Towards a consolidation of worldwide journal rankings — A classification using random forests and aggregate rating via data envelopment analysis

Grigory Pishchulov • Heinz Tüselmann • Rudolf R. Sinkovics

June 2014

Authors and contact information:

Grigory Pishchulov
Assistant Professor in Supply Chain Management
Leverhulme Trust Overseas Visiting Fellow
Faculty of Business, Economics and Social Sciences
TU Dortmund University
Martin-Schmeißer-Weg 12, 44227 Dortmund, Germany
grigory.pishchulov@tu-dortmund.de
www.wiso.tu-dortmund.de/scm/
Phone: 0049 231 755 3234

Heinz Tüselmann
Professor of International Business
Chair of the Academy of International Business UK & Ireland Chapter
Centre for International Business and Innovation (CIBI)
Manchester Metropolitan University Business School
All Saints Campus, Oxford Road, Manchester M15 6BH, UK
h.tuselman@mmu.ac.uk
www.business.mmu.ac.uk/cibi
Phone: 0044 161 247 3908

Rudolf R. Sinkovics
Professor of International Business
Centre for Comparative & International Business Research (CIBER)
The University of Manchester, Manchester Business School
Booth Street West, Manchester M15 6PB, UK.
Rudolf.Sinkovics@manchester.ac.uk
www.manchester.ac.uk/research/rudolf.sinkovics
Phone: 0044 161 305 8980

Towards a consolidation of worldwide journal rankings — A classification using random forests and aggregate rating via data envelopment analysis

Grigory Pishchulov, Heinz Tüselmann, Rudolf R. Sinkovics

Abstract

The question of how to assess research outputs published in journals is now a global concern for academics. Numerous journal ratings and rankings exist, some featuring perceptual and peer-review-based journal ranks, some focusing on objective information related to citations, some using a combination of the two. This research consolidates existing journal rankings into an up-to-date and comprehensive list. Existing approaches to determining journal rankings are significantly advanced with the application of a new classification approach, 'random forests', and data envelopment analysis. As a result, a fresh look at a publication's place in the global research community is offered. While our approach is applicable to all management and business journals, we specifically exemplify the relative position of 'operations research, management science, production and operations management' journals within the broader management field, as well as within their own subject domain.

Key Words

Citation indices, Journal rankings, Journal lists, Research assessment, Data envelopment analysis

1 Introduction and objectives

The ranking of academic journals is a highly contentious element of research assessment, and thus a widely debated foundation stone for the ranking of individual research outputs and university rankings [1, 2]. As it affects people's careers and aspirations, the issue is one of perennial topicality and debate. Findings are repeatedly challenged as lists arguably bear non-intended consequences, skew scholarship and foster academic monoculturalism [3], and the methodologies underpinning the various approaches are contested as they are open to non-intended use [4, 5]. Within business and management, in recent years we have witnessed an increasing proliferation of rankings, listings and productivity indicators, drawing the attention of a wide range of academic disciplines, including accounting, economics, finance, international business and marketing [6], of associations such as the Association of Business Schools (ABS and the Association to Advance Collegiate Schools of Business (AACSB), among others, but also that of dominant industry players such as Thomson Reuters' Web of Science, Elsevier's Scopus, and Google Scholar. These various parties are distinguished by unique interests. The commercial providers have started to monetize a rapidly expanding and lucrative global intelligence information business by building on the academic 'gift economy' [7] — collecting institutional profile information and then selling it back to the institutions for strategic-planning purposes [8]. However, the aim of this paper is not to go into aspects of 'use and abuse' or epistemological positions regarding journal rankings [2, 4]. Instead, given their broad adoption in today's academic practice, we address some distinct methodological shortcomings of the previous attempts to rank journals and contribute to the development of a more suitable methodology, which in turn, can be used to gauge the relative standing of individual journals more realistically.

There are three conventional ways of assessing journal quality: (i) subjective (perceptual), (ii) objective (citation-based) and (iii) a combination thereof (hybrid). All three feature well-known methodological limitations [9-11]. Recently, a fourth approach has gained momentum — the 'meta'-ranking approach — which, like the hybrid approach, is intended to provide a balanced view by delivering a composite journal ranking [cf. 12, 13]. In contrast to the hybrid studies, which usually combine a few rankings or ratings and often involve the hand-collection of perceptual data, meta-analyses typically rely on a comprehensive selection of existing, in many cases reputable, rankings or ratings, and aim to deliver a reproducible outcome (cf. Table 1). As outlined, the existence of journal rankings is often — justifiably — contested on philosophical grounds, and there is the fundamental question whether possible distortions in terms of scholarship and unintended consequences of ranking exercises [see e.g. 2] may offset the advantages of increased manageability of scholarly outputs. Indeed, the emergence of meta-rankings can be seen as a result of the sheer volume and range of diverse lists that are — counter to the original motivation for developing them, which was to improve academic resource 'management' — proving to be unmanageable outside their respective academic institutions and often include different selections of journals. Within the academic community there seems to be agreement that *if* rankings are being used, the agenda should be the pursuit of a rigorous and objective perspective, based on state-of-the-art methodologies, free of individual stakeholder interests in this contentious area.

However, despite the advances made by meta-studies, a number of shortcomings remain. These include, (i) arbitrary inclusion or datedness of journal lists; (ii) over-reliance on citation data; (iii) limited coverage in terms of disciplinary focus, number of journals and number of lists included; (iv) inadequate treatment of missing data and unsophisticated imputation methods; (vi) treatment of ordinal rank data as metric; (vii) choice of ranking categories.

In the present study, we elaborate an approach that addresses these shortcomings while combining the strong features of existing studies, extending these and adding novel features. Therefore, we substantiate the methodological underpinnings to the current debate on journal rankings. We (i) extend recent work and offer an aggregate journal ranking based on a comprehensive number of journals, (ii) cover a significant number of disciplines within business and management, and (iii) deploy a unique methodological approach and integrate subjective and objective rankings with a focus on systematism and the production of comprehensive journal rankings. Specifically, this is the first meta-ranking to feature both the random forests framework (a non-parametric state-of-the-art predictive learning method) for missing data imputation and data envelopment analysis (DEA) (an established non-parametric approach to performance evaluation of peer entities) for the aggregation of rankings. This paper is decidedly focused on the methodological advancement of existing journal rankings. Thus, our final aggregate journal ranking outcomes (see and Table 5) can be seen as frame of reference for a substantive discussion and objectification of journal rankings, which is otherwise rather politicized.

The paper is organized as follows. The next section provides a critical review of objective, subjective and hybrid approaches to journal ranking and rating. Following this, Section 3 provides an overview of the major meta-ranking studies. Subsequently, in Sections 4 to 6, we present our novel meta-approach to journal ranking and rating, discuss its specific methodological advancements and apply it to our data set of journal rankings and ratings. This involves dealing with issues of database compilation, data missingness and imputation methods, classification trees, random forests and the subjection of the data to DEA. Section 7 concludes with a discussion of main results of our study and their implications. Appendices provide full modeling and computational details.

With particular emphasis on operations research, management science, production and operations management (OR/MS/POM), we apply the method to ascertain the relative positions of journals within the broader business and management discipline, as well as the relative position within the OR/MS/POM field.

2 Review of objective, subjective and hybrid approaches to journal ranking and rating

With regard to *objective* ranking, issues arise around the analysis of citation data. The Impact Factor delivered by the Journal Citation Reports [14] — defined as the number of cites received in the given year by an average article published in the given journal within the preceding years — is the most widely accepted citation-based measure for “significance and performance of scientific journals”. It is widely acknowledged for its comprehensibility, robustness and availability [15]. Yet, it has received a considerable amount of criticism in the literature, connected to the accuracy problem in collecting citation data, undifferentiated treatment of citations, biases due to different maturing of published work across different journals, inaccurate definition of citable work and differing citation habits across different sub-disciplines. Further criticism includes biasedness towards journals with lengthy articles [15, see also 16]; and a selective disciplinary and geographical coverage [17, 18]. Some of these deficits have recently been addressed by introducing a newer, prestige-oriented metric called Eigenfactor Score [19] which augments the Journal Citation Reports, and the emergence of Scopus — a citation database by Elsevier which offers a broader journal coverage together with new citation indices SNIP (Source-Normalized Impact per Paper) and SJR (SCImago Journal Rank). These aim to account for discipline-related citation habits and the prestige of the citing journals, respectively [20, 21]. Yet, and despite these advancements, extensive discussions of the underlying methodological issues raise concern of the sole reliance on citation-based analysis in journal ranking exercises. This is because important work may be considered as "common knowledge" and is sometimes left uncited — with acknowledgement given to other work or citation counts frequently representing simply fashion and herding within the academic community which implicates that citing does not necessarily imply influence [9, 22, 23]. There are also problems of selective citations and the opportunity for self- and mutual citations, a poor association between the quality of a journal and that of individual articles in it, as well as possible subjectivity which can be pertinent to the analysis based on the objective citation data [5, 24, 25]. Regardless of these shortcomings, the citation impact factor remains an important indicator in the academic community to assess journal quality.

Subjective, or perceptual, rankings are developed via opinion surveys among the experts within an institution, a society, or a research network and may be motivated by the needs to elaborate a basis for institutional decision making and evaluation purposes as well as to provide guidance within particular disciplines [1, 26, 27]. For these reasons, a variety of rankings exist which are tailored to the needs of a particular institution or a discipline [10, 26-28]. Generally, perceptual rankings alleviate the problems pertinent to citation data, and explicitly capture the *perceived* quality of journals [5, 29]. On the other hand, they are prone to biasedness in the experts' judgments — due to the institutional focus or self-identification with particular journals [11, 26]. Furthermore, the coverage of perceptual lists is often restricted to a particular discipline or by institutional preferences [26].

Due to the shortcomings of the above two approaches, the *hybrid* lists — which in some way combine subjective and/or objective data — have gained attention in the literature [e.g. 13, 29, 30]. Indeed, pooling data that originates from different sources helps to produce a more balanced view and is seen as a desired approach [13, 27, 31]. However, hybrid ranking lists typically have a particular disciplinary or geographical focus; they usually combine a few rankings or ratings and involve hand-collection of perceptual data, and, with a few exceptions, use unsophisticated and less principled techniques for data aggregation [cf. 1].

Because objective, subjective and hybrid approaches have attracted the above criticisms, the meta-approach to journal ranking and rating has recently received a substantial development, being intended to overcome the drawbacks of the hybrid approaches by relying on a comprehensive selection of existing, in many cases reputable, rankings or ratings, and aiming to deliver a reproducible outcome.

3 Overview of journal meta-rankings and ratings

Table 1 offers a compilation of the main journal meta-ranking studies. As can be seen, most of these studies focus on particular sub-disciplines, with the exception of Mingers and Harzing [1] and Halkos and Tzeremes [22] who take a cross-disciplinary approach. The journal coverage ranges from 25 to 229, with the exception of Mingers and Harzing [1] who cover over 800 journals. In terms of rankings used, most of the studies draw on a combination of subjective and objective rankings. Two thirds of the meta-rankings are based on journal rankings contained in Harzing's broadly accepted Journal Quality List (JQL) [32].

The number of underlying rankings is often 10 or less. There is quite a spread in terms of the recentness of the rankings, with only two studies covering recent years. As for data missingness, which arises because of selective coverage of journals, either this is not addressed, or it is not dealt with properly in these meta-rankings (see Section 5.1). For Theußl et al. [33] and Cook et al. [12], data missingness is not an issue. They effectively adopt the perspective that only the observed rank data can determine the ultimate ranking. There are a few, varied, attempts to impute missing data: for example, Bancroft et al. [34] employ a maximum likelihood approach, while Mingers and Harzing [1] implement a form of chained regression.

Insert Table 1 about here

As for the aggregation method for rating/ranking journals, the main approaches used are scoring methods, cluster analysis and consensus ranking via integer programming, with only one study, that of Halkos and Tzeremes [22], featuring the state-of-the-art DEA. While scoring is attractive due to its simplicity, it is rather subjective in its application. Cluster analysis offers a more advanced approach, but usually only delivers a limited set of categories. DEA, in contrast, is a methodologically profound and objective approach that helps to reduce manipulation, over-interpretation and bias. The integer programming approach deployed by Theußl et al. [33] and Cook et al. [12] is very effective at producing a consensus ranking, yet it works within the confines of treating missing data as non-existent. Further, it cannot deliver an interval or ratio scale outcome.

4 Compiling a database for journal meta-ranking

In view of the limitations and shortcomings of meta-rankings described above, we proceed to develop a comprehensive journal database, which will subsequently be subjected to our rating and ranking exercise.

The primary databases are the journal quality ranking lists contained in the 49th edition of Harzing's Journal Quality List (JQL49) [32] and the Thomson Reuters Journal Citation Reports [14, various years]:

- The ranking lists contained in JQL49 are dated in the range from 2001 to 2013.
- To reflect an up-to-date, rather than historical, journal status, we select the 10 most recent ranking lists (out of the 22 contained in the JQL49 database), covering a 6-year time span (2008 to 2012).[1]

[1] The Wirtschaftsuniversität Wien Journal Rating 2008 (WIE 2008) list was excluded because it now publishes only its A+ and A ratings and no longer its B, C and D ratings. If we had included WIE 2008,

- We update and correct a number of the journal lists in JQL49 based on information in the most recent publicly available editions of the respective ranking lists.[2]
- In order to capture a comprehensive quantity of journals, all journals listed in JQL49, a total of 939 journals, are considered. This provides a broad and cross-disciplinary coverage.

The 10 ranking lists selected for aggregation by means of DEA (Section 6) are labeled 'target lists', as shown in Table 2. In an additional step, these rankings are further augmented by including 2011 Impact Factor data from the Journal Citation Reports [14][3]. Thus, we use 11 rankings in total.

Insert Table 2 about here

Most of the journal quality lists rank the journals on an ordinal scale, using differing numbers of scale gradations (ranks) and their designations. Thus, we relabeled the ranks in each of the lists as 1, 2, etc., from highest to lowest. The length of the original scale is maintained in all lists. This overcomes the problems related to adjusting original scale lengths to a common scale length, and the resulting subjectivity/arbitrariness [31].

In addition, all journals with an Impact Factor are ranked and divided into quintiles, with 1 denoting the top quintile, 5 the lowest quintile, and a value of 6 being assigned to journals that are not indexed in the 2011 Journal Citation Reports. This procedure helps to

journals that it ranked below an A would have been wrongly recorded as missing cases. We also excluded Den 2011 (Danish Ministry Journal List) because it has only two categories: top journal and others, it is thus lacks differentiation. We further excluded FNEGE (Foundation National pour l'Enseignement de la Gestion des Entreprises) 2011 because it merely replicates the CNRS (Centre National de la Recherche Scientifique) 2011 ratings for management and business journals. Finally, we excluded AERES (Agence d'évaluation de la recherche et de l'enseignement supérieur) 2012 because it mainly maps CNRS 2011 ratings to a scale with fewer gradations and does not substantially add to the existing data.

[2] We have in particular made corrections in the ranking lists ABS 2010, CNRS 2011, UQ 2011 and HEC 2011. These and other adjustments of the JQL can be obtained from the authors on request.

[3] We use the two-year average of the 2011 Impact Factor [14]. An alternative would have been to use the five-year average. However, for a number of journals, no five-year average data exists. If we had used a five-year average, these journals would have received a non-entry, despite being included in the citation list. The same rationale applies to the exclusion of alternative measures such as the article influence score.

alleviate several of the well-known shortcomings of using the Thomson Reuters metric score in analyses [cf. 18], as well as the problems with conventional normalization procedures [26].

5 Resolving the data missingness problem in journal rankings and ratings

5.1 Data missingness and imputation approaches

A significant problem pertinent to journal meta-ranking approaches is the considerable amount of missing data. In our database of 939 journals, target ranking lists from 1 to 10 (see Table 2) contain 4,770 entries out of the 9,390 possible. This corresponds to an overall missingness rate of nearly 50%. The pattern of missingness varies across journals, and coverage rates range from approximately 28% to 88% across lists. As can be seen from Figure 1, three strategies for dealing with data missingness can be identified in the existing journal ranking studies:

*1) C*ompleting *the data set.* This can be achieved either by the *removal* of records with missing data — which would however lead to an undesirable loss of information — or by *imputation*. The latter involves replacing missing entries with artificially generated values (see below).

2) Averaging. For example, Rainer and Miller [35] calculate the average score from the ranks that are available. However, this may lead to biased results [see 34]. The same criticism applies to the work of Franke and Schreier [31] and Steward and Lewis [27] who use a form of weighting to replace the missing data.

3) Reliance on stated preferences. Cook et al. [12] and Theußl et al. [33] employ an integer programming method that seeks to find the consensus ranking that exhibits the least total deviation from the underlying rankings. They thus neither extrapolate nor disregard existing data. Instead, their approach relies purely on the pairwise preference relations between the journals, effectively stated by the underlying ranking lists. Despite its advantages, we do not use the approach in this paper, instead favoring imputation for the following reasons: Firstly, as Mingers and Harzing [1] point out, lists can be biased in their selective coverage and imputation reduces this bias. Tse [36] supports this, referring to humans' limited information-processing capability [cf. 12]. Secondly, imputation enables us to extend lists while retaining their original spirit [37]. Therefore, this paper considers imputation to be the most viable strategy for dealing with missingness in journal lists.

Insert Figure 1 about here

In line with Farhangfar et al. [38] and Gheyas and Smith [39], three approaches to missing data imputation can generally be identified (see Figure 1):

1) *Data-driven imputation methods* [38]. Missing items are replaced with artificial values, for example the mean, median or mode of the respective variable, or with a random draw from the observed values [39, 40]. However, these methods distort the association between variables [40]. In the context of the journal ranking problem, the approach would lead to the distortion of the aggregate ranks of individual journals. While this is partly overcome by Benati and Stefani [10], who associate missing rank data with a separate category, their approach is not tailored to offer a rank ordering of journals.

2) *Parametric imputation methods.* These methods assume an explicit data model, such as the *regression imputation* [40, 41] or the *maximum likelihood* approach featured by the *expectation–maximization* (*EM*) *algorithm* [42]. The *multiple imputation* methodology [see e.g. 40, 43] represents a further advancement but its reliance on the assumed data model can lead to incorrect inferences [e.g. 41, 44] and it should be used with caution [43, 45]. With regard to journal rankings, Bancroft et al. [34] use the maximum likelihood approach to arrive at rank estimates for 25 journals related to business policy/strategic management research, previously ranked in a longitudinal study with censoring. Mingers and Harzing [1] use a form

of chained regression imputation to estimate missing ranks for a restricted subset of journals drawn from seven ranking lists of the JQL (17th ed.). Similarly, Schulze et al. [37] carry out repeated imputation through a sequential univariate regression and a single imputation through a sequential multivariate regression, while utilizing a number of additional ranking lists as predictor variables (yet they deal only with imputation and do not attempt to derive an aggregate rating or ranking). However, parametric methods have received criticism regarding potential model misspecification and validity concerns [39, 41, 46].

3) For the purposes of our study, we pursue the branch of *non- and semi-parametric imputation methods*, as these do not (or do not fully) rely on a data model [39, 41]. A major advancement within this branch is the group of *machine learning* approaches [46], which we draw on for our study [see e.g. 39, 47][4]. In particular, the work by Twala et al. [48] demonstrates the competitiveness of tree-based methods compared to parametric imputation methods in terms of predictive accuracy, see also Hapfelmeier et al. [49]. More specifically, we utilize the *random forests* method [50] which represents a recent and remarkable advancement in non-parametric classification and regression. This method employs an ensemble of classification or regression trees (see Section 5.2) for predicting the response variable as a committee, while the process of constructing the individual trees in the ensemble involves randomness. This approach results in a superior prediction accuracy that compares favorably or competitively ‘to the best statistical and machine learning methods’ [51-53]. At the same time, the random forests method is deemed more versatile than the conventional statistical methods and can flexibly accommodate a wide range of prediction problems — even those that are ‘nonlinear and involve complex interactions’ [53], while being

[4] Gheyas and Smith [39] provide an overview of imputation approaches, and in particular those featuring neural networks. However, we do not consider this group of methods in our study, preferring instead a methodology which is more straightforward in its application.

acknowledged, among others, for robustness and ease of training as compared to other machine learning methods [52, 53].

5.2 Classification trees and random forests and their application

Classification and regression trees (CART) represents a well-established and widely used non-parametric predictive learning method [46, 52], which has been developed with a strong emphasis on the possibility of missing data among the variables. It seeks to determine the association between the response and predictor variables via recursive, data-driven, partitioning of the predictor space and exhibits a degree of accuracy comparable to the best of the classical statistical methods [54], while producing highly interpretable models and exhibiting other strong advantages [52]. Breiman [50] has advanced CART to produce the random forests framework which effectively reduces variability of individual tree predictions by de-correlating and aggregating them across a tree ensemble, offering as a result a remarkably high prediction accuracy and a number of other advantages [52, 53]. Random forests are particularly easy to train, basically requiring to fine tune a few parameters only. Sections A.1.1 and A.2.1 in Appendix A give more detailed overviews of CART and random forests methodologies.

Drawing on the random forest framework, we proceed towards imputing the missing data in each of the target journal-ranking lists[5]. Imputation in each individual list is based on predictor variables which are comprised of: (i) journals' subject areas as per JQL49; (ii) the remaining target lists[6], (iii) other journal ranking lists included in JQL49, and (iv) Citation Impact Factors from the Journal Citation Reports (see Table 2 and Table 3). Specifically, we utilize ranking lists from 2001 onwards (see Table 3). Although these are older than the cut-

[5] Table 2 exhibits 11 target lists. Imputation has to be carried out in 10 of these.

[6] As indicated in Section 4, target ranking list no. 11 is based on 2011 Impact Factor data [14] and features an ordinal rank scale with a few gradations for the purposes of aggregate ranking. When acting as a predictor variable for missing data imputation, this ranking list however maintains original ratio scale data of the 2011 Impact Factor if the latter is available, and indicates a missing value otherwise.

off date for the target lists, and are therefore based on more historical data, their inclusion is warranted to improve imputation accuracy.[7]

Insert Table 3 about here

The first step in the application of random forests is to (i) *pre-impute missing entries* in each single predictor.[8] This task is necessary as the predictor variables themselves have missing values. While random forests have a built-in mechanism for this step, we use CART to accomplish this task.[9] The second step involves (ii) *checking the imputation accuracy* in the target lists using cross-validation [see e.g. 52]. We find differences in the accuracy of the imputations for different ranking lists. For instance, missing values for Ast 2008 are found to be more difficult to predict than missing values in other lists. Additionally, we perform numeric experiments using different settings for CART and random forests to determine the optimal parameter settings for the imputation engine. The third step is (iii) *the actual imputation of missing data* in the target lists. Having regard for misprediction rates in all of the target lists in step *ii*, we find that it would be inappropriate to stick to the point estimates of missing rank data; instead, the uncertainty involved must be reflected in rank predictions. We therefore adopt, similarly to Zhou et al. [30], a fuzzy rank approach — by letting each journal belong to two or more different ranks within the same ranking list, while the respective degrees of rank membership are required to sum up to unity (e.g. in ABS 2010, journal X is 60% associated with rank ‘1’ and 40% with rank ‘2’). A particular advantage of

[7] Although VHB 2011 and UQ 2011 are included in the primary list, VHB 2003 and UQ 2007 are also used for imputation purposes because they use different methodologies, scoring systems or ranking procedures from the newer versions of the lists [for details and a discussion of the VHB and UQ lists, see 32].

[8] All necessary computations have been conducted in R software environment (version 3.0.0). We have used CART implementation delivered by the R package *rpart* (version 4.1-1) and the implementation of the random forest method delivered by the R package *randomForest* (version 4.6-7).

[9] This approach had to be adopted because *randomForest* package (see footnote 8) does not allow for missing data when predicting an unknown response. Handling such situations is however an inherent feature of CART, thus the said approach has been adopted (see e.g. Hastie et al. [52, p. 333]). After pre-imputing the missing values in the predictor variables, we add one dummy variable per each such predictor to indicate whether the respective predictor value is original or has been pre-imputed.

this approach is that our aggregate ranking method (see Section 6 below) accommodates fuzzy rank membership in a natural way.

Notably, random forests have a built-in mechanism for estimating individual rank probabilities when making a prediction. We accordingly adopt these probabilities as the respective degrees of rank membership predicted for the given journal in the given ranking list. Random forests exhibited a superior performance in producing such estimates [55]; however, that performance can be further improved by means of *calibration* techniques. For this purpose we have employed the calibration method suggested by Boström [56] and similarly used the *Brier score* (mean squared deviation of the predicted rank probabilities from the true ones) as performance measure, while the calibration data set has been comprised of all test data samples which had been formed in the course of cross-validations conducted in step ii. In our experience, calibration has yielded only a marginal improvement of the *Brier score*, which is in line with Niculescu-Mizil and Caruana [55]. By completing this step we have produced a comprehensive and complete data set, which is then subjected to DEA. Appendix A provides details to the individual steps of the above imputation procedure.

6 Rating and ranking journals by DEA

DEA [57] represents an established management science approach to multi-attribute rating of peer entities [58-60], in our case journals. A typical DEA setup involves measuring the efficiency of a number of peer entities called *decision-making units*, or DMUs (e.g. universities) that have a number of common inputs (e.g. budgets, number of staff) and outputs (e.g. research outputs, teaching quality). These inputs and outputs constitute the basis for evaluating the efficiency of the DMUs. There are no *a priori* weights attached to the inputs and outputs. Instead, DEA offers each DMU an opportunity to cross-evaluate and apply input and output weights that most favorably express its own efficiency. Essentially, DEA determines ‘frontiers rather than central tendencies’ in the data [58], [61]. As a non-parametric method, it requires no a priori assumptions on the interaction between the variables in the data set [58].

Conventionally, the DEA methodology is applied to metric data, but it has been extended to cover a variety of settings with ordinal rank data [see 62 for a recent discussion]. Cook et al. [see e.g. 63] further addressed settings with a differentiated treatment of individual rankings — an approach that particularly suits the aggregate journal rating purposes. Against this background, DEA treats rank positions in individual ranking lists as outputs of the DMUs (i.e., journals) while assuming away any variable inputs. It then allows each journal to attach weights to the individual rank positions in each target ranking list. These rank weights should represent the respective journal in the best possible light or, more specifically, provide it with the maximum possible weighted average rank, representing the journal’s self-rating of its own performance. Furthermore, the weights chosen by the journal also determine performance ratings of all other journals from its perspective. Thus, by choosing its own rank weights, each journal explicitly evaluates itself vis-à-vis all other journals. In this way, a cross-evaluation matrix is obtained, from which the ultimate ratings of the individual journals can be derived [64, 65].

Due to the DEA’s advantage of avoiding *a priori* assumptions and subjective bias, we adopt the above approach to derive an aggregate journal rating and ranking. To this end, we employ the DEA framework for aggregation of ordinal preferences by Green et al. [64] while further extending it to include a rank discrimination threshold in line with Noguchi et al. [65] and a differentiated treatment of individual rankings as in Cook et al. [63]. In addition to that, we enforce convexity constraints on the rank weights in line with Hashimoto [66]. Further, we use the aggressive form of cross-evaluation [64] to give each journal the opportunity to appear most strongly against its peers, and derive the ultimate journal ratings from the cross-evaluation matrix using the arithmetic means so that all journals have an equal say in determining the final result. Section B.1 in Appendix B provides specific details to the DEA model adopted in our study and to the cross-evaluation method.

As explained in Section 4, we subject 11 target ranking lists to the above aggregation procedure, while the missing rank data has to be imputed in these lists by means of the random forests method as per Section 5.2, and supplied to DEA in the form of fuzzy membership degrees to which the respective journal is associated with the individual ranks of the respective ranking list. This represents a distinctive feature of our model as compared to DEA approaches to ordinal rank data [62-66]. Random forests method can impute the fuzzy rank membership in a natural way and ordinal DEA can also accommodate fuzzy rank data. Thus, these two approaches are complementary to each other for the purposes of aggregate journal rating.

Before proceeding with DEA, we exclude from the final list of journals used in this study those journals with ranks available for less than 25% of the 11 target lists (see Table 2). This reduces the list from 939 to 786 journals, representing around 84% of all journals in JQL49. This approach is taken because the ranks that are available for sparsely ranked journals may not be representative enough, and it also ensures that the imputations are ‘pluralistic’ enough rather than being based on just one or two rankings. Our conservative choice of this lower limit of 25% for the number of original rankings per journal is in line with previous related studies, such as Cook et al. [12] and Theuβl et al. [33].[10]

A particular problem in attaching weights to the individual journal ranks in our DEA exercise is the arbitrary choice of a rank discrimination threshold to separate the weights of any two consecutive ranks [see 62, 64, 65, 67]. If the threshold is virtually ‘0’, this leads to the undesirable suggestion that there may be no difference between any pair of journal ranks.

[10] We found that the final results remain robust when this lower limit is set to a higher value, e.g. 35%.

If the threshold value is set to the maximum, this infringes on the spirit of DEA, since it largely restricts the freedom of choice in determining the rank weights [64]. We resolve this dilemma by setting up the process so that journals settle on an intermediate value of the threshold via Nash bargaining [68][11]. Section B.2 in Appendix B provides specific details to the implementation of this procedure. Accordingly, we find the compromise value of the threshold to be 31.3% of the maximal possible value. We then use DEA to rate the journals, producing in effect rating scores in the range from 0.55705 to 1, which yield 729 unique ranks, with 786 tied ranks. Table 4 and 5 offer a selection of the results.

We also conduct a series of tests to address the sensitivity of the final rating to the choice of the rank discrimination threshold. We find that the results differ across the entire range of feasible threshold values — with Pearson correlations among the corresponding ratings ranging from 80.4% to 100% and Spearman rank correlations from 79.7% to 100%. At the same time, the rating remains robust in the proximity of the selected threshold value; neither of the above two correlation measures falls below 99.97% within the range of ±10% around the selected threshold value. The final rating exhibits a Pearson correlation of 88.2% and a Spearman rank correlation of 89.9% with the rating produced by means of the Borda count — a points-based system that specifies equidistant weights for the individual ranks in each of the ranking lists.

Insert Table 4 and Table 5 around here

[11] To be specific, we consider a bargaining problem with *n* = 786 players [68] where journals are acting as players. The utility that a journal attaches to a particular threshold value is taken to be its own *standing* in the DEA rating that arises under this threshold value. A journal's standing is defined as the difference between this journal's rating score and the average one across the list, normalized to account for the length of the rating scale. The analytic form of each journal's utility function is obtained by fitting a cubic polynomial to 10 equally spaced data points computed for each journal within the feasible range of the threshold. Further, instead of using the disagreement point in the sense of the original Nash bargaining problem, we refer to the minimum utility point [69] — where, accordingly, a journal's minimum utility is its lowest possible standing throughout the entire feasible range of the threshold. The bargaining solution is then determined as the threshold value that maximizes the Nash product over the entire feasible range [see also 70].

7 Conclusion and implications

The debates over the use and abuse of journal rankings are heated and have recently heightened in their intensity. Much of the effort in the scholarly exchange regarding these rankings is concerned with the construction and publication of list data. However, fundamental issues related to epistemological positions and their implications for scholarly exchange and the scientific production system [71] are still to be resolved [72]. This paper empathizes with these concerns and criticisms in relation to issues such as the homogenization of research cultures, the reduction of pluralism, the skewness of scholarship and the polarization and entrenchment of orthodoxies [2], to mention just a few. Notwithstanding the importance of the wider and philosophical discourse, the main contribution of this paper is a methodological one, driving the advancement of journal rankings. Our position is that, if journal rankings are here to stay, we better pursue a rigorous perspective based on state-of-the-art methodologies that transcend the individual stakeholder interests in this contested field.

With this paper we provide a meta-ranking that overcomes some of the specific shortcomings of the existing meta-rankings in terms of the construction of the underlying database, the treatment of missing data and the ranking approach. To the best of our knowledge, this is the first study to go beyond previous ranking snapshots, and to uniquely feature a combined application of the random forest framework and DEA, two established non-parametric methods, in the construction of the aggregate list. This makes our study wholly non-parametric and therefore free from subjective a priori assumptions about the interaction between the various ranking and rating data included in the study. In this process, we ensure that we retain the strong features of existing and relevant methods, extend them and add novel features (such as fuzzy rank membership and rank discrimination via Nash bargaining) so as to arrive at a ‘state-of-the-art’ meta-ranking. Confidence in our findings is established through a series of extensive robustness checks, and reliability and cross-validation procedures. However, despite the recency of our methodological approach, future work may still direct its attention towards some possible extensions. For example, it could be explored whether a form of ‘discounting’ or weighting should be introduced for the imputed journal ranks, due to their omission from the original ranking studies. In our research, they are treated on an equal basis to the existing ranks.

Table 4 offers a selection of the final aggregate journal ranks. We deliberately refrain from making any judgement as to the quality of various ranks, or the ‘star-rating’ of certain journals, as is frequently found in other ranking lists. We simply provide a rank-ordering of the journals along with their numerical ratings, leaving stakeholder or user groups to arrive at their own subjective judgments regarding the cut-off points for quality grades. There are also a number of useful applications of this list. It allows for the relative standing of a particular journal to be ascertained vis-à-vis all other journals, as well as within its own subject area.

Based on our meta-ranking, Table 5 highlights the ranking order of journals within the OR/MS/POM domain. *Management Science, Journal of Operations Management* and *Operations Research* occupy the top-three positions. This discipline is represented over proportionally well when looking at the top 50% of all the journals within the business and management area. Overall, the journals in this subject area perform well vis-à-vis other disciplines that are included in our meta-ranking. Table 5 also offers a look at the relative position of journal outlets from a disciplinary perspective. OR/MS/POM journals account for about 10% of all the 786 journals in the final list of our meta-analysis. They account for around 10% of the top 5% of the whole journal list (*Management Science*, *Journal of Operations Management* and *Operations Research*, *Journal of the Royal Statistical Society: Series B*) and for around 12% of the top 10% and top 20%. Journals up to a tied rank of 19 in Table 5, including *Decision Sciences, Risk Analysis, European Journal of Operational*

Research and *Omega* fall within the top quartile of their own subject discipline and within the top 20% of all management and business journals (see Tables 4 and 5). On the other hand, OR/MS/POM journals account for only around 7% of both, the lower third and lower quartile of journals in our meta-ranking list.

Besides, our meta-ranking may also serve as a reference point onto which the grade and/or star-rating of a particular journal or the population of journals in other lists (e.g. ABS, VHB, Cranfield) can be mapped (see Table 5). This allows pinpointing whether there is congruence between the journal grading of those lists and the results of our meta-approach.

Since we have deliberately refrained from attaching grade categories to our journal rankings, the interpretation of such a comparison lies in the eye of the beholder. However, if we were to find gross discrepancies between our meta-ranking and other journal ranking lists, this would not be easy to argue away. Instead, it may serve as an invitation to the authors of the journal list in question to revisit their assessment and ameliorate such discrepancies. While our list is certainly not a panacea, we introduce a ‘dose of objectivity’ into some of the issues picked up in the wider debates on journal rankings, such as vested interests, gamesmanship and politicking. To this end, we hope to contribute to shifting the discussion back towards the essence of scholarly endeavours, namely the development of interesting and relevant contributions.

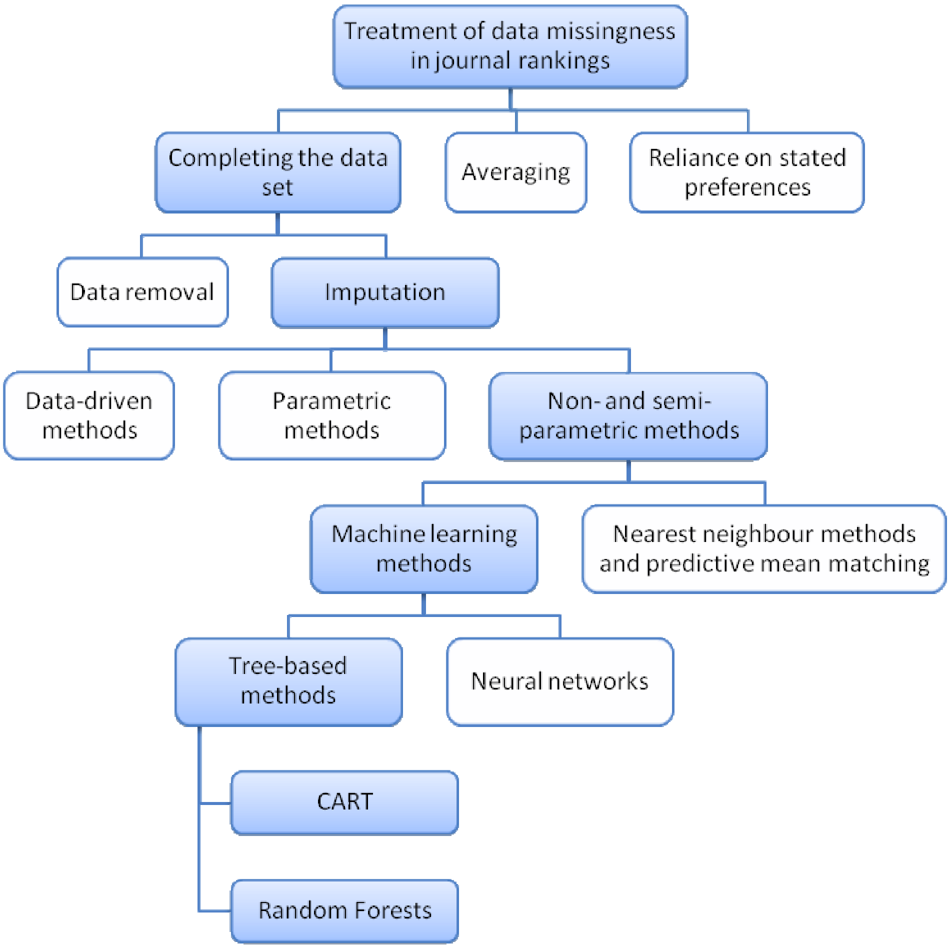
Tables and Figures

Table 1: Overview of journal meta-ranking studies and selected hybrid ranking studies

	Disciplinary focus	No. of journals	No. of rankings used	Rankings used	Age of rankings	Treatment of missing data	Imputation method	Rating / ranking by	Outcome: rating or ranking (scale type, feasible range, meaning)
Meta-Rankings									
Bancroft et al. 1999 [34]	Business policy / strategy	25	5	P	1987–1994	Imputation	Maximum likelihood	Mean rank	Rating (interval 1–4, 4 = outstanding as a publication outle ranking (ordinal 1–23, 1 = top rating)
Benati & Stefani 2011 [10]	Mathematics & Economics	138	7	JQL, OI	2002–2005	Imputation	Separate category	Cluster analysis	Classification (nominal, 4 classes)
Theußl et al. 2014 [33]	Marketing	62	12	JQL	2001–2009	Ignoring		IP	Ranking (ordinal 1–5[†], 1 = top quality)
Cook et al. 2010 [12]	Accounting	140	26	P, JQL, OI, C, U	2002–2007	Ignoring		IP	Ranking (ordinal 1–33[†], 1 = top quality)
Mingers & Harzing 2007 [1]	Business, management and related disciplines	834	10	JQL, C	1994–2005	Partial imputation	Chained regression	Cluster analysis	Ranking (ordinal 1–4, 4 = top quality)
Halkos & Tzeremes 2011 [22]	Business, management and related disciplines	229	8	C, JQL, OI	2009	Does not apply		DEA	Rating (ratio 0–1, 1 = maximum performance in terms of citedness), ranking (ordinal A–D, A = top tier rating)
Franke & Schreier 2008 [31]	Tech. & Innovation / Entrepreneurship	43	37	C, JQL, P, U, OI	1989–2004	Ignoring		Scoring	Rating (ratio 0–10, 10 = max. quality), ranking (ordinal 1–39, 1 = top rating; ordinal A–D, A = top tier rating)
Rainer & Miller 2005 [35]	Management Information Systems	50	9	P, C	1991–2001	Ignoring		Scoring	Rating (ratio 0–1, 0 = maximum quality), ranking (ordinal 1–47, 1 = top quality rating)
Steward & Lewis 2010 [27]	Marketing	100	11	P, C, U	1993–2006	Ignoring		Scoring	Rating (ratio 0–1, 0 = maximum quality), ranking (ordinal 1–49[†], 1 = top quality rating)
Hybrid Rankings (selected)									
Zhou et al. 2001 [30]	Cross-disciplinary (Hong Kong RAE)	285 *	3	C, OI, OS	1996–2000	Does not apply		Fuzzy inference	Ranking (ordinal A–C, A = top quality)
Morris et al. 2009 [26]	Business, management and related disciplines	1039 **	9	OI, C, OS	2003–2008	Ignoring		Modal score & Delphi	Ranking (ordinal 0*–4*, 4* = top quality)
Crookes et al. 2010 [73]	Nursing & midwifery	144	3	OS, C	2006–2007	Does not apply		Scoring	Rating (interval 0–100, 100 = maximum quality), ranking (ordinal 1–4, 1 = top tier rating)
DuBois & Reeb 2000 [74]	International Business	30	5	C, OS	1995–1998	Does not apply		Scoring	Ranking (ordinal 1–23, 1 = top quality)
Bauerly & Johnson 2005 [29]	Marketing (mainly US background)	252	1	U, OS	2001	Does not apply		Does not apply	Rating (ratio 1–1434, citations in doctoral program syllabi ranking (ordinal 1–34[†], 1 = top rating)
Kao et al. 2008 [13]	Management (Taiwanese journals)	46	5	C, OI, OS	2003–2005	Does not apply		DEA & scoring	Rating (ratio 0–1, 1 = maximum quality), ranking (ordinal 1–46, 1 = top rating; ordinal A–E, A = top tier rating)

Notes:	**Legend:**	**Legend:**
* Ranking of just a single journal within a single discipline with 285 journals is provided as an illustration ** As of 21 April 2009 † As for the number of journals in the reported ranking	P — Perceptual rankings published in academia OS — Opinion survey as source of perceptual data JQL — Cross-disciplinary rankings present in the JQL OI — Other institutional cross-disciplinary rankings C — Citation data or citation-based rankings U — Rankings featuring other usage data (e.g. download counts, citations in syllabi etc.) The symbols are in descending order of the respective rankings' share in the data set.	IP — integer programming DEA — data envelopment analysis

Figure 1: Approaches to treatment of missing data in journal rankings and methods of completing the data set



Note: Shaded boxes represent the approach followed in this paper.

Table 2: Target lists

No.	*Title*	*Year*	*Abbreviation*
1	Aston	2008	Ast 2008
2	Australian Business Deans Council Journal Ranking List	2010	ABDC 2010
3	Association of Business Schools Academic Journal Quality Guide	2010	ABS 2010
4	Centre National de la Recherche Scientifique	2011	CNRS 2011
5	Hautes Études Commerciales de Paris Ranking List	2011	HEC 2011
6	University of Queensland Adjusted ERA Ranking List	2011	UQ 2011
7	Association of Professors of Business in German-speaking countries	2011	VHB 2011
8	Cranfield University School of Management	2012	Cra 2012
9	ERASMUS Research Institute of Management Journal Listing	2012	EJL 2012
10	ESSEC Business School Paris	2013	ESS 2013
11	Impact Factor from the Thomson Reuters Journal Citation Reports	2011	Thomson Reuters 2012

Table 3: Additional lists used for imputation purposes

No.	*Title*	*Year*	*Abbreviation*
1	Wirtschaftsuniversität Wien Journal Rating	2001	WIE 2001
2	Association of Professors of Business in German-speaking countries	2003	VHB 2003
3	British Journal of Management (Business & Management RAE rankings)	2001	BJM 2004
4	Theoharakis et al.	2005	Theo 2005
5	Hong Kong Baptist University School of Business	2005	HKB 2005
6	European Journal of Information Systems	2007	EJIS 2007
7	European Journal of Information Systems (including citation impact factors)	2007	EJIS–CI
8	University of Queensland Journal Rating	2007	UQ 2007
9 *to* 14	Impact Factor from the Thomson Reuters Journal Citation Reports	2005 to 2010	Thomson Reuters 2006 to 2011

Table 4: Aggregate journal ranks: selected results (N = 786)

Journal	*Subject area*	*Rating*	*Ranking*	*Tied Rank*
Academy of Management Review	General Management & Strategy	1	1	1
Administrative Science Quarterly	General Management & Strategy	1	1	1
Journal of Finance	Finance & Accounting	1	1	1
Journal of Marketing	Marketing	1	1	1
Quarterly Journal of Economics	Economics	0.99853	2	5
Journal of Political Economy	Economics	0.99832	3	6
Econometrica	Economics	0.99701	4	7
American Economic Review (The)	Economics	0.99411	5	8
Accounting Review (The)	Finance & Accounting	0.98425	6	9
MIS Quarterly	MIS, KM	0.98425	6	9
Strategic Management Journal	General Management & Strategy	0.98425	6	9
Academy of Management Journal	General Management &Strategy	0.98163	7	12
Information Systems Research	MIS, KM	0.98163	7	12
Journal of Consumer Research	Marketing	0.98163	7	12
Journal of Financial Economics	Finance & Accounting	0.98163	7	12
Marketing Science	Marketing	0.98163	7	12
Review of Financial Studies	Finance & Accounting	0.98163	7	12
Journal of Economic Literature	Economics	0.98101	8	18
Journal of Applied Psychology	Psychology	0.97685	9	19
Accounting, Organizations and Society	Finance & Accounting	0.96588	10	20
Journal of Accounting & Economics	Finance & Accounting	0.96588	10	20
Journal of Accounting Research	Finance & Accounting	0.96588	10	20
Organization Science	OS/OB,HRM/IR	0.96588	10	20
American Journal of Sociology	Sociology	0.96549	11	24
Annual Review of Psychology	Psychology	0.95627	12	25
Management Science	OR,MS,POM	0.95627	12	25
Journal of Marketing Research	Marketing	0.95361	13	27
American Sociological Review	Sociology	0.95354	14	28
American Political Science Review	Public Sector Management	0.95274	15	29
Journal of International Business Studies	International Business	0.95187	16	30
Organizational Behavior and Human Decision Processes	OS/OB,HRM/IR	0.95187	16	30
Journal of Operations Management	OR,MS,POM	0.94574	17	32
American Journal of Public Health	Economics	0.94227	18	33
Review of Economic Studies	Economics	0.93991	19	34
Journal of Economic Perspectives	Economics	0.9382	20	35
Operations Research	OR,MS,POM	0.9379	21	36
Journal of the American Statistical Association	Economics	0.93394	22	37
Organization Studies	OS/OB,HRM/IR	0.93087	23	38
American Psychologist	Psychology	0.93029	24	39
Journal of the Royal Statistical Society, Series B	OR,MS,POM	0.92495	25	40
Research Policy	Economics	0.92386	26	41
Journal of Financial & Quantitative Analysis	Finance & Accounting	0.92215	27	42
Annals of Statistics	OR,MS,POM	0.91961	28	43
Journal of Management Studies	General Management & Strategy	0.91688	29	44

Journal of Retailing	Marketing	0.91512	30	45
Journal of Personality & Social Psychology	Psychology	0.91436	31	46
Review of Economics & Statistics	Economics	0.91409	32	47
Annual Review of Sociology	Sociology	0.90852	33	48
Journal of Development Economics	Economics	0.90058	34	49
Journal of Monetary Economics	Economics; Finance & Accounting	0.89924	35	50
First ranked journals within subject areas, outside top 50				
Business History	Business History	0.6963	214	126
Journal of Communication	Communication	0.7647	111	126
Journal of Business Venturing	Entrepreneurship	0.88189	45	60
Journal of Product Innovation Management	Innovation	0.80578	83	98
Annals of Tourism Research	Tourism	0.83548	65	78

Abbreviations: OR,MS,POM = Operations Research, Management Science, Production & Operations Management; MIS, KM = Management Information Systems, Knowledge Management; OS/OB,HRM, IR = Organisation Behavior/Studies, Human Resource Management, Industrial Relations

Table 5: Ranking position of journals within the subject area Operations Research, Management Science, Production & Operations Management (OR/MS/POM)

Journal	*Rating*	*Ranking*	*Tied Rank*	*Rank within Subject Area*	*ABS 2010*
Management Science	0.95627	12	25	1	4
Journal of Operations Management	0.94574	17	32	2	4
Operations Research	0.9379	21	36	3	4
Journal of the Royal Statistical Society, Series B	0.92495	25	40	4	4
Annals of Statistics	0.91961	28	43	5	
Transportation Research Part B: Methodological	0.88095	46	61	6	4
Decision Sciences	0.83024	68	83	7	3
Transportation Research Part A: Policy & Practice	0.81931	72	87	8	3
Risk Analysis	0.77628	102	117	9	4
Mathematical Programming	0.77052	106	121	10	3
Annals of Probability	0.76941	107	122	11	
European Journal of Operational Research	0.76367	116	131	12	3
IEEE Transactions on Intelligent Transportation Systems	0.76316	117	132	13	
Journal of the Royal Statistical Society, Series A	0.75692	123	138	14	3
SIAM Journal on Control & Optimization	0.74447	134	149	15	
Transportation Science	0.74422	136	151	16	3
IEEE Transactions on Engineering Management	0.73753	137	152	17	3
OMEGA - International Journal of Management Science	0.73751	138	153	18	3
Production and Operations Management	0.73667	139	155	19	3
Biometrika	0.73	150	166	20	
Mathematics of Operations Research	0.7298	152	168	21	3
International Journal of Production Research	0.72265	162	178	22	3
International Journal of Operations & Production Management	0.71567	176	194	23	3
Journal of Business Logistics	0.70445	197	215	24	2
Manufacturing and Service Operations Management	0.70421	200	218	25	3
Transportation Research Part C: Emerging Technologies	0.70373	201	219	26	
Journal of the Operational Research Society	0.69812	211	229	27	3

Transportation Research Part E: Logistics	0.69586	216	236	28	3
Journal of Scheduling	0.69383	222	242	29	3
International Journal of Production Economics	0.69378	223	243	30	3
Journal of Optimization Theory & Applications	0.69018	238	259	31	
Journal of Transport Geography	0.68921	242	263	32	2
Transportation Research Part D: Transport & Environment	0.68609	253	277	33	2
Reliability Engineering & System Safety	0.68578	254	278	34	3
Journal of Supply Chain Management	0.68438	260	284	35	1
Computers & Operations Research	0.68327	263	287	36	2
Service Industries Journal	0.68288	265	289	37	2
Supply Chain Management: An International Journal	0.67861	282	307	38	3
OR Spectrum	0.67193	300	329	39	2
Advances in Applied Probability	0.66825	319	349	40	
Operations Research Letters	0.666	324	355	41	2
Journal of Productivity Analysis	0.66502	327	358	42	3
Naval Research Logistics	0.66481	329	360	43	3
INFORMS Journal on Computing	0.66231	342	376	44	3
International Journal of Human-Computer Studies	0.66154	344	378	45	3
Annals of Operations Research	0.65699	359	393	46	2
Applied Statistics: Journal of the Royal Statistical Society Series C	0.6569	360	396	47	
Transportation	0.65667	361	397	48	2
Theory and Decision	0.64877	391	430	49	2
American Statistician	0.64865	393	432	50	
Production Planning & Control	0.64473	417	458	51	3
Interfaces	0.6436	425	468	52	2
Journal of Combinatorial Optimization	0.64295	429	473	53	1
Transport Reviews	0.64287	432	476	54	2
Research Technology Management	0.64176	435	479	55	
Queueing Systems	0.641	440	484	56	
International Journal of Project Management	0.63986	445	490	57	2
Journal of Multivariate Analysis	0.63637	462	507	58	
Computers & Industrial Engineering	0.63589	464	509	59	2
International Journal of Physical Distribution & Logistics Management	0.63518	470	515	60	2
Journal of Manufacturing Systems	0.62404	527	579	61	
Mathematical Methods of Operations Research	0.62394	529	581	62	
Quality & Quantity	0.62121	543	596	63	
Journal of Purchasing and Supply Management	0.62039	546	599	64	2
Journal of Service Management	0.61963	549	602	65	2
Industrial Management and Data Systems	0.61871	553	606	66	1
International Journal of Flexible Manufacturing	0.61455	574	630	67	2
International Journal of Logistics: Research and Applications	0.61056	596	652	68	2
International Journal of Logistics Management	0.60808	609	665	69	2
Journal of Manufacturing Technology Management	0.60211	636	692	70	2
Quality Management Journal	0.60151	638	694	71	
International Transactions in Operational Research	0.59995	642	698	72	2
Journal of Multi-Criteria Decision Analysis	0.59936	648	704	73	
Business Process Management Journal	0.59595	658	714	74	1

International Journal of Quality & Reliability Management	0.5937	669	725	75	2
Total Quality Management & Business Excellence	0.59064	674	730	76	2
International Journal of Manufacturing Technology & Management	0.5893	677	733	77	
Benchmarking: An International Journal	0.57895	704	761	78	1
Knowledge and Process Management	0.57747	707	764	79	1

Note: *1.The ABS ranking scale has four quality ratings ranging from 4 to 1. The 4 category comprises journals that publish the most original and best executed research, the 3 category journals that publish original and well executed research papers and are highly regarded, the 2 category journals that publish original research of acceptable standards and the 1 category journals that publish research of recognized standards. For a full specification of the journal quality grades, see ABS – Academic Journal Quality Guide, Version 4, 2010.*

References

[1] Mingers, J, Harzing, A-W. Ranking journals in business and management: A statistical analysis of the Harzing data set. European Journal of Information Systems. 2007;16(4):303-316.

[2] Willmott, H. Journal list fetishism and the perversion of scholarship: Reactivity and the ABS list. Organization. 2011;18(4):429-442.

[3] Adler, NJ, Harzing, A-W. When knowledge wins: Transcending the sense and nonsense of academic rankings. Academy of Management Learning & Education. 2009;8(1):72-95.

[4] Rowlinson, M, Harvey, C, Kelly, A, Morris, H. The use and abuse of journal quality lists. Organization. 2011;18(4):443-446.

[5] Hult, GTM, Reimann, M, Schilke, O. Worldwide faculty perceptions of marketing journals: Rankings, trends, comparisons, and segmentations. globalEDGE Business review. 2009;3(3):1-23.

[6] Bruton, GD, Lau, C-M. Asian management research: Status today and future outlook. Journal of Management Studies. 2008;45(3):636-659.

[7] Bollen, J, Van de Sompel, H, Hagberg, A, Chute, R. A principal component analysis of 39 scientific impact measures. PLoS ONE. 2009;4(6):e6022.

[8] Hazelkorn, E. Rankings and the reshaping of higher education : The battle for world-class excellence. Houndmills, Basingstoke, UK: Palgrave Macmillan; 2011.

[9] Frey, BS, Rost, K. Do rankings reflect research quality? Journal of Applied Economics. 2010;13(1):1-38.

[10] Benati, S, Stefani, S. The academic journal ranking problem: A fuzzy-clustering approach. Journal of Classification. 2011;28(1):7-20.

[11] Fam, K-S, Shukla, P, Sinha, A, Luk, C-L, Parackal, M, Chai, JCY. Rankings in the eyes of the beholder: A vox populi approach to academic journal ranking. Asian Journal of Business Research. 2011;1(1):1-17.

[12] Cook, WD, Raviv, TAL, Richardson, AJ. Aggregating incomplete lists of journal rankings: An application to academic accounting journals. Accounting Perspectives. 2010;9(3):217-235.

[13] Kao, C, Lin, H-W, Chung, S-L, Tsai, W-C, Chiou, J-S, Chen, Y-L, et al. Ranking Taiwanese management journals: A case study. Scientometrics. 2008;76(1):95-115.

[14] Thomson Reuters. 2011 Journal Citation Reports. Philadelphia: Thomson Reuters; 2012.

[15] Glänzel, W, Moed, HF. Journal impact measures in bibliometric research. Scientometrics. 2002;53(2):171-193.

[16] Leydesdorff, L. Caveats for the use of citation indicators in research and journal evaluations. Journal of the American Society for Information Science and Technology. 2008;59(2):278-287.

[17] Bordons, M, Fernández, MT, Gómez, I. Advantages and limitations in the use of impact factor measures for the assessment of research performance. Scientometrics. 2002;53(2):195-206.

[18] Mahdi, S, D'Este, P, Neely, A. Citation counts: Are they good predictors of RAE scores? A bibliometric analysis of RAE 2001. Bedford: Cranfield University School of Management & Advanced Institute of Management (AIM), 2008.

[19] Bergstrom, CT, West, JD, Wiseman, MA. The Eigenfactor metrics. Journal of Neuroscience. 2008;28(45):11433-11434.

[20] González-Pereira, B, Guerrero-Boteb, VP, Moya-Anegón, Fl. The SJR indicator: A new indicator of journals' scientific prestige. 2009;arXiv:0912.4141v1.

[21] Moed, HF. Measuring contextual citation impact of scientific journals. Journal of Informetrics. 2010;4(3):265-277.

[22] Halkos, GE, Tzeremes, NG. Measuring economic journals' citation efficiency: A data envelopment analysis approach. Scientometrics. 2011;88(3):979-1001.

[23] Jones, MJ, Brinn, T, Pendlebury, M. Journal evaluation methodologies: A balanced response. Omega. 1996;24(5):607-612.

[24] Baum, JAC. Free-riding on power laws: Questioning the validity of the Impact Factor as a measure of research quality in organization studies. Organization. 2011;18(4):449-466.

[25] Albers, S. Misleading rankings of research in business. German Economic Review. 2009;10(3):352-363.

[26] Morris, H, Harvey, C, Kelly, A. Journal rankings and the ABS journal quality guide. Management Decision. 2009;47(9):1441-1451.

[27] Steward, MD, Lewis, BR. A comprehensive analysis of marketing journal rankings. Journal of Marketing Education. 2010;32(1):75-92.

[28] Meredith, JR, Steward, MD, Lewis, BR. Knowledge dissemination in operations management: Published perceptions versus academic reality. Omega. 2011;39(4):435-446.

[29] Bauerly, RJ, Johnson, DT. An evaluation of journals used in doctoral marketing programs. Journal of the Academy of Marketing Science. 2005;33(3):313-329.

[30] Zhou, D, Ma, J, Turban, E. Journal quality assessment: An integrated subjective and objective approach. IEEE Transactions on Engineering Management. 2001;48(4):479-490.

[31] Franke, N, Schreier, M. A meta-ranking of technology and innovation management/entrepreneurship journals. Die Betriebswirtschaft. 2008;2008(2):185-216.

[32] Harzing, A-W. Journal Quality List. 49th ed. http://www.harzing.com, 8 June 2013.

[33] Theußl, S, Reutterer, T, Hornik, K. How to derive consensus among various marketing journal rankings? Journal of Business Research. 2014;67(5):998-1006.

[34] Bancroft, DRE, Gopinath, C, Kovács, ÁM, Rejtö, LdK. A new methodology for aggregating tables: Summarizing journal quality data. Journal of Business Venturing. 1999;14(3):311-319.

[35] Rainer, RK, Miller, MD. Examining differences across journal rankings. Communications of the ACM. 2005;48(2):91-94.

[36] Tse, ACB. Using mathematical programming to solve large ranking problems. Journal of the Operational Research Society. 2001;52(10):1144-1150.

[37] Schulze, GG, Warning, S, Wiermann, C. Zeitschriftenrankings für die Wirtschaftswissenschaften – Konstruktion eines umfassenden Metaindexes. Perspektiven der Wirtschaftspolitik. 2008;9(3):286-305.

[38] Farhangfar, A, Kurgan, LA, Pedrycz, W. A novel framework for imputation of missing values in databases. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans. 2007;37(5):692-709.

[39] Gheyas, IA, Smith, LS. A neural network-based framework for the reconstruction of incomplete data sets. Neurocomputing. 2010;73(16-18):3039-3065.

[40] Schafer, JL, Graham, JW. Missing data: Our view of the state of the art. Psychological Methods. 2002;7(2):147-177.

[41] Durrant, GB. Imputation methods for handling item-nonresponse in practice: Methodological issues and recent debates. International Journal of Social Research Methodology. 2009;12(4):293-304.

[42] Dempster, AP, Laird, NM, Rubin, DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B. 1977;39(1):1-38.

[43] van Buuren, S. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research. 2007;16(3):219-242.

[44] Nielsen, SF. Proper and improper multiple imputation. International Statistical Review. 2003;71(3):593-607.

[45] Paul, C, Mason, WM, McCaffrey, D, Fox, SA. A cautionary case study of approaches to the treatment of missing data. Statistical Methods & Applications. 2008;17(3):351-372.

[46] Breiman, L. Statistical modeling: The two cultures. Statistical Science. 2001;16(3):199-231.

[47] Farhangfar, A, Kurgan, L, Dy, J. Impact of imputation of missing values on classification error for discrete data. Pattern Recognition. 2008;41(12):3692-3705.

[48] Twala, BETH, Jones, MC, Hand, DJ. Good methods for coping with missing data in decision trees. Pattern Recognition Letters. 2008;29(7):950-956.

[49] Hapfelmeier, A, Hothorn, T, Ulm, K. Recursive partitioning on incomplete data using surrogate decisions and multiple imputation. Computational Statistics and Data Analysis. 2012;56(6):1552-1565.

[50] Breiman, L. Random forests. Machine Learning. 2001;45(1):5-32.

[51] Biau, G, Devroye, L, Lugosi, G. Consistency of random forests and other averaging classifiers. Journal of Machine Learning Research. 2008;9:2015-2033.

[52] Hastie, T, Tibshirani, R, Friedman, J. The elements of statistical learning. 2nd ed. New York: Springer; 2009.

[53] Strobl, C, Malley, J, Tutz, G. An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests. Psychological Methods. 2009;14(4):323-348.

[54] Lim, T-S, Loh, W-Y, Shih, Y-S. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Machine Learning. 2000;40(3):203-229.

[55] Niculescu-Mizil, A, Caruana, R. Predicting good probabilities with supervised learning. 22nd International Conference on Machine Learning: ACM, 2005. p. 625-632.

[56] Boström, H. Calibrating random forests. Proceedings of the 2008 Seventh International Conference on Machine Learning and Applications: IEEE, 2008. p. 121-126.

[57] Charnes, A, Cooper, WW, Rhodes, E. Measuring the efficiency of decision making units. European Journal of Operational Research. 1978;2(6):429-444.

[58] Cooper, WW, Seiford, LM, Tone, K. Data envelopment analysis. Second ed. New York: Springer; 2007.

[59] Liu, JS, Lu, LYY, Lu, W-M, Lin, BJY. A survey of DEA applications. Omega. 2013;41(5):893-902.

[60] Liu, JS, Lu, LYY, Lu, W-M, Lin, BJY. Data envelopment analysis 1978–2010: A citation-based literature survey. Omega. 2013;41(1):3-15.

[61] Cook, WD, Tone, K, Zhu, J. Data envelopment analysis: Prior to choosing a model. Omega. 2014;44(0):1-4.

[62] Llamazares, B, Peña, T. Preference aggregation and DEA: An analysis of the methods proposed to discriminate efficient candidates. European Journal of Operational Research. 2009;197(2):714-721.

[63] Cook, WD, Doyle, J, Green, R, Kress, M. Multiple criteria modelling and ordinal data: Evaluation in terms of subsets of criteria. European Journal of Operational Research. 1997;98(3):602-609.

[64] Green, RH, Doyle, JR, Cook, WD. Preference voting and project ranking using DEA and cross-evaluation. European Journal of Operational Research. 1996;90(3):461-472.

[65] Noguchi, H, Ogawa, M, Ishii, H. The appropriate total ranking method using DEA for multiple categorized purposes. Journal of Computational and Applied Mathematics. 2002;146(1):155-166.

[66] Hashimoto, A. A ranked voting system using a DEA/AR exclusion model: A note. European Journal of Operational Research. 1997;97(3):600-604.

[67] Park, KS, Jeong, I. How to treat strict preference information in multicriteria decision analysis. Journal of the Operational Research Society. 2011;62(10):1771-1783.

[68] Osborne, MJ, Rubinstein, A. Bargaining and markets. London: Academic Press; 1990.

[69] Diskin, A, Felsenthal, DS. Individual rationality and bargaining. Public Choice. 2007;133(1-2):25-29.

[70] Wang, Y-M, Chin, K-S. Some alternative models for DEA cross-efficiency evaluation. International Journal of Production Economics. 2010;128(1):332-338.

[71] Whitley, R. Changing governance and authority relations in the public sciences. Minerva. 2011;49(4):359-385.

[72] Clark, T, Wright, M. Reviewing journal rankings and revisiting peer reviews: Editorial perspectives. Journal of Management Studies. 2007;44(4):612-621.

[73] Crookes, PA, Reis, SL, Jones, SC. The development of a ranking tool for refereed journals in which nursing and midwifery researchers publish their work. Nurse Education Today. 2010;30(5):420-427.

[74] DuBois, FL, Reeb, D. Ranking the international business journals. Journal of International Business Studies. 2000;31(4):689-704.

[75] R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/, 2013.

[76] Conversano, C, Siciliano, R. Incremental tree-based missing data imputation with lexicographic ordering. Journal of Classification. 2009;26(3):361-379.

[77] Breiman, L, Friedman, JH, Olshen, RA, Stone, CJ. Classification and regression trees. New York: Chapman and Hall; 1984.

[78] Friedman, JH. Recent advances in predictive (machine) learning. PHYSTAT 2003. Stanford, Calfornia. http://www-spires.slac.stanford.edu/cgi-wrap/getdoc/slac-pub-10321.pdf, 2003.

[79] Loh, W-Y. Improving the precision of classification trees. Annals of Applied Statistics. 2009;3(4):1710-1737.

[80] Therneau, TM, Atkinson, EJ. Introduction to recursive partitioning using the rpart routines. Technical Report Series: Section of Biostatistics, Mayo Foundation, USA. http://www.mayo.edu/hsr/techrpt/61.pdf, 1997.

[81] Yohannes, Y, Webb, P. Classification and regression trees, CART. Washington, D.C.: International Food Policy Research Institute; 1999.

[82] Steinberg, D. Cart: Classification and regression trees. In: Wu, X, Kumar, V, editors. The top ten algorithms in data mining. Boca Raton: CRC Press; 2009. p. 179-201.

[83] Ripley, BD. Pattern recognition and neural networks: Cambridge University Press; 1996.

[84] Kim, H, Loh, W-Y. Classification trees with unbiased multiway splits. Journal of the American Statistical Association. 2001;96(454):598-604.

[85] Loh, W-Y. Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011;1(1):14-23.

[86] Hothorn, T, Hornik, K, Zeileis, A. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics. 2006;15(3):651-674.

[87] Therneau, TM, Atkinson, B, R port by Brian Ripley. Rpart: Recursive partitioning. R package version 4.1-1. http://cran.r-project.org/web/packages/rpart/, 2013.

[88] Piccarreta, R. Classification trees for ordinal variables. Computational Statistics. 2008;23:407-427.

[89] Archer, KJ. Rpartordinal: An R package for deriving a classification tree for predicting an ordinal response. Journal of Statistical Software. 2010;34(7):1-17.

[90] Liaw, A, Wiener, M. Classification and regression by randomForest. R News. 2002;2(3):18-22.

[91] Breiman, L. Bagging predictors. Machine Learning. 1996;24(2):123-140.

[92] Cutler, DR, Thomas C. Edwards, J, Beard, KH, Cutler, A, Hess, KT, Gibson, J, et al. Random forests for classification in ecology. Ecology. 2007;88(11):2783-2792.

[93] Strobl, C, Boulesteix, A-L, Zeileis, A, Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics. 2007;8(25).

[94] Strobl, C, Boulesteix, A-L, Kneib, T, Augustin, T, Zeileis, A. Conditional variable importance for random forests. BMC Bioinformatics. 2008;9:307.

[95] Strobl, C, Hothorn, T, Zeileis, A. Party on! A new, conditional variable importance measure for random forests available in the party package. The R Journal. 2009;1/2:14-17.

[96] Hapfelmeier, A, Hothorn, T, Ulm, K, Strobl, C. A new variable importance measure for random forests with missing data. Statistics and Computing. 2012;(forthcoming).

[97] Liaw, A, Wiener, M. Randomforest: Breiman and Cutler's random forests for classification and regression. R package version 4.6-7. http://cran.r-project.org/web/packages/randomForest/, 2012.

[98] Hothorn, T, Hornik, K, Strobl, C, Zeileis, A. Party: A laboratory for recursive partytioning. R package version 1.0-9. http://cran.r-project.org/web/packages/party/index.html, 2013.

[99] Hothorn, T, Hornik, K, van de Wiel, MA, Zeileis, A. A lego system for conditional inference. The American Statistician. 2006;60(3):257-263.

[100] Schapire, RE, Freund, Y, Bartlett, P, Lee, WS. Boosting the margin: A new explanation for the effectiveness of voting methods. The Annals of Statistics. 1998;26(5):1651-1686.

[101] Reyzin, L, Schapire, RE. How boosting the margin can also boost classifier complexity. In: Cohen, WW, Moore, A, editors. Twenty-Third International Conference on Machine Learning: ACM New York, 2006. p. 753-760.

[102] Brier, GW. Verification of forecasts expressed in terms of probability. Monthly Weather Review. 1950;78(1):1-3.

[103] Cook, WD, Kress, M. A data envelopment model for aggregating preference rankings. Management Science. 1990;36(11):1302-1310.

[104] Cook, WD, Kress, M. An extreme-point approach for obtaining weighted ratings in qualitative multicriteria decision making. Naval Research Logistics. 1996;43(4):519-531.

[105] Stein, WE, Mizzi, PJ, Pfaffenberger, RC. A stochastic dominance analysis of ranked voting systems with scoring. European Journal of Operational Research. 1994;74(1):78-85.

[106] Cook, WD, Kress, M. A linear value function in mixed MCDM problems with incomplete preference data: An extreme point approach. INFOR. 2002;40(4):331-346.

[107] Wang, Y-M, Chin, K-S, Yang, JB. Three new models for preference voting and aggregation. Journal of the Operational Research Society. 2007;58(10):1389-1393.

[108] Oral, M, Kettani, O, Lang, P. A methodology for collective evaluation and selection of industrial R&D projects. Management Science. 1991;37(7):871-885.

[109] Lim, S. Context-dependent data envelopment analysis with cross-efficiency evaluation. Journal of the Operational Research Society. 2012;63(1):38-46.

[110] Wu, J, Liang, L, Zha, Y, Yang, F. Determination of cross-efficiency under the principle of rank priority in cross-evaluation. Expert Systems with Applications. 2009;36(3):4826-4829.

[111] Wu, J, Liang, L, Yang, F. Determination of the weights for the ultimate cross efficiency using Shapley value in cooperative game. Expert Systems with Applications. 2009;36(1):872-876.

[112] Wu, J, Liang, L, Yang, F, Yan, H. Bargaining game model in the evaluation of decision making units. Expert Systems with Applications. 2009;36(3):4357-4362.

[113] Mehrabian, S, Jahanshahloo, GR, Alirezaee, MR, Amin, GR. An assurance interval for the non-Archimedean epsilon in DEA models. Operations Research. 2000;48(2):344-347.

[114] Felsenthal, DS, Diskin, A. The bargaining problem revisited: Minimum utility point, restricted monotonicity axiom, and the mean as an estimate of expected utility. Journal of Conflict Resolution. 1982;26(4):664-691.

[115] Peters, H, Vermeulen, D. WPO, COV and IIA bargaining solutions for non-convex bargaining problems. International Journal of Game Theory. 2012;41(4):851-884.

[116] Zhou, L. The Nash bargaining theory with non-convex problems. Econometrica. 1997;65(3):681-685.

Appendix A: Missing data imputation

As indicated in Sections 4 and 5.2, our data set is comprised of 939 journals scoring in 11 target lists (Table 2) and 14 additional lists (Table 3), while each individual journal may only be scoring in some but not necessarily all of the 25 lists. Let $J = \{1, \ldots, 939\}$ denote the set of journals and $L = \{1, \ldots, 25\}$ comprise the lists in the order of their appearance in Tables 2–3. For convenience, we may interchangeably refer to $j \in J$ as *cases* and to $\ell \in L$ as *variables*.

Let $r_{j\ell}$ ($j \in J$, $\ell \in L$) represent the score of *j*-th journal in ℓ-th list[12] when it is available and let $r_{j\ell} = \mathrm{M}$ when not — what defines a missing value. For the purposes of aggregate journal rating and ranking pursued in this study, missing values have to be imputed in the target lists from 1 to 10 (while it does not need to be done in the 11th target list by its construction as explained in Section 4). Let accordingly $I = \{1, \ldots .10\} \subset L$ comprise the lists which require imputation. For imputation purposes, we further augment the data set with two additional variables no. 26 and 27 which respectively indicate journals' primary and secondary subject areas as per the JQL49 database [32]. If a particular journal is not assigned a secondary subject area in that database, then we let its primary subject area serve as the secondary too. By construction, both variables do not have missing values. Let $V = L \cup \{26, 27\}$ accordingly comprise all of the variables in the data set.

Imputation of missing values is then conducted by taking each single $\ell \in I$ as dependent variable (*response*) and $V \setminus \{\ell\}$ as independent variables (*predictors*), and making inference about the missing values in ℓ from the predictor values using the random forests method. As indicated in Section 5.2, this is accomplished in three basic steps, which we

[12] The score represents the journal's *rank* if the respective list ranks journals on an ordinal scale, and its *rating* if the list rates the journals on an interval or ratio scale.

describe below in detail. All necessary computations have been conducted in R software environment [75] (version 3.0.0).

A.1 Pre-imputation of missing data in predictors

As predictors from L exhibit missing values on their own, we first pre-impute missing data in every $\ell \in L$ while treating $V \setminus \{\ell\}$ as predictors. Following Hastie et al. [52] (see also [76]), we employ classification and regression trees (CART) to accomplish this task.

A.1.1 Overview of CART

Classification and Regression Trees (CART) [77] represents a widely used non-parametric method of *supervised learning* — i.e., learning from data about how do certain input variables (*predictors*) take effect on certain output data (*response*), for the purpose of correctly predicting or estimating the response from the predictors' values [46, 52, 78]. In this context, prediction of a numerical response (measured on an interval or ratio scale) is being termed *regression*, whereas *classification* deals with a categorical response (measured on a nominal sale). The CART method is capable of doing either type of learning and, in addition to that, has been developed with a strong emphasis on the possible data missingness in the predictors. It exhibits at the same time a degree of accuracy comparable with the best of the classical statistical methods [54] while producing highly interpretable models — without requiring to make *a priori* distributional assumptions for the data. Further, it does not require transformations of predictor variables, allows any mixture of the variables, is resistant to the presence of outliers and irrelevant variables, and is fast to train [52, 78]. For these reasons, CART has been adopted in many applied areas and is a most popular predictive learning method used in data mining [52, 78].

CART produces a data model in the form of a binary tree which is grown in the top-down fashion by *recursively partitioning* the data. The construction of the tree (*tree fitting*) proceeds starting from the root node — which is associated with all of the observations contained in the data set. A node is split by selecting a particular predictor variable and partitioning its range into two subsets — which respectively define the left and the right branch descending from that node; the observations attached to that node are accordingly separated into two groups which become associated with the respective child nodes. The choice of the variable and its partition is made in a way that would maximize the efficiency of the split. For a categorical response, this corresponds to the greatest possible reduction of heterogeneity of the response among the observations at the node — called the *node impurity*, for which there are several different measures available [52, 79]. Specifically, a best split achieves the greatest possible impurity reduction — which is measured by averaging the impurity among the two child nodes. In simpler words, the goodness of a split is determined by the extent to which the discrimination between the predictor values helps to discriminate the response. For a numerical response, the node impurity is the sum of squared deviations of the response from its mean value at that node. The branching of nodes continues either until a zero impurity is achieved or until there are only a few observations arrived at a node. The generated tree can then be used to predict the response from the new predictor values: each such case is run down the tree by applying the branching rules generated during tree fitting; the terminal node at which the given case arrives determines the prediction for this case — as the majority value of the response among those observations which have ended up at that node during the construction of a classification tree, and respectively its mean value when doing a regression [52, 53, 80-82].

However, the tree grown to its maximum size may overfit the training data and not generalize well; on the other hand, a too small tree may not capture enough dependencies in the data [52, 53]. *Tree pruning* is accordingly undertaken to strike a balance between the complexity and predictive capability of the tree by successively pruning its branches and estimating its resulting predictive accuracy via *N*-fold *cross-validation* — which proceeds by partitioning the observations in the data set into *N* approximately equal groups and then successively removing each group from the data set, fitting a new fully-grown tree, pruning it to the complexity level in question, and using it to predict the response in the removed observations. Predictions are then compared to the true responses to obtain the *cross-validation error* over all groups of observations. In this way, a sequence of trees of different complexity (ranging from the fully grown tree to the single-node one) is evaluated, and the complexity level with the smallest cross-validation error is ultimately chosen; alternatively, the smallest tree with the error within 1 standard deviation from the minimum can be chosen as well [52, 80-83].

Classification trees further allow to specify a *prior distribution* for the response categories and *misclassification costs matrix* to distinguish between the severity of wrongly classifying response categories; these settings take effect on the evaluation of the node impurity, the prediction at a terminal node and the prediction errors. Furthermore, CART implements a mechanism that flexibly accommodates missing data in the predictors. This is being accomplished by looking for *surrogate variables* at every node split: specifically, after a node split has been produced with a particular predictor variable (*primary splitter*) and its range partition (*split point*), another predictor is being sought with a suitable split point which would most closely mimic the split achieved with the primary splitter at this node; this defines the 1st *surrogate*. In the same way the 2nd best surrogate is being determined, and so on. Then each time when an observation requiring a prediction is lacking the value of the primary splitter at a particular node when being run down the tree, the 1st surrogate will be utilized to properly send this observation further down; if the value of the 1st surrogate is missing as well, then the 2nd surrogate is used, and so on. Hence this mechanism is trying to benefit from the correlations within the data to universally allow missingness while effectively compensating for it [52, 80-82].[13] For the above reasons, CART is suggested to be an ideal choice for the imputations of missing values in the data set [52].

The CART approach to learning has however the following drawbacks: 1) predictions being sharply discontinuous across the individual regions of the predictors' space due to recursive partitioning; 2) instability with respect to small variations in the data and, by that, a high variability of predictions; 3) difficulties in capturing additive structures in the association between the response and the predictors; 4) fragmentation of data — which may cause certain relevant predictors to be disregarded if there are relatively many of them, resulting in a lower accuracy as compared to the best available methods, and 5) potential bias in variable selection towards variables with many distinct realisations and those with many missing values [52, 53, 78, 84]. Still, the above indicated advantages of CART, in particular the high interpretability of the tree models and its non-parametric approach [52, 53, 78], have secured its broad adoption in many applications [46]. A substantial research effort has further been undertaken to address some of the limitations of CART [85, 86].

A.1.2 Application of CART

As indicated above, the CART method has been adopted to pre-impute missing values in the variables $\ell \in L$. We use for this purpose a CART implementation delivered by the R package *rpart* [80, 87] — "the de-facto standard in open-source recursive partitioning software" [86]. Note that the following variables in L represent journal rankings and are therefore measured on an ordinal scale: $L_0 = I \cup \{11+\ell \mid \ell = 1, 2, 5, 6, 7, 8\}$, whereas the remaining variables $L_1 = L \setminus L_0$ rate the journals on an interval or ratio scale.[14]

[13] Notably, CART can also implement node splits which are based on a linear combination of the variables instead of just a single one; in this case, however, the predictors are not allowed to have missing data.

[14] We interpret the scale of the journal ranking list BJM 2004 (see entry no. 3 in Table 3) as an interval one.

Missing values are accordingly imputed for variables in L_1 by means of regression trees. Each $\ell \in L_1$ is successively treated as the response variable with predictors $V \setminus \{\ell\}$, and a regression tree is grown given all such cases j in the data set for which $r_{j\ell} \neq \mathrm{M}$. By the same approach, missing values are imputed for variables in L_0 using classification trees. Regarding the latter, we employ the *rpart*'s default Gini node impurity measure (see also Section A.1.3 below for a further discussion). Following Loh [79], we further specify the costs of misclassifying rank r to rank r' as a loss matrix C with $C_{rr'} = |r - r'|$, to account for the ordinal nature of the response variable. Furthermore, higher ranks in ℓ are typically less populated than middle ranks (e.g. the journals with the highest rank typically represent a small fraction of all journals ranked in ℓ), while their misprediction should be taken more seriously. To account for the underrepresented ranks and thus redistribute the misclassification error between the ranks, we employ case weighting; the weight attached to cases with rank r is taken to be $n_{\ell,r} \cdot (n_\ell - n_{\ell,r})$, where $n_\ell = |\{ j \,|\, r_{j\ell} \neq \mathrm{M}\}$ is the number of cases with non-missing values in ℓ and $n_{\ell,r} = |\{ j \,|\, r_{j\ell} = r\}|$ is the number of those with the value of r. The minimum weight is normalized to unity.

While growing a tree of either kind, we allow as many surrogate variables at node splits as many are present in L apart from the response and the primary splitter. Each tree is initially grown to the maximum depth by setting *rpart*'s control parameter *cp* to 0; splits of nodes with less than 10 observations are not attempted. Then, tree pruning is conducted by means of the 10-fold cross-validation, while we prefer to stick to the tree with the smallest cross-validation error [84]. Finally, the tree is used to predict the response in all cases where its value is missing.

A.1.3 Notes

A more appropriate choice of the node impurity measure for classification trees would be the one that respects the ordinal nature of variables in L_0. The original CART monograph introduces two such ones: *ordered twoing* and *symmetric Gini* [82, 88]. However, none of them is implemented by the *rpart* package. This purpose serves *rpartOrdinal* — an R package by Kellie J. Archer [89] that implements ordered twoing and an ordinal variant of Gini impurity measure suggested by Piccarreta [88]. However, this implementation does not accommodate missing data in predictors [Archer, personal communication]. Further, Twala et al. [48] introduced a novel approach to handling missingness at node splits that has exhibited an excellent performance. However, its implementation has not been available to us. Exploring these options represents an interesting opportunity for the future work.

A.2 Imputations with random forests and accuracy validation

Having completed the data set by means of CART, we now re-impute those values which have been originally missing in the variables $\ell \in I$. These imputations are accomplished by means of random forests — a novel predictive learning method that delivers, among a number of other strong features, a superior predictive accuracy.

A.2.1 Overview of random forests

Random forests [50] represent an *ensemble learning* method in which a number of classification or regression trees (depending on the task) comprise an ensemble that predicts the response as a committee — by the majority principle in classification tasks, or by averaging over the individual predictions of the committee members in regression tasks [52, 90]. Tree-growing in such ensemble involves randomization: firstly, the training data for an individual tree represents an equal-sized *bootstrap sample* of the original data set, obtained by a random draw from the latter with replacement. This approach to building a tree ensemble is known as *bootstrap aggregation*, or *bagging* [91]. As individual trees exhibit a high variability (cf. Section A.1.1), bagging can remarkably improve on their predictive accuracy — by reducing the variance via aggregation of predictions within a tree ensemble [52, 91]. Secondly, in addition to bagging, random forests inject a further randomness to the process of tree growing — by taking only a random selection of predictor variables into consideration when making a node split. This approach helps to reduce correlation between individual trees in the ensemble; if their prediction strength is not restricted by that too far, then this leads to a significant improvement of predictive accuracy of the tree ensemble [50] — what makes random forests "competitive with the best available methods and superior to most methods in common use" [92], [52, 53]. The number of predictors to be selected randomly for node-splitting in classification tasks (our primary concern) is recommended to be $\lfloor \sqrt{m} \rfloor$, where m is the total number of predictors [52, 90]; however, the performance of random forests remains quite insensitive to this choice over a wide range of values and can be excellent with a random selection of just 1 or 2 predictors, either [50, 52, 90]. As random forests benefit from the variability of individual trees, all trees are grown full and thus require no pruning (we refer the reader to [52, 53] for a more detailed discussion of this strategy with regard to the possibility of overfitting).

Apart from having a strong predictive accuracy, random forests offer a built-in measure of the prediction error — the *out-of-bag* (*OOB*) error estimate — which is computed on-the-fly during the construction of the forest and makes the user free from the need to additionally validate the prediction error: since bootstrap sampling leaves out each single case about 36% of all times, the predictions by those trees for which the given case did not enter the bootstrap sample can be aggregated and compared with the true response value — thus producing an estimate of the forest's prediction error rate by averaging over all cases in the data set. As soon as the OOB error rate stabilizes with the growing number of trees, it represents an unbiased estimate of the generalization error [50, 52, 90].

Furthermore, random forests are robust with respect to the noise in the response variable [50] and deliver a number of further advantages: a case proximity measure, outlier detection, clustering, and a novel variable importance measure, among others [92]. At the same time, random forests are particularly fast and easy to train, requiring to fine tune a few parameters only [90], and can effectively deal with a large number of predictor variables — large even when compared to the number of cases in the data set. Being at the same time a non-linear and non-parametric technique, random forests allow application to a wide range of problems, even if they are "nonlinear and involve complex high-order interaction effects" [93]. For the reasons indicated, random forests have gained a fast adoption in many areas since their introduction [53, 93]. We refer the reader for more recent studies of the variable importance measure to Strobl et al. [93-95] and Hapfelmeier et al. [96].

A.2.2 Application of random forests

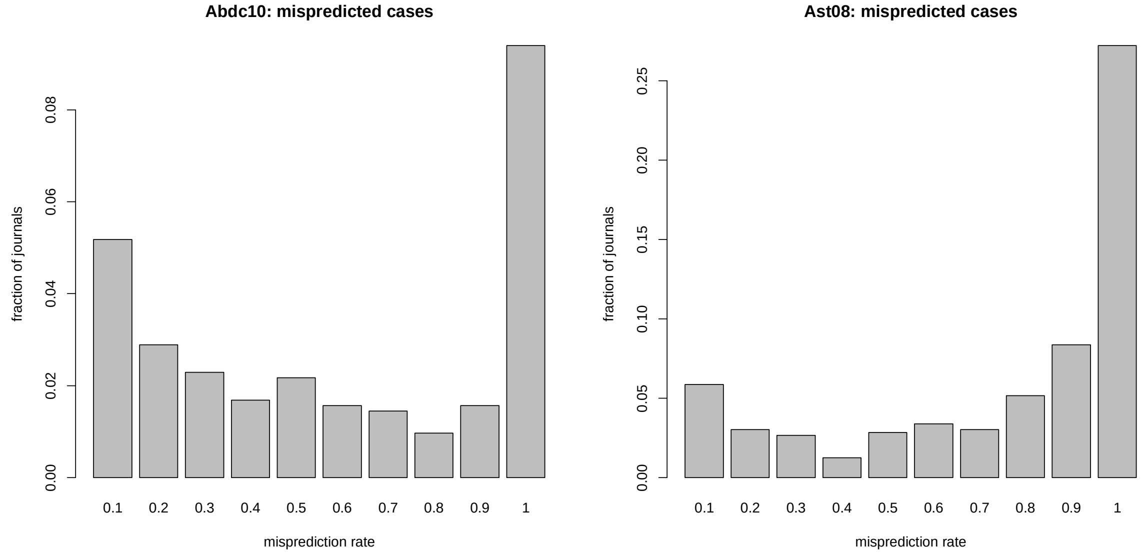
As indicated above, we now re-impute the originally missing values in each of the variables $\ell \in I$ by means of random forests. We use for this purpose an implementation of the method delivered by the R package *randomForest* [90, 97]. For each $\ell \in I$, we utilize variables in $V \setminus \{\ell\}$ as predictors and, in addition to that, we introduce one dummy variable per each predictor with pre-imputed values — indicating whether the respective predictor value is original or imputed. Hence there are each time altogether $m = |V| - 1 + |L| - 1 = 48$ predictor variables. We use the following parameters in constructing the forests: number of trees equal to 500, number of randomly selected predictors equal to $\lfloor \sqrt{m} \rfloor$, and the minimum node size equal to 1 observation. These settings are default for *randomForest*. Furthermore, we utilize *class weights* to balance the misprediction error between the individual ranks by the same approach as we used in Section A.1.2 for case weights.

One of the features of random forests is a novel mechanism for handling missing data in predictors that is based on the random forests' case proximity measure [97]. However, the implementation of the method by the *randomForest* package does not allow for missing data when predicting the response. Mainly for this reason we stick to the strategy of pre-imputing missing values in the predictors by means of CART as described in Section A.1. We cannot thus rely on the forests' built-in OOB error rate to estimate the accuracy of imputations, and have conducted 10-fold cross-validations of prediction error (see Section A.1.1) delivered by the combination CART + *randomForest* in each variable $\ell \in I$. We have repeated these cross-validations 10 times and averaged the misprediction error rates over the trials. Table A.1 presents this average error for each response variable $\ell \in I$, overall and per individual rank. As one can see, the overall error varies remarkably across journal rankings — from about 17% for ABDC 2010 to about 47% for Ast 2008. Figure A.1 shows the fractions of journals which were mispredicted 10% to 100% of times in the ranking lists ABDC 2010 and Ast 2008 during the above 10 trials. Both graphs reveal a pattern characteristic of all 10 ranking lists $\ell \in I$, exhibiting a long bar on the right which indicates that a relatively large fraction of journals consistently cannot be ranked same as they appear in the respective ranking lists — at least with the data underlying the present study.

Table A.1: Rank prediction errors as per 10-fold cross-validations, averaged over 10 trials

Rank	Ast ‘08	ABDC ‘10	ABS ‘10	CNRS ‘11	HEC ‘11	UQ ‘11	VHB ‘11	Cra ‘12	EJL ‘12	ESS ‘13
1	0.315	0.136	0.287	0.407	0.171	0.256	0.300	0.376	0.281	0.119
2	0.357	0.083	0.198	0.447	0.419	0.192	0.525	0.180	0.463	0.747
3	0.575	0.175	0.190	0.335	0.384	0.141	0.292	0.473	0.500	0.140
4	0.570	0.459	0.501	0.412	0.518	0.448	0.384	0.497	0.087	0.348
5	0.808			0.576			0.758			1.000
6							1.000			
Overall:	0.469	0.171	0.245	0.422	0.368	0.212	0.419	0.331	0.210	0.288

Figure A.1: Fractions of journals mispredicted at different rates in 10 cross-validation trials



A.2.3 Notes

An alternative random forest method to use for imputations would be the *conditional inference forests* offered by the R package *party* [98]. These forests are comprised of trees which implement a conditional inference approach to node splitting [86, 99] and can accommodate missing data in the predictors by using surrogate variables as in CART (cf. Section A.1.1). Hence unlike the random forests implementation by the *randomForest* package, conditional inference forests allow for data missingness when predicting the response. They can therefore be applied to impute missing values in the variables $\ell \in I$ without the need to pre-impute missing data in the predictors, and have performed favorably in a series of tests in [49]. Furthermore, they can naturally treat ordinal response variables [86, 98] and offer an unbiased variable importance measure [93-95]. However, the latter issue has not been a concern in our study, thus we chose the combination CART + *randomForest* for a better predictive accuracy. For comparison, Table A.2 shows the overall error exhibited in the repeated 10-fold cross-validations by our method (as per Table A.1) and by the conditional inference forests (which have been run with their default parameter settings[15] and, in addition, with a maximum possible number of surrogate variables allowed at node splits, and case weights assigned as described in Section A.1.2).

Table A.2: Overall rank prediction errors as per 10-fold cross-validations, exhibited by the combination CART + *randomForest* (CART+RF) and by conditional inference forests (CF), averaged over 10 trials

Method	Ast ‘08	ABDC ‘10	ABS ‘10	CNRS ‘11	HEC ‘11	UQ ‘11	VHB ‘11	Cra ‘12	EJL ‘12	ESS ‘13
CART+RF	0.469	0.171	0.245	0.422	0.368	0.212	0.419	0.331	0.210	0.288
CF	0.496	0.407	0.330	0.501	0.435	0.295	0.475	0.332	0.233	0.309

A.3 Actual imputations

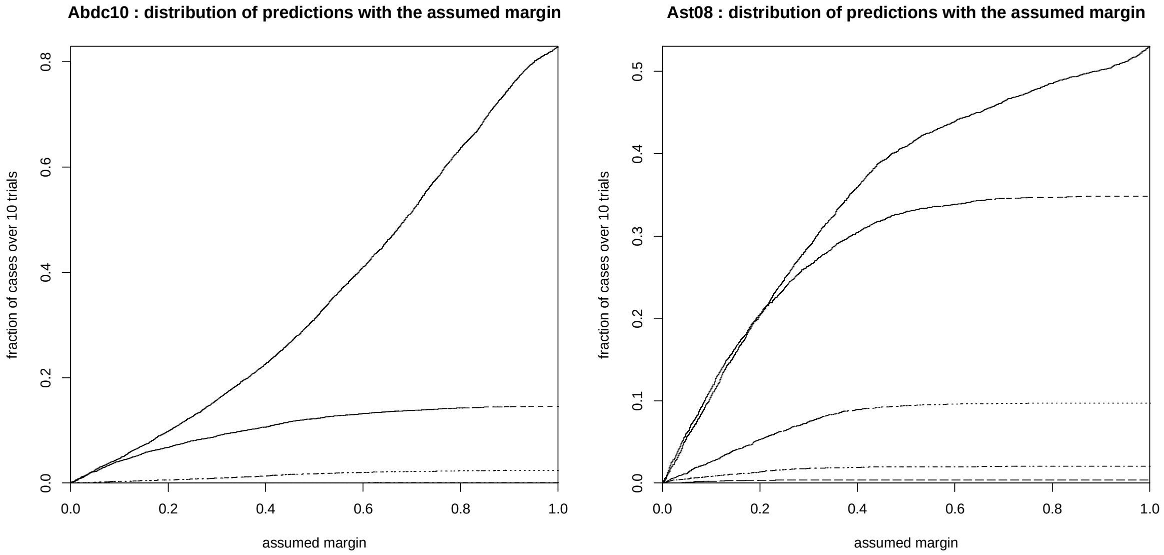
Having tested the accuracy of imputations on the existing rank data, we now turn to actually imputing those values which have been originally missing in the variables $\ell \in I$. Given however the magnitude of the cross-validated error rates reported in Section A.2.2, the accuracy of the predictions to be obtained should be questioned. Random forests predict a categorical response by the majority principle — i.e., by choosing that response category which is being predicted (*voted for*) by the largest fraction of trees in the ensemble. If the true response is actually known for the predicted case (e.g. when dealing with a test data sample), then the difference between the fraction of correct votes and the largest fraction of votes for any other response category defines the *margin* of the prediction delivered by the forest [50,

[15] Since missing values need not be pre-imputed in the predictors when using conditional inference forests, we leave dummy variables out (cf. Section A.2.2) and thus have each time $|V|-1=26$ predictors in the course of imputations. The default number of predictors to be randomly selected for node-splitting is set in the *party* package to 5. This number coincides at the same time with the setting recommended for random forests (cf. Section A.2.1) and we therefore stick to the default.

100]. A positive margin means a correct prediction; the greater its value, the stronger is the confidence in the prediction. The average value of the margin attained on the test data determines how well will the ensemble generalize — i.e., predict the response variable when its true value is unknown [101].

In the latter case, the difference between the largest and the second largest fraction of votes in the ensemble can be assumed to serve as the margin of the prediction. While conducting the cross-validations in 10 independent trials as explained in Section A.2.2, we have tracked how this *assumed margin* is associated with the probability of a true prediction. Figure A.2 displays the results obtained for the ranking lists ABDC 2010 and Ast 2008. The graphs in both panels represent the cumulative frequencies of correct vs. incorrect predictions as follows: the topmost graph shows the frequency of true predictions — i.e., when the category predicted by the largest fraction of votes happens to be the true response; the second graph from above shows the frequency of wrong predictions such that the true response happens to be voted for by the second largest fractions of votes; the third graph from above shows the frequency of wrong predictions such that the true response happens to be voted for by the third largest fractions of votes, and so on. The data sample used to produce these graphs has been comprised of all predictions attempted during the 10 trials of the above cross-validations; consequently, the sample size represents each time a 10-fold of the number of cases originally available in the respective variable. Note that the maximum value attained by the respective topmost graph on the vertical axis represents the overall rate of the true prediction — which is the complement of the overall error rate reported in Table A.1. By examining both panels of Figure A.2, we observe that the first and the second graphs from above almost coincide over the lower ranges of the assumed margin — meaning in particular that for such low margins, the true prediction is as likely to be associated with the largest fraction of votes as with the second largest one. Also for higher values of the assumed

Figure A.2: Cumulative frequency distributions of correct predictions with the assumed margin, over 10 cross-validation trials



Note: In each panel, the *k*-th graph from above represents the cumulative frequency at which the correct prediction is being voted for by the *k*-th largest fraction of trees in the ensemble.

margin, there is still a chance of such misprediction. The graph patterns look similar for all variables $\ell \in I$.

Given this prediction uncertainty, it would not be consistent to stick to the point estimates of the missing journal ranks as predicted by the forests; instead, the uncertainty must be reflected in predicted ranks. We therefore adopt, similarly to [30], a fuzzy rank approach — by letting each journal belong to two or more different ranks within the same ranking list. We accordingly define the rank membership as the probability of the given journal belonging to the respective rank. Notably, random forests provide a built-in estimate for such probability as the fraction of trees predicting the respective rank.

As indicated in Section 5.2, random forests exhibited a superior performance in producing such estimates [55]. The *Brier score* — defined as the mean squared deviation of the predicted rank probabilities from the true ones — is commonly used as the respective quality measure for predictions given in terms of probability estimates [56]. Specifically, let $\ell \in I$ be a particular target list and $J_\ell \subseteq \{j \in J \mid r_{j\ell} \neq \mathrm{M}\}$ represent a subset of journals scoring in the list ℓ. Let R_ℓ denote the number of rank gradations in the ranking list ℓ, and let $f_{jk\ell}$ be the probability estimate for journal $j \in J_\ell$ to belong to rank k in this ranking list — so that $\sum_{k=1}^{R_\ell} f_{jk\ell} = 1$. Let further the respective true probabilities be represented by $c_{jk\ell} \in \{0,1\}$ with $c_{jk\ell} = 1$ if and only if $r_{j\ell} = k$ (i.e., if the true rank of this journal in this ranking list is k). Assume that estimates $f_{jk\ell}$ of rank probabilities have been obtained for all $j \in J_\ell$. They then accordingly determine the Brier score over J_ℓ defined as (cf. [102]):

$$P_\ell = \frac{1}{|J_\ell|} \cdot \sum_{j \in J_\ell} \sum_{k=1}^{R_\ell} (f_{jk\ell} - c_{jk\ell})^2 . \tag{A.1}$$

Although random forests have been shown to deliver a very good performance in terms of the Brier score [55], suggestions have been made in the literature on how to improve that performance by means of *calibration* techniques — which attempt to adjust the predicted probabilities based on the results of predictive learning. In particular, Boström [56] suggested two such techniques for multi-categorical response variables, which we designate as Boström's calibration methods no. 1 and 2 and delineate them below along with their application to our imputation procedure.

Boström's calibration method no. 1 and its application

This method suggests adjusting the predicted rank probabilities $f_{jk\ell}$ for journal j in the ranking list ℓ as follows:

$$\hat{f}_{jk\ell} = \begin{cases} f_{jk\ell} + p \cdot (1 - f_{jk\ell}) & \text{if } k = \arg\max_k \{f_{jk\ell}\} \\ f_{jk\ell} \cdot (1 - p) & \text{otherwise} \end{cases} \tag{A.2}$$

where $p \in [0,1]$ is the calibration parameter. In essence, this method increases the probability estimate for the most probable rank [56] while properly reducing the probability estimates for all other ranks — whose values decrease by $p \cdot 100$ per cent. The optimal value of p is to be determined by replacing $f_{jk\ell}$ in (A.1) with the calibrated rank probabilities $\hat{f}_{jk\ell}$ and minimizing the Brier score P_ℓ as a function of p. The set J_ℓ of journals defines then the *calibration data set*.

We implement this calibration method for each target list $\ell \in I$ by conducting the 10-fold cross-validation of imputations obtained in the respective list by means of the CART+RF method as described in Section A.2.2 and accordingly deriving $f_{jk\ell}$ as the fraction of trees in the respective random forest that predict rank k for journal j. Thus $J_\ell = \{j \in J \mid r_{j\ell} \neq \mathrm{M}\}$ encompasses all journals scoring in the list ℓ. To achieve more balanced results, we utilize the results of all 10 cross-validation trials described in Section A.2.2 by including each journal $j \in J_\ell$ in equation (A.1) 10 times with its true rank probabilities and attaching the rank probabilities estimates from the *t*-th cross-validation trial to the *t*-th instance of that journal ($t = 1,\ldots,10$). To simplify the presentation, we will still refer to the index set of the outer sum in (A.1) as J_ℓ.

The optimal value of parameter p for the target list $\ell \in I$ is then determined as follows (to simplify the presentation, we will below suppress the subscript ℓ in the notation if it remains unambiguous). Let

$$k^*(j) = \arg\max_k \{f_{jk}\}$$

represent the rank of journal j which has been voted for by the majority of trees in the respective random forest; the ties are broken by picking the highest rank. By substituting $\hat{f}_{jk}$ for f_{jk} into (A.1) we accordingly express the Brier score as a function of p as follows:

$$\hat{P}(p) = \frac{1}{|J_\ell|} \cdot \sum_{j \in J_\ell} \left(\left(f_j^* + p(1 - f_j^*) - c_j^* \right)^2 + \sum_{k \neq k^*(j)} \left(f_{jk}(1-p) - c_{jk} \right)^2 \right), \tag{A.3}$$

where $f_j^* = f_{j,k^*(j)}$ and $c_j^* = c_{j,k^*(j)}$. Expressing the first derivative of $\hat{P}(p)$ and setting it equal to zero yields the first-order condition for the optimality of p:

$$\frac{d\hat{P}}{dp} = \frac{2}{|J_\ell|} \cdot \sum_{j \in J_\ell} \left(\left(f_j^* + p(1-f_j^*) - c_j^* \right)(1-f_j^*) - \sum_{k \neq k^*(j)} \left(f_{jk}(1-p) - c_{jk} \right) \cdot f_{jk} \right) = 0 .$$

Solving this equation for p yields:

$$\tilde{p} = \frac{\sum_{j \in J_\ell} \left(\sum_{k \neq k^*(j)} f_{jk}(f_{jk} - c_{jk}) - (f_j^* - c_j^*)(1-f_j^*) \right)}{\sum_{j \in J_\ell} \left(\sum_{k \neq k^*(j)} f_{jk}^2 + (1-f_j^*)^2 \right)} = \frac{\sum_{j \in J_\ell} \left(\sum_{k=1}^{R_\ell} f_{jk}(f_{jk} - c_{jk}) + c_j^* - f_j^* \right)}{\sum_{j \in J_\ell} \left(\sum_{k=1}^{R_\ell} f_{jk}^2 + 1 - 2 f_j^* \right)} .$$

With $r_j \in \{1, \ldots, R_\ell\}$ representing the true rank of journal j in the ranking list ℓ, $c_{j,r_j} = 1$ holds true, and therefore the above solution can be rewritten in the following form:

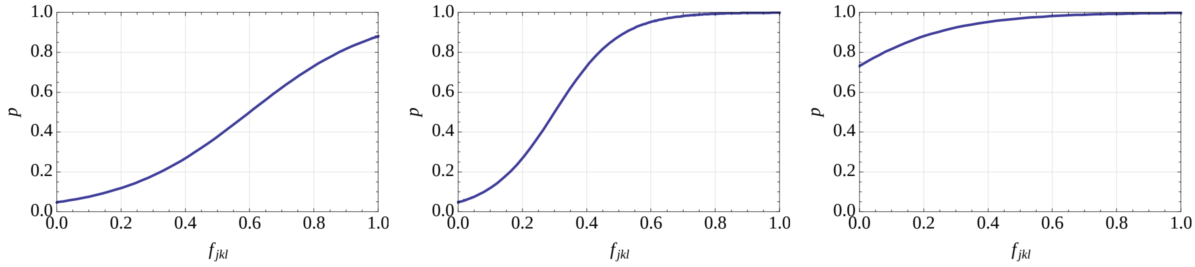
$$\tilde{p} = \frac{\sum_{j \in J_\ell} \left(\sum_{k=1}^{R_\ell} f_{jk}^2 + c_j^* - f_j^* - f_{j,r_j} \right)}{\sum_{j \in J_\ell} \left(\sum_{k=1}^{R_\ell} f_{jk}^2 + 1 - 2 f_j^* \right)} . \tag{A.4}$$

It is straightforward to verify that the second derivative of (A.3) is positive, what proves $\hat{P}(p)$ to be strictly convex and thus $\tilde{p}$ to be its global minimizer. The positivity of the second derivative implies at the same time that the denominator in (A.4) is positive. Furthermore, it is easy to see that the numerator in (A.4) is no greater than the denominator, what proves $\tilde{p} \le 1$ to hold. It may however happen that $\tilde{p} < 0$ — in which case $p^* = 0$ is the global minimizer of $\hat{P}(p)$ on the feasible range $0 \le p \le 1$ by the strict convexity. Thus,

$$p^* = \max\{0, \tilde{p}\}$$

is the optimal value of the calibration parameter for the ranking list $\ell \in I$. This value has to be used in equation (A.2) to adjust the rank probability estimates after imputing them for journals $j \in J \setminus J_\ell$ with missing rank data.

Figure A.3: Graph of the calibration function under different parameter values



(a) $A=-5, B=3$ (b) $A=-10, B=3$ (c) $A=-5, B=-1$

Boström’s calibration method no. 2 and its application

This method suggests to replace the constant calibration parameter p with a non-decreasing function of probability $f_{j,k^*(j),\ell}$. We follow Boström [56] and use a sigmoid function which we define as

$$p(f_{jk\ell}) = \frac{1}{1+e^{A\cdot f_{jk\ell}+B}}, \qquad \text{(A.5)}$$

where A and B are function parameters subject to optimization. The function is non-decreasing if and only if $A \leq 0$ and has its values within the range $(0,1]$. We stick to Boström’s [56] grid search approach for determining (sub-)optimal values of parameters A and B. To determine the ranges for the parameter values over which the search should be performed, it is instrumental to observe the following properties of $p(f_{jk\ell})$:

- a higher absolute value of A leads to a steeper initial increase of the function (cf. panels **a** and **b** in Figure A.3), with reasonable values of A being found in the range $-100 \leq A \leq 0$;
- pushing B in the negative direction makes the graph start from a higher ordinate (cf. panels **a** and **c** in Figure A.3), with reasonable values of B being found in the range $-5 \leq B \leq 0$;
- pushing B in the positive direction lowers the starting point of the graph and leads to a longer interval of a slower initial decrease of the function (cf. panels **a** and **c** in Figure A.3), with reasonable values of B being found in the range $0 \leq B \leq |A|+5$.

Table A.3: Brier score before and after calibration with Boström's methods 1 and 2

Brier score	Ast '08	ABDC '10	ABS '10	CNRS '11	HEC '11	UQ '11	VHB '11	Cra '12	EJL '12	ESS '13
Uncalibrated	0.5513	0.2662	0.3642	0.5287	0.4642	0.3331	0.5479	0.4541	0.3019	0.3940
With method 1	0.5513	0.2545	0.3573	0.5287	0.4641	0.3218	0.5475	0.4533	0.3009	0.3937
With method 2	0.5511	0.2489	0.3563	0.5287	0.4635	0.3202	0.5476	0.4533	0.3009	0.3929

We then proceed with each ranking list $\ell \in I$ by evaluating the Brier score for each combination of $A \in \{-100, \ldots, 0\}$ and $B \in \{-5, \ldots, |A| + 5\}$ over the sample of journals produced in 10 cross-validation trials as explained above for the calibration method no. 1, while replacing p with $p(f_{j,k^*(j),\ell})$ in equation (A.2) and $f_{jk\ell}$ with $\hat{f}_{jk\ell}$ in equation (A.1). The minimum value of the Brier score determines then the best combination of parameter values A and B for that ranking list. These values have to be used in equation (A.5) to adjust the imputed rank probability estimates for journals $j \in J \setminus J_\ell$ by means of equation (A.2) with $p(f_{j,k^*(j),\ell})$ in place of p.

Of the above two calibration methods, the second one offers more flexibility in calibrating rank probabilities, however at the expense of more intensive computations; furthermore, in contrast to the first method, the second one obtains very likely only a sub-optimal solution. Table A.3 shows the Brier score before and after calibration with the above two methods, as measured over the sample of journals constructed in the course of 10 cross-validation trials as explained above in detail for the calibration method no. 1. As one can see from the table, calibration can offer only a marginal improvement to the Brier score attained by random forests — what confirms the insight by Niculescu-Mizil and Caruana [55]. We can also see that calibration method no. 2 exhibits a slightly better performance (which is never worse than the performance of the first method).

We accordingly perform the calibration of the imputed rank probabilities in all of the ranking lists $\ell \in I$. The actual imputations are then conducted in the list ℓ as follows. Having pre-imputed all missing journal ranks in the data set by means of CART as explained in Section A.1.2, we utilize the subset $J_\ell = \{ j \in J \mid r_{j\ell} \neq \mathrm{M} \}$ of all journals scoring in the list ℓ as the training data set with the response variable ℓ to construct a random forest as explained in Section A.2.2, and then use this random forest to impute missing rank data for all journals $j \in J \setminus J_\ell$. We then derive the rank probability estimates $f_{jk\ell}$ for these journals as fractions of trees in the random forest that predict rank k for journal j in the given list. The imputation is repeated over 10 trials, and the values of $f_{jk\ell}$ are averaged over these 10 trials. They get finally adjusted to $\hat{f}_{jk\ell}$ by means of formula (A.2) where $p(f_{j,k^*(j),\ell})$ supplants p and the best combination of parameter values A and B as determined by the calibration method no. 2 is utilized in equation (A.5).

This completes the imputation of missing rank data in the ranking lists $\ell \in I$. Consider now the set of all target lists $T = I \cup \{11\}$. By construction, the 11th list does not have missing values (cf. Section A). Thus the set of journals J and the set of variables T comprise a complete data set with $\hat{f}_{jk\ell}$ representing the membership grade of journal j to rank k in the list ℓ, where $\hat{f}_{jk\ell} := c_{jk\ell}$ if the respective journal has been originally scoring in the respective list, and is derived by the above described imputation procedure otherwise. This complete data set is then subjected to DEA for the purpose of producing an aggregate journal rating and ranking.

Appendix B: Aggregate rating by data envelopment analysis

This appendix details the aggregate rating procedure by means of data envelopment analysis delineated in Section 6. We will utilize notation introduced throughout Appendix A, with the following modification. As explained in Section 6, we exclude from the entire set of journals J those ones which have original ranks available in less than 25% of the 11 target lists $\ell \in T$. To simplify the exposition, we accordingly re-define the set of journals J as follows:

$$J = \left\{ j \in \mathbb{N} \,\middle|\, 1 \le j \le 939 \;\wedge\; \sum_{\ell \in T} 1_{r_{j\ell} \neq \mathrm{M}} \ge 0.25 \cdot |T| \right\},$$

where 1_C is the indicator function of condition C taking on the value of 1 if C holds true and otherwise 0. As indicated in Section 6, this reduces the number of journals in J from 939 to 786, representing around 84% of all journals in JQL49.

B.1 DEA model and cross-evaluation

The DEA approach adopted in the present work to produce an aggregate rating of journals in J is based on the approach suggested by Green, Doyle and Cook [64] for aggregation of voters' preferences over a set of candidates defined in the form of preference orders. Their work is in turn based on the seminal work by Cook and Kress [103]. The reader is referred to [62] for a recent critical review of this and the follow-up research. In the present work we introduce a number of modifications to the approach of Green et al. [64] which will be explained subsequently in detail.

Consider a journal $j \in J$ whose grade of membership to the rank $k \in \{1,\dots,R_\ell\}$ in the ranking list $\ell \in T$ is given by $\hat{f}_{jk\ell}$. In the spirit of DEA [64, 65, 103], this journal is given the opportunity to determine rank weights $w_{k\ell}$ ($k \in \{1,\dots,R_\ell\}$, $\ell \in T$) that would maximize its own *rating* defined in terms of the weighted average rank:

$$\theta_{jj} := \max_{w_{k\ell}} \sum_{\ell \in T} \sum_{k=1}^{R_\ell} w_{k\ell} \hat{f}_{jk\ell} \tag{B.1}$$

subject to the constraints:

$$\sum_{\ell \in T} \sum_{k=1}^{R_\ell} w_{k\ell} \hat{f}_{ik\ell} \leq 1 \qquad \forall i \in J \tag{B.2}$$

$$w_{k\ell} - w_{k+1,\ell} \geq w_{k+1,\ell} - w_{k+2,\ell} \qquad \forall \ell \in T,\ k = 1, \ldots, R_\ell - 2 \tag{B.3}$$

$$w_{\ell, R_\ell - 1} - w_{\ell, R_\ell} \geq \varepsilon \qquad \forall \ell \in T \tag{B.4}$$

$$w_{\ell, R_\ell} \geq \varepsilon \qquad \forall \ell \in T \tag{B.5}$$

Rank weight $w_{k\ell}$ can be interpreted as the "worth of being ranked in [k]th place" in the ranking list ℓ [104] or the "importance accorded [a journal] that is ranked in [k]th place" in the ranking list ℓ [63]. Constraints (B.2) represent the usual DEA constraints expressing in their left-hand sides the respective rating score of each individual journal in J under the rank weights chosen by the given journal j and therefore requiring that none of the journals can attain a score higher than 1. Constraints (B.3)–(B.5) are the weak convexity constraints imposed on the rank weights in each ranking list $\ell \in T$ which essentially require that the difference between two consecutive ranks expressed in terms of their weights is at least as large as the difference between two respectively lower consecutive ranks. Including convexity constraints in the DEA model of Green et al. [64] has been suggested by Noguchi et al. [65]; however, their variant of convexity constraints has received criticism from Llamazares and Peña [62] whose argument we share and therefore add constraints (B.3)–(B.5) to the original model of Green et al. [64] in the weak form as advocated by Hashimoto [66] and Stein et al. [105]. These constraints also ensure that rank weights are nonnegative and non-decreasing from the lowest rank (R_ℓ) to the highest (1) in each ranking list $\ell \in T$. The nonnegative constant ε represents the *rank discrimination threshold* which particularly determines the minimum amount by which the weights of any two consecutive ranks have to differ.

The choice of the rank discrimination threshold has received a substantial discussion in the literature [see e.g. 62, 64, 65, 67, 103, 106] as it is likely to severely affect the results produced. However, to our best knowledge, no universal and satisfactory solution has been suggested. Specifically, Cook and Kress [103] suggested to use the maximum possible value of ε, however their approach has been invalidated by Green et al. [64] as infringing on the fundamental principle of DEA. They have in turn suggested to use $\varepsilon = 0$ — what has been criticized by Noguchi et al. [65] as contradicting the basic purpose of ranking — the argument which we share as well (see also [67]). As a remedy for this problem, Noguchi et al. [65] have suggested their own formula for calculating the value of ε — which has however been criticized for its arbitrariness (see [107]). We share this criticism and adopt in the present work a novel game-theoretical approach to determining the value of ε which is presented in detail in Section B.2 below.

Note that model (B.1)–(B.5) represents a further departure from the approach adopted in [64-66, 103] in the following two important aspects. Firstly, following [63, 104, 106], we keep the voters' preferences (in our case rank membership grades $\hat{f}_{jk\ell}$) in their disaggregate form — what is justified by the existence of different number of ranks in different ranking lists and different meaning attached to them. In contrast, the reference model of Green et al. [64] would aggregate the membership grades of the given journal to the given rank across all ranking lists. Secondly, our model allows for fuzzy rank memberships by letting each journal belong to two or more ranks in each ranking list — to accommodate the uncertainty associated with imputations of missing rank data (cf. Section A.3). As a result, (B.1)–(B.5) comprise a linear optimization problem with 51 variables and 837 constraints.

Following Green et al. [64], the aggregate rating list of journals $j \in J$ is then derived by means of model (B.1)–(B.5) and using the *cross-evaluation* approach as follows. As previously indicated, solving the model with a fixed $j \in J$ gives the journal j the opportunity to determine the most favorable values $w_{k\ell}^{(j)}$ for rank weights $w_{k\ell}$ to accord itself the highest possible rating score. This *self-rating* of journal j is accordingly denoted by θ_{jj} as per (B.1). At the same time, rank weights $w_{k\ell}^{(j)}$ determine the rating scores of all journals $i \in J$ from the perspective of the journal j when being substituted for $w_{k\ell}$ in the left-hand sides of constraints (B.2). Denote these rating scores by θ_{ji} — i.e.,

$$\theta_{ji} = \sum_{\ell \in T} \sum_{k=1}^{R_\ell} w_{k\ell}^{(j)} \hat{f}_{ik\ell} . \tag{B.6}$$

Solving now model (B.1)–(B.5) successively with each single $j \in J$, we let the journals in this way *cross-evaluate* each other and produce by that the *cross-evaluation matrix* $\Theta = (\theta_{ji})_{i,j \in J}$. By construction, its i-th column contains the self-rating θ_{ii} of the i-th journal in the row i along with its *peer-ratings* θ_{ji} in the remaining rows. The aggregate rating A_i of journal $i \in J$ is accordingly derived as the arithmetic mean of the i-th column in Θ [cf. 70]:

$$A_i = \frac{1}{|J|} \cdot \sum_{j \in J} \theta_{ji} . \tag{B.7}$$

Note that if problem (B.1)–(B.5) has multiple optimal solutions then picking an arbitrary one would lead to arbitrariness in deriving the peer-ratings θ_{ji} of journals $i \in J \setminus \{j\}$ by means of (B.6) [64]. For this reason we employ a secondary goal in determining rank weights that puts the *aggressive* form of cross-evaluation into effect. By its virtue, each journal $j \in J$ picks such optimal solution of problem (B.1)–(B.5) which is the least beneficial one for all other journals in their totality — so that each journal $j \in J$ is given the opportunity to appear most strongly against its peers. This is achieved by first solving problem (B.1)–(B.5) in order to obtain the optimal objective value θ_{jj}, and then solving another linear program with the objective function

$$\min_{w_{k\ell}} \sum_{\substack{i\in J,\\ i\neq j}} \sum_{\ell\in T} \sum_{k=1}^{R_\ell} w_{k\ell}\, \hat{f}_{ik\ell} \tag{B.8}$$

and constraints (B.2)–(B.5) where constraint (B.2) is replaced for $i = j$ with:

$$\sum_{\ell\in T} \sum_{k=1}^{R_\ell} w_{k\ell}\, \hat{f}_{jk\ell} = \theta_{jj} \,.$$

An optimal solution of this linear program defines then rank weights $w_{k\ell}^{(j)}$ which have to be substituted into equation (B.6) for obtaining peer-ratings θ_{ji} [cf. 64, 70, 108]. Note that an alternative form of cross-evaluation (the *benevolent* one) would replace the objective in (B.8) with maximization, what we however find less suitable for the purposes of journal ranking.[16]

The aggregate rating scores A_i accordingly comprise the ultimate rating list of journals $i \in J$ and further determine their aggregate ranking, on which both Section 6 provides further details.

Notes

Cross-evaluation is deemed a powerful extension of DEA and has attracted much interest in research and application, being praised for its ability to rank order the subjects (in our case journals) — a capability not offered by DEA per se [70, 109, 110]. Several different approaches to cross-evaluation have received discussion in the literature. In particular, an alternative approach suggested by Green et al. for deriving aggregate rating scores A_i is the eigenvector method [64, p. 467] which would compute a weighted average in (B.7) by giving more weight to higher rated journals $j \in J$. This would however treat different journals unequally, whereas the arithmetic means approach adopted in (B.7) allows journals have an equal say in determining the final result. Wu et al. [111] have on the other hand pointed out that simple averaging must not necessarily be Pareto optimal, what raises concerns with

[16] The reader is referred to [70] for an overview of these and other possible formulations of the secondary goal discussed in the literature.

regard to its acceptability from the individual subjects' perspective. As a remedy for this issue they have suggested a game-theoretic approach to averaging the rating scores with weights determined via the subjects' Shapley value in a coalitional game. Adopting their approach would however render computations in our setting intractable; we therefore maintain the commonly adopted aggregation approach as per (B.7) while letting journals determine the final outcome in a cooperative fashion by choosing the rank discrimination threshold via *n*-person Nash bargaining — as explained in Section B.2 below. Furthermore, Wu et al. [110] pursued maximization of the subject's ranking position as the secondary goal in cross-evaluations. While likely being computationally prohibitive in our setting, their approach represents an interesting opportunity for the future research. Our approach to obtaining the rank discrimination threshold in Section B.2 is interrelated with theirs in that we express a journal's utility in the bargaining game via its relative standing in the rating list. We refer the reader to [70] for a recent overview of other existing approaches to cross-evaluation which should not be treated here in a greater detail.

As a final note, we have conducted all computations necessary in this appendix using MATLAB (version 7.11.0) and its Optimization toolbox (version 5.1).

B.2 Determining the rank discrimination threshold

As discussed in Section B.1 above, the problem of determining a proper value for the rank discrimination threshold (denoted by ε in model (B.1)–(B.5)) has received a substantial discussion in the related literature, however none of the approaches suggested have proven to be suitable for the purposes of the present work (cf. Section B.1). The analysis of aggregate rating scores in our setting has revealed that the choice of the rank discrimination threshold affects different journals differently; in particular, increasing the value of ε improves the relative standing of some journals in the resulting rating list while worsening that of the others. Hence setting the value of ε exogenously would inevitably lead to arbitrariness in the results produced. To avoid such arbitrariness, we propose a novel game-theoretic approach to choosing the rank separation threshold that lets the journals in J jointly determine the value of ε via bargaining. We model the respective bargaining situation in terms of *n*-person Nash bargaining problem [68] and determine its outcome as follows below.[17]

In the first step, we obtain the maximum value $\varepsilon_{\max}$ for the rank separation threshold under which problem (B.1)–(B.5) is still feasible. This is achieved by solving the following linear program (cf. [113]):

$$\max \varepsilon \qquad \text{s.t. (B.2)–(B.5)}$$

whose optimal solution defines the feasible range $[0, \varepsilon_{\max}]$ for the rank separation threshold. We have found $\varepsilon_{\max}$ to amount to approximately 0.01961 in our setting.

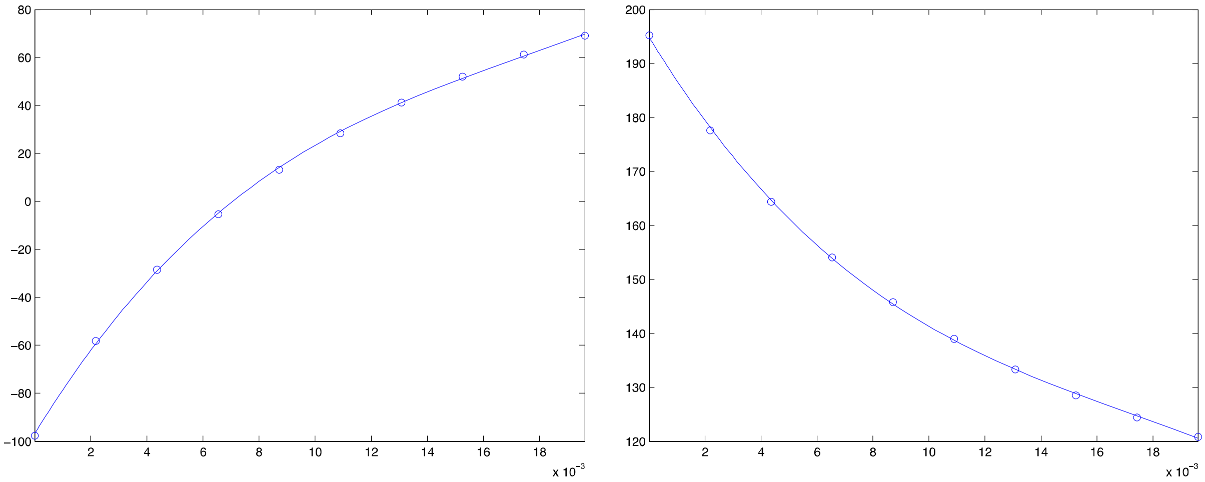
In the next step we obtain the utility functions $u_j(\varepsilon)$ which should respectively represent the utility that the journal $j \in J$ extracts from a particular value $\varepsilon \in [0, \varepsilon_{\max}]$. Let $A_i(\varepsilon)$, $i \in J$, denote the aggregate rating score (B.7) of journal i produced by the approach described in Section B.1 with the given value of ε. We accordingly define the utility $u_j(\varepsilon)$, $j \in J$, as the *standing* of the journal j in the aggregate rating list comprised of rating scores $A_i(\varepsilon)$, $i \in J$:

$$u_j(\varepsilon) = \frac{\sum_{i \in J} \left(A_j(\varepsilon) - A_i(\varepsilon) \right)}{\max_i A_i(\varepsilon) - \min_i A_i(\varepsilon)}. \tag{B.9}$$

In simpler words, a journal's standing represents its position relative to the average journal on the list, normalized by the length of the rating scale. Note that the normalization is necessary since different values of ε lead to different lengths of the rating scale.

[17] In a similar fashion, Wu et al. [112] consider a cooperative game approach in which the subjects (in our case journals) seek to determine a set of common weights (in our case rank weights) via *n*-person Nash bargaining, and utilize the weights so obtained to calculate each subject's rating score.

Figure B.1: Fitting the utility function to data points for two exemplary journals



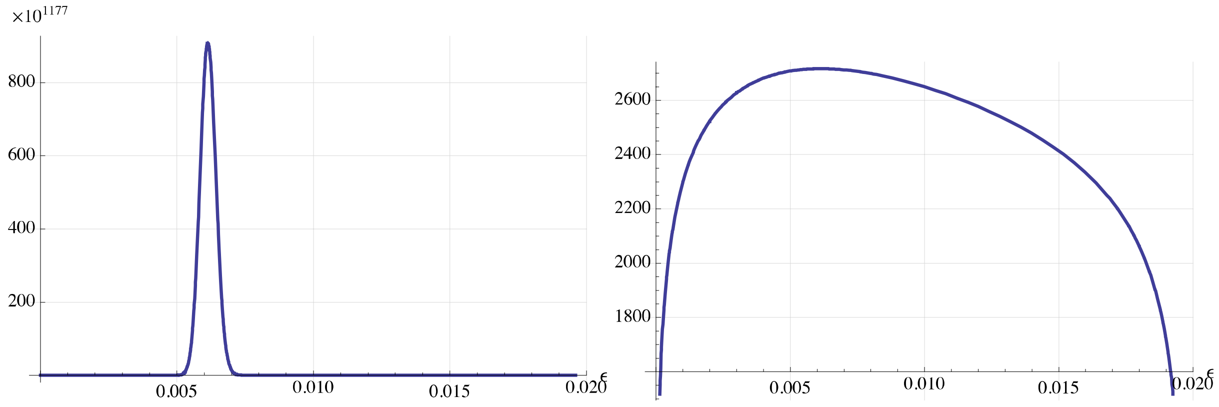
(a) Abacus

(b) Academy of Management Annals

We then repetitively produce the aggregate rating list at each of 10 equally spaced values of ε ranging from 0 to $\varepsilon_{\max}$. This yields, for each $j \in J$, a series of 10 data points that represent the utility function $u_j(\varepsilon)$ by virtue of (B.9). This function finally gets interpolated on the entire feasible range $[0, \varepsilon_{\max}]$ by fitting a cubic polynomial $\hat{u}_j(\varepsilon)$ to these 10 data points using the least squares method. The polynomial approximations obtained in this way for the utility functions of the individual journals $j \in J$ exhibit a strong fit with the data. Figure B.1 illustrates this on the example of the first two journals in J. The mean absolute percentage error of fitting comes above 1% for only two out of 786 journals, and has the maximum value of 1.65%. The maximum absolute percentage error comes above 1% for 11 our 786 journals, and has the maximum value of 3.53%. In all of the latter cases, however, the respective absolute error is negligibly low (below 0.03). We adopt for these reasons the polynomial functions $\hat{u}_j(\varepsilon)$ as an excellent representation of utility functions of the respective journals $j \in J$.

In the next step, we describe the bargaining situation in terms of the *n*-person bargaining problem [cf. 68] with $n = |J| = 786$ and the *bargaining set*

Figure B.2: Plot of the Nash product (left) and its natural logarithm (right) as a function of the rank discrimination threshold



$$U = \left\{ \left(\hat{u}_j(\varepsilon) \right)_{j \in J} \,\middle|\, 0 \le \varepsilon \le \varepsilon_{\max} \right\} \subset \mathbb{R}^n ,$$

while the *disagreement point* is set to be

$$d = \left(\min_{0 \le \varepsilon \le \varepsilon_{\max}} \hat{u}_j(\varepsilon) \right)_{j \in J} \in \mathbb{R}^n .$$

Note that the bargaining set is comprised of all *n*-dimensional utility vectors induced by the feasible values of the rank discrimination threshold, whereas the disagreement point is the vector whose elements represent the minimum utilities possible for the respective journals [cf. 69, 114]. Note further that the bargaining set is by construction connected and closed while being at the same time non-convex. Replacing it with its convex hull as in classical Nash bargaining [68] does not however prove to be a satisfactory approach in our setting because the nature of the given bargaining game does not allow its players (i.e., journals $j \in J$) to treat a lottery over U — or, equivalently, a randomized choice of ε — as a viable bargaining outcome [cf. 115]. In simpler words, the journals cannot be assumed to be expected utility maximizers in the given bargaining game; instead, they maximize their utilities $\hat{u}_j(\varepsilon)$ induced by a deterministic choice of ε — which therefore should represent the bargaining outcome.

Building on Zhou's [116] generalization of Nash bargaining to non-convex problems, we determine in the final step the bargaining outcome $\hat{\varepsilon}$ by maximizing the Nash product

$$\prod_{j\in J}\left(\hat{u}_j(\varepsilon)-d_j\right) \tag{B.10}$$

as a function of ε on the feasible range $[0,\varepsilon_{\max}]$. Note that (B.10) happens to be a polynomial of degree 3^{786} whose graph is plotted in Figure B.2. The optimal solution $\hat{\varepsilon}$ can be efficiently obtained by maximizing the natural logarithm of the Nash product with the Newton method and has been found to amount approximately to 0.0061322 — what comprises about 31.3% of its maximal possible value. The value of $\hat{\varepsilon}$ has been then accordingly utilized as the rank discrimination threshold to produce the aggregate rating by means of the approach presented in Section B.1 above.