Quantifying Bias and Uncertainty in Historical Data Collections with Probabilistic Programming

Leo Lahti[a], Eetu Mäkelä[b] and Mikko Tolonen[b]

[a] *University of Turku, Turku, Finland*
[b] *University of Helsinki, Helsinki, Finland*

Abstract

The enhanced access to ever-expanding digital data collections and open computational methods have led to the emergence of new research lines within the humanities and social sciences, bringing in new quantitative evidence and insights. Any data interpretation depends critically on understanding of the scope and limitations in data collection, as well as on reliable downstream analysis. Quantitative analysis can complement qualitative research by providing access to overlooked information that is accessible only through systematic discovery and analysis of latent patterns underlying the available data collections. Probabilistic programming is an expanding paradigm in machine learning that provides new statistical tools for intuitive interpretation of complex data sets. This new paradigm stems from Bayesian analysis and emphasizes explicit modeling of the data generating processes and associated uncertainties. Despite its remarkable application potential, probabilistic programming has so far received little attention in computational humanities. We use a brief case study in computational history to demonstrate how probabilistic programming can be incorporated in reproducible data science workflows in order to detect and quantify bias in a widely studied historical text collection, the Eighteenth Century Collections Online.

Keywords
bias, computational history, probabilistic programming, uncertainty, bibliographic data science

1. Introduction

Research questions in computational humanities often deal with very similar quantitative challenges than the natural or social sciences, which have a long tradition in data-driven research. Techniques from other fields can be often readily borrowed in new application fields with small adaptations. This is enabling the translation of well-established methodological paradigms from other disciplines, such as ecology, econometrics, or physics into the emerging field of computational humanities. Research in computational humanities benefits from a rich mixture of data science techniques that range from database management to data harmonization, computational modeling and visualization. Reproducible data science workflows that unify the complementary steps of data analysis have become a standard tool to facilitate collaborative research [15, 11].

Understanding biases and uncertainty is fundamental to research. Even the most perfectly harmonized and clean research data sets contain subtle selection and other biases. Such biases can, however, be potentially detected and treated through explicit formal analysis. In

addition to questions of raw data quality, one needs to critically assess and understand data representativeness and consider the limitations of the chosen methodology. Whereas much of the recent technical efforts that have been aimed at ensuring data quality have focused on the development of automated procedures for data processing and harmonization, data curation is only the stepping stone towards critical statistical analysis and modeling [14, 6].

Here, we consider probabilistic programming as a flexible tool that can be used to detect and characterize bias and uncertainties in research data collections that are relevant to computational humanities. Probabilistic programming is a statistical paradigm that has increased in popularity in various fields of applied computational science, leading to the emergence of a variety of alternative probabilistic programming languages such as Stan [4]. Probabilistic programming can be used to build explicit models and compare evidence between alternative hypotheses on the data generating processes. We show how this can help to bridge the gap between qualitative and quantitative interpretations, and provides promising tools for hypothesis-driven, data-intensive research in computational humanities.

We provide a brief overview of probabilistic analysis and demonstrate its application on detecting specific biases in the composition of Eighteenth Century Collections Online (ECCO). This full text collection of British literature has become one of the most central digital resources for studying printed eighteenth-century texts in English language. However, the contents of ECCO are fully derived from the information contained in the English Short Title Catalogue (ESTC) [2, 1], whose original purpose was to become a 'machine-readable union catalogue of books, pamphlets and other ephemeral material printed in English-speaking countries from 1701 to 1800', although ESTC has been subsequently extended to include additional material from earlier time periods. ECCO specifically aims to represent the full texts contained in ESTC, and all titles in ECCO are hence included in ESTC by design. However, a key to our analysis is the observation that a notable fraction of ESTC titles remain missing from ECCO. We anticipate that the missing coverage is not random but instead a variety of factors may influence the coverage of specific titles. A quantitative analysis can reveal such biases. As such, ESTC provides the natural reference point for studying biases in ECCO and the comparison of these two sources provides a well-defined case study on detecting and quantifying bias in historical document collections. Our work is thus highlighting the importance of modeling and interpreting bias as part of a standard data science workflow, and demonstrates how such analysis can be supported by the emerging paradigm of probabilistic programming.

2. Materials and Methods

2.1. Data

Our case study is based on the systematic and ongoing data collection and harmonization that has already been carried out and reported previously [14]. Here, we briefly describe the background work that forms the foundation for statistical modeling that is the focus in the present work.

Eighteenth Century Collections Online (ECCO)

The Eighteenth Century Collections Online (ECCO) is a digitized full text collection of printed eighteenth-century texts in English language[1]. The ECCO collection of full texts has been

---

[1]https://www.gale.com/primary-sources/eighteenth-century-collections-online

constructed based on the ESTC library catalogue (see below) and, subsequently, analysis of ECCO is often based on the implicit assumption that it has a comprehensive and unbiased coverage of the English titles catalogued in the ESTC. In this report, we assess the validity of this assumption by comparing the coverage of the ESTC titles in ECCO and formally quantifying specific key factors that potentially influence the coverage. We have obtained access to ECCO confidentially for research purposes upon a separate agreement. We do not analyze the full text contents of the ECCO, and hence the identification of the matching records between the two collections is sufficient for our current research purpose and could be readily done based on shared document identifiers. By design, all titles in ECCO are being covered in ESTC, whereas some titles catalogued in ESTC remain missing from ECCO.

English Short-Title Catalogue (ESTC)

The English Short-Title Catalogue (ESTC) focuses on British early modern publishing and includes bibliographic information on mainly English documents from the eighteenth century. The cataloguing work has been coordinated by the British Library [2, 1], and we have obtained access to this data for research purposes upon a separate agreement. Here, we focus on the bibliographic information that is available for over 227 000 titles in the period 1701-1800. We have carried out initial harmonization of various bibliographic fields in this data collection as described previously [14, 12, 16]. This has provided us with manually curated information on author identities and gender, publication types (books versus pamphlets), and physical aspects of the documents, such as the standard sizes, as well as information on the publishing country. We deduplicated author names based on a combination of automation and careful manual curation as described earlier [10], and derived additional gender, liftime, and other information by matching the uniquely detected author names in ESTC with those reported in Virtual International Authority File (VIAF), a centralized collection of widely-used library authority files. The analyses in this paper rely on the earlier harmonization efforts that provided curated high-quality metadata on authors and documents, thus forming the basis of our current analysis.

2.2. Bibliographic data science

We have previously implemented the background data harmonization as a collection of reproducible bibliographic data science workflows [14] that help to monitor and improve the overall quality of the raw bibliographic records, thus facilitating a more reliable and accurate detection of large-scale patterns. Our initial quality-controlled version of the ESTC is now allowing us to take the next steps towards systematic statistical analyses of its information contents.

Whereas the ECCO and ESTC data collections have been obtained from external memory organizations based on a separate agreement and they are thus not openly available, our own data harmonization and analysis workflows are open source and available online[2]. The overall data processing workflow consists of a collection of custom Python scripts, R packages, and other components [13].

2.3. Binomial model for coverage

Our task is to analyse how often titles listed in ESTC library catalogue are covered by the collection of full texts included in ECCO. This well-defined task for formal analysis is motivated by the need to understand the correspondence between these two sources. In particular, we analyse how specific background factors influence ESTC coverage in ECCO and contribute potential bias to the subsequent interpretation of this full text collection.

The coverage of the ESTC titles in ECCO can be formally modeled as a repeated coin-tossing experiment, where each title in the ESTC have a certain probability of being covered in ECCO. This probability may be then influenced by a number of factors, such as author gender or publication decade. Joint analysis of a large number of documents can reveal systematic differences in coverage associated with such factors. A common approach to modeling such probabilities is based on logistic regression coefficients in a standard binomial model:

$$
\begin{aligned}
k &\sim Bin(n, p) \\
p &\sim logit^{-1}(\beta_0 + \Sigma_{i=1}^{J} \beta_i x_i),
\end{aligned}
\tag{1}
$$

Here, n and k are the numbers of titles included in the ESTC and ECCO, respectively, and p is the probability that an ESTC title with the given properties will be included in ECCO. Here, we use $J = 3$ covariates that include author gender (x_1), publication type (x_2), and publishing time (x_3). These covariates are encoded as binary indicator variables that indicate female authors ($x_1 = 1$), pamphlets ($x_2 = 1$), and publication time ($x_3 = 1$ for publications after the year 1750). We have excluded authors with an unknown gender (5.8% of unique authors in our harmonized version of ESTC). For publication time, we use a binary encoding into early and late century (i.e. later than 1750) since our preliminary experiments indicated that the main temporal variation is associated with this split, rather than publication year or decade. The covariates influence the log-odds of the binomial probability through the regression coefficients β; where β_0 represents the standard intercept term. This will allow us to assess the effect of each covariate on the estimated probability p through the logit link. Here, the k, n and the covariate information encoded in x are observed variables, whereas the regression parameters β are to be inferred from the data and can be used to retrieve posterior estimates of the latent binomial probability p. In the probabilistic formulation, we can additionally set priors for the model parameters. We have used a Gaussian prior $\beta_0 \sim N(0, 5)$ for the intercept term and $\beta_i \sim N(0, 2.5)$ for the regression terms $i \in \{1, 2\}$. The prior distributions that we have chosen to use here are relatively uninformative; they cover the expected range of plausible parameter values, with a negligible effect on the posterior distributions that are essentially determined by the data in our experiments.

2.4. Probabilistic programming

The standard binomial regression model in Equation 1 could be fitted based on classical statistical techniques as well as through probabilistic programming. We have chosen this relatively standard model for our case study in order to draw analogies and facilitate further comparison between the two approaches. In standard statistical analysis, one would find the maximum-likelihood solution for the equation 1 by identifying a single optimal point estimate for β. Probabilistic programming allows us to not only implement this standard model but additionally allows us to incorporate prior information and quantify uncertainty in the inferred

parameters in the form of a posterior distribution, and compare the evidence that the data is providing between alternative models.

Probabilistic programming is an emerging paradigm that bridges the gap between Bayesian probabilistic analysis and its practical application. Our case study has been implemented in Stan, which is one of the several available options and has an active user community and a dedicated collection of tutorials [4]. We promote the use of probabilistic programming as a tool for analysing biases and uncertainty in digitized materials that are common in humanities research.

Probabilistic analysis differs from classical statistical analysis in certain important ways. First, it allows a flexible construction of complex models that do not need to be analytically tractable, in contrast to common classical alternatives. Second, probabilistic models emphasize the need to assess the uncertainty in the inferred model and model parameters. Third, the framework allows the incorporation of prior information whenever this is available to support the modeling. Whereas probabilistic programs provide access to an increasingly large family of statistical models, a reliable inference is often slower compared to classical statistical analysis. For more details, we refer to the original publications on Stan [4].

We implemented the probabilistic version of the binomial regression model of Equation 1 with the *rstanarm* R package [9] that provides shortcuts to the probabilistic programming variants of common statistical models. The source code of the analyses are available in the online repository linked above.

3. Results

3.1. Observed biases

The first step in understanding potential biases in the ECCO full text collection with respect to the ESTC library catalogue is to quantify and compare the coverage frequencies between different groups of published titles.

Altogether, 62.6% of the unique titles catalogued in our harmonized version of ESTC are covered in ECCO. We excluded titles where author gender is unknown (5.8% of all harmonized ESTC titles). Of the remaining ESTC titles 4.8% have been authored by a woman, and we also see that the ECCO coverage is in general somewhat lower for women (Table 1). However, the lower coverage of female authors might be partially explained by other, confounding differences. For instance, a closer view reveals that women also author pamphlets considerably more often than men do (54.2% versus 43.6% of all titles for women and men, respectively). Since pamphlets have a lower coverage in general (Table 2), the lower coverage of female authors in ECCO might be readily explained by their different publishing patterns, rather than gender. Also other factors, such as the publishing time, may play a role as we demonstrate in the next section. Such general comparisons of publication frequencies indicate that certain groups of publications are over- or underrepresented in ECCO with respect to ESTC.

Such preliminary observations of publication frequencies or other broad patterns have obvious limitations. Mere quantification of the coverage frequencies can tell us little about the uncertainties about these observations, and distinguishing between the overlapping effects such as author gender, genre, and time is not always straightforward. Whereas the overall frequencies shown in Figure 1 can be readily estimated from data, a a full statistical analysis is necessary when we are willing to draw conclusions on the overall influence of gender or other factors on ECCO coverage and quantify the uncertainty associated with such factors.

Table 1
The frequency of male and female authors in ESTC and the corresponding ECCO coverage.

Gender	Frequency (%)	Coverage (%)
Male	95.2	62.8
Female	4.8	58.9

Table 2
The frequency of books and pamphlets in ESTC and the corresponding ECCO coverage.

Genre	Frequency	Coverage
Books	55.8%	70.9%
Pamphlets	44.2%	52.1%

3.2. Separating the effects of genre and gender

The log-odds for the model parameters - the intercept and the regression coefficients associated with genre (pamphlet), gender (female), and time period (latter half of the eighteenth century) were estimated by probabilistic programming (Figure 2).

Negative log-odds indicate a negative impact on ECCO coverage. For instance, according to our results a female author will reduce the probability that a title is covered in ECCO. When we include the gender effect as the only explanatory factor in the binomial model, female gender is associated with the log-odds of -0.16. This corresponds to the difference in the ECCO coverage probabilities between female and male authors that we can also see in Table 1 (58.9% vs. 62.8%, respectively). However, when we take genre (books versus pamphlets) into account in the model, being a female author reduces the log-odds of ECCO coverage by only -0.08. Thus, the gender effect is less striking when we also model differences in the typical publishing patterns between male and female authors. Moreover, the publishing time period shows additional major influence; in general, the ECCO drops over time from 67% in 1700's to 58% in 1790's. This drop is observed specifically at the latter half of the century. According to our binomial model, a publication date after 1750 has a larger negative impact on ECCO coverage than author gender (Fig. 2). The evidence for a negative gender effect remains also after controlling this time factor, however ($\beta_1 = -0.06$) with a high certainty indicated by the 99% posterior intervals being below zero. Thus, our quantitative analysis reveals gender-specific differences and supports the earlier arguments that the publication activity of female authors may have been underestimated due to their biased coverage in common research data collections [3, 8]. The further step in such analysis would be to perform systematic model criticism and comparisons in order to balance the complexity of the models with the available information in the data [7]. In our binomial model, the full model outperformed all submodels with an expected log predictive density of -10 or more, as estimated with the leave one out cross-validation (loo-cv) method as implemented in the R package *loo* [19].

The analysis is also allowing us to compare the relative effects of the different document and author properties on ECCO coverage. The effect of genre is notably larger than the effect of gender. Pamphlets have the estimated log-odds of -0.80, which roughly translates into the 45% odds of ECCO coverage for pamphlets as compared to books (Figure 2). The odds of ECCO coverage for titles published during the latter half of the century is 73% compared to the first half of the century. The odds for female authors in the full model including both genre and publication time is 94%.
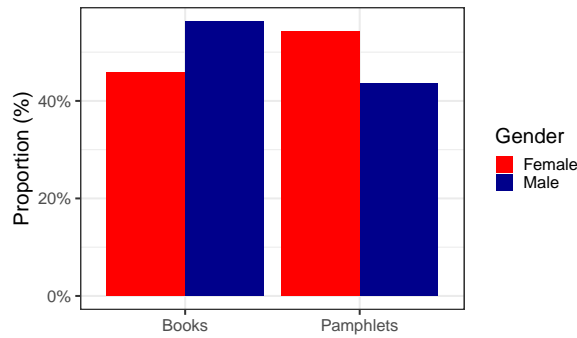
Figure 1: The proportion of books and pamphlets compared to all publishing activity by men and women.
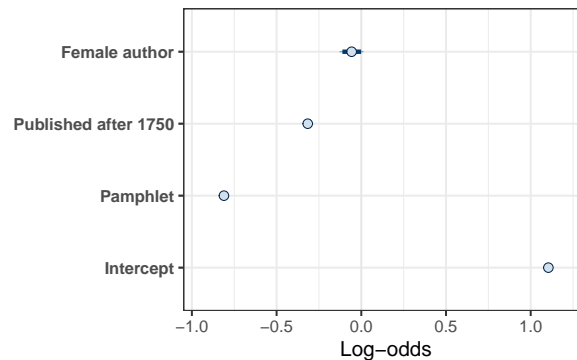

Figure 2: The effect of gender, publication type, and time period on ECCO coverage. The estimated log-odds are shown with 99% posterior intervals The posterior is based on a binomial model estimated with the stan_glm function in the *rstanarm* R package. Female author, publishing during the latter half of the century, and pamphlet publications affect negatively the log-odds, and hence probability, that a title is included in ECCO.

4. Discussion

The improved availability of digital resources has led to the emergence of new research lines that complement classical humanities and social sciences. Quantitative analysis can provide access to insights that are only obtained through systematic large-scale analysis. When applied thoughtfully and critically, this has the potential to enrich and complement qualitative research. Interpreting the available data depends critically on understanding of the inherent limitations in data collection and analysis, as well as on the availability of reliable quantitative methods. Such considerations are at the core of statistical practice [18, 5].

We have previously demonstrated the benefits of large-scale data harmonization in computational history, and in particular shown how data curation can be scaled up by semi-automated means in order to analyse metadata collections covering millions of titles [14, 17]. Ensuring data quality and commensurability is only the first step towards research use.

We have now demonstrated how probabilistic programming can be incorporated into reproducible data science workflows in computational humanities. Whereas similar results could be obtained with classical statistical techniques for our specific case study, the probabilistic framework can open up access into a wider set of extensions and the use of prior information.

Limited data availability leads to uncertainty in the parameter estimates, which are naturally captured by the posterior distributions of the probabilistic models. Also our case study highlighted some remaining uncertainty in the estimated gender effect. As such, probabilistic analysis can facilitate the detection of biases and uncertainty that influence data interpretation.

We provided supporting evidence for the hypothesis that the Eighteenth Century Collections Online (ECCO) may contain previously unreported biases, and that specific aspects of literature are being underrepresented in the data collection. In particular, female authors, pamphlets, and later publication dates led to reduced odds of ECCO coverage. The negative impact of later publishing might be partially explained by differences in the frequencies in first editions and reprints but this would warrant a more careful analysis. Such biases have obvious implications for any conclusions that would be potentially drawn from the data. Whereas in this technically oriented work we have highlighted probabilistic programming as a new method in computational humanities, we are now carrying out complementary work that will yield a more comprehensive critical historical analysis and interpretation. A more thorough benchmarking of the alternative bias detection methods and a wider set of case studies could further help to understand the overall range of applications and pragmatic limitations in such modeling tasks.

Bayesian analysis and probabilistic programming have obvious limitations that are slowing down their wider adoption. In addition to slower model inference compared to classical alternatives, effective construction, use and interpretation of probabilistic models requires a robust understanding of modern statistics as well as statistical programming. The emergence of probabilistic programming languages and standardization of common modeling tasks is now lowering these barriers, however. A wider adoption and open sharing of these workflows can help to standardize the analysis of biases as well as other specific approaches that are common in computational humanities [11]. Probabilistic programming provides a natural extension to the already ubiquitous data science workflows that are being used to harmonize and interpret research data in computational humanities.

Limitations on data availability are common in the humanities. Our case study is based on a joint analysis of the ECCO and ESTC data collections that we have obtained for research purposes. We hope that this can inspire related research on such data collections, and help to demonstrate the benefits for wider research collaborations that could be obtained if the open data science workflows could be combined with fully open historical data resources.

The currently emerging ecosystems of computational programming languages continues to facilitate the translation of methodology from other data-intensive fields of research such as ecology, bioinformatics, or econometrics. Probabilistic programming places emphasis on formal analysis of uncertainty in inference and models, and on continuous model criticism. As such, it can be extended far beyond our brief case study in order to construct more comprehensive hierarchical models of the data-generating processes that characterize various aspects of spatial, temporal, seasonal, and other variation in the quantitative aspects of social sciences and humanities.

Acknowledgments

References

[1]   R. Alston. "The history of ESTC". In: *Age of Johnson* 15 (2004), pp. 269–329.

[2]   R. Alston and M. Jannetta. "Bibliography, machine-readable cataloguing and the ESTC". In: *The Library Quarterly* 50 (2 1980), pp. 273–274.

[3]   M. Bell. "Women writing and women written". In: *The Cambridge History of the Book in Britain*. Vol. IV. Cambridge University Press, 2002, pp. 431–452.

[4]   B. Carpenter et al. "Stan: A probabilistic programming language". In: *Journal of Statistical Software* 76.1 (2017).

[5]   D. Donoho. "50 years of data science". In: *Journal of Computational and Graphical Statistics* 26 (4 2017), pp. 745–766.

[6]   M. Eetu et al. "Wrangling with Non-Standard Data". In: *Proc. Digital Humanities in the Nordic Countries*. Vol. 2612. CEUR Workshop Proceedings. Riga, Latvia, Oct. 2020, pp. 81–96.

[7]   J. Gabry et al. "Visualization in Bayesian workflow". In: *Journal of the Royal Statistical Society A* 182 (2 2019), pp. 389–402.

[8]   C. Gallagher. *Nobody's story: the vanishing acts of women writers in the marketplace, 1670-1820*. Oxford University Press, 1995.

[9]   B. Goodrich et al. *rstanarm: Bayesian applied regression modeling via Stan*. R package version 2.21.1. 2020. URL: https://mc-stan.org/rstanarm.

[10] M. J. Hill et al. "Reconstructing intellectual networks: from the ESTC's bibliographic metadata to historical material". In: *Proc. Digital Humanities in the Nordic Countries*. Ed. by C. Navarretta, M. Agirrezabal, and B. Maegaard. Vol. 2364. CEUR Workshop Proceedings. Copenhagen, Mar. 2019, pp. 201–219.

[11] L. Lahti. "Open Data Science". In: *Advances in Intelligent Data Analysis XVII. Lecture Notes in Computer Science 11191*. Vol. 11191. India: Springer Nature, Oct. 2018, pp. 31–39.

[12] L. Lahti, N. Ilomäki, and M. Tolonen. "A quantitative study of history in the English short-title catalogue (ESTC) 1470-1800". In: *LIBER Quarterly* 25.2 (Dec. 2015), pp. 87–116.

[13] L. Lahti et al. "Best practices in bibliographic data science". In: *Proc. Research data and humanities (RDHUM) 2019 conference: data, methods and tools. Studia Humaniora Ouluensia 17*. Ed. by J. Jantunen et al. Oulu: University of Oulu, Aug. 2019, pp. 57–65.

[14] L. Lahti et al. "Bibliographic Data Science and the History of the Book (c. 1500–1800)". In: *Cataloging & Classification Quarterly* 57.1 (Jan. 2019), pp. 5–23.

[15] E. Mäkelä et al. "Interdisciplinary collaboration in studying newspaper materiality". In: *Proc. Twin Talks workshop. Digital Humanities in the Nordic Countries*. Ed. by S. Krauwer and D. Fišer. Vol. 2365. CEUR Workshop Proceedings. Copenhagen, Mar. 2019, pp. 55–66.

[16] M. Tolonen et al. "Data Visualization in Enlightenment Literature and Culture". In: Palgrave Macmillan, 2020. Chap. Examining the Early Modern Canon: The English Short Title Catalogue and Large-Scale Patterns of Cultural Production, In press.

[17]   M. Tolonen et al. "Scaling up bibliographic data science". In: *Proc. Digital Humanities in the Nordic Countries*. Ed. by C. Navarretta, M. Agirrezabal, and B. Maegaard. Vol. 2364. CEUR Workshop Proceedings. Copenhagen, Mar. 2019, pp. 450–456.

[18]   J. Tukey. "The Future of Data Analysis". In: *Annals of mathematical statistics* 33 (1 1962), pp. 1–67.

[19]   A. Vehtari, A. Gelman, and J. Gabry. "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". In: *Statistics and Computing* 27 (5 2017), pp. 1413–1432.