Almelhem, Ali; Iyigun, Murat; Kennedy, Austin; Rubin, Jared

Working Paper

Enlightenment Ideals and Belief in Progress in the Run-up to the Industrial Revolution: A Textual Analysis

IZA Discussion Papers, No. 16674

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Almelhem, Ali; Iyigun, Murat; Kennedy, Austin; Rubin, Jared (2023) : Enlightenment Ideals and Belief in Progress in the Run-up to the Industrial Revolution: A Textual Analysis, IZA Discussion Papers, No. 16674, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/282801

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16674

Enlightenment Ideals and Belief in Progress in the Run-up to the Industrial Revolution: A Textual Analysis

Ali Almelhem
Murat Iyigun
Austin Kennedy
Jared Rubin

DECEMBER 2023

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16674

Enlightenment Ideals and Belief in Progress in the Run-up to the Industrial Revolution: A Textual Analysis

Ali Almelhem
The World Bank

Murat Iyigun
University of Colorado at Boulder and IZA

Austin Kennedy
University of Colorado at Boulder

Jared Rubin
Chapman University

DECEMBER 2023

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Enlightenment Ideals and Belief in Progress in the Run-up to the Industrial Revolution: A Textual Analysis*

Using textual analysis of 173,031 works printed in England between 1500 and 1900, we test whether British culture evolved to manifest a heightened belief in progress associated with science and industry. Our analysis yields three main findings. First, there was a separation in the language of science and religion beginning in the 17th century. Second, scientific volumes became more progress-oriented during the Enlightenment. Third, industrial works—especially those at the science-political economy nexus—were more progress-oriented beginning in the 17th century. It was therefore the more pragmatic, industrial works which reflected the cultural values cited as important for Britain's takeoff.

JEL Classification: C81, C88, N33, N63, O14, Z11

Keywords: language, religion, science, political economy, progressiveness, Enlightenment, industrial revolution

Corresponding author:
Murat Iyigun
Department of Economics
University of Colorado at Boulder
Boulder, CO 80309
USA

E-mail: murat.iyigun@colorado.edu

* We thank Joel Mokyr and Naci Mocan for extensive comments on earlier drafts. Aaron Berman provided excellent research assistance. We also thank Sascha Becker, Carola Frydman, Oded Galor, Walker Hanlon, Phil Hoffman, Noel Johnson, Mark Koyama, Stelios Michalopoulos, Petra Moser, Luigi Pascali, Louis Putterman, Jean-Laurent Rosenthal, Francesca Trivellato, Felipe Valencia, Nico Voigtländer, David Weil, Yiling Zhao, and participants at seminars at Brown, Cal Tech, Northwestern, Oxford, Peking University School of Economics, and University of São Paulo for incredibly helpful comments. All errors are ours.

1 Introduction

Economists have generated substantial empirical evidence in the last decade suggesting that cultural values can play a central role in economic growth and that these values have deep roots in a society's historical past (Spolaore and Wacziarg 2013; Enke 2019; Giuliano and Nunn 2021). Several studies have sought historical episodes that may have affected a society's cultural trajectory and proceeded to test the implications for growth therein (see, for instance, Nunn and Wantchekon (2011), Alesina, Giuliano and Nunn (2013), Grosfeld, Rodnyansky and Zhuravskaya (2013), and Schulz et al. (2019)).[1] This literature has yielded important insights regarding the way that cultural residues from historical events continue to impinge on economic growth. Yet, due to data limitations inherent in historical studies, works in this literature are rarely able to capture how and when culture changes *over time*.

This is not a trivial issue, empirically or conceptually. Empirically, it is often impossible to derive panel data on cultural phenomena, thus making it exceedingly difficult to trace cultural change. Conceptually, it is not always clear what might even constitute cultural data, let alone how one would collect it in the absence of historical surveys. One promising solution to this problem is studying *language*. There is much cultural information embedded in language, such as gender norms, religious norms, which occupations societies value, attitudes towards risk and human capital attainment, and much more (Chen 2013; Galor, Özak and Sarid 2020; Michalopoulos and Xue 2021; Erikson 2021; Giorcelli, Lacetera and Marinoni 2022). This makes language an incredibly useful tool for analyzing cultural differences across societies. Yet, analyzing a cross-section of language does not provide insight into *when* language or culture changed. The timing of such changes is important, particularly in studies which seek to identify the deep roots of growth.

This paper tackles these issues directly by providing empirical evidence linking cultural change—as embedded in language—to one of the most important episodes in economic history: Britain's industrialization. We employ textual analysis methods to the universe of digitized printed volumes contained in the Hathitrust Digital Library, written in English in England between 1500 and 1900, in order to shed new light on cultural changes that took place in Britain both prior to and after its industrialization.

There are many aspects of cultural change that may have impacted Britain's industrialization, any of which may be detectable in the corpus of works written in English. In order to narrow the scope of our study, we build on an insightful recent argument put forth by Joel Mokyr (2016) linking a progress-oriented view of science promoted by great Enlightenment

[1] For reviews of the literature at the intersection of culture and historical persistence, see Nunn (2014), Voth (2021), Cirone and Pepinsky (2022), Acharya, Blackwell and Sen (2024), and Lowes (2024).

thinkers, such as Francis Bacon and Isaac Newton, with what would become the "Industrial Enlightenment," and ultimately Britain's Industrial Revolution. This progress-oriented view centered around the idea that science and our understanding of the natural world could be used to improve the lot of humankind. Mokyr argues that such a progress-oriented view of science gave birth to a pan-European "culture of growth." This culture was, in turn, sustained and supported by the social norms of elite intellectuals that fostered the free flow, dissemination, and discussion of new ideas across Europe. Accordingly, it was these cultural values in combination with Britain's abundance of skilled craftsmen and artisans that made its industrialization possible.[2]

The evidence Mokyr presents is abundant and convincing. Yet, there are two margins on which evidence is lacking. First, while the attitudes of elite thinkers and scientists were clearly becoming more progress-oriented in this period, it is far from clear that these ideas spread to those artisans and craftsmen who ended up becoming the driving force of Britain's Industrial Revolution. It is certainly important that elite thinkers held progress-oriented views—upper-tail human capital has been shown to be an important precursor of industrialization (Squicciarini and Voigtländer 2015)—but the "Industrial Enlightenment" was largely advanced by artisans and those with closer ties to industry.[3] Second, qualitative evidence by construction cannot account for the hundreds of thousands of works produced in this period. Is it possible to marshal quantitative evidence that the language of science became more progress-oriented in this period? If so, such evidence may provide insights that elude qualitative studies.

Our textual analysis addresses both of these issues. We analyze textual data gathered from the Hathitrust Digital Library, which comprises digital scans and optical character recognition (OCR) output. Once we account for duplicates and volumes that cannot be read via OCR, this yields 173,031 unique volumes published between 1500 and 1900. We begin the analysis by employing a technique popularized in recent machine learning and applied statistics literatures called Latent Dirichlet Allocation (LDA; see Blei, Ng and Jordan (2003)). This method extracts latent topics in a large corpus of text, allowing us to analyze the distribution and evolution of these topics across time. The LDA views individual volumes as a "bag of words," looking for words that commonly appear together within the same volume regardless of the order they appear. Importantly, this method does so in an unsupervised fashion, divorcing our results from any prior beliefs or scholarly interpretations on the history of economic development in Europe during the period.

[2]For the recent debate on the causes of Britain's Industrial Revolution, see Mokyr (2009, 2016), Allen (2009), Koyama and Rubin (2022, ch. 8), Kelly, Mokyr and Ó Gráda (2023), and Heblich, Redding and Voth (2022).
[3]By the 17th century, British literacy rates were above 50% (Buringh and Van Zanden 2009, Table 9). Hence, even most of those outside of the upper-tail human capital had access to the written word.

The LDA yields 60 *topics* based on words that frequently co-exist with each other and are unique from other topics. We then employ an algorithm to determine the sets of topics that most frequently co-exist with each other. The top three sets of unique topics clearly relate to three different *categories*: science, religion, and political economy.[4] Using these categories, we are able to derive time-varying categorical weights for all sixty topics with respect to science, religion, and political economy. We are able to create similar weights for each volume in our dataset.

We proceed to create an "industrial score" for each volume. To do so, we transcribe the detailed indexes of five volumes of *Appleby's Illustrated Handbook of Machinery* (Appleby 1877–1903). These industrial manuals, published in the late 19th century, cover nearly all aspects of industrial production. Each volume in the corpus is given an industrial score based on the occurrence within each volume of the (weighted) list of root words associated with industrialization.

This process yields three sets of results. First, we quantify how the relative weights of the corpus of volumes produced in English changed over time. We calculate these weights for each volume in the corpus. The results indicate that as early as 1600, and certainly by the mid-17th century, there was little overlap of scientific and religious topics within volumes. In other words, the language of science was secularized by the early Enlightenment. However, there was a shared language for scientific and political economy volumes throughout the period, and there was likewise a shared (though different) language for religious and political economy volumes. These trends are largely stable between the period 1650 and 1900.

Second, we proceed to test the theory espoused by Mokyr (2016) that the language of science became more progress-oriented during the Enlightenment. To this end, we assign each volume a "progress-oriented sentiment" score based on the presence of progress-oriented words (obtained from various dictionaries and thesauruses; see Section 4.1) contained in the volume. We proceed to attach these sentiment scores to the volume's categorical weight (i.e., science, religion, political economy). This exercise yields sentiment scores by category over the period 1500 to 1900. We find that the language of science started to become more progress-oriented in the 18th century and this persisted throughout the period in question. However, there is an important caveat to this finding: works of "pure" science were largely neutral with respect to progress-oriented language. The most progress-oriented works were at the *nexus* of science and political economy.

Third, we test the more specific implication from Mokyr (2016) that works associated with *industrialization* became more progress-oriented during the Enlightenment. We find

[4]The algorithm yielded distinct enough categories that we could clearly, yet subjectively, label them as science, religion, and political economy. See Table 1 and the related discussion for more on the categorization process.

that, beginning in the 18th century, works with higher industrial scores were more progress-oriented. This result is strongest for works related to industrialization that were at the nexus of science and political economy.

These findings have significant implications for the way we conceptualize the role of language in industrial, scientific, technological, and economic development. In particular, they suggest that progress-oriented views were imbued in the types of industrial volumes that sought to reach both a scientific and non-scientific audience; those with some type of political or economic (but not religious) interest. This is highly consistent with the idea of "Industrial Enlightenment," espoused in Mokyr (2009), which emphasized that Enlightenment ideals diffused into mechanical and artisanal applied pursuits. Our results suggest that it was precisely at this nexus where language became more progress-oriented in the period preceding the Industrial Revolution. As Mokyr (2009, 2016) suggests, these Enlightenment ideals—diffused to elite artisans and skilled craftsmen—likely played a central role in increasing the rate of technological innovation, especially technology related to industry. The idea that science and technology could be used for the betterment of mankind was a key cultural component of the massive economic and technological changes characterizing 18th and 19th century Britain.[5]

This paper relates closely to many recent works that take advantage of new computing techniques, stronger computing power, and advances in OCR to use "words as data" (Grimmer and Stewart 2013; Gentzkow, Kelly and Taddy 2019). Topic modeling has recently been used in a wide variety of contexts in political science, economics, and the humanities. Erikson (2021) applies LDA and sentiment analysis to political and economic tracts written in England from 1550 to 1720. She finds that the language of economics increasingly spoke to a wider audience—moving away from appeals to religion—as trade expanded and appeals to the "good of the nation" to justify economic privileges and charters became more common. We similarly find a move away from the language of religion in works of science and political economy in this period, although our focus is more on the language of science, industrialization, and progress. In another work closely related to our study, Grajzla and Murrell (2019) apply topic modeling to the set of Francis Bacon's works to study the features and origins of his ideas, and how they led to the political and economic development of England.[6]

[5]These insights are also related to the insights of McCloskey (2006, 2010, 2016), who argues that changes in rhetoric favoring "bourgeois virtues"—specifically, the way people spoke about work, profit, and industry—played a key role in northwestern Europe's takeoff. Although we do not test McCloskey's theory directly, a clear implication of this theory is that the language of political economy should have become more progress-oriented in the 17th and 18th centuries (at least, for works written in English and Dutch). White (1978) argued that such attitudes favoring hard work and industry had medieval roots. This theory is outside the scope of our paper given the coverage of our data.

[6]Other works using similar techniques abound. For instance, Blei and Lafferty (2009) apply LDA to 100 years of articles from the journal *Science* to demonstrate the effectiveness of topic modeling in uncovering

Similarly Giorcelli, Lacetera and Marinoni (2022) use text analysis to show that Darwin's ideas diffused throughout public discourse after the publication of *On the Origin of Species*. Our paper likewise draws a connection between scientific advances and cultural change, but for an earlier period.

The paper proceeds as follows. Section 2 describes the data and the data extraction methodology. Section 3 describes our method for classifying each topic into three categories (science, religion, and political economy) and presents the results for each topic and each volume over time. Section 4 lays out our strategy for classifying volumes as "progress-oriented," reporting how these results change over time. Section 5 presents how language related to industrialization changed over the period under study. Section 6 presents qualitative examples of industrial volumes that used progress-oriented language. Section 7 concludes.

2 Data and Methodology

2.1 Data from the Hathitrust Digital Library

We collected data from the Hathitrust Digital Library (HDL), a collaboration between major universities in the US (now the Big Ten Academic Alliance) and the University of California public system. The HDL aims to establish a shared repository of digitized works from member universities for archival and non-consumptive research purposes. This includes materials digitized by Google, Microsoft, or the Internet Archive that exist in both the copyrighted and public domain. Additionally, HDL provides the computational infrastructure for large-scale text mining and algorithmic analysis. The HDL repository consists of over 17 million volumes from over 150 universities worldwide and allows access to this corpus for search and discovery to the fullest extent possible. Our data set covers all 173,031 unique works

macroscopic features and dynamics over time. In the social sciences, two prevalent examples are Blaydes, Grimmer and McQueen (2018), who use topic models to compare Muslim and Christian political advice texts and show how these texts evolved over time, and Hanson, McMahon and Prat (2018), who apply LDA to the FOMC meeting minutes to uncover general communication patterns and the impact of greater transparency on member behavior.

originally printed in England and written in English in the HDL over the period 1500–1900.[7] Figure B.1 reports the distribution of volumes in our data set by year.

We are interested in the content of the volumes. Due to copyright law, we could not simply download each volume. Instead, we used HDL's Extracted-Features dataset, which models each volume as a "bag of words". A bag of words model is a representation of textual data. It simplifies a document to a multiset of its words, disregarding the order of words while retaining word multiplicity. The bag of words model is only concerned with how often words occur in the document, without regard to where they occur. This permits insight into the underlying topics, sentiment, and keywords without the text being comprehensible from a syntactic point of view.

2.2 Data Processing

In order to meaningfully analyze the HDL data we first condense the vocabulary to a set of terms that is most likely to reveal the underlying content of each volume. We begin by grouping each volume by quarter-century beginning in 1500 and ending in 1900 (i.e. 1500–1524, 1525–1549, and so on).[8] We do this because language changes over time, and we do not want language employed at the end of our sample, for which there are many more books, to determine results for previous periods.

We cull the list by removing any duplicate volumes and volumes printed in Latin. We use the Online Computer Library Center (OCLC), the world's largest online public access catalog, to identify duplicate volumes based on OCLC catalog number. In the case of English corpus, we begin with 420,081 volumes. This is reduced to 173,031 after removing duplicates.

[7] There is a potential bias in the HDL data: the libraries from which the HDL has digitized books may be biased towards the predilections of librarians or professors. While the HDL data are the best available in terms of fully digitized, machine-readable tracts, in order to properly analyze these data it is necessary to know what the biases are. We address this issue in Appendix C, where we compare the HDL data with the data available in the English Short Title Catalog, which is a "comprehensive, international union catalogue listing early books, serials, newspapers and selected ephemera printed before 1801." The results reported in Appendix C suggest that there are no biases in the HDL data with respect to scientific works, although the HDL data has relatively fewer religious works and (slightly) relatively more political economy works. As noted in the appendix, these biases can mostly be accounted for by the ESTC containing ephemera such as sermons, whereas the HDL data has little ephemera. In short, biases in the HDL data are small, and if they exist at all are in the direction of under-counting religious documents. To the extent that the progress-oriented documents would be the most likely to survive in both data sets—which we believe is the most likely case—our results can be interpreted as an upper bound on progress-oriented sentiment for religious works. There also may be bias in the HDL data towards more recent works, since these volumes are more likely to both still exist and be in good enough shape to be digitized. We do in fact find slight biases towards more recent books in the HDL data. If anything, these biases should favor us finding more progress-oriented volumes early in the period under question, as these volumes are more likely to survive. This works against the hypothesis of finding a rise in progress-orientation in the build-up to industrialization.

[8] The publication year is retrieved from the Library of Congress, which we view as the most accurate source for volume metadata.

We then "clean" each volume. The first step of cleaning involves tokenization, which is the process of converting each word, or any string of characters separated by whitespace, into a separate token. From these tokens, we remove punctuation, numbers, or any other non-alphanumeric characters (such as parenthesis, dashes, or pound signs).

Because of the nature of our data, comprising of scanned pages and the associated OCR output of extremely old books, we encountered some obvious errors that needed manual correction. One example is the "long-S" correction, in which books printed prior to 1650 often had the character 's' incorrectly identified as an 'f' because of the type of font used. To remedy this issue, we resort to manually identifying unambiguous word corrections, such as 'juftice' to 'justice', and replace them.[9]

We proceed to remove stop words such as 'the' and 'of' that appear frequently in all volumes and do not convey any meaningful information. We then remove any short words less than three characters long or words that occur less than twice in the volume. Words with these features are likely to either not provide much contextual meaning or be an OCR error. We convert each of the remaining words to their respective roots by removing any inflectional affixes such as suffixes or prefixes (e.g., playing, player, and played all map to play).

Finally, we follow the suggestion by Blei and Lafferty (2009) and rank the remaining words by their term frequency-inverse document frequency (tf-idf) score. This metric is a measure of informativeness that boosts the ranking of words that occur frequently in one volume, and less frequently in all other volumes, indicating the importance of this word in a particular volume.

The tf-idf score is calculated as follows. Term-frequency (tf) is:

$$\text{tf}_v = 1 + \log(n_v),$$

where n_v is the frequency with which the term v appears in a specific volume. The inverse document frequency (idf) score is:

$$\text{idf}_v = \log(D_v/D),$$

[9] Another common error is the use of Greek instead of English letters. Examples are the use of β instead of the letter 'B', or accented vowels. Again, we manually replaced each of these characters with its nearest equivalent in the English language.

where D_v is the number of volumes in which term v appears, and D is the total number of volumes. Finally, the tf-idf score is:

$$\text{tf-idf}_v = \frac{\text{tf}_v}{\text{idf}_v}.$$

For each volume, we drop the bottom 20% of all words ranked by their tf-idf score. The final result is a bag of words representation of each volume, printed to text files to be used as input for the LDA model described in the next section.

2.3 Data Extraction via Latent Dirichlet Allocation

We proceed to produce a set of *topics* from the bag of words corpus described above. A topic is a set of root words that commonly co-occur within volumes. In order to produce a set of topics, we employ the Latent Dirichlet Allocation (LDA). The LDA is a generative statistical model developed to extract macroscopic features of a given corpus comprised of many individual documents (Blei, Ng and Jordan 2003). Consistent with the HDL data, the algorithm does not view a document as an ordered set of words, but rather as a bag of words where only the word and its corresponding frequency matters. The model assumes the data generating process is modeled as a Dirichlet distribution, where each document is a multinomial distribution over topics, and each topic is another multinomial distribution over words. Each document in our corpus is generated by repeatedly sampling from this distribution, given the proportion of topics present in each document.

For a set of observed documents, the algorithm derives the optimal Dirichlet distribution such that the observed corpus would be generated by repeated sampling from this distribution. As a result, each topic is neither semantically nor epistemologically defined, but rather is identified by groups of words that tend to co-occur.

The corpus is initially modeled as a document-term matrix $D \times V$, where D is the number of volumes and V is the number of words in the vocabulary (V is very large). After estimating the LDA model, the result is a new representation of each volume as a mixture of topics, rather than a mixture of words. This reduces the dimensionality of a corpus to a $D \times T$ matrix, where T is the number of topics. In short, the algorithm reduces the dimensionality of the data set from many thousands of words to T topics, where T is determined by a process described in Section 2.4.

The following example helps clarify what the algorithm does. Say that we train an LDA model on a set of documents taken from two journals, one in chemistry and the other in sociology. The algorithm will identify the type of vocabulary used in each subset by discovering words that frequently co-occur. These frequently co-occuring words are then

organized into topics, likely ones specific to jargon used in chemistry and sociology. The topics themselves are a multinomial distribution over a vocabulary, which was repeatedly sampled to produce this set of observed documents. Each word is not restricted to one topic. Words can appear in multiple topics in various proportions (for instance "equilibrium" may appear in both chemistry and sociology topics).

2.4 Model Selection

Two challenges in unsupervised machine learning algorithms are judging model quality and parameter tuning. Our data are unlabeled, meaning that we have no clearly identifiable way of determining if the LDA model is a fair representation of our dataset.[10] Moreover, it is not clear ex ante which model to select between those with different parameters. We address these issues with the *perplexity* measure frequently used in the machine learning literature to determine statistical goodness-of-fit.

In information theory, perplexity calculates how well a certain probability distribution predicts a given sample. Specifically, it computes the probability that an unobserved sample is generated from a given probability distribution. This allows us to compare between different probability distributions, with a lower perplexity score suggesting that a model is better at predicting the sample. We calculate perplexity in combination with another technique from the machine learning literature, cross-validation, which partitions the data into K folds (in our case $K = 4$).[11] In each fold, 75% of the data is used to generate the probability model (training data), and the remaining 25% is used to measure how accurate the model performs on this unseen data (testing data). The training and testing data are rotated in each fold to balance any bias in the selection of data for all folds. We repeat this procedure for each of the K folds to determine the average perplexity across each parameter setting, and choose the parameters of our model that minimize this metric. The parameter we tune

[10] Since the true distribution of topics within the corpus is unknown, we cannot compare our model against the true model. The true model is the correct representation of the corpus with respect to both topic composition and distribution. The LDA model creates groupings of topics and volumes without any idea of what is considered a correct or incorrect grouping. If the data had labels, we could compare our model to those true labels and test whether our LDA model is a fair representation of our corpus.

[11] K-fold cross validation is used for two main purposes: to tune hyper-parameters and to better evaluate the performance of a model. K is therefore selected to ensure that the training set and testing set are drawn from the same distribution, and that both sets contain sufficient variation such that the underlying distribution is represented. For example, in a 10-fold cross validation with only 10 instances, there would only be 1 instance in the testing set. This instance does not properly represent the variation of the underlying distribution. Selecting K is not an exact science, as it is hard to estimate how well a fold represents the overall dataset. 4-fold and 5-fold cross validation means that 25% and 20% of the data, respectively is used for testing. This is typically pretty accurate for data sets of the size used in this paper.

is the number of topics T, in addition to the Dirichlet priors alpha and beta. As seen in Figure B.2, this process yields an optimal number of topics at $T = 60$.[12]

3 Classifying Volumes by Topic

The purpose of this paper is to uncover how, if, and when the language of science changed in early modern and industrial England. To do so, we first need to classify volumes by topics. We can then explore how the topic content of the entire corpus evolved over time.

Appendix A lists each of the 60 topics derived from the LDA described in Section 2.3. Some topics are clearly science-based. For instance, topic 7 {fig water iron engin pressur steam electr} and topic 44 {plant flower stem genus yellow calyx bot} are both science-based. Many topics do not obviously fall into one category. In order to achieve a more systematic categorization, we discern how often topics *co-exist* in the same volume. The goal is to find *categories* of topics (e.g., science, religion, political economy) that have a high relative importance in the corpus and are distinct from each other. We can then use these categories as the basis for categorizing all other topics based on how often they co-exist with the topics used for categorization.

3.1 Categorization

The categorization process proceeds as follows. We first identify the distributions of all 60 topics for each volume in our corpus. For each volume, each topic has a weight representing its occurrence in the volume, and these weights sum to one per volume. We use these weights to discern how often two topics co-exist in the same volume. This is found by multiplying, within each volume, each weight by every other topic weight within the volume to get topic-pair weights per volume. This yields $\frac{60!}{2!(60-2)!} = 1770$ topic-pair weights. Unlike the frequency of the sixty topics—which adds up to one per volume—the topic-pair weights per volume do not sum up to one.

In order to place the topics into categories, we proceed to identify the most-frequently occurring topic-pairs for all volumes. First, we calculate the share of each topic-pair for the entire corpus over time. Next, we identify the share of any given topic weight across all volumes expressed as a fraction of all of the topic-pair weights summed across all volumes.

[12] Due to the large number of volumes and extremely high dimensionality of the data, we used computing resources provided by the Rocky Mountain Advanced Computing Consortium (RMACC). We used the RMACC Summit supercomputer, which is supported by the National Science Foundation (awards ACI-1532235 and ACI-1532236), the University of Colorado Boulder, and Colorado State University. The Summit supercomputer is a joint effort of the University of Colorado Boulder and Colorado State University.

Letting w_{iv} denote the weight corresponding to a given topic-pair $i \in \{1, ..., I\}$ with $I = 1,770$ and volume $v \in \{1, ..., V\}$, we have:

$$Share_i = \frac{\sum_{v=1}^{V} w_{iv}}{\sum_{i=1}^{I} \sum_{v=1}^{V} w_{iv}}. \quad (1)$$

In order to categorize the topics, we identify the most frequent and distinct topic-pairs that appear across all volumes over time. We chose to establish three categories of topics.[13] This was a subjective choice, based on our reading of the topics, many of which seemed to fall into religion, science, or political economy. A similar algorithm as the one described below could be used to break the corpus into more (or less) categories.

We use multiple topics per category rather than using individual topics in order to more accurately place topics in relation to the categories. For example, topics such as botany and chemistry may not share enough of the same language or appear together frequently enough for their topic-pair *Share* to identify them as similar. Yet, most would agree that both of these topics fit into a similar broad category, i.e. hard science. Thus, including more topics to represent a category increases the chance of accurate categorization.

The topics that form the basis for each category should have two features: high relative importance in the corpus and independence from topics in other categories. To determine which topics satisfy these criteria, we generate every possible combination of three topics, i.e. every possible category. This gives us $\frac{60!}{3!(60-3)!}$ potential categories.

We then established the relative incidence of each category. To do this, for each category we summed the topic-pair shares of all three topics, i.e. $Incidence_c = \sum_{i \in c} Share_i$ for topic-pairs i in category c.[14] For example, if the category is the topics 1, 2, and 3, then we sum the value of *Share* for the topic pairs 1 and 2, 2 and 3, and 1 and 3. We then ranked categories by their incidence.

The value of *Share* is generally high for topic-pairs in which both topics occur relatively frequently in the corpus. Thus, a category having a high value for *Incidence* indicates that its topics have a high relative importance in the corpus individually. Moreover, the value of *Share* is high for a topic-pair only if the topics frequently occur together within volumes. Therefore, if *Incidence* takes on a high value, it indicates that the three topics within the category occur together often throughout the corpus.

We proceed to rank each possible category from highest *Incidence* to lowest. To determine the three categories we use in the analysis, we ensure that the categories are distinct from each other. We therefore select the three categories with the highest *Incidence* value

[13] Results are similar when using four or five categories.
[14] We derive similar results using a product rather than a sum.

that have none of the same topics as the other two categories. We also exclude categories that contain innocuous topics (i.e., topics that commonly occur alongside others without providing additional meaning to them, see Appendix A). The categories (labeled manually) and their topics produced by this process are presented in Table 1.

Table 1: Categories

Category	Topics and associated words
"Political Economy"	33 - law lord show public evid opinion fact 34 - govern nation polit parliament constitut war parti 47 - trade amount labour money price cent increas
"Religion"	4 - church christian christ bishop holi paul doctrin 12 - god christ lord thi faith holi sin 52 - hath fame religion men shew virtu likewis
"Science"	7 - fig water iron engin pressur steam electr 8 - acid solut heat carbon water sulphur iron 41 - line angl equal equat sin sun plane

3.2 Placement and Evolution of Topics

We proceed to place individual topics relative to the three categories laid out in Table 1. The objective is to understand how close topics are to one category or another and to observe how this changes over time.

We first start by recalculating equation (1), using a 20-year (+/- 10 years) moving bin of volumes instead of the entire corpus. We use this moving bin due to the low number of volumes in early years. This gives each of the 1770 topic-pairs a *Share* value for each year between 1510 and 1890.

Next, for a given topic and category, we sum together the topic-pair shares of the topic itself and the topics in the category. For example, taking topic 1 and the political economy category, we sum together the topic-pair shares of 1 and 33, 1 and 34, and 1 and 47.[15] We perform this process for every topic and all three categories for each year. Thus, for each topic we have a yearly "score" for each category. Within a topic-year these scores are

[15] If the topic is in one of the categories, for this calculation we sum together the topic-pair shares of the topic and the other two topics in the category, then multiply by 1.5. We again use a sum here, as a product gives an outsized weight to low shares. In this example, topic 1 may be very close to topic 33 but not topics 34 or 47; we still want topic 1 to be considered close to the political economy category. Summing the shares gives this result, whereas using a product would over-weight the fact that topic 1 is not close to 34 and 47, thus underestimating its closeness to the political economy category.

meaningful. A higher score for one category over the other indicates that the given topic co-occurs more frequently with the three topics listed in Table 1 in that category.[16]

We then divide the raw category scores by the sum of all three category scores for each topic in each year. This provides, for each topic, a convex combination where the coefficients represent the extent to which the topic corresponds to each category. We plot these coefficients within a unit simplex with the categories as vertices. We present the results for each half-century in Figure 1.[17]

There are at least three salient facts to note based on Figure 1. First and foremost, our corpus is fairly thin in the earlier eras, especially in 1550. Hence, the conclusions based on or driven by the data from earlier periods need to be interpreted with caution. Second, there is a clear trend which started to take hold in first half of the 18th century whereby the scientific and religious topics become "purified" of each other. In particular, one can see that the frequency of publications or volumes in our corpus that combine topics of science with religion start to thin out after 1750, setting a trend which continues and holds through the end of our sample period. By contrast, there is a visible and steady shift toward publications that combine religious topics with political economy as well as those that involve science and political economy. Finally, and bearing in mind the caveat we expressed at the outset of this paragraph, the separation of scientific output from religious is a trend that predates the onset of the Industrial Revolution. This observation supports the influential ideas espoused by Mokyr (2016).

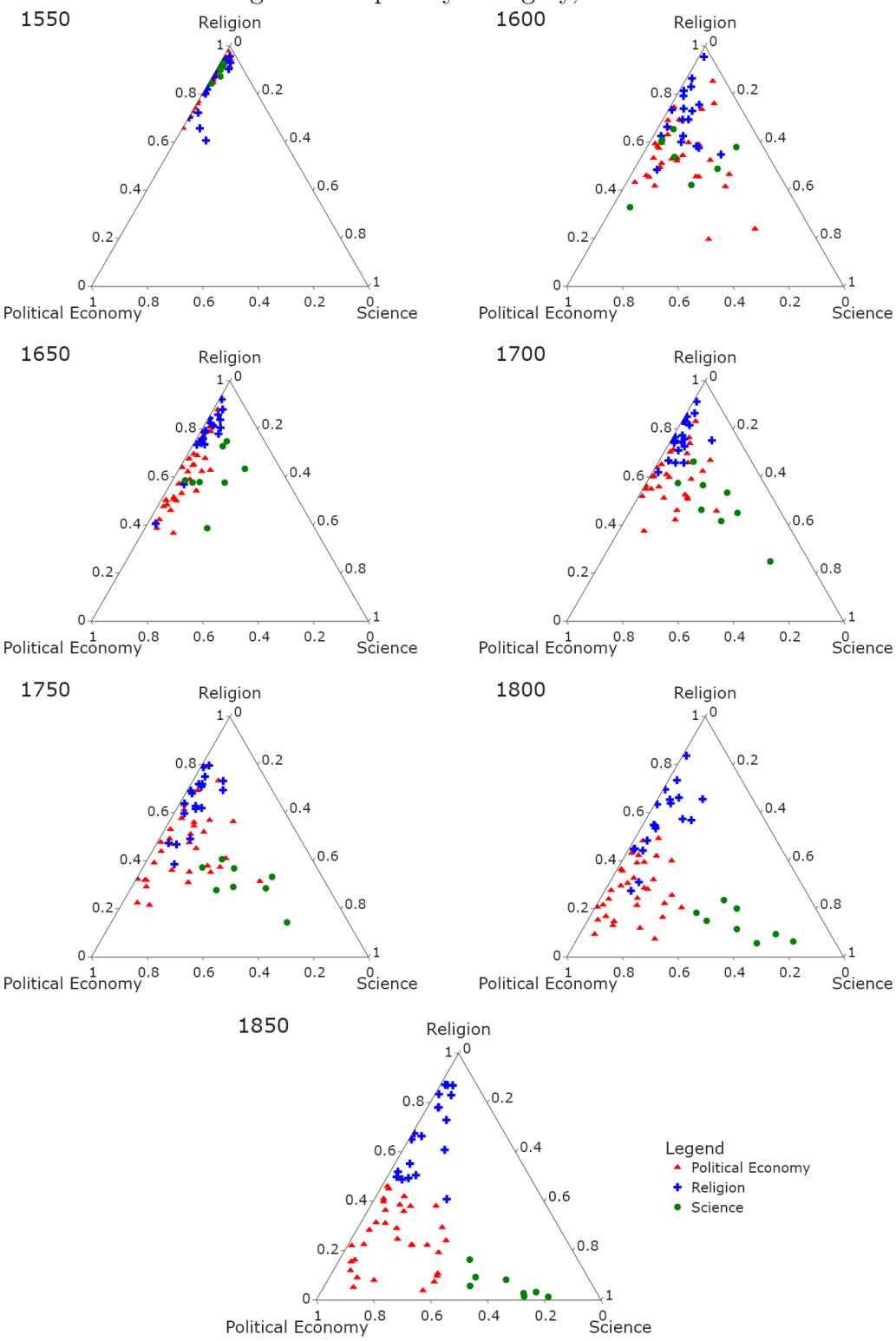
3.3 Volume Classification

We proceed to classify individual volumes into the three categories: science, political economy, and religion. This permits us to examine how categories evolved in relation to each other within volumes over time. We begin with the convex combinations constructed in the previous section. For each year and each topic, we have three coefficients which represent the weight of each category for that topic and year. We also have the original topic weights for each volume from the output of the LDA model.

[16] The raw category scores on their own do not allow us to compare topics directly to each other or to themselves across time. Given that they are calculated by adding together *Share* values, those topics that occur more frequently in general have higher *Share* values, and therefore will have higher scores for *all* categories than those topics that occur less frequently. Instead we want a higher relative category score to indicate that a topic is closer to a category than another topic is.

[17] Yearly results are available upon request. Note that the topics which make up the categories are not exactly in the corner of the triangle which represents them. This is to be expected, as the categories were chosen over the entire corpus, but the topics evolve individually over time. For example, the language used to describe a scientific subject may have been more intertwined with religious language earlier in the corpus, and become less so later in the corpus.

Figure 1: Topics by Category, 1550–1850



Note: Categorization into "Science", "Political Economy" or "Religion" based on topics' placement in 1850.

We take each volume and multiply the weight of each topic by the category coefficients for the corresponding topic and year. This scales the category weights by the topic weights within the volume. If a topic heavily represents one of the categories but does not occur much in the volume, it will be reflected in this calculation. We can therefore create the following category coefficients for each volume v:

$$\text{Science}_v = \sum_{t=1}^{60} \alpha_{t,v}\beta_{t,\text{Science}}, \tag{2}$$

$$\text{Political Economy}_v = \sum_{t=1}^{60} \alpha_{t,v}\beta_{t,\text{Political Economy}}, \tag{3}$$

$$\text{Religion}_v = \sum_{t=1}^{60} \alpha_{t,v}\beta_{t,\text{Religion}}, \tag{4}$$

where $\alpha_{t,v}$ is the weight of topic t in volume v and $\beta_{t,c}$ is the category coefficient of topic t for category $c \in \{\text{Science}, \text{Political Economy}, \text{Religion}\}$. Note that for each volume, $\text{Science}_v + \text{Political Economy}_v + \text{Religion}_v = 1$.

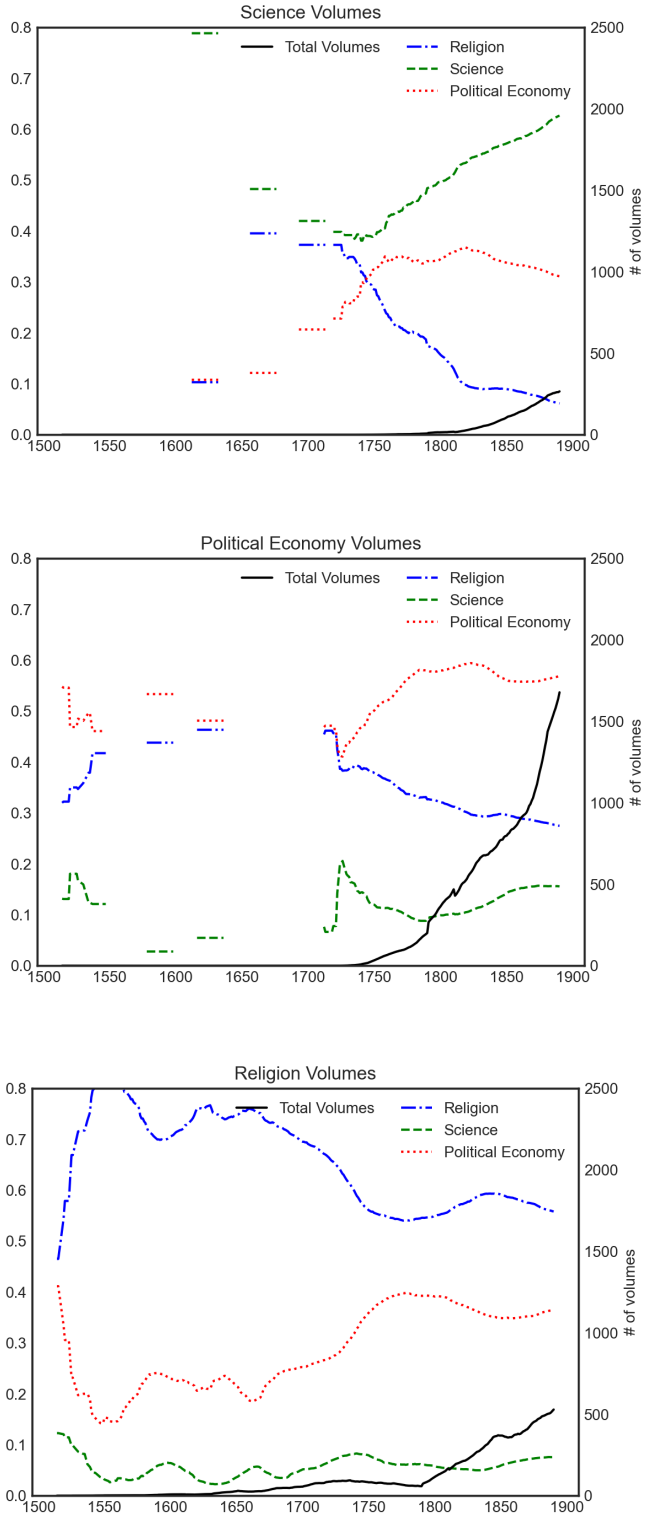
These weights allow us to classify each volume as predominantly science, political economy, or religion, based on the volume's highest weight derived from equations (2)–(4). Meanwhile, the weights also permit a measurement of how closely related a volume is to the other two categories.

Our categorization allows us to analyze how volumes in a particular category evolve over time in their relation to the other categories. Figure 2 reports the relation between the different categories over time. In these figures, we classify each volume as one of the three categories, and given this classification take the weights placed on each category. We sum these weights for each category and each year (smoothed over 20-year intervals).

As seen in Figure 2, volumes classified as science saw their science language increase significantly beginning in the early 18th century, starting with around 40 percent scientific language in 1700 and culminating with over 60 percent similar language by 1850. This increase came at the expense of religious language. Around 1700, science volumes used on average 40 percent religious language, while by 1850 they were only comprised of only 10 percent religious language. The final panel of Figure 2 indicates that beginning in the late 17th century, the language of political economy became more frequently used in works of religion. However, the language of science was rarely used in religious works throughout the entire period in question.

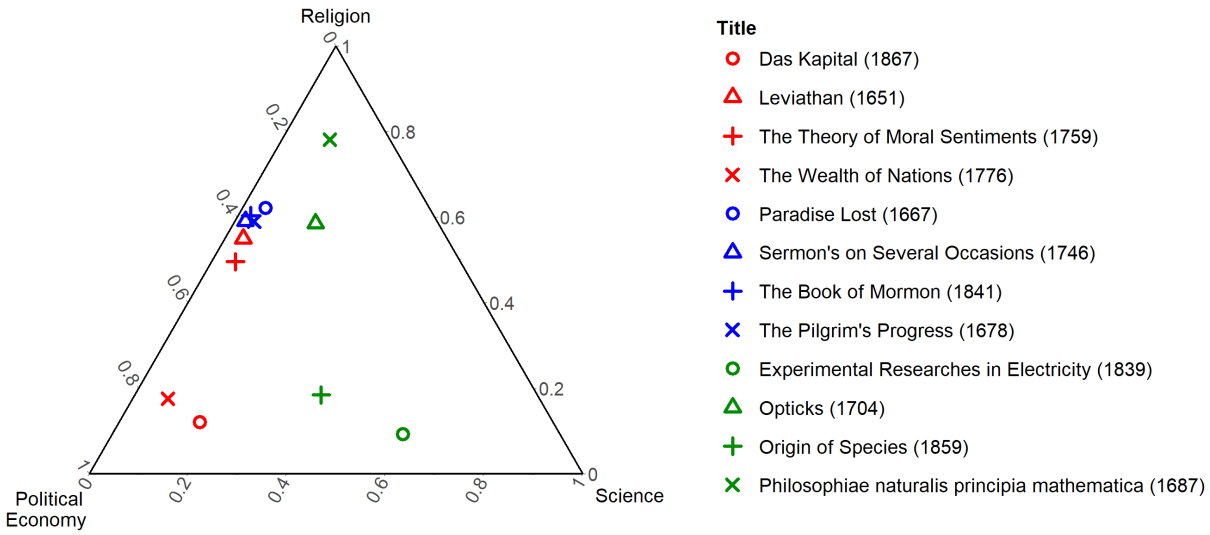
This finding finds qualitative support in Figure 3, which places 12 famous works in the simplex. We chose four well-known works of religion, political economy, and science from

Figure 2: Relationship between Categories over time, within volumes



various dates in the period in question. Perhaps surprisingly, the work with the highest religion score is Isaac Newton's *Philosophiæ Naturalis Principia Mathematica*. Yet, this is indicative of the language used in the 17th century in works that we would now clearly recognize as science. Newton was a deeply religious man. In fact, when writing his masterpiece, Newton claimed in private correspondance that "when I wrote my treatise about our system, I had an eye upon such principles as might work with considering men, for the belief of a deity, and nothing can rejoice me more than to find it useful for that purpose" (Janiak and Newton 2004, p. 94). Erikson (2021, p. 45–49) notes that moralistic tones were likewise invoked in early economic writings, which were written in the scholastic tradition and were more concerned with justice and sinfulness than in general welfare. Later books of science, such as Faraday's 1839 *Experimental Researches in Electricity* or Darwin's 1859 classic *On the Origin of Species* barely use the language of religion, although both invoke the language of political economy to some degree.

Figure 3: Selected Famous Volumes Categorized



4 Did the Language of Science become more Progress-Oriented prior to Industrialization?

4.1 Sentiment Analysis

The purpose of sentiment analysis is to measure the emotions and feelings of the writer, which are generally expressed in positive or negative tones. In our case, we are interested in looking beyond negative or positive tones, but rather with sentiment related to "progress." To do this, we employ dictionary techniques from the Natural Language Processing literature, which rely on lists of words (*dictionaries*) that comprise of synonyms for each category.

To create our dictionaries, we gather the list of synonyms for "progress" from the website www.thesaurus.com. Next, we manually reviewed each list to flag any words that may have double meanings or are related to religious or scientific language. For example, the word *positive* may be used to denote a progress-oriented outlook, but could also be used in a mathematics or chemistry sense. As a result, this word may be incorrectly picked up as belonging to the *progress* category when the book is in reality discussing mathematics. This identification procedure was done by each of the authors independently, and words were only removed if all authors unanimously voted on removing a word from the list. We then removed all words that according to the Oxford English Dictionary were not known prior to 1643 (the year of Newton's birth).[18] We do this to remove bias favoring words that would not have been in volumes written during the Enlightenment. The final list of words retained in each dictionary is shown in Table 2.

Table 2: Progress Dictionary Word List

progress	advance
improvement	rise
stride	amelioration
betterment	

Phrases were also removed from the word lists since our volumes are represented as a bag of words. This simplifying representation does not consider the order of words in the volume, but focuses on word counts only. Hence, phrases such as "step forward" are compared against the words "step" and "forward" separately. Therefore, phrases are removed from the word-

[18] Although this cutoff date is arbitrary, we view it as conservative, given that the OED only reports the first *known* usage in text. The three words removed by this criterion are development, headway, and boost. We report the results in which these three words are included in the "progress dictionary" in Figures B.3 and B.4. The remaining removed words are shown in Table B.1.

lists as they would not match with any single words. As a final step, each remaining word in the dictionary is converted to its respective root to match the volume cleaning procedure described in Section 2.2.

To gauge sentiment, we use a simple count of word occurrences for a given volume, normalized by the total number of words in each volume. Formally, $w_{i,\ell}$ is the count of word ℓ in dictionary list L in volume i, and W_i is the total number of words in volume i:

$$\text{Sentiment}_i = \frac{\sum_{\ell \in L} w_{i,\ell}}{W_i}. \tag{5}$$

In this case, the numerator represents the absolute score for each sentiment category, while the denominator acts as a deflator that controls for the size of the volume. Together, they measure the percent of words in each volume that are progress-oriented. This procedure is repeated for each volume individually.[19]

In Figures B.5 and B.6, we re-run the analysis using words related to progress and progression from the *Dictionarium Anglo-Britannicum* (Kersey 1708), a 1708 English-language dictionary.[20] This produces an alternative dictionary of words we know were used prior to industrialization. Results are similar to those reported here.

4.2 Volume Sentiment over Time

Each volume now has a sentiment score for words related to "progress".[21] Figure 4 reports the average progress score over time, as a percentile of all volumes in the corpus. Consistent with Mokyr (2016), volumes appear to have become more progress-oriented during the Enlightenment of the 17th century.

Yet, the hypothesis we are testing is not simply that language became more "progress-oriented" over time. We seek to uncover whether the *language of science* became more progress-oriented in the build-up to Britain's Industrial Revolution. To address this issue, we now plot each volume in a unit simplex in Figure 5, along with each volume's sentiment.[22]

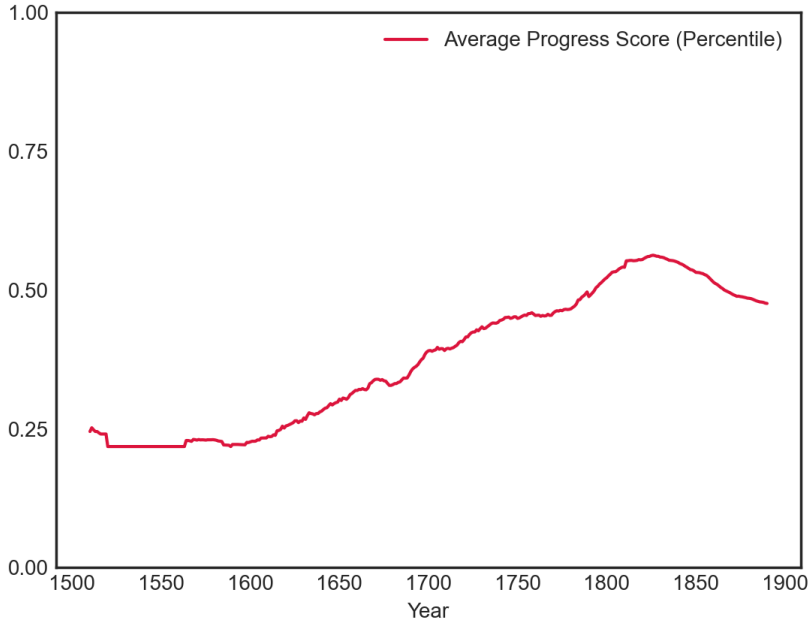
[19] One caveat is that the size of the volumes are uneven and may generate bias between them. Naturally, larger volumes will return more word counts than volumes with only a few words in them.

[20] We extracted words from the dictionary by first looking up the definitions of progress and progression. These definitions included the words "proceed(ing)," "forward," and "advance(ment)." We proceeded to look up these terms, which gave the additional term "further." These terms comprise the progress dictionary that we use in Figures B.5 and B.6.

[21] In a similar exercise reported in Figure B.7, we add together a volume's "progress" score and subtract its "regress" score, which is made up of synonyms of the word "regression". This yields an overall sentiment score, with a positive score representing an overall positive sentiment, a negative score representing an overall negative sentiment, and zero representing a neutral sentiment.

[22] For an easier interpretation, we plot each volume's percentile in the entire corpus in terms of sentiment, rather than the raw score.

Figure 4: Average Progress Score (percentile), 1500–1900



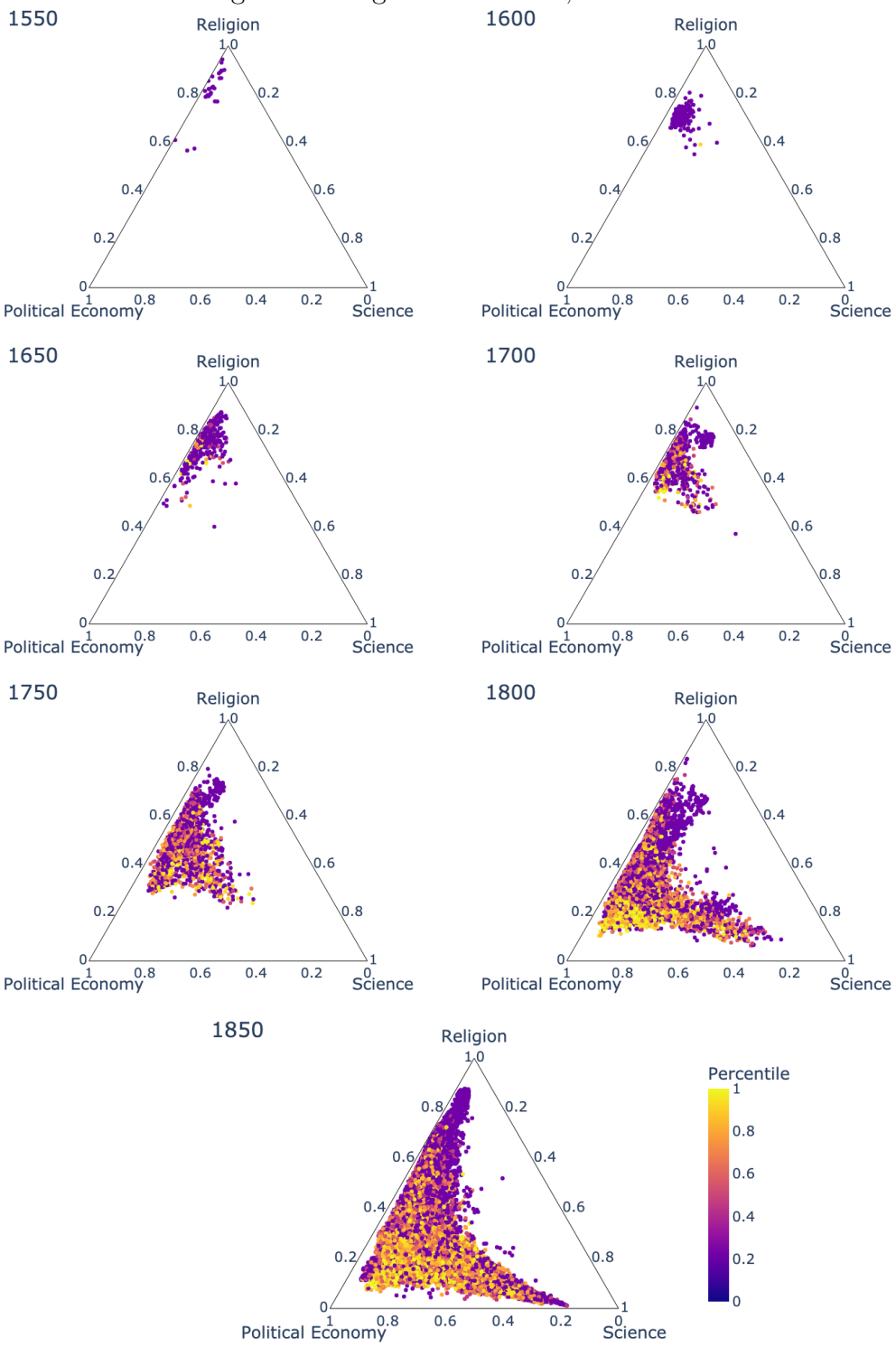
Note: average progress score taken using a 20-year moving average.

We show plots for every 50 years from 1550–1850 to observe how the sentiment towards science, religion, political economy, and their combinations evolved over time.

Two outcomes are of note in Figure 5. First, consistent with the topics plotted in Figure 1, volumes show similar "purification" in terms of the broad topic categories they fall into beginning in the first half of the 18th century. Even with the high number of volumes published in the eighteenth and nineteenth centuries, the religion-science axis is essentially devoid of volumes, whereas most volumes are published on the political economy-science or religion-political economy axis. Second and more importantly, it appears that volumes published along the political economy-science axis became increasingly progress-oriented over time, as represented by the increasing presence of lighter-colored dots. Additionally, volumes along the religion-political economy axis appear to be less progress-oriented, especially as one moves closer to the pure religion vertex.

These conclusions are supported by Figure B.9. Instead of showing individual volumes, Figure B.9 shows the average sentiment of all volumes within sub-triangles of the overall simplex. Sentiment is represented by the color of the sub-triangles, with darker shades indicating more progress-oriented sentiment. Additionally, the number of volumes in each sub-triangle is represented by the size of the white dot in the middle of each sub-triangle. As in Figure 5, areas along the political economy-science axis became increasingly progress-oriented over time, especially relative to areas close to the political economy vertex.

Figure 5: Progress Sentiment, 1550–1850



Note: Each dot represents a volume. For each year, all volumes +/- 10 years are included (i.e., for 1800 all volumes from 1790 to 1810 are included). The color of each dot represents the sentiment of that volume, with lighter colors representing more progressive sentiment. A grayscale version is available in Figure B.8.

4.3 Regression Analysis

Figure 5 presents visual evidence that works along the science-political economy nexus started to become more progress-oriented around 1700. But visual evidence can be deceiving. In this section, we confirm this visual evidence with quantitative support from regression analyses. These regression analyses are not meant to imply a causal relationship, as omitted variable biases and reverse causation may be present. Instead, this exercise is simply meant as an accounting exercise that clarifies the conditions under which progress-oriented language in correlated with the language of science, political economy, and religion over time.

We first place volumes (v) into 20 year bins based on date of publication (t).[23] We estimate the regression in equation (6), with standard errors clustered by the year (i.e., not the bin) of publication.[24]

$$\begin{aligned} Sentiment_{v,t} &= \alpha_1 + \alpha_2 Science_v + \alpha_3 PolitEcon_v + \alpha_4 Science_v \times PolitEcon_v \\ &+\alpha_5 Science_v \times Religion_v + \alpha_6 Religion_v \times PolitEcon_v + \lambda_t + \lambda_t \mathbf{A}_{v,t}\boldsymbol{\alpha} + \varepsilon_{v,t}, \end{aligned} \tag{6}$$

where $Sentiment_{v,t}$ represents the progress-oriented sentiment score in terms of percentile over the whole corpus for volume v published in bin t; *Science*, *Religion*, and *PolitEcon* represent the volume's category weights as derived in section 3.3; and λ_t are bin fixed effects.[25] We also include interactions between each category, to take into account that practically all volumes fall within a combination of categories. $\mathbf{A}_{v,t}$ is a vector of all of the variables and their interactions already included in equation (6). These latter interactions permit an analysis of how the coefficients change over time.

Full results are included in Appendix Table B.3. We plot the marginal effects of *Science* from equation (6) in panel A of Figure 6. These marginal effects are plotted over time for volumes of varying weights of science, religion, and political economy. The results suggest that volumes containing equal parts scientific and political economy topics became more progress-oriented as they became more scientific. This marginal effect is greater than anywhere else in the simplex throughout the 18th and 19th centuries. For volumes that contain equal parts scientific, religious, and political economy topics, contain only scientific topics, or contain equal parts religion and political economy, the marginal effect is near zero or

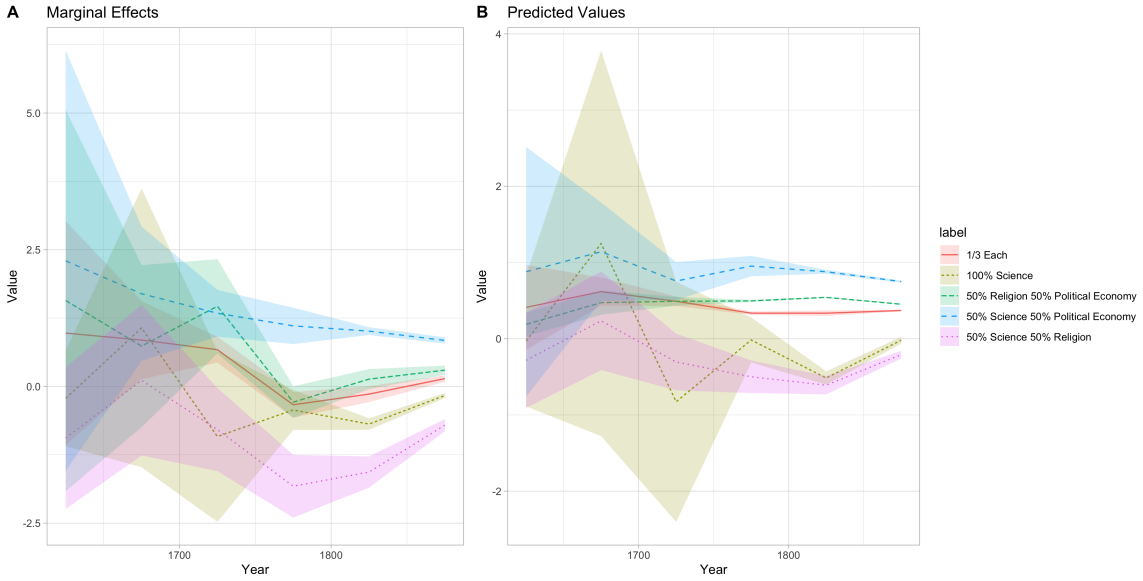
[23] We use 20 year (+/- 10 years) bins, for the years 1610, 1630, 1650, etc. up until 1890. We exclude the 16th century due to the low amount of volumes digitized in that period.

[24] Clustering standard errors by year addresses the possibility that errors are correlated within years. Robust standard errors yield very similar results.

[25] Recall that the three category weights for each volume add up to one. Hence, we exclude *Religion* as an independent variable.

even slightly negative throughout the period. The marginal effect of *Science* on volumes that contained equal measures scientific and religious topics is negative, although (as shown before) very few volumes are located at this nexus.

Figure 6: Marginal Effects and Predicted Values, Progress Sentiment Regressions



These results are further supported by panel B of Figure 6, which shows the predicted sentiment (in terms of percentile over the entire corpus) of volumes with varying weights of science, political economy, and religion. The predicted values tell a similar story. Volumes containing equal parts scientific and political economy topics show the highest level of progress-oriented sentiment beginning in the mid-18th century. In fact, most of the growth in predicted sentiment of these volumes occurred in the 18th century and remained stable after this point. Meanwhile, volumes at the religion-political economy nexus or the nexus of all three categories show slightly positive progress-oriented sentiment, and this is constant throughout the period. Volumes of pure science and those at the science-religion nexus have, on average, negative progress-oriented sentiment throughout most of the period in question.

In short, the language of science started to become more progress-oriented in the 18th century for those volumes located at the science-political economy nexus, and it maintained this progress orientation throughout the period under study. Meanwhile, volumes of "pure" science were largely neutral (or even negative) with respect to progress-oriented language. The timing of these findings aligns with that of Mokyr's "Industrial Enlightenment" hypothesis: as Britain commenced its industrialization in the mid-18th century, works of *applied* science—those at the nexus of science and political economy—became more progress-oriented.

4.4 Placebo Test: "Optimistic" Sentiment and the Language of Science

It is possible that our analysis thus far has picked up sentiment that is not necessarily more progress-oriented, but is more broadly optimistic in nature. These are distinct concepts, and they have significant implications for the theory we are testing. The idea espoused in Mokyr (2009, 2016) is that the key cultural change associated with the Enlightenment was in how our understanding of the world could be used to improve the lot of humankind. It was not that people spoke of science in "happier" terms. Yet, optimistic language is close enough to progress-oriented language that a change in the former could lead to spurious correlations regarding the latter.

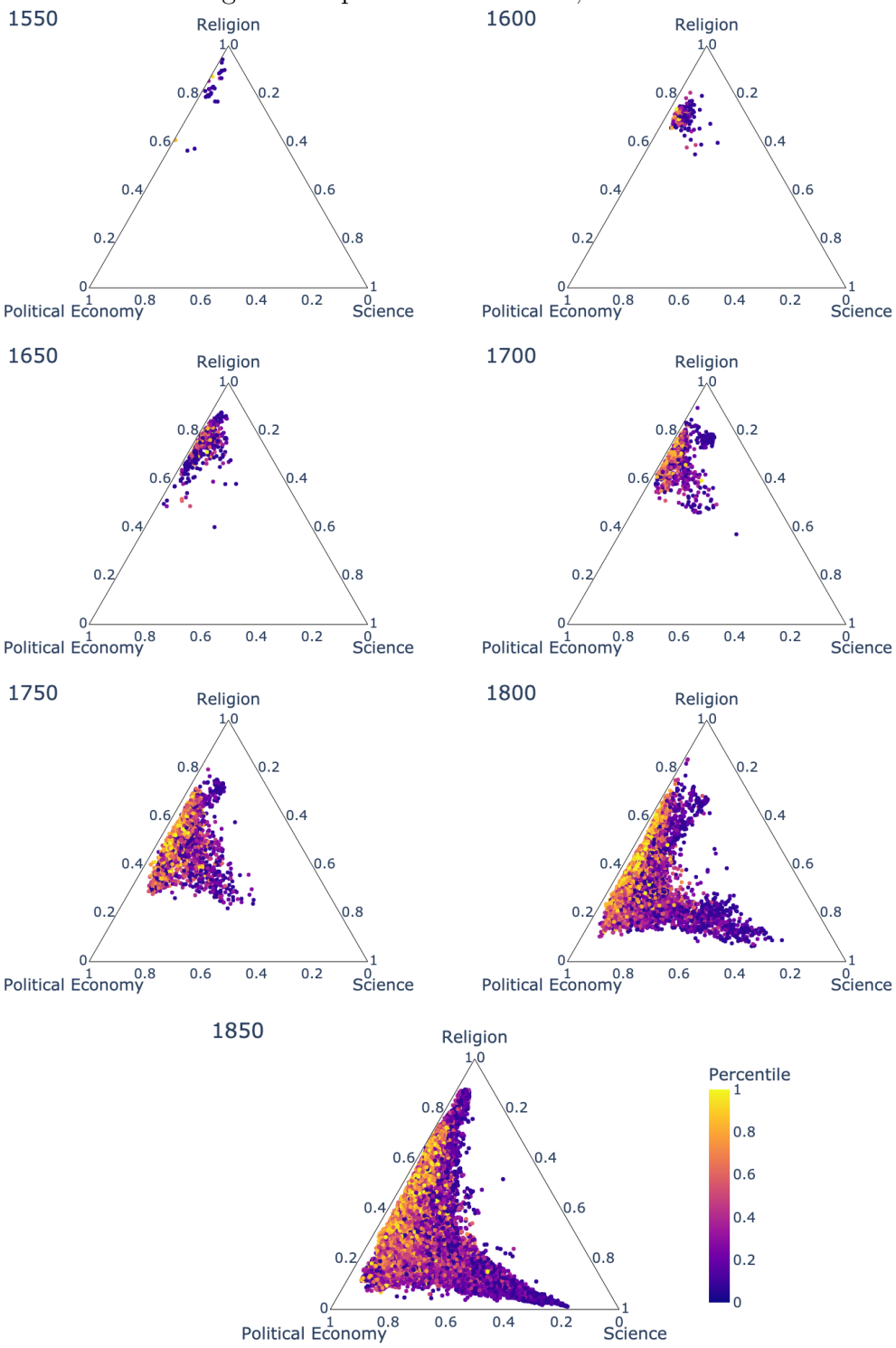
We address this issue by creating a "dictionary" of optimistic sentiment using the same methodology we used to create the progress dictionary. Using synonyms for optimistic and optimism from www.thesaurus.com yields the set of words listed in Table 3. These words are used to calculate sentiment in the same manner as we calculated progress-oriented sentiment (i.e., using equation (5)).

Table 3: Optimism Dictionary Word Lists

optimistic	optimism	anticipation
assurance	assured	calmness
cheer	cheerful	cheerfulness
cheering	confidence	confident
easiness	elation	encouraged
encouragement	enthusiasm	exhilaration
expectant	happiness	happy
hopeful	hopefulness	hoping
idealism	idealistic	merry
promising	rosy	sanguine
sanguineness	sureness	trust
trusting	utopian	

Each volume is assigned an optimism sentiment score. Figure 7 shows each volume's optimism sentiment in the unit simplex. There are two outcomes to note. First, volumes along the science-political economy nexus are much *less* optimistic than almost anywhere else on the simplex. This is especially true of volumes that approach the science nexus. Meanwhile, it appears that the most optimistic language is used in volumes at the religion-political economy nexus. Second, and more importantly, these results are nearly the mirror opposite of those found for progress-oriented sentiment in Figure 5. Those results indicated

Figure 7: Optimism Sentiment, 1550–1850



Note: Each dot represents a volume. For each year, all volumes +/- 10 years are included (i.e., for 1800 all volumes from 1790 to 1810 are included). The color of each dot represents the sentiment of that volume, with lighter colors representing more optimistic sentiment.

that the most progress-oriented language was employed at the science-political economy nexus, especially between 1700 and 1850.

These results suggest that the analysis is not merely picking up some broader change in optimistic language. Volumes at the science-political economy nexus became more progress-oriented in this period, but not more optimistic. These results greatly reduce the likelihood that we have picked up some spurious change in language that is correlated with, but not specific to, progress and the betterment of humankind.

5 The Language of Industrialization

The results presented thus far suggest that volumes at the intersection of science and political economy were more progress-oriented than other volumes, and this was the case since the early 18th century. This is consistent with the concept of the "Industrial Enlightenment" espoused by Mokyr (2009). Yet, the hypothesis put forth by Mokyr (2016) implies that volumes related to *industrial production* should have been particularly progress-oriented about the future. According to Mokyr, this is why cultural changes brought about by the Enlightenment ultimately resulted in the massive economic transformation associated with industrialization.

We test this hypothesis in this section by focusing on volumes associated with industrialization. In order to derive a list of words associated with industrialization, we digitized the detailed indexes of *Appleby's Illustrated Handbook of Machinery*, volumes 1–5 (Appleby 1877–1903). These handbooks, published between 1877 and 1903, provide schematics, mechanical details, measurements, prices, etc. for a wide range of industrial machines. They range from "prime movers" (volume 1), "hoisting machinery" (volume 2), "pumping machinery" (volume 3), "machine and hand tools" (volume 4), and "steam and electric plant" (volume 5).[26] These volumes cover all types of industrial machinery and their indexes are extremely detailed. We focus on the indexes of these books, rather than the entire content of the books, so that the degree of progress-oriented language in these books is immaterial to our results.

[26]The subtitle of the Prime Movers volume is "fixed, portable and machine engines, boilers, locomotives, steam launches, heated air, gas and water engines, turbines, and water wheels." The subtitle on the Hoisting Machinery volume is "winding engines, hydraulic, steam, and hand cranes, winches, and jacks." The subtitle of the Pumping Machinery volume is "pumping engines, centrifugal, steam and hand pumps." The subtitle of the Machine and Hand Tools volume is "workshop construction, with plans, sections and descriptions of engineering shops, and their equipments; machine tools for working metals, wood, etc., and their accessories, mechanics' tools, shafting, pulleys, belting &c., files, saws, and engineering stores." The subtitle of the Steam and Electric Plant volume is "employed in the construction and equipment of harbours, docks, canals, railways, &c., excavators, dredgers, conveyors and plant for handling coal and other materials, iron structures, bridges, and appliances for erection, quarrying and stone working machinery."

We derive a list of industrial words by transcribing the index of each of the five Appleby's volumes. As before, we then omit words that, according to the Oxford English Dictionary, were not in use prior to 1643.[27] Each word is weighted by the number of times it appears in the indexes. The top 10 industrial words are reported in Table 4, along with the number of times they appear in the Appleby's indexes, and the top 51 words are reported in Appendix Table B.2.[28]

Table 4: Top 10 Industrial Words

Word/Prefix	Count
crane	51
electr	42
weight	37
rope	27
cost	27
water	25
machin	24
coal	23
iron	22
steel	21

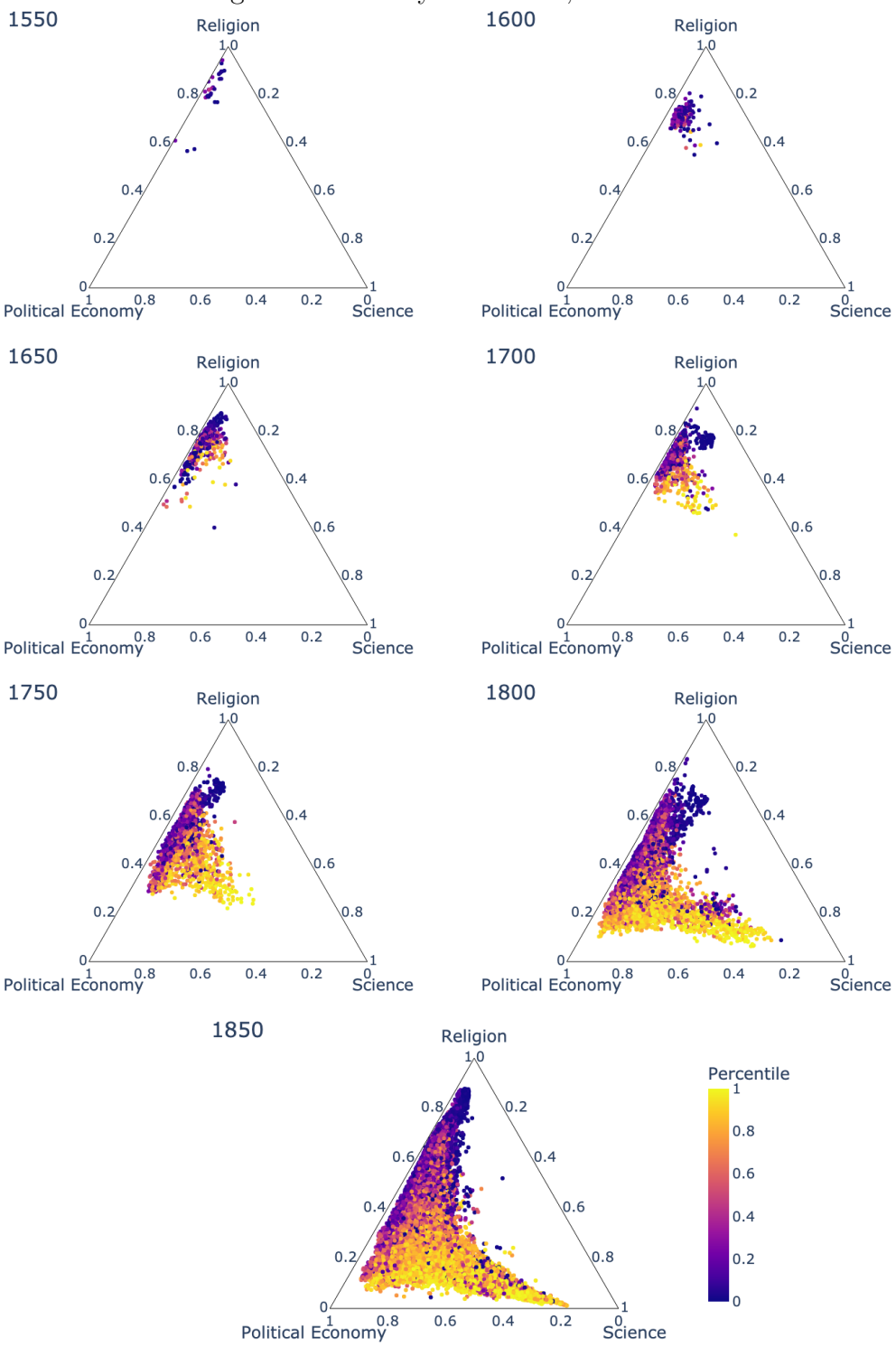
We proceed to derive an *industrial score* for each volume in the corpus. The industrial score is calculated by multiplying the count of each word in a volume by its corresponding weight, summed across all words with positive industrial weights. This sum is normalized by dividing by the total length of the volume. Each volume therefore has an industrial score between 0 and 1. The ranking of industrial scores for each volume within the unit simplex (i.e., with respect to the religion, science, and political economy categories) are reported in Figure 8.

Two results are immediately apparent from Figure 8. First, volumes using industrial terminology appear overwhelmingly on the science-political economy axis. This is particularly true beginning around 1750, when volumes first appear at this nexus. Second, volumes of "pure science"—those in the bottom right corner of the triangles—appear to be the most related to industrialization (i.e., the brightest yellow), while volumes of "pure religion" appear to be the least related to industry. This is true across all time periods for which there are volumes close to these axes.

[27] In Figure B.10, we report the industry sentiment scores using words in existence after 1643. Results are similar.

[28] We omitted words from the index list that were either innocuous or clearly had meanings unrelated to industrialization. These omitted terms are "note", "skip", "british", "foreign", "ga", "bear", "rel", and "men". The four coauthors independently went through the entire word list and omitted words that at least 3 of the 4 agreed should be omitted. Results are similar using a 2 out of 4 or 4 out of 4 threshold.

Figure 8: Industry Sentiment, 1550–1850



Note: Each dot represents a volume. For each year, all volumes +/- 10 years are included (i.e., for 1800 all volumes from 1790 to 1810 are included). The color of each dot represents the sentiment of that volume, with lighter colors representing more industrial sentiment. A grayscale version is available in Figure B.11.

Classifying volumes by their "industrial score" permits a test of the "Industrial Enlightenment" thesis laid out by Mokyr (2009, 2016). According to this thesis, views on applied, industrial pursuits using scientific principles became much more progress-oriented in the build-up to Britain's industrialization. In our framework, this indicates that volumes at the science-political economy nexus (i.e., those related to "Applied Enlightenment" principles) on topics related to industry should have been particularly progress-oriented in the period prior to and during Britain's industrialization.

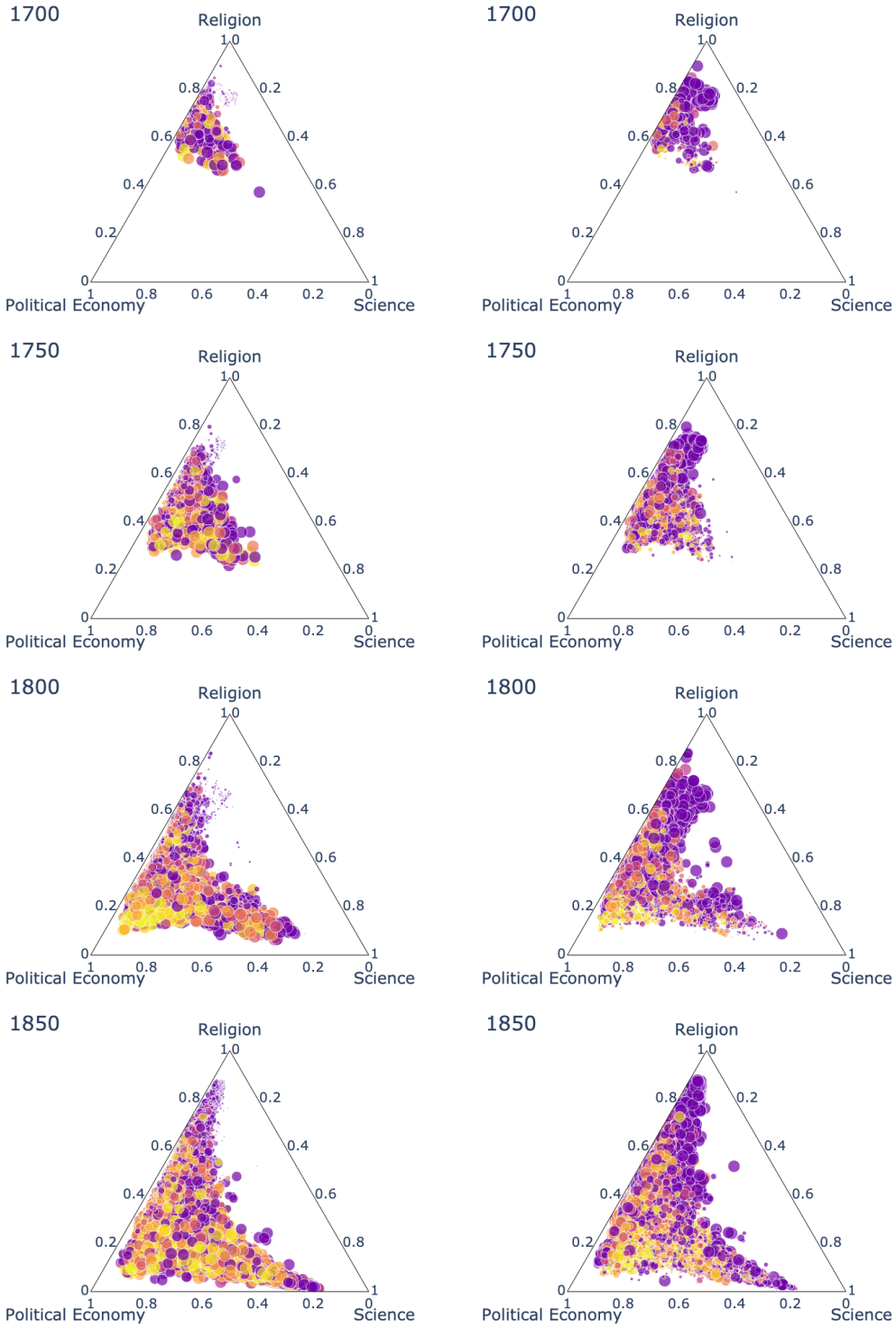
We first test the hypothesis with a visual representation of the relationship between progress-oriented sentiment and industry sentiment. Figure 9 re-plots progress sentiment (as in Figure 5) with larger circles representing higher industry sentiment in the left column and larger circles representing lower industry sentiment in the right column (i.e., the industry sentiment reported in Figure 8).[29] It is apparent from Figure 9 that beginning in the mid-18th century, and especially in the early 19th century, volumes at the science-political economy nexus were, on average, high in both industry and progress-oriented score. While there are some volumes with high industry scores that have low progress-oriented scores (i.e., big purple circles in the left column of the figure), these volumes tend to have higher religion scores and were published in earlier periods. These insights are supported by the right-most column, which highlights the location in the simplex of volumes with low industry scores. It is clear that these volumes are both located closer to the religion axis and have lower progress scores. The "yellow belt" of high progress volumes at the science-political economy nexus is much less visible in the right column, as most of these high progress volumes have a low industry score.

An econometric analysis further supports these findings. This requires an analysis of three dimensions: a volume's industrial score, its placement in the science-religion-political economy simplex, and its progress-oriented sentiment. To clarify these relationships, we present results from an OLS regression that includes interactions of all three dimensions along with time bin interactions. Specifically, we run a regression of the form:

$$\begin{aligned} Sentiment_{v,t} = \beta_1 + \beta_2 Science_v + \beta_3 PolitEcon_v + \beta_4 Industry_v \\ + \beta_5 Science_v \times PolitEcon_v + \beta_6 Science_v \times Religion_v + \beta_7 Religion_v \times PolitEcon_v \\ + \beta_8 Science_v \times Industry_v + \beta_9 PolitEcon_v \times Industry_v \\ + \beta_{10} Science_v \times Religion_v \times Industry_v + \beta_{11} Science_v \times PolitEcon_v \times Industry_v \\ + \beta_{12} Religion_v \times PolitEcon_v \times Industry_v + \lambda_t + \lambda_t \mathbf{B}_{v,t} \boldsymbol{\beta} + \varepsilon_{v,t}. \end{aligned} \tag{7}$$

[29] The circle sizes are increasing in industry-sentiment score. We only show 1700–1850 in this figure in order for the entire figure to fit on one page while being legible. Results for 1550–1650 are available upon request.

Figure 9: Progress Sentiment, with larger circles for higher (left) or lower (right) Industry Sentiment, 1700–1850

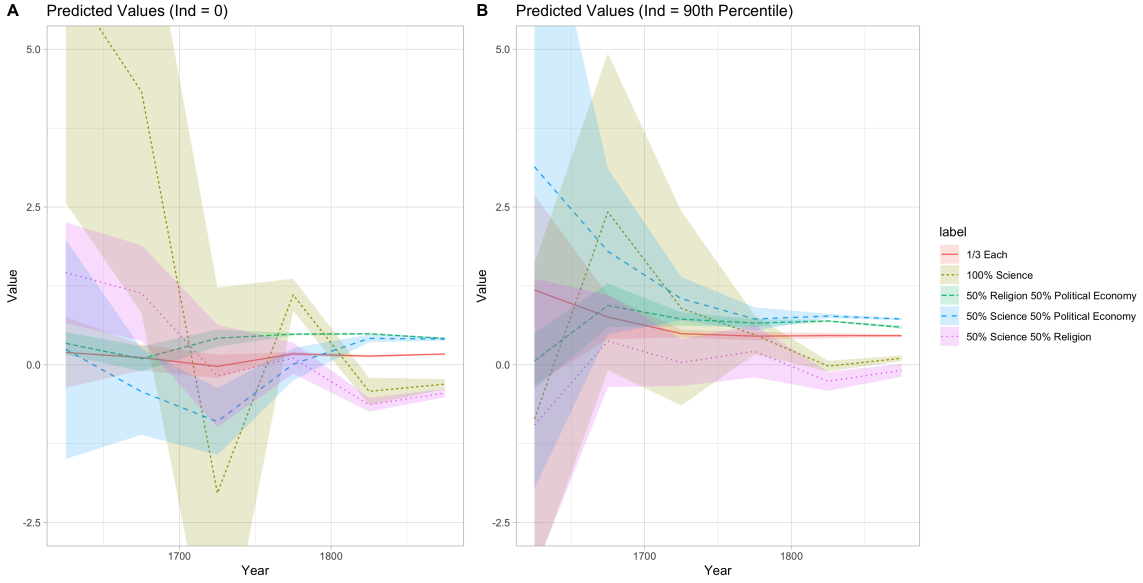


Note: Each dot represents a volume. For each year, all volumes +/- 10 years are included (i.e., for 1800 all volumes from 1790 to 1810 are included). The color of each dot represents the progress sentiment of that volume, with lighter colors representing greater progress sentiment. In the left column, larger circles entail greater industry sentiment, while in the right column, larger circles entail smaller industry sentiment.

where, as in our previous regression, $\mathbf{B}_{v,t}$ is a vector of all of the variables and their interactions already included in equation (7) and where the inclusion of $\lambda_t \mathbf{B}_{v,t}\boldsymbol{\beta}$ allows an analysis of how the coefficients change over time.

As in the previous regression analysis, this one is not meant to imply a causal relationship between industrial language and progress-oriented sentiment; it is simply meant as an accounting exercise that clarifies the conditions under which industrial and progress-oriented language are correlated. Appendix Table B.4 reports the regression results. Figure 10 reports the predicted progress-oriented sentiment scores for volumes at various locations in the unit simplex when the industrial score is 0 (panel A) and when it is at the 90th percentile (panel B).

Figure 10: Predicted Values of Progress Sentiment at 0 and 90 Industry Percentile



Several results follow directly from this analysis. First, volumes at the science-political economy nexus were more progress-oriented in the 18th century if they *also* had a high industry score. Volumes at this nexus were similarly progress-oriented by the mid-19th century at both high and low industry scores, but they were more progress-oriented earlier (mid-18th versus early-19th century) when their industry score was higher. In fact, the predicted progress sentiment is negative for volumes at this nexus with zero industry sentiment until the mid-18th century. The predicted progress sentiment is always positive for volumes at this nexus at the 90th industry sentiment percentile. Second, volumes at the religion-political economy nexus were more progress-oriented than those at the science-political economy nexus for volumes with a zero industrial score, but are less progress-oriented at the 90th percentile industrial score. This result is consistent with Figure 9, which reveals that most of the

progress-oriented, high-industry volumes are not located on the religion-political economy nexus, even if there were progress-oriented volumes at this nexus (see Figure 5).

In sum, these findings provide strong evidence in support of Mokyr's "Industrial Enlightenment" and "Culture of Growth" theses. The results indicate that volumes employing industrial language that were also at the science-political economy nexus became more progress-oriented in the mid-18th century. After this point, these volumes were, on average, the most progress-oriented volumes in the corpus.

6 Examples of Progress-Oriented Industrial Volumes

What exactly were the "progress-oriented" cultural values that emerged in industrial volumes in the 18th and 19th centuries? While the exercise thus far has been quantitative in nature, some insight can be gleaned from a qualitative account of the language used in industry-based volumes from the period.

To this end, we provide examples of the language used in a set of volumes that scored particularly high in both industry sentiment and progress-oriented sentiment. Such volumes can provide qualitative insight into the type of progress-oriented language that was used in the 18th and 19th centuries. Consider first *The Motion of Fluids, Natural and Artificial*, a 1735 book by Martin Clare (1735). This is a lengthy book on the science of fluid motion, including chapters on hydrostatic principles, gravity, cohesion, siphons, pumps, engines, and much more. It would certainly be recognized in the present as a book of science, although the language it used placed it at 43.7% science, 28.6% political economy, and 27.7% religion according to our algorithm. Its industrial score is in the 99th percentile of all volumes in our data. Like many books of the time, it had a very long subtitle. In this case, the subtitle is particularly telling (italics ours): "In particular that of Air and Water, In a familiar Manner, proposed and proved, by evident and conclusive Experiments with many useful Remarks. *Done with Plainness and Perspicuity, as that they may be understood by the Unlearned.*" This book was meant to be read by any literate person, not just the human-capital elite. The author, Martin Clare, clarifies in the preface that the book was meant so that humankind could benefit from its insights (p. vii, italics ours):

> The young Philosopher may be assisted hereby, in his first Searches after truth: Besides which Advantage, his Mind will be better prepared for receiving Lectures in Natural and Experimental Philosophy; which, with proper Encouragement, might easily be introduced into Societies, and *made of singular Use and Benefit to Mankind.*

This is precisely the type of progress-oriented language Mokyr (2016) suggests became more common on the eve of Britain's industrialization.

In the same year, Edward Saul (1735) published the second edition of his book *An Historical and Philosophical Account of the Barometer or Weather-Glass*. Like *The Motion of Fluids*, this book would be recognized in the present as a book of science, although it shared much language with religious works.[30] The central focus of the book is the science of barometers and how they can be used to predict weather patterns. Like Clare, Saul (1735) wrote for a general audience, not the human capital elite (p. 12):

> My design therefore in these papers, is not to write for the Entertainment of Philosophers, or of those Gentlemen, who by the Advantage of a learned Education, or of a Course of Experiments, have had better Opportunities of improving themselves in Speculations of this Nature: But for the Satisfaction of many of my inquisitive Countrymen; who having given themselves and their Parlours an Air of Philosophy, by the Purchase of a Barometer, may be willing to know the Meaning of it, and desirous of exerting now and then a Superiority of Understanding, by talking clearly and intelligibly upon it.

Much of the book focuses on the science of barometers and atmospheric pressure. Towards the end of this relatively short book, Clare argues for the usefulness of the study, suggesting that barometers can be used in the service of humanity by shedding light on a natural phenomenon (weather) which had mystified humans throughout history (p. 100, italics ours):

> It wou'd often be of great Consequence to form a probable Judgment some few Hours before hand, of the ensuing State of the Weather; whether it may be likely to continue, or liable to a sudden Alteration: But altho' in such an Enquiry (by the peculiar Situation and Uncertainty of our Climate) we can arrive at little more than bare Conjectures; yet even here, *a good Barometer will be of Service to us, in giving us some Light and Intimation.*

The two examples above were from the period just prior to Britain's industrialization. In the early 19th century, similar progress-oriented language was used in several tracts on an invention that promised increased prosperity: the railroad. Many of the volumes that scored highest on both our industrial score and progress score metrics concerned railways. These include volumes with titles like *Account of a patent improved metallic railway wheel with*

[30] The algorithm gives this volume category weights of 33.7% science, 24.8% political economy, and 41.5% religion. Its industrial score is in the 98th percentile of all volumes in the data.

wood-faced tyre ... (1840), *Railway rescue: a letter addressed to the directorates of Great Britain* (1848), *A practical treatise on rail-roads and interior communications in general* (1830), and *What will Parliament do with the railways* (1836). There was understandably much interest in how railways worked and what their practical utility was. Such concerns—and a progress-oriented response to these concerns—is exemplified in a short treatise by the famous engineer George Stephenson (1831), whose *A Report on the Practicability and Utility of the Limerick and Waterford Railway* described the technical issues and benefits associated with a proposed railroad connecting two southern Irish cities located approximately 130 km apart.[31] Stephenson argues that the railroad would benefit Ireland, which was part of the UK at the time, by employing underutilized capital and labor while connecting rural areas to markets. In a work largely devoted to laying out the costs and revenues associated with the railroad, Stephenson discusses how the railway will improve general well-being (p. 8–9): "[a] direct and obvious gain would then it appears be assured to Ireland, by the general introduction of Railways ... through the instrumentality of a cheap and expeditious means of transit, will be assured to Ireland, by allowing her people to reciprocate with England and with other nations, the products of industry; and by enabling her to take amongst nations that standing to which her natural capabilities, with her free government and institutions, entitle her."

Such language was common in discussion of railways. By this time, Britain had already industrialized, and the idea that industry could enable progress was well-entrenched, as indicated by the results reported in previous sections. It is therefore of little surprise that those who wrote about the early railways—arguably the most economically important innovation of the 19th century—would do so in such a progress-oriented manner.

7 Conclusion

The role of cultural attitudes—specifically, of Enlightenment ideals that had a progress-oriented view of scientific and industrial pursuits—in Britain's economic takeoff and industrialization has been emphasized by leading economic historians. Foremost amongst them is Joel Mokyr (2016), who states that the progress-oriented view of science promoted by great Enlightenment thinkers, such as Francis Bacon and Isaac Newton, among many others, was central to what would become the "Industrial Enlightenment," and ultimately Britain's Industrial Revolution.

[31] The algorithm gives this volume category weights of 13.3% science, 75.9% political economy, and 10.8% religion. It's industrial score is in the 99.5 percentile of all volumes in the data.

In this paper, we test these claims using quantitative data from 173,031 works printed in England in English between 1500 and 1900. A textual analysis resulted in three salient findings. First, there is little overlap in scientific and religious works in the period under study. This indicates that the "secularization" of science was entrenched from the beginning of the Enlightenment. Second, while scientific works did become more progress-oriented during the Enlightenment, this sentiment was mainly concentrated in the nexus of science and political economy. We interpret this to mean that it was the more pragmatic works of science—those that spoke to a broader political and economic audience, especially those literate artisans and craftsmen at the heart of Britain's industrialization—that contained the cultural values cited as important for Britain's economic rise. Third, while volumes at the science-political economy nexus were progress-oriented for the entire time period, this was especially true of volumes related to industrialization. Thus, we have unearthed some inaugural quantitative support for the idea that a cultural evolution in the attitudes towards the potential of science accounts in some part for the British Industrial Revolution and its economic takeoff.

The tools of textual analyses and the dataset we have constructed can be further utilized to study and test other hypotheses regarding European economic, political, and cultural history. For instance, there is a strand in economic history which postulates that European political fragmentation and the competition among its sovereigns—coupled with the Enlightenment belief in freedom of thought and expression—fostered and sustained a vibrant marketplace of ideas essential for economic development. In future work, we intend to apply textual analyses techniques on the corpus of work we have assembled in order to investigate if volumes written in English did indeed begin to reflect more freedom of expression and thought in the run-up to the Britain's economic takeoff. Likewise, similar techniques can be applied to the corpus of works in other languages. For instance, works by McCloskey (2006, 2010, 2016) suggest that similar results should be found in the corpus of works written in Dutch. Meanwhile, this was the period where the Spanish economy began to lag behind the leaders of Europe, while Spain was also the vanguard of the Counter-Reformation. Whether these economic and political phenomena are reflected in the cultural attitudes regarding progress and science remains a fruitful avenue for future work.

References

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2024. Historical Persistence. In *The Oxford Handbook of Historical Political Economy*, ed. Jeffery A. Jenkins and Jared Rubin. Oxford: Oxford University Press pp. 117–141.

Alesina, Alberto, Paola Giuliano and Nathan Nunn. 2013. "On the Origins of Gender Roles: Women and the Plough." *Quarterly Journal of Economics* 128(2):469–530.

Allen, Robert C. 2009. *The British Industrial Revolution in Global Perspective*. Cambridge: Cambridge University Press.

Appleby, Charles James. 1877–1903. *Appleby's Illustrated Handbook of Machinery*. Vol. 1–5 London: E. & F.N. Spon.

Blaydes, Lisa, Justin Grimmer and Alison McQueen. 2018. "Mirrors for Princes and Sultans: Advice on the Art of Governance in the Medieval Christian and Islamic Worlds." *Journal of Politics* 80(4):1150–1167.

Blei, David M., Andrew Ng and Michael Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3:993–1022.

Blei, David M. and John D. Lafferty. 2009. Topic Models. In *Mining: Classification, Clustering, and Applications*, ed. Ashok N. Srivastava and Mehran Sahami. Boca Raton, FL: Taylor and Francis pp. 71–94.

Buringh, Eltjo and Jan Luiten Van Zanden. 2009. "Charting the "Rise of the West": Manuscripts and Printed Books in Europe, a long-term Perspective from the Sixth through Eighteenth Centuries." *Journal of Economic History* 69(2):409–445.

Chen, M. Keith. 2013. "The Effect of Language on Economic Behavior: Evidence from Savings Rates, Health Behaviors, and Retirement Assets." *American Economic Review* 103(2):690–731.

Cirone, Alexandra and Thomas B. Pepinsky. 2022. "Historical Persistence." *Annual Review of Political Science* 25:241–259.

Clare, Martin. 1735. *The Motion of Fluids, Natural and Artificial*. London: Edward Symon.

Enke, Benjamin. 2019. "Kinship, Cooperation, and the Evolution of Moral Systems." *Quarterly Journal of Economics* 134(2):953–1019.

Erikson, Emily. 2021. *Trade and Nation: How Companies and Politics Reshaped Economic Thought*. New York: Columbia University Press.

Galor, Oded, Ömer Özak and Assaf Sarid. 2020. "Linguistic Traits and Human Capital Formation." *AEA Papers and Proceedings* 110:309–313.

Gentzkow, Matthew, Bryan Kelly and Matt Taddy. 2019. "Text as Data." *Journal of Economic Literature* 57(3):535–74.

Giorcelli, Michela, Nicola Lacetera and Astrid Marinoni. 2022. "How Does Scientific Progress affect Cultural Changes? A Digital Text Analysis." *Journal of Economic Growth* 27(3):415–452.

Giuliano, Paola and Nathan Nunn. 2021. "Understanding Cultural Persistence and Change." *Review of Economic Studies* 88(4):1541–1581.

Grajzla, Peter and Peter Murrell. 2019. "Toward Understanding 17th Century English Culture: A Structural Topic Model of Francis Bacon's Ideas." *Journal of Comparative Economics* 47(1):111–135.

Grimmer, Justin and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3):267–297.

Grosfeld, Irena, Alexander Rodnyansky and Ekaterina Zhuravskaya. 2013. "Persistent Antimarket Culture: A Legacy of the Pale of Settlement after the Holocaust." *American Economic Journal: Economic Policy* 5(3):189–226.

Hanson, Stephen, Michael McMahon and Andrea Prat. 2018. "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach." *Quarterly Journal of Economics* 133(2):801–870.

Heblich, Stephan, Stephen J. Redding and Hans-Joachim Voth. 2022. Slavery and the British Industrial Revolution. Technical report NBER Working Paper 30451.

Janiak, Andrew and Isaac Newton. 2004. *Correspondence with Richard Bentley [1692–3].* Cambridge Texts in the History of Philosophy Cambridge: Cambridge University Press p. 94–105.

Kelly, Morgan, Joel Mokyr and Cormac Ó Gráda. 2023. "The Mechanics of the Industrial Revolution." *Journal of Political Economy* 131(1):59–94.

Kersey, John. 1708. *Dictionarium Anglo-Britannicum: or, a general English Dictionary...* London: J. Wilde.

Koyama, Mark and Jared Rubin. 2022. *How the World Became Rich: The Historical Origins of Economic Growth.* Cambridge: Polity Press.

Lowes, Sara. 2024. Culture in Historical Political Economy. In *The Oxford Handbook of Historical Political Economy*, ed. Jeffery A. Jenkins and Jared Rubin. Oxford: Oxford University Press pp. 887–924.

McCloskey, Deirdre N. 2006. *The Bourgeois Virtues: Ethics for an Age of Commerce*. Chicago: University of Chicago Press.

McCloskey, Deirdre N. 2010. *Bourgeois Dignity: Why Economics Can't Explain the Modern World*. Chicago: University of Chicago Press.

McCloskey, Deirdre N. 2016. *Bourgeois Equality: How Ideas, not Capital or Institutions, Enriched the World*. Chicago: University of Chicago Press.

Michalopoulos, Stelios and Melanie Meng Xue. 2021. "Folklore." *Quarterly Journal of Economics* 136(4):1993–2046.

Mokyr, Joel. 2009. *The Enlightened Economy: An Economic History of Britain, 1700-1850*. New Haven, CT: Yale University Press.

Mokyr, Joel. 2016. *A Culture of Growth: The Origins of the Modern Economy*. Princeton, NJ: Princeton University Press.

Nunn, Nathan. 2014. "Historical Development." *Handbook of Economic Growth* 2:347–402.

Nunn, Nathan and Leonard Wantchekon. 2011. "The Slave Trade and the Origins of Mistrust in Africa." *American Economic Review* 101(7):3221–3252.

Saul, Edward. 1735. *An Historical and Philosophical Account of the Barometer or Weather-Glass*. London: A Bettesworth and C. Hitch.

Schulz, Jonathan F, Duman Bahrami-Rad, Jonathan P Beauchamp and Joseph Henrich. 2019. "The Church, Intensive Kinship, and Global Psychological Variation." *Science* 366(6466):eaau5141.

Spolaore, Enrico and Romain Wacziarg. 2013. "How Deep are the Roots of Economic Development?" *Journal of Economic Literature* 51(2):325–369.

Squicciarini, Mara P. and Nico Voigtländer. 2015. "Human Capital and Industrialization: Evidence from the Age of Enlightenment." *Quarterly Journal of Economics* 130(4):1825–1883.

Stephenson, George. 1831. *A Report on the Practicability and Utility of the Limerick and Waterford Railway*. London: Walton and Mitchell.

Voth, Hans-Joachim. 2021. Persistence: Myth and Mystery. In *The Handbook of Historical Economics*, ed. Alberto Bisin and Giovanni Federico. London: Elsevier pp. 243–267.

White, Lynn. 1978. *Medieval Religion and Technology: Collected Essays*. Berkeley: University of California Press.

Appendices for Online Publication

A Topics

1 - paint pictur artist music engrav painter colour
2 - town road church build built river stone
3 - franc pari french loui madam duke count
4 - church christian christ bishop holi paul doctrin
5 - love heart beauti soul sweet dark night
6 - india chines china nativ indian bengal govern
7 - fig water iron engin pressur steam electr
8 - acid solut heat carbon water sulphur iron
9 - exist refer period similar consist occur connect
10 - vol lond fol folio calf copi pari
11 - thou thi hath sir doth duke ladi
12 - god christ lord thi faith holi sin
13 - diseas blood patient treatment medic pain fever
14 - tho adj tlie lat tbe hut arc
15 - ofth sor differ juft sufficient suffer hath
16 - scotland quot edinburgh scottish highland burn dougla
17 - note latin verb greek languag comp text
18 - quod cum est sed quam qui aut
19 - parish esq counti street ditto rev park
20 - thoma john william richard robert mari henri
21 - henri bishop edward earl reign william archbishop
22 - poet poetri play poem genius johnson literari
23 - roman greek rome athen caesar greec senat
24 - court defend plaintiff estat properti bill contract
25 - esq jan oct dec nov feb juli
26 - morn river arriv distanc travel kill wild
27 - earl duke queen parliament majesti sir lord
28 - fig surfac develop genus structur upper geolog
29 - thou thi israel hath ver david jew
30 - fish black white bird colour tail brown
31 - armi enemi command march french attack captain
32 - quod cum vel regi anno est qui

33 - law lord show public evid opinion fact
34 - govern nation polit parliament constitut war parti
35 - miss ladi mother look room father woman
36 - compani court offic appoint counti act board
37 - indian river island coloni south america american
38 - excit punish display indulg circumst alarm digniti
39 - railway messr committe street liverpool manufactur cent
40 - king citi war princ armi england command
41 - line angl equal equat sin sun plane
42 - arab egypt greek egyptian persian sultan turk
43 - edit cloth crown illustr vol svo rev
44 - plant flower stem genus yellow calyx bot
45 - boil wine water salt sugar butter mix
46 - moral human exist scienc idea principl develop
47 - trade amount labour money price cent increas
48 - linn genus nat margin lin hab var
49 - par qui est che pour sur nous
50 - thi thou heaven sweet hath thine joy
51 - plant soil garden hors dri cultiv winter
52 - hath fame religion men shew virtu likewis
53 - school colleg societi educ church rev instruct
54 - fame fee fever cafe fet ufe feem
55 - letter dear ladi friend father wife repli
56 - hym doe hath bee sayd doth own
57 - kal kai tov yap rov occ masc
58 - ship island sea captain vessel coast sail
59 - emperor itali spain german germani franc duke
60 - cri ladi repli gentleman door boy captain

Topics eliminated during category selection: {5,9,22,26,35,46,50,55,60}

B Additional Figures and Tables

Table B.1: Removed "Progress" Word List

breakthrough	evolution	growth
increase	momentum	movement
pace	process	break
buildup	course	flowering
impetus	motion	progression
rate	way	anabasis
evolvement	step forward	

Table B.2: Top 51 Industrial Words

Word/Prefix	Count	Word/Prefix	Count
crane	51	fix	13
electr	42	variou	12
weight	37	gaug	12
rope	27	locomot	12
cost	27	float	11
water	25	price	11
machin	24	metal	11
coal	23	storag	11
iron	22	store	11
steel	21	strength	11
pile	21	grab	11
tool	19	pave	11
portabl	18	dock	10
work	18	pipe	10
steam	17	chain	10
block	17	differ	10
bridg	16	oil	10
hand	16	capac	9
materi	16	construct	9
speed	14	marbl	9
effici	14	road	9
light	14	wagon	9
stone	13	elev	9
system	13	navi	9
build	13	test	9
girder	13		

Table B.3: Dependent Variable: Progress Percentile

	Reference	1675	1725	1775	1825	1875
(Intercept)	2.026***	0.048	0.014	0.395***	0.211***	0.240***
	(0.232)	(0.074)	(0.059)	(0.078)	(0.056)	(0.056)
Science	−0.218	1.288	−0.698	−0.212	−0.470	0.052
	(0.446)	(1.374)	(0.913)	(0.483)	(0.450)	(0.447)
PolitEcon	−0.915*	1.145	0.913+	0.888+	1.315**	1.283**
	(0.454)	(0.716)	(0.482)	(0.459)	(0.455)	(0.454)
Science × Religion	−1.451*	−0.466	1.706	−1.339*	−0.304	0.364
	(0.582)	(1.463)	(1.109)	(0.652)	(0.615)	(0.587)
Science × PolitEcon	5.030	−3.780	−0.520	−1.956	−1.631	−3.017
	(4.784)	(5.493)	(5.034)	(4.820)	(4.787)	(4.785)
Religion × PolitEcon	1.832*	−1.113	−0.216	−1.444+	−1.153	−1.338+
	(0.726)	(1.234)	(0.791)	(0.760)	(0.728)	(0.728)
Num.Obs.	162.446					
R2	0.127					

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Volumes are placed into 50 year ((+/-) 25 year) bins. Columns represent interactions between bin fixed effects and the variables of interest (rows). Observations prior to 1600 are dropped. Standard errors are clustered by year of publication.

Figure B.1: Distribution of Volumes, with 20-year smooth

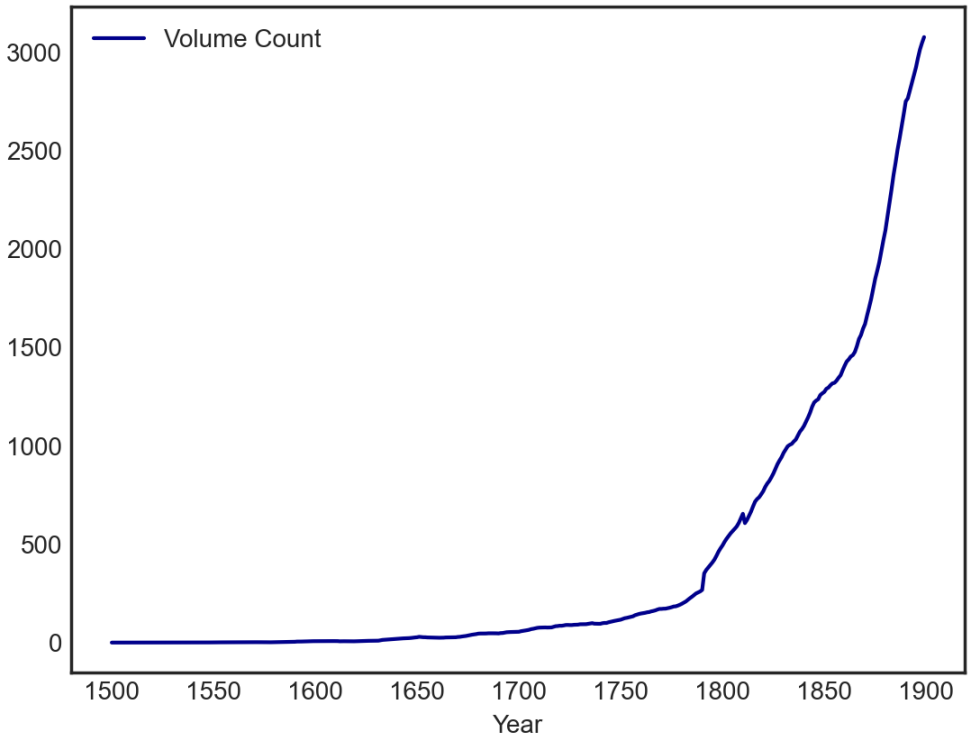


Figure B.2: Model Selection and Topic Optimization

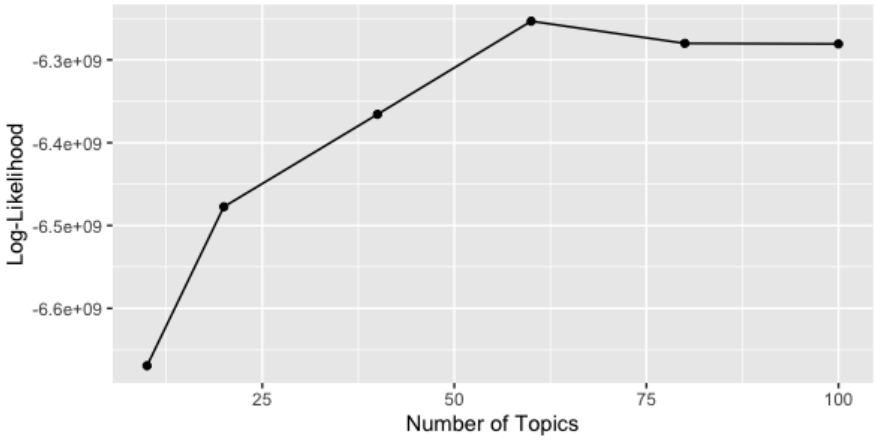
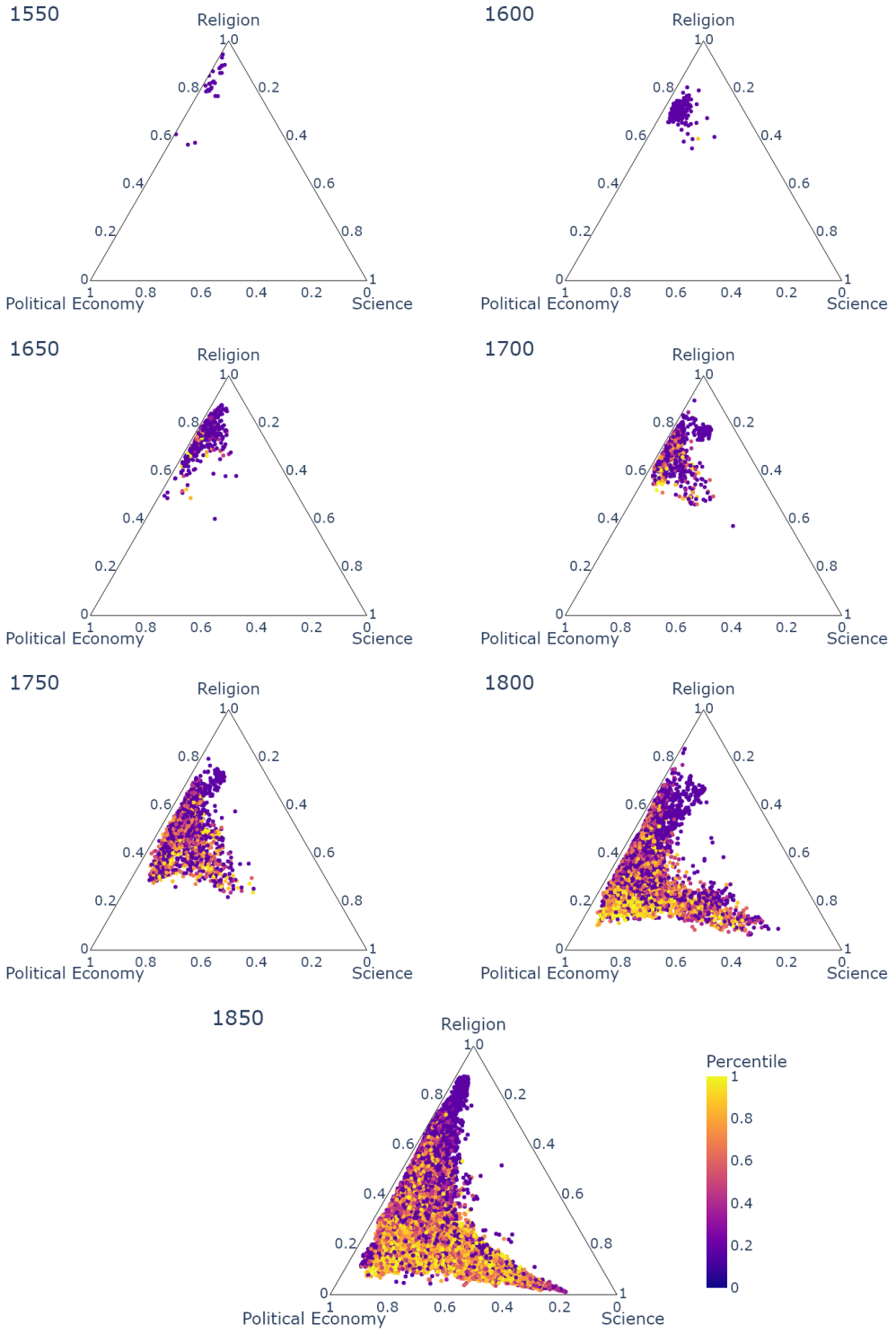


Table B.4: Dependent Variable: Progress Percentile

	Reference	1675	1725	1775	1825	1875
(Intercept)	1.585***	−0.139+	−0.218**	−0.245+	0.072	0.074
	(0.212)	(0.083)	(0.079)	(0.143)	(0.069)	(0.069)
Industry	0.175	0.309	0.761+	0.983*	−0.276	−0.149
	(0.383)	(0.457)	(0.421)	(0.481)	(0.389)	(0.386)
Science	5.936**	−1.630	−7.869**	−4.658*	−6.460***	−6.306***
	(1.811)	(2.530)	(2.472)	(1.820)	(1.816)	(1.812)
PolitEcon	−0.122	−2.061*	−0.040	0.205	0.195	0.282
	(0.556)	(0.883)	(0.697)	(0.565)	(0.558)	(0.557)
Science × Religion	−6.838**	2.685	10.397***	5.390*	4.938*	5.532*
	(2.475)	(3.249)	(3.041)	(2.576)	(2.478)	(2.478)
Science × PolitEcon	−11.452**	5.421	12.447*	9.453*	13.608**	13.253**
	(4.349)	(5.206)	(5.472)	(4.389)	(4.363)	(4.350)
Religion × PolitEcon	0.796	3.872**	1.620	1.661+	0.607	0.304
	(0.921)	(1.475)	(1.159)	(0.982)	(0.941)	(0.927)
Interaction with Industry						
Science	−7.949***	5.350	10.267**	6.087**	8.494***	8.379***
	(2.316)	(3.484)	(3.101)	(2.348)	(2.322)	(2.317)
PolitEcon	−1.557	6.530**	2.198	1.572	2.192	1.939
	(1.527)	(2.492)	(1.694)	(1.548)	(1.530)	(1.528)
Science × Religion	4.476	−4.587	−11.905*	−4.908	−3.512	−3.833
	(4.911)	(6.022)	(5.437)	(5.187)	(4.922)	(4.919)
Science × PolitEcon	31.169+	−27.971	−32.176+	−28.926	−31.548+	−31.489+
	(17.933)	(18.912)	(18.474)	(17.971)	(17.939)	(17.935)
Religion × PolitEcon	1.176	−9.303*	−4.861	−5.054+	−1.142	−1.256
	(2.956)	(4.705)	(3.293)	(3.044)	(2.967)	(2.969)
Num.Obs.	162.446					
R2	0.187					

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Volumes are placed into 50 year ((+/-) 25 year) bins. Columns represent interactions between bin fixed effects and the variables of interest (rows). Observations prior to 1600 are dropped. *Industry* represents the industry score by percentile over the whole corpus. Standard errors are clustered by year of publication.

Figure B.3: Progress Sentiment, 1550–1850, words first used after 1643 included



Note: Each dot represents a volume. For each year, all volumes +/- 10 years are included (i.e., for 1800 all volumes from 1790 to 1810 are included). The color of each dot represents the sentiment of that volume, with lighter colors representing more progressive sentiment.

Figure B.4: Marginal Effects and Predicted Values, Progress Sentiment Regressions, words first used after 1643 included

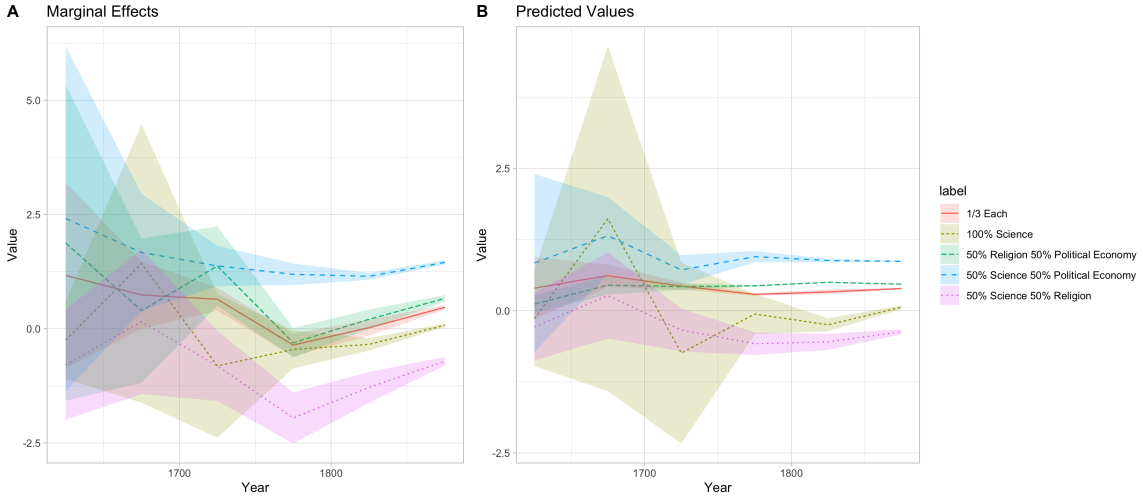
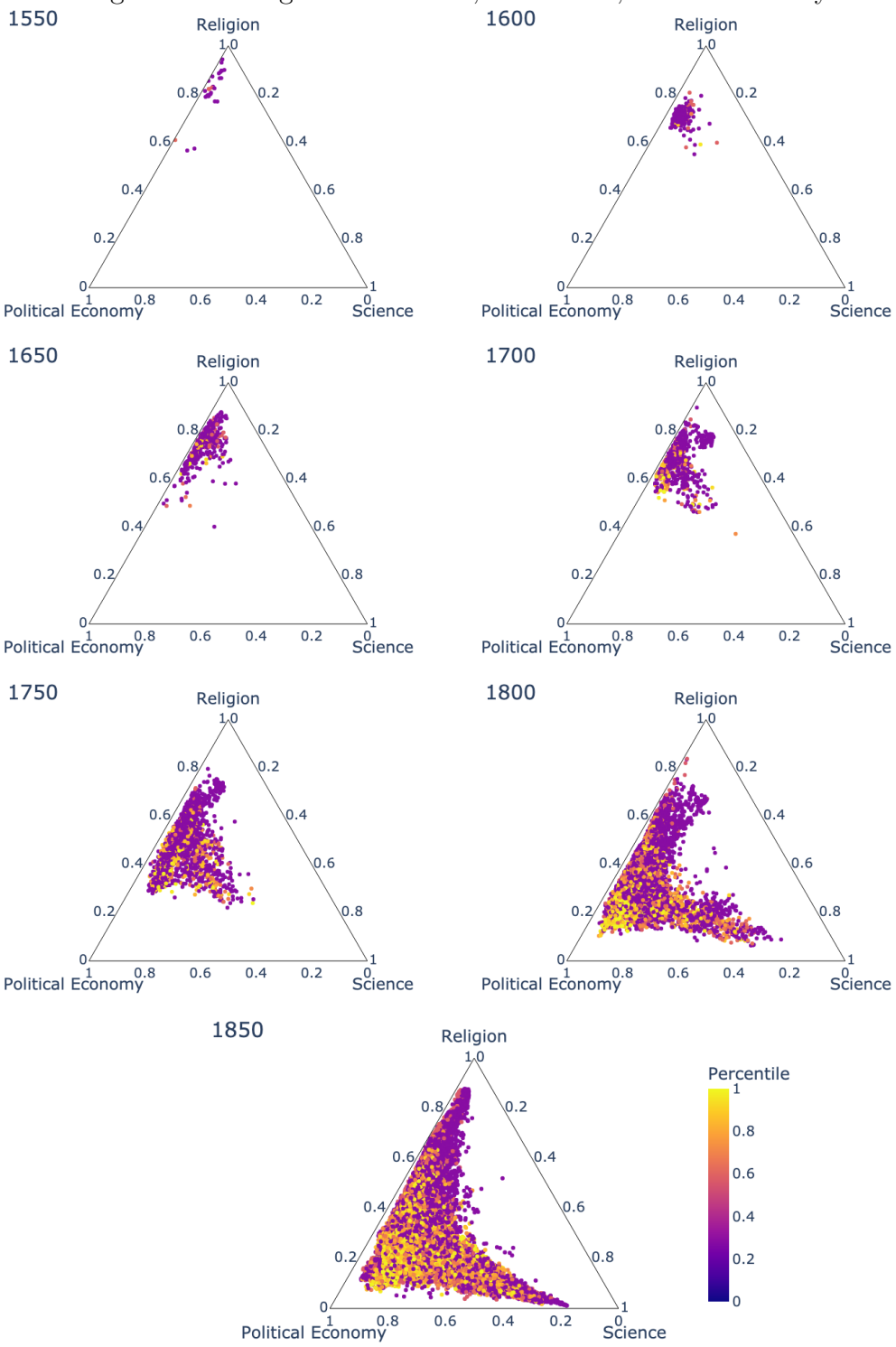


Figure B.5: Progress Sentiment, 1550–1850, 1708 Dictionary



Note: Each dot represents a volume. For each year, all volumes +/- 10 years are included (i.e., for 1800 all volumes from 1790 to 1810 are included). The color of each dot represents the sentiment of that volume, with lighter colors representing more progressive sentiment.

Figure B.6: Marginal Effects and Predicted Values, Progress Sentiment Regressions, 1708 Dictionary

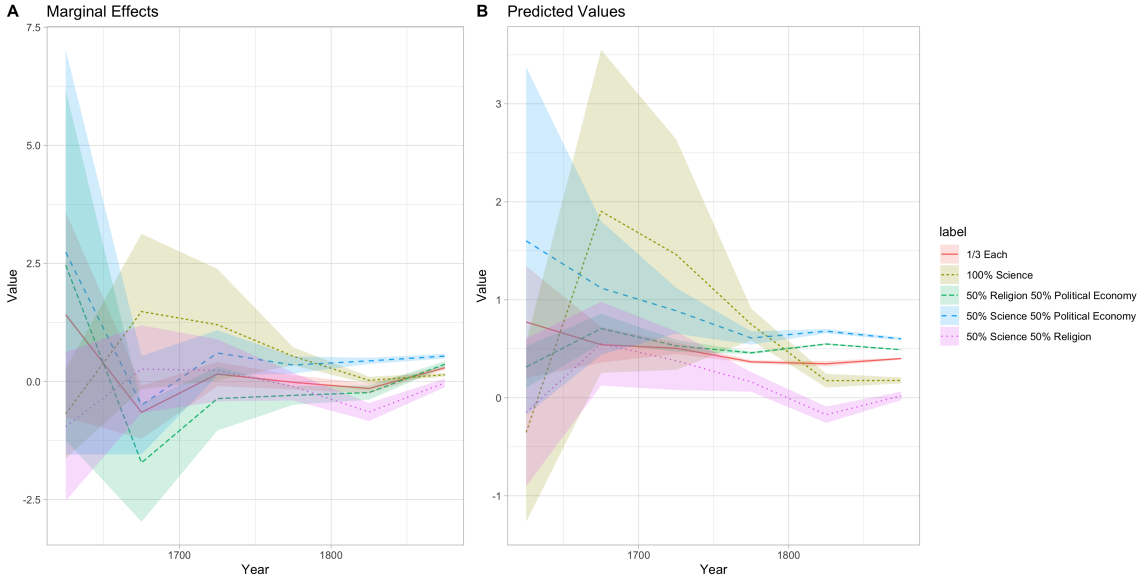
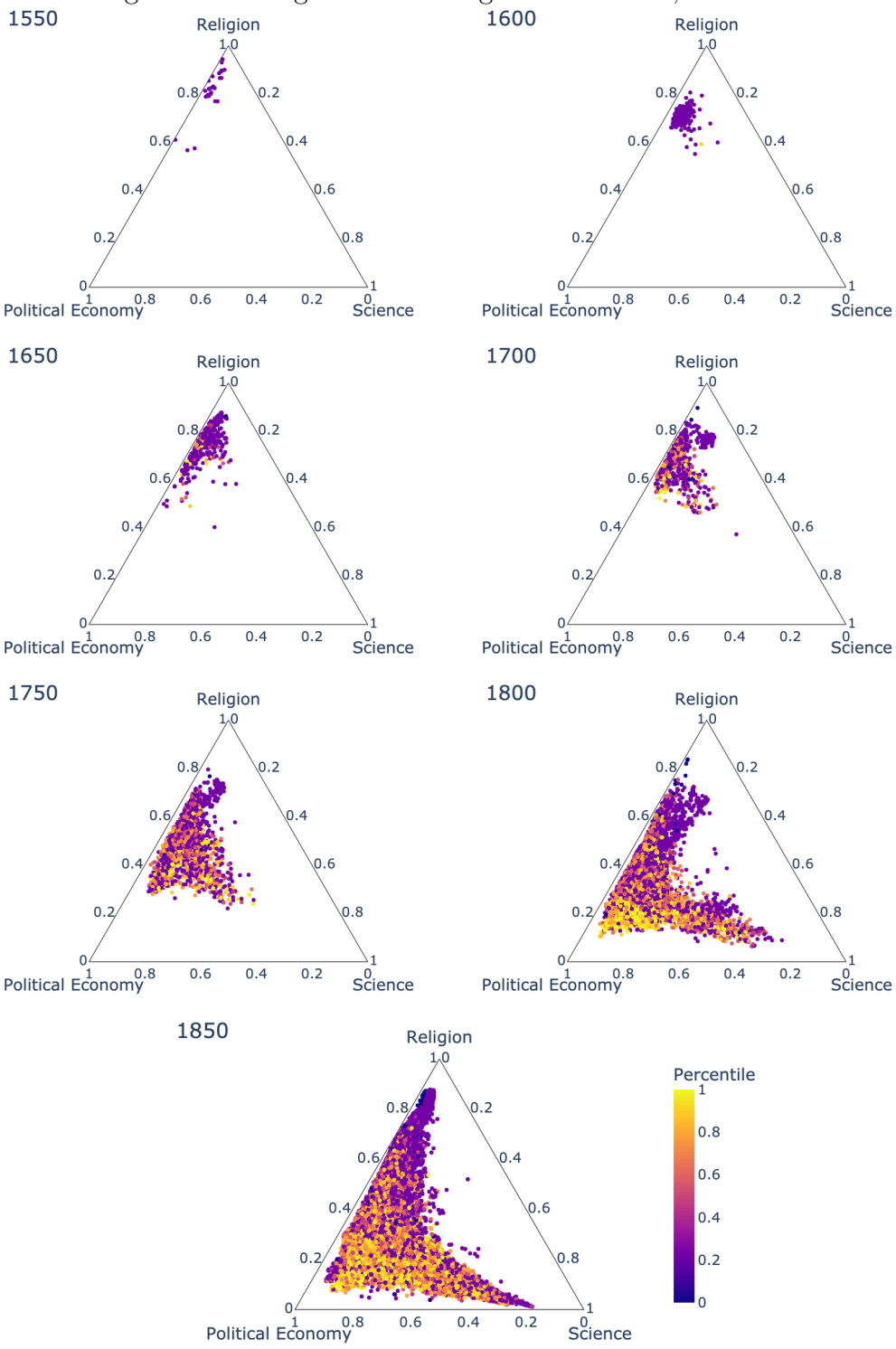
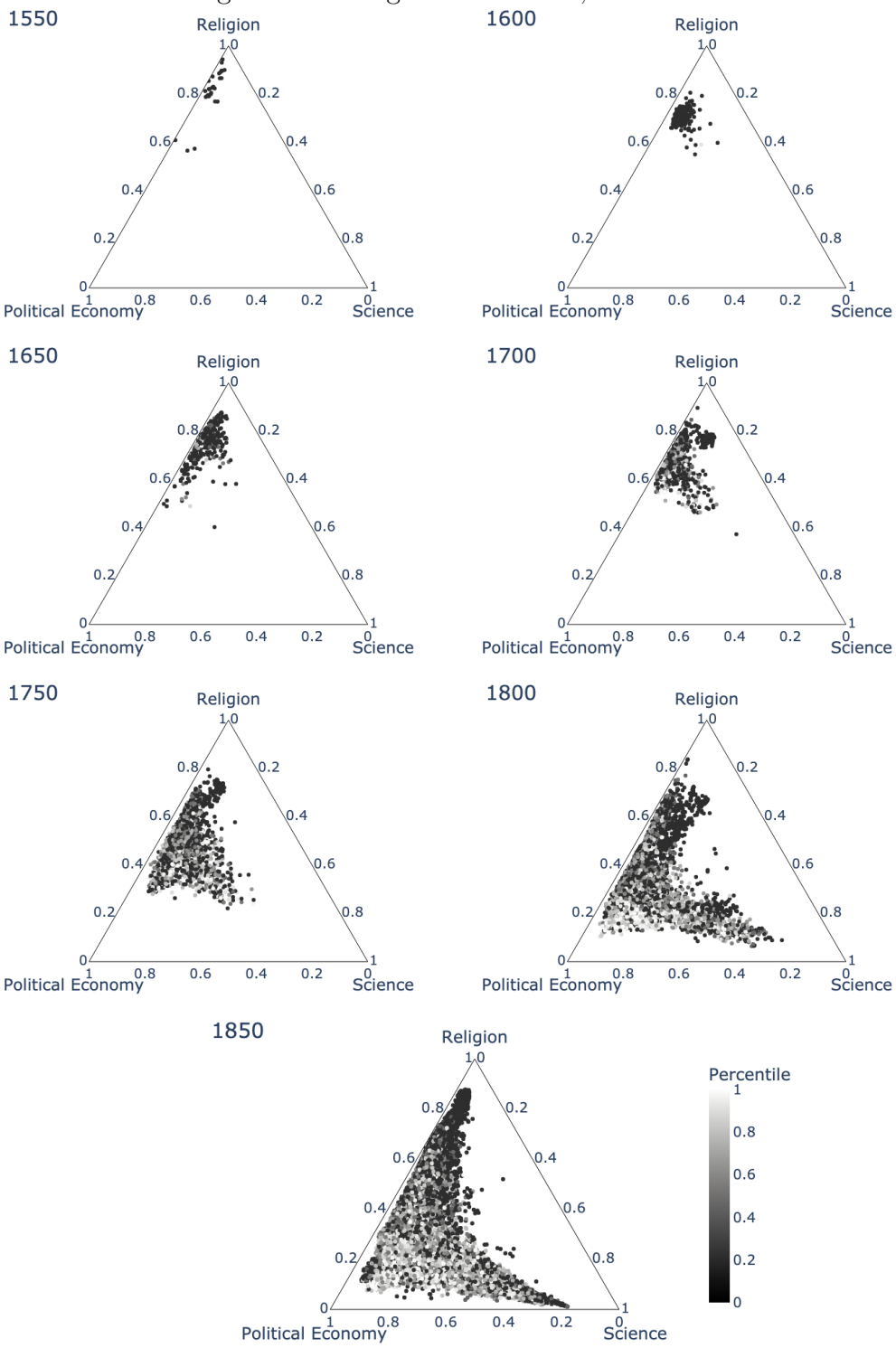


Figure B.7: Progress minus Regress Sentiment, 1550–1850



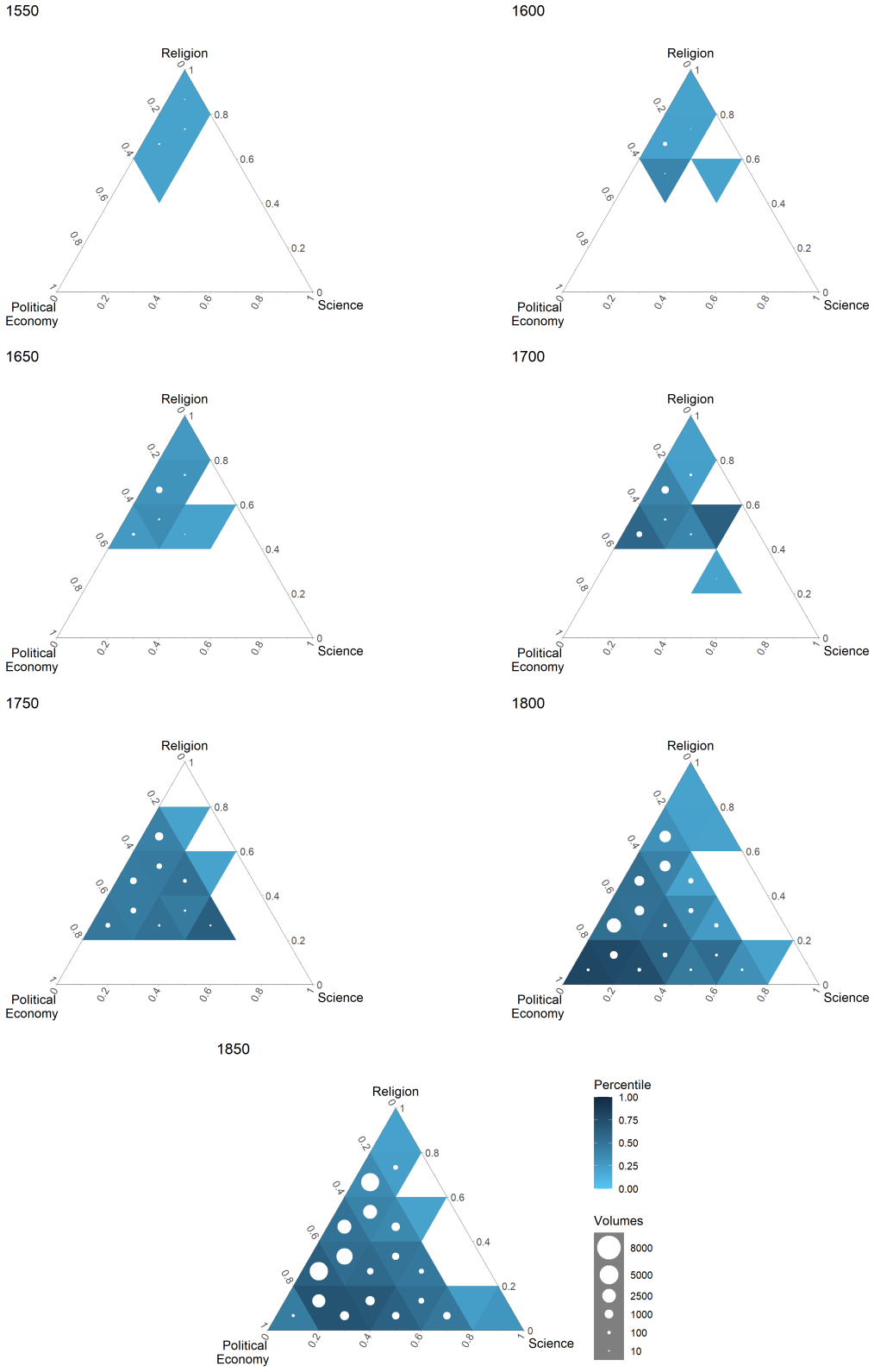
Note: Each dot represents a volume. For each year, all volumes +/- 10 years are included (i.e., for 1800 all volumes from 1790 to 1810 are included). The color of each dot represents the progress sentiment subtracted by the regress sentiment of that volume, with lighter colors representing greater sentiment.

Figure B.8: Progress Sentiment, 1550–1850



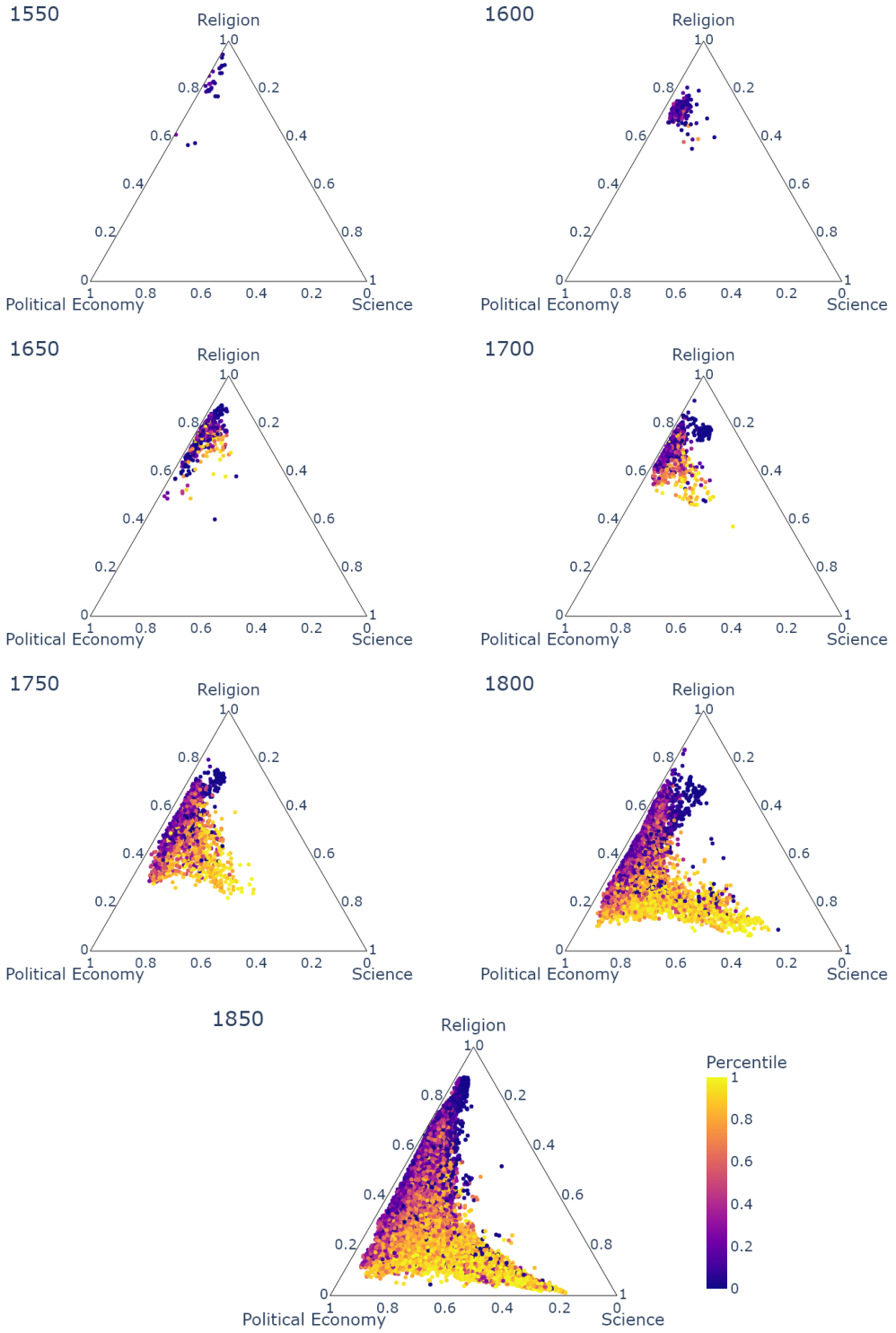
Note: Each dot represents a volume. For each year, all volumes +/- 10 years are included (i.e., for 1800 all volumes from 1790 to 1810 are included). The shade of each dot represents the sentiment of that volume, with lighter shades representing more progressive sentiment.

Figure B.9: Progress Sentiment in Triangles



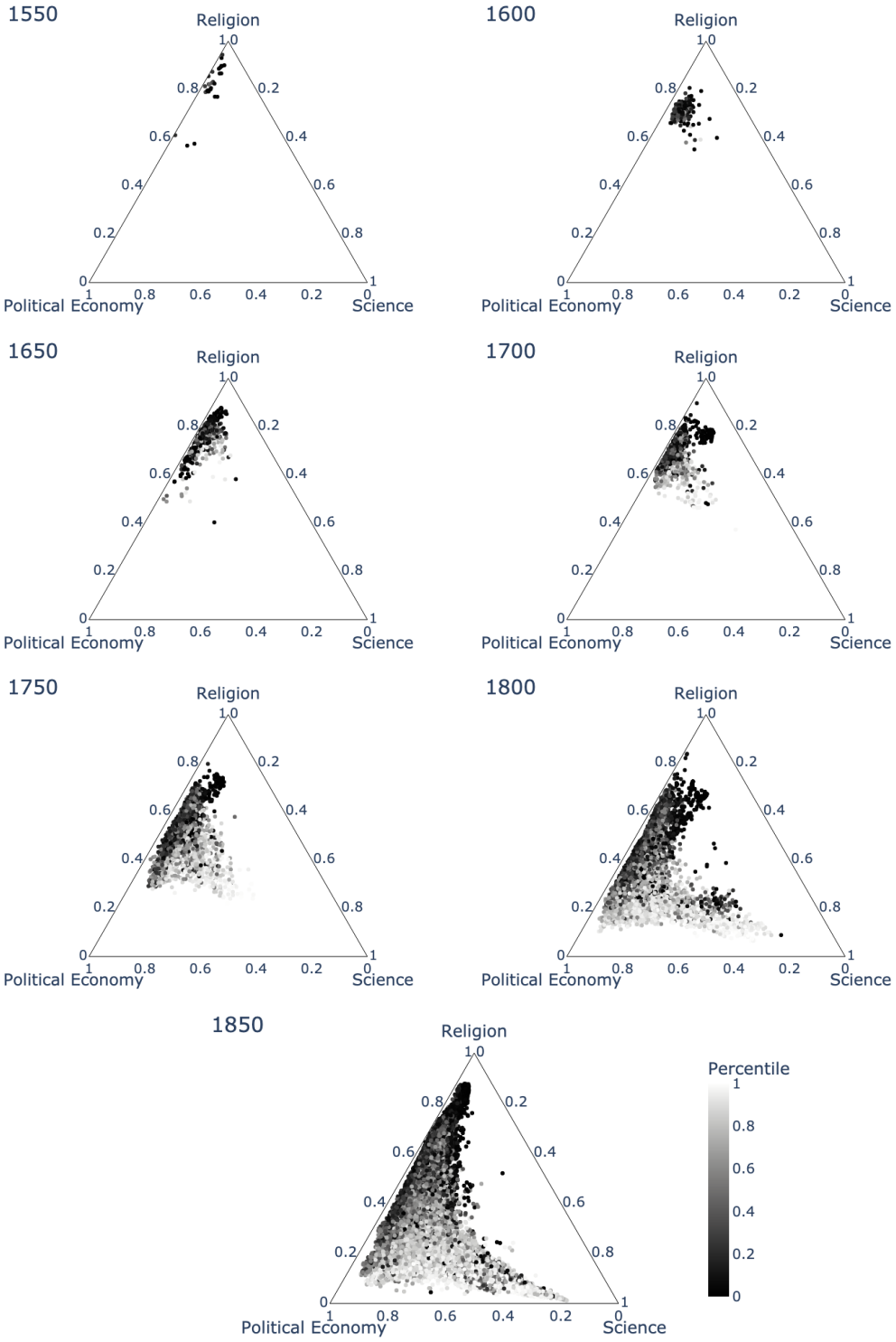
Note: Each sub-triangle shows the average sentiment by percentile of all volumes that fall within that sub-triangle. The size of the white dot in each sub-triangle represents the amount of volumes published within the sub-triangle. For each year, all volumes +/- 10 years are included (i.e., for 1800 all volumes from 1790 to 1810 are included).

Figure B.10: Industry Sentiment, 1550–1850, words first used after 1643 included



Note: Each dot represents a volume. For each year, all volumes +/- 10 years are included (i.e., for 1800 all volumes from 1790 to 1810 are included). The color of each dot represents the sentiment of that volume, with lighter colors representing more industrial sentiment.

Figure B.11: Industry Sentiment, 1550–1850



Note: Each dot represents a volume. For each year, all volumes +/- 10 years are included (i.e., for 1800 all volumes from 1790 to 1810 are included). The color of each dot represents the sentiment of that volume, with lighter colors representing more industrial sentiment.

C Checking for Bias in the Hathitrust Data

The Hathitrust Digital Library (HDL) data by construction only includes books that are currently fully available to be scanned. This means there may be two sources of bias in our data. The first is that these data do not include books that are no longer in existence. The second is that the libraries from which the HDL has digitized books may be biased towards the predilections of librarians or professors. While the HDL data are the best available in terms of fully digitized, machine-readable tracts, in order to properly analyze these data, it is necessary to identify the extent to which such omissions may positively or negatively bias the estimates we present in this paper.

We address these issues by comparing the HDL data to the data collected in the English short title catalogue (ESTC). The ESTC is a "comprehensive, international union catalogue listing early books, serials, newspapers and selected ephemera printed before 1801. It contains catalogue entries for items issued in Britain, Ireland, overseas territories under British colonial rule, and the United States ... The database contains over 480,000 entries, and represents the holdings of some 2,000 libraries world-wide." While the ESTC cannot shed light on books that are no longer in existence, it does help us understand the second source of bias, i.e., books that are selected to be in the libraries that have been digitized by HDL. The ESTC is much more comprehensive, containing the contents of an order of magnitude more libraries than the HDL data. The ESTC also includes metadata for each entry; importantly for our purposes, it provides a subject for each entry. However, the ESTC data could not be used in place of the HDL data, since the ESTC includes neither a full digitization of the entries nor publications after 1801.

Unsurprisingly, there are many more entries in the ESTC data than in the HDL data. There are two reasons for this. One is that the ESTC data are comprised of holdings from many more libraries. The second is that the ESTC data include serials, newspapers, and ephemera that are rarely included in the HDL dataset. Overall, there are 17,692 volumes in the HDL data printed up to 1800, whereas the ESTC data include 343,185 titles printed in England and written in English up to and including the year 1800.

To discern any potential bias in the HDL data, we first scraped the ESTC website of all titles printed in England and written in English up to 1800. To do this, we utilized web scraping techniques in *Python*; i.e., packages of *BeautifulSoup4*, *Selenium*, *Chrome Driver*, and *requests_html*. The code we employed has two successive functions: (1) interacting with the parameter entry interface on ESTC and (2) clicking through and saving all the responses to each query we ask in an automated fashion.

For (1), each iteration has several search parameters that act as constants; language code is 'eng,' country is 'enk,' and document type is 'alldocuments.' Some parameters change with each iteration; each iteration includes one year in the range of 1500 to 1800. After the algorithm enters the search parameters for the current year, it interacts with the 'Go' button on the page, waits for the page to refresh, and clicks on the hyperlinked number of results pop-ups. If there are non-zero documents with the requested criteria, the algorithm proceeds to go to function (2). With the existence of some content based on (2), the above code would be satisfactory to garner access to the entire corpus of texts with our desired parameters. However, ESTC limits the number of search results that one can access with one search at the industry standard of 1,001 (ESTC may report there are 3,432 results for a given year, but one can only access the first 1,001). This means that each inquiry is capped to produce a maximum of 1,001 results to be scraped by (2), which is a problematic feature, especially for the later years in the range. To circumvent this, we increase our number of iterations by shifting our unit of measure from the year to sub-year intervals. We achieve this through the logical parameters that ESTC allows users to add to their search inquiries. We produce a multi-level depth-based logic decision tree that acts as follows: if the number of works for a given year exceeds 1,001, then apply the first tree-level of logic with AND, and repeat with NOT. If the first logic term, with AND or NOT, exceeds 1,001, add an additional logic level and consider both AND and NOT. We incorporate 4 levels to this decision tree. For our first level, we start with using where it was published; AND London or NOT London. To further partition the search results, where necessary, at lower levels, we utilize some of the most common words in the English language such as "be," "an," "I," "in," "on," "by," and "more." With these additional logic considerations repeating for multiple iterations each year, we are able to access 99.8% of the works that satisfy our desired parameters (i.e. 1400–1800, eng, enk, alldocuments).

For each iteration, as described above, we feed its output (i.e., the hyperlink that produces the search results) to an algorithm that interacts with each search result on their own page and saves the relevant information into a useful data structure (i.e., appends each result to this data structure). We implement an algorithm that does this with *bs4*, *selenium* and *requests_html*. Once equipped with this search results link, the scraping code proceeds as follows; (a) clicks on the first search result, (b) stores the metrics of interest of this search result (title, publisher, author, year, meta-data, etc.), (c) finds the 'Next button,' (d) clicks on the 'Next Button' if it is live, and (e) repeats steps (a) through (d) until there is no

live next result button. Steps (a) through (e) occur for each iteration, each year with its subsequent run's logic.[32]

To compare the ESTC scraped data to the HTD data requires an additional algorithm. Since our use case is to compare book titles (strings) to one another with varying degrees of conventional, modern, spelling over the evolution of the English language, the standard computer scientific notion of exact equivalence is not satisfactory. There exists packages that address this exact use case. We chose to utilize the *fuzzywuzzy* library that handles these 'fuzzy' or weak string matches we desire. In particular, we utilize *rapidfuzz*, as its implementation is meant for larger data sets and produces greater efficiency in these cases.

Both the ESTC and HTD data sets contain a column of titles, IDs, authors, etc., with each row in each data set corresponding to a different book title. To conduct a string comparison, we perform standard pre-processing techniques, such as removing leading and lagging punctuation, removing capitalization, etc. We then take a row (book title) in the ESTC data and compare it against all titles in the HTD data. In each comparison, we calculate the match score (a metric from 0 to 100% of how similar the two strings are) and produce a best match by searching for the maximum of these numbers. We then create a new dataframe corresponding to the title in the ESTC data, its best match found in the HTD data, their respective IDs in each, and the Match Score.

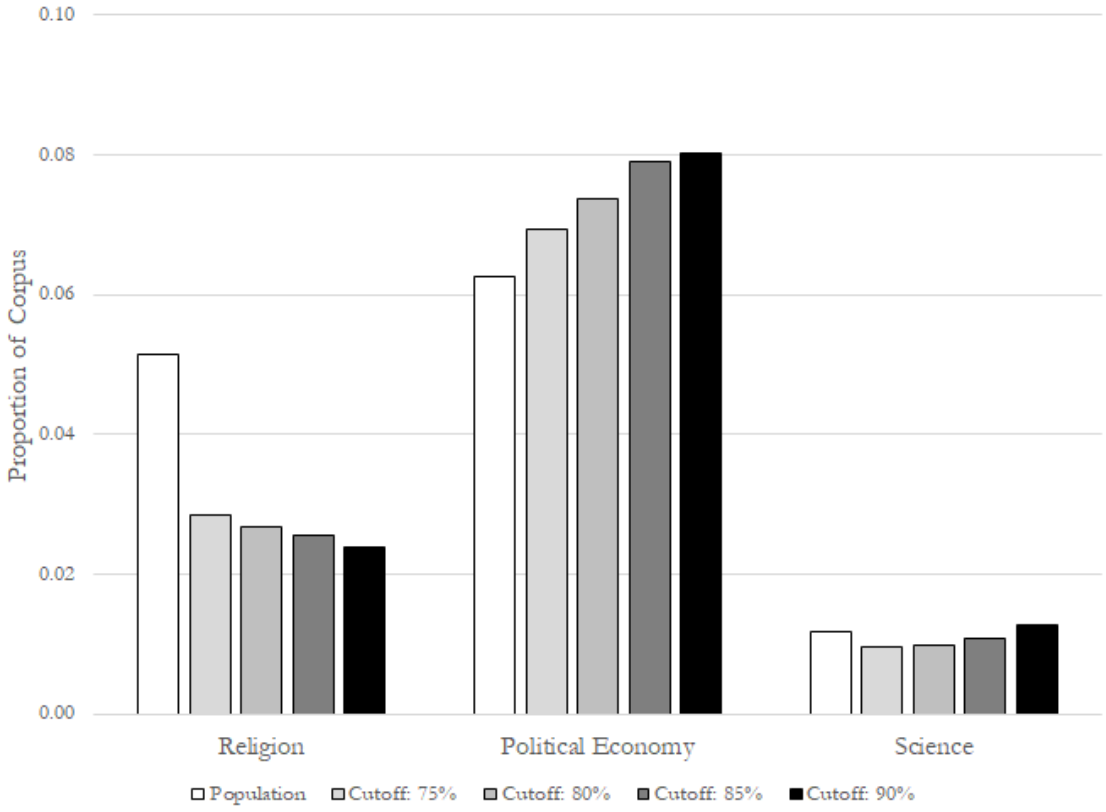
With this dataframe constructed, we can create an intersection and set difference by partitioning the data based on some real number threshold value. For our case we utilize 75%, 80%, 85%, and 90%. That is, a book is included in the intersection of the two data sets if their match is greater than or equal to the threshold value; otherwise they are in the set difference dataframe. We find that the ESTC data accounts for 29.61% of the HTD data at the 75% threshold, 20.4% of the HTD data at the 80% threshold, 13.73% of the HTD data at the 85% threshold, and 9.76% of the HTD data at the 90% threshold.

We proceed to use the metadata associated with the ESTC data that classifies each entry by subject. The ESTC metadata has several subject columns: 'subject' 'corporate subject', 'person as a subject', 'title as a subject', and 'conference as a subject'. We count the occurrence of words across the different subject groups. We then take the word counts and divide them by the total word count so they are comparable across groups. We do this for all the ESTC data as well as the matched HDL data at the various thresholds noted above (75%, 80%, 85%, and 90%).

[32]Since the search parameters do not produce perfect partitions, we ended up running additional scraping, i.e. rescraping a book multiple times. However, in 99.8% of works scraped, we already removed duplicate scrapings from our data set. This means that 99.8% of titles that were scraped are not biased and are a true 99.8% collection of the universe of documents in the parameter set as of October 2023.

Our primary interest with respect to these data is whether the HDL data is overly representative of religion, science, or political economy. To address this issue, we manually assigned all words that were in the subjects of at least 0.01% of the ESTC data as science, religion, political economy, or none of the above. For instance, the most common religious words are 'sermons' and 'church', the most common political economy words are 'government' and 'politics', and the most common science words are 'almanacs' and 'medicine'. We then summed up the total share for each group to derive the percentage of works in each data set that are religion, science, and political economy. Figure C.1 reports the results.

Figure C.1: ESTC vs. Hathitrust Subject Key Words by Category

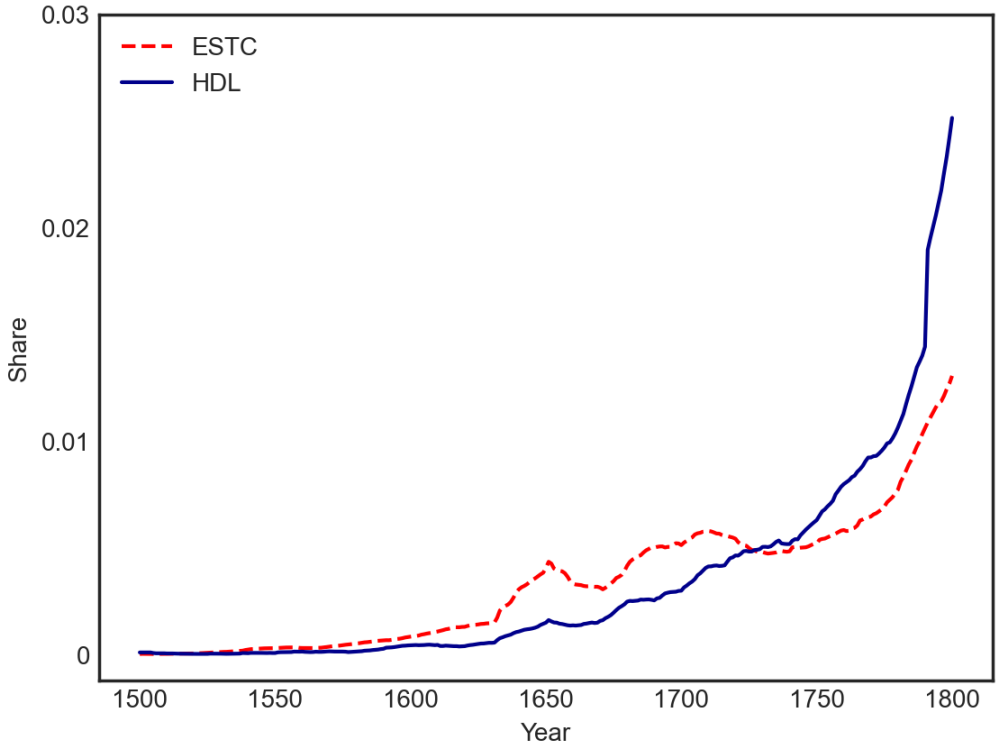


There are several features of this figure that are of interest for the present exercise. First, it appears that the HDL data is *not biased with respect to science*. This is reassuring, as works of science are the primary focus of the paper. However, the HDL data do appear to under-represent religion. At first glance, this may seem problematic. However, this is almost certainly due to the ESTC data containing "selected ephemera," including sermons. Works in which 'sermon' is in the topic comprise 18.9% of religious works in the ESTC data, whereas they only comprise 4.5% of the religious works in the HDL data (at the 90%

threshold). In other words, works labeled as sermons alone account for around 1/3 of the difference between the ESTC data and HDL data.

Likewise, the HDL data seems to show a slight bias in favor of political economy works, especially when higher matching thresholds are employed (at the 75% threshold the difference is small: 6.25% of the ESTC data are political economy, whereas 6.92% of the HDL data are political economy). This difference is mostly driven by topics that include the words 'politics', 'government', 'revolution', and 'political'. This result thus also appears to be a result of the ESTC including ephemera such as sermons. Although politics and government were certainly the subject of ephemera, they were also clearly the subject of books, as this topic is by far the most common in the HDL data. This is less true of sermons and other ephemera, which would show up in the ESTC data but not the HDL data. In any case, the difference between the ESTC and HDL data are small with respect to political economy, showing at most a small bias in favor of political economy volumes in the HDL data.

Figure C.2: ESTC vs. Hathitrust Data by Year of Publication (PDF)



We further test whether the time distribution of publications is similar between the ESTC and HDL data sets. Figure C.2 reports the distribution by year for each data set. It is readily apparent that the ESTC data set has relatively more works from the 17th century while the HDL data has relatively more works from the 18th century over the relevant time

span (1500–1800). This is not surprising given that books must be machine-readable to enter into the ESTC data set. Rare books or books that are too damaged to be digitized can end up in the ESTC data but not the HDL data. This biases the HDL data to contain more popular books from the 16th and 17th centuries—books that had large print runs were more likely to be in good enough condition to digitize. Since it is precisely these books that should have had the greatest impact in disseminating beliefs ("progress-oriented" or not), we do not believe this bias affects the implications of our analysis.

Finally, the exercise described above cannot account for books that are no longer in existence. This is a potential source of bias, particularly if those books were widely read and contributed to the type of language people used at the time. We believe this to be unlikely, however. The heroic efforts by those in the digital humanities to preserve and digitize the known corpus of writing from this period means that the books excluded from the ESTC data base are mostly those that are truly lost forever. While it is certainly possible that some of these works had influence in specific times and places, the very fact that they are lost forever—and never reprinted—suggests that these were works of relatively low value or impact. However, we are happy to qualify all results in this paper as "based on volumes of enough importance to have at least one copy available in a 21st century library."