approach.

4.3. Methods

Ordinary least-squares regression is one of the simplest methods to model the functional linear relationship between the WoS JIF and the predictors, namely the JIF estimated with other data sources, ResearchGate Scores and further engineered features.

With $\mathbf{y} = (y_1, y_2, ..., y_n)^T$ and $\mathbf{X}$ the design matrix of predictors with n observations and p variables, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. The model assumptions state that the error term $symbol\epsilon = (\epsilon_1, \epsilon_2, .., \epsilon_n$ is normally distributed with equal variance of errors. The residuals are given by $\hat{\boldsymbol{\epsilon}} = r_i(\boldsymbol{\beta}) = y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_q x_{iq})$. For ordinary least-squares regression, the coefficients $\boldsymbol{\beta}$ are estimated by minimizing the sum of squared residuals, $\hat{\boldsymbol{\beta}}_{OLS} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} r_i^2(\boldsymbol{\beta})$. The estimated coefficient can then be used to predict $\mathbf{y}$ given $\mathbf{X}$ by $\hat{\mathbf{y}} = \hat{\boldsymbol{\beta}}_{OLS}\mathbf{X}$. For the formulas to estimate confidence and predition intervals we refer to any textbook on linear regression. Ordinary least-squares regression serves only as a benchmark because of it's wide use in science.

Robust MM-regression: For an robust estimate of the regression coefficients, the residual squares r_i^2 are replaced by another function of the residuals $\rho(r_i(\boldsymbol{\beta}))$, where ρ is a symmetric function of the residuals with minimum at 0. This leads to a minimization problem $\boldsymbol{\beta}_M = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(r_i(\boldsymbol{\beta}))$ and by differentiating to β we get $\frac{\delta}{\delta\beta} [\sum_{i=1}^{n} \rho(r_i(\boldsymbol{\beta}))] = -\sum_{i=1}^{n} \psi(r_i(\boldsymbol{\beta})) \cdot \mathbf{x}_i = 0$ with $\psi = \rho'$ and $\mathbf{x}_i$ as i-th observation. To account for scale changes in $\mathbf{y}$, the residuals are scaled so that $\sum_{i=1}^{n} \psi(r_i(\boldsymbol{\beta})/\hat{\sigma}(\boldsymbol{\beta})) \cdot \mathbf{x}_i = 0$. Scaling must be estimated simulatively, and should be estimated robustly, e.g. with the median of absolute value $\hat{\sigma}_{MAV} = \frac{med_i(|r_i|)}{0.6745}$. The resulting estimator is called S-estimator that has maximum break point (Maronna, Martin, and Yohai 2006a), but this estimator is inefficient. The MM-estimator uses the solution of the S-estimator as initial value ($\hat{\sigma}_S$ and $\boldsymbol{\beta}_S$), and then using an M-estimate with so-called redescending ψ function. This estimator is the most efficient and robust estimator known in the literature. As redescending function we used the Tukeys biweight function (Tukey 1960). For more information also on confidence and prediction intervals, we refer to Maronna *et al.* (2006a). Robust MM-regression is the main method used in the following.

Generalized additive models: The model has the form $y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i \quad , \quad i = 1, \ldots, n$ observations and p variables and $\text{E}(\epsilon_i) = 0$. Using a kubic natural b-spline smoothing function $f(\mathbf{x})$, $y_i = f(\mathbf{x}_i) + \epsilon_i \quad , \quad i = 1, \ldots, n$. g is obtained through minimization of

$||\mathbf{y} - \mathbf{f}||^2 + \lambda J_{md}(f)$. J_{md} is a function that measures the smoothness of f, and λ is again a smoothness parameter that controls the trade-off between smoothness and data fit of f. For the solution of this minimization problem (by rang-k approximations) we refer to the theory of thin-plate regression splines in (Wood 2003, 2006). Generalized additive models serves as a benchmark for non-linear fits in the following.

Artificial deep neural networks (ANN): A neural network is just a non-linear statistical model Hastie, Tibshirani, and Friedman (2009), which is based on weighted linear combinations of the sample values. Within an ANN, the aim is to find millions of weights to get the best possible output from input data and multiple layers. The weights are updated iteratively. In the first iteration, the weights are random and the result is accordingly bad. For each iteration, we then go to "direction optimum", whereby a gradient method is used for this. It requires backpropagation with an optimizer and a loss function. The quality of the predictions is evaluated based on a selected metric on validation and training data. After tuning the parameters the following parameter setting turned out to be optimal. For the optimizer Adam (Kingma and Ba 2014) was used with activation reLU (He, Zhang, Ren, and Sun 2015). For the loss function the mean squared error was used and for the evaluation of the predictions the mean absolute error. Ten layers were chosen, the first layer with 1000 neurons, the second layer with 900, up to the last hidden layer with 100 neurons. The output's layers activation was selected to be linear. A drop-out of 10% was selected in each layer to avoid overfitting. For the number of epochs 500 was chosen with a stopping criterion of 50 (if after 50 epochs no improvement: stopp). The ANN is used as a benchmark in the paper as a fully automated non-linear method that does not require modelling.

4.4. Comments on model fitting

Figure 3 shows the journal impact factor that was estimated from the Web of Science database in compare to the predicitions from the Google Scholar database. One feature to note on this Figure, that outliers were present in the databases. For example, the Journal of Statistical Software is a huge outlier that influences the least-squares fit (regression line in red) and also a generalized non-robust additive model (blue line). The robust method using an MM-estimator (line in magenta) (Maronna, Martin, and Yohai 2006b) is not influenced from this outlier. This has a huge impact not only on the R^2 (0.14 for the least-squares fit compared to 0.88 from the MM-fit), but also on the coefficients and standard errors. The Figure 3 also suggests that it is questionable to fit the model at original scale. In all models, we tested different transformations of the variables, such as log-transformation and square-root transformation, but the model results did not improve.

The huge outlier in the Web of Science can be explained by the article "*Fitting Linear Mixed-Effects Models Using lme4*" that was written by Bates, Mächler, Bolker, and Walker (2015) and published in the Journal of Statistical Software. This article was cited over 2700 times in the Web of Science. The data fetching from Google Scholar is restricted to 1000 results for each IP address and thus the estimated value (using the Google Scholar model (see Section 5.1) - 5,288 - represents about a quarter of the value of the Web of Science - 22,737. One approach could be to start fetching the data sets from different IP addresses including several weeks of work (see, Martín-Martín *et al.* 2018a). Another approach is to down-weight those outliers by using robust methods in order to bound the influence of such an outlier. This allows to estimate a model that works well for the majority of journals and outliers can be flagged and further be analysed.

Generally, it could be found that all residual diagnostic plots from any of the fitted models (in Table 3 and 4) looks perfect for MM-regression (results are not shown but available upon request). The residuals of the majority of the observations are approx. normal, the variances are approx. equal for the whole range of fitted values and no particular pattern is visible in the Tukey Ascombe plot. Huge outliers are visible, such as the Journal of Statistical Software,

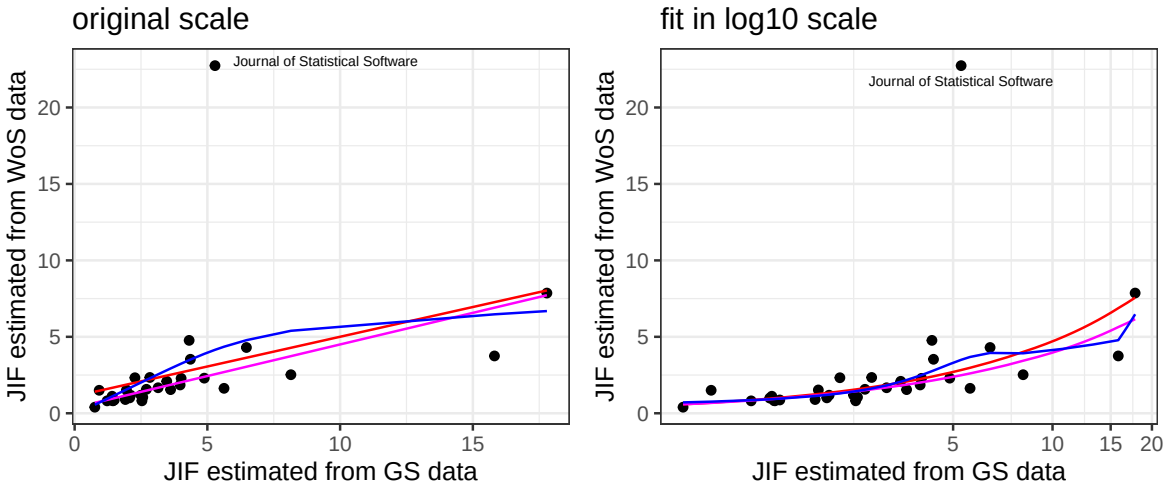

Figure 3: Journal Impact Factor obtained from Web of Science (WoS) and Google Scholar (GS) data. In the right plot the fits are done in log10 scale. The red line corresponds to the linear model using the least-squares method, the magenta one is an MM-estimation and the blue corresponds to a generalized additive model (gam) with thin-plate regression splines.

but these outliers are down-weighted by the robust MM-regression. For example, the Journal of Statistical Software received a weight of 0.

It should be also noted that no serious violation regarding multicolinearity was found.

4.5. Evaluation measures

Measures to determine the accuracy of predictions that are often used in practice are the (root) mean squared error, the mean absolute error or relative mean absolute error. For robust estimation, these measures are not suitable, because outliers - even they have bounded influence on the estimated values - have dramatic impact on these measures. Thus the following robust-adequate measure of prediction error (rMedAPE) was used:

$$\text{rMedAPE} = median\,|(\mathbf{y} - \hat{\mathbf{y}})/\mathbf{y}| \quad , \tag{2}$$

whereby $\mathbf{y} = (y_1, y_2, ..., y_n)^T$ and $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)^T$ represents the observed n values of the response variable of a model and its estimates from the model, respectively. It reports the relative median distance of the predicted values to its original values. Note that in Figure 6 the MedAPE is used, thus the absolute median prediction error without the normalization by $\frac{1}{\mathbf{y}}$.

In this study, the rMedAPE was estimated using cross-validation with 5 folds and 1000 repetitions. It represents an out-of-sample error, because using cross-validation, rMedAPE is only calculated for the corresponding test data sets.

For all models, the model assumptions were carefully checked if they are fulfilled. This includes the normality of the residuals, the homogeneity of variance of the residuals and outlier diagnostics. These results are not presented in this article, but available upon request.

The R^2 was used to check the overall model quality. The R^2 measures the proportion of the variance in $\mathbf{y}$ that is predictable from the independent (predictor) variable(s), i.e. the explained variability of the response variable. It is thus a measure of the overall fit of the model and it is normalized to $[0, 1]$, whereby the higher the R^2 the better the fit.

5. Results

As mentioned in Section 4.4, no model assuption violations have been observed for any robust fit of any model. The models and results of the models are discussed in the following.

The main focus of this paper was to present results about the statistics journals (Sections 5.1 and 5.2). However, other areas of research papers are discussed as well in Section 5.3.

5.1. Results for SCI-listed journals statistics journals

Table 3 shows different models to estimate the WoS impact factor using different sources and predictors. The first model, for example, estimates the WoS journal impact factor using Google Scholar citation statistics by JIF_WoS $1.11 + 0.39 * JIF\_GS$. This model has an R^2 of 0.139 and in average (median) the prediction is 58.8% off the true value. Note that for the artificial deep neural network all information was used, i.e. estimated JIFs from Google Scholar, Scopus, ResearchGate Scores and all features that were engineered (kind of journal, size of journal, origin of journal).

Model	$\text{JIF}_{WoS} \sim$	β_0	β_1	R^2	RMedAE
1	JIF_{GS} (least-squares)	1.11	0.39	0.139	58.8%
2	JIF_{GS} (gam)			0.160	72.4%
3	JIF_{GS} (MM)	0.37	0.41	0.881	28.0%
4	JIF_{Scopus} (MM)	0.73	0.36	0.710	28.1%
5	$\text{JIF}_{ResearchGate}$ (MM)	0.57	0.67	0.870	34.1%
6	JIF_{GS} + JIF_{Scopus} + RG (MM)			0.889	39.5%
7	JIF_{GS}*size + JIF_{Scopus} (MM)			0.93	**27.2**%
8	JIF_{GS} + JIF_{Scopus} + RG+$kind$ + $origin + size\,(MM)$			0.876	41.0%
9	JIF_{GS}*size + JIF_{Scopus} + RG+$kind + origin\,(MM)$			**0.946**	39.6%
10	ANN**				48.89%

$*$ interaction between JIF_{GS} and size of a journal. $**$ Using all available information.

Recommended model

From Table 3 it can be observed that the Google Scholar based simple model (model 3), which estimates the WoS journal impact factor, is favorable when the estimation method is robust (MM-regression). In comparison, the least-squares fit (model 1) and the generalized additive model (with thin-plate splines, model 2) provide the worst results because outliers have a big influence on them. When using more sophisticated models, the R^2 can be increased, but the relative median absolute prediction error (rMedAPE) may raise up. Note that all models referred in Table 3 provide nice diagnostic plots (not shown here), except the least-squares model and the generalized additive model using thin-plate regression splines (gam). Thus, only the robust fitted models fulfill the model assumptions. Also note that transformations of the target variable and/or explanatory variables do not improve the results. Also, note that in the following we prefer the simple model (third model reported in Table 3), because the quality is almost as good as any competing models, and in practice, the data collection and data processing is much easier than for any other model. For example, for the last model

not only data from Google Scholar must be collected and pre-processed, but also data from Scopus, ResearchGate and each journal to predict must be also classified on origin (the US versus Europe) and kind (theoretical, applied, or a mix of both).

When estimating the journal impact factor obtained from Web of Science using the third model - the simple model that estimates the WoS impact factor from Google Scholar - the journal impact factor of a journal obtained from Web of Science can be estimated with about 28% accuracy in average just by using (limited) open-access data from Google Scholar. 94.6% of the variation of the journal impact factor calculated with the Web of Science is explained by the last model (model 9) in Table 3.

For the journal Statistical Science (SS), for example, the number of citations for reference year 2017 is around 4.767 in Google Scholar. The estimated journal impact factor of Web of Science using the recommended Google Scholar model 3 of Table 3 is then $\widehat{JIF}_{SS} = 0.37 + 0.41 \cdot 5.169 = 2.271$ (confidence interval [2.25, 2.76]), while the journal impact factor from Web of Science is $JIF_{SS} = 2.324$.

Overrepresentation issues

The overrepresentation of Google Scholar citation statistics is implicitly expressed by the regression coefficients. This overrepresentation is not easy to interprete, because it represents those citations in Google Scholar that are not within the two-years period for the two-years journal impact factor. A second factor is that Google Scholar citation statistics includes citations from non-SCI journals, i.e. citations of research reports and similar contributions are counted as citation, while they are not accounted for using the definition of the Web of Science impact factor.

Thus we interprete it as how much citations are overrepresented in Google Scholar when estimating the journal impact factor with Google Scholar data.

The recommended model reports that if the journal impact estimated with Google Scholar data increases one unit, then the journal impact factor from Web of Science increases by 0.41.

In average this means that the overrepresentation of Google Scholar can be approx. estimated by a weighted means for $i = 1, ..., n$ journals by $\frac{\sum_{i=1}^{n} w_i JIF\_GS_i}{\sum_{i=1}^{n} w_i JIF\_WoS_i} = 1.897$, with w_i the robustness weights from the MM-regression model estimates.

This means that in average 1.897 times more citations are reported by Google Scholar then in Web of Science for articles in 2017. The models introduced in this article take this into account.

Further interpretations

One aim was to fit a model to estimate the impact factor from Web of Science and reported by Clarivate Analytics. An increase of one unit of the impact factor estimated by Google Scholar, increases the journal impact factor of Web of Science by 0.41 (model 3 of Table 3). The point estimate of slope for predicting the WoS journal impact factor with a model using Scopus data only (the fourth model in Table 3) is only slightly lower, but the confidence intervals overlap ([0.37; 0.46] versus [0.31; 0.42]). Thus we cannot observe a difference between the *coverage* of Google Scholar and Scopus (extracted through Publish or Perish). The parameter from ResearchGate expresses not the journal impact factor measured with Equation 1, because it represents a ResearchGate statistics for the journals. If the ResearchGate value increase one unit, in average the journal impact factor increases by 0.67.

It should be mentioned that also for the largest model, the explanatory variables are (mostly) significant. We skip the interpretation of the larger models, because of correlated explanatory variables this is almost impossible.

5.2. Prediction of the JIF of non-SCI covered journals

Using the model fits resulting from the SCI-indexed journals (see the appendix for a full list), the aim is also to predict non-SCI journals, for example, SCIE journals which are fully covered in Web of Science, but where Clarivate Analytics is not reporting an impact factor.

In case - and this we cannot proof - the non-SCI covered journals behave similar to the SCI covered journals in terms of impact factors estimated by Google Scholar or Scopus data and impact factor calculated through Web of Science, the model would predict the impact factor of any non SCI covered journal without any bias. Because the Web of Science journal impact factor is not known for the non-SCI covered journals, we cannot give any figure about a potential bias in our estimates.

The quality of the estimation as a mean squared error include the variance and squared bias. The average variance of the predictions is already expressed by the rMedAPE. In addition, we also provide confidence intervals for any journal.

Illustrative example

This article was motivated by the daily business as the editor-in-chief of the Austrian Journal of Statistics. Said that, this results can be obtained for any other non SCI-indexed journal in a straight-forward manner.

The Austrian Journal of Statistics is included in all major bibliometrics databases. Figure 4 shows results from extracting Google Scholar data. The number of articles published in each year can be seen on the left upper part. The peaks can be explained by the publication of special issues with a lot of articles. The peaks for the cites per year and cites per paper and its relative measures per author and article are explainable by the citation statistics of a few articles. For example, the peak in 2013 is related to the article Taheri (2013), which is cited 158 times. We rank all articles according to their number of citations. Naturally, the year of publication of an article plays a central role, thus the Google Scholar Rank of articles is lower for articles published in the very past and raise constantly over time in case all articles are cited with the same number each year. We can see that the citation statistics was low until about 2004, because in average the articles published before 2004 has a high mean Google Scholar Rank. After 2004 we see the natural increase of the mean Google Scholar Rank.

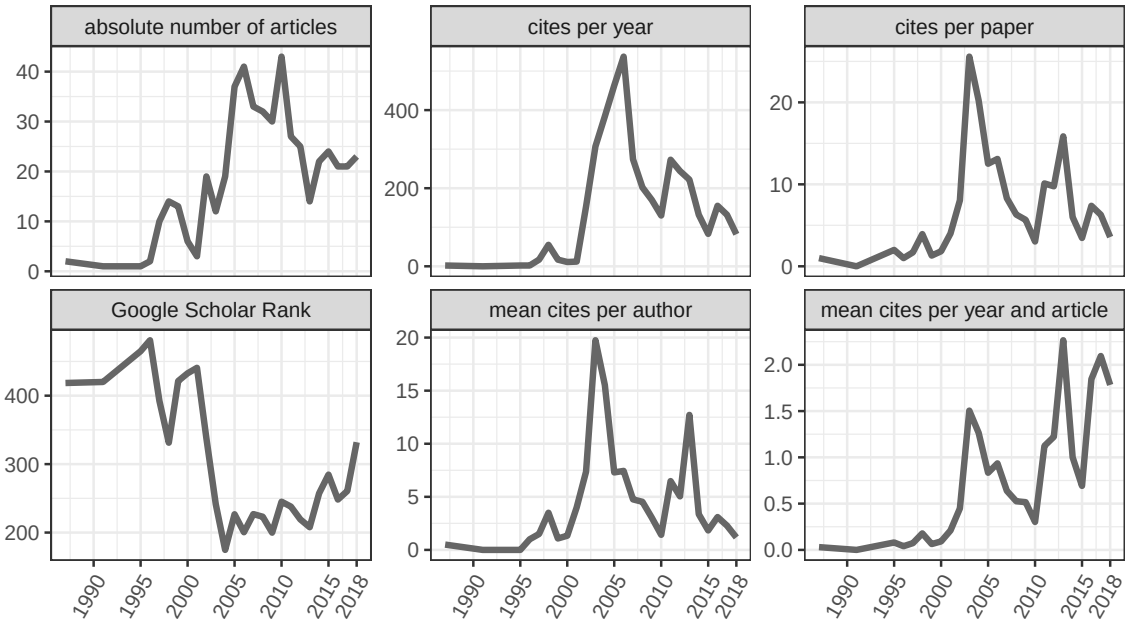

Figure 4: Estimates using extracted citation data from Google Scholar.

The same statistics (except the Google Scholar Rank) is visualized in Figure 5 for Scopus

data.

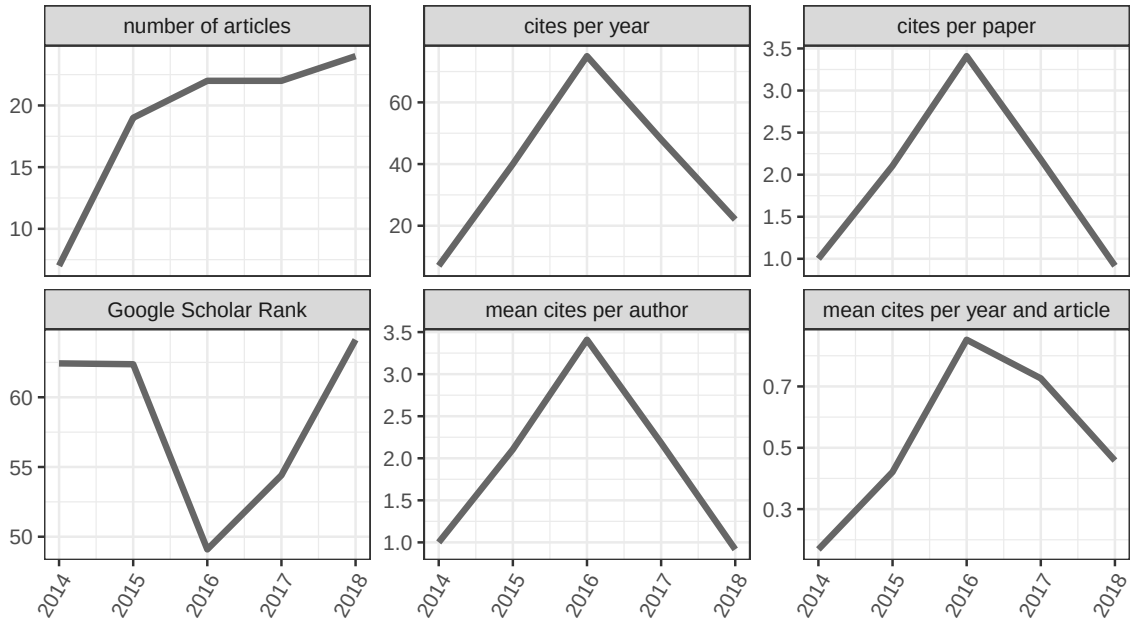

Figure 5: Estimates using extracted citation data from Scopus.

A peak in 2016 is visible and it can be explained with the high number of citations (28, through Scopus) of the article Barcelo-Vidal and Martín-Fernández (2016), which has more than twice number of citations compared to any other article.

The estimated WoS SCI impact factor using Google Scholar and Scopus citation data is shown in Figure 6. The left hand side graphics shows the 90% confidence intervals (boarder of the shaded area) and the point estimates as middle black line. Note that these boarders does not represent prediction intervals, because we assume that the individual variation is not relevant here, but only the uncertainty of the model fit counts. This can be motivated by the fact that the values for the average citations per year obtained from the universe of Google Scholar is fixed (population statistics) and not a random variable, and thus only the model uncertainty counts. The right hand side graphics shows the median error estimated by the cross-validation procedure. The point estimates of Google Scholar model estimates of the WoS impact factor for 2018 is slightly larger than for the Scopus model, but a ranking is not possible, because the intervals overlaps. Actually, the WoS impact factor of the Austrian Journal of Statistics in year 2018 is estimated with 1.105 and the 90% confidence interval is [0.92; 1.29] using the Google Scholar model, and 0.90 ([0.72; 1.08]) using the Scopus model.

5.3. Results for other areas of research

Journals from two other areas, namely a selection of journals in food science & technology and sport science were also inspected. Since the main part of the work is the evaluation of the statistical journals, only the model quality related to the other two areas is presented in Table 4. It can be observed that the model results are better than for the statistics journals. In fact the relative median absolute error of the best model is only 10.6 % for food science & technology journals, and 14.6% for sport science. The R^2 is very high (0.957 and 0.911), thus the variation of the journal impact factor calculated by Clarivate Analytics using the Web of Science database is almost fully explainable with the statistical models using Google Scholar, Scopus and ResearchGate citation data.

While the simple Google Scholar model works best for journals in food science & technology, more complex models works better than simple models in case of journals from sport science.

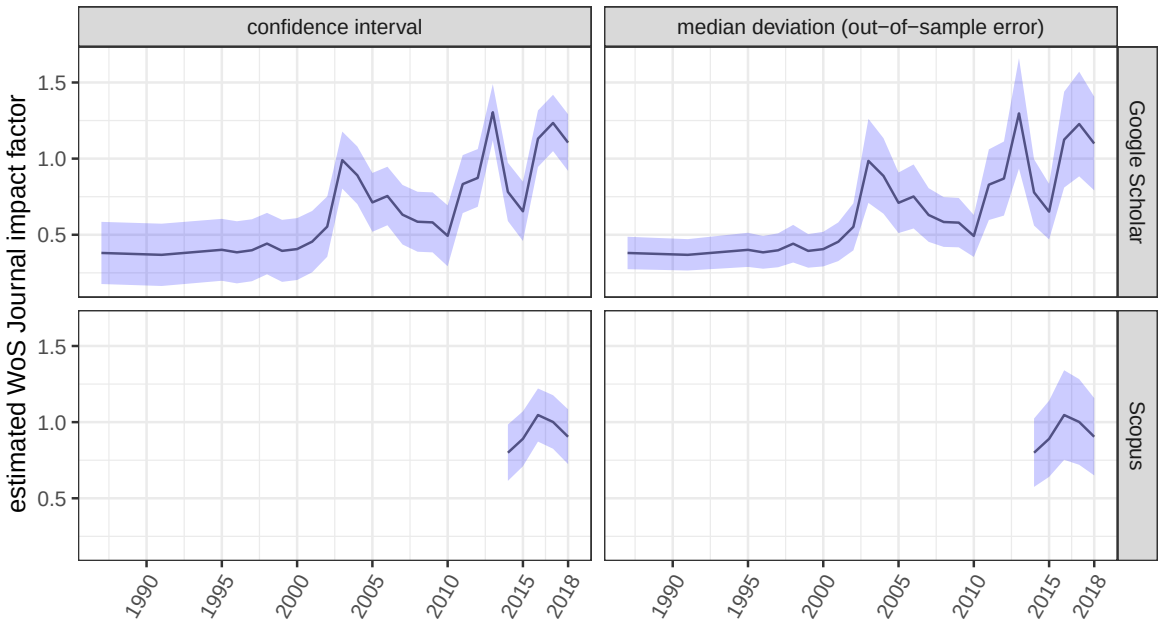

Figure 6: Fitted SCI Web of Science journal impact factor for the Austrian Journal of Statistics using our models for Google Scholar (model 3) and Scopus data (model 4) and the related uncertainties of the estimates, expressed as 90 percent confidence intervals (left) and a cross validated median prediction error, MedAPE (right).

Table 4: Results from classical least-squares, generalized additive models, and robust MM-regression of different models for the food science & technology journals and sports journals. The R^2 and the relative median absolute error (rMedAPE) of different models and methods are reported. The results are based on 5-fold cross-validation with 1000 replications.

Models for food science & technology	R^2	rMedAPE
$\text{JIF}_{WoS} \sim \text{JIF}_{GS}$ (least-squares)	0.843	16.4 %
$\text{JIF}_{WoS} \sim \text{JIF}_{GS}$ (gam)	0.891	49.2 %
$\text{JIF}_{WoS} \sim \text{JIF}_{GS}$ (MM)	0.950	**10.6** %
$\text{JIF}_{WoS} \sim \text{JIF}_{Scopus}$ (MM)	0.768	25.1 %
$\text{JIF}_{WoS} \sim \text{JIF}_{ResearchGate}$ (MM)	0.911	18.1%
$\text{JIF}_{WoS} \sim \text{JIF}_{GS} + \text{JIF}_{Scopus} + RG$ (MM)	**0.957**	12.9 %
$\text{JIF}_{WoS} \sim \text{JIF}_{GS}$*size + JIF_{Scopus} + RG(MM)	0.889	14.7 %
ANN**		25.2 %
Sport science	R^2	rMedAPE
$\text{JIF}_{WoS} \sim \text{JIF}_{GS}$ (least-squares)	0.025	32.6%
$\text{JIF}_{WoS} \sim \text{JIF}_{GS}$ (gam)	0.738	30.7%
$\text{JIF}_{WoS} \sim \text{JIF}_{GS}$ (MM)	0.558	27.0%
$\text{JIF}_{WoS} \sim \text{JIF}_{Scopus}$ (MM)	0.835	15.3%
$\text{JIF}_{WoS} \sim \text{JIF}_{ResearchGate}$ (MM)	0.187	22.4%
$\text{JIF}_{WoS} \sim \text{JIF}_{GS} + \text{JIF}_{Scopus} + RG$ (MM)	0.861	18.5%
$\text{JIF}_{WoS} \sim \text{JIF}_{GS} + \text{JIF}_{Scopus}$ + RG + size (MM)	**0.911**	**14.6**%
ANN**		44.9 %

$*$ interaction between JIF_{GS} and size of the journal. $**$ Using all available information.

6. Conclusion and discussion

While many journal metrics alternative to JIF (like Google Scholar Metrics, SNIP, SJR, ...) are becoming more and more important and may have some advantages regarding coverage, access and methodical aspects for the calculation of a journal metric, the WoS SCI journal impact factor is still one of the most important proxy to measure the quality of a journal. Although the following is strongly criticized by the San Francisco Declaration on Research Assessment, for example, (unfortunately) the JIF of a journal in practice also plays a role in the assessment of authors. The JIF is not only important for journal editors and authors looking for journals to publish their work, but also universities often rely on the JIF of journals to assess their researchers whether PhD or habilitation candidates have enough papers in journals with (relatively) high JIF, how much research an institute has done in terms of publications and related JIF of journals, etc. They often look not only on citation statistics of an article, but where the article is published and use the corresponding journal metrics for their evaluation.

The Web of Science database, maintained by Clarivate Analytics is usually used to estimate the JIF. Fortunately, Clarivate Analytics keeps care of the consistency and coverage of Web of Science and carefully estimates the journal impact factor *in-house*. However and unfortunately, this database is restricted and has closed-access and therefore it is impossible to re-estimate the journal impact factor *out-of-house* for any journal. Furthermore, journals has not even a WoS impact factor, which are listed among the SCIE or non-SCI(E) journals. This is disadvantageous and unfortunate for new journals or for independent journals that has difficulties to receive a SCI listing. If these journals would be listed as SCI journals, more authors would send their higher quality articles to publish in these journals, thus more citations would happen, which would naturally increase the JIF. It is therefore of interest for many scientists to calculate a journal impact factor from alternative sources to the Web of Science.

So far, many articles are focused on coverage comparisons (Gorraiz 1992; Stock and Stock 2003; Holmberg 2015; Mongeon and Paul-Hus 2016; Meho and Yang 2007; Van Eck, Waltman, Lariviere, and Sugimoto 2018) or correlations Harzing and van der Wal (2009); López-Cózar and Cabezas-Clavijo (2013); Martín-Martín *et al.* (2018b) between citation statistics based on different data sources. The general view is that the Google Scholar database for instance can not be used, because even the correlations are high, the non-existent programming interface is not a valualbe source of information, and it is too time and man-power intensive to scrap all citation data needed for the calculations Else (2018); Martín-Martín *et al.* (2018a).

The aim of this study was to estimated the WoS journal impact factor using alternative data sources based on the functional linear dependency between these sources and the WoS journal impact factor. Citation data were accessed from the following alternative databases: Google Scholar, ResearchGate, Scopus and CrossRef through Publish or Perish (Harzing and Van der Wal 2008). It was showed that the least-squares model, and the general additive model performs not well in predicting the impact factor because of the limitations arise from the number of accessible articles per year and per journal through Publish or Perish. The fits and predictions with ordinary least-squares regression, generalized additive models, and artificial deep neural networks are strongly influenced by outlier journals, and thus these methods havn't performed well. It was demonstrated that the JIF can be really well estimated using robust regression methods and predictors based on alternative citation data sources and variables obtained by feature engineering. However, the information gained with feature engineering showed that the size of a journal as well as the orientation (application-oriented to very theoretical) and origin of a journal did not provide a significant additional level of explanation and thus the vague conclusion can be drawn that Google Scholar, Scopus, and WoS have proportionally good coverage for these different journal types. The prediction of robust models performed well (high explained variance) except those journals, which publish more than 1000 articles per year (as it was shown in the case of the Journal of Statistical

Software). It was also found that the out-of-sample error was low. Therefore the dependency on the Web of Science database could be heavily reduced because the impact factors can be estimated well without access to the restricted Web of Science database. It could lead universities and the research community to accept the presented modeling approach as well as artificial deep neural networks for impact factor estimation even it is connected with small prediction uncertainties.

Note that different models are optimal for journals in different research areas, although the very simple model with the information from Google Scholar only did well for both research areas. The differences may not be great, but by fine-tuning a model, the prediction errors might be reduced slightly. As an alternative, we showed the use of deep artificial neural networks. Here the choice of model is not important, but the results are worse than with the robust regression using MM estimators. The reason is that ANN's are very sensitive to outliers.

Moreover, the introduced model allows estimating the WoS impact factor for not only those journals, which are SCIE indexed, but also for those that are not SCIE nor SCI listed. This was done for the case of the Austrian Journal of Statistics, which is an SCIE indexed journal, but Clarivate Analytics does not calculate its impact factor. These calculations using the introduced approach can be repeated for any other journal in a straight-forward manner.

Note that the models formulated were not intended to make predictions about future JIFs of a journal. This was also not the goal of this paper. This would require a much more extensive data collection to collect data for, e.g., the last 15 years and to make advanced and different modelling.

Future work may include advanced text mining approaches to extract additional features about the journals. The computational costs and efforts of this may be very high and it is questionable if it results in a different functional linear dependency between the citation statistics gathered from alternative databases and the impact factor. However, first attempts (results not shown) did not show any significant effects when for example the counts of different key words per an article was used.

Appendix: List of journals investigated

Data from the following SCI listed journals were extracted, processed and analysed to fit the models.

Statistics (32) Annals of Applied Statistics (AoAS), Annals of Mathematics, Annals of Statistics (AoS), Bayesian Analysis (BA), Biometrical Journal (BJ), Biometrics (Biometrics), Biometrika (Biometrika), Biostatistics (Biostat), Computational Statistics Data Analysis (CSDA), Computational Statistics (CS), Econometrica (Eco), Electronic Journal of Statistics (EJoS), Journal of Econometrics, Journal of Multivariate Analysis (JoMA), Journal of Official Statistics, Journal of Statistical Computation and Simulation, Journal of Statistical Planning and Inference (JoSPaI), Journal of Statistical Software, Journal of the American Statistical Association, Pharmaceutical Statistics, Psychometrika, Quarterly Journal of Economics, Scandinavian Journal of Statistics, Statistical Applications in Genetics and Molecular Biology, Statistical Methods in Medical Research, Statistical Science, Statistics and Computing (SaC), Statistics and its Interface, Stochastic Processes and Their Applications, Structural Equation Modeling, Technometrics, The American Statistician

Food Science & Technology (32) American Journal of Enology and Viticulture (AJoEaV), Annual Review of Food Science and Technology (ARoFSaT), Bioscience Biotechnology and Biochemistry (BBaB), Biotechnology Progress (BP), Cereal Chemistry (CC), Cereal Foods World (CFW), Chemical Senses (CheS), Critical Reviews in Food Science

and Nutrition (CRiFSaN), Dairy Science Technology (DST), Deutsche Lebensmittel-Rundschau (DLR), European Food Research and Technology (EFRaT), European Journal of Lipid Science and Technology (EJoLSaT), Food and Chemical Toxicology (FaCT), Food Hydrocolloids (FH), Food Microbiology (FM), Food Policy (FP), Food Quality and Preference (FQaP), Food Research International (FResearchI), Food Reviews International (FReviewsI), International Dairy Journal (IDJ), International Journal of Dairy Technology (IJoDT), International Journal of Food Microbiology (IJoFM), Journal of AOAC International (JoAI), Journal of Cereal Science (JoCS), Journal of Dairy Research (JoDR), Journal of Food Biochemistry (JoFB), Journal of Food Engineering (JoFE), Journal of Food Protection (JoFP), Journal of Medicinal Food (JoMF), Journal of Sensory Studies (JoSenStu), Journal of Texture Studies (JoTS), Journal of the American Oil Chemists Society (JotAOCS)

Sport Science (23) American Journal of Physical Medicine Rehabilitation (AJoPMR), American Journal of Sports Medicine (AJoSM), Archives of Physical Medicine and Rehabilitation (AoPMaR), British Journal of Sports Medicine (BJoSM), Clinical Biomechanics (CB), Clinical Journal of Sport Medicine (CJoSM), Clinics in Sports Medicine (CiSM), European Journal of Applied Physiology (EJoAP), Exercise and Sport Sciences Reviews (EaSSR), Gait Posture (GP), Human Movement Science (HMS), International Journal of Sport Nutrition and Exercise Metabolism (IJoSNaEM), International Journal of Sports Medicine (IJoSM), Journal of Applied Physiology (JoAP), Journal of Athletic Training (JoAT), Journal of Electromyography and Kinesiology (JoEaK), Journal of Motor Behavior (JoMB), Journal of Orthopaedic Sports Physical Therapy (JoOSPT), Journal of Orthopaedic Trauma (JoOT), Journal of Shoulder and Elbow Surgery (JoSaES), Knee Surgery Sports Traumatology Arthroscopy (KSSTA), Medicine and Science in Sports and Exercise (MaSiSaE), Scandinavian Journal of Medicine Science in Sports (SJoMSiS)

References

Aksnes D, Sivertsen G (2019). "A Criteria-based Assessment of the Coverage of Scopus and Web of Science." *Journal of Data and Information Science*, **4**(1), 1 – 21.

Andrade A, Gonzaelez-Jonte R, Campanario J (2009). "Journals that increase their impact factor at least fourfold in a few years: The role of journal self-citations." *Scientometrics*, **80**(2), 515–528. doi:10.1007/s11192-008-2085-9. URL https://doi.org/10.1007/s11192-008-2085-9.

Barcelo-Vidal C, Martín-Fernández JA (2016). "The Mathematics of Compositional Analysis." *Austrian Journal of Statistics*, **45**(4), 57–71. doi:10.17713/ajs.v45i4.142. URL https://www.ajs.or.at/index.php/ajs/article/view/vol45-4-4.

Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software*, **67**(1), 1–48. doi:10.18637/jss.v067.i01.

Chavarro D, Ráfols I, Tang P (2018). "To what extent is inclusion in the Web of Science an indicator of journal 'quality'?" *Research Evaluation*, **27**(2), 106–118. ISSN 0958-2029. doi:10.1093/reseval/rvy001.

Copiello S, Bonifaci P (2018). "A few remarks on ResearchGate score and academic reputation." *Scientometrics*, **114**(1), 301–306. doi:10.1007/s11192-017-2582-9.

Delgado López-Cózar E, Orduña-Malea E, Martín-Martín A (2018). "Google Scholar as a data source for research assessment." *CoRR*, **abs/1806.04435**. 1806.04435, URL http://arxiv.org/abs/1806.04435.

Else H (2018). "How I scraped data from Google Scholar." doi:10.1038/d41586-018-04190-5. URL https://www.nature.com/articles/d41586-018-04190-5.

Goodman D (2005). "Web of Science (2004 version) and Scopus." *The Charleston Advisor*, **6**, 5.

Gorraiz J (1992). *Die unertraegliche Bedeutung der Zitate*, volume 42. Biblos.

Harzing A, Van der Wal R (2008). "Google Scholar as a new source for citation analysis?" *Ethics in Science and Environmental Politics*, **8**, 62–71. doi:10.3354/esep00076.

Harzing AW, Alakangas S (2016). "Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison." *Scientometrics*, **106**, 787–804.

Harzing AW, van der Wal R (2009). "A Google Scholar h-index for journals: An alternative metric to measure journal impact in Economics Business?" *Journal of the American Society for Information Science and Technology*, **60**(1), 41–46.

Hastie T, Tibshirani R, Friedman J (2009). *The Elements of Statistical Learning*. 2nd edition. Springer, New York. ISBN 978-0-387-84857-0.

He K, Zhang X, Ren S, Sun J (2015). "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification." 1502.01852.

Hirsch J (2005). "An index to quantify an individual's scientific research output." *Proceedings of the National Academy of Sciences*, **102**(46), 16569–16572. doi:10.1073/pnas.0507655102.

Holmberg K (2015). *Altmetrics for Information Professionals: Past, Present and Future*. Elsevier Science. ISBN 9780081002773. URL https://books.google.ch/books?id=GhdiBQAAQBAJ.

Kaltenborn K, Kuhn K (2003). "Der Impact-Faktor als Parameter zu Evaluation von Forscherinnen/Forschern und Forschungen." *Medizinische Klinik*, **98**(3), 153–169. doi:10.1007/s00063-003-1240-6.

Kingma D, Ba J (2014). "Adam: A Method for Stochastic Optimization." *CoRR*, **abs/1412.6980**.

Kraker P, Lex E (2015). "A Critical Look at the ResearchGate Score as a Measure of Scientific Reputation." In *In Proceedings of the Quantifying and Analysing Scholarly Communication on the Web workshop (ASCW'15)*. Oxford, UK. doi:10.5281/zenodo.35401.

Li J, Burnham J, Lemley T, Britton R (2010). "Citation Analysis: Comparison of Web of Science, Scopus, SciFinder, and Google Scholar." *Journal of Electronic Resources in Medical Libraries*, **7**(3), 196–217. doi:10.1080/15424065.2010.505518.

López-Cózar E, Cabezas-Clavijo A (2013). "Ranking journals: could Google Scholar Metrics be an alternative to Journal Citation Reports and Scimago Journal Rank?" *Learned Publishing*, **26**(2), 101–114. doi:10.1087/20130206.

Maronna R, Martin D, Yohai V (2006a). *Robust Statistics: Theory and Methods*. Wiley Series in Probability and Statistics. Wiley. ISBN 9780470010921. URL https://books.google.ch/books?id=iFVjQgAACAAJ.

Maronna R, Martin R, Yohai V (2006b). *Robust Statistics: Theory and Methods*. John Wiley, New York.

Martín-Martín A, Costas R, van Leeuwen T, López-Cózar ED (2018a). "Evidence of open access of scientific publications in Google Scholar: A large-scale analysis." *Journal of Informetrics*, **12**(3), 819 – 841. ISSN 1751-1577. doi:https://doi.org/10.1016/j.joi.2018.06.012.

Martín-Martín A, Orduna-Malea E, Thelwall M, Delgado López-Cózar E (2018b). "Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories." *Journal of Informetrics*, **12**(4), 1160–1177. doi:10.1016/j.joi.2018.09.002.

Meho L, Yang K (2007). "A New Era in Citation and Bibliometric Analyses: Web of Science, Scopus, and Google Scholar." *Journal of the American Society for Information Science and Technology*, **58**(13), 2105–2125.

Mongeon P, Paul-Hus A (2016). "The journal coverage of Web of Science and Scopus: a comparative analysis." *Scientometrics*, **106**(1), 213–228. doi:10.1007/s11192-015-1765-5.

Orduna-Malea E, Martín-Martín A, Thelwall M, Delgado López-Cózar E (2017). "Do ResearchGate Scores Create Ghost Academic Reputations?" *Scientometrics*, **112**(1), 443–460. doi:10.1007/s11192-017-2396-9.

Pentz E (2001). "CrossRef: A Collaborative Linking Network." URL http://webdoc.sub.gwdg.de/edoc/aw/ucsb/istl/01-winter/article1.html.

Prins A, Costas R, van Leeuwen T, Wouters P (2016). "Using Google Scholar in research evaluation of humanities and social science programs: A comparison with Web of Science data." *Research Evaluation*, **25**(3), 264–270. doi:10.1093/reseval/rvv049.

Seeger T, Kohlen R, Strauch D (2004). *Grundlagen der praktischen Informationen und Dokumentationen*. 5th edition. KG Saeur, München. ISBN 3-598-11674-8.

Simons K (2008). "The Misused Impact Factor." *Science*, **322**(4899), 165. doi:10.1126/science.1165316.

Stock M, Stock W (2003). "Die Wissenschaftliche Artikel, Patente und deren Zitationen. Der Wissenschaftsmarkt im Fokus." *Password*, **10**, 30–37.

Taheri S (2013). "Trends in Fuzzy Statistics." *Austrian Journal of Statistics*, **32**(3), 239–257. doi:10.17713/ajs.v32i3.459.

Thelwall M, Kousha K (2016). "ResearchGate articles: Age, discipline, audience size, and impact." *JASIST*, **68**, 468–479. doi:10.1002/asi.23675.

Tukey J (1960). "A survey of sampling from contaminated distributions." *Contributions to Probability and Statistics*, pp. 448–485.

Van Eck N, Waltman L, Lariviere V, Sugimoto C (2018). "Crossref as a new source of citation data: A comparison with Web of Science and Scopus." Blog, last checked 28.01.2020, URL https://www.cwts.nl/blog?article=n-r2s234&title=crossref-as-a-new-source-of-citation-data-a-comparison-with-web-of-science-and-scopus.

Vanclay J (2012). "Impact factor: outdated artefact or stepping-stone to journal certification?" *Scientometrics*, **92**(2), 211–238. doi:10.1007/s11192-011-0561-0.

Wood S (2003). "Thin plate regression splines." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**(1), 95–114. doi:10.1111/1467-9868.00374.

Wood S (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Affiliation:

Matthias Templ
Institute of Data Analysis and Process Design
Zurich University of Applied Sciences
CH-8400 Winterthur, Switzerland
E-mail: matthias.templ@gmail.com
URL: https://www.zhaw.ch/de/ueber-uns/person/teml/

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 49
June 2020

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2019-12-30
Accepted: 2020-06-19