Probabilistic Analysis of Early Modern British Book Prices

Iiro Tiihonen[1,2], Mikko Tolonen[1] and Leo Lahti[2]

[1]*Faculty of Humanities, University of Helsinki, Finland*
[2]*Faculty of Technology, University of Turku, Finland*

Abstract

Books are a valuable exception to the general rule that quantitative information about early modern history is scarce, as their survival rate during the period has varied between low and high tens of percents, and descriptive information summarizing their properties has been collected to library catalogues. However, one critical element that is essential for the numeric characterisation of a print product is most often missing - its price. In this paper, we use an exceptionally large data set of price information extracted from the English Short Title Catalogue (ESTC) for the early modern period to train a probabilistic model that predicts the price of a print product based on its physical properties. Our results suggest that just the simple physical properties of the print products can explain a significant proportion of the variation in prices. We use the model to quantitatively address the debated question about development of print product prices in eighteenth century Britain. We interpret the predictions of the model as a data driven narrative, and many of the developments it brings up can be readily linked with the relevant historical literature.

Keywords

bibliographic data science, book history, book prices, statistical modeling, early modern period

1. Introduction

Bibliographic data gives insight to various historical phenomena from vernacularisation and political crises to the effect of books themselves to early modern society. As bibliographic data is transformed into machine readable form, the application of modern data science becomes possible [10, 15, 8]. Price is recognised as an important aspect of bibliographical information in various branches of historical and social scientific literature, as it can be used to approximate the accessibility of information and culture by wealth and functioning of the book trade [5, 13, 14, 12, 3, 1]. As questions about unequal access to information and culture and the way (book) trade worked are relevant for the understanding of early modern societies, price of a print product is a valuable piece of information. Unfortunately it is often not available, and even analyses that make use of relatively large bibliographical data sets become much more limited when the price is concerned [6].

Our starting point is one of the historical debates for which the question of book prices is essential. The core of this debate is about the effect of the 1774 affirmation of the Statute of Anne that transformed copyrights of books from perpetual ownership to last 14 years in Great Britain. By applying a quantitative model that describes book price formation, we could

---

✉ iiro.tiihonen@helsinki.fi (I. Tiihonen); mikko.tolonen@helsinki.fi (M. Tolonen); leo.lahti@utu.fi (L. Lahti)
🆔 0000-0003-0703-4556 (I. Tiihonen); 0000-0003-2892-8911 (M. Tolonen); 0000-0001-5537-637X (L. Lahti)

untangle the variation related basic physical attributes, and obtained results that suggest that this decrease of prices did indeed happen. The model itself is the most important result of the paper, as expanding on it opens the possibility for further historical research and extrapolation of data.

There have been attempts to establish overviews about book prices and their variation in different periods and regions of early modern Europe [5, 13, 3]. Economists have been at the forefront of formal quantitative work on early modern book prices, and have used mathematical models for purposes like derivation of general level time series [4, 16] and hypothesis testing regarding the impact of specific variables of the model like indicators of competitive business environment and logistic costs [6]. We use our model for hypothesis testing as well. However, we also stress that our model has more general value as a tool for extrapolation and further hypothesis testing. We argue that carefully designed quantitative models open up a much wider array of possibilities for a detailed systematic characterization of the factors underlying the observed price variation [9]. Moreover, noteworthy patterns in price formation can be also detected by identifying systematic patterns that cannot be explained by the quantitative model.

This article describes and implements a model that aims to quantify the influence of physical attributes on the observed price variation. These factors already give significant insight to the formation of price for the majority of print products, and validation in leave-out data indicates that our model can also predict price variation for new documents with a relatively good accuracy.

The data set we use is based on the English Short Title Catalogue (ESTC), an Union Catalogue of early modern print products of British Isles and North America comprised of the metadata information of various libraries, most notably the British National library. The tabular version used in this paper is the result of long parsing, harmonisation and enrichment process of the Helsinki Computational History Group (COMHIS). Critically, this version of the ESTC includes price information for roughly 30000 records as well. Of these, 28274 records with a parsed price were deemed reliable enough in their information for modeling.[1]

The data collection that we use is an order of magnitude larger than most data sets on early modern book prices. A recent study used a data set of 3000 purchase prices for books from the entirety of reformation era Europe, and characterised it as an unique data for the study of book markets [6]. Another study considers an inventory of a Venetian book shop with 12 000 entries "one of the most extensive and significant sources for the study of book prices in early modern Europe." [1]. Our data set provides exceptionally good coverage for eighteenth century Great Britain and additionally includes remarkably rich background data on the documents' dimensions, genres, and other properties [8]. In addition, British economic history and related background information is better known than perhaps that of any other European country, allowing us to normalise the prices with respect to price index.

---

[1]We removed categories of advertisements, pilot guides and prospectus as the prices mentioned in these records was most often for another product. We also removed periodicals and many kinds of public records, as multiple issues and parts are often collected together in the ESTC, but it was often unclear if the price was for a single issue or part only. We also removed cases where we found that the harmonised ESTC's information for relevant variables like page count or plates was wrong. When found, erroneous prices were removed or corrected if possible.
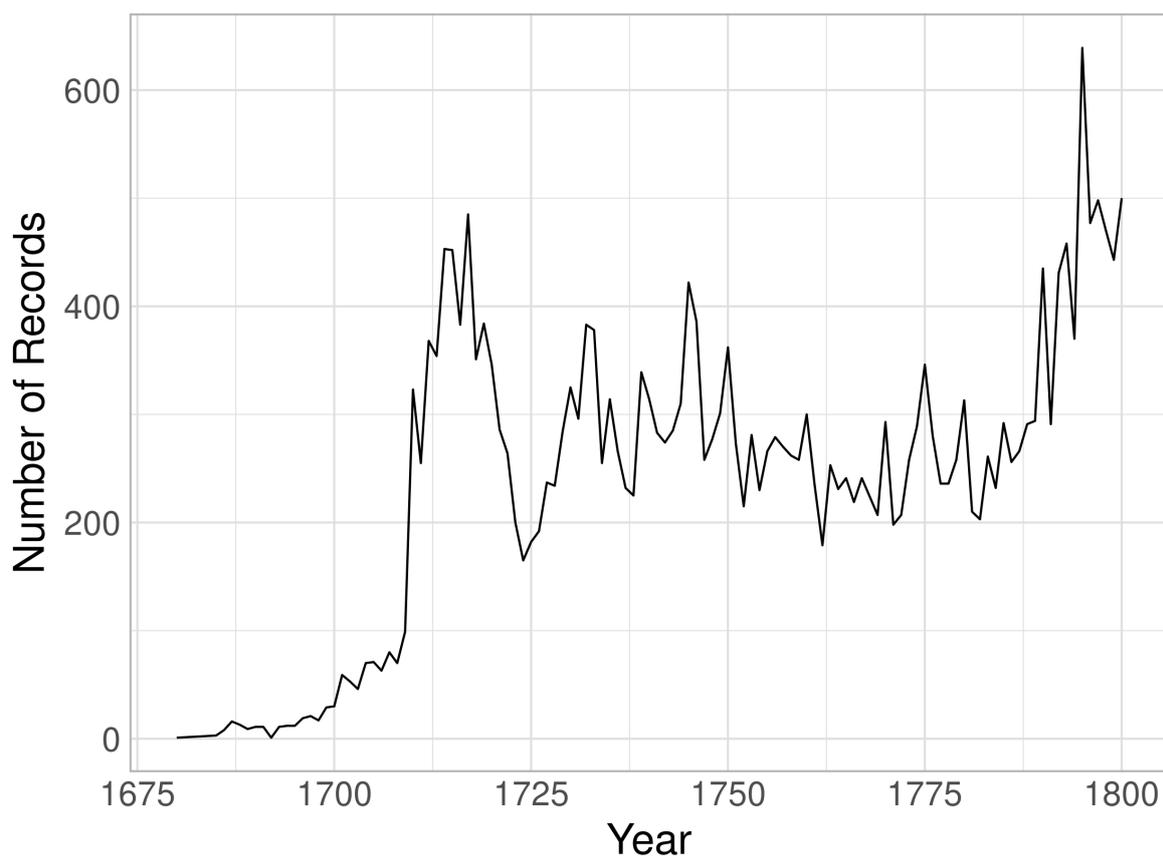
Figure 1: The observations used in this article, 28251 in total, by publication year. Most of the price information is from the eighteenth century.

2. Materials and Methods

2.1. Data and Preprocessing

The price information within the ESTC is scattered across the general notes field 500a of this MARC-annotated catalogue. This means that the price information had to be parsed amidst natural language. Roughly 6 percent of the half million ESTC records have a parsed price. In some cases, multiple prices are provided for different bindings and discounts for those that brought multiple copies, and sometimes the word price appears in other contexts than the price of the record. These more complicated cases sometimes led to errors in the extraction of price. The most suspicious prices (e.g very high) were manually checked and corrected if necessary. Unfortunately there is no information on why the note about price was made for these specific records, making it hard to evaluate how representative our data set is.

Prior to modeling, we made choices of data processing and selection to focus our analysis. We selected smallest price, adjusted it to the general price level and focused on the years 1680-1800. The choice of the smallest price might matter, as variants with a greater price might follow a different logic of pricing. To adjust prices to the overall price level of the publishing year and to be able to use the same price for normal linear and Poisson models, we regularised

each book price with the following formula:

$$y_{ij} = int(\frac{y_{ij}^*}{p_j}) \tag{1}$$

Where y_{ij} is the regularised price of the ith print product published in the year j, y_{ij}^* the original price in decimals of pence, p_j is the price index of London for the year j from the Prices and Wages in London and Southern England, 1259-1914 data set[2] and the operator int rounds the obtained value to an integer. We limited our modeling to the observations from the period 1680-1800 as data prior to that is very scarce, our main interest is the eighteenth century and the exchange rate of our main monetary units (penny, shilling, pound) stayed stable during the selected period [7]. This selection dropped the number of observations to 28251.

2.2. Quantitative Model

Existing literature made it reasonable to assume that the amount of paper and other materials used in a print product would have connection to its price [11, 14]. As the harmonised ESTC data collection also has estimates of these attributes, they made for a good starting points for modeling prices. The paper and plate size of the print product are both fixed effects in our model. They affect the price in a linear fashion: an increase of a fixed size (e.g addition of a sheet of paper) to a print product always adds the same amount of price to it. Linearity is both convenient to model and roughly corresponds with the notion that materials formed a significant part of the production costs of books in the early modern period.

Another question to address was that we had multiple reasons to assume that this general trend would not hold at all for some works. The phenomenon of luxury printing [13] is known in the literature of book history. The high price of a luxury item might not be strongly related to its basic physical properties, but to prestige or some other quality hard to measure. The very highly priced print products were in general very difficult for our model. Thus, our attempt was to capture a general association of price and size, but we are aware that it is prone to large errors. Additionally, both the price and the predictor variables include instances of false information, that can lead to severe errors. These considerations motivated us to use a Bayesian regression model, where the error term is Student's t-distributed. As this kind of model has a wider tail than the normal distribution usually used in regression models, it is less affected by deviations from the general trend. The posterior predictions of a Bayesian model are also more robust, as they include the uncertainty in the parameters.

Our model is defined as follows: Let Y be the price of a print product, x the vector of constant 1 and the associated features of physical properties (size in terms of paper and plates, paper and plate pages are assumed to be of the same size for the same observation) and $\beta = (\beta_0, \beta_1, \beta_2)$ the vector of the effects of the physical features and of the constant. Now, our regression model (note the 'vectorised' notation) is defined as
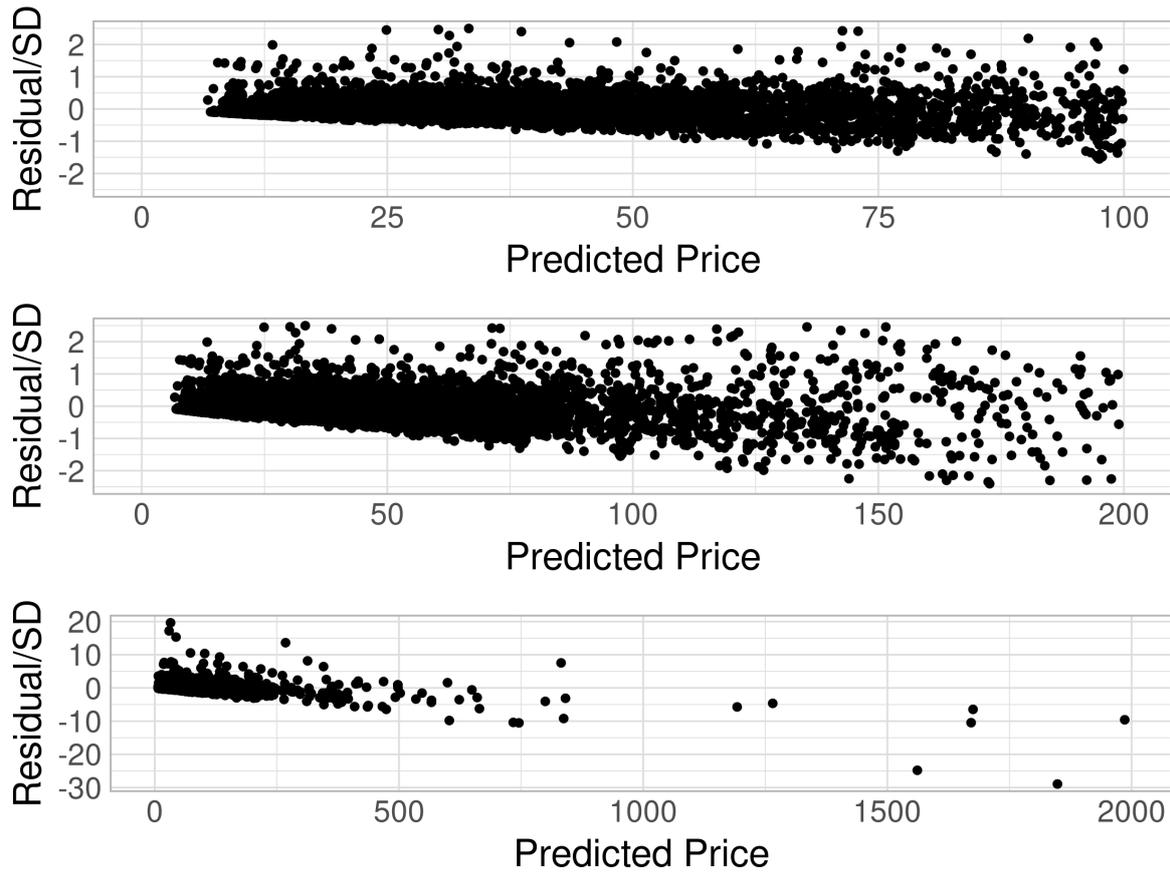
$$Y|\beta, \sigma, X \sim T(2.02, x^T\beta, \sigma) \tag{2}$$

$$\beta \sim T(1, 0, 1) \tag{3}$$

$$\sigma \sim IG(1, 1). \tag{4}$$

---

[2]Availabe at https://gpih.ucdavis.edu/Datafilelist.htm

Figure 2: Predicted price (mean of the posterior predictive distribution) and residuals divided with the standard deviation of the test data. The visualisations were done with different X and Y axis to demonstrate the importance of variation by price and the effect of outliers. 814 Observations of the total test data are missing from the top panel and 243 from the middle panel.

Where T marks the Student's t-distribution and IG inverse gamma distribution. The model fitting was implemented with STAN. The paper and plate variables were standardised prior to fitting.

The posterior distribution of parameters and evaluation metrics of predictions are summarised in table[3] 1. Standard deviation's change in the amount of paper has a larger effect than standard deviations change in plates. However, as majority of observations do not have plates, a standard deviations increase in plates should not be directly compared with a standard deviations increase in paper, as later corresponds to a larger physical increase. As plate estimates are also more prone to large errors than paper estimates, the posterior estimate for the effect of plates is less reliable. The data was split into 50% training and 50% test sets.

Table 1
Posterior Distribution of Parameters and Model Performance Evaluation Metrics.

Parameter	1st Quantile	Median	3rd Quantile	R^2	EMAE
				0.46	0.4
β_0 (constant)	38	38	38		
β_1 (paper)	56	56	56		
β_2 (plate)	11	11	12		
σ	9.4	9.4	9.5		

3. Results

We applied the model to evaluate the price trends of early modern Great Britain. These predictions are illustrated in figure 3, which plots the residuals of our model by year and publication place for roughly averaged sized print products. Here the residual of an observation is defined as

$$\epsilon_i = Y_i - \hat{Y}_i. \tag{5}$$

Where Y_i is the real and $\hat{Y}$ the predicted price of a print product. If the residuals of certain period differ significantly from 0, we can get insight to potential temporal trends of prices without modeling them explicitly. Negative residuals indicate that the model overpredicts prices and vise versa for positive residuals. As the residuals increase as a function of the print products size, we limited the analysis to print products that were of average (half a standard deviation from zero on standardised scale) size in terms of paper and plates. In the case of our model, we can say that variation in residuals by time should not be explained by (possible) associated changes in the physical composition of print products. We focused on the median and middle half of residuals to decrease the influence of outliers.

Our main question to address was the disputed effect of the affirmation of the Statute of Anne in 1774 that transformed copyrights of books from perpetual to the length of 14 years in Great Britain. Some have seen this as a revolutionary turning point that affected the prices and hence the accessibility of print products and their contents significantly [2], while others have challenged the impact of legislation to prices and the related claim about over priced printing prior to 1774 [14, pp. 25–30]. The larger context for which the question is relevant is how we understand eighteenth century British history with respect to access to culture and information: did the changes in legislation create a phenomenon of greater access to books? Regarding this question, the national character of the ESTC (it is overwhelmingly made of British print products) is ideal. Additionally, the question is important for understanding argued differences in printing (price and quality) between Britain and continental Europe.

The first decades of the eighteenth century at figure 3 could be interpreted as an increase of prices until 1730's. after which a long relatively stationary period follows. This increase is interesting, as Licensing of the Press Act lapsed in 1695. This lapse created a situation where the monopoly of the Stationer's Company - the guild responsible for printing - ceased. Our time series does not reflect this change that arguably opened the way for cheaper competing printing.[4]

---

[3]EMAE=Explained Mean Absolute Error. Defined as the proportion of the Mean Absolute Error explained by the model compared to the Mean Absolute Error of always predicting the median of the test data.

[4]This absence of any noticeable effect caused by the change of legislation might be explained by the possibilities for cheap competition being exploited in other places than London.
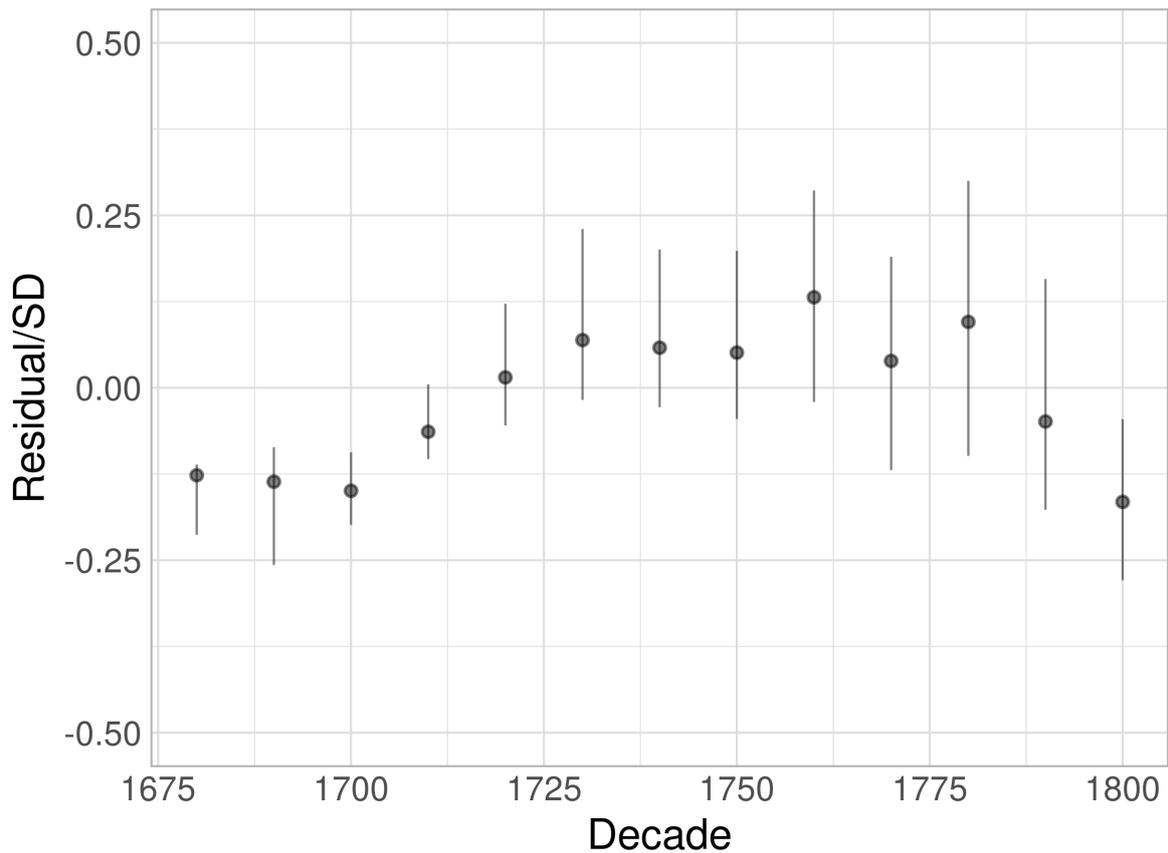
Figure 3: The median residuals of the test data for roughly average sized (standardised plates and paper $\pm$ 0.5 SD from Zero) print products 1680-1800 in London. Confidence Interval from 1st quantile of residuals to the 3rd quantile. Standardised by dividing with the standard deviation of the test data which was used to make this figure.

The increase is followed by stagnation that lasts until the last decade of the eighteenth century. It is interesting considering the claims about the structure of the British book trade of the period that make the affirmation of the Statute of Anne a noteworthy event. The change that the Statute arguably caused was to break the monopolistic arrangement of book trade centered around control of valuable copyrights that gave those holding them control over the price. This control meant that prices would (mostly) not fall it the copyright was not free from copyright restrictions [13]. The stagnation could be interpreted as stability of high prices made possible by the power that publishers (copyright holders) had over the prices.

This stable state is followed by a significant decrease in the 1790's and the first year of the nineteenth century. The decrease would align well with views that claim that monopolies and artificially high prices started crumbling at the end of the century. However, the change in residuals does not align perfectly with the change in legislation. One possible explanation is that the effect of the change in legislation increased in time, as larger quantities of print products became free from copyright. Although we can not argue that we observed the causal workings of copyright monopoly and its collapse, we can say that if that narrative was true, the variation we observed would make sense.

Our initial results differ from the results of earlier quantitative approaches. One index of the book purchasing power of eighteenth century's average Britton suggests that books became significantly harder to acquire for the average person during the later half of the eighteenth century [13]. This result is contrary to ours, and Gregory Clark's Productivity in Book Production in England, 1470–1860 index has a pattern of late eighteenth century decrease in the ratio of wages to book prices as well [3].

The models performance was evaluated both for the sake of the specific historical research case and to see how much potential it has as a more general tool of quantitative historical research. Figure 2 demonstrates that the model captures the trend of the majority of works quite well, but struggles significantly with some cases, overpredicting or underpredicting significantly. The model explains 46 percent of the variation in terms of R^2 and 40 percent of the Mean Absolute Error compared to the median of the test data. Later metric is more reliable, as R^2 is more heavily affected by outliers, some of which were removed because of unreliable paper or plate information in the harmonised ESTC. As is it is most likely that the data still contains erroneous information, we argue that our robust model captures an association of physical properties and price, and the strength of this association might be stronger than indicated by the evaluation metrics. At the same time, some outliers are most likely the products of different logic of pricing, e.g luxury products.

4. Conclusions

The model succeeded in the task that motivated its creation, as it provided initial insight to the question of eighteenth century book price trends in Great Britain. The disputed decrease of prices was present in our analysis. Interestingly, our results differed from some of the earlier time series that approximate the relative prices of print products. A more thorough comparison and standardisation of these different quantitative approaches - including ours - could help the discourse about eighteenth century book prices to converge.

Meaningful application of the model in historical research requires evaluation of its performance. Our evaluation suggests that the model has significant predictive and explanatory power. This encourages us to develop the model further. The incorporation of spatiotemporal effects and addition of publisher and edition (first or re-publication) information could improve the models performance and usefulness significantly. It is also important to remember that there were significant outliers, and these might represent separate domain(s) of pricing. Detection and analysis of these cases in a quantitative manner is a possible topic for further research.

The most definite result is the proof of concept. We have demonstrated the usefulness of our model both in terms of technical measures and as a tool of which predictions can be interpreted as a historical narrative. There are various other applications for the model as well. For example, it can be used for extrapolating prices for the much larger set of works in the ESTC that do not have the price information but do have the features used in the prediction. Hopefully both the technical development and the application of the model will continue in the future.

Acknowledgments

References

[1] F. Ammannati and A. Nuovo. "Investigating Book Prices in Early Modern Europe: Questions and Sources". In: *JLIS.it* 8.3 (2017).

[2] W. S. Clair. "The Political Economy of Reading". In: *The John Coffin Memorial Lecture in the History of the Book*. 2005. URL: https://ies.sas.ac.uk/sites/default/files/files/Publications/StClair%5C%5FPolEcReading%5C%5F2012.pdf.

[3] G. Clark. *A Farewell to Alms: A Brief Economic History of the World*. Princeton University Press, 2007.

[4] G. Clark. "Lifestyles of the Rich and Famous: Living Costs of the Rich versus the Poor in England, 1209-1869". In: *Towards a global history of prices and wages*. 2004. URL: http://www.iisg.nl/hpw/papers/clark.pdf.

[5] J. Dittmar. "Book Prices in Early Modern Europe: an Economic Perspective". In: *Buying and Selling: The Business of Books in Early Modern Europe*. Ed. by S. Graheli. Brill, 2019, pp. 72–87.

[6] J. Dittmar and S. Seabold. "New Media and Competition: Printing and Europe's Transformation After Gutenberg ". In: *Conditionally Accepted to the Journal of Political Economy* (2021).

[7] C. Emsley, T. Hitchcock, and R. Shoemaker. *London History - Currency, Coinage and the Cost of Living*. 2021. URL: www.oldbaileyonline.org,.

[8] 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 (2015), pp. 87–116. DOI: http://doi.org/10.18352/lq.10112.

[9] L. Lahti, E. Mäkelä, and M. Tolonen. "Quantifying Bias and Uncertainty in Historical Data Collections With Probabilistic Programming". In: *Proc. CEUR Workshop Proceedings on Computational Humanities Research*. 2020, pp. 80–89. URL: http://ceur-ws.org/Vol-2723/short46.pdf.

[10] L. Lahti, J. Marjanen, H. Roivanen, and M. Tolonen. "Bibliographic Data Science and the History of the Book (c. 1500–1800)". In: *Cataloging & Classification Quarterly* 57.1 (2019), pp. 5–23. DOI: 10.1080/01639374.2018.1543747.

[11] D. McKenzie. "Printing and Publishing 1557–1700: Constraints on the London Book Trades". In: *The Cambridge History of the Book in Britain*. Ed. by J. Barnard and D. McKenzie. Vol. 4. The Cambridge History of the Book in Britain. Cambridge University Press, 2002, pp. 553–567. DOI: 10.1017/chol9780521661829.028.

[12] S. Pinker. *The Better Angels of Our Nature: Why Violence Has Declined*. Viking Books, 2011.

[13] J. Raven. "Book as a Commodity". In: *The Cambridge History of the Book in Britain*. Ed. by M. Suarez and M. Turner. Vol. 5. The Cambridge History of the Book in Britain. Cambridge University Press, 2009, pp. 83–117. DOI: :10.1017/chol9780521810173.005.

[14] R. Sher. *The Enlightenment and the Book : Scottish Authors and Their Publishers in Eighteenth-Century Britain, Ireland, and America*. The University of Chicago Press, 2006.

[15] M. Tolonen, L. Lahti, H. Roivainen, and J. Marjanen. "A Quantitative Approach to Book-Printing in Sweden and Finland, 1640–1828". In: *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 52.1 (2019), pp. 57–78. DOI: 10.1080/01615440.2018.1526657.

[16] J. van Zanden. *The Long Road to the Industrial Revolution : The European Economy in a Global Perspective, 1000-1800*. Brill, 2009.

A. Computational Reproducibility

The script used to produce the analyses in this article is available at:
https://github.com/COMHIS/article_2021_probabilistic_analysis_of_early_modern_british_book_prices.