Digital Scholarship in the Humanities, 2024, **39**, 13–29
https://doi.org/10.1093/llc/fqad072
Advance access publication 13 October 2023
Full Paper

OXFORD

A quantitative window on the history of statistics: topic-modelling 120 years of *Biometrika*

Nicola Bertoldi [1], Francis Lareau[2], Charles H. Pence [3], Christophe Malaterre [1,*]

[1]Philosophy Department & Centre Interuniversitaire de Recherche sur la Science et la Technologie, Université du Québec à Montréal (UQAM), Montréal, QC H3C 3P8, Canada
[2]Computer Science Department, Université du Québec à Montréal (UQAM), Montréal, QC H2X 3Y7, Canada
[3]Institut Supérieur de Philosophie, Université Catholique de Louvain, Louvain-la-Neuve B-1348, Belgium

*Corresponding author. Département de Philosophie, Centre Interuniversitaire de Recherche sur la Science et la Technologie, Université du Québec à Montréal, 455 Boulevard René-Lévesque Est, Montréal, Québec H3C 3P8, Canada. E-mail: malaterre.christophe@uqam.ca

Abstract

As one of the oldest continuously publishing journals in statistics (published since 1901), *Biometrika* provides a unique window onto the history of statistics and its epistemic development throughout the 20th and the beginning of the 21st centuries. While the early history of the discipline, with the works of key figures, such as Karl Pearson, Francis Galton, or Ronald Fisher, is relatively well known, the later (and longer) episodes of its intellectual development remain understudied. By applying digital tools to the full-text corpus of the journal articles ($N = 5{,}596$), the objective of this study is to provide a novel quantitative exploration of the history of the statistical sciences via an all-encompassing view of 120 years of *Biometrika*. To this aim, topic-modelling analyses are used and provide insights into the epistemic content of the journal and its evolution. Striking changes in the thematic content of the journal are documented and quantified for the first time, from the decline of Pearsonian and Weldonian biometrical research and the journal's tight connection to biology in the 1930s to the rise of modern statistical methods beginning in the 1960s and 1970s. Newly developed approaches are used to infer author networks from publication topics. The resulting network of authors shows the existence of several communities, well-aligned with topic clusters and their evolution through time. It also highlights the role of specific figures over more than a century of publishing history and provides a first window onto the foundation, development, and diverse applications of the statistical sciences.

Keywords: biometrika; pearson; history and philosophy of statistics; topic model; text-mining; LDA.

1. Introduction

The history of statistics, as with any such broad domain of study in the history of science, is a wide and varied field containing a host of different approaches to understanding its source material. A number of particular aspects of that history are now particularly well understood: the development of early statistical methodology (Porter 1986; Stigler 1986, 1999; Hacking 1990); the 'probabilistic' or 'statistical' revolution in the context of 20th-century science, especially early debates over evolutionary theory, eugenics, and statistical physics (Krüger, Daston and Heidelberger 1987; Krüger, Gigerenzer and Morgan 1987; Gigerenzer et al. 1989; Salsburg 2001); and the relationship between statistics and the modern state (Patriarca 1996; Desrosières 2002; Igo 2007; Didier 2020; Ghosh 2020) are among the most impressive examples. A few figures, especially in early statistics, have been the subject of scientific bibliographies, including Adolphe Quetelet (Donnelly 2015; Droesbeke 2021), Francis Galton (Bulmer 2003), and Karl Pearson (Porter 2004), as well as peripheral figures in conflicts over statistical science like William Bateson (Cock and Forsdyke 2022).

These stories are vital and have drastically expanded our understanding of both statistics itself and its uptake across the natural and social sciences. It is our goal in this article to offer a complement to these studies, one which we hope can help scholars in resolving some of the blind spots inherent to the kind of historical methodology that they employ. By relying on close reading, archival work, and, often, a detailed understanding of the social and political contexts of particular periods, classic approaches to the history of science share a common set of 'invisibles'—to borrow the phrasing of J. T. Burman in the history of psychology (Burman 2018)—or features that will tend to be

difficult to detect following their methodology. With respect to the traditional history of science, for example, it can be hard to notice very broad-scale trends or to appreciate the contributions of minor and often-unread figures.

To that end, we turn here to a reconstruction of the history of statistics drawn from an analysis of its published literature (Stigler 1986, 1999; Hacking 1990; Salsburg 2001). The journal *Biometrika* is (along with the *Journal of the Royal Statistical Society*) one of the first publications in the history of statistics and one of the longest running. It has played host to many of the dramatic and sweeping changes that have shaped the discipline of professional statistics in the hundred and twenty years since its founding. *Biometrika* can thus be an invaluable resource for those looking to comprehend the rich and varied history of statistical enquiry, giving us a way to see that history by looking directly at the ways in which the field itself has been practised over time.

To be sure, digital analysis comes with its own set of 'invisibles'. Connections between article content and external pressures on statistics (which, as we know from the literature above, are often significant) may be harder to see. As we will discuss below, our choice of journal may create other blind spots. But an analysis of more than a century of statistical articles will nonetheless, we claim, help us both extend our view of the history of statistics toward the present (many of the extant studies we cited above only cover the 19th or early 20th centuries) and shed light on a host of 'minor' figures in the literature. If nothing else, we hope to help researchers answer the following question: 'How can these tools help you to see what you are interested in such that you can then make better judgments about what to select for further research?' (Burman 2018: 300).

The story of the first years of *Biometrika* is, by now, well known (Elderton 1951; Cox 2001; Aldrich 2013; Pence 2022: 1, 79–80). Karl Pearson and W. F. R. Weldon, two of the pioneers of the use of statistics in the life sciences, had begun to feel as though their usual journals were too constraining. First, those journals were more and more occupied by their opponents in a bitter dispute over evolutionary theory (Froggatt and Nevin 1971; Provine 1971). Furthermore, they would never leave for their authors the space required to discuss statistical methodology, much less to publish tables of statistical constants or the raw data that supported the biometricians' analyses (Editorial: The Scope of Biometrika 1901). Thus began *Biometrika* under the editorship of Karl Pearson, which ran from 1901 until 1936. As Aldrich (2013) has noted, this was not a sense of 'journal editor' familiar to readers today: Pearson not only directly wrote several hundred articles, but he also was almost always the sole reviewer, often freely edited received contributions, sought submissions from others in the field, or inspired the choice of topic or methods used. This means that these first decades of *Biometrika* bear, as we will see below, the indelible mark of Karl Pearson's unique interests: eugenics, craniometry, the biological theory of inheritance, the method of curve fitting by moments, and so forth. This came to an end with Karl Pearson's death in 1936, at which point the editorship passed to Karl's son Egon S. Pearson (Elderton 1951). As we will again see confirmed below, this change in editorship marks a shift in the stated focus of the journal. What had begun as a journal dedicated to 'problems which depend for their solution on a study of the differences between individual members of a race or species' (Editorial: The Scope of Biometrika 1901: 1) would quickly become a journal of statistical method—now, one written by a second generation of scholars, many of whom had learned statistics in the first place from articles published in *Biometrika* (Aldrich 2013: 13). In 1966, the editorship passed to David R. Cox, who held the position until 1990. (Shorter periods of editorship commenced after Cox stepped aside from the position, with five editors since 1991.) The journal celebrated its 100th anniversary in 2001, with a series of papers published concerning not only the journal's general history, but also targeted discussions of particular areas as they were found both within and beyond the journal's pages, including general methodology (Davison 2001), time series (Tong 2001), survival analysis (Oakes 2001), and others (Atkinson 2001; Hall 2001; Smith 2001).

Put briefly, *Biometrika* is itself a window into the historical development of statistics as an independent discipline.[1] The challenge, of course, is that a quantitative look at the content of *Biometrika* requires engaging with the nearly nine thousand articles that have been published between 1901 and the present day. Our goal here is to surmount this issue by turning to the tools of the digital humanities, which, we argue, will give us a way in which to begin to build a quantitatively grounded history of statistics derived directly from the content of *Biometrika*'s archives. Particularly useful for this purpose are topic modelling algorithms, that is 'statistical methods that analyse the words of the original texts [contained in large archives of documents] to discover the themes that run through them, how those themes are connected to each other, and how they change over time' (Blei 2012). It is precisely for this reason that, in the last 15 years, topic modelling algorithms have been applied to analysing the archives of a wide range of scholarly journals, such as *Science* (Blei and Lafferty 2006, 2007), *Cognition* (Cohen Priva and Austerweil, 2015), the *Journal of the*

History of Biology (Peirson et al. 2017), the *Proceedings of the Cognitive Science Society* (Rothe, Rich and Li 2018), as well as groups of journals, for instance in philosophy of science (Malaterre et al. 2021) or in bioethics (Bystranowski, Dranseika and Żuradzki 2022).

We thus aim here to provide a 'distant reading' of *Biometrika*'s first 120 years through synchronic, diachronic, and author-based topic modelling with a view to answering such questions as: What kinds of topics can be identified over the history of *Biometrika*? How have they changed over time? How can we use these topics to understand the communities of authors who have written about them? This article is structured as follows. Section 2 presents the corpus and the methods that were used. Section 3 describes the different topics that were found in the corpus as revealed by the topic-modelling analyses we conducted. The diachronic evolution of these topics throughout *Biometrika*'s 120 years is the focus of Section 4. Section 5 then analyses topic-based author correlations and reveals the existence of different communities of authors around specific shared research interests that parallel the temporal evolution of the journal topics. Section 6 summarizes the insights on the history of statistics that were gained by applying topic-modelling approaches to 120 years of *Biometrika* and offers perspectives for further research.

2. Methods

Computational textual analyses take as a starting point the fact that words are not used at random when mobilized to convey meaning in texts. Instead, they usually form recognizable associative patterns, hence the intuition that analysing these patterns may provide information about the semantic content of the texts in which they occur. As one of the pioneers of distributional semantics wrote, 'you shall know a word by the company it keeps' (Firth 1957: 11). Algorithmic text-mining methods that build on this intuition and search for word patterns in digitized texts have been found to be extremely effective (e.g. Srivastava and Sahami 2009; Aggarwal 2015). One such method, called 'topic-modelling', explores the conjoined presence of sets of terms across texts in a given corpus, making it possible to retrieve information about the thematic content of specific texts (Blei and Lafferty 2009). Adding metadata, such as publication dates and author names, then makes it possible to carry out diachronic analyses about topic changes over time and to build author networks based on topic similarities in their respective publications. Such is the overall approach that we used to analyse the textual content of *Biometrika*. More specifically, the research design we implemented consists of five major steps (Fig. 1).

2.1 Corpus assembly and cleaning

The corpus of full-text *Biometrika* articles was assembled from two sources: JSTOR for all articles from 1901 to 2013, and the publisher's website (Oxford University Press) for later articles. A total of 8,242 documents were thereby collected. Removal of editorials, book reviews, errata, front and back matters, as well as any document shorter than 4,000 characters resulted in a corpus of 5,596 articles that we considered to consist only of research articles. All documents were cleaned with standard procedures, including the removal of HTML tags, mathematical formulas, abstracts, footnotes, and lists of references. With the assistance of language detection algorithms, twenty-one articles were identified as written in French or German. These articles were machine-translated into English with DeepL by chunks of approximately 5,000 characters (https://www.deepl.com/translator). Machine translation tools have indeed been shown to be very reliable for bag-of-words text-mining approaches, all the more so when terms are lemmatized as in step 2 below (Lucas et al. 2015; de Vries, Schoonvelde and Schumacher 2018; Malaterre and Lareau 2022). In parallel, authors were manually disambiguated and curated (in particular when first names had been abbreviated, possibly under different forms). This resulted in a list of 4,490 unique authors.

2.2 Preprocessing

Preprocessing was done in a standard way so as to reduce noise and optimize the size of the data retained for analysis. Numbers were removed, words were tokenized and those that included at least one non-ascii letter were removed (reducing noise due to numerous tables and formulas). Stop-words, such as determinants, prepositions, conjunctions, or pronouns, were also removed with the assistance of a part-of-speech tagging algorithm for identifying the morphosyntactic category of every word in the corpus. Only nouns, verbs, modals, adjectives, adverbs, proper nouns, and foreign words were kept. All textual data were then lemmatized. To carry out these operations, we used TreeTagger (Schmid 1994) along with Penn TreeBank (Marcus, Marcinkiewicz and Santorini 1993). In addition, words shorter than three characters as well as words that occurred in fewer than twenty articles in the corpus were removed. These operations resulted in a total of 7,578,267 word tokens distributed among the 5,596 research articles of the corpus, corresponding to a lexicon of 9,073 unique terms. The resulting preprocessed corpus was then vectorized, resulting in the construction of a term-document frequency matrix.

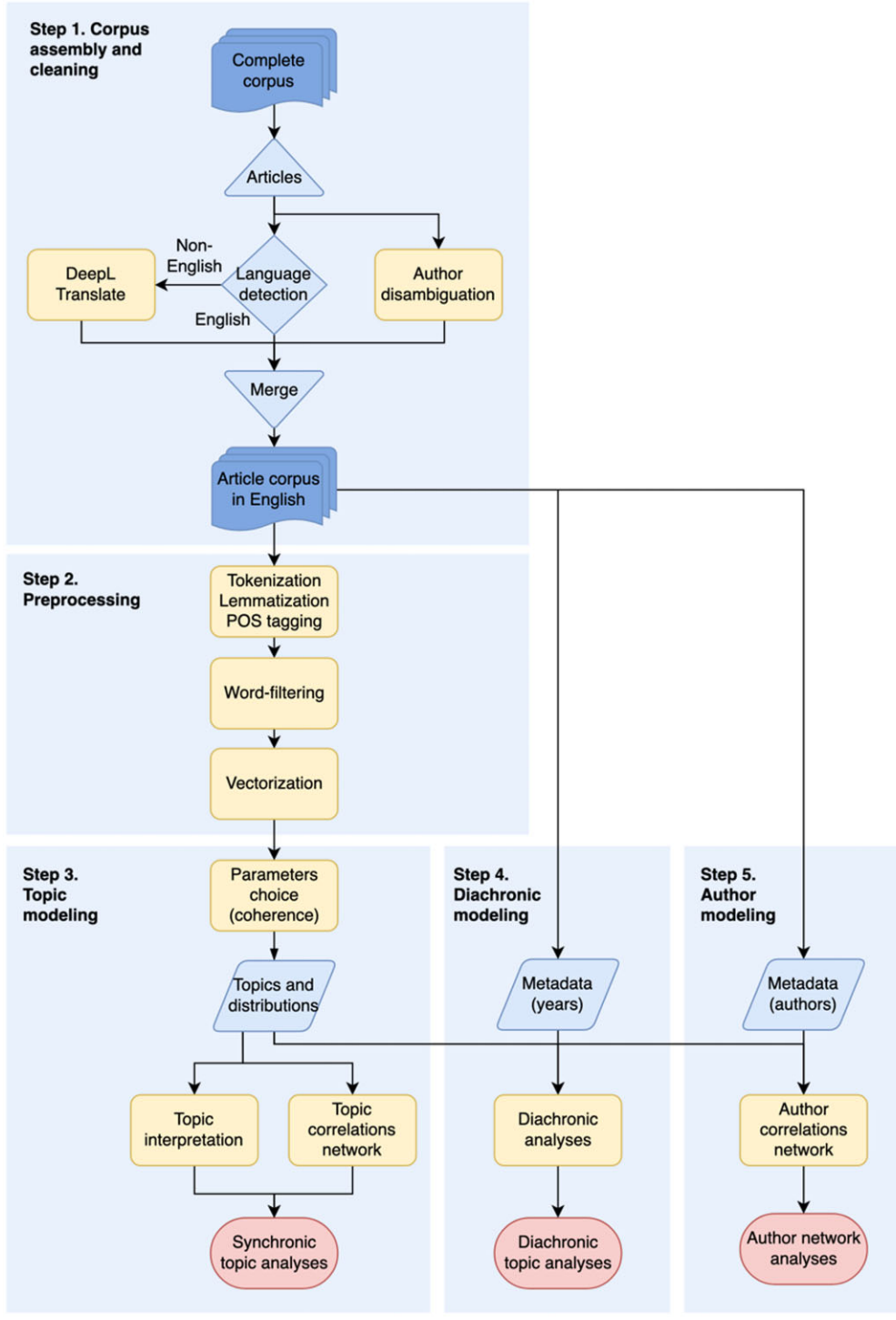

Figure 1. Research design. Five major steps, from corpus assembly to topic and author modelling (textual corpora in dark blue, data in light blue, operations in orange, results in red).

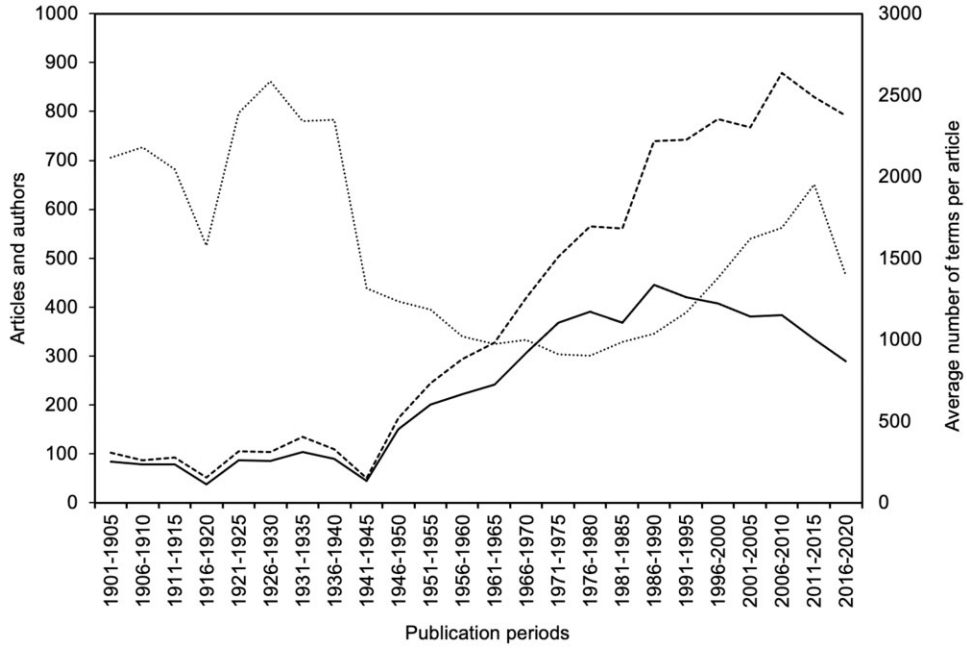

Figure 2. The *Biometrika* corpus. The graph shows the number of articles (solid) and authors (dash) per time period (left-side *y*-axis) as well as the average number of terms per article (after preprocessing, dot, right-side *y*-axis).

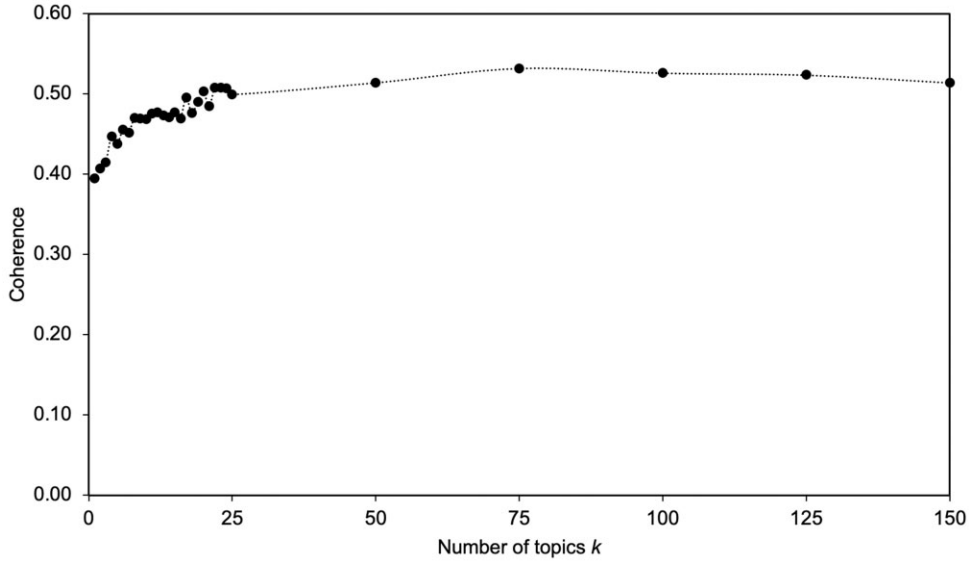

Figure 3. Topic model coherence as a function of the number k of topics. For $k \leq 25$, coherence was calculated for all k values; a local optimum was found for $k = 23$ with a coherence $c = 0.51$. For $k \geq 25$, coherence was calculated by k increments of 25; an optimum was found for $k = 75$ with a coherence $c = 0.53$. A trade-off between maximizing coherence and facilitating interpretation of a smaller number of topics led to a choice of $k = 23$ (in the elbow region of the curve).

2.3 Topic-modelling

Topic-modelling was done using the well-tested latent Dirichlet allocation (LDA) algorithm, following Blei, Ng and Jordan (2003) and Griffiths and Steyvers (2004) and performed through an API for Python (https://pythonhosted.org/lda/api.html). LDA is a generative probabilistic computational method that infers topics from the distribution frequencies of terms within documents. Topics are modelled as probability distributions over all terms of the lexicon, while documents are modelled as probability distributions over all topics. As a heuristic for choosing hyperparameter values, notably the number of topics (k) and the Dirichlet priors (alpha, beta), we explored the space of topic coherence measures following Röder, Both and Hinneburg (2015), as shown on Fig. 3. This led us to favour a model with high coherence yet a relatively low number of topics in order to facilitate topic interpretation ($k = 23$, $\alpha = 0.22$, $\beta = 0.01$). The number of topics $k = 23$ was further validated by expert analyses of the top words and top articles (comparing with different models up to 150 topics).[2]

Carrying out the LDA modelling on this basis thereby led to twenty-three topics, each one defined by a specific word probability distribution, and 5,596 probability distributions over the twenty-three topics, each of these later distributions corresponding to one article. All twenty-three topics were interpreted by looking at the most probable words within each topic, as well as by examining the original articles in which the topic was the most likely. Since this interpretation sometimes resulted in lengthy phrases hard to summarize in catchy labels without risking misrepresenting the topic, we preferred to give each topic a short label composed of the two to three most meaningful terms among its top-10 words, alongside a more detailed interpretation. In doing so, we exercised as little personal judgement as possible by starting from the most strongly associated term and proceeding down the list. We eliminated terms whenever we could safely assume that they very likely formed compounds with preceding terms or that they did not provide any additional information about the content of the topic relative to preceding terms. For instance, the topic that we ended up labelling 'Design-block' included top-10 words (such as 'design', 'block', 'treatment', 'factor', and 'effect'), which appeared to be about randomized block design. Inspection of the articles in which the topic had the strongest probability of being present revealed research on various aspects of randomized block designs and how to control for the influences of multiple factors on the effects of alternative experimental treatments. This corroborated our interpretation. All topics were interpreted following this approach. Finally, to assess whether some topics tended to simultaneously occur in articles, topic correlation in documents was calculated (Pearson coefficient). The topic correlation network was built with Gephi (Bastian, Heymann and Jacomy 2009). We then used Louvain community detection for partitioning the topics into clusters, following (Blondel et al. 2008) as implemented in Gephi. This resulted in identifying four distinct topic clusters (each identified by a capital letter added to topic labels).

2.4 Diachronic modelling

In order to provide a diachronic view of topic evolution over time, article publication dates were taken into account. These dates were grouped into time periods of 5 years so as to average out yearly fluctuations. The corpus was thus segmented into twenty-four periods from 1901 to 2020. For each period, topic probabilities were averaged over all articles published in that period. This resulted in twenty-four averaged topic probability distributions depicting the relative abundance of all topics in each one of the twenty-four time periods.

2.5 Author modelling

Author topic profiles were computed by averaging the topic probability distributions of their respective articles. When an article was written by several authors, its topic probability distribution was assigned a weight of 1 over the number of authors (in other words, its topic probability distribution was evenly shared among co-authors). Author correlations were then calculated on the basis of their topic profiles (Pearson coefficient), and the author correlation network was built on Gephi (for author publication weight ≥ 1 and $r \geq 0.7$, so as to reduce noise and clutter). The network thus inferred depicts the closeness of the most prolific authors on the basis of their topic profiles, revealing latent groups or 'hidden communities' of authors with shared research interests as inferred from the topic profiles of their respective publications (note that despite exhibiting similar terminological patterns these authors need not agree with one another).

3. The twenty-three topics of *Biometrika*

The topics that resulted from our topic-modelling analyses reveal the diversity of research work that has been published in *Biometrika*, from studies in biometry to articles on formal statistics. Table 1 provides the list of the twenty-three topics, with their top-10 words and

Table 1. List of topics with their top-10 words (ordered by decreasing order of probability in topic).

Topic label	Interpretation	Top-10 words
A-Age-population	From vital statistics to mathematical statistics	age; population; year; number; individual; time; period; rate; death; total
A-Colour-plant	Heredity of Mendelian (discrete and qualitative) traits	colour; plant; number; offspring; parent; eye; white; character; hair; result
A-Correlation-mean	Study of correlations within varieties and between physical or psychological characters and physical or mental capacities	correlation; total; length; mean; age; difference; weight; character; series; variation
A-Pearson-biometrika	History of statistics (including Biometrika) and archaeological application of statistics	pearson; year; biometrika; head; man; time; find; know; statistical; theory
A-Skull-measurement	Craniometry, that is, the study of the relations among measurements of different components of the human skull	skull; series; suture; measurement; bone; sagittal; character; type; male; close
A-Value-sample-mean	Mathematical statistics in the Pearsonian tradition	value; sample; curve; mean; population; frequency; distribution; find; deviation; standard
B-Design-block	Randomized block designs	design; block; treatment; factor; effect; column; row; interaction; square; balanced
B-Distribution-approximation	Statistical approximation	distribution; value; approximation; function; term; obtain; moment; normal; probability; point
B-Distribution-dependence	Analysis of heterogeneous data or extreme values	distribution; pair; bivariate; rank; dependence; correlation; measure; normal; datum; outlier
B-Sample-population	Sample selection	sample; sampling; probability; population; size; procedure; value; rule; sequential; trial
B-Test-statistic	Formal properties of statistical tests	test; statistic; hypothesis; distribution; power; sample; null; level; alternative; value
B-Value-estimate	Estimation of quantitative relations between observable variables	value; estimate; number; method; observation; example; set; problem; probability; form
C-Likelihood-parameter	Methods for estimating parameters of distributions and models through various types of likelihood functions, especially in the presence of nuisance parameters	likelihood; log; parameter; maximum; function; distribution; model; conditional; estimate; density
C-Matrix-covariance	Methods for analyzing multidimensional associations	matrix; vector; covariance; element; component; correlation; linear; analysis; column; variable
C-Method-function	Statistical methods for estimating functions, especially in machine learning	method; function; datum; rate; smooth; kernel; estimator; algorithm; estimate; propose
C-Model-regression	Regression models, especially linear ones	model; regression; linear; variable; parameter; fit; residual; error; function; datum
C-Prior-posterior-bayesian	Bayesian methods for building and selecting models, estimating their parameters and fitting them to empirical data	prior; distribution; posterior; model; density; bayesian; probability; algorithm; parameter; bayes
C-Process-time-series	Methods for studying the properties of stochastic processes in time	process; time; series; model; state; stationary; autoregressive; order; estimate; function
C-Region-distance	Intersections between geometry and statistics	point; region; line; distance; space; plane; shape; direction; area; sin
C-Theorem-function	Theorems on probability distributions, random variables, statistics as functions of data, statistical models and algorithms	theorem; function; condition; follow; probability; proof; distribution; result; define; variable
D-Estimate-variance	Parametric and nonparametric methods for estimating the values of parameters and constructing confidence intervals	estimator; estimate; variance; sample; mean; error; bias; method; asymptotic; confidence
D-Model-effect	Statistical analysis of causal relations in data from observational studies or randomized trials	model; effect; datum; study; variable; cluster; treatment; response; covariates; assumption
D-Time-censor	Methods for studying time-dependent processes in the case of censored data, especially with respect to survival analysis	time; censor; function; model; hazard; survival; datum; failure; estimate; estimator

interpretation (more details about each topic, notably their top articles, can be found in Supplementary Table S1). The twenty-three topics were grouped into four clusters as a result of the modularity analyses that we conducted on the graph of topic correlations as depicted in Fig. 4.

3.1 Pearsonian mathematical statistics

Cluster A is composed of topics that relate to various aspects of biometry, which was defined by Francis Galton as 'the application to biology of the modern methods of statistics' (Galton 1901: 7–8), and was one major area of investigation of Karl Pearson. As such, these topics are concerned with what can be called Pearsonian mathematical statistics and its applications to the study of heredity, demography, craniometry, etc.

Among the six topics of this cluster, two appear to cover the biological aspects of biometry. First, the topic 'A-Colour-plant' depicts articles about the study of Mendelian inheritance (anchored in a discrete and qualitative conception of heritable traits), concerning, for instance, Mendel's theory (see typical examples of *Biometrika* articles: A-Colour-plant.9, 12, 17 as listed in Supplementary Table S1),[3] the transmission of traits such as iris or skin pigmentation in humans and animals (A-Colour-plant.2) or the outcomes of hybridization experiments (A-Colour-plant.1, 3). Secondly, the topic 'A-Skull-measurement' concerns the statistical aspects of craniometry, be they be about paleo-anthropological findings (A-Skull-measurement.1, 2, 3) or about race differences in cranial types (A-Skull-measurement.6, 9, 13), underlining the tight connection

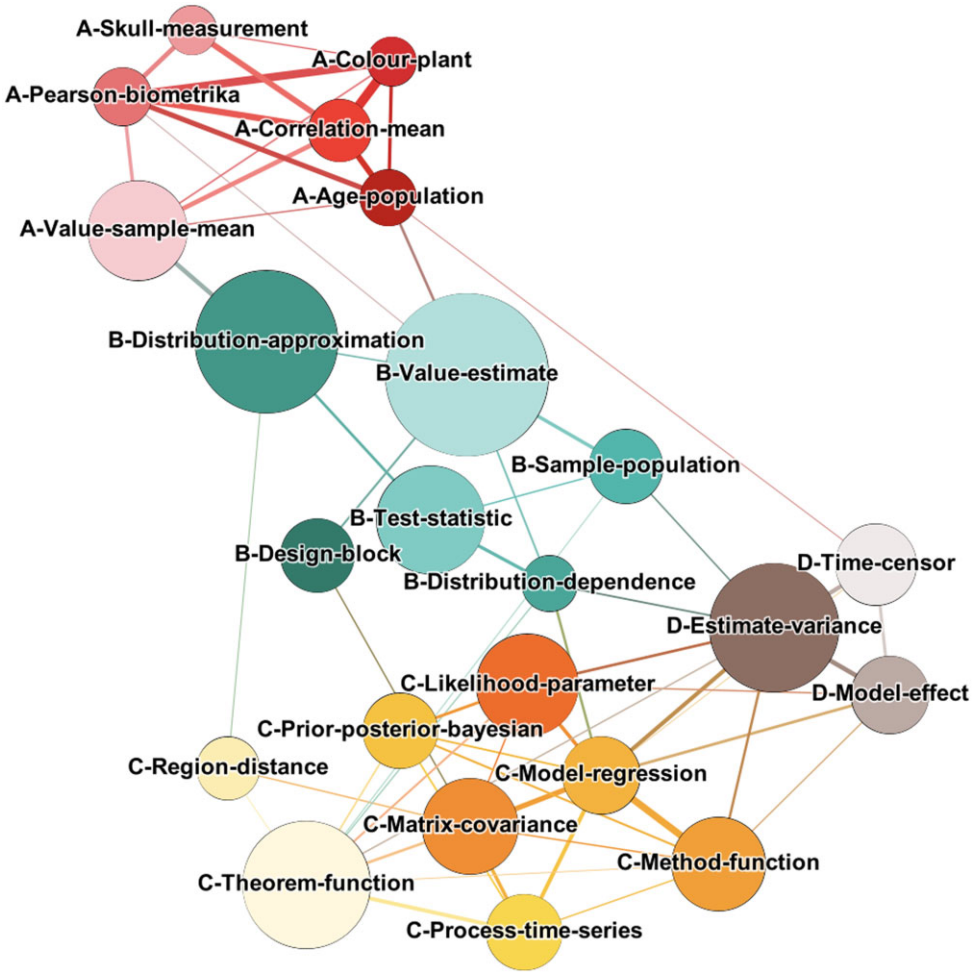

Figure 4. Topic correlation network. Node size proportional to topic probability across the whole corpus; edge thickness proportional to topic correlations in articles; clusters denoted by letters preceding topic labels and four sets of colour shades attributed to topics in alphabetic order within each cluster [graph made with Gephi (Bastian, Heymann and Jacomy 2009) with ForceAtlas 2 for rendering]. Note that the same shades of colours are assigned to the same topics in the following figures.

between the early days of statistics and research into eugenics.

Note how the topic 'A-Pearson-biometrika' denotes articles of a quite distinctive historical nature. These papers concern the history of statistics and probability theory, including the lives and legacies of prominent figures and the journal itself.

The three remaining topics of cluster A are related to the more mathematical facets of the biometrical school. The topic 'A-Correlation-mean' concerns the study of trait correlations between characters or varieties of given species. The topic 'A-Value-sample-mean' denotes articles that address distributions drawn from samples of different sizes (e.g. A-Value-sample-mean.2) and the analysis of the degree of association between pairs of variables of various kinds (A-Value-sample-mean.1, 3). As for the topic 'A-Age-population', it reveals articles that apply statistical methods to analyse variations of different characteristics of a given population (e.g. death and fertility rates) as a function of population age structure.

Altogether, topics of cluster A characterize the development of mathematical statistics in the Pearsonian tradition and its application to a wide range of domains, from heredity to epidemiology, including demography or even craniology, as has been studied elsewhere (Stigler 1986; Magnello 2009). The structure of the topic correlation network (see Fig. 4) shows that topics about the mathematical and statistical facets of the biometrical school ('A-Value-sample-mean', 'A-Age-population') occupy a more central position compared to topics about the biological aspects of biometry ('A-Colour-plant', 'A-Skull-measurement'), in the sense that the latter appear to connect cluster A to cluster B.

3.2 Transition to modern statistics

Topics in cluster B denote a transition from Pearsonian statistics to what may be called 'modern statistics'. Topics 'B-Distribution-approximation' and 'B-Value-estimate', which sit at the borderline between clusters A and B, may be interpreted as developments from the earlier biometrical school. Articles most correlated with 'B-Distribution-approximation' discuss mathematical transformations (B-Distribution-approximation.2–4, 9) or recurrence relations (B-Distribution-approximation.20). In turn, articles most strongly associated with 'B-Value-estimate' discuss mathematical problems about estimating the quantitative relationships between observable variables in experimental settings, especially agricultural (B-Value-estimate.4, 8, 15), psychological (B-Value-estimate.5), and biological (B-Value-estimate.8, 13, 14) experiments.

At the centre of cluster B, the topic 'B-Test-statistic' likely characterizes the important place occupied, in modern statistics, by the theory of hypothesis testing, especially as developed by Egon S. Pearson and Jerzy Neyman (Hall 2001). Some of the most representative articles for this topic deal precisely with the formal properties of procedures for testing alternative hypotheses (B-Test-statistic.2, 6, 9, 16) or normality (B-Test-statistic.5, 7).

Topics 'B-Design-block' and 'B-Sample-population' occupy a less central place within the cluster. While 'B-Design-block' concerns various aspects of randomized block design and questions about how to control for the influences of multiple factors on the effects of alternative experimental treatments, the topic 'B-Sample-population' relates to the selection of populations and samples according to certain preferred characteristics (e.g. optimal size in clinical trials). Finally, the topic 'B-Distribution-dependence' appears to specifically concern analyses of heterogeneous data (B-Distribution-dependence.12) or data with extreme values (B-Distribution-dependence.3, 4, 6, 8, 15, 16). Quite removed from Pearsonian mathematical statistics and cluster A, this latter topic provides multiple connection points between cluster B on the one hand, and clusters C and D on the other, indicating the directions for further developments in modern statistics.

3.3 Modern statistics and probability theory

Cluster C contains topics that are closely related to the process of mathematization that has characterized the history of modern statistics and probability theory. Some of these topics correspond to the generalization of methods that were already part of Pearsonian statistics, such as the topic 'C-Model-regression', which concerns issues about the choice and formulation of regression models, parameter estimation, and model fitting, as well as the topic 'C-Matrix-covariance' for which the most representative articles generally deal with multivariate analysis methods.

In contrast, other topics, such as 'C-Method-function', 'C-Prior-posterior-Bayesian', and 'C-Likelihood-parameter' capture themes that were initially marginal within *Biometrika* but became more prevalent starting in the late 1970s and early 1980s (Davison 2001; Hall 2001). Representative articles of 'C-Method-function' concern the application of statistical methods for estimating various kinds of mathematical functions, such as regression functions (C-Method-function.1, 11, 14–16) and functions describing networks (C-Method-function.3, 7, 19). Some recent articles also tend to cover machine learning applications. Articles strongly associated with 'C-Prior-posterior-Bayesian' generally concern the application of Bayesian methods to building, selecting, estimating, and fitting statistical models. As for the 'C-Likelihood-parameter', the topic denotes articles about the properties of various kinds

of likelihood functions that can be used for estimating distribution and model parameters in maximum likelihood estimation frameworks.

Among the three remaining topics in cluster C, 'C-Theorem-function' appears to characterize articles that present theoretical results in the fields of statistics and probability theory, including various theorems and proofs, for instance, about the use of Student's law to identify the necessary and sufficient conditions for a certain random variable to be normally distributed or about the existence of maximum likelihood estimates for some generalized linear models (C-Theorem-function.2, 11). The topic 'C-Process-time-series' captures works about the properties of stochastic processes that evolve in time, notably through time series analysis (Tong 2001). Finally, 'C-Region-distance' is associated with articles discussing various subthemes whose common feature is the interplay between statistics and geometry, for instance, statistical analysis of spatial patterns or the use of statistics and probability theory for solving geometric problems (C-Region-distance.18–20).

3.4 Further refinements

Last, topics in cluster D deal with very specific subjects and methods that can be regarded as further refinements of those addressed in clusters B and C. Most fundamentally, they constitute methods that are newer than those found in cluster C and were thus innovations that appeared later in the history of the journal. For instance, the topic 'D-Estimate-variance' corresponds to articles that focus on parametric and non-parametric methods for finding good parameter estimators, a domain of research described by Hall (2001). The topic 'D-Model-effect' generally denotes articles that are concerned with problems related to causal inference, for instance using causal diagrams or instrumental variables (D-Model-effect.7, 8, 9). As for 'D-Time-censor', the topic is mostly associated with articles studying time-dependent processes, especially in the context of survival analyses as identified by Oakes (2001).

More generally, the fact that clusters C and D are tightly connected with one another seems to indicate that the topics of both clusters capture more up-to-date (relative to current standards) issues compared to those of clusters A and B. The relations of these topics to one another are thus much tighter than those to their more historical counterparts in clusters A and B. Note how the topic that constitutes their shared barycentre is 'C-Model-regression'. This could indicate the continued relevance of regression models as a subject of discussion in *Biometrika*.

4. Evolution of the topics over time

Starting in 1901, the corpus of *Biometrika* provides a large window of 120 years to study the evolution of research in statistics. Having segmented the corpus into twenty-four time periods of 5 years, we quantified the relative probability of finding the twenty-three topics of the topic-model within each time period. The results, which are synthesized in Fig. 5, show drastic changes in topic prevalence over time: Most of the dominant topics in the early decades of *Biometrika* have nearly totally disappeared by the end of the 20th century, being replaced by others in the 1940s, which suffered a quite similar fate in the 1980s, leaving room to yet other sets of topics in the last 40–50 years.

The topics which largely defined the content of *Biometrika* throughout the entirety of Karl Pearson's editorship (1901–1936) are those of cluster A. Their aggregate probability was roughly 90% in each of the three time periods from 1901 to 1915, then started to wane, especially between 1936 and 1940. These topics concern various aspects of biometry and the application of Pearsonian mathematical statistics to biology. As can be seen in Fig. 5, the 1936–1940 period appears to have constituted a turning point in the evolution of topics associated with cluster A, such as (among others) 'A-Colour-plant', which captures themes concerning Mendelian heredity, but also 'A-Skull-measurement' about craniometric studies. A possible explanation could be that Egon Pearson may have been trying to minimize the eugenic connotations of his father's research, given that Karl Pearson's death in 1936 occurred at the same time that eugenics was becoming politically unpalatable (Porter 2004). Concomitantly, we can highlight an increase in the average probability of the topic 'A-Value-sample-mean', most strongly correlated with articles discussing themes in Pearsonian mathematical statistics, the only cluster-A topic to outlast (for a decade or so) the period of Karl Pearson's narrower focus. By the end of the 1950s, it was clear that *Biometrika* had irretrievably changed course from the one Karl Pearson had charted (Aldrich 2013).

Topics in cluster B have followed a trajectory that is mainly complementary to the one defined by the topics of cluster A. Their prevalence in the corpus started to grow precisely when the topics of cluster A began to dwindle, that is in the late 1930s. By the end of World War II, cluster B surpassed cluster A as the largest topic cluster and prevailed throughout the 1960s and 1970s. Two topics have clearly dominated cluster B ever since the late 1930s: 'B-Distribution-approximation' and 'B-Value-estimate', which can be interpreted as developments from the more mathematical and statistical aspects of the biometrical school. The topic 'B-Test-statistic' emerged later in the 1950s. The years in which

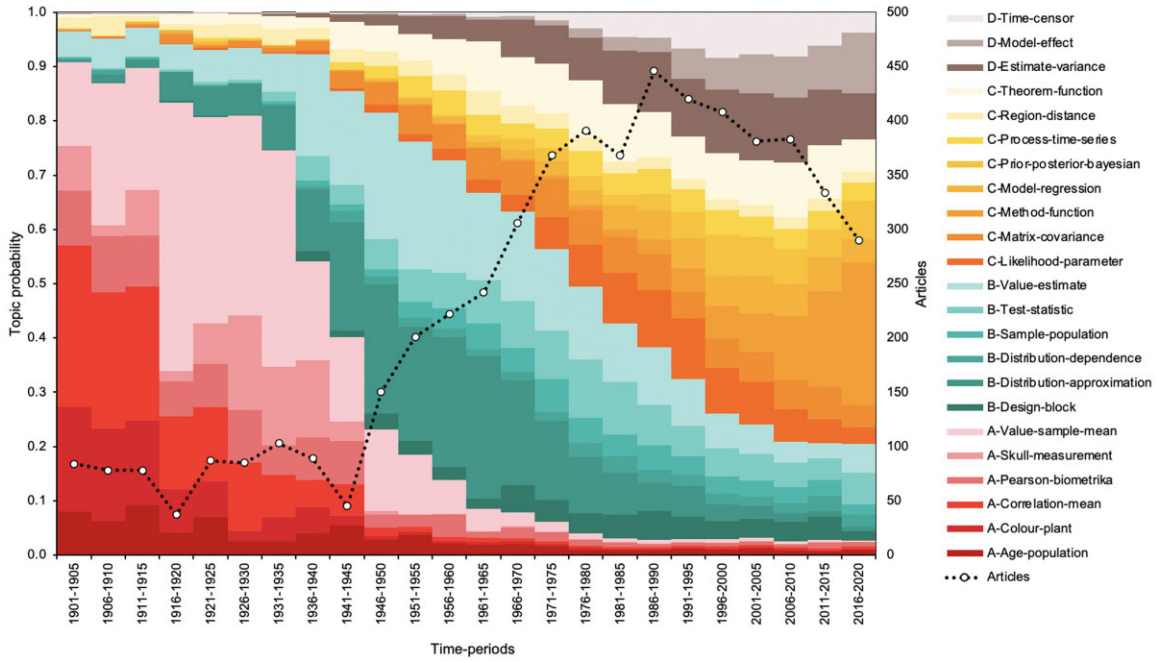

Figure 5. Time evolution of the topics of *Biometrika* (1901–2020). Time periods of 5 years on the *x*-axis; averaged topic probability over all publications of given time periods on the left *y*-axis (colour shades); number of *Biometrika* research articles per time period on the right *y*-axis (dotted line).

cluster B gained and retained prominence coincided with Egon Pearson's editorship and (roughly) the first half of D. R. Cox's. This feature of the diachronic topic model is thus consistent with Cox's claim that 'the character of *Biometrika* changed appreciably' during this period (Cox 2001: 8). However, the aggregate probability of topics in cluster B has been slowly but steadily declining since the end of the 1970s, and, in the first decades of the 21st century, those topics accounted for no more than 20% of the total. Research interests moved on to other questions, notably those depicted by topics in cluster C, which have been dominant in *Biometrika* since the 1980s.

By the 1970s, three topics of cluster C had become most significant: first, 'C-Theorem-function' and 'C-Matrix-covariance', then 'C-Likelihood-parameter'. In addition, three distinct trends seem to have characterized the period from the 1980s onwards: (1) The continued rise of topics 'C-Likelihood-parameter' and 'C-Theorem-function' that took place between 1981 and 2000, denoting a shift towards more theoretical questions, as already noted by Davison (2001: 16). (2) The emergence of 'C-Prior-posterior-Bayesian' reveals the growing importance of Bayesian themes in *Biometrika*, especially during the second half of Cox's editorship (Cox 2001). (3) The growing prominence of the topic 'C-Method-function', which has become the overall dominant topic in *Biometrika* since the 2000s, plausibly indicating a shift towards empirical applications and machine learning.

The fourth cluster, cluster D, has only grown relevant in the last 30 years. Topic 'D-Estimate-variance', which is about parametric and non-parametric methods for selecting parameter estimators, can be connected to the shift towards theoretical issues mentioned above. The topic 'D-Time-censor' was especially prevalent in the 1990s and 2000s and denotes themes associated with survival analysis (Oakes 2001). As for 'D-Model-effect', its prevalence has been growing since the 1990s and is now one of the most dominant topics in *Biometrika*. This may also have to do with the shift towards empirical topics discussed above.

The overall increase in the probability of clusters C and D from the 1970s onwards can be seen as reflecting the challenges that the growing availability of large bodies of data has come to pose to statistical analysis. This topic trend captures *Biometrika's* shift away from what might be called 'specifically biometrical themes and problems', represented by cluster A, to more general statistical and mathematical ones, such as, for instance, the formal development of maximum likelihood estimation and Bayesian inference. As noted by Davison (2001), throughout the 20th century, *Biometrika* has steadily moved away from its initial focus on the collection and statistical analysis of biological data to become 'a journal of statistics in which

emphasis is placed on papers containing original theoretical contributions of direct or potential value in application' (Davison 2001: 13). In this context, topics of cluster B may have played a pivotal role, both thematically and chronologically, in bringing about such a shift.

5. Authors and their topic networks

Biometrika would not be *Biometrika* without the hundreds of authors who have contributed their research papers since its launch in 1901. Topic analysis helps shed light not only on the content of the journal but also on the profiles of these contributing authors and their relatedness. Since topic probabilities have been computed for every article, and since article authors are also known, specific topic profiles can be calculated for each author. Measuring the correlations between author topic profiles then reveals the proximity of authors to one another. Applying this approach to the 4,490 authors of the corpus resulted in a large correlation network. Figure 6 is a subset of this network in which only the most prolific 1,924 authors are represented and coloured depending on their dominant topics (authors contributing together up to 80% of articles).[4] Two features are notable. First, the four topic clusters identified above (Section 3) correspond well with the author clusters. This indicates that authors not only tended to address just a few select topics in their research but also that they tended to address in a group-like fashion the same sets of topics as those that were identified as forming clusters. Furthermore, some authors stand out as playing a pivotal role in linking topics and clusters. Bearing in mind how topics have changed over time in the corpus (Section 4), one witnesses parallel changes in authorship, moving from the era of Pearsonian biometry and mathematical statistics (topics of cluster A) to groups of authors having addressed or still addressing topics of more contemporary relevance (clusters C and D, after transitioning through cluster B). The second striking feature is the significance of Karl Pearson's contribution and, though to a lesser degree, those of Egon Pearson, D. R. Cox, and Peter Hall. These four authors indeed stand out for their productivity. Karl Pearson's contribution to *Biometrika* is represented by the largest node in the entire network. This feature reflects Pearson's leading role during the first 35 years of the journal (Cox 2001; Aldrich 2013). Egon Pearson and Cox, who served as editors from 1936 to 1966 and from 1966 to 1990, respectively, also left a significant legacy in terms of publications in the journal, as did Peter Hall who, contrary to Pearson and Cox, was not an editor of the most prolific and influential statisticians of his generation, whose impact 'has had a profound effect on much of modern mathematical statistics' (Robinson and

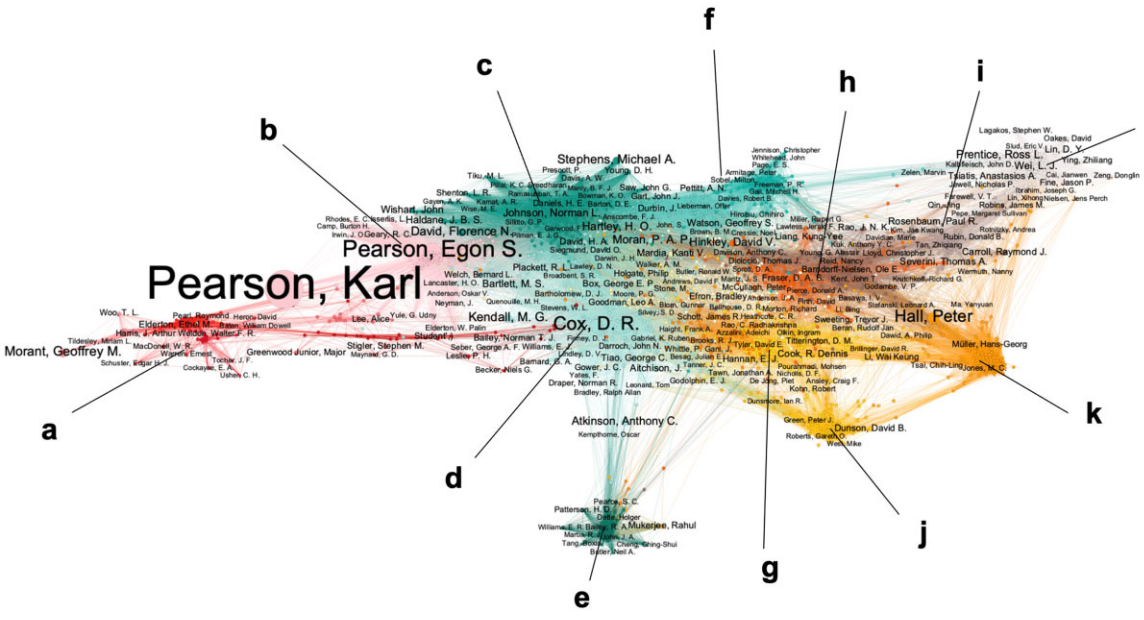

Figure 6. Authors and their topic networks in *Biometrika* (1901–2020). Nodes represent authors; node size and name size are proportional to author contribution to articles; only authors with contribution ≥ 1 are represented (1,924 authors = top 42% of authors responsible for 4,477 articles or 80% of all corpus articles); only authors with contribution ≥ 4 are labelled (224 authors = top 5% of authors); node colour corresponds to author dominant topic (averaged over all of their publications). Edge thickness is proportional to author correlations in terms of topics; correlation threshold ≥ 0.7 (graph made with Gephi with ForceAtlas 2 for rendering).

Welsh 2018: 209). In part, these changes capture the changing nature of the role of scientific journal editor throughout the 20th century, whereas Karl Pearson used *Biometrika* as the main outlet for his own publications, the following editors tended to blend into an increasing number of contributing authors (see Fig. 2).

A first group of authors—the left-side portion of the network, colour-coded in shades of red—consists of authors whose dominant topics are those of cluster A. Two sub-communities are apparent, one on each side of Karl Pearson. On the left can be found authors [community (a) in Fig. 2] whose interests clearly correspond to *Biometrika*'s initial focus on data and applied statistical problems related to the life, health, and anthropological sciences (see Supplementary Table S2 for each author's topic profile). In contrast, to the right of Pearson, the second sub-community (b) includes later authors whose research themes appear closer to Pearsonian mathematical statistics, and which seems to foreshadow *Biometrika*'s turn away from the life sciences after Karl Pearson's death. Note how the node corresponding to Karl Pearson constitutes the centre of gravity binding together both sub-communities. Also note the position of Egon Pearson in the vicinity of authors more related to cluster B topics (colour-coded in shades of green): this is fully consistent with what is known of his editorial role in aligning *Biometrika* with broader trends in the history of statistics. In Cox's own words, 'on K. Pearson's death, E. S. Pearson published a long appreciation of his father and then totally changed the emphasis of the journal, making it a prime place for publication of contemporary research, which it has remained' (Cox 2016: 754).

A second group—the central section of the network, colour-coded in shades of green—is composed of authors whose dominant topics belong to cluster B. Although this portion of the network is more compact than the one associated with cluster A, four sub-communities stand out. A first sub-community (c), located in the upper central part of the network in darker green, comprises authors whose dominant topics include topics such as 'B-Distribution-approximation'. A second sub-community (d) lies at the core of the network, around the editor Cox, and is composed of authors whose dominant topics are lighter green ones, especially the topic 'B-Value-estimate'. These first two sub-communities constitute the bulk of the part of the network that is associated with cluster B topics. They also form a bridge between the authors of cluster A and those of clusters C and D, confirming our interpretation of the topics above. The other two sub-communities appear less central: at the bottom centre (e), we find an 'island', a close-knit cluster of authors whose dominant topic is 'B-Design-block' (indicating an interesting independence of authors researching this topic from their colleagues), and on the upper right of the cluster-B portion, a sub-community (f) whose dominant topic is 'B-Sample-population'.

A third group—the right-side section of the network, colour-coded in shades of orange and brown—includes authors whose dominant topics are part of clusters C and D. This group includes at its centre three sub-communities which, despite being associated with different dominant topics, appear to seamlessly merge with each other. The first one (g) is found on the lower left, in shades of bright orange and yellow-orange, and corresponds to authors whose dominant topics are 'C-Matrix-covariance', 'C-Model-regression', 'C-Process-time-series', or 'C-Theorem-function'. Secondly, above this first sub-community and in dark orange and dark brown (h), one finds a cluster of authors whose dominant topics are 'C-Likelihood-parameter' and 'D-Estimate-variance'. It might be argued that the proximity between authors associated with those two topics is not surprising since they are both mostly about (respectively, maximum likelihood and nonparametric) estimation methods (see Section 3). Thirdly, this second sub-community merges, in turn, into a smaller and less compact third sub-community (i), situated at centre-right in medium brown, between the darker brown of 'D-estimate-variance' and the grey of 'D-Time-censor'; this group contains authors with 'D-Model-effect' as the dominant topic.

Finally, one finds three other more peripheral but still close-knit sub-communities. The first one (j), beneath the core of authors associated with topics of clusters C and D, in gold, is strongly characterized by the topic 'C-Prior-posterior-Bayesian'. The second one (k), on the far centre-right in orange, is composed by authors with 'C-Method-function' as the dominant topic. Peter Hall connects this second sub-community with the sub-community related to the topic 'D-Estimate-variance', which is, once again, not surprising, given the statistician profile of this author (Robinson and Welsh 2018: 221–3). Finally, the third sub-community (l) can be found in the upper rightmost section of the diagram in pale grey and is closely associated with the topic 'D-Time-censor'.

The author network thereby reveals specific sub-communities of authors who have tended to address similar research topics over the course of their publications in the journal. In so doing, it also highlights aspects of the social dynamics that underlie the development of knowledge in this specific domain of science. Our analysis shows the contribution of prolific authors (such as Karl and Egon Pearson, D.R. Cox, Peter Hall, etc.), how they fit into communities sharing common interests, and how these communities position themselves in relation to each other according to those interests. Put differently, the structure of this network offers

us a way to understand how individual authors are responsible for generating the structure of the network of topics as described in Section 3.

6. Conclusion

As we incidentally mentioned above, there is a fair amount of secondary literature covering *Biometrika*'s trajectory (Pearson 1936, 1938; Cox 2001, 2016; Aldrich 2013). According to this literature, *Biometrika* was initially founded as an outlet for the biometrical school and, under Karl Pearson's tenure as principal editor, aimed to discuss a broad range of topics related to the application of statistical methods to the life, health, and anthropological sciences. However, Weldon's untimely death in 1906 and Pearson's embroilment in various mathematical controversies caused a first shift in the journal's orientation (away from biometrical topics and from life-science and eugenics questions in general, to more 'purely' mathematical and statistical ones), which was further amplified by E. S. Pearson's decision, after 1936, to redirect *Biometrika*'s focus, building a forum for discussing more focused subjects in pure and applied statistics (such as, for instance, the Neyman–Pearson framework for hypothesis testing) which often departed from his father's view of the discipline (Cox 2001, 2016; Davison 2001). By the end of its first 100 years, the journal's profile had thus profoundly changed as a result of welcoming discussions on topics further and further removed from the core interests of its founders, such as, in recent years, the theory of maximum likelihood estimation and Bayesian methods of inference.

The first notable contribution of the present topic-modelling analysis of *Biometrika* is to confirm this narrative. This is not a facile or simplistic confirmation, however. It is noteworthy that the story as told in the hagiographic biographies of Weldon (by Karl Pearson) or Karl Pearson (by his son Egon) printed in *Biometrika*, along with much of the secondary literature on *Biometrika* to date, has only focused on 'major' figures: Weldon, Karl and Egon Pearson, R. A. Fisher, Jerzy Neyman, and perhaps a few 'second-tier' characters like Raymond Pearl or Arthur Darbishire [for a critique of this approach; see Kim (1994)]. Our analysis thus shows that, even when we take into account the full gamut of authors who published in *Biometrika* over its first 40 years, our current historical story is robust. This is a contingent claim that could quite clearly have been otherwise.

While establishing a quantitative basis for such claims about the general direction of the journal, the results of our study also provide a comprehensive and quantitative view of the broad research themes that have retained—and still retain to date—the attention of generations of statisticians (and, in its early years, biologists or eugenicists) who have published in the journal. The diachronic topic model strikingly reveals the major topic changes that the journal underwent, not just in its early days but throughout the 20th century and today. Specific topics have been identified, with illustrative publications and authors, and the mapping of their prevalence at specific times over the course of the past 120 years chronicles the history of *Biometrika* and the development of basic and applied statistics. Furthermore, the author network which was inferred from the topic-model provides insights about the underlying sub-communities of researchers, their composition, their structure as well as their relationships.

Zooming in on specific results can reveal further findings of interest or help us generate novel hypotheses about the history of statistics, the epistemic development of some of its major theories or the interactions between leading figures of the field. For instance, the author network that was built hints at the existence of two distinct sub-communities within the biometrical school of the early decades of *Biometrika*: one associated with more biologically oriented topics, and another with more mathematically oriented topics. Another conjecture stemming from the results concerns the theoretical turn taken by *Biometrika* during E. S. Pearson's tenure and its description as a transition from the research issues captured by topics, such as 'A-Value-sample-mean' to those represented by topics, such as 'B-Distribution-approximation' and 'B-Value-estimate'. Yet another hypothesis could be made, this time about the last 20 years of *Biometrika*, when analysing in more detail the two topics whose averaged probabilities have grown the most: 'D-Model-effect' and 'C-Method-function'. It could indeed be argued that their prevalence might have reflected an 'empirical counter-turn' in *Biometrika*'s orientation, fuelled by the challenges that the growing availability of large bodies of data has posed to statistical analysis. In turn, this can lead to speculations about possible future trends in the evolution of statistics. All in all, we urge that our paper attests to both the richness of *Biometrika*'s history within the broader history of statistics and the possibilities revealed by the use of topic-modelling—itself a particular type of statistical approach—for analysing it.

Acknowledgements

The authors are grateful to JSTOR and Oxford University Press for providing access to *Biometrika* articles for text-mining purposes. The authors thank the audiences of a 2021 TEC seminar at UQAM, and of the 2021 SPS congress for comments on an earlier version of the manuscript.

Author contributions

Nicola Bertoldi (Conceptualization, Data curation, Formal analysis, Investigation, Software, Validation, Writing—original draft, Writing—review and editing), Francis Lareau (Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing—review and editing), Charles Pence (Conceptualization, Funding acquisition, Investigation, Resources, Validation, Writing—original draft, Writing—review and editing), Christophe Malaterre (Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing—original draft, Writing—review and editing)

Supplementary data

Supplementary materials include Table S1 (List of topics with their top-20 words and top-20 articles), Table S2 (List of authors and their topic profiles), Figure S1 (Author correlation graph with Louvain-based community detection), data-for-graph file, code for topic analyses with input datasets (document-term matrix, document metadata file). Available on https://doi.org/10.5281/zenodo.8368810.

Funding

N.B was supported by fellowships from the Canada Research Chair in Philosophy of the Life Sciences at UQAM and the CIRST. F.L. acknowledges funding from the Fonds de recherche du Québec—Société et culture (FRQSC-276470) and the Canada Research Chair in Philosophy of the Life Sciences at UQAM. C.H.P acknowledges funding from the Fonds de la Recherche Scientifique—FNRS (Grant no. F.4526.19). C.H.P. and C.M. acknowledge funding from the XIe Commission mixte permanente Québec–Wallonie-Bruxelles (Grant no. 11.805). C.M. acknowledges funding from Canada Foundation for Innovation (Grant no. 34555) and Canada Research Chairs (CRC-950-230795).

Notes

1. Constructing a corpus for examining the history of statistics is no easy feat, in part because there are many fewer journals that cover its entire history when compared with other scientific disciplines. Our choice of *Biometrika* was thus in part a pragmatic one, as it offers us the extent needed to tell a synthetic story about the history of statistics without having to perform a (more challenging) multi-journal analysis. As noted here, however, *Biometrika* is also considered by practising statisticians to have treated many of the field's most important issues, as was made particularly clear by the contributions to the journal's centenary celebration. We thus claim that *Biometrika* constitutes one of the best available corpora for the type of analysis that we perform in the present article. We thank an anonymous reviewer for encouraging us to make our reasoning explicit.
2. Though the choice of number of topics is always debatable, we used both an agnostic measure—coherence—and an expert viewpoint as heuristics to assess the relative merit of different models, bearing in mind the objectives of the present research (e.g. DiMaggio, Nag and Blei 2013). Though models with high values of k can reveal more details about the thematic landscape of a corpus, they also tend to exhibit the so-called junk or jargon topics that gather common terms left aside by the other topics, resulting in interpretation issues. Furthermore, these models may lead to the appearance of redundant topics: this is notably the case when dominant themes in a corpus are split into several topics by the model. For these reasons, when investigating the overall thematic scope of a discipline or a set of journals, we usually prefer to opt for models with low k values that still score high in terms of coherence. On the other hand, models with higher k values may prove useful for more narrow and specific research questions (e.g. about the role of a given theory in a specific domain of science). For an example of a corpus alternatively analysed with low and high k values, see Malaterre, Chartier and Pulizzotto (2015, 2022) and Malaterre et al. (2021). In the present case, $k = 23$ led to a model with a relatively high coherence value while exhibiting well-formed and interpretable topics. This model was notably compared and found superior to the $k = 75$ model (which, despite a slightly higher coherence measure, turned out to exhibit numerous jargon as well as redundant topics).
3. Supplementary Table S1 includes examples of articles in which specific topics have a high probability of being expressed. These examples are referenced using the topic name and a sequential number.
4. A similar graph was constructed using Louvain community detection instead of dominant topics to identify author clusters (see Supplementary Fig. S1). This resulted in twenty clusters extremely well aligned with the dominant-topic approach. The latter was kept for the sake of interpretability. The striking similarity between those two sets of networks constitutes an argument for the robustness of the findings.

References

Aggarwal, C. C. (2015) *Data Mining*. Cham, Springer.

Aldrich, J. (2013) 'Karl Pearson's Biometrika: 1901-36', *Biometrika*, **100**: 3–15.

Atkinson, A. C. (2001) 'One Hundred Years of the Design of Experiments On and Off the Pages of Biometrika', *Biometrika*, **88**: 53–97.

Bastian, M., Heymann, S., and Jacomy, M. (2009) 'Gephi: an open source software for exploring and manipulating networks', in International AAAI Conference on Weblogs and Social Media. San Jose, CA, AAAI.

Blei, D. M. (2012) 'Probabilistic Topic Models', *Communications of the ACM*, **55**: 77–84.

Blei, D. M., and Lafferty, J. D. (2006) 'Dynamic topic models', in *Proceedings of the 23rd International Conference on*

Machine Learning (ICML'06), pp. 113-120. Pittsburgh, ACM.

Blei, D. M., and Lafferty, J. D. (2007) 'A Correlated Topic Model of Science', *Annals of Applied Statistics*, **1**: 17–35.

Blei, D. M., and Lafferty, J. D. (2009) 'Topic Models'. In: Srivastava, A. N. and Sahami, M. (eds) *Text Mining: Classification, Clustering, and Applications*, pp. 71–94. London, Chapman and Hall/CRC.

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

Blondel, V. D. *et al.* (2008) 'Fast Unfolding of Communities in Large Networks', *Journal of Statistical Mechanics: Theory and Experiment*, **2008**: P10008.

Bulmer, M. (2003) *Francis Galton: Pioneer of Heredity and Biometry*. Baltimore, MD, Johns Hopkins University Press.

Burman, J. T. (2018) 'Digital Methods Can Help You… If You're Careful, Critical, and Not Historiographically Naïve', *History of Psychology*, **21**: 297–301.

Bystranowski, P., Dranseika, V., and Żuradzki, T. (2022) 'Half a Century of Bioethics and Philosophy of Medicine: A Topic-modeling Study', *Bioethics*, **36**(9): 902–25.

Cock, A. G., and Forsdyke, D. R. (2022). *Treasure Your Exceptions: The Science and Life of William Bateson*, 2nd edn. New York, Springer.

Cohen Priva, U., and Austerweil, J. L. (2015) 'Analyzing the History of Cognition using Topic Models', *Cognition*, **135**: 4–9.

Cox, D. R. (2001) 'Biometrika: The First 100 Years', *Biometrika*, **88**: 3–11.

Cox, D. R. (2016) 'Some Pioneers of Modern Statistical Theory: A Personal Reflection', *Biometrika*, **103**: 747–59.

Davison, A. C. (2001) 'Biometrika Centenary: Theory and General Methodology', *Biometrika*, **88**: 13–52.

Desrosières, A. (2002) *The Politics of Large Numbers: A History of Statistical Reasoning*. Cambridge, MA, Harvard University Press.

Didier, E. (2020) *America by the Numbers: Quantification, Democracy, and the Birth of National Statistics*. Cambridge, MA, The MIT Press.

DiMaggio, P., Nag, M., and Blei, D. (2013) 'Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding', *Poetics*, **41**: 570–606.

Donnelly, K. (2015) *Adolphe Quetelet, Social Physics and the Average Men of Science, 1796–1874*. Pittsburgh, PA, University of Pittsburgh Press.

Droesbeke, J.-J. (2021) *Adolphe Quetelet: Passeur d'idées*. Bruxelles, Académie Éditions.

'Editorial: The Scope of Biometrika' (1901) *Biometrika*, **1**: 1–2.

Elderton, W. P. (1951) 'Biometrika 1901-1951', *Biometrika*, **38**: 267–68.

Firth, J. R. (1957) 'A Synopsis of Linguistic Theory 1930–1955'. In Firth, J. R. (ed.) *Studies in Linguistic Analysis*, pp. 1–32. Oxford, Blackwell.

Froggatt, P., and Nevin, N. C. (1971) 'The 'Law of Ancestral Heredity' and the Mendelian-Ancestrian Controversy in England, 1889-1906', *Journal of Medical Genetics*, **8**: 1–36.

Galton, F. (1901) 'Biometry', *Biometrika*, **1**: 7–10.

Ghosh, A. (2020) *Making It Count: Statistics and Statecraft in the Early People's Republic of China*. Princeton, NJ, Princeton University Press.

Gigerenzer, G. *et al.* (1989) *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge, Cambridge University Press.

Griffiths, T. L., and Steyvers, M. (2004) 'Finding Scientific Topics', *Proceedings of the National Academy of Sciences*, **101**: 5228–35.

Hacking, I. (1990) *The Taming of Chance*. Cambridge, Cambridge University Press.

Hall, P. (2001) 'Biometrika Centenary: Nonparametrics', *Biometrika*, **88**: 143–65.

Igo, S. E. (2007) *The Averaged American: Surveys, Citizens, and the Making of a Mass Public*. Cambridge, MA, Harvard University Press.

Kim, K.-M. (1994) *Explaining Scientific Consensus: The Case of Mendelian Genetics*. New York, The Guilford Press.

Krüger, L., Daston, L., and Heidelberger, M. (eds) (1987) *The Probabilistic Revolution, Volume 1: Ideas in History*. Cambridge, MA, Bradford Books.

Krüger, L., Gigerenzer, G., and Morgan, M. S. (eds) (1987) *The Probabilistic Revolution, Volume 2: Ideas in the Sciences*. Cambridge, MA, Bradford Books.

Lucas, C. *et al.* (2015) 'Computer-assisted Text Analysis for Comparative Politics', *Political Analysis*, **23**: 254–77.

Magnello, M. E. (2009) 'Karl Pearson and the Establishment of Mathematical Statistics', *International Statistical Review*, **77**: 3–29.

Malaterre, C., Chartier, J.-F., and Pulizzotto, D. (2015) 'What is this Thing called Philosophy of Science? A Computational Topic-modeling Perspective 1934-2015', *HOPOS*, **9**: 215–49.

Malaterre, C., Chartier, J.-F., and Pulizzotto, D. (2022) 'Topic Modeling in HPS: Investigating Engaged Philosophy of Science throughout the 20th Century'. In: Ramsey, G. and De Block, A. (eds) *The Dynamics of Science: Computational Frontiers in History and Philosophy of Science*, pp. 164–85. Pittsburgh, PA, University of Pittsburgh Press.

Malaterre, C., and Lareau, F. (2022) 'The Early Days of Contemporary Philosophy of Science: Novel Insights from Machine Translation and Topic-modeling of Non-parallel Multilingual Corpora', *Synthese*, **200**: 242.

Malaterre, C. *et al.* (2021) 'Eight Journals over Eight Decades: A Computational Topic-modeling Approach to Contemporary Philosophy of Science', *Synthese*, **199**: 2883–2923.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993) 'Building a Large Annotated Corpus of English: The Penn Treebank', *Computational Linguistics*, **19**: 313–30.

Oakes, D. (2001) 'Biometrika Centenary: Survival Analysis', *Biometrika*, **88**: 99–142.

Patriarca, S. (1996) *Numbers and Nationhood: Writing Statistics in Nineteenth-century Italy*. Cambridge, Cambridge University Press.

Pearson, E. S. (1936) 'Karl Pearson: An Appreciation of Some Aspects of His Life and Work. Part I: 1857–1906', *Biometrika*, **28**: 193–257.

Pearson, E. S. (1938) 'An Appreciation of Some Aspects of His Life and Work. Part II: 1906–1936', *Biometrika*, **29**: 161–248.

Peirson, B. R. E. *et al.* (2017) 'Quantitative Perspectives on Fifty Years of the Journal of the History of Biology', *Journal of the History of Biology*, **50**: 695–751.

Pence, C. H. (2022) *The Rise of Chance in Evolutionary Theory: A Pompous Parade of Arithmetic*. London, Academic Press.

Porter, T. M. (1986) *The Rise of Statistical Thinking, 1820–1900*. Princeton, NJ, Princeton University Press.

Porter, T. M. (2004) *Karl Pearson: The Scientific Life in a Statistical Age*. Princeton, NJ, Princeton University Press.

Provine, W. B. (1971) *The Origins of Theoretical Population Genetics*. Princeton, NJ, Princeton University Press.

Robinson, J., and Welsh, A. H. (2018) 'Peter Gavin Hall. 20 November 1951—9 January 2016', *Biographical Memoirs of Fellows of the Royal Society*, **64**: 207–29.

Röder, M., Both, A., and Hinneburg, A. (2015) 'Exploring the Space of Topic Coherence Measures', in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* Shanghai, China, ACM Press. pp. 399–408.

Rothe, A., Rich, A. S., and Li, Z. (2018) 'Topics and Trends in Cognitive Science', in *Proceedings of the 40th Annual Conference of the Cognitive Science Society*. Austin, TX, Cognitive Science Society.

Salsburg, D. (2001) *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. New York, W. H. Freeman and Company.

Schmid, H. (1994) 'Probabilistic Part-of-speech Tagging Using Decision Trees', in *Proceedings of International Conference on New Methods in Language Processing*, Manchester, Association for Computational Linguistics, pp. 44–49.

Smith, T. M. F. (2001) 'Biometrika Centenary: Sample Surveys', *Biometrika*, **88**: 167–243.

Srivastava, A. N., and Sahami, M. (2009). *Text Mining: Classification, Clustering, and Applications*. Boca Raton, FL, CRC Press.

Stigler, S. M. (1986) *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA, Harvard University Press.

Stigler, S. M. (1999) *Statistics on the Table: The History of Statistical Concepts and Methods*. Cambridge, MA, Harvard University Press.

Tong, H. (2001) 'A Personal Journey through Time Series in Biometrika', *Biometrika*, **88**: 195–218.

de Vries, E., Schoonvelde, M., and Schumacher, G. (2018) 'No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications', *Political Analysis*, **26**: 417–30.