Journal of Informetrics 16 (2022) 101284

Contents lists available at ScienceDirect

Journal of Informetrics

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

Is it all bafflegab? – Linguistic and meta characteristics of research articles in prestigious economics journals

Julian Amon*, Kurt Hornik

Institute for Statistics and Mathematics, Wirtschaftsuniversität Wien, Welthandelsplatz 1, Vienna 1020, Austria

ARTICLE INFO

Keywords:
Research impact
SJR indicator
NLP
Readability
Gradient boosting
GLMLSS

ABSTRACT

In competitive research environments, scholars have a natural interest to maximize the prestige associated with their scientific work. In order to identify factors that might help them address this goal more effectively, the scientometric literature has tried to link linguistic and meta characteristics of academic papers to the associated degree of scientific prestige, conceptualized as cumulative citation counts. In this paper, we take an alternative approach that instead understands scientific prestige in terms of the rankings of the journals that the articles appeared in, as such rankings are routinely used as surrogate research quality indicators. For the purpose of determining the most important drivers of suchlike prestige, we use state-of-the-art text mining tools to extract 344 interpretable features from a large corpus of over 200,000 journal articles in economics. We then estimate beta regression models to investigate the relationship between these predictors and a cross-sectionally standardized version of *SCImago Journal Rank (SJR)* in multiple topically homogeneous clusters. In so doing, we also reinvestigate the *bafflegab theory*, according to which more prestigious research papers tend to be less readable, in a methodologically novel way. Our results show the consistently most informative predictors to be associated with the length of the paper, the span of coreference chains in its full text, the deployment of a personal and moderately informal writing style, the "density" of the article in terms of sentences per page, international and institutional collaboration in research teams and the references cited in the paper. Moreover, we identify various linguistic intricacies that matter in the association between readability and scientific prestige, which suggest this relationship to be more complicated than previously assumed.

1. Introduction

For any scholar, reporting research results in a concise and easily comprehensible fashion constitutes a fundamental goal within the scientific process, ideally set out to facilitate both the dissemination of knowledge and academic progress in the field in general. At the same time, in the process of publication, researchers are also incentivized to have their manuscripts accepted in the most prestigious journal possible, as the reputation of the publication outlet serves as an important research quality indicator in an increasingly competitive scholarly environment. Interestingly, according to the *bafflegab theory* (Armstrong, 1980a; 1989; Hartley et al., 1988), those two goals might be at odds. The theory posits that scientists can actually gain prestige and, ultimately, impact by writing in a less readable manner, so much so that "an obtuse writing style" (Seligman, 1986) may even be beneficial in the publication process. A common explanation for this is that the readers' idea of scientific writing tends to be at variance with the type of writing that scores

* Corresponding author.
E-mail address: julian.amon@wu.ac.at (J. Amon).

https://doi.org/10.1016/j.joi.2022.101284
Received 15 March 2021; Received in revised form 7 March 2022; Accepted 23 March 2022
Available online 5 April 2022

highly on various readability measures (short words and sentences), particularly when the primary audience of the paper is destined to be academically inclined.

In order to examine the degree to which this is true, the particular role of linguistic characteristics, such as readability (Didegah and Thelwall, 2013; Gazni, 2011; Van Wesel et al., 2014), syntactic and lexical complexity (Lu et al., 2019a) or the use of keywords (Uddin and Khan, 2016), in determining scientific impact has been empirically studied by various contributions to the literature. Other authors have also investigated the role of meta-level article properties, such as the number of involved authors, institutions or countries (Didegah and Thelwall, 2013; Gazni and Didegah, 2011) or the number and quality of references (Boyack and Klavans, 2005), yet all these studies only concentrate on a relatively small and/or isolated set of features. They typically apply standard statistical tools such as correlation tests or linear regression to study the relationship of these features and scientific impact, notably operationalized in terms of subsequent citation counts.

On an intuitive level, the accumulation of citations certainly does positively contribute to the *scientific prestige* of an article, abstractly defined in the sense of Hartley et al. (1988) and Armstrong (1989). However, we argue that the ranking of the journal that the manuscript appears in – more precisely, its *SCImago Journal Rank (SJR)* – is a more immediate measure and, thus, a better proxy of this inherently latent variable. This is not only because the paper by González-Pereira et al. (2010) introducing the SJR explicitly declares this ranking to be a "metric of [...] scientific prestige", but also because journal rankings are frequently used as a straightforward means to judge the academic merits of a publication, for example in university-level recruitment or for the purposes of determining publication bonuses (e.g., Abritis and McCook, 2017). As a surrogate indicator of research quality, journal rankings hence bypass the need to wait for the scientific impact of a paper to materialize in terms of the number of citations it receives over time.

This study therefore seeks to reinvestigate the link between linguistic and meta characteristics of journal articles and the degree of scientific prestige associated with them in various methodologically novel ways: first of all, we use a large data set of article full texts published in the scientific domain of economics that we segment into 100 topically homogeneous clusters based on the outcome of an unsupervised topic model employing latent Dirichlet allocation (LDA) (Blei et al., 2003). In each cluster, we then model a cross-sectionally standardized version of the SJR index of aggregate *meta-articles* – averaging the features of all articles published in a given year, journal and subject area – as a function of a wide array of linguistic, bibliometric and other predictors via variable dispersion beta regression (Ferrari and Cribari-Neto, 2004). In this setup, we perform variable selection with the help of a boosting algorithm for Generalised Linear Models for Location, Scale and Shape (GLMLSS) originally developed by Mayr et al. (2012) in order to identify the most important article-level drivers of scientific prestige in each topical segment.

2. Related work

2.1. Scientific impact and linguistics

At the intersection of scientometrics, computational linguistics and text mining, there is a major branch of literature investigating the particularities of scientific writing and their relation to the perceived prestige and subsequent impact of research publications. Clearly, an important criterion through which language influences the perceived quality of a manuscript is the ease with which the reader can extract the relevant information about the results of the academic research in the field. This kind of comprehension is a difficult concept to measure however, since it depends not only on the text itself, but also on the target recipients (cf. Stevens et al., 1992). Instead, readability, defined as "the ease of understanding or comprehension based on the style of writing" (Klare, 1963), is often used as a proxy. A survey of 132 editors conducted by Wolff (1970) found writing style and readability to be the fifth most important among 15 criteria for assessing journal manuscripts, thereby confirming that linguistic characteristics matter greatly in the communication of research outcomes. A later survey by Sternberg and Gordeeva (1996) confirmed "quality of presentation" as a crucial attribute of impactful scientific papers.

While superior writing skills are known to be vital for producing high-impact research (Zimmerman, 1989), the aforementioned bafflegab theory actually questions whether maximizing readability, as quantified by traditional formulas like the *Flesch Reading Ease* (Flesch, 1948) or *Gunning Fog Index* (Gunning et al., 1952), should indeed constitute a central goal in the creation of manuscripts for researchers seeking to maximize the perceived reputation of their communications. Relatively recent empirical evidence (Didegah and Thelwall, 2013; Gazni, 2011; Van Wesel et al., 2014) appears to largely confirm the alleged negative association between readability and citation count that was alluded to by Hartley et al. (1988). Notwithstanding these findings, the exact scientific field under investigation seems to matter greatly, as there are also studies that found this relationship not to be very strong in various sub-fields of social sciences (Hartley et al., 2002; Stremersch et al., 2007).

Traditionally, readability measurement based on formulas like the ones mentioned above has relied on shallow text characteristics like average word and sentence length. Recent advances in Natural Language Processing (NLP) allow us to move significantly beyond this relatively superficial conceptualization of readability by incorporating more complex text properties. Several empirical studies found the combination of supervised learning with modern-day NLP tools for feature extraction to be significantly better suited than classical "shallow" formulas (e.g. De Clercq & Hoste, 2016; Feng, Jansche, Huenerfauth, & Elhadad, 2010; Schwendinger, Vana, & Hornik, 2020). This is due to major advances in computational linguistics that allow the construction of more complex linguistic features related to the syntactic and semantic characteristics of textual data. Feng et al. (2010), for example, propose the use of up to two hundred features for predicting readability and found several levels of quantitative linguistic analysis to have significant predictive power for readability. In our study, we thus feed a similarly diverse array of features into our model, in order to make sure that we conceptualize readability as broadly as possible in the attempt to link it to scientific prestige.

In the scientometric domain, there are also several contributions that use modern NLP tools to linguistically analyze corpora of journal articles. A major part of the literature in this regard focuses on writing styles and how they differ between academic disciplines (Lei, 2016) as well as between native and non-native English speakers (Lu et al., 2019b) or how they reflect the use of different scientific methodology (Argamon et al., 2008). Regarding the analysis of scientific impact, there are a few papers that deployed more sophisticated NLP tools for predicting citation count. Lu et al. (2019a), for instance, find no relationship of practical significance between 12 variables of linguistic complexity, sub-categorised into measures of syntactic and lexical complexity, and three classes of high-, medium- and low-impact journal articles created on the basis of normalized citation count. Gerrish and Blei (2010) on the other hand, develop a dynamic topic model that serves as a basis for the calculation of a bespoke corpus-level textual statistic called the *influence score* that is indeed shown to significantly correlate with citation count.

2.2. Scientific impact and meta features

Individually, several categories of meta features of journal articles have been shown to be related to their scientific impact in terms of citations received. For example, regarding the overall structure of an article, shorter titles (Sienkiewicz and Altmann, 2016), longer abstracts (Didegah and Thelwall, 2013) and the number of figures and tables (Haslam et al., 2008) all have been positively associated with the number of citations it received. However, the empirical evidence is somewhat ambiguous in this regard, as the demonstrated levels of correlation tend to be relatively low and studies with opposing results can usually be found. One such study, conducted by Lee et al. (2018), conversely suggests a negative association between the number of tables and scientific impact.

When it comes to characteristics of the authors and the institutions that they work at, on the other hand, scholars have rather consistently highlighted that the number of authors, as well as international and institutional collaboration among them, positively contribute to the number of citations their research is expected to receive (Gazni and Didegah, 2011; Haslam et al., 2008; Leimu and Koricheva, 2005; Sooryamoorthy, 2009). The comprehensive analysis by Larivière et al. (2015) also historically contextualizes the degrees of interinstitutional and international collaboration which they find to be necessary to realize higher impact in both natural and social sciences, supposedly as a consequence of the complexity of modern-day scientific problems. Conversely, Mubin et al. (2018) did not find that the number of authors or average years of academic experience significantly differed between particularly prestigious papers that won an award at a computer science conference and those that did not.

Another important category of article meta features that has been empirically linked to citation counts concerns bibliometrics. It has been repeatedly demonstrated in the literature that articles with a greater number of references tend to be cited more often (Bordons et al., 2013; Chen, 2012; Rigby, 2013). Besides the mere amount of references, there are several other bibliometric indicators that have been empirically studied: Didegah and Thelwall (2013) and Peng and Zhu (2012) showed that publications with high-impact references have a tendency to accumulate higher citation counts themselves. The same holds true *ceteris paribus* for articles whose references are more contemporary, i.e., have been published more recently (Haslam et al., 2008). The results of Didegah and Thelwall (2013) also revealed that the internationality of references was among the variables that significantly correlated with citation rates in the field of nanotechnology. Furthermore, Larivière and Gingras (2010) examined whether bibliometric interdisciplinarity – defined as the percentage of references from subject areas other than that of the article – mattered in subsequent citation impact. They found a non-linear relationship that suggested there to be a sweet spot of interdisciplinarity, as citation counts dropped at either end of the scale. Finally, how often references are cited in the full text also seems to be of relevance for citation-based measures of scientific prestige (Boyack and Klavans, 2005).

Importantly though, all of the aforementioned studies directly or indirectly seek to establish a link between article-level characteristics and scientific prestige by correlating these features with citation counts. There is therefore, to the best of our knowledge, no previous work on how these variables are linked to the scientific prestige that is associated with the journal that the articles are published in, although journal ranking is arguably a more widely used and immediate indicator of the quality and (potential) impact of the research performed. While rankings and their use as a proxy for research quality are certainly not without critique (see e.g., Frey and Rost, 2010), their practical relevance implies that virtually every scholar has an interest in publishing in the most prestigious journal possible. Furthermore, existing contributions to the literature focus predominantly on shallow text characteristics and/or isolated elements of metadata routinely computed on the basis of open access articles or even abstracts only. This is despite the fact that modern technology enables computational access to large representative corpora of full texts in virtually any scientific discipline, which can be analyzed on a large scale through recent advances in NLP and quantitative text analysis. Thus, we endeavor to shed further light on the driving factors of successful academic writing by filling precisely this conceptual and methodological gap in the literature.

3. Data and methodology

In the following, we describe all steps of our research procedure, starting from obtaining raw data all the way to the estimation of regression coefficients in all topical clusters.

3.1. Raw data

Although publishing research results in prestigious journals constitutes an important goal for scholars in almost all areas of science, we chose the topical focus for this study to be the field of economics and its various sub-disciplines. The economic sciences offer a large degree of intra-disciplinary linguistic variation, comprising very concise mathematical papers in econometrics as well as much

more verbose papers in business administration, for instance. While this variation posits interesting questions for a differentiated scientific analysis, we have chosen the area of economics in particular because it maintains a sufficient level of topical homogeneity that should prevent our conclusions from becoming too general and high-level. Moreover, journal rankings are of particular relevance for academic practice in economics, as tenure decisions typically measure scientific excellence only via the numbers of publications in the top journals in the fields.

The compilation of a corpus of full-text journal articles of the dimension required for performing detailed textual analyses in all sub-disciplines of this field is enabled through the specific Text and Data Mining (TDM) APIs of the publishing houses Elsevier[1] and Springer Nature. These APIs enable programmatic access to a vast library of open- and closed-access journal articles and corresponding meta data in the form of fully annotated XML files under the terms of the corresponding TDM licenses. Specifically, we started by obtaining important journal meta information from the Elsevier Serial Title API for all the economics journals accessible at our university library. We then collected a list of all available articles in each of these journals from the Elsevier Article Metadata API until the end of January 2020 and finally requested the corresponding XML files from either the Elsevier or the Springer Nature full-text API. Through this process, we obtained a data set of 255,644 articles that we split into groups of 20 in order to run the computationally demanding feature extraction pipelines (described in Section 3.2) in parallel.

3.2. Feature extraction and data cleaning

For each document in our corpus, we performed the following four tasks: first, we extracted the text in the title, abstract, full text, footnotes and appendices. For each of those sections, we then separately performed various cleaning steps that, among other things, involved removing double punctuation, expanding abbreviations and trimming white space. Second, we obtained the NLP annotations of all these distinct parts of each paper by performing tokenization, lemmatization, part-of-speech (POS) tagging, named entity recognition, constituency parsing and coreference resolution using the open-source `Java`-based toolkit *Stanford CoreNLP* (Manning et al., 2014) via its `R` interface in form of the `StanfordCoreNLP` package (Hornik, 2020). The annotation objects generated by this procedure then allow the straightforward calculation of a wide range of linguistic features which, for our purposes, can be categorized into seven groups: linguistic base, POS, entity, parse tree, word overlap, semantic and coreference. In total, we use 127 linguistic variables, whereby different subsets of this feature universe are used for different parts of the paper, i.e., the full 127 features are only computed for the full text, whereas all the other parts of a given paper are characterized via far fewer amounts of features (e.g., 27 for the title), simply as a consequence of their relative shortness. A description of all linguistic features is contained in Table S.1 in the supplementary materials.

Third, we computed 30 meta features for each article from the comprehensive information contained in the XML format that the articles are retrieved in from the API. This set of meta variables can also be grouped into several subcategories, namely base, author, bibliometric and mathematical. We refer the reader again to Table S.1 for a description of all the meta features employed.

Since we plan to stratify the document collection into topically homogeneous sub-fields of economics, the fourth and final step then involved recording the frequencies of all (lemmatized) unigrams, bigrams and trigrams that appear at least twice in the full text of a given article, in order to combine them later to form a term-document matrix for the entire corpus (see Section 3.4). Through this four-stage procedure, each journal article is therefore mapped to a vector of 344 (linguistic and meta) characteristics and a table of term frequencies, thereby bringing a previously unseen range of interpretable linguistic features to the area of scientometrics.

After gathering the feature vectors from all documents in the corpus, we performed various data cleaning steps. We stipulated that each article in the data set have a publication date, at least one author, a title and abstract, at least one full-text section and at least one reference in its bibliography. A total of 29,950 documents did not meet these requirements and were thus removed. Moreover, we had to remove 4520 papers for which we could not accurately compute certain features, mostly because their XML files deviated slightly from the otherwise standardized structure of the publisher, which occurs, for instance, with errata and editorial notes. Finally, a further 166 articles were discarded as they were published more than once in exactly the same form and thus also appeared multiple times in our data set.

In the end, we are hence left with 221,008 articles from 290 different journals (see Table S.3 in the supplementary materials for a full list). Approximately 80% of those articles appeared in a journal pertaining to Elsevier and the remaining 20% in an outlet belonging to Springer Nature. The top 25 of those journals with the most articles in the corpus already cover a wide range of different sub-fields of economics (Table 1). In terms of publication date, the earliest papers in our sample are from the late 1990's (Table 2). This is a consequence of the fact that the XML files obtainable from the publishers with earlier publication dates were created based on optical character recognition (OCR) of print-outs only, which is error-prone and renders the extraction of meta information impossible. Nevertheless, we believe our corpus to be more representative than others have been due to its size, comprehensiveness and, importantly, the pervasiveness of closed-access journals in it.

3.3. Article-level measurement of scientific prestige

So far, we have mapped each article to a high-dimensional feature vector. In order to now associate each publication with a numeric indicator of scientific prestige, we assigned to each of those vectors a standardized ranking of the journal that the article

[1] Some rights reserved. This work permits non-commercial use, distribution, and reproduction in any medium, provided the original author and source are credited.

Table 1
The number of articles in the corpus by the top 25 most frequent journals.

Journal	Freq.	Journal	Freq.
European Journal of Operational Research	12,247	Research Policy	2546
Journal of Environmental Management	8969	Journal of Economic Dynamics and Control	2538
International Journal of Production Economics	5507	Journal of Public Economics	2284
Journal of Business Research	5465	Industrial Marketing Management	2257
Economics Letters	4746	European Economic Review	2115
Ecological Economics	4712	Journal of Financial Economics	2110
Journal of Banking & Finance	4106	Transportation Research Part B: Methodological	2090
Economic Modelling	3991	International Journal of Hospitality Management	1960
Journal of Economic Behavior & Organization	3515	Journal of Econometrics	1858
Energy Economics	3462	Journal of International Money and Finance	1837
Journal of Business Ethics	3445	Omega	1822
Tourism Management	2644	Journal of Development Economics	1771
Decision Support Systems	2556		

Table 2
The number of articles in the corpus by publication year.

Year	Freq.	Year	Freq.	Year	Freq.	Year	Freq.	Year	Freq.
2020	5801	2014	14,422	2008	9250	2002	4313	1996	1
2019	21,829	2013	13,579	2007	8418	2001	4170	1995	2
2018	17,787	2012	11,809	2006	6858	2000	3917		
2017	16,444	2011	10,138	2005	5832	1999	3626		
2016	15,481	2010	9838	2004	5411	1998	2583		
2015	14,545	2009	9503	2003	5128	1997	323		

was published in. While there are numerous indicators for assessing journal quality, a full discussion of their relative merits and problems is beyond the scope of this paper and can instead be found in the literature (e.g., Serenko and Bontis, 2013; Tüselmann et al., 2015). We focus on the annually updated SJR as a commonly used and easily obtainable citation-based journal ranking based on eigenvector centrality. The version we employ was introduced by Guerrero-Bote and Moya-Anegón (2012) and is calculated as the average number of weighted citations received in a given year by the documents published in a given journal in the three previous years.

The prestige of a given article in the corpus is therefore represented by the SJR value of its publication outlet in the year of publication. However, as the shape of the cross-sectional distribution of these values can vary quite significantly from year to year, we do not use these values as a response variable directly, but rather normalize them using the cross-sectional empirical distribution function (ECDF) in each given year. More precisely, let J_t be the set of all economics journals, for which an SJR ranking in year t is available and denote these rankings by s_{jt}.[2] Furthermore, let y_{ijt} be the SJR in year t of the journal j that article i appeared in, where t is the year of publication. Notably, since all articles published in the same year and journal receive the same ranking (see Section 3.5 for more details on this issue), we have, for a given year t and journal $j \in J_t$,

$$y_{ijt} = s_{jt} \quad \text{for all articles } i. \tag{1}$$

We now compute the non-parametric probability integral transform of y_{ijt} using the cross-sectional ECDF of the SJRs of all the $n_t := |J_t|$ economics journals, for which s_{jt} is available in a given year. In this way, we define

$$y^*_{ijt} = F_t(y_{ijt}), \qquad \text{where } F_t(x) = \frac{1}{n_t} \sum_{j \in J_t} \mathbb{1}_{\{s_{jt} \le x\}} \tag{2}$$

as the nicely interpretable response variable to be assigned to each corpus element. With this specification, each article's prestige is quantified by the percentage of all economics journals that had lower SJRs than the journal it was published in.

3.4. Topical segmentation

Even within the field of economics, there is expected to be a rather large variation in the writing styles and linguistic characteristics of the various sub-fields, so that the response should not be modeled jointly over all disciplines. Evidently, a very mathematical paper in financial econometrics has to be written significantly differently than a theoretical paper describing, say, global ramifications of protectionism in order to be well-received by editors and referees of top journals in the respective fields. These differences therefore necessitate a modeling approach that considers each topic area separately.

In order to avoid having to rely on potentially unavailable meta data on the subject area of a given article, we hence segment our data set into topically homogeneous clusters in the following data-driven way: the sheer size of the data set, first of all, compels

[2] A complete list of all mathematical notation used in this paper is contained in Table A.1 in the appendix.

us to reduce the dimensionality of the term-document matrix that holds the frequencies of distinct terms for each document in the corpus by i) removing 272 different function words, ii) requiring that a given term has to be contained in at least 150 articles to be considered and iii) removing all bi-/trigrams that appear less than four times in all documents. The resulting matrix, which is still of dimension 108,303 × 221,008 with a sparsity of 99.4%, subsequently serves as a basis for the estimation of an LDA topic model using the Gibbs sampling algorithm by Phan et al. (2008) interfaced to R via the `topicmodels` package (Grün and Hornik, 2011). LDA assumes documents to be generated from a probability distribution over K topics, which are themselves just distributions over a fixed vocabulary (i.e., all terms in the term-document matrix). Based on a user-set value of K, this algorithm thus identifies the different themes that pervade the document collection via the parameters of the posterior word distribution for each topic. At the same time, it also discovers how much of each topic is present in each document via the parameters of the posterior topic distribution.

To maintain parsimony, we estimate a topic model with $K = 30$ topics and use each of those two sets of parameters for different purposes: first, we investigate the posterior word distributions to find the most frequent terms in each topic and manually label these topics accordingly. The labels given to each topic are displayed in Table A.2 (in the appendix). Moreover, Figures S.1 and S.2 in the supplementary materials contain word cloud plots for all 30 of those topics: for example, the words most representative of topic 11 are “bank”, “risk”, “financial”, “credit” and “loan”, so that we labeled this topic “Financial System”.

While these topics would in principle allow us to stratify the corpus in terms of the most dominant topic in each article, we deem such an approach too inaccurate, as many research articles in economics combine insights from various sub-disciplines, so that a uni-dimensional clustering would not yield a satisfactory degree of topical homogeneity within each group. Instead, we apply k-means clustering (Hartigan and Wong, 1979; MacQueen, 1967) to the second set of parameters, namely the posterior “topic loadings” of all papers in the corpus. In this way, each document is mapped to a 30-dimensional vector representing a discrete probability distribution over the 30 topics, which we use to find groups of articles with similar topic proportions via the classical k-means algorithm. Regarding the number of clusters, we choose $k = 100$ as this seemed to strike a balance between having enough groups to achieve high amounts of topical consistency measured in terms of average Jensen-Shannon divergence within each cluster, but not too many for the interpretation of the results in each cluster to become unwieldy.

3.5. Regression model specification

Finally, we come to the process of linking the wide range of features extracted from each article to its scientific prestige by estimating a model that establishes such a relationship in each of the 100 topically homogeneous clusters. Before we describe the structure of our model in detail, we have to address the issue that, in a given cluster, all publications in journal j and year t yield the same continuous response, i.e., we cannot distinguish these articles in terms of their prestige and therefore are only able to observe an average effect of the features on the response for them. To mitigate the statistical consequences that emerge from this repeated measurement structure, we aggregate all those “response-invariant” documents in a given cluster, journal and year to so-called *meta-articles* by averaging all their predictors and assigning the normalized SJR indicator that they share to this individual covariate vector. More precisely, for a given journal j, year t and topical cluster c, the response variable y^*_{ijtc} is the same for all articles i. Consequently, we marginalize by averaging each feature in the p-dimensional covariate vector $\mathbf{x}_{ijtc}$ over the index i

$$\widetilde{\mathbf{x}}_{jtc} = \frac{1}{n_{jtc}} \sum_{i=1}^{n_{jtc}} \mathbf{x}_{ijtc}, \tag{3}$$

where n_{jtc} is the number of articles published in journal j and year t appearing in cluster c with $c = 1, \dots, 100$. The size of our data set and our clustering based on topical similarity rather than journal allows us to perform this marginalization step, whilst maintaining a decent sample size in virtually all clusters (see Table 3).

Our goal is therefore to model the relationship between the response variable y^*_{jtc} that takes values in (0,1) and the vector of meta-article predictors $\widetilde{\mathbf{x}}_{jtc}$ separately in each cluster c. For continuous data on open intervals like this, variable dispersion beta regression models (Ferrari and Cribari-Neto, 2004; Simas et al., 2010) are a natural choice, especially as a consequence of the variety of density shapes they can accommodate. In fact, beta regression can be seen as a special case of a GLMLSS, originally developed in a slightly more general form (also allowing for effects other than linear) by Rigby and Stasinopoulos (2005). In contrast to traditional regression approaches that model only the conditional expectation of the response, the GLMLSS framework allows to model various distribution parameters simultaneously as functions of the predictors. Regarding our specific application, the beta distribution is traditionally parameterized in terms of positive shape parameters θ and τ, yet, in the spirit of Ferrari and Cribari-Neto (2004), we instead employ the mean-precision parameterization that uses a mean parameter $\mu := \theta/(\theta + \tau)$ and a precision parameter $\phi := \theta + \tau$. Assuming that $y^*_{jtc} \sim \mathcal{B}(\mu_{jtc}, \phi_{jtc})$, we thus seek to estimate the following model that expresses both these parameters as linear functions of the predictors, modulo suitable link functions, separately in each cluster c:

$$g_1(\mu_{jtc}) = \eta_{1jtc} = \widetilde{\mathbf{x}}'_{jtc}\beta_c \tag{4}$$

$$g_2(\phi_{jtc}) = \eta_{2jtc} = \widetilde{\mathbf{x}}'_{jtc}\gamma_c \tag{5}$$

for unknown regression coefficients $\beta_c = (\beta_{1c}, \dots, \beta_{pc})'$ and $\gamma_c = (\gamma_{1c}, \dots, \gamma_{pc})'$ and strictly increasing and twice differentiable link functions $g_1 : (0,1) \to \mathbb{R}$ and $g_2 : (0,\infty) \to \mathbb{R}$, typically the logit and log function, respectively.

Table 3
Information on the clusters and quality of the model fits.

Cluster	Size	IDs of top 3 topics	Pseudo-R^2	Cluster	Size	IDs of top 3 topics	Pseudo-R^2	Cluster	Size	IDs of top 3 topics	Pseudo-R^2
1	689	30, 27, 22	37.0%	35	1030	25, 12, 13	37.7%	69	973	22, 8, 20	28.4%
2	1107	11, 27, 12	35.4%	36	173	24, 14, 29	59.3%	70	468	9, 1, 6	37.9%
3	1246	15, 1, 22	36.3%	37	409	26, 20, 8	34.6%	71	348	18, 1, 4	33.5%
4	734	26, 16, 10	40.8%	38	844	11, 1, 13	43.3%	72	1146	1, 23, 13	33.9%
5	55	21, 26, 16	25.9%	39	598	3, 4, 8	34.4%	73	432	19, 8, 3	35.0%
6	739	9, 20, 8	29.3%	40	778	27, 14, 29	41.8%	74	566	3, 4, 28	39.9%
7	974	12, 1, 25	43.2%	41	537	14, 17, 29	33.9%	75	597	25, 2, 7	40.1%
8	1228	23, 1, 13	39.6%	42	500	27, 29, 18	48.6%	76	262	29, 2, 25	62.8%
9	547	4, 20, 8	30.7%	43	653	26, 1, 6	36.8%	77	252	24, 30, 29	46.6%
10	497	8, 28, 20	41.2%	44	461	9, 1, 6	34.2%	78	591	28, 1, 23	29.9%
11	847	1, 6, 26	38.1%	45	232	14, 2, 25	33.8%	79	384	16, 23, 13	51.1%
12	597	2, 29, 25	45.3%	46	241	21, 26, 16	37.3%	80	530	6, 1, 17	43.5%
13	684	28, 20, 8	24.8%	47	964	7, 1, 25	39.8%	81	545	10, 24, 14	46.2%
14	1019	18, 1, 4	26.2%	48	721	22, 1, 13	38.8%	82	777	16, 1, 13	35.7%
15	195	19, 24, 30	53.2%	49	875	15, 1, 7	42.0%	83	562	4, 20, 8	37.1%
16	349	18, 14, 27	26.1%	50	669	15, 7, 1	40.8%	84	530	12, 17, 23	49.5%
17	762	8, 4, 3	38.3%	51	518	28, 3, 1	32.8%	85	307	24, 14, 10	36.9%
18	773	8, 20, 22	30.8%	52	557	6, 1, 9	40.8%	86	686	22, 20, 13	26.6%
19	979	4, 20, 8	39.3%	53	605	15, 3, 1	28.4%	87	598	20, 8, 22	40.1%
20	242	29, 2, 18	49.8%	54	721	15, 11, 1	33.2%	88	543	17, 6, 1	46.1%
21	651	7, 1, 15	43.3%	55	294	19, 29, 27	46.6%	89	652	17, 14, 29	45.2%
22	461	22, 18, 27	34.8%	56	480	20, 5, 8	28.9%	90	812	20, 8, 22	27.5%
23	348	26, 21, 1	27.9%	57	859	3, 28, 20	31.8%	91	528	8, 10, 20	39.9%
24	436	29, 24, 17	53.4%	58	211	19, 24, 14	63.3%	92	582	2, 1, 10	26.4%
25	685	8, 20, 4	39.0%	59	304	29, 18, 27	35.1%	93	1188	17, 23, 14	35.6%
26	993	6, 1, 23	34.0%	60	607	25, 7, 2	46.7%	94	787	25, 7, 2	28.1%
27	23	21, 16, 26	79.5%	61	727	18, 1, 27	31.9%	95	822	2, 10, 29	29.1%
28	460	27, 14, 18	25.2%	62	339	30, 24, 2	37.1%	96	989	13, 23, 22	33.4%
29	720	3, 8, 20	34.3%	63	983	20, 22, 8	33.5%	97	512	5, 4, 20	37.3%
30	527	5, 4, 20	40.4%	64	181	14, 24, 18	19.9%	98	572	7, 2, 29	43.6%
31	1087	10, 2, 1	26.8%	65	1339	27, 18, 1	33.8%	99	605	4, 5, 20	42.0%
32	669	20, 22, 8	42.5%	66	690	12, 25, 1	48.2%	100	999	5, 1, 8	38.2%
33	266	24, 29, 30	48.0%	67	476	20, 6, 4	32.6%				
34	971	8, 5, 10	34.5%	68	1004	13, 1, 23	40.7%				

Table 3 reports several pieces of information characterizing the clusters and the quality of the beta regression model fits in them. The top three topics are the IDs of the topics that have the highest posterior probabilities when averaged over all articles in a given cluster. Please refer to Table A.2 (in the appendix) for the labels given to those IDs based on LDA output. Moreover, we computed a pseudo-R^2 as the in-sample reduction in mean-squared error (MSE) (in %) attributable to our model from Eqs. (4) and (5) to assess model fit.

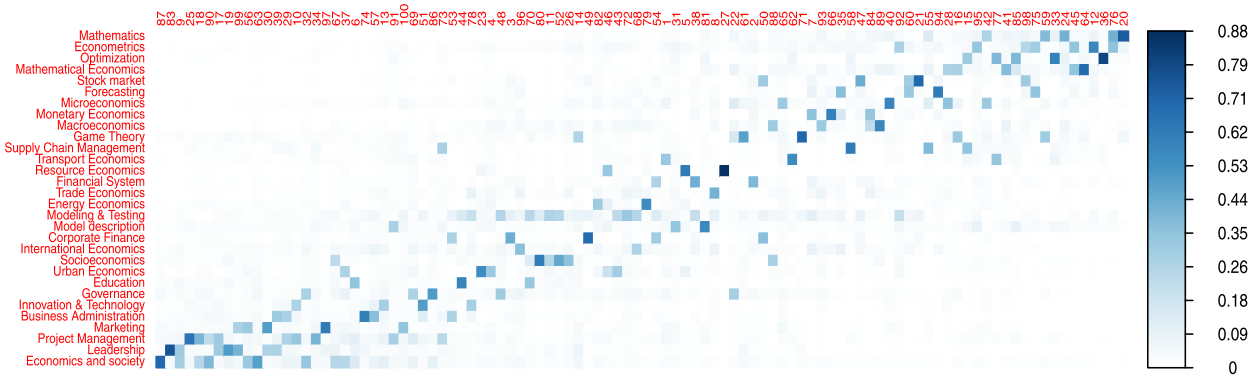

Fig. 1. Seriated heat map of average posterior topic probabilities in the clusters (topics in rows, clusters in columns).

While this model could be estimated at relative ease with maximum likelihood methods, Groll et al. (2018) pointed out that such procedures can become rather unstable in data set-ups as high-dimensional as ours, thus requiring alternative approaches that perform variable selection. Doing so would moreover boost interpretability, which constitutes a major goal in our modeling efforts.

The literature offers various options for feature selection in GLMLSSs (e.g., Bayer and Cribari-Neto, 2014; Liu and Kong, 2015; Umlauf et al., 2018; Zhao et al., 2014), yet we choose a component-wise gradient boosting algorithm originally due to Mayr et al. (2012) and later refined by Thomas et al. (2018). This approach has been shown to perform well in preventing overfitting in high-dimensional settings, deals very effectively with multicollinearity and is implemented highly efficiently in the R package `gamboostLSS` (Hofner et al., 2016). It therefore appeared particularly appropriate for our purposes, considering that correlations between linguistic features that can be thought of as similar measures of latent text characteristics are expected to be high. Very briefly, this algorithm iteratively fits simple base learners to the negative gradient of the loss function – in our case the negative log beta likelihood of the response – and then uses only the best-performing one (in terms of residual sum of squares) to update the additive predictors η_1 and η_2 as defined in Eqs. (4) and (5). When left to run until infinity, this algorithm can be shown to converge to the same solution as classical maximum likelihood estimation, but stopping it early induces sparsity as less important predictors are never updated.[3] Hence, the number of iterations until stopping, which is always a scalar denoted by m_{stop} under the algorithmic revision by Thomas et al. (2018) that we employ, becomes a central hyperparameter to be tuned before model estimation. In application, we specifically determine a suitable value for m_{stop} in each cluster based on 10-fold cross validation and then run the algorithm until the pre-specified number of iterations. Notably, the mean parameter μ_{jtc} and the precision parameter ϕ_{jtc} are allowed to depend on possibly different sets of predictors, so that the results presented in Section 4 are estimates of β_c and γ_c that will generally differ in terms of the degrees of sparsity that they possess.

4. Results

In the presentation of our results, we proceed as follows: we commence by briefly characterizing the clusters in terms of their topical orientation resulting from the segmentation procedure explained in Section 3.4. Subsequently, we analyze the goodness-of-fit of the regression models in these clusters based on their in-sample mean-squared-error (MSE) and then go into more detail by investigating which features were found to be relevant in which clusters based on the estimates of the corresponding regression coefficients. The discussions of these regression outputs are grouped by the main types of feature (i.e., linguistic and meta) for better orientation.

4.1. Topical segments

In order to characterize the subject matter that the clusters predominantly deal with, we averaged the posterior topic probabilities over all meta-articles in the cluster and stacked those 30-dimensional probability vectors together into a 100×30 matrix. To illustrate similar clusters, we seriated this matrix of prototypic topic probabilities using an algorithm based on principal components implemented in the `seriate` package (Hahsler et al., 2008) and displayed its transpose in Fig. 1. From this figure, we conclude that, while a relatively small number of topics drives each cluster, the segmentation procedure based on posterior topic probabilities has indeed allowed us to reach a more fine-grained understanding of the topical composition of the corpus. For example, both clusters 76 and 98 deal with econometrics, yet the former mostly concentrates on theoretical papers, since it is combined with the mathematics topic, and the latter more on applications of econometrics to finance, courtesy of its high weight on the stock market topic. A simple topic model approach would not have been able to differentiate such nuances. As an additional illustration of the usefulness of this segmentation approach, we have performed an external validation of the topical consistency of the clusters. The results for this are contained in Appendix C.

[3] As the methodological particularities are beyond the scope of this paper, we refer the interested reader to the original papers by Mayr et al. (2012) and Thomas et al. (2018) and the references therein for further details.

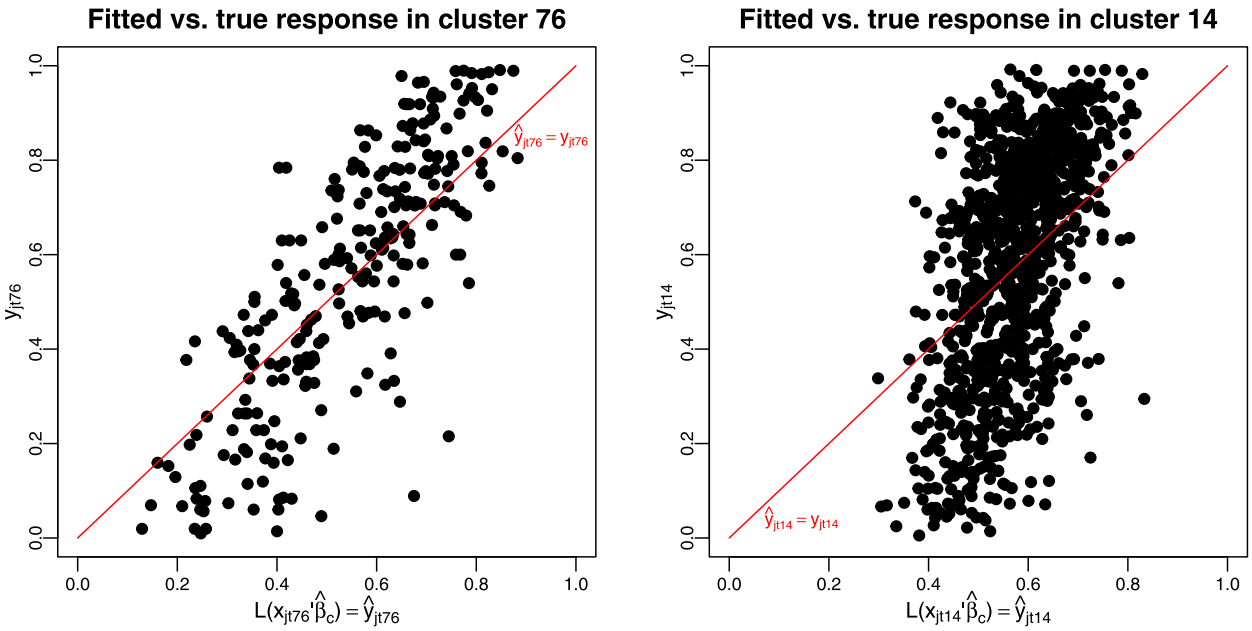

Fig. 2. Fitted vs. true response plots in two clusters (L is the standard logistic function).

In the following, we shall thus use the information on the topical focus of the different clusters to relate our findings to the corresponding sub-fields of economics. In order to characterize the clusters in a more compact fashion, we associate each LDA topic with an ID as contained in Table A.2 in the appendix.

4.2. Goodness-of-fit of regression models

We start the discussion of results by quantitatively investigating the goodness-of-fit of the models. More specifically, we compare their in-sample MSE to a base value that results from guessing the unconditional mean of y^*_{jtc}, i.e., its unconditional variance. Besides cluster sizes and the top three topic IDs, Table 3 displays this goodness-of-fit measure based on the reduction of in-sample MSE achieved by the corresponding beta regression fit. As can be seen from the table, there is a considerable degree of variability among the clusters in terms of how well the models fit the data, yet all of them exhibit at least a 20% reduction in MSE, with around 38% on average and 79% at most. The highest value is achieved in the very small and topically homogeneous cluster 27 that seems to largely deal with environmental issues. The worst value, on the other hand, was realized in the relatively small and mathematics-heavy cluster 64, which already indicates that the quality of the model fit is not necessarily related to the sample size.

In some clusters, the signal-to-noise ratio seems to be particularly low, so that even a larger sample size does not lead to a significantly better fit, while the data sets in other clusters do not have this problem to such a high degree. Figure 2 further illustrates this by juxtaposing plots of fitted vs. true response values in clusters 76 and 14. The former deals mostly with mathematical econometrics and belongs to the clusters where we can best explain the scientific prestige associated with the articles based on our features (MSE reduction approximately 63%). However, for the latter – largely occupied with game theory and socioeconomic modeling – the plot clearly shows that our model does not do much better than just naively predicting the unconditional mean despite having around four times as many observations as in cluster 76 at its disposal. Even so, there is still a clearly recognizable tilt to the right in the point cloud indicating that even in such a very low signal-to-noise scenario, our model does possess some utility, entailing a 26% reduction in MSE in this specific case.

4.3. Regression output – Linguistic features

Next, we examine the concrete outcome of the boosting-based estimation process and review which kinds of variables are the most important correlates of scientific prestige in the various clusters. To maintain brevity, we focus on the estimates of β_c in the conditional mean Eq. (4), the most important of which are illustrated graphically in Fig. 3. We therefore concentrate our interpretations on variables that are consistently relevant in many clusters or stand out from the others whilst offering interesting insights into specific sub-groups. For further details, Figures S.3 and S.4 in the supplementary materials display the estimates of all β_c coefficients. Of course, all predictors were standardized before model estimation, so that their coefficients can actually be compared in terms of their relative magnitude.

First of all, only a small number of title-based features seems to be positively associated with article-level prestige. While 17 out of 22 such predictors were relevant in at most one cluster, the semantic category “Difficult” has a negative impact in eight clusters, with a particularly strong influence in rather quantitative fields related to mathematical economics. Together with a similarly negative

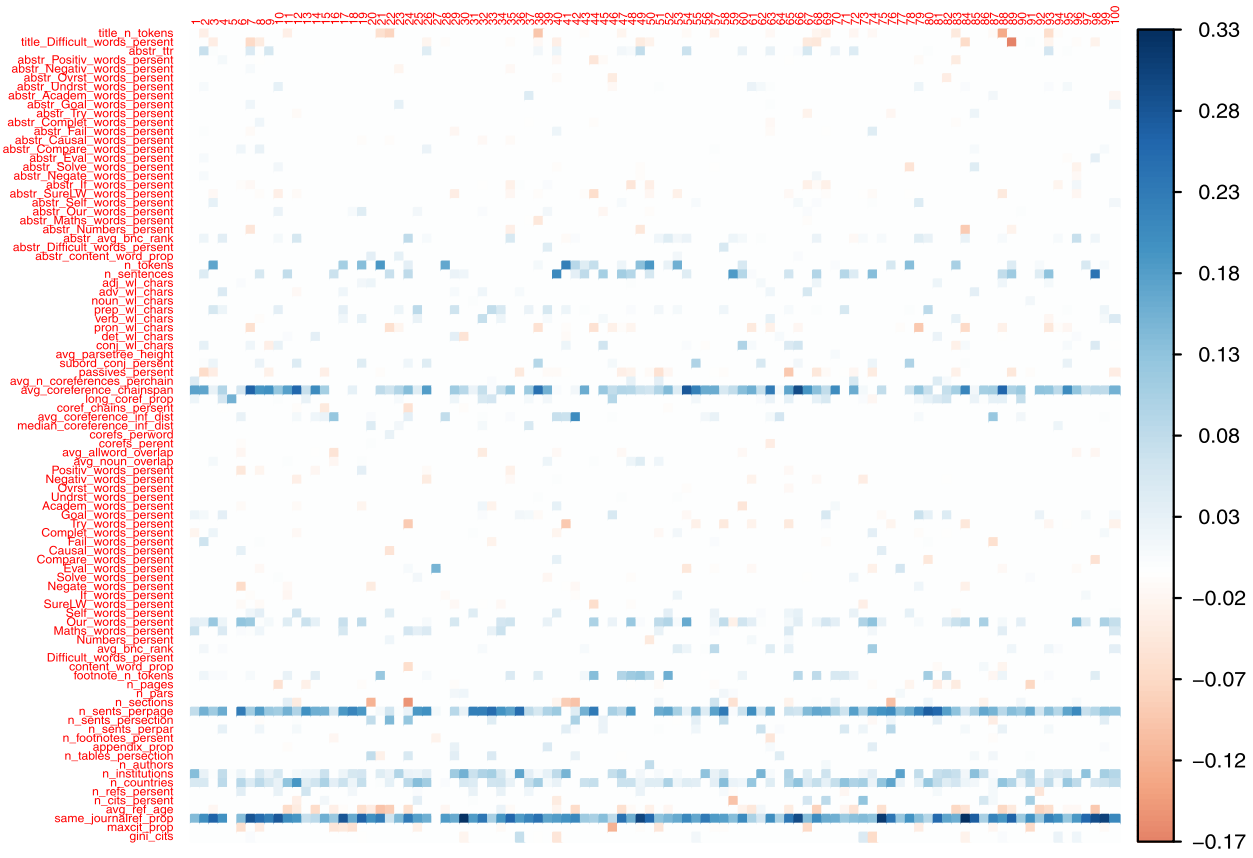

Fig. 3. Estimates of regression coefficients β_c for relevant variables in all clusters.

contribution of the title length (measured by number of tokens), this seems to suggest that in such sub-fields of economics, short and concise titles are positively associated with the level of prestige of the corresponding research.

The relatively low informativeness of linguistic features in the title is generally confirmed in the abstract. 58 out of the 119 abstract-related covariates were not relevant in any cluster and a further 33 only in one, notably only with very small coefficients in absolute value. However, in almost every cluster, there is at least one semantic category whose increased use in the abstract is (positively or negatively) associated with scientific prestige, which points to the importance of writing style and word choice in the abstract. For instance, in the environmental economics clusters 46 and 82, excessive use of words in the tag category “Ovrst” (Overstated) is indicative of articles published in lower-prestige journals, which could be a consequence of the sensationalist and unscientific impression that a paper written in an excessively lurid style creates. Furthermore, our results seem to give additional tentative evidence to the negative relationship between abstract readability and scientific prestige that the bafflegab theory poses and that various papers found (e.g., Lu et al., 2019a; Van Wesel et al., 2014). More precisely, both the type-token-ratio and the average British National Corpus (BNC) rank of the words in the abstract – measures of lexical complexity and diversity – are consistently positively associated with article prestige in eleven and nine clusters, respectively, covering a wide range of topics from corporate finance and monetary economics to behavioral and social economics.

When it comes to full-text features, we can start by confirming Didegah and Thelwall (2013) and Haslam et al. (2008) in finding a positive link between the overall length of the paper (measured by the number of both tokens and sentences) and article prestige. These variables have a consistently positive sign and are of relatively big magnitude in more than 20 clusters of varying topical backgrounds, where the effect size is particularly considerable in quantitative domains like mathematical economics and finance. POS and entity features, on the other hand, are not very good predictors of normalized journal ranking, as we tend to observe coefficients different from zero only in very few isolated clusters without obvious explanations. A slight exception are the average word lengths of different POS tags that typically have an effect on article prestige in between 6 and 10 clusters. Since word length is another important factor in readability assessment (e.g., Flesch, 1948), these results allow us to analyze its association with scientific prestige in more detail. Indeed, the fact that the length of words in some POS tags impacts the response consistently positively (e.g., verbs and prepositions), while the length of others does so negatively (e.g., pronoun), indicates that this relationship is more complicated than what had been found in previous studies (e.g., Gazni, 2011).

Additional light on these more intricate aspects of the bafflegab theory can be shed by parse tree features that allow us to investigate linguistic complexity along its second dimension besides lexical, namely *syntactic* complexity (Kormos, 2011; Lu et al., 2019b). Our results in this regard show that the number of subordinating conjunctions per sentence has a consistently positive impact

on the prestige associated with journal articles in eight clusters. This provides additional evidence in favor of the hypothesis that a verbose, hypotactic writing style with complex syntactic structures tends to leave the reader with an impression of sophistication, thereby positively contributing to the expected level of scientific prominence that he (consciously or unconsciously) assigns to it. With respect to the syntactic dimension of "bafflegab", we have hence collected further evidence that scientists can indeed gain prestige by writing in a more complex fashion. However, certain types of constructions that are generally also considered to decrease readability, such as passive voice (variable `passives_persent`), are consistently negatively associated with the response, which re-emphasizes the importance of paying attention to the linguistic subtleties in this regard.

In fact, the adverse effects of a too clinical writing style are confirmed even further by several other variables: the number of pronouns per sentence as well as an increased use of words in the semantic categories "Our" and "Self" all have a consistently positive impact on the normalized journal ranking in between 10 and 40 clusters, which suggests that scholars might be able to gain prestige by describing their research procedures from a personal perspective, rather than in an impersonal fashion characterized by the use of few pronouns and an excessive deployment of passive voice. This conclusion ties back in an interesting way to a study by Hyland and Jiang (2017) who found the use of first person pronouns to have increased in scientific writing over time, albeit primarily in natural rather than social sciences.

Besides those two semantic categories, there are several others that possess a certain level of informativeness in various clusters. Notably, though, their signs often differ between clusters, since, evidently, different research areas within the realm of economics use distinct terminologies and standard vocabulary. Consequently, for a paper in, say, cluster 75 (dealing mostly with forecasting and econometrics) the increased use of terms in a tag category like "Causal" is beneficial for the overall linguistic impression that it creates, whereas the effect might be adverse in articles on public policy and governance (cluster 22). For the "Maths" category, on the contrary, the effect is consistently positive in 12 clusters and often rather large in magnitude. Interestingly, the majority of those research areas are not core mathematical fields, but relate to other disciplines of economics such as transport economics (cluster 1), project management (17, 25 and 34) or socioeconomics (26 and 88). This could possibly be a consequence of the fact that in traditionally less quantitative fields, the increased use of mathematical vocabulary is associated with higher degrees of methodological sophistication, which might be particularly appreciated in precisely such domains where this cannot necessarily be considered commonplace.

The persistently most important category of linguistic full-text features were interestingly found to be coreference-related. Especially the variable `avg_coreference_chainspan` – as the average distance in number of tokens between the first and the last mention of a regularly appearing phrase – is one of the most informative predictors overall with consistently positive coefficients in 77 clusters. This variable can be interpreted as a measure of how well an article frames its content, i.e., picks up points made in the introduction again in the conclusion. Papers that adopt this approach tend to tell a very consistent story, clearly communicate their key ideas and thus create a longer-lasting positive impression on the reader, which evidently appears to occur particularly frequently among articles published in highly prestigious journals. This interpretation also aligns well with the fact that the proportion of long coreferences (i.e., those that span more than half of the document) among all of them or the average inference distance in number of characters are both also positively associated with article prestige in 16 and 10 clusters, respectively.

Footnote- and appendix-related characteristics of meta-articles are, in general, no major factors in determining scientific prestige; 37 out of the 46 variables that we measure on those two parts of scientific papers were relevant in only two or fewer clusters. The length of footnotes in number of tokens, in contrast, is consistently positively associated with the response in 18 clusters, all of which have the "Modeling & Testing" topic among their top 3. This is not very surprising as the use of footnotes is particularly pervasive in empirical fields, where they provide additional information to the reader without cluttering up the full text. Indeed, the inclusion of meaningful footnotes in the right places appears to be particularly common among articles published in high-quality journals, so that authors seeking to publish in such outlets might want to embrace this idea and set up their articles accordingly. The fact that the word lengths of certain POS tags as well as the proportion of long words in the appendix also have a persistently positive sign in several clusters seems to point into a similar direction.

4.4. Regression output – meta features

Lastly, we turn our attention to meta features, among which we find several of the strongest predictors of normalized journal rankings. The number of sentences per page is one of the most important drivers overall, appearing in 87 clusters with a consistently positive sign. This variable is of course strongly influenced by editorial decisions on the layout of the journal, so it comes at no surprise that it is somewhat reflective of the publication outlet and its associated prestige, when measured at the article level. The persistent positivity is interesting, nonetheless, as it implies that high-quality journals, on average, prefer a more compact setup that puts many sentences on a page, rather than stretching a paper out over many more pages than strictly necessary. The positive appearance of the number of sentences both per section and paragraph points into a similar direction, although breaking down the structure of an article into these constituent elements is of course largely down to the author. Consequently, the previously found general positive impact of article length might also play a role here.

In fact, we found the characteristics of the authors themselves to also matter greatly: both the number of institutions and the number of countries involved in the research being communicated were very informative predictors in our models of scientific prestige in 47 and 67 clusters, respectively. This aligns well with previous findings by Gazni and Didegah (2011), Larivière et al. (2015) and others in confirming the relevance of institutional and international collaboration for producing high-grade research outcomes and suchlike papers. Judging by our results, however, even more important for the purpose of publishing these papers in a top journal is to ensure that they are an excellent fit for this particular publication outlet. This is evidenced by the fact that the proportion of the references that go to the same journal (variable `same_journalref_prop`) is the single most consistent and important driver of

article prestige with consistently positive coefficients, typically of the largest relative magnitude, in virtually all clusters. Embedding the premises of a research project into a web of preexisting literature that is primarily spun by articles from the same journal that the authors seek to publish in establishes a high degree of alignment with the journal's topical focus and is therefore indicative of potentially interesting extensions of topics that had previously been discussed in it. This seems to be particularly appreciated in high-quality journals, where such a procedure additionally implies that the authors predominantly cite high-quality papers.

In a similar vein, the recency of the references is also positively associated with scientific prestige (i.e., the more recent, the better), albeit only in 22 clusters. In agreement with previous literature (Boyack and Klavans, 2005; Haslam et al., 2008), we thus found bibliometric features to be persistently high-performing covariates in our models. This is in contrast to the final category of meta features, namely mathematical characteristics, which we found not to be majorly associated with scientific prestige measured by normalized journal ranking.

The previous paragraphs only discussed the results for the mean Eq. (4) of our model, even though we additionally have information on which predictors drive the conditional uncertainty associated with the mean "guess" that our beta regression models make. We refer to Figures S.5 and S.6 in the supplementary materials for a complete display of the estimates of the γ_c coefficients, as a full discussion of the corresponding interpretations would transcend the scope of this paper. Nevertheless, it is interesting to note that a variable is estimated to be different from zero in only 1.9 clusters on average in Eq. (5), while this number is 3.3 in Eq. (4). Moreover, the signs of these coefficients are much more inconsistent between clusters than what is the case in the mean equation. While `same_journalref_prop`, for instance, still tends to be the strongest overall predictor when it comes to conditional precision, its sign is now positive in some clusters and negative in others.

5. Discussion

With our variable dispersion beta regression model, specified in Eqs. (4) and (5) and estimated in each topical cluster, the previous section has identified the most important characteristics of academic articles in terms of their explanatory power for an SJR-based proxy of scientific prestige. In the following, we discuss some of our modeling choices by pointing out various issues that we encountered in the process and how we resolved them. Moreover, we briefly sketch the outcomes that ensued from implementing several alternatives to these choices as robustness checks and highlight some of the limitations of our study.

5.1. Modeling choices and robustness checks

A first critical step in our modeling procedures was the segmentation of the corpus into topically homogeneous clusters. Even though our topic model identified highly consistent and easily interpretable topics that cover all areas of economics, the widely known limitations of LDA (e.g. Koltcov et al., 2014) still imply the danger of these topics potentially covering too broad content. We therefore increased the accuracy of our topical subdivisions by feeding the topic vectors into a clustering algorithm to group articles of similar topical composition together. Indeed, the vast majority of scientific writing in economics is likely to be well represented by this approach that clusters together articles with similar weightings of different topic-specific vocabularies, such as finance and econometrics or optimization and supply chain management. To achieve this, we did try several clustering methodologies, namely hierarchical clustering (Johnson, 1967), DBSCAN (Ester et al., 1996) – both based on Jensen-Shannon divergence – and model-based clustering (Fraley and Raftery, 2002), yet only k-means converged in reasonable amounts of computation time and delivered consistent enough outcomes when run with different starting values.

In order to review the robustness of this segmentation approach, we varied the number of LDA topics K in steps of five between 20 and 40. The word cloud plots emerging from a topic model with 20 or 40 topics revealed that the resulting topics were either too vague and broad, or, respectively, started to become repetitive if a too high number was chosen for K. Nonetheless, building our subsequent clustering and regression pipeline on top of either of those less suitable topic models prompted very similar conclusions to the ones presented in Section 4. More precisely, the most informative predictors were consistently related to paper length, coreference chain spans, the use of personal pronouns, the "density" of the article in terms of sentences per page, international and institutional collaboration in the research team and the sources and recency of the references cited in the paper.

The next central issue in our research procedure is the question of how to address the repeated measurement issue briefly sketched in Section 3.5 that results from assigning the same continuous response to all articles published in the same journal, year and topical area. In order to acknowledge that we can only observe an average effect of the covariates on the response, we aggregated the observations to meta-articles as discussed previously. Nonetheless, as a robustness check, we also estimated the model based on individual articles, implying much bigger sample sizes in all clusters, and found results to be virtually identical, even down to individual coefficient estimates in some clusters.

Two further aspects of our modeling choices are worth discussing: firstly, one could argue that the articles published in a given journal jointly have a delayed *direct* impact on SJR, since the citations that they receive influence such citation-based measures of journal ranking in subsequent years. However, there is an inherent uncertainty in this delay as the time for citations to materialize evidently varies greatly between papers. It is thus important to highlight that with the given model, we did not seek to make claims on how the features of academic articles published in journal j and year t subsequently affect the ranking of this journal through their impact on citations. Instead, we consider the normalized SJR in year t to be a proxy for the latent variable of "article prestige" as conceived by Hartley et al. (1988) and Armstrong (1989), thereby seeking to express this article-specific measurement as a function of the predictors. In this, the current academic prestige of an article is assumed to be best reflected in the most recent available ranking of the journal it was published in, which is why we specified this relationship in the model concurrently.

Secondly, the toolbox of Generalized Additive Model for Location, Scale and Shape (GAMLSS) would in principle allow a move beyond our naturally restrictive linearity assumptions with the potential benefit of substantially improving the model fits. We therefore also specified our beta regression model with smooth effects in all covariates based on P-splines (Eilers and Marx, 1996). The boosting algorithm by Thomas et al. (2018) easily accommodates this generalization, so that we were able to perform variable selection in a similar way as in the linear case. However, the increased flexibility of the model only marginally decreased the in-sample MSE in all clusters compared to the GLMLSS specification, if at all. On the contrary, we encountered severe overfitting issues even when running the algorithm for several dozen iterations only, especially in clusters with small sample sizes, where the number of covariates was up to an order of magnitude larger than the number of observations. Due to the substantially higher levels of interpretability in the linear model specification, the more stable progression of the boosting algorithm and the comparable fitting performance, we thus decided to stick with a GLMLSS for the main part of the paper.

5.2. Limitations

There are several limitations to our study: first of all, it is important to emphasize that the type of observational data that we base our conclusions on allow us to make only associative rather than causal claims. Therefore, some of the proposed relationships might be spurious or consequence of an omitted variable bias. For instance, in the relationship between linguistics and scientific prestige, there might be a mediating effect of the proof-reading and editing process of top journals that cannot be adequately captured by our model. While actual direct causalities as in "this article was published in a top econometrics journal *because* it is written in a certain way" can thus not be reasonably ascertained, our work significantly expands on the type of correlative studies done in the literature on the relationship between scientific prestige and article features.

Furthermore, our approach is limited in its novel and yet incomplete conceptualization of scientific prestige through normalized SJR. How much academic prestige and even quality scholars ascribe to a specific research paper is indubitably influenced by the overall reputation of its publication outlet. However, using journal rankings for this purpose in an isolated fashion certainly does not paint a complete picture, either, especially since such rankings have problems of their own (cf. Willmott, 2011). Novelty, structure, methodological consistency, practical relevance and citation counts are just some of the many other qualitative and quantitative factors that influence the perceived quality of a journal article. It is therefore worth reiterating that we do not claim that the response variable in our regression models actually measures the scientific merit of a publication. Instead, our goal was to identify linguistic and meta correlates of an indicator that is of tremendous practical importance in many scientific communities, particularly so in economics.

Additional limitations of our study concern the selection of the sample. We focus on a single academic discipline, which inherently limits all of our conclusions to the realm of economics. Moreover, our corpus comprises a large amount of articles in journals whose rankings span the whole range of the scientific literature, yet there are of course several dimensions in which it is not fully representative. For example, we have only obtained contributions from two publishing houses, so that several outstanding academic journals in economics are missing from our sample. However, based on the representativeness of our data set, we believe that our conclusions are likely to generalize to journals from publishers that currently do not yet offer standardized programmatic access to their articles.

6. Conclusion and future work

This paper took a novel approach in analyzing the relationship between linguistic and meta characteristics of scientific texts and the prestige associated with the journal they were published in. We deployed state-of-the-art NLP techniques to extract these characteristics from a large corpus of articles in economics, segmented these documents by topic and then estimated a variable dispersion beta regression model in each topically homogeneous cluster after aggregating individual constituents to meta-articles. Through the use of a variable selection procedure based on gradient boosting, we were able to determine the most important drivers of scientific prestige in 100 granular economic sub-disciplines. While the majority of high-resolution NLP predictors were not found to be majorly associated with the reputation of the publication outlet, we did identify (at least) six groups of covariates that have the potential to separate high-quality papers from lower-quality ones, where *quality* is understood merely in terms of the ranking of the journals they were published in. Those groups are related to the length of the paper in number of tokens and/or sentences, the span of coreference chains in its full text, the deployment of a personal writing style involving little passive voice and many first person pronouns, the average number of sentences per page, collaboration in research teams and the references cited in the paper. Even though one must be very cautious in concluding any sort of direct causality from individually measured effects, we deem the consistency of the influence of this particular feature set across topics to be a clear indication of an underlying association and more than sufficient grounds for further investigation.

Our empirical results moreover revealed a complex relationship between readability and scientific prestige that sheds additional light on the previously hypothesized bafflegab theory and contained interesting avenues for future research. More precisely, we found the POS tag of the words to matter when examining the relationship between word length as a traditional readability dimension and normalized journal ranking, since different parts of speech often had opposing signs in their regression coefficients. Furthermore, the impact of the number of subordinating conjunctions – also negatively associated with readability as a measure of syntactic complexity – was frequently found to be in agreement with the bafflegab theory, whereas other factors in the syntactic dimension of readability, such as the number of passive constructions per sentence, were found to be negatively associated with article-level prestige.

By the means of this study, we thus provided further evidence to support previous theoretical and empirical findings on how language matters in the communication of research outcomes. In the scientific domain, language serves both as a proxy for research quality (cf. Armstrong, 1980b) and as a simple, yet ideally cautiously designed means to the end of disclosing advances in a given field to the community and the general public. Precisely the fact that certain meta-article-level covariates are strong predictors in separating articles published in lower-quality journals from those published in top-tier outlets regardless of the topics discussed in the paper strongly supports this claim.

An important direction for future investigation relates to alternatives for the conceptualization of scientific prestige. While previous papers have had the tendency to do so via citation counts, we argued that the use of normalized journal rankings as a more immediate indicator of research reputation is an equally valid choice, although it comes with the inevitable downside of using journal-level measurements to answer article-level questions. A promising idea would hence be to "marry" those two approaches and construct an article-level indicator of scientific prestige that has its foundations in the reputation of its publication outlet, but is additionally informed by the number of citations it received over time. Modeling such a response would arguably be an even more comprehensive way of operationalizing scientific prestige, precisely because even articles published in lower-tier journals can gain substantial prestige over time by being well-received by the academic community, which manifests in the amount of citations received. As our current approach does not accommodate such scenarios, we thus believe there to be tremendous potential in enriching our procedures by citation counts.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Julian Amon: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft, Visualization. **Kurt Hornik:** Validation, Writing – review & editing, Supervision.

Acknowledgments

The authors would like thank an anonymous reviewer for the suggestion of an explicit validation of our topic modeling approach, which was ultimately included in Appendix C. We would also like to express our gratitude to all other reviewers for their helpful suggestions for improving the quality of this paper.

Appendix A. Notation table

Table A.1
Notation table.

Var.	Definition
K	Number of topics in the LDA topic model
k	Number of clusters for the k-means algorithm
p	Number of covariates
J_t	Set of all economics journals in our sample, for which an SJR in year t is available
n_t	Number of elements in J_t
s_{jt}	Value of the SJR indicator of journal j in year t
F_t	Cross-sectional empirical distribution function of s_{jt} for a given t
y_{ijt}	Value of the SJR indicator for article i published in journal j in year t
y^*_{ijt}	Probability integral transform of y_{ijt} under F_t, i.e., $y^*_{ijt} = F_t(y_{ijt})$
y^*_{ijtc}	Value of y^*_{ijt} for article i published in journal j and year t and assigned to topical cluster c
$\mathbf{x}_{ijtc}$	Covariate vector for article i published in journal j in year t and assigned to topical cluster c
n_{jtc}	Number of articles in cluster c that were published in journal j and year t
$\bar{\mathbf{x}}_{jtc}$	Value of $\mathbf{x}_{ijtc}$ averaged over all articles $i = 1, \dots, n_{jtc}$ in cluster c, published in journal j and year t
y^*_{jtc}	Value of y^*_{ijtc} for a given journal j, year t and cluster c and any i (all the same for j, t and c given)
μ_{jtc}	Mean parameter of the beta distribution for y^*_{jtc}
ϕ_{jtc}	Precision parameter of the beta distribution for y^*_{jtc}
g_1	A link function $g_1 : (0,1) \to \mathbb{R}$ for the mean parameter of the beta distribution, typically logit
g_2	A link function $g_2 : (0,\infty) \to \mathbb{R}$ for the precision parameter of the beta distribution, typically log
β_c	A coefficient vector for the mean parameter of the beta distribution for y^*_{jtc} in cluster c
γ_c	A coefficient vector for the precision parameter of the beta distribution for y^*_{jtc} in cluster c
η_{1jtc}	The linear predictor for the mean parameter of the beta distribution for y^*_{jtc}, i.e., $\eta_{1jtc} = \bar{\mathbf{x}}'_{jtc}\beta_c$
η_{2jtc}	The linear predictor for the precision parameter of the beta distribution for y^*_{jtc}, i.e., $\eta_{2jtc} = \bar{\mathbf{x}}'_{jtc}\gamma_c$
m_{stop}	Hyperparameter in the boosting algorithm by Thomas et al. (2018)

Appendix B. Topic labels

Table A.2
Labels of the topics from LDA output.

Topic ID	Topic label	Topic ID	Topic label
1	Modeling & testing	16	Energy economics
2	Econometrics	17	Macroeconomics
3	Business administration	18	Game theory
4	Leadership	19	Supply chain management
5	Marketing	20	Economics and society
6	Socioeconomics	21	Resource economics
7	Stock market	22	Governance
8	Project management	23	Trade economics
9	Education	24	Optimization
10	Model description	25	Forecasting
11	Financial system	26	Urban economics
12	Monetary economics	27	Microeconomics
13	International economics	28	Innovation & technology
14	Mathematical economics	29	Mathematics
15	Corporate finance	30	Transport economics

Table A.2 matches the topic IDs used in Table 3 to their labels. From the LDA algorithm, we obtain posterior word distributions for each topic and use this information to assign suitable labels to the topics. Word clouds of all topics that illustrate these posterior word distributions are contained in Figures S.1 and S.2 in the supplementary materials.

Appendix C. Analysis of cluster consistency

We have segmented our corpus into 100 topically homogeneous clusters by apply k-means to the posterior LDA topic loadings of all papers in the corpus. In order to investigate whether this approach does a reasonable job for the overall subject area of economics, we performed an external validation using the web-based service Connected Papers. For a given seed paper, this service finds the 40 most related articles and arranges the result in a force-directed graph using a similarity metric that is based on co-citation and bibliographic coupling. In this way, two papers that apply similar methodology to related topical areas are identified as being closely connected, even if they do not cite each other. In application, we found this to work very well, as the graph easily identified papers that essentially investigated the same problem, but for two different geographies, for example. Unfortunately, an API for this service is not (yet) available, which is why we could only manually analyze a subset of our data. More specifically, we randomly chose five clusters and sampled 50 papers from each of them. For each of those papers, we then verified which of the most related articles found on Connected Papers are contained in our sample. Finally, among this subset, we computed the proportion of all papers that are part of the same cluster as the original seed paper. The results are displayed in Table A.3. We see that in all five clusters investigated, at least around 80% of all similar articles are contained in the same cluster. In total, this was true for 1766 of the 2122 connected papers identified ($\approx 83.2\%$). This is remarkable, since these related papers span up to 6.6 journals on average, implying that we have successfully clustered together similar articles even if they were published in different outlets. We have made the DOIs and cluster affiliations of all seed papers and connected papers publicly available here.

Table A.3
Analysis of connected papers.

Cluster	# papers sampled	# journals covered	# connected papers in our sample	% of connected papers in original cluster	Avg. # connected journals
21	50	19	449	92.4%	6.1
23	50	11	227	81.5%	1.9
38	50	26	502	82.1%	6.6
57	50	21	627	79.3%	5.6
98	50	28	317	81.1%	4.2

Table A.3 displays the result from the analyses of cluster consistency using Connected Papers. The table is to be read in the following manner (we use the first row as an example): we randomly sampled 50 papers from cluster 21. Those 50 papers were published in 19 different journals. Of the 2000 (40×50) similar papers found on Connected Papers, 449 were contained in our sample. Of those 449 similar papers, 92.4% are also included in cluster 21, showing a high degree of topical consistency, even though the connected papers were published in 6.1 different journals on average.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.joi.2022.101284.

References

Abritis, A., & McCook, A. (2017). Cash bonuses for peer-reviewed papers go global. *Science*. https://www.sciencemag.org/news/2017/08/cash-bonuses-peer-reviewed-papers-go-global. Accessed: 2021-01-28

Argamon, S., Dodick, J., & Chase, P. (2008). Language use reflects scientific methodology: A corpus-based study of peer-reviewed journal articles. *Scientometrics, 75*(2), 203–238. 10.1007/s11192-007-1768-y.

Armstrong, J. S. (1980). Bafflegab pays. *Psychology Today*, 12.

Armstrong, J. S. (1980). Unintelligible management research and academic prestige. *Interfaces, 10*(2), 80–86.

Armstrong, J. S. (1989). Readability and prestige in scientific journals. *Journal of Information Science, 15*, 123–124.

Bayer, F. M., & Cribari-Neto, F. (2014). Model selection criteria in beta regression with varying dispersion. arXiv preprint arXiv:1405.3718.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Bordons, M., Aparicio, J., & Costas, R. (2013). Heterogeneity of collaboration and its relationship with research impact in a biomedical field. *Scientometrics, 96*, 443–466. 10.1007/s11192-012-0890-7.

Boyack, K. W., & Klavans, R. (2005). Predicting the importance of current papers. In *Proceedings of the 10th international conference of the international society for scientometrics and informetrics: vol. 1* (pp. 335–342). Karolinska University Press Stockholm.

Chen, C. (2012). Predictive effects of structural variation on citation counts. *Journal of the American Society for Information Science and Technology, 63*(3), 431–449. 10.1002/asi.21694.

De Clercq, O., & Hoste, V. (2016). All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics, 42*(3), 457–490. 10.1162/COLI_a_00255.

Didegah, F., & Thelwall, M. (2013). Which factors help authors produce the highest impact research? Collaboration, journal and document properties. *Journal of Informetrics, 7*(4), 861–873. 10.1016/j.joi.2013.08.006.

Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science, 11*(2), 89–121.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd international conference on knowledge discovery and data mining* (pp. 226–231). AAAI Press.

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010). A comparison of features for automatic readability assessment. In *Proceedings of the 23rd international conference on computational linguistics: Posters*. In *COLING '10* (pp. 276–284). USA: Association for Computational Linguistics.

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics, 31*(7), 799–815. 10.1080/0266476042000214501.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association, 97*(458), 611–631. 10.1198/016214502760047131.

Frey, B. S., & Rost, K. (2010). Do rankings reflect research quality? *Journal of Applied Economics, 13*(1), 1–38. 10.1016/S1514-0326(10)60002-5.

Gazni, A. (2011). Are the abstracts of high impact articles more readable? Investigating the evidence from top research institutions in the world. *Journal of Information Science, 37*(3), 273–281. 10.1177/0165551511401658.

Gazni, A., & Didegah, F. (2011). Investigating different types of research collaboration and citation impact: A case study of harvard university's publications. *Scientometrics, 87*(2), 251–265. 10.1007/s11192-011-0343-8.

Gerrish, S. M., & Blei, D. M. (2010). A language-based approach to measuring scholarly impact. In *Proceedings of the 27th international conference on international conference on machine learning*. In *ICML'10* (pp. 375–382). Madison, WI, USA: Omnipress.

González-Pereira, B., Guerrero-Bote, V. P., & Moya-Anegón, F. (2010). A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of Informetrics, 4*(3), 379–391. 10.1016/j.joi.2010.03.002.

Groll, A., Hambuckers, J., Kneib, T., & Umlauf, N. (2018). Lasso-type penalization in the framework of generalized additive models for location, scale and shape. *Working Papers in Economics and Statistics*. Universtität Innsbruck. https://www2.uibk.ac.at/downloads/c4041030/wpaper/2018-16.pdf

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software, 40*(13), 1–30. 10.18637/jss.v040.i13.

Guerrero-Bote, V. P., & Moya-Anegón, F. (2012). A further step forward in measuring journals' scientific prestige: The SJR2 indicator. *Journal of Informetrics, 6*(4), 674–688. 10.1016/j.joi.2012.07.001.

Gunning, R., et al. (1952). *Technique of clear writing*. New York: McGraw-Hill.

Hahsler, M., Hornik, K., & Buchta, C. (2008). Getting things in order: An introduction to the R package seriation. *Journal of Statistical Software, 25*(3), 1–34. 10.18637/jss.v025.i03.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means clustering algorithm. *Applied Statistics, 28*(1), 100–108. 10.2307/2346830.

Hartley, J., Sotto, E., & Pennebaker, J. (2002). Style and substance in psychology: Are influential articles more readable than less influential ones? *Social Studies of Science, 32*(2), 321–334. 10.1177/0306312702032002005.

Hartley, J., Trueman, M., & Meadows, A. J. (1988). Readability and prestige in scientific journals. *Journal of Information Science, 14*(2), 69–75. 10.1177/016555158801400202.

Haslam, N., Ban, L., Kaufmann, L., Loughnan, S., Peters, K., Whelan, J., & Wilson, S. (2008). What makes an article influential? Predicting impact in social and personality psychology. *Scientometrics, 76*, 169–185. 10.1007/s11192-007-1892-8.

Hofner, B., Mayr, A., & Schmid, M. (2016). gamboostLSS: An R package for model building and variable selection in the GAMLSS framework. *Journal of Statistical Software, 74*(1), 1–31. 10.18637/jss.v074.i01.

Hornik, K. (2020). StanfordcoreNLP: Stanford coreNLP annotation. R package version 0.1-6, https://datacube.wu.ac.at.

Hyland, K., & Jiang, F. K. (2017). Is academic writing becoming more informal? *English for Specific Purposes, 45*, 40–51. 10.1016/j.esp.2016.09.001.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*, 241–254.

Klare, G. R. (1963). *The measurement of readability*. Ames, IA: Iowa State University Press.

Koltcov, S., Koltsova, O., & Nikolenko, S. (2014). Latent Dirichlet allocation: Stability and applications to studies of user-generated content. In *Proceedings of the 2014 ACM conference on web science*. In *WebSci '14* (pp. 161–165). New York, NY, USA: Association for Computing Machinery. 10.1145/2615569.2615680.

Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing, 20*(2), 148–161. 10.1016/j.jslw.2011.02.001.

Larivière, V., & Gingras, Y. (2010). On the relationship between interdisciplinarity and scientific impact. *Journal of the American Society for Information Science and Technology, 61*(1), 126–131. 10.1002/asi.21226.

Larivière, V., Gingras, Y., Sugimoto, C. R., & Tsou, A. (2015). Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology, 66*(7), 1323–1332. 10.1002/asi.23266.

Lee, P., West, J. D., & Howe, B. (2018). Viziometrics: Analyzing visual information in the scientific literature. *IEEE Transactions on Big Data, 4*(1), 117–129. 10.1109/TBDATA.2017.2689038.

Lei, L. (2016). When science meets cluttered writing: Adjectives and adverbs in academia revisited. *Scientometrics, 107*(3), 1361–1372. 10.1007/s11192-016-1896-3.

Leimu, R., & Koricheva, J. (2005). What determines the citation frequency of ecological papers? *Trends in Ecology & Evolution, 20*(1), 28–32. 10.1016/j.tree.2004.10.010.

Liu, F., & Kong, Y. (2015). zoib: An R package for Bayesian inference for beta regression and zero/one inflated beta regression. *The R Journal, 7*(2), 34–51. 10.32614/RJ-2015-019.

Lu, C., Bu, Y., Dong, X., Wang, J., Ding, Y., Larivière, V., Sugimoto, C. R., Paul, L., & Zhang, C. (2019). Analyzing linguistic complexity and scientific impact. *Journal of Informetrics, 13*(3), 817–829. 10.1016/j.joi.2019.07.004.

Lu, C., Bu, Y., Wang, J., Ding, Y., Torvik, V., Schnaars, M., & Zhang, C. (2019). Examining scientific writing styles from the perspective of linguistic complexity. *Journal of the Association for Information Science and Technology, 70*(5), 462–475. 10.1002/asi.24126.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability, volume 1: Statistics* (pp. 281–297). Berkeley, Calif.: University of California Press. https://projecteuclid.org/euclid.bsmsp/1200512992

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The stanford CoreNLP natural language processing toolkit. In *Association for computational linguistics (ACL) system demonstrations* (pp. 55–60). http://www.aclweb.org/anthology/P/P14/P14-5010

Mayr, A., Fenske, N., Hofner, B., Kneib, T., & Schmid, M. (2012). Generalized additive models for location, scale and shape for high dimensional data'a flexible approach based on boosting. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 61*(3), 403–427. 10.1111/j.1467-9876.2011.01033.x.

Mubin, O., Tejlavwala, D., Arsalan, M., Ahmad, M., & Simoff, S. (2018). An assessment into the characteristics of award winning papers at CHI. *Scientometrics, 116*(2), 1181–1201. 10.1007/s11192-018-2778-7.

Peng, T.-Q., & Zhu, J. J. H. (2012). Where you publish matters most: A multilevel analysis of factors affecting citations of internet studies. *Journal of the American Society for Information Science and Technology, 63*(9), 1789–1803. 10.1002/asi.22649.

Phan, X.-H., Nguyen, L.-M., & Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on world wide web*. In *WWW '08* (pp. 91–100). New York, NY, USA: Association for Computing Machinery. 10.1145/1367497.1367510.

Rigby, J. (2013). Looking for the impact of peer review: Does count of funding acknowledgements really predict research impact? *Scientometrics*, 57–73. 10.1007/s11192-012-0779-5.

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 54*(3), 507–554. 10.1111/j.1467-9876.2005.00510.x.

Seligman, L. (1986). The manuscript evaluation process used by AACD journals. *Journal of Counseling & Development, 65*(4), 189–192. 10.1002/j.1556-6676.1986.tb01311.x.

Serenko, A., & Bontis, N. (2013). Global ranking of knowledge management and intellectual capital academic journals: 2013 update. *Journal of Knowledge Management, 17*(2), 307–326. 10.1108/JKM-11-2016-0490.

Sienkiewicz, J., & Altmann, E. G. (2016). Impact of lexical and sentiment factors on the popularity of scientific papers. *Royal Society Open Science, 3*(6), 160140. 10.1098/rsos.160140.

Simas, A. B., Barreto-Souza, W., & Rocha, A. V. (2010). Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis, 54*(2), 348–366. 10.1016/j.csda.2009.08.017.

Sooryamoorthy, R. (2009). Do types of collaboration change citation? Collaboration and citation patterns of south african science publications. *Scientometrics, 81*, 177–193. 10.1007/s11192-009-2126-z.

Sternberg, R. J., & Gordeeva, T. (1996). The anatomy of impact: What makes an article influential? *Psychological Science, 7*(2), 69–75. 10.1111/j.1467-9280.1996.tb00332.x.

Stevens, K. T., Stevens, K. C., & Stevens, W. P. (1992). Measuring the readability of business writing: The cloze procedure versus readability formulas. *The Journal of Business Communication (1973), 29*(4), 367–382.

Stremersch, S., Verniers, I., & Verhoef, P. C. (2007). The quest for citations: Drivers of article impact. *Journal of Marketing, 71*(3), 171–193. 10.1509/jmkg.71.3.171.

Thomas, J., Mayr, A., Bischl, B., Schmid, M., Smith, A., & Hofner, B. (2018). Gradient boosting for distributional regression: Faster tuning and improved variable selection via noncyclical updates. *Statistics and Computing, 28*(3), 673–687. 10.1007/s11222-017-9754-6.

Tüselmann, H., Sinkovics, R. R., & Pishchulov, G. (2015). Towards a consolidation of worldwide journal rankings - a classification using random forests and aggregate rating via data envelopment analysis. *Omega, 51*, 11–23. 10.1016/j.omega.2014.08.002.

Uddin, S., & Khan, A. (2016). The impact of author-selected keywords on citation counts. *Journal of Informetrics, 10*(4), 1166–1177. 10.1016/j.joi.2016.10.004.

Umlauf, N., Klein, N., & Zeileis, A. (2018). BAMLSS: Bayesian additive models for location, scale, and shape (and beyond). *Journal of Computational and Graphical Statistics, 27*(3), 612–627. 10.1080/10618600.2017.1407325.

Van Wesel, M., Wyatt, S., & ten Haaf, J. (2014). What a difference a colon makes: How superficial factors influence subsequent citation. *Scientometrics, 98*(3), 1601–1615. 10.1007/s11192-013-1154-x.

Willmott, H. (2011). Journal list fetishism and the perversion of scholarship: Reactivity and the ABS list. *Organization, 18*(4), 429–442. 10.1177/1350508411403532.

Wolff, W. M. (1970). A study of criteria for journal manuscripts. *American Psychologist, 25*(7), 636–639.

Zhao, W., Zhang, R., Lv, Y., & Liu, J. (2014). Variable selection for varying dispersion beta regression model. *Journal of Applied Statistics, 41*(1), 95–108. 10.1080/02664763.2013.830284.

Zimmerman, J. L. (1989). Improving a manuscript's readability and likelihood of publication. *Issues in Accounting Education, 4*(2), 458–466.

Schwendinger, F., Vana, L., & Hornik, K. (2020). Readability prediction: How many features are necessary? Preprint.