
12-11-2023

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

Ali Almelhem

The World Bank, ali.almelhem@colorado.edu

Murat Iyigun

University of Colorado, Boulder, iyigun@colorado.edu


Austin Kennedy

University of Colorado, Boulder, austin.kennedy@colorado.edu

Jared Rubin

Chapman University, jrubin@chapman.edu

Follow this and additional works at: https://digitalcommons.chapman.edu/esi_working_papers

 Part of the [Econometrics Commons](#), [Economic Theory Commons](#), and the [Other Economics Commons](#)

Recommended Citation

Almelhem, A., Iyigun, M., Kennedy, A., & Rubin, J. (2023). Enlightenment ideals and belief in progress in the run-up to the Industrial Revolution: A textual analysis. *ESI Working Paper 23-13*. https://digitalcommons.chapman.edu/esi_working_papers/393/

This Article is brought to you for free and open access by the Economic Science Institute at Chapman University Digital Commons. It has been accepted for inclusion in ESI Working Papers by an authorized administrator of Chapman University Digital Commons. For more information, please contact laughtin@chapman.edu.

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

Comments

ESI Working Paper 23-13

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[¶]

May 7, 2024

Abstract

We trace the evolution of the language of science, religion, and political economy in the centuries leading to the British Industrial Revolution. Using textual analysis of 173,031 works printed in England between 1500 and 1900, we test whether British culture manifested a 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 late-17th century. Second, volumes using language at the nexus of science and political economy became more progress-oriented during the Enlightenment. Third, volumes using industrial language—especially those at the science-political economy nexus—were more progress-oriented beginning in the 17th century.

Keywords: language, religion, science, political economy, progressiveness, Enlightenment, Industrial Revolution

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

*We thank Joel Mokyr and Naci Mocan for extensive comments on earlier drafts. Aaron Berman provided excellent research assistance. We also thank Ran Abramitzky, Sascha Becker, Gabriele Cristelli, Vicky Fouka, Carola Frydman, Oded Galor, Walker Hanlon, Florencia Hnilo, 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, Hoover Institution, Northwestern, Oxford, Peking University School of Economics, UCLA, UCSD, University of São Paulo, Wyoming, Yale, the 2023 EHA Meetings, and the 2024 ASREC Meetings for incredibly helpful comments. All errors are ours.

[†]The World Bank; ali.almelhem@colorado.edu

[‡]University of Colorado, Boulder & IZA; iyigun@colorado.edu

[§]University of Colorado, Boulder; austin.kennedy@colorado.edu

[¶]Chapman University; jrubin@chapman.edu

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; Giorelli, Lacetera and Marinoni 2022). This makes language a useful tool for analyzing cultural differences across societies.

This paper tackles these issues by providing empirical evidence linking cultural change—as embedded in language—to one of the most important episodes in economic history: Britain’s industrialization. In doing so, we trace for the first time in the literature the evolution in the language of science, religion, and political economy in the centuries leading to the British Industrial Revolution. We employ textual analysis methods to the universe of digitized printed volumes contained in the Hathitrust Digital Library, published 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 thinkers, such as Francis Bacon and Isaac Newton, with what would become the “Industrial

¹For 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).

Enlightenment,” and ultimately Britain’s Industrial Revolution. This progress-oriented view of science was nascent and revolutionary in 17th century Europe, while it was more or less absent in Classical Antiquity, medieval Europe, the Middle East, South Asia, or China. As such, it encapsulated the idea that science and our understanding of the natural world could be used to improve the lot of humankind. For instance, [Friedel \(2010\)](#) illustrates how the West developed a “culture of improvement” over the last millennium—slowly through the end of the 15th century, but at a much more rapid pace afterward. Along these lines, [Slack \(2015\)](#) makes the case that the idea of “improvement” was primarily confined to increases in productivity and efficiency in agriculture alone until the end of the 17th century, after which point it began to be widely used and applied in all other facets of human endeavour.² [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.³

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.⁴ 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

²The idea of progress as a driver of improvement in practical applications of “useful knowledge” among high human capital mechanics and artisans has also been emphasized to a great extent in [Mokyr \(2002, 2016\)](#) and developed further in [de la Croix, Doepke and Mokyr \(2018\)](#).

³For 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 \(2022\)](#).

⁴By 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.

read via OCR,⁵ this yields 173,031 unique volumes published between 1500 and 1900. These include works of fiction (e.g., Shakespeare, Austen, Dickens), poetry, scientific manuals, religious texts, and much more. 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 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.

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.⁶ 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. To be clear, we are *not* classifying volumes as ones of science, religion, or political economy; we are classifying volumes based on the *language* they used. In fact, the corpus includes works of fiction that would not obviously fall into one of these categories. What matters for our purposes is the *language* such works employed. Moreover, some volumes we would clearly recognize as works of science may predominantly use the language of religion—this is especially true of volumes from the 16th and 17th centuries, when it was common to invoke God and the heavens in discussions of ostensibly scientific topics.

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 manuals, published in the late 19th century, cover nearly all aspects of machine-related production and much more. 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.

⁵The fact that some books cannot be read means that there may be selection bias against older volumes, which are more likely to be unreadable. There is certainly survivor bias among these volumes as well. For these reasons, we show robustness to dropping volumes published prior to 1650.

⁶The 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.

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 in the language of science and religion. In other words, the language of science was secularized by the early Enlightenment. However, there was a shared language for science and political economy throughout the period, and there was likewise a shared (though different) language for religion and political economy. 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 became more progress-oriented beginning in the 18th century and this persisted throughout the period in question. However, there is an important caveat to this finding: works using the language of “pure” science were largely neutral with respect to progress-oriented language. The most progress-oriented works were at the *nexus* of the languages 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 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 the languages 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.⁷

At the outset, it is incumbent upon us to note some limitations of our approach. The changes in English language we shall present below could reflect a host of demand and supply factors including scientific progress driven by the age of discoveries, the expansion of higher-learning institutions, selection into book writing and different genres, and selection of authors writing in English instead of other languages, such as Latin. However, [Mokyr \(2016\)](#) provides the essential scaffolding upon which a compelling case can be made that the evolving culture of Britain—specifically, as it relates to the progress-oriented view of scientific works with industrial applications—was one of the likely drivers of the patterns we shall identify in what follows.

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 works closely related to our study, [Grajzl and Murrell \(2019, 2021\)](#) 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. A more related recent contribution is [Grajzl and Murrell \(2023\)](#), which uses English textual data about the pre-industrial era and also speaks to the relationship between religion and science. Their data span 1530–1700, resulting in around a third of the volumes as in our sample (57,863 versus 173,031). They find no evidence in this period of secularization of the languages of science and institutional thought. We similarly find that, through at least the mid-17th century, most works—regardless of topic—used the language of religion. However, by extending the data to 1900, we find that the languages of science and religion became

⁷These 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.

increasingly distinct in the late 17th through 19th centuries. Finally, [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.⁸

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

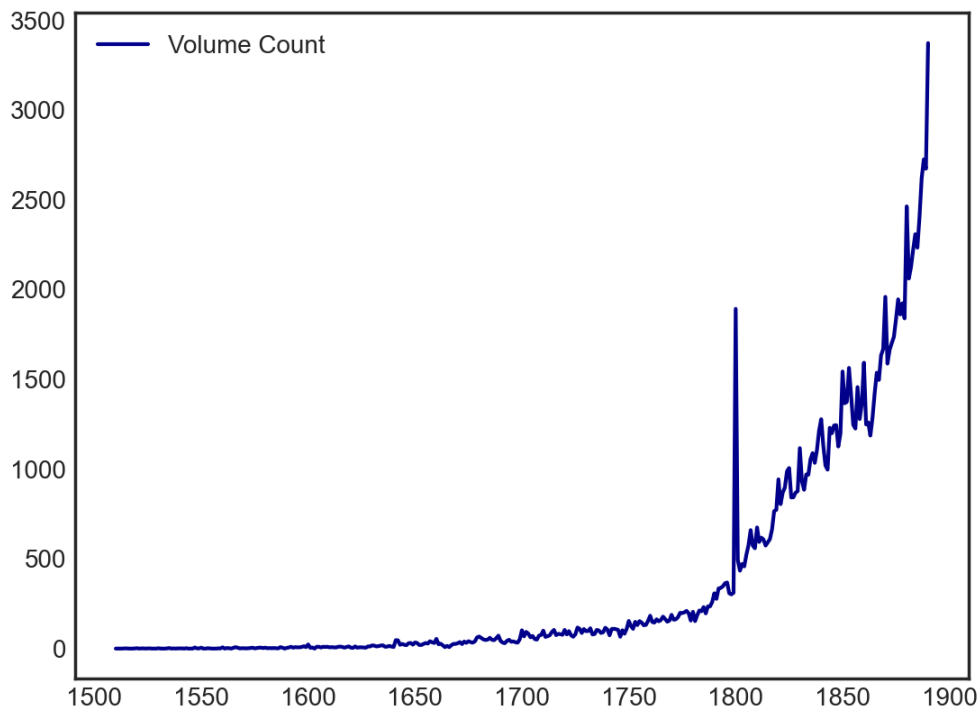
We collected data from the Hathitrust Digital Library (HDL), a collaboration between major universities in the US (now the Big Ten Academic Alliance) and the University of California public system. The HDL aims to establish a shared repository of digitized works from member universities for archival and non-consumptive research purposes. This includes materials digitized by Google, Microsoft, and 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 published in England and written in English in the HDL over the period 1500–1900.⁹ Figure 1 reports the distribution of volumes in our data set by year.¹⁰

⁸Other 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 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.

⁹We requested that HDL provide us with all books in their repository printed in England between 1300 and 1900. There was no constraint with respect to language, and several volumes in the early part of the corpus are in Latin.

¹⁰There 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

Figure 1: Distribution of Volumes



Note: the distribution of volumes smoothed over 20-year intervals are available in Figure B.1.

We are interested in the content of the volumes. While we have access to word order in our dataset, the most suitable procedure for describing intra-document structure, Latent Dirichlet Allocation (described in Section 2.3), does not require word order.¹¹ Hence, we used HDL’s Extracted-Features dataset, which models each volume as a “bag of words”.

know what the biases are. We address this issue in Appendix D, 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 D 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.

¹¹We chose LDA because we wanted to look at the evolution of topics, so a topic-model is the most suitable algorithm to employ. We did not choose LDA simply for computational reasons—we had a supercomputer at our disposal.

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

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 culling 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.¹²

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.¹³

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).

¹²Hathitrust is a consortium of university libraries. The reason that so many volumes are dropped is that many popular volumes appear in multiple libraries in many editions (e.g., Darwin’s *On the Origin of Species* appears in multiple libraries).

¹³Another 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.

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. As an example, the terms “whale” or “Ahab” would have a high TF-IDF score in the novel *Moby Dick* because they are important for the book itself but not common across all other books. In contrast, terms with a low TF-IDF score do not appear frequently in particular volumes, even if they appear regularly in many volumes throughout the corpus, which suggests that such words have low informational content.

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:

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

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.¹⁴ 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 anywhere within the same volume. In

¹⁴There are two reasons we drop the bottom 20% of words. First, this is a standard technique in NLP preprocessing to reduce computational demand, as using the entire vocabulary would be extremely resource intensive ([Blei and Lafferty 2009](#)). Second, TF-IDF scores measure the informativeness of a given word in a document. A high TF-IDF score suggests a given word appears frequently in this particular document (Term Frequency), and not in many other documents (Inverse Document Frequency), suggesting that a given word is commonly and distinctively used in this document. By removing the bottom 20% of words ranked by their TF-IDF score, we are removing words that do not appear much in this document while also appearing regularly in all other documents, indicating a given word does not contain unique information about the content of this document.

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. LDA is an *unsupervised* machine learning algorithm. This means that the researcher does not have labels for the data for the algorithm to learn from, and we do not set any variables in the objective function from which the LDA algorithm attempts to maximize. As a result, unsupervised algorithms discover patterns in the underlying data without any explicit guidance or instruction. 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-occurring 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.¹⁵ 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. That is, perplexity captures how “surprised” a model is of new data it has not seen before. The smaller the perplexity, the more likely it is that a model can guess the value which will be drawn from the distribution. Specifically, perplexity 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 $K = 4$).¹⁶ 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 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 $T = 60$.¹⁷

¹⁵Since 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.

¹⁶ 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.

¹⁷Due 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-

An alternative metric to determine goodness-of-fit is *coherence* (Mimno et al. 2011). This measure is used to evaluate the interpretability and semantic meaningfulness of topics identified by the LDA model. It assesses how well the top words within a topic relate to each other thematically. A higher coherence score indicates that the words within a topic are more semantically similar and the topic itself is easier for humans to understand. This helps researchers choose the optimal number of topics and avoid nonsensical topics generated by the LDA model that are artifacts of statistical inference. This coherence score metric yields 80 topics. In Appendix C, we re-run the analysis in Sections 3 and 4 and find very similar results to those reported in those sections.

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.1 lists each of the 60 topics derived from the LDA described in Section 2.3. Some topics are clearly related to the language of science. For instance, topic 7 {fig water iron engin pressur steam electr} and topic 44 {plant flower stem genus yellow calyx bot} are both the language of science. 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 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

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.

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. Letting w_{iv} denote the weight corresponding to a given topic-pair $i \in \{1, \dots, I\}$ with $I = 1,770$ and volume $v \in \{1, \dots, V\}$, 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.¹⁸ This was a subjective choice, based on our reading of the topics, many of which seemed to fall into the language of 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 .¹⁹ 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

¹⁸Results are similar when using four or five categories.

¹⁹We derive similar results using a product rather than a sum.

Share is high for a topic-pair only if the topics frequently co-exist within the same volume. Therefore, if *Incidence* takes on a high value, it indicates that the three topics within the category co-exist within the same volume often throughout the corpus.

We proceed to rank each possible category from highest *Incidence* to lowest. We seek categories with the highest *Incidence* value that i) have analytic meaning as a category, and ii) do not include the same topics as other categories. We begin by excluding topics that are commonly found among categories with high *Incidence* but are broadly used in writing regardless of the subject. These topics—5, 9, 22, 26, 35, 50, and 55—use words commonly found in literature. If one were to re-do this exercise to analyze the evolution of the language of love, some of these topics would not be excluded. However, for our purposes, their inclusion would entail that the chosen categories have little analytical meaning. We also omit topic 46, which has both “scienc” and “religion” as root words.²⁰ It is not surprising that these two words would be found together in the same topic, but including this topic would not yield a category that is analytically distinct. The highest *Incidence* category that does not have a topic in the omitted topics list is {33,34,47}. All three of these topics are broadly “political economy” in nature and thus have analytic meaning. We hence choose this as a category. We proceed to seek categories with the highest *Incidence* score that do not contain topics from the omitted topic list or 33, 34, or 47. This yields the category {4,12,52}, all three of which are associated with religion, and we therefore use this as our second category. We proceed to seek a third category that does not contain topics from the omitted topic list or the other two categories. This yields the category {7,8,41}, all three of which are associated with science, and we therefore use this as our third category. The categories (labeled manually) and their topics produced by this process are presented in Table 1.

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 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

²⁰Nonetheless, our results are robust to the “religion” category containing topics {4,12,46}, which would be the case had we not omitted topic 46. These results are available upon request.

Table 1: Categories

Category	Topics and associated words
“Political Economy”	33 - law lord show public evid opinion fact suppos respect observ
	34 - govern nation polit parliament constitut war parti british civil declar
	47 - trade amount labour money price cent increas bank capit rate
“Religion”	4 - church christian christ bishop holi paul doctrin rome gospel pope
	12 - god christ lord thi faith holi sin heaven jesus thou
	52 - hath fame religion men shew virtu likewis doth tho design
“Science”	7 - fig water iron engin pressur steam electr air heat weight
	8 - acid solut heat carbon water sulphur iron gas oxid metal
	41 - line angl equal equat sin sun plane distanc circl earth

Note: only the first ten roots are included in this table. The full list of roots for each topic are available in Appendix [A.1](#).

volumes in early years and to address measurement error with respect to year of publication. This gives each of the 1770 topic-pairs a *Share* value for each year between 1510 and 1890.²¹

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 47.²² 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 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.²³

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

²¹One issue with binning the data is that doing so may obscure the timing of changes in sentiment we find in our analysis. We address this issue in Appendix [E](#), which re-analyzes the data without using bins. The main results hold.

²²If 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.

²³The 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.

these coefficients within a unit simplex with the categories as vertices. We present the results for each half-century in Figure 2, where each topic is labeled based on its value in 1850.²⁴

There are at least three salient facts to note based on Figure 2. First and foremost, our corpus is fairly thin in the earlier eras, especially in 1550. Hence, the conclusions 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 languages of science and religion became increasingly distinct. In particular, one can see that the frequency of topics using both the languages of science and 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 the language of religion with that of political economy as well as those that involve the languages of science and political economy. Finally, the separation of the languages of science and religion 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

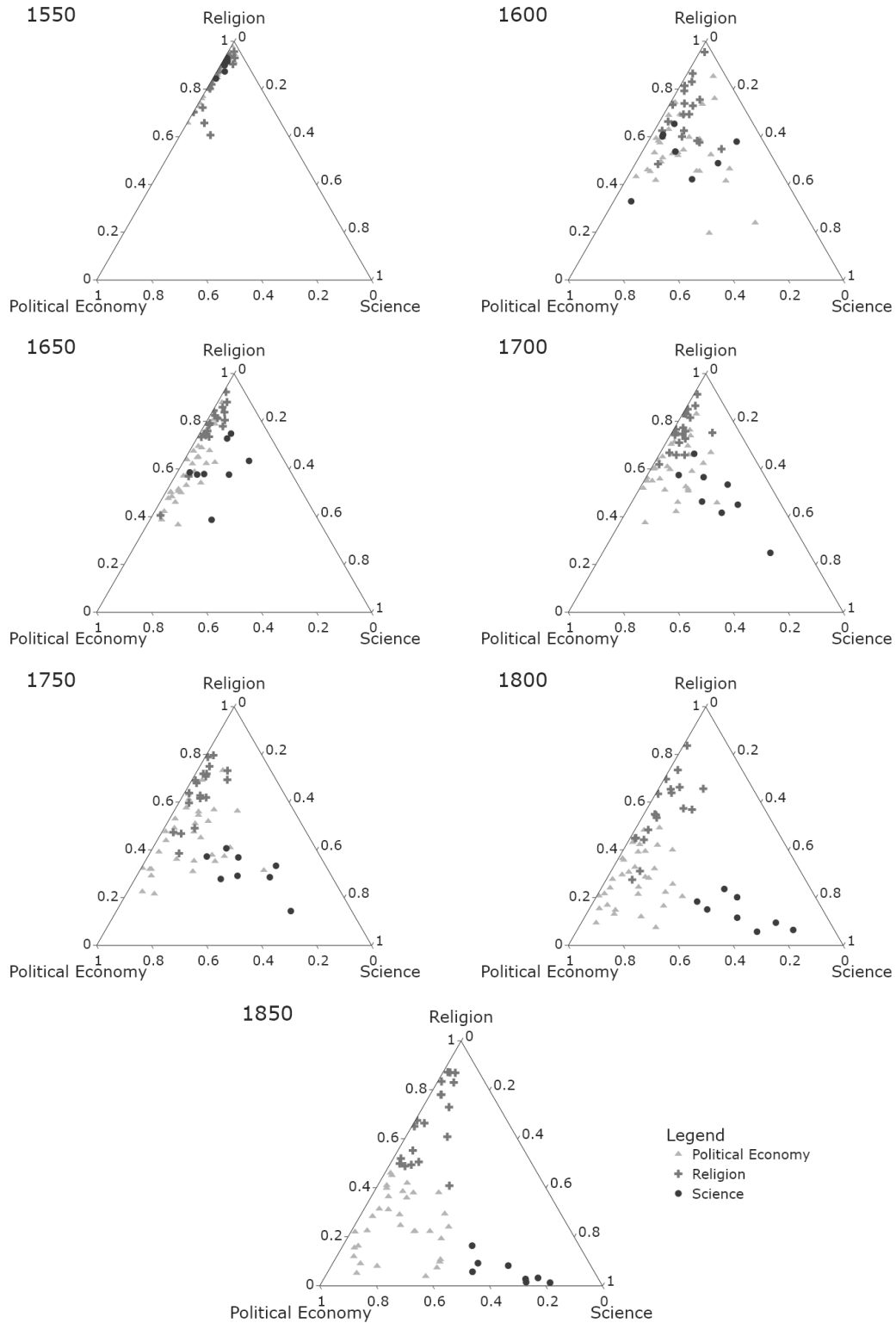
We proceed to classify individual volumes into the three categories based on the language they contain: science, political economy, and religion. This permits us to examine how categories evolved in relation to each other within volumes over time. We begin with the convex combinations constructed in 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.

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 :

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

²⁴Yearly 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 of science may have been more intertwined with religious language earlier in the corpus, and become less so later in the corpus.

Figure 2: Topics by Category, 1550–1850



Note: Categorization into “Science”, “Political Economy” or “Religion” based on topics’ placement in 1850. A color version is available in Figure B.4.

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

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

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 $c \in \{\text{Science, Political Economy, Religion}\}$. Note that for each volume, $\text{Science}_v + \text{Political Economy}_v + \text{Religion}_v = 1$.

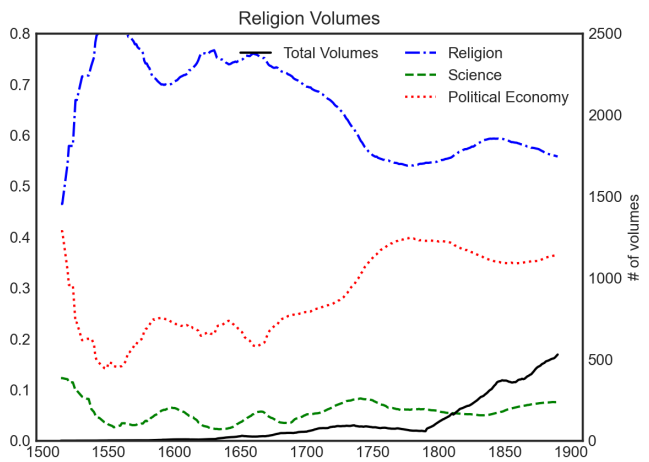
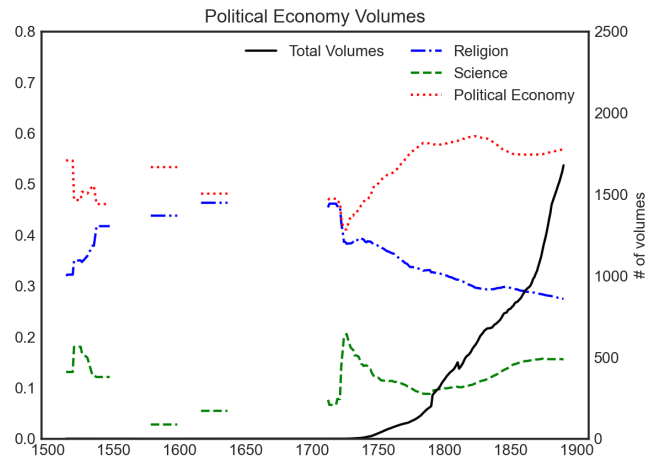
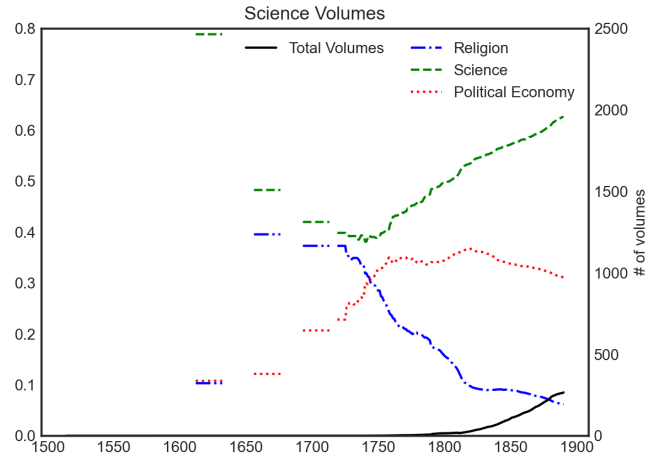
These weights allow us to classify each volume as predominantly using the language of science, political economy, or religion, based on the volume’s highest weight derived from equations (2)–(4). Meanwhile, the weights also permit a measurement of how closely related a volume is to the other two categories.

Our categorization allows us to analyze how the language contained in volumes in a particular category evolve over time in their relation to the other categories. Figure 3 reports the relation between the different categories over time. In these figures, we classify each volume as one of the three categories (i.e., each volume is classified in a category if that category has the highest category coefficient as laid out in equations (2)–(4)), 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).²⁵

As seen in Figure 3, volumes classified as using the languages of science saw their science-specific 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-language 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 3 indicates that beginning in the late 17th century, the language of political economy became more frequently used in works that used the language of religion. However, the language of

²⁵These categorizations permit us to test whether the language used in books that were *translated* differed systematically from books originally written in English. There was certainly selection bias in which books were translated, but such a test can provide some preliminary insight into what was produced in English versus foreign languages of high enough value to warrant a translation. While the metadata associated with our data do not allow us to distinguish translations, we found 2,557 volumes ($\sim 1.5\%$ of volumes) have the string “transl” in their title. We sampled these volumes and they all appear to be translations, although surely there is some measurement error. We compare these translated volumes with the rest of the corpus in Figure B.3. Prior to 1750, the distribution of translations and non-translations appears to be similar across all three categories. After 1750, relatively more volumes using the language of religion were translated than those originally written in English, while relatively fewer volumes using the language of political economy and science were translated. The figure suggests that the “secularization” of language was happening more rapidly in England than abroad in the 18th century.

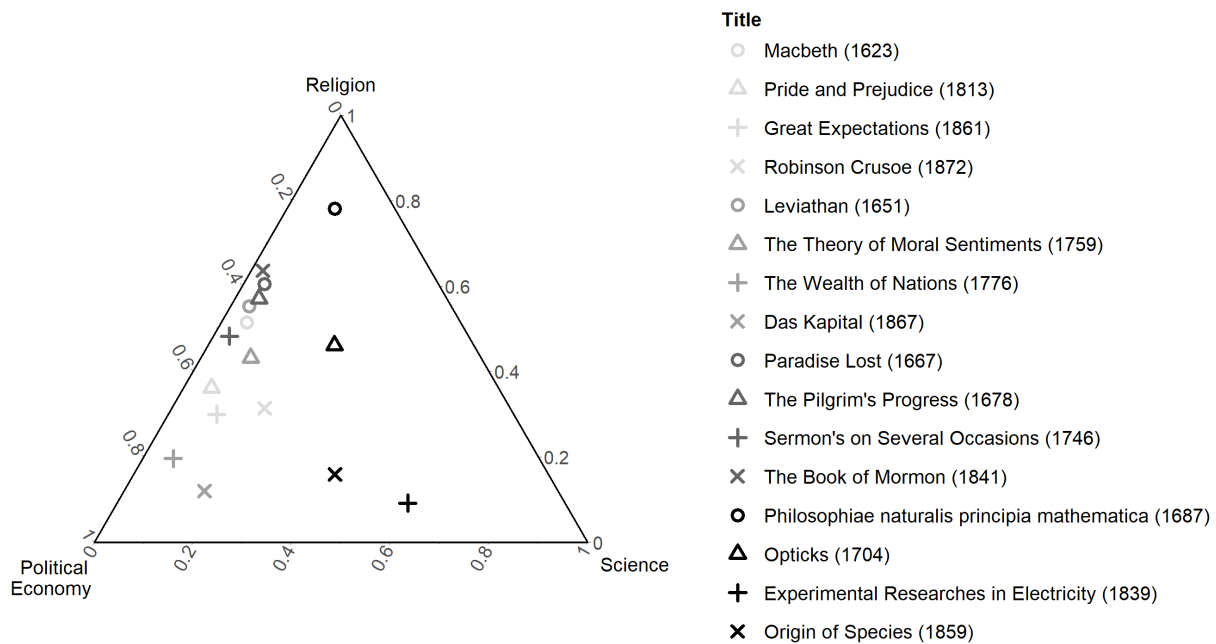
Figure 3: Relationship between Categories over time, within volumes



science was rarely used in works that used the language of religion throughout the period in question.

This finding is supported qualitatively in Figure 4, which places 16 famous works in the simplex. We chose four well-known works indisputably classified as religion, political economy, and science from various dates in the period in question, as well as four famous works of British fiction. 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 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 4: Selected Famous Volumes Categorized



Note: this figure is available in color in Figure B.5.

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 “progress-oriented” words.

To create our list of “progress-oriented” words, we gather the list of synonyms for “progress” from the website www.thesaurus.com. The overall word list is available in Appendix A.2. We manually removed several words from this list for reasons noted in Appendix A.2. First, many synonyms of progress are associated with “movement” or “taking a trip”—this is not the definition of progress of interest to this analysis. Second, several synonyms had alternative meanings. Including such words would introduce severe bias when the words are commonly used in science, such as “evolution” or “momentum.” Third, phrases were 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. Fourth, we removed all words that according to the Oxford English Dictionary were not known prior to 1643 (the year of Newton’s birth).²⁶ We do this to remove bias favoring words that would not have been in volumes written during the Enlightenment. As a final step, each remaining word in our list of “progress-oriented” words is converted to its respective root to match the volume cleaning procedure described in Section 2.2. The final list of “progress-oriented” words is shown in Table 2.

Table 2: Progress Dictionary Word List

progress	advance
improvement	rise
stride	amelioration
betterment	

²⁶Although 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 our list of “progress-oriented” words in Figures B.9 and B.10.

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 :

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

In this case, the numerator represents the absolute score for the words in the progress dictionary, 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.²⁷

In Figures B.11 and B.12, we re-run the analysis using words related to progress and progression from the *Dictionarium Anglo-Britannicum* (Kersey 1708), a 1708 English-language dictionary.²⁸ This produces an alternative list of “progress-oriented” 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”.²⁹ Figure 5 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.³⁰

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 6, along with each volume’s sentiment.³¹

²⁷One 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.

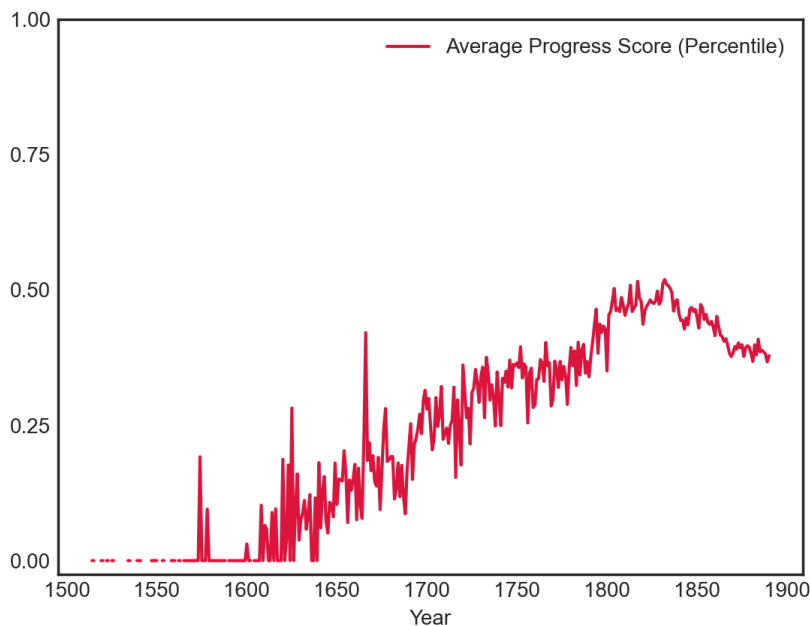
²⁸We extracted words from the dictionary by first looking up the definitions of progress and progression. These definitions included the words “proceed(ing),” “forward,” and “advance(ment).” We proceeded to look up these terms, which gave the additional term “further.” These terms comprise the progress dictionary that we use in Figures B.11 and B.12.

²⁹In a similar exercise reported in Figure B.14, 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.

³⁰This finding is also consistent with a much simpler Google n-gram search of the word “progress”, as reported in Figure B.8. In Figure B.7, we compare the average progress score of translations and non-translations. Prior to the first quarter of the 18th century, there is no discernible difference between the two. Beginning around the third decade of the 18th century, volumes originally written in English appear to have a higher progress score, and this pattern continues throughout the remainder of the period in question.

³¹For an easier interpretation, we plot each volume’s percentile in the entire corpus in terms of sentiment, rather than its raw score.

Figure 5: Average Progress Score (percentile), 1500–1900

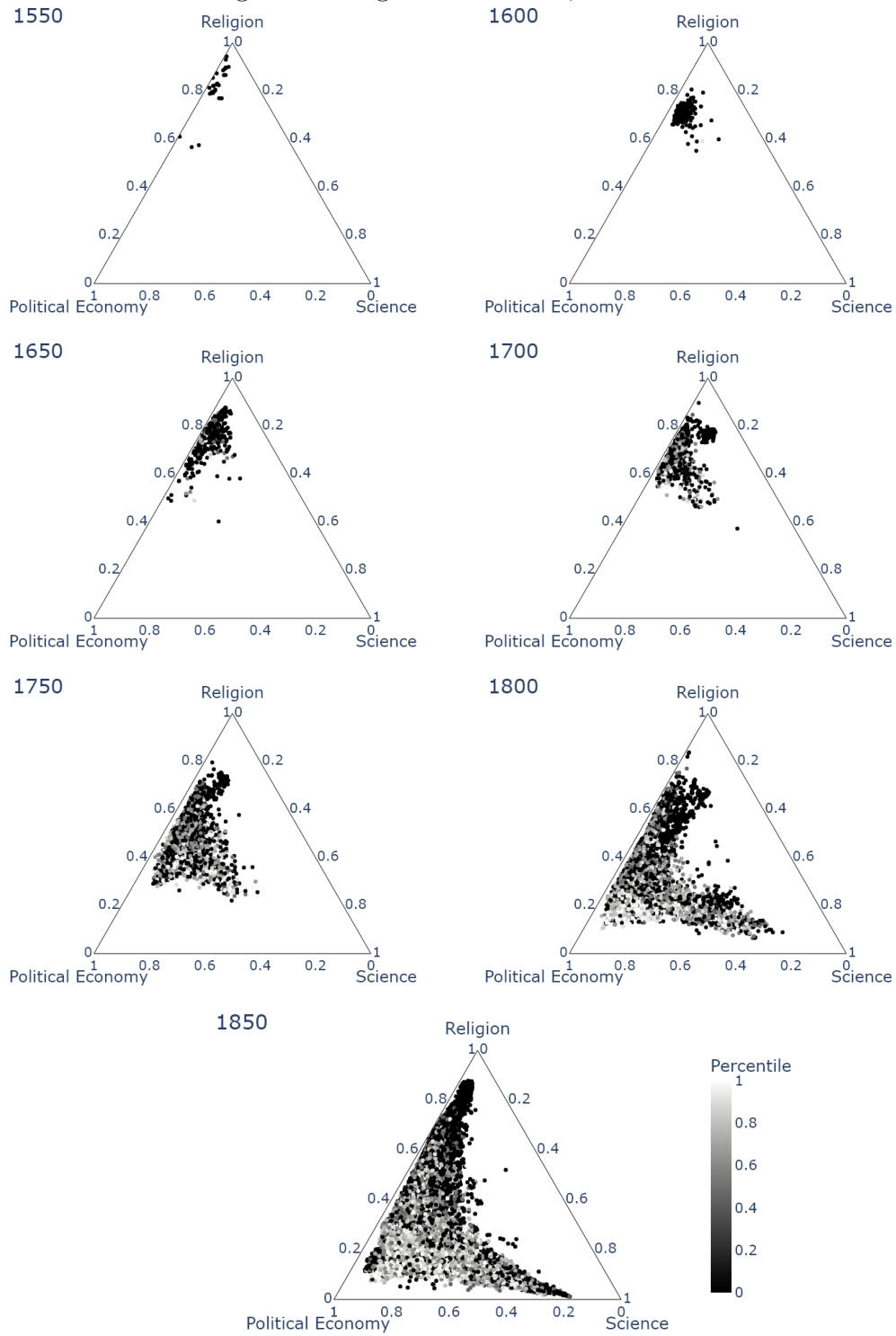


Note: the average progress score taken using a 20-year moving average available in Figure B.6.

Two outcomes are of note in Figure 6. First, consistent with the topics plotted in Figure 2, volumes show similarly distinct language 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.16. Instead of showing individual volumes, Figure B.16 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 6, areas along the political economy-science axis became increasingly progress-oriented over time, especially relative to areas close to the political economy vertex.

Figure 6: 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. A color version is available in Figure B.15.

4.3 Regression Analysis

Figure 6 presents visual evidence that works using language at 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 is correlated with the languages of science, political economy, and religion over time.

We first place volumes (v) into 20 year bins based on date of publication (t).³² We estimate the regression in equation (6), with standard errors clustered by year (i.e., not the bin) of publication.³³

$$\begin{aligned} \text{Sentiment}_{v,t} = & \alpha_1 + \alpha_2 \text{Science}_v + \alpha_3 \text{PolitEcon}_v + \alpha_4 \text{Science}_v \times \text{PolitEcon}_v \\ & + \alpha_5 \text{Science}_v \times \text{Religion}_v + \alpha_6 \text{Religion}_v \times \text{PolitEcon}_v + \lambda_t + \lambda_t \mathbf{A}_{v,t} \boldsymbol{\alpha} + \varepsilon_{v,t}, \end{aligned} \quad (6)$$

where $\text{Sentiment}_{v,t}$ 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.³⁴ 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.1. We plot the marginal effects of Science from equation (6) in panel A of Figure 7. 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 language 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 language, contain only scientific lan-

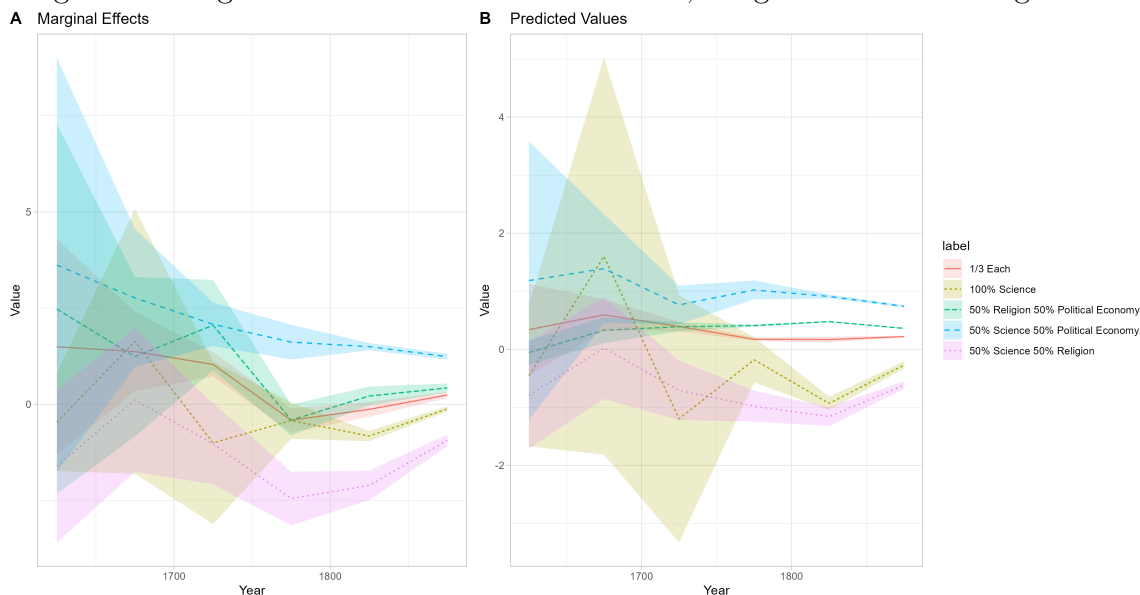
³²We use 20 year (+/- 10 years) bins, for the years 1610, 1630, 1650, etc. up until 1890. We exclude the 16th century due to the low amount of volumes digitized in that period. Unbinned results are available in Appendix E. Since the data are sparse prior to 1650, we re-run the regressions dropping all data from prior to 1650. These results, reported in Figure B.13, are similar to those reported in Figure 7.

³³Clustering standard errors by year addresses the possibility that errors are correlated within years. Robust standard errors yield very similar results.

³⁴Recall that the three category weights for each volume add up to one. Hence, we exclude Religion as an independent variable.

guage, or contain equal parts religion and political economy language, the marginal effect is near zero or even slightly negative throughout the period. The marginal effect of *Science* on volumes that contained equal measures scientific and religious language is negative, although (as shown before) very few volumes are located at this nexus.

Figure 7: Marginal Effects and Predicted Values, Progress Sentiment Regressions



These results are further supported by panel B of Figure 7, 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 language 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 remains 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 using pure scientific language and those at the science-religion nexus have, on average, negative progress-oriented sentiment throughout most of the period in question. Although the result on volumes of pure science may seem surprising, this could reflect the more technical audience that such volumes sought to reach—not those artisans and tinkerers that were so essential for Britain’s industrialization.

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 using the language 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 using the language of *applied* science—those at the nexus of science and political economy—became more progress-oriented.³⁵

5 The Language of Industrialization

The results presented thus far suggest that volumes using language 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. 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 using the language of 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).³⁶ These volumes cover all types of industrial machinery and their indexes are extremely detailed. While they do not cover every field in which progress was made during the Industrial Revolution (e.g., medicine, domestic lighting), they include most aspects of

³⁵It is possible that our analysis thus far has picked up sentiment that is not necessarily progress-oriented, but is broadly optimistic in nature. We address this issue by creating a “dictionary” of optimistic sentiment using the same methodology we used to create the progress dictionary. We report the results in Appendix F. The results are nearly the mirror opposite of those found for progress-oriented sentiment in Figure 6. These results suggest that the analysis is not merely picking up some broader change in optimistic language.

³⁶The 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.”

industrialization, including engines, mining, railways, iron structures, pumps, boilers, workshops, metallurgy, and much more. 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.

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 1643.³⁷ Each word is weighted by the number of times it appears in the indexes. The top 10 industrial words are reported in Table 3, along with the number of times they appear in the Appleby’s indexes, and the top 51 words are reported in Appendix Table B.2.³⁸

Table 3: 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

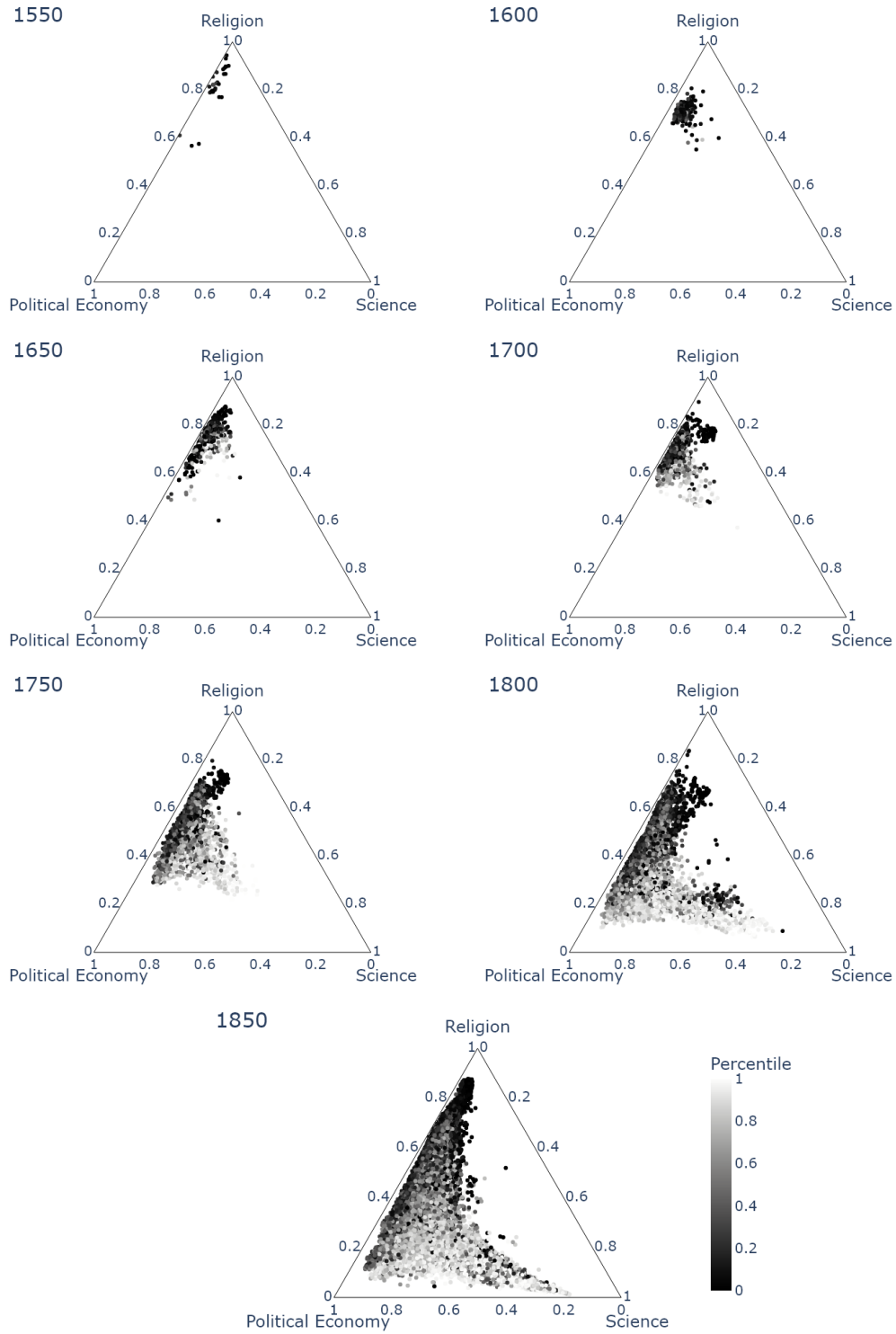
We proceed to derive an *industrial score* for each volume in the corpus. The industrial score is calculated by multiplying the count of each word in a volume by its corresponding weight, summed across all words with positive industrial weights. This sum is normalized by dividing by the total length of the volume. 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 using the language of “pure science”—those in the bottom right corner of the triangles—appear

³⁷In Figure B.17, we report the industry sentiment scores using words in existence after 1643. Results are similar.

³⁸We 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 shade of each dot represents the sentiment of that volume, with lighter shades representing more industrial sentiment. A color version is available in Figure B.18.

to be the most related to industrialization (i.e., the lightest shade), while volumes using the language 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.

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 employing language 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 plots progress and industry sentiment together, with darker red shading indicating a higher progress-oriented score and darker blue shading indicating a higher industry score (and thus darker purple indicating a high score for both). It is apparent from Figure 9 that beginning in the mid-18th century, and especially in the early 19th century, volumes using language at the science-political economy nexus were, on average, high in both industry and progress-oriented score.

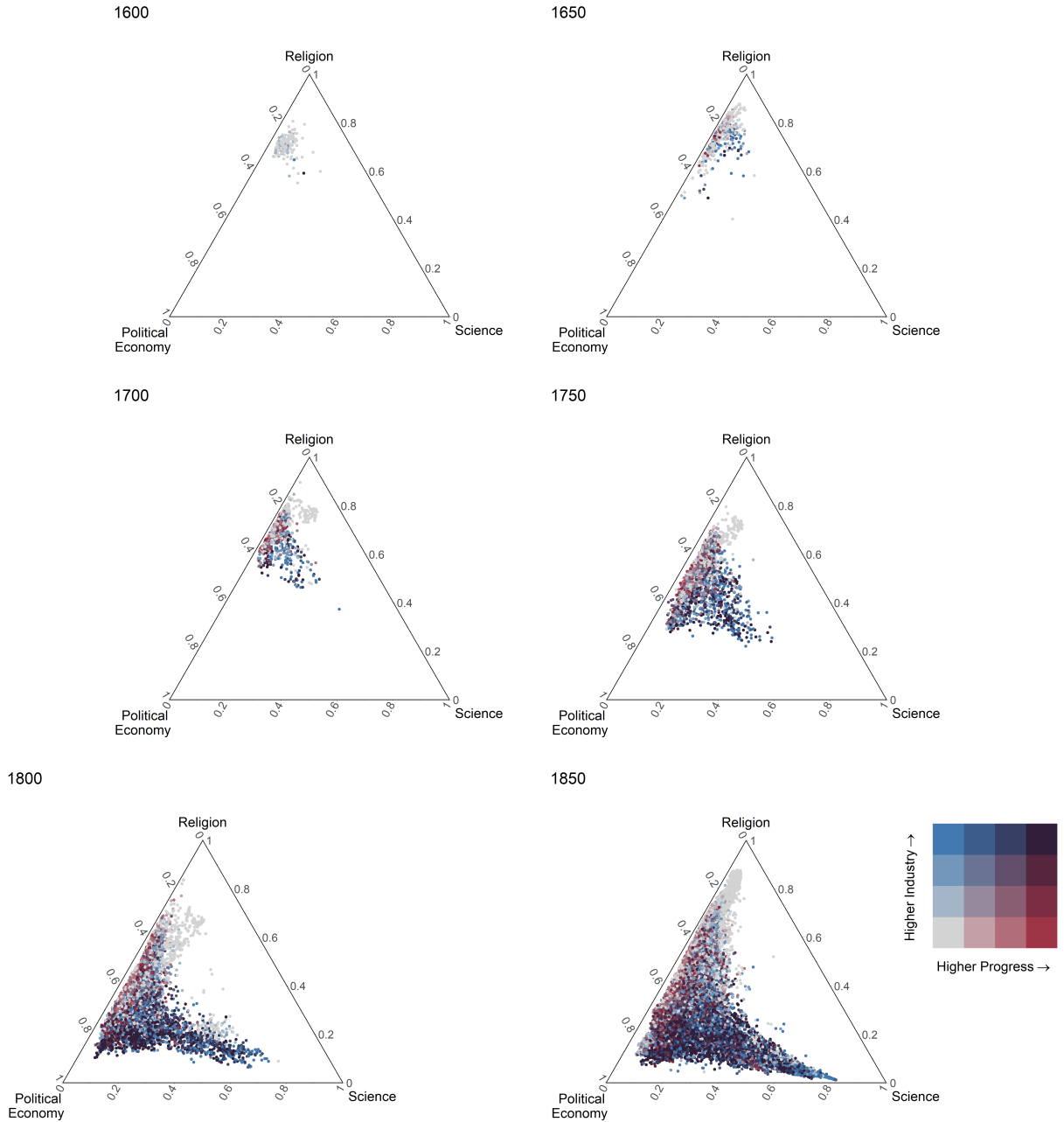
An econometric analysis further supports these findings. This requires an analysis of three dimensions: a volume’s industrial score, its placement in the science-religion-political economy simplex, and its progress-oriented sentiment. To clarify these relationships, we present results from an OLS regression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

$$\begin{aligned}
 \textit{Sentiment}_{v,t} = & \beta_1 + \beta_2 \textit{Science}_v + \beta_3 \textit{PolitEcon}_v + \beta_4 \textit{Industry}_v \\
 & + \beta_5 \textit{Science}_v \times \textit{PolitEcon}_v + \beta_6 \textit{Science}_v \times \textit{Religion}_v + \beta_7 \textit{Religion}_v \times \textit{PolitEcon}_v \\
 & + \beta_8 \textit{Science}_v \times \textit{Industry}_v + \beta_9 \textit{PolitEcon}_v \times \textit{Industry}_v \quad (7) \\
 & + \beta_{10} \textit{Science}_v \times \textit{Religion}_v \times \textit{Industry}_v + \beta_{11} \textit{Science}_v \times \textit{PolitEcon}_v \times \textit{Industry}_v \\
 & + \beta_{12} \textit{Religion}_v \times \textit{PolitEcon}_v \times \textit{Industry}_v + \lambda_t + \lambda_t \mathbf{B}_{v,t} \boldsymbol{\beta} + \varepsilon_{v,t},
 \end{aligned}$$

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.3 reports the regression results. Figure 10 re-

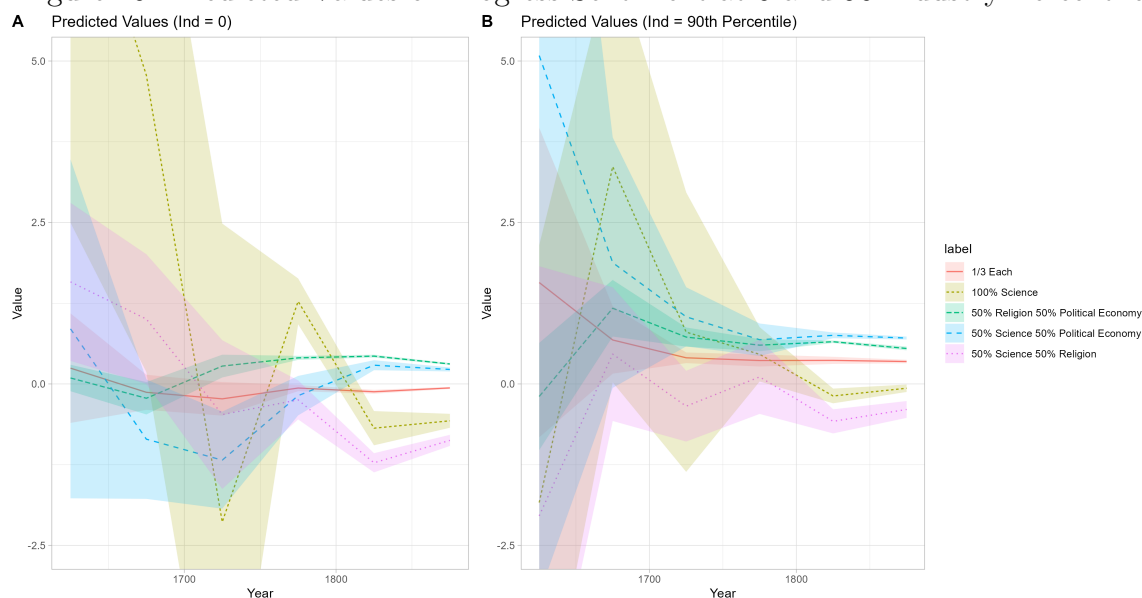
Figure 9: Progress and Industry Sentiment, 1600–1850



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 along both the industry and progress axes. A version with the size of circle representing industry sentiment is available in Figure B.19.

ports 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).³⁹

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



Several results follow directly from this analysis. First, volumes using language at the science-political economy nexus were more progress-oriented in the 18th century if they *also* had a high industry score. 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 using language 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 were 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 were not located on the religion-political economy nexus, even if there were progress-oriented volumes at this nexus (see Figure 6).

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 also employed language 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.

³⁹Due to the relatively few volumes published prior to 1650 in our data set, we re-run these regressions dropping all pre-1650 data. The results, reported in Figure B.20, are similar to those reported in Figure 10.

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.⁴⁰ The central focus of the book is the science of

⁴⁰The 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. Although we would now recognize it as a book of science, as its name suggests (“An Historical and Philosophical Account...”) it is not surprising that it used the *language* of religion slightly more than that of science.

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 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.⁴¹ [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.

In this paper, we test these claims using quantitative data from 173,031 works printed in England between 1500 and 1900. We trace for the first time in the literature the evolution of the languages of science, religion, and political economy in the centuries leading to the British Industrial Revolution. We document the following three results. First, there is little overlap in works using the language of science and religion in the period under study. This indicates that the “secularization” of science was entrenched from the beginning of the Enlightenment. Second, while works using the language of science did become more progress-oriented during the Enlightenment, this sentiment was mainly concentrated in the nexus of science and political economy. Third, those volumes using language at the nexus of science and political

⁴¹The 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.

economy that *also* used the language of industrialization were particularly progress-oriented. We interpret these findings to mean that it was the more pragmatic, industrial works using the language 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.

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 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 in which 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.
- de la Croix, David, Matthias Doepke and Joel Mokyr. 2018. “Clans, Guilds, and Markets: Apprenticeship Institutions and Growth in the Preindustrial Economy.” *The Quarterly Journal of Economics* 133:1–70.
- 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.
- Friedel, Robert. 2010. *A Culture of Improvement: Technology and the Western Millennium*. Cambridge, MA: The MIT 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.
- Grajzl, 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.
- Grajzl, Peter and Peter Murrell. 2021. “Characterizing a legal–intellectual culture: Bacon, Coke, and seventeenth-century England.” *Cliometrica* 15(1):43–88.
- Grajzl, Peter and Peter Murrell. 2023. A Macroscope of English Print Culture, 1530-1700, Applied to the Coevolution of Ideas on Religion, Science, and Institutions. Technical report SSRN Working Paper 4336537.
- 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.
- Mimno, David, Hanna M. Wallach, Edmund Talley Miriam Leenders and Andrew McCallum. 2011. “Optimizing Semantic Coherence in Topic Models.” *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* pp. 262–272.
- Mokyr, Joel. 2002. *The Gifts of Athena: Historical Origins of the Knowledge Economy*. Princeton, NJ: Princeton University Press.
- 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.

- Slack, Paul. 2015. *The Invention of Improvement: Information and Material Progress in Seventeenth-Century England*. London: Oxford University Press.
- 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 and Progress-Oriented Words

A.1 Topics

- 1 - paint pictur artist music engrav painter colour figur repres portrait centuri style art execut beauti gold ornament design collect plate
- 2 - town road church build built river stone wall citi erect north south hill tower situat bridg villag ancient valley castl
- 3 - franc pari french loui madam duke count napoleon princ monsieur emperor charl assembl queen bonapart revolut grand convent duchess henri
- 4 - church christian christ bishop holi paul doctrin rome gospel pope scriptur cathol apostl divin religion roman council faith clergi epistl
- 5 - love heart beauti soul sweet dark night earth voic sun bright heaven thi child happi mother joy thou fear rose
- 6 - india chines china nativ indian bengal govern european calcutta bombay khan british madra hindu hindoo provinc emperor opium rajah japanes
- 7 - fig water iron engin pressur steam electr air heat weight surfac inch plate construct current diamet resist cylind temperatur machin
- 8 - acid solut heat carbon water sulphur iron gas oxid metal colour oxygen quantiti precipit alcohol hydrogen dissolv liquid copper temperatur
- 9 - exist refer period similar consist occur connect instanc adopt establish system distinct examin previous extent peculiar consequ result probabl remark
- 10 - vol lond fol folio calf copi pari par morocco gilt sermon sur catalogu neat cum print memoir translat tom von
- 11 - thou thi hath sir doth duke ladi pray exit scene nay hast madam fool sweet marri exeunt mistress ant henri
- 12 - god christ lord thi faith holi sin heaven jesus thou spirit father love soul divin christian grace bless glori hath

- 13 - diseas blood patient treatment medic pain fever skin oper brain membran stomach organ nerv surfac nervous urin occur bone affect
- 14 - tho adj tlie lat tbe hut arc lit sax tin tliat aud ihe tiie tlic sec dryden ger tlio tor
- 15 - ofth sor differ juft sufficient suffer hath thousand eftat offic effect offer affect underftand apoftl tbe loft ofa amongft
- 16 - scotland quot edinburgh scottish highland burn dougla ane ing jame loch bonni con ava hae auld glasgow andrew mari avith
- 17 - note latin verb greek languag comp text plural deriv translat compar denot sentenc refer passag iii vowel noun root english
- 18 - quod cum est sed quam qui aut hoc vel ess atqu cic sunt enim etiam quid autem qua nec ita
- 19 - parish esq counti street ditto rev park lane market york borough kent bridg joseph yorkshir castl township durham georg bristol
- 20 - thoma john william richard robert mari henri elizabeth ann edward daughter jame esq georg manor marri wife parish buri roger
- 21 - henri bishop edward earl reign william archbishop norman saxon richard abbot centuri castl abbey church canterburi roman pope ireland york
- 22 - poet poetri play poem genius johnson literari literatur wrote publish critic theatr writer poetic stage style shakespear music vers translat
- 23 - roman greek rome athen caesar greec senat cicero templ athenian itali alexand augustus homer emperor persian plini consul asia jupit
- 24 - court defend plaintiff estat properti bill contract statut action law judgment entitl payment tenant writ execut parti purchas sale debt
- 25 - esq jan oct dec nov feb juli aug june hon sept dau april rev coll earl mar capt bart dublin
- 26 - morn river arriv distanc travel kill wild reach even night journey walk start parti visit wind hors cover black water
- 27 - earl duke queen parliament majesti sir lord charl henri jame william princ georg edward scotland thoma ladi royal elizabeth bishop

- 28 - fig surfac develop genus structur upper geolog shell thick limeston section rock seri
anterior fossil lower format stratum occur coal
- 29 - thou thi israel hath ver david jew jerusalem mose luke hebrew jesus gen prophet job
christ behold matt egypt judah
- 30 - fish black white bird colour tail brown anim take game fli nest size yellow femal male
food wing play upper
- 31 - armi enem command march french attack captain regiment colonel british war militari
cavalri offic arriv wound battl advanc artilleri corp
- 32 - quod cum vel regi anno est qui rex domini apud hoc dei ejus quia fuit quam johann
super idem dominus
- 33 - law lord show public evid opinion fact suppos respect observ question practic case
effect requir relat land express england learn
- 34 - govern nation polit parliament constitut war parti british civil declar ireland minist
polici establish propos bill system franc foreign irish
- 35 - miss ladi mother look room father woman ask dear knew talk girl felt door child cri
moment wife voic walk
- 36 - compani court offic appoint counti act board council paid majesti aforesaid committe
report parish grant sum justic prison date charg
- 37 - indian river island coloni south america american coast lake africa north canada provinc
popul british cape west bay governor trade
- 38 - excit punish display indulg circumst alarm digniti dread violenc assembl perceiv exert
temper retir conceal contempt disposit abandon solemn reflect
- 39 - railway messr committe street liverpool manufactur cent tho manchest associ engin
secretari patent district exhibit york chairman presid birmingham local
- 40 - king citi war princ armi england command court queen kingdom brother english peac
battl reign nation enem march crown royal
- 41 - line angl equal equat sin sun plane distanc circl earth motion parallel centr axi moon
posit squar surfac forc perpendicular

- 42 - arab egypt greek egyptian persian sultan turk ancient turkish east christian desert Nile
persia city eastern temple asia russian pasha
- 43 - edit cloth crown illustr vol svo rev price post translat extra histori volum seri revis
paper map print fcap essay
- 44 - plant flower stem genus yellow calyx bot fruit seed juli smooth purpl root corolla leaf
linn oblong fig base ovate
- 45 - boil wine water salt sugar butter mix dri pound oil cut colour milk hot cold white dish
pour liquor meat
- 46 - moral human exist scienc idea principl develop univers philosophi conscious religion
individu sens theori divin system physic religi christian influenc
- 47 - trade amount labour money price cent increas bank capit rate gold manufactur total
pay cost system industri popul silver averag
- 48 - linn genus nat margin lin hab var folii fig syst ent apic bot brit gen abdomen tab thorax
apex prod
- 49 - par qui est che pour sur nous tout vous une fait sont quil comm bien cett dit avec con
aux
- 50 - thi thou heaven sweet hath thine joy thee song hast oft breast fate breath ere beneath
youth behold fli mighti
- 51 - plant soil garden hors dri cultiv winter tree sheep grow crop fruit growth seed season
manur food grass farm weather
- 52 - hath fame religion men shew virtu likewis doth tho design pretend discours punish
farther mankind oblig liberti contrari publick fee
- 53 - school colleg societi educ church rev instruct bishop visit christian religi preach mis-
sionari appoint sunday oxford institut mission teach week
- 54 - fame fee fever cafe fet ufe feem obferv defir confider fay occafion ifland juft fuppof feen
purpof fort efq hall
- 55 - letter dear ladi friend father wife repli happi convers acquaint famili honour visit told
busi brother pleas arriv woman compani

56 - hym doe hath bee sayd doth own wee kyng hem tyme ben all ther wold down thou
sonn self again

57 - kal kai tov yap rov occ masc xai fut sing mid inf hann gen aor avrov til ovk comp kat

58 - ship island sea captain vessel coast sail shore bay north board wind port boat voyag
south cape harbour west river

59 - emperor itali spain german germani franc duke pope russia russian spanish italian rome
austria charl empir count europ don napl

60 - cri ladi repli gentleman door boy captain laugh tom girl hors jack fellow miss exclaim
talk doctor smile dog dress

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

A.2 Progress-Oriented Words

Words used in Progress Dictionary

advance

amelioration

betterment

improvement

progress

rise

stride

Words omitted because only in use after 1643

boost

development

headway

Terms omitted because they are multiple words

build-up (hyphenated term in context of “progress”)

step forward

Terms omitted because of alternative meaning

anabasis (military)

break (many more common meanings, especially “to separate or sever”)

breakthrough (often used in context of “scientific breakthrough”)
evolution/evolvment (science)
flowering (science)
growth (economics)
increase (math)
momentum (science)
motion (science)
process (science)
proficiency (most commonly related to “skill”, not “progress”)
promotion (business)
rate (math)

Terms omitted because definition associated with “movement” or “taking a trip”

course
dash
expedition
hike
impetus
journey
lunge
march
movement
ongoing
pace
passage
procession
tour
unfolding
voyage
way

B Additional Figures and Tables

Table B.1: Dependent Variable: Progress Percentile

	1625	1675	1725	1775	1825	1875
(Intercept)	2.402*** (0.322)	0.028 (0.115)	-0.058 (0.093)	0.453*** (0.114)	0.177* (0.090)	0.209* (0.091)
Science	-0.469 (0.647)	2.104 (1.871)	-0.537 (1.257)	0.043 (0.691)	-0.359 (0.651)	0.345 (0.648)
PolitEcon	-1.273* (0.615)	1.358 (0.980)	1.109+ (0.653)	1.213+ (0.620)	1.807** (0.615)	1.809** (0.615)
Science \times Religion	-2.283* (0.933)	-0.769 (2.064)	2.234 (1.571)	-1.768+ (1.007)	-0.284 (0.973)	0.654 (0.939)
Science \times PolitEcon	8.168 (6.750)	-5.914 (7.697)	-2.003 (7.086)	-4.088 (6.790)	-3.521 (6.754)	-5.447 (6.751)
Religion \times PolitEcon	2.273* (1.037)	-0.997 (1.735)	0.380 (1.122)	-1.501 (1.081)	-1.015 (1.040)	-1.320 (1.040)
Num.Obs.	162 236					
R2	0.124					

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Volumes are placed into 20 year ((+/-) 10 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.

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

	1625	1675	1725	1775	1825	1875
(Intercept)	1.896*** (0.297)	-0.206+ (0.115)	-0.285** (0.104)	-0.271 (0.194)	0.055 (0.093)	0.019 (0.095)
Industry	0.430 (0.598)	0.157 (0.706)	0.472 (0.640)	0.942 (0.720)	-0.711 (0.603)	-0.340 (0.602)
Science	8.450** (3.019)	-3.421 (3.802)	-10.201** (3.844)	-6.740* (3.026)	-8.972** (3.023)	-8.763** (3.020)
PolitEcon	-0.587 (0.630)	-2.816** (1.077)	0.065 (0.819)	0.644 (0.644)	0.689 (0.633)	0.868 (0.631)
Science \times Religion	-10.632* (4.149)	5.592 (4.980)	13.784** (4.849)	7.949+ (4.265)	7.445+ (4.153)	8.793* (4.151)
Science \times PolitEcon	-12.369* (5.627)	6.696 (6.792)	13.744+ (7.302)	9.831+ (5.664)	15.022** (5.644)	14.360* (5.630)
Religion \times PolitEcon	1.489 (1.049)	5.430** (1.840)	2.210 (1.342)	1.735 (1.142)	0.683 (1.075)	0.216 (1.058)
<i>Interaction with Industry</i>						
Science	-11.874** (3.733)	9.716+ (5.150)	14.239** (4.726)	9.591* (3.763)	12.709*** (3.739)	12.344** (3.736)
PolitEcon	-1.237 (2.880)	9.464* (3.834)	2.247 (3.042)	1.447 (2.901)	2.280 (2.882)	1.763 (2.880)
Science \times Religion	5.931 (8.113)	-6.376 (9.363)	-13.692 (8.740)	-5.271 (8.440)	-3.627 (8.123)	-5.104 (8.119)
Science \times PolitEcon	43.306 (28.702)	-45.647 (29.929)	-43.803 (29.382)	-40.796 (28.735)	-43.884 (28.708)	-43.498 (28.704)
Religion \times PolitEcon	-0.517 (4.960)	-12.091+ (7.051)	-3.116 (5.336)	-4.515 (5.059)	0.544 (4.970)	0.172 (4.973)
Num.Obs.	162 236					
R2	0.190					

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Volumes are placed into 20 year ((+/-) 10 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.1: Distribution of Volumes, with 20-year smooth

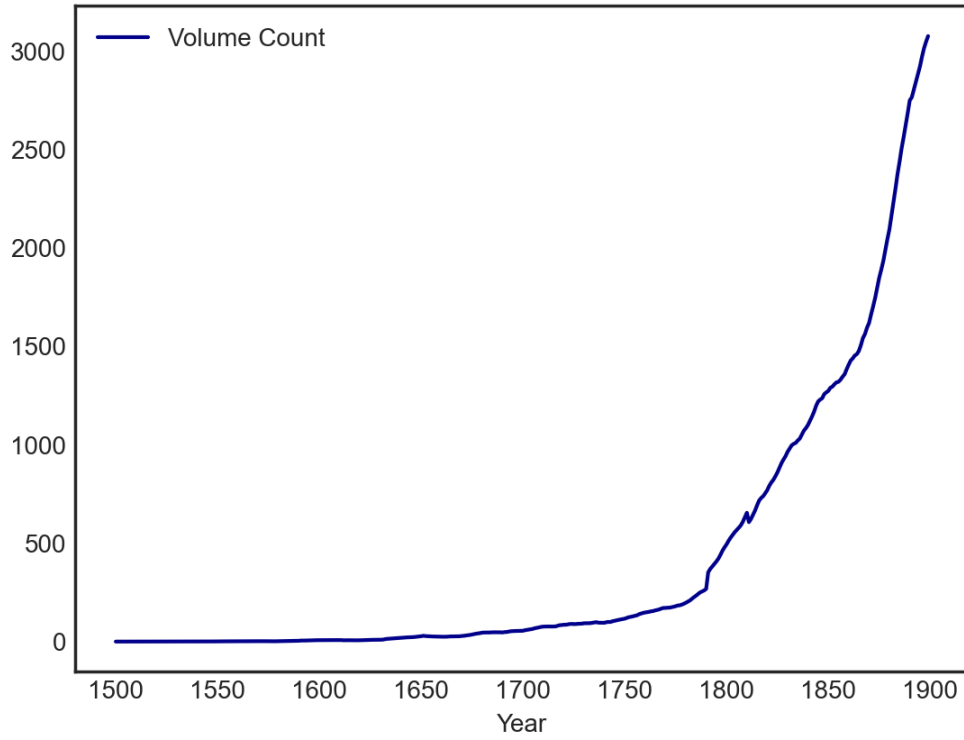


Figure B.2: Model Selection and Topic Optimization

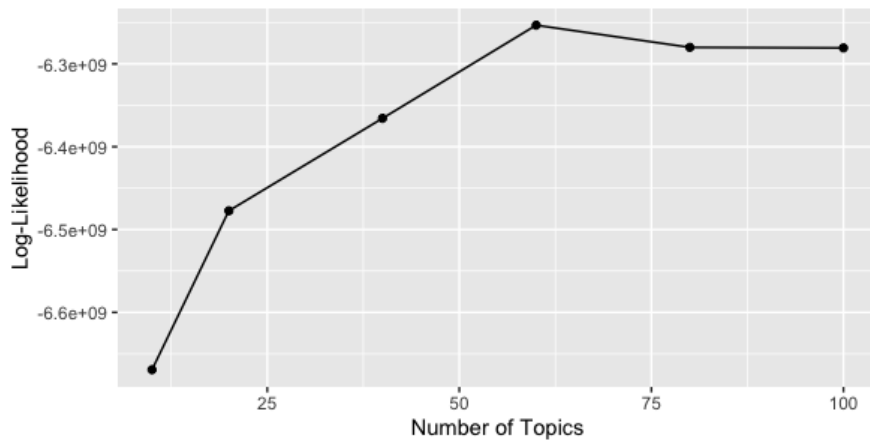


Figure B.3: Distribution of Categories over time, translations and non-translations

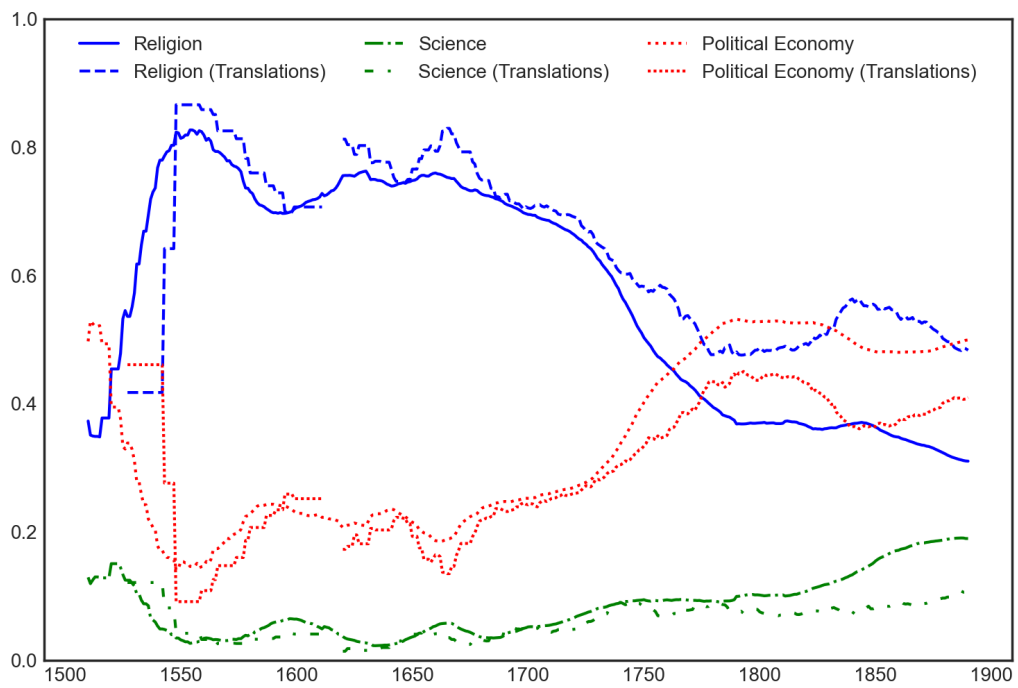
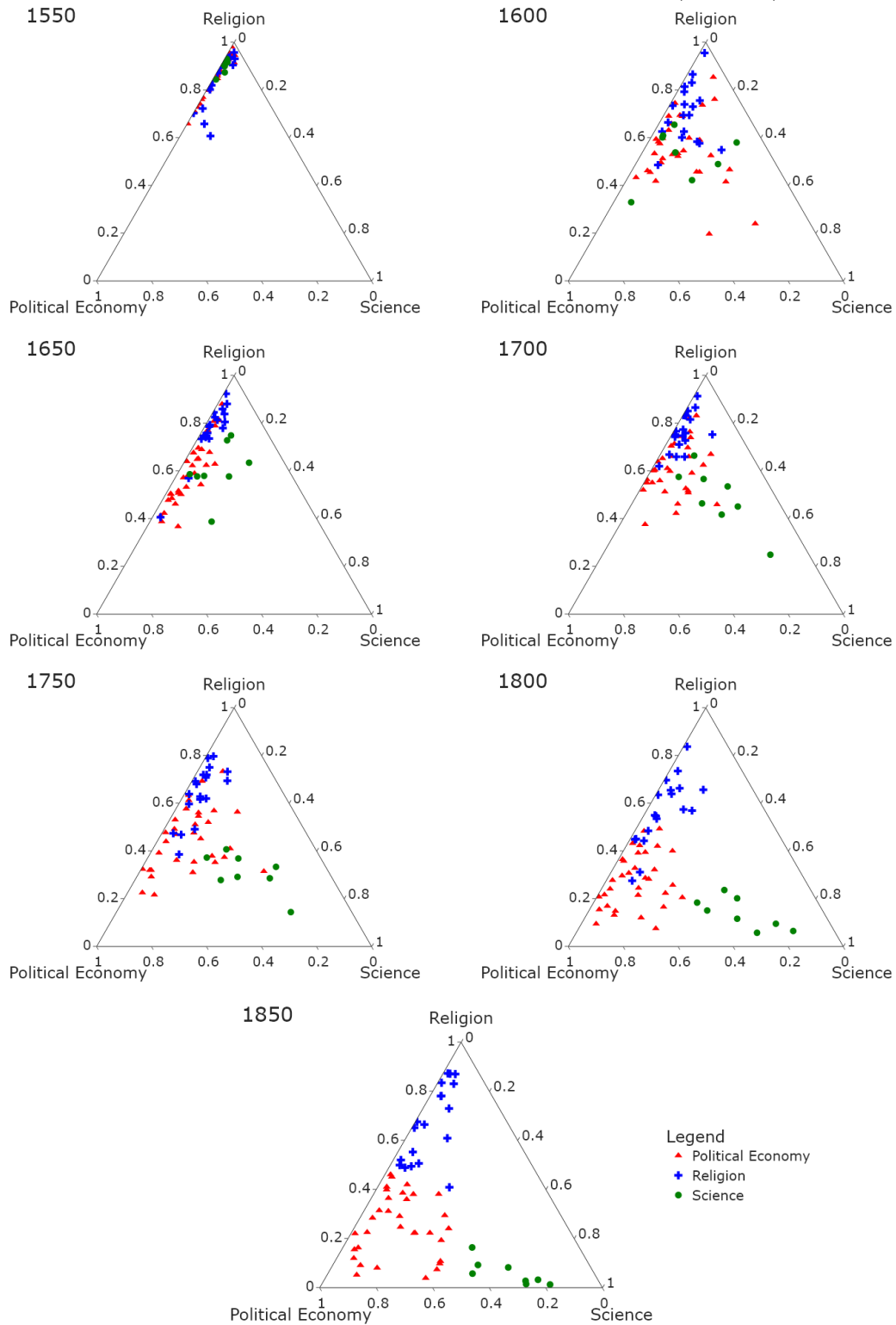


Figure B.4: Topics by Category, 1550–1850 (in color)



Note: Categorization into “Science”, “Political Economy” or “Religion” based on topics’ placement in 1850.

Figure B.5: Selected Famous Volumes Categorized (in color)

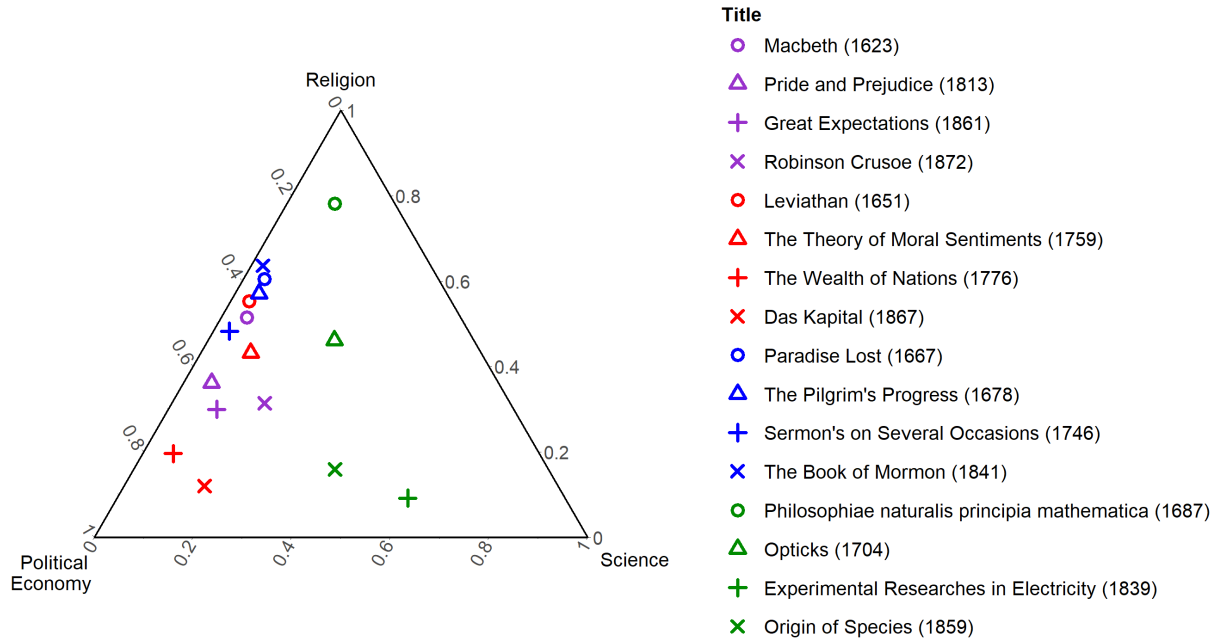


Figure B.6: Average Progress Score (percentile), 1500–1900, using 20-year moving average

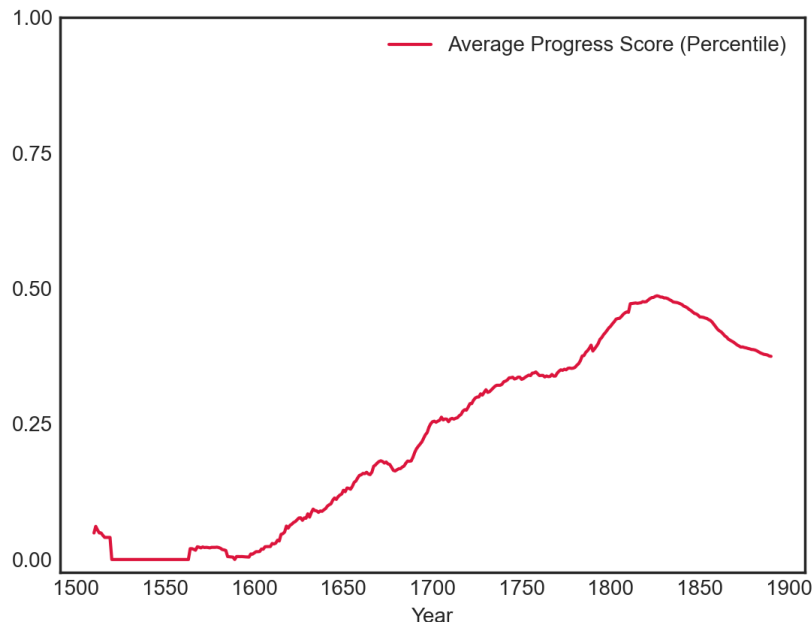


Figure B.7: Average Progress Score (percentile), translations and non-translations

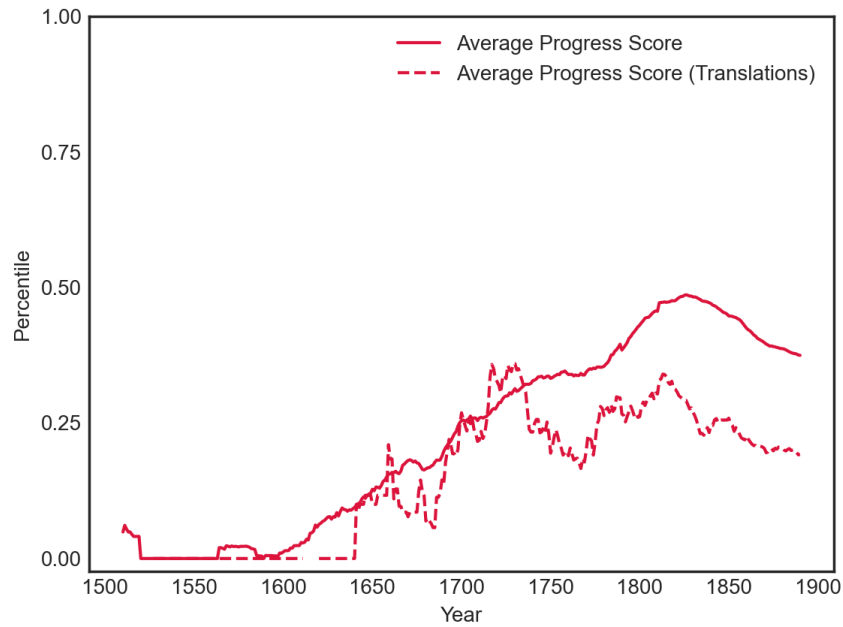


Figure B.8: Google n-grams of “Progress”

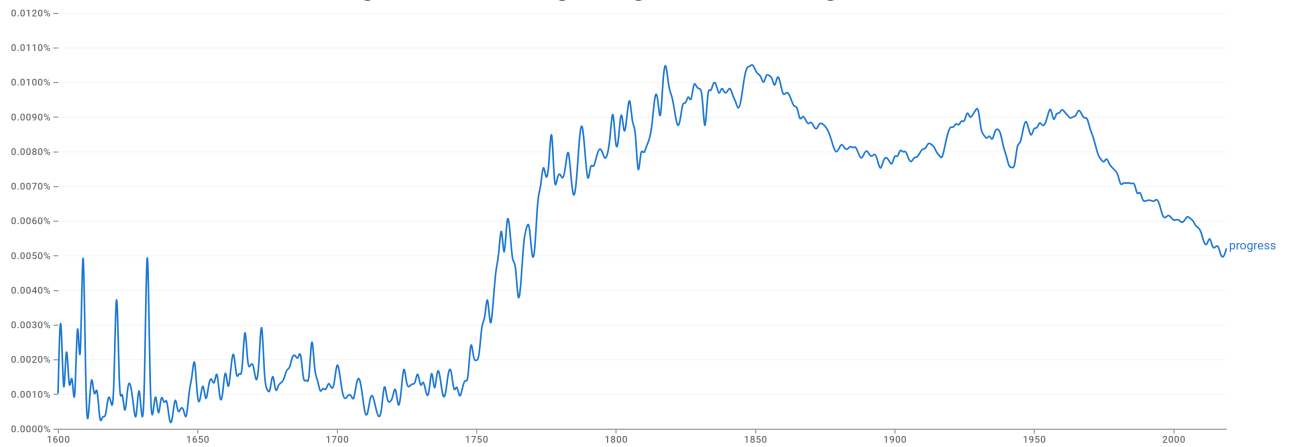
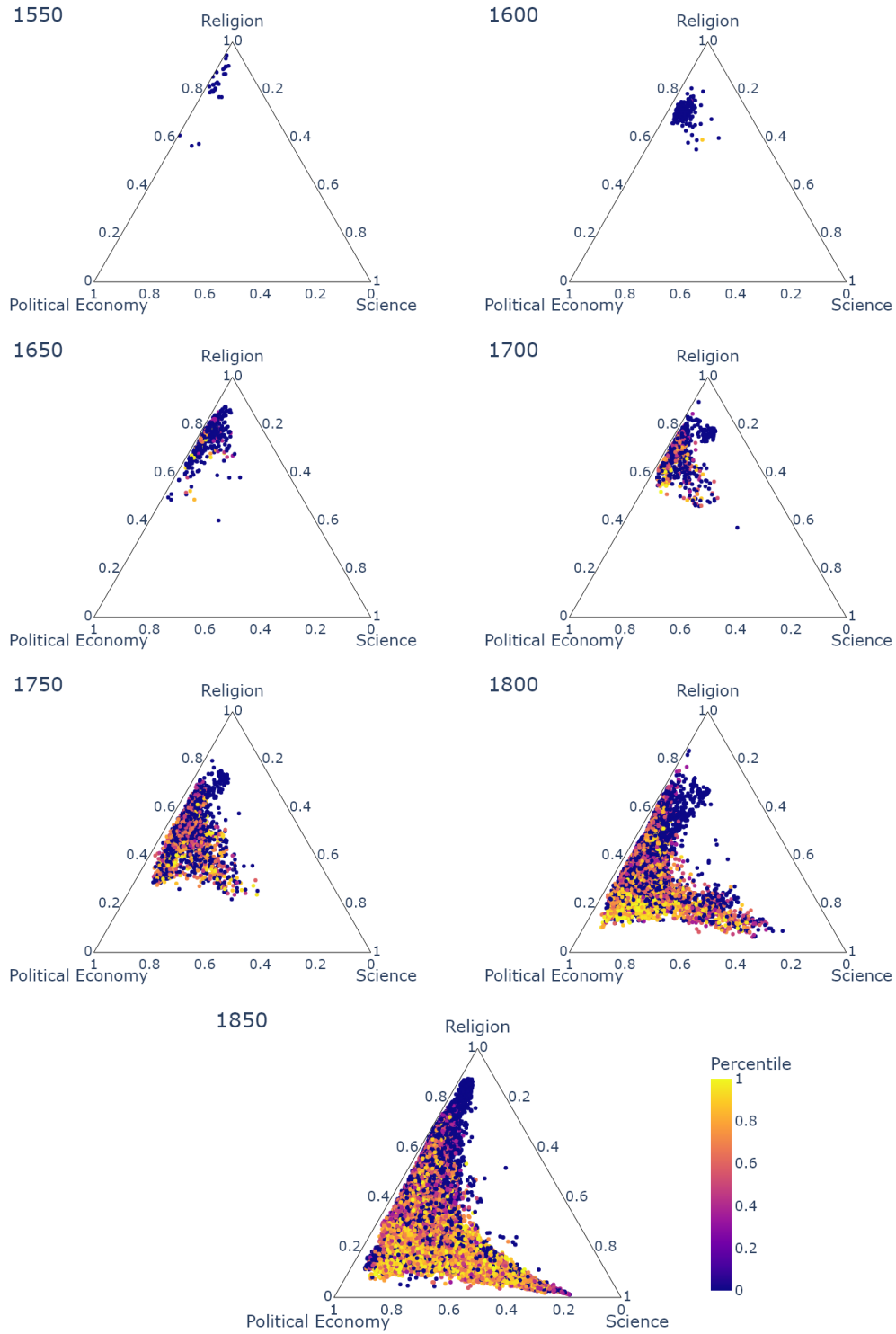


Figure B.9: Progress Sentiment, 1550–1850, words first used after 1643 included



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

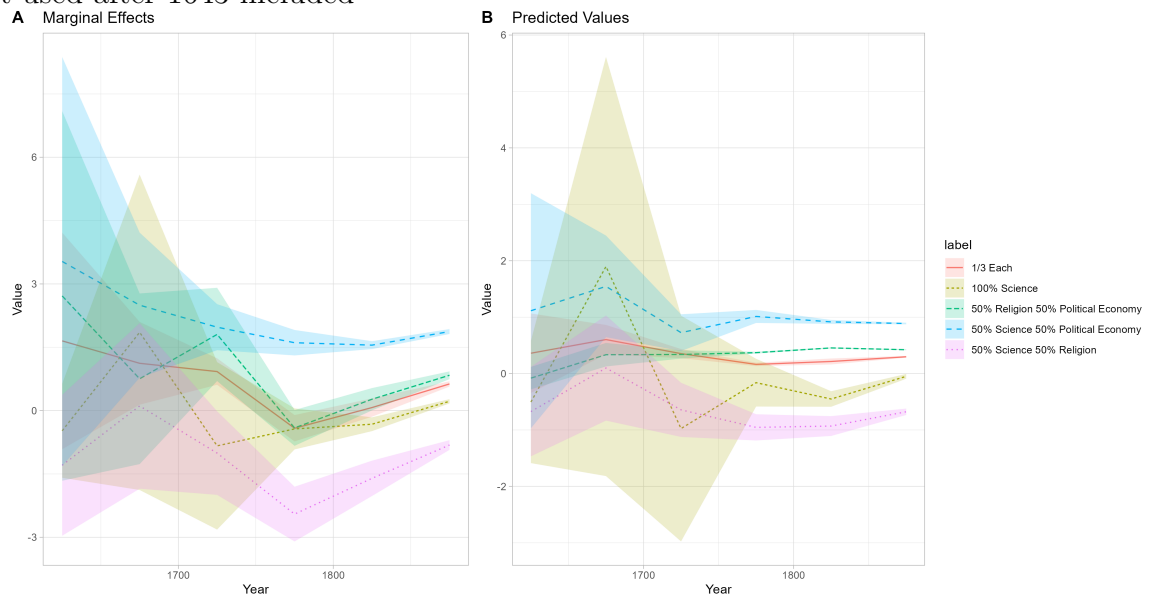
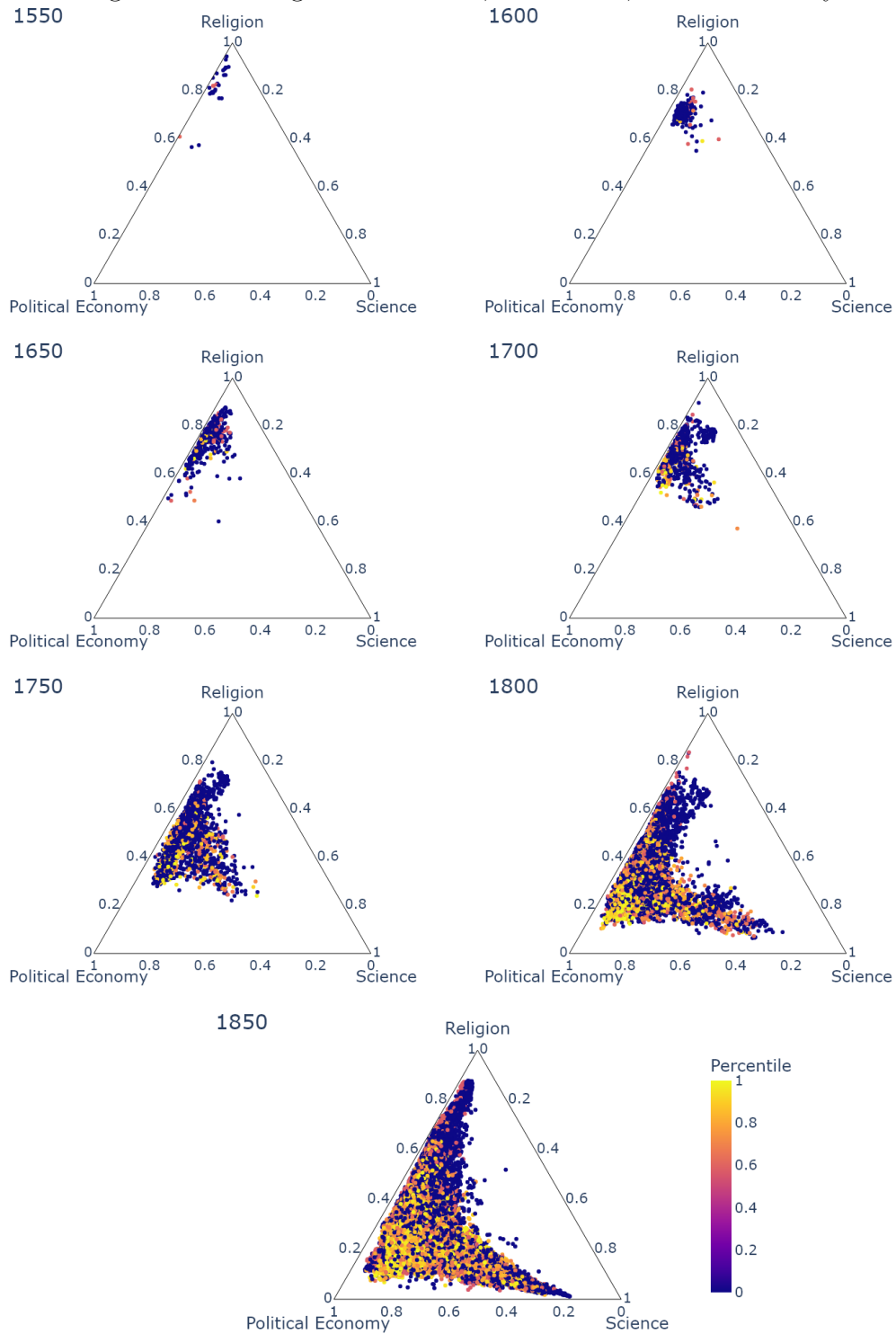


Figure B.11: Progress Sentiment, 1550–1850, 1708 Dictionary



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.12: Marginal Effects and Predicted Values, Progress Sentiment Regressions, 1708 Dictionary

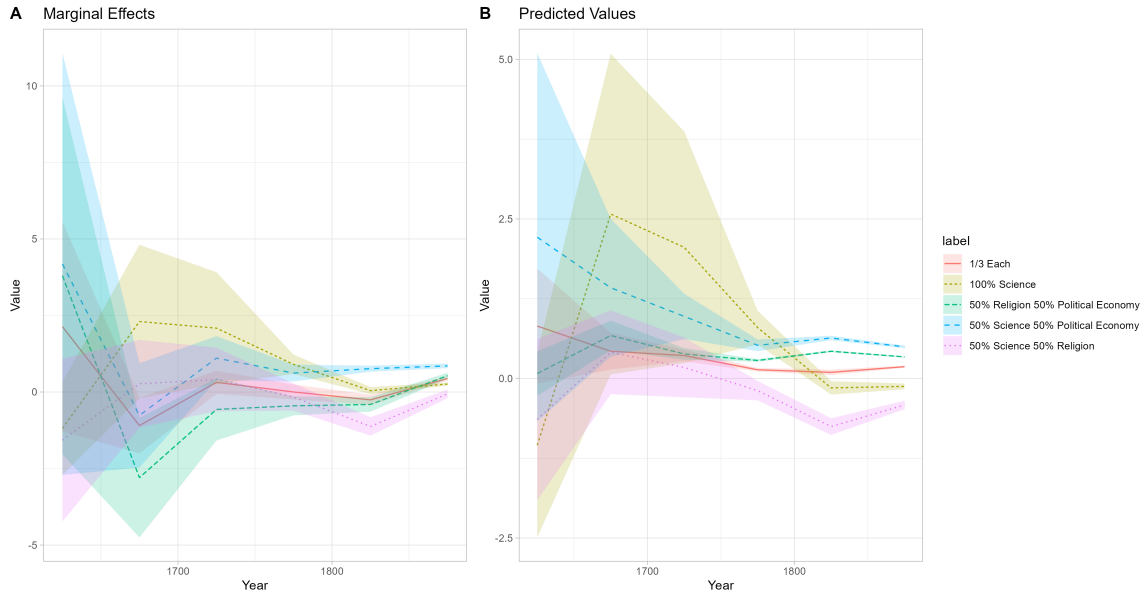


Figure B.13: Marginal Effects and Predicted Values, Progress Sentiment Regressions, dropping observations prior to 1650

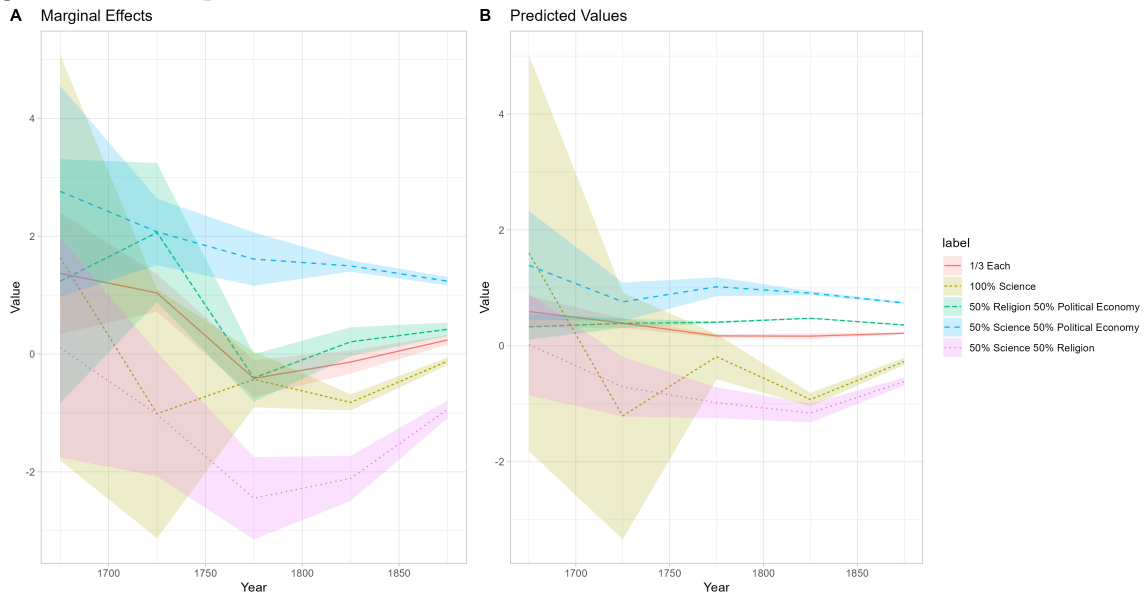
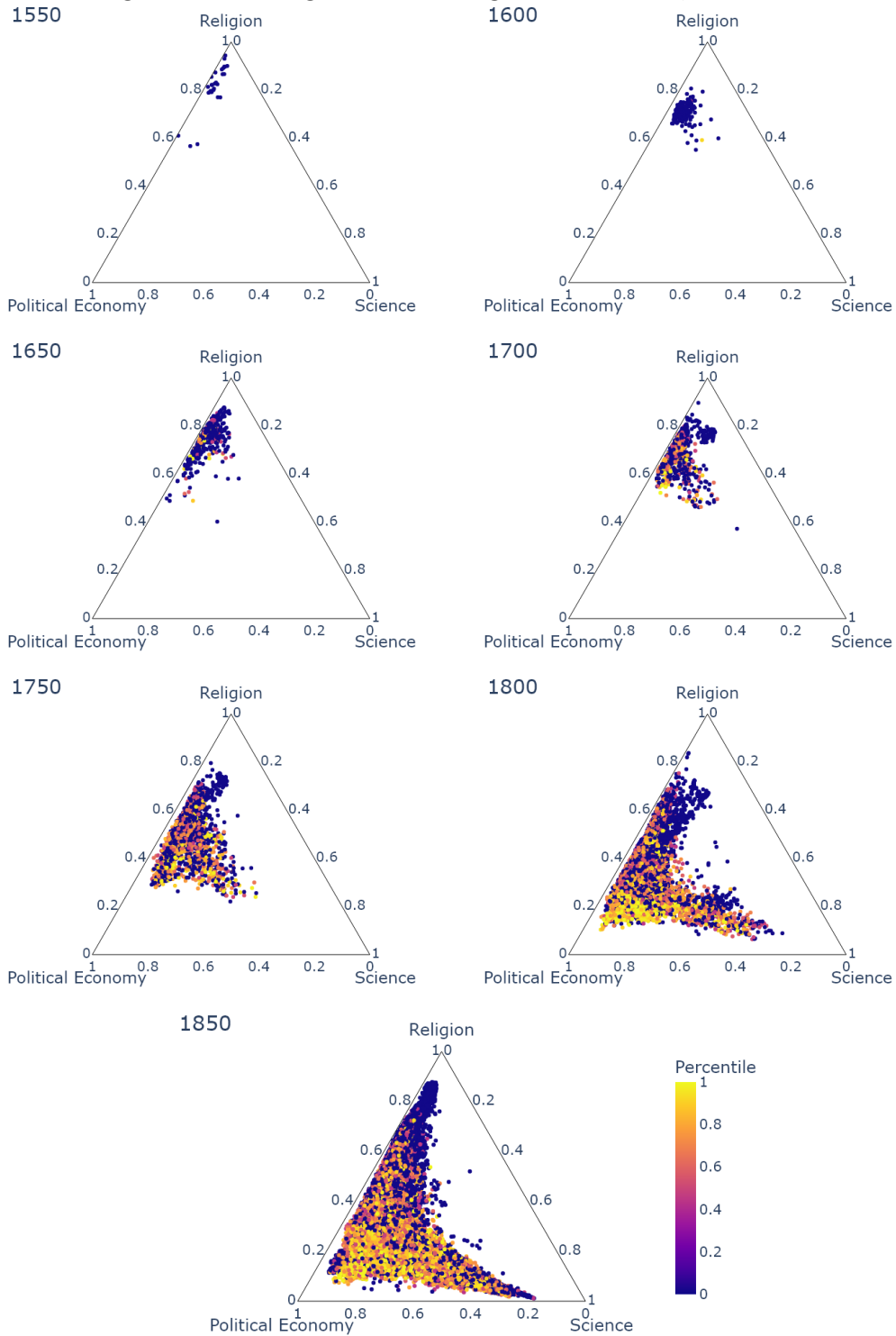
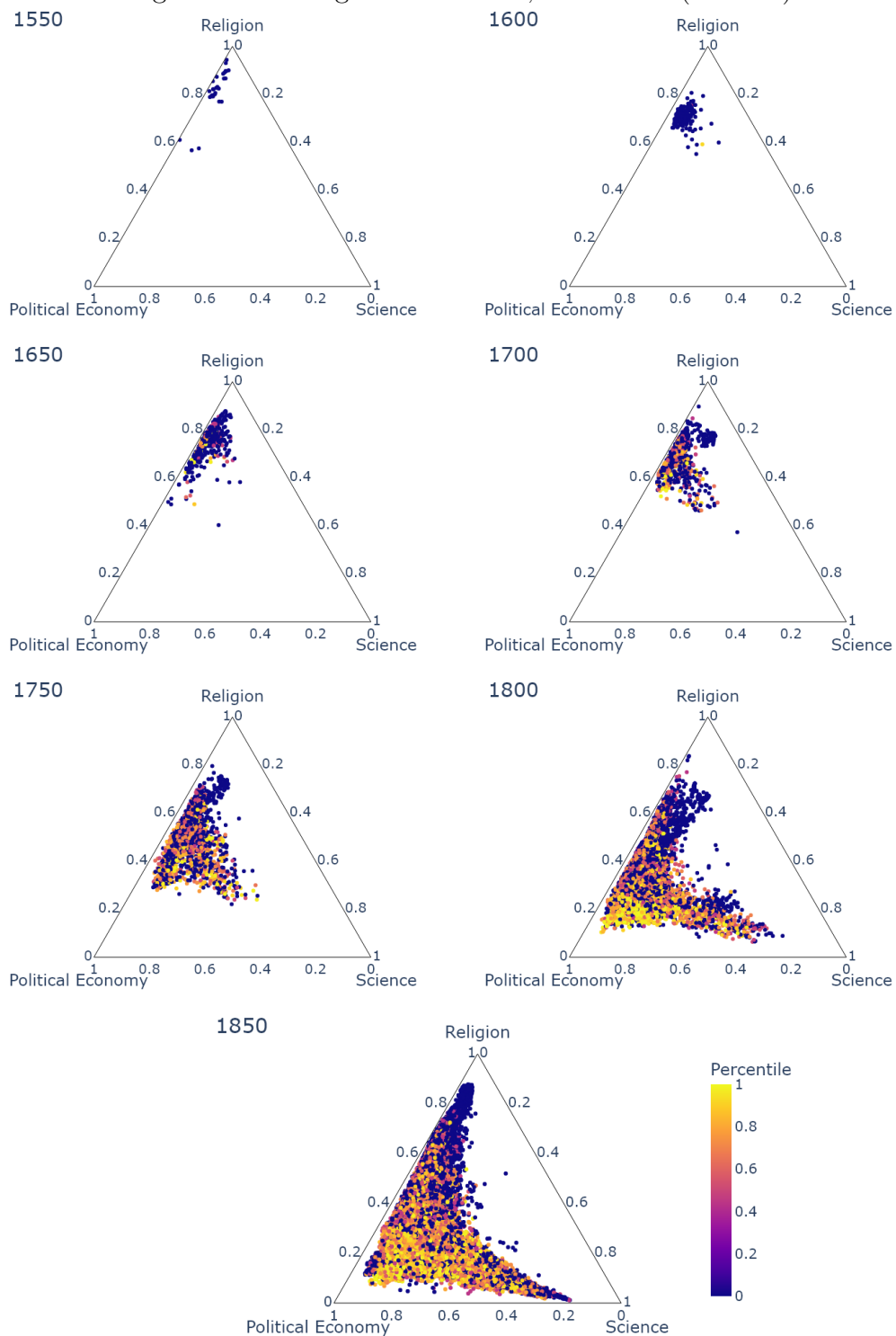


Figure B.14: Progress minus Regress Sentiment, 1550–1850



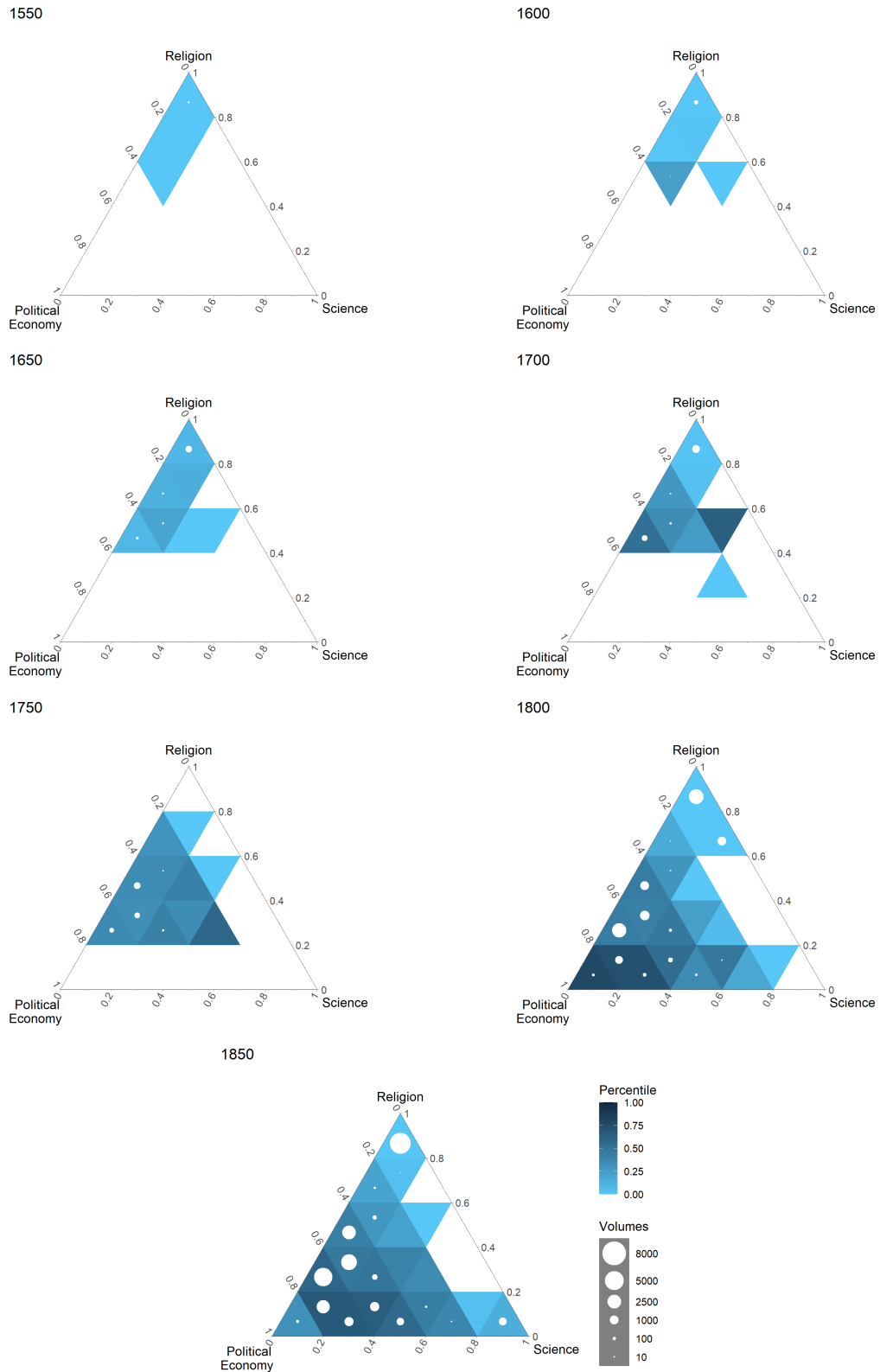
Note: Each dot represents a volume. For each year, all volumes +/- 10 years are included (i.e., for 1800 all volumes from 1790 to 1810 are included). The color of each dot represents the progress sentiment subtracted by the regress sentiment of that volume, with lighter colors representing greater sentiment.

Figure B.15: Progress Sentiment, 1550–1850 (in color)



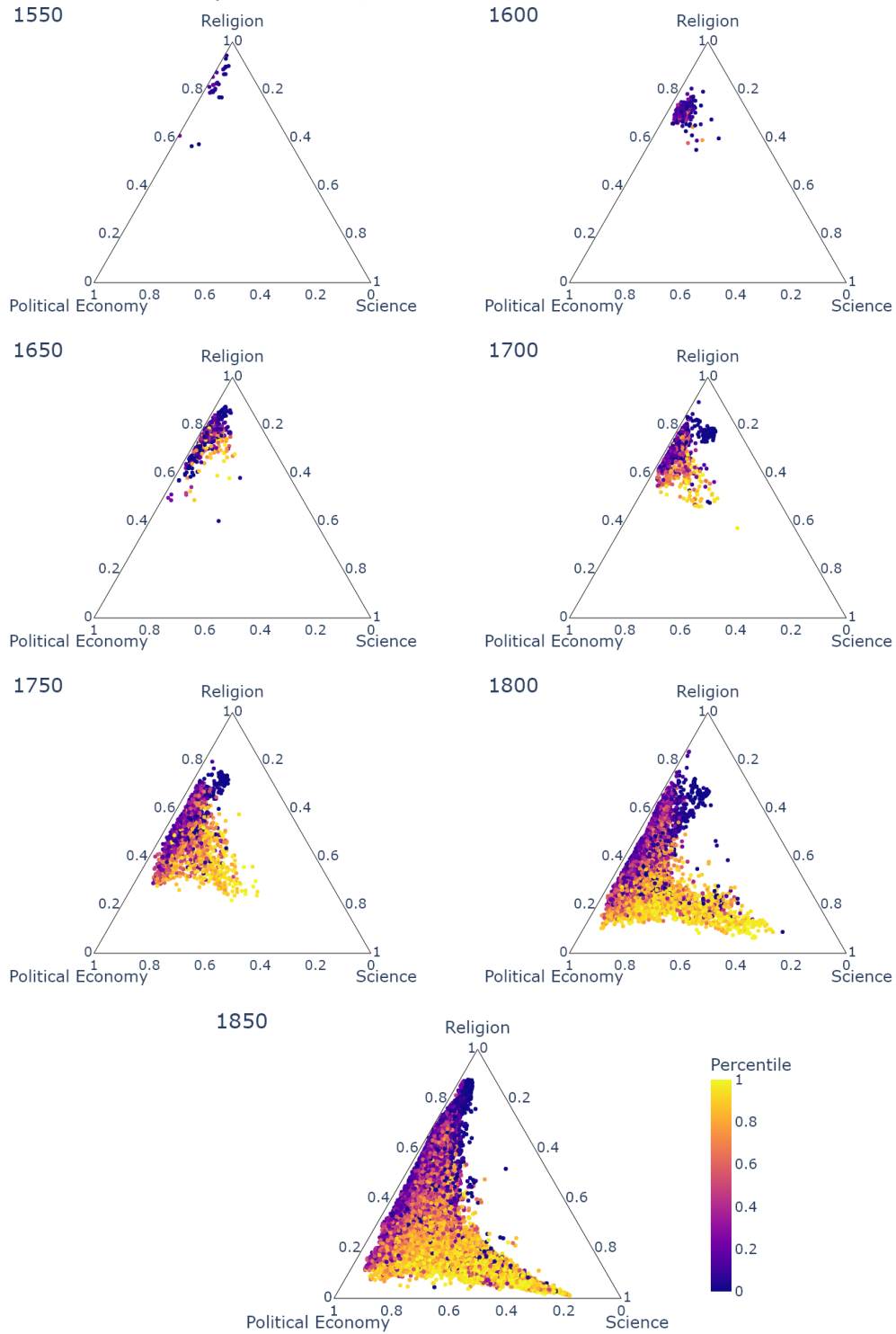
Note: Each dot represents a volume. For each year, all volumes +/- 10 years are included (i.e., for 1800 all volumes from 1790 to 1810 are included). The color of each dot represents the sentiment of that volume, with lighter colors representing more progressive sentiment. A grayscale version is available in Figure 6.

Figure B.16: Progress Sentiment in Triangles



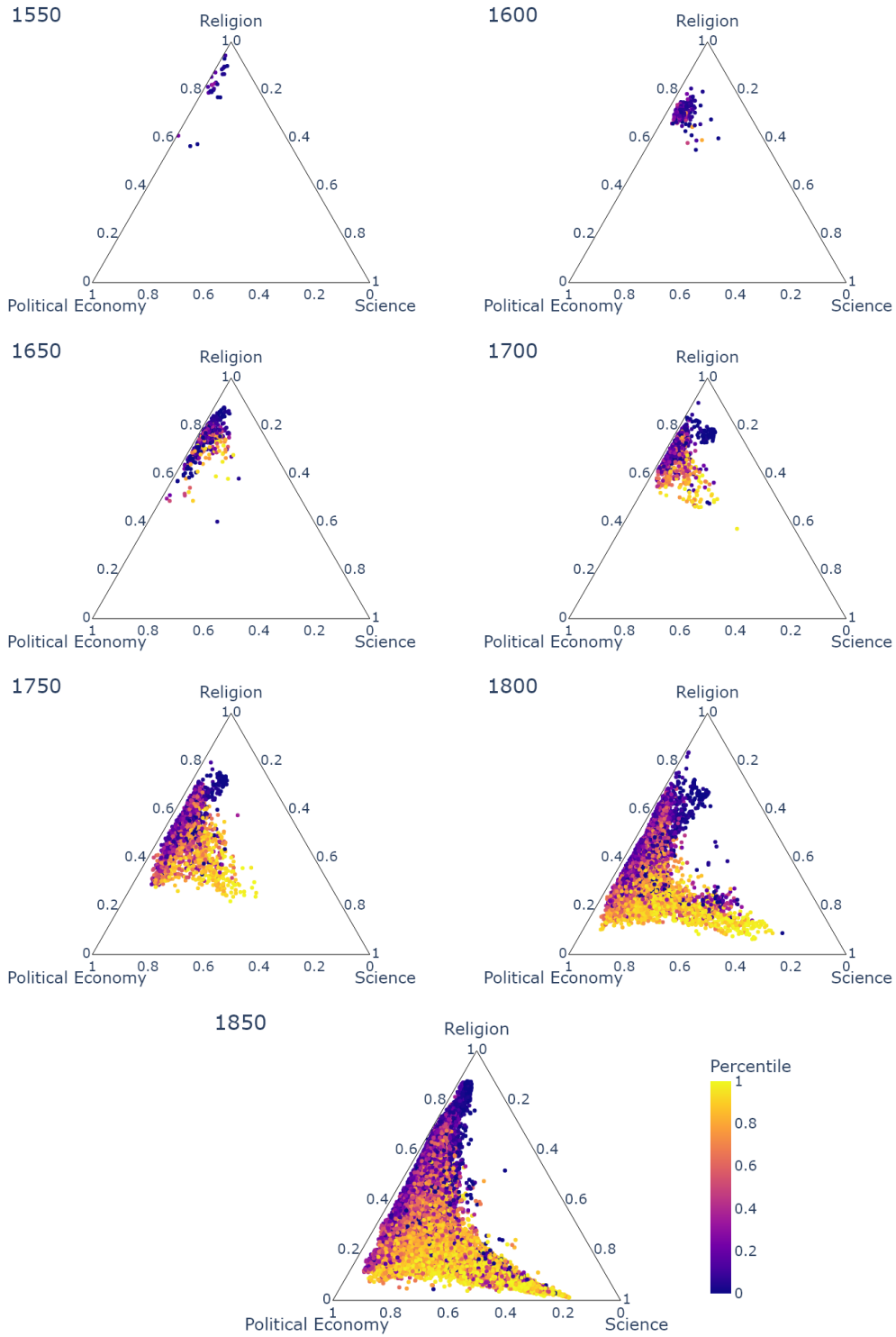
Note: Each sub-triangle shows the average sentiment by percentile of all volumes that fall within that sub-triangle. The size of the white dot in each sub-triangle represents the amount of volumes published within the sub-triangle. For each year, all volumes +/- 10 years are included (i.e., for 1800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

Figure B.17: Industry Sentiment, 1550–1850, words first used after 1643 included



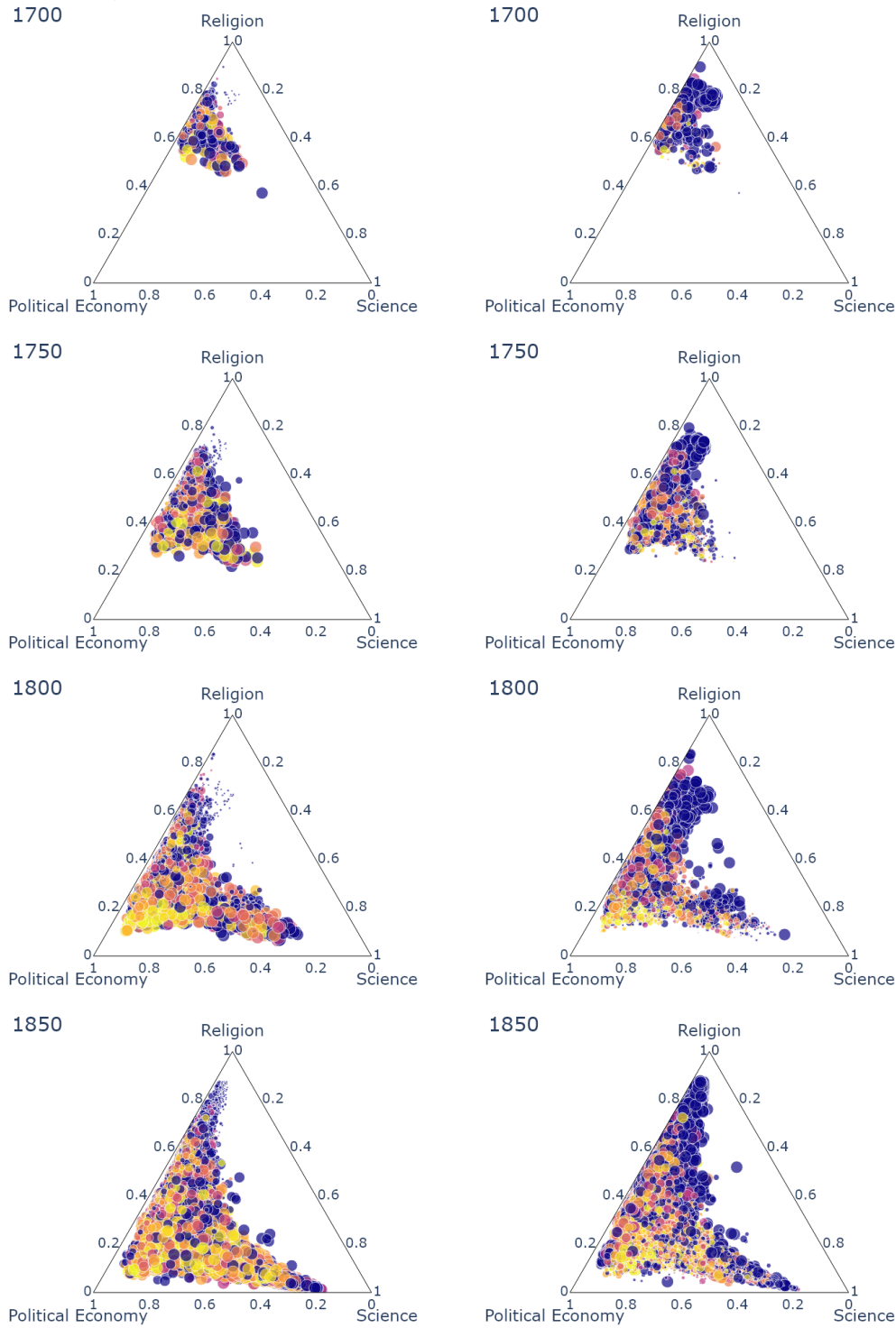
Note: Each dot represents a volume. For each year, all volumes +/- 10 years are included (i.e., for 1800 all volumes from 1790 to 1810 are included). The color of each dot represents the sentiment of that volume, with lighter colors representing more industrial sentiment.

Figure B.18: Industry Sentiment, 1550–1850



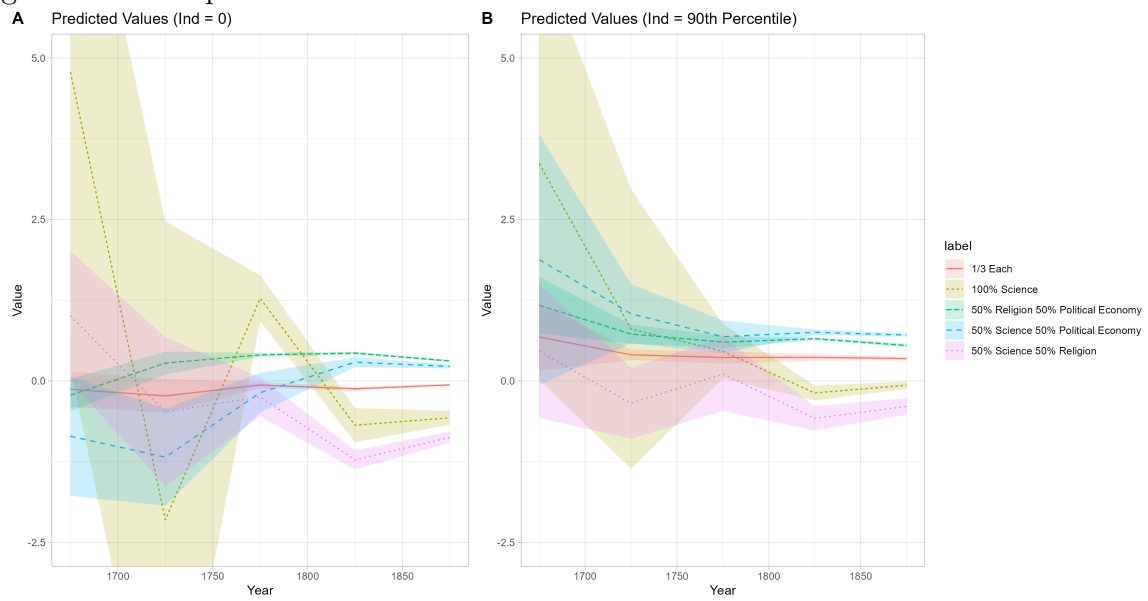
Note: Each dot represents a volume. For each year, all volumes +/- 10 years are included (i.e., for 1800 all volumes from 1790 to 1810 are included). The color of each dot represents the sentiment of that volume, with lighter colors representing more industrial sentiment.

Figure B.19: 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.

Figure B.20: Predicted Values of Progress Sentiment at 0 and 90 Industry Percentile, dropping observations prior to 1650



C Re-analyzing the Data Using the Coherence Score

The coherence score is an alternative to the perplexity score for determining the optimal number of topics (Mimno et al. 2011). To calculate coherence, we use the *UMass-Coherence* score, which measures whether words in a topic tend to co-occur together. The advantages of using the *UMass-Coherence* score is twofold. The first is that it is well-suited for our bag-of-words representation of documents, as it depends on document-level occurrence rather than a sliding window distance at the sentence level. The second is that the *UMass-Coherence* score computes these counts over the original corpus used to train the topic models, rather than an external corpus. This suggests that this metric is more intrinsic in nature, as it attempts to confirm that the models learned data known to be in the corpus. The *UMass-Coherence* score is defined as:

$$C_v = \sum_i \sum_{j < i} \log \left(\frac{D(w_j, w_i) + \beta}{D(w_i)} \right)$$

where $D(w_i)$ is the number of documents that contain at least one instance of the word w_i , and $D(w_j, w_i)$ is the number of documents that contain at least one instance of the word w_i and w_j . The equation includes a β smoothing parameter in the numerator to avoid log zero errors. Large negative scores suggest that words in a given topic do not co-occur appear frequently in the same document, while scores closer to zero suggest words tend to co-occur more often.

In our data, the coherence score metric yields 80 topics as the optima. Using the same data processing techniques as in Section 2 yields the topics listed in Table C.1, where four of the topics (13, 35, 44, and 53) consist of characters that are remnants of OCR failures and are therefore omitted.

We proceed to establish categories, as we did in Section 3.1. As before, we begin by excluding topics that are commonly found among categories with high *Incidence* but are broadly used in writing regardless of the subject. These topics—6, 19, 68, and 71—use words commonly found in literature. The three categories with the highest *Incidence* score that do not have any of these topics are {8,28,43}, {8,28,33}, and {8,28,69}. Since topics 8 and 28 are clearly religious topics, we matched it with a topic that is also a religion topic. Topic 43 has to do with law, and topic 33 contains the root word “scienc,” so neither of these work to complete the category. Topic 69, however, contains both “religi” and “religion,” so we chose {8,28,69} as our first category. We proceed to seek categories with the highest *Incidence* score that do not contain topics from the omitted topic list or 8, 28, or 69. This yields the category {2,43,67}, which appears to be, broadly speaking, related to political economy. We

Table C.1: Topics Using Coherence Score

1	quot	chines	sanskrit	worship	hindu	japanes	languag	buddha
2	trade	amount	money	labour	cent	increas	price	bank
3	court	offic	counti	compani	aforesaid	appoint	board	council
4	court	defend	plaintiff	estat	bill	properti	contract	action
5	fame	fee	cafe	fever	fay	fet	ufe	feem
6	tom	cri	gentleman	jack	boy	captain	miss	fellow
7	gold	colour	plate	white	black	silver	print	take
8	god	christ	holi	faith	jesus	sin	christian	heaven
9	parliament	bill	ireland	majesti	irish	duke	committe	vote
10	church	bishop	christian	holi	pope	christ	rome	cathol
11	water	air	produc	lower	cover	show	posit	increas
12	tho	street	railway	ditto	messr	liverpool	esq	manchest
13	OMITTED							
14	boil	wine	salt	sugar	butter	mix	pound	milk
15	india	chines	nativ	china	bengal	govern	british	indian
16	agus	ari	mid	dia	ami	swa	izay	mac
17	ship	island	captain	sea	vessel	sail	board	coast
18	indian	america	american	island	coloni	canada	british	york
19	conduct	happi	occas	societi	affect	enjoy	consequ	influen
20	sun	electr	earth	magnet	current	motion	moon	heat
21	adj	lat	sax	latin	verb	adv	dryden	signifi
22	acid	solut	carbon	heat	sulphur	gas	oxid	precipit
23	line	equal	angl	sin	equat	plane	circl	parallel
24	fish	hors	bird	anim	food	fli	tail	dog
25	thi	thou	israel	hath	ver	jew	david	jerusalem
26	coloni	south	africa	nativ	cape	australia	chief	zealand
27	don	por	para	del	spain	spanish	como	sus
28	love	thi	heart	thou	soul	god	lord	earth
29	virtu	religion	punish	mankind	pretend	fame	esteem	likewis
30	genus	brown	black	margin	fig	speci	shell	pale
31	vol	lond	fol	folio	calf	copi	pari	par
32	thou	thi	hath	doth	duke	nay	pray	scene
33	exist	human	refer	distinct	idea	moral	sens	principl
34	welsh	efe	wale	oedd	gan	arglwydd	fel	wrth
35	OMITTED							
36	roman	greek	rome	athen	greec	caesar	senat	cicero
37	franc	french	pari	duke	emperor	loui	spain	germani
38	arab	egypt	greek	sultan	turk	russian	turkish	persian
39	diseas	patient	treatment	blood	medic	pain	fever	oper
40	church	build	ancient	stone	built	erect	wall	citi

Table C.1: Topics Using Coherence Score (cont.)

41	tlie	hut	tliat	tho	tlic	ava	tlio	lii
42	school	colleg	societi	educ	rev	instruct	institut	committe
43	law	public	evid	act	respect	lord	suppos	question
44	OMITTED							
45	che	della	dell	del	gli	nel	piÃ ¹	signor
46	saxon	norman	britain	anglosaxon	roman	dane	harold	alfr
47	kai	kal	verb	tov	comp	yap	greek	latin
48	armi	enemi	command	march	french	attack	captain	regiment
49	plant	soil	garden	cultiv	grow	tree	fruit	seed
50	tbe	arc	ihe	tin	tor	ano	lit	ofth
51	cic	eft	vel	quod	cum	idem	vide	plin
52	writ	car	pur	cafe	dit	ceo	statut	fait
53	OMITTED							
54	scotland	ireland	pop	irish	counti	dublin	castl	lat
55	bot	stem	genus	calyx	linn	corolla	flower	plant
56	von	und	ber	der	unb	herr	ein	bie
57	esq	jan	oct	dec	nov	feb	aug	juli
58	parish	thoma	esq	counti	manor	richard	robert	rev
59	edit	cloth	crown	illustr	vol	svo	rev	post
60	ladi	repli	dear	miss	madam	woman	husband	daughter
61	paint	pictur	painter	artist	rome	florenc	portrait	italian
62	hym	kyng	hem	sayd	all	tyme	ben	ther
63	fig	surfac	develop	structur	anterior	posterior	organ	bone
64	river	road	town	north	valley	south	mountain	lake
65	coal	geolog	rock	limeston	surfac	clay	occur	stratum
66	cum	quod	est	sed	quam	qui	aut	hoc
67	govern	nation	war	england	establish	constitut	civil	foreign
68	thi	thou	sweet	heaven	bright	song	smile	breath
69	moral	social	religi	christian	influnc	develop	polit	popular
70	men	ofth	ing	hath	con	tho	sor	shew
71	mother	father	knew	told	look	miss	room	love
72	henri	william	earl	thoma	edward	john	duke	jame
73	quod	cum	regi	vel	anno	domini	rex	apud
74	qui	par	nous	vous	tout	sur	une	quil
75	hath	doe	doth	bee	wee	own	one	again
76	iron	engin	fig	steam	pressur	construct	inch	machin
77	poet	music	play	poetri	poem	johnson	theatr	literari
78	sir	letter	lord	write	john	honour	friend	send
79	morn	arriv	visit	travel	even	river	journey	beauti
80	king	princ	citi	queen	war	armi	court	kingdom

Note: only the first eight roots are included in this table due to space constraints. Topics eliminated during category selection: {6,19,68,71}

proceed to seek a third category that does not contain topics from the omitted topic list or the other two categories. This yields the category {11,20,76}, which is clearly science-related. These categories (labeled manually) and their topics produced by this process are presented in Table C.2.

Table C.2: Categories, using Coherence Score

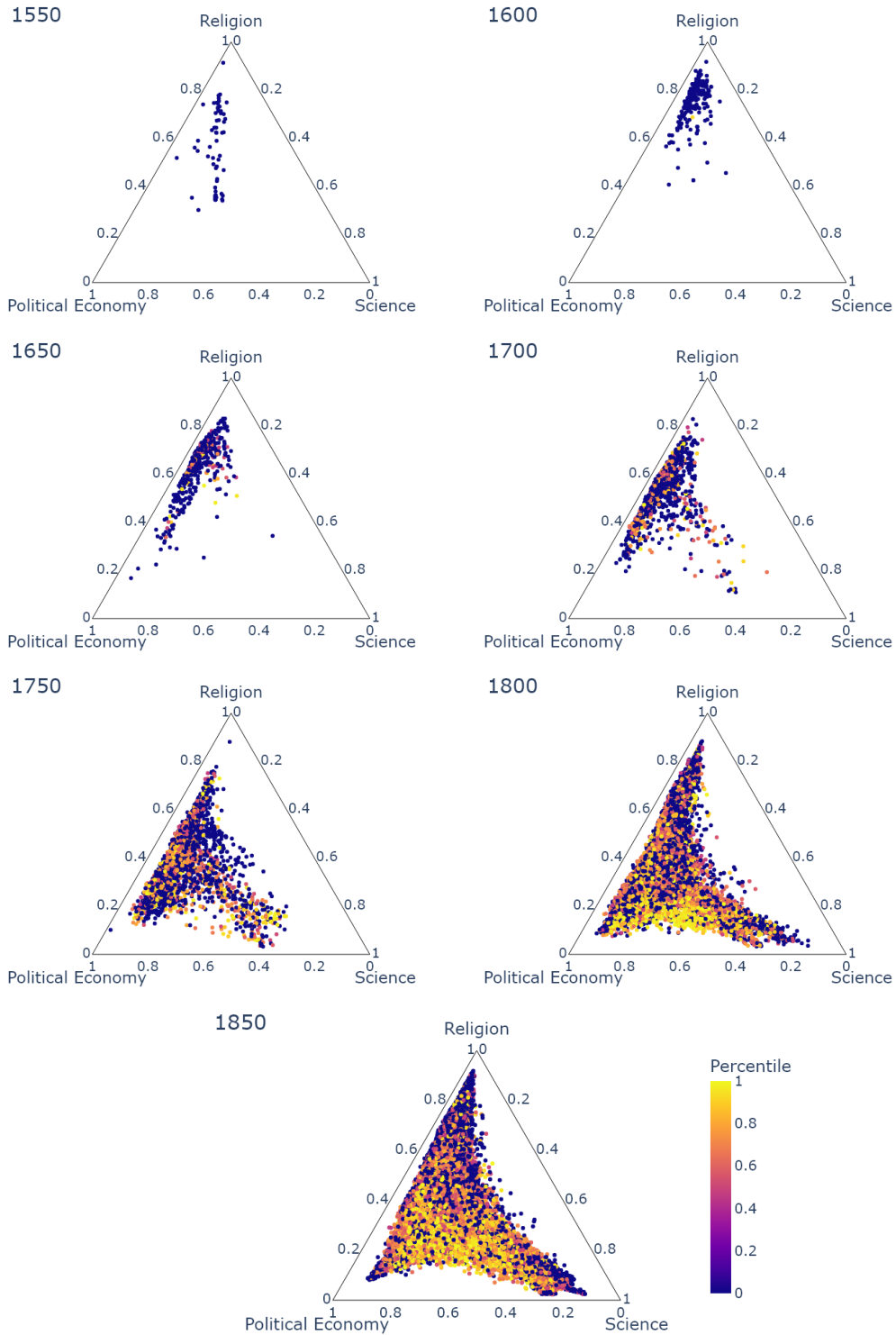
Category	Topics and associated words
“Political Economy”	2 - trade amount money labour cent increas price bank capit rate
	43 - law public evid act respect lord suppos question case son
	67 - govern nation war england establish constitut civil foreign british english
“Religion”	8 - god christ holi faith jesus sin christian heaven divin hath
	28 - love thi heart thou soul god lord earth heaven father
	69 - moral social religi christian influenc develop polit popular centuri educ
“Science”	11 - water air produc lower cover show posit increas employ effect
	20 - sun electr earth magnet current motion moon heat surfac distanc
	76 - iron engin fig steam pressur construct inch machin heat cylind

Note: only the first ten roots are included in this table due to space constraints.

We proceed, as in Sections 3.2 and 3.3, to place each topic and volume in the simplex. Each volume has the same progress-oriented sentiment score as in Section 3.3, since we have not changed the progress dictionary. Here, we simply replicate Figure 6, which shows how volume sentiment changed over time within the simplex. The results are available in Figure C.1.

The primary results hold using the coherence score metric. First, as in the exercise presented in the body of the paper, there is a clear trend throughout the time period whereby the languages of science and religion became increasingly distinct. Second, as before, it appears that volumes published along the political economy-science axis became increasingly progress-oriented over time. This finding is supported by Figure C.2, which plots the marginal effects of *Science* and the predicted sentiment of volumes with varying weights of science, political economy, and religion. Even more so than in Figure 7, it appears that the marginal effect and predicted values of progress sentiment are much higher at the political economy-science nexus than anywhere else. The spike in sentiment appears (according to predicted values) to have been strongest in the 18th century, similar to our finding in the body of the paper. Also like Figure 7, these results suggest that works of “pure” science were not necessarily more progress-oriented.

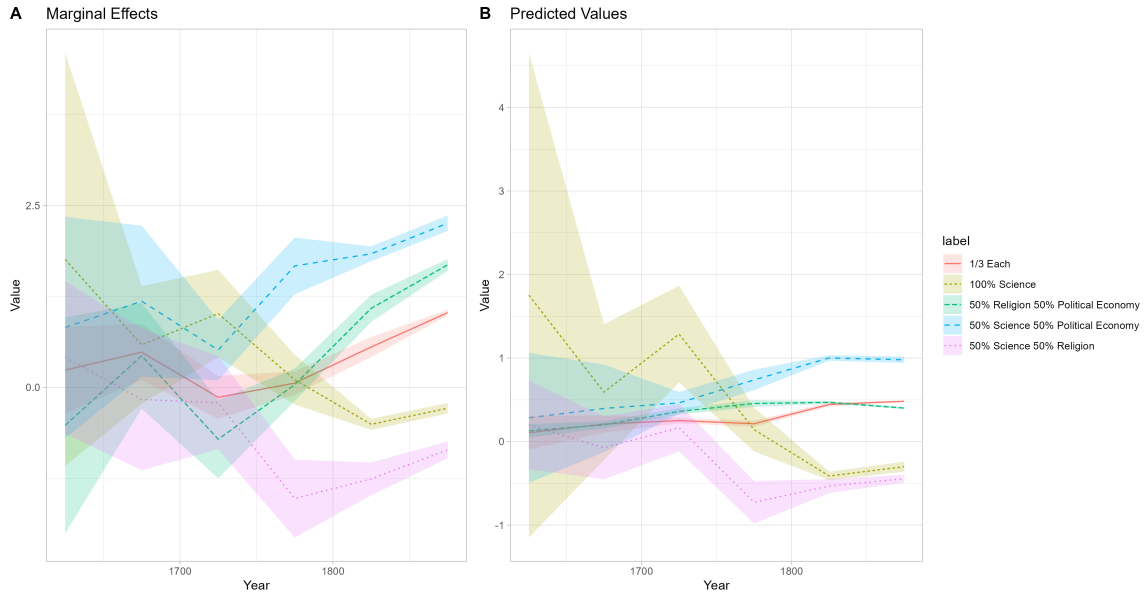
Figure C.1: Progress Sentiment, 1550–1850, using Coherence Score



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 yellow shades representing more progressive sentiment.

In short, using the coherence score instead of perplexity score to determine the optimal number of topics barely changes the results. If anything, one of the key results shown in Figure C.2 is stronger using this metric.

Figure C.2: Marginal Effects and Predicted Values, Progress Sentiment Regressions



D 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.⁴²

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.

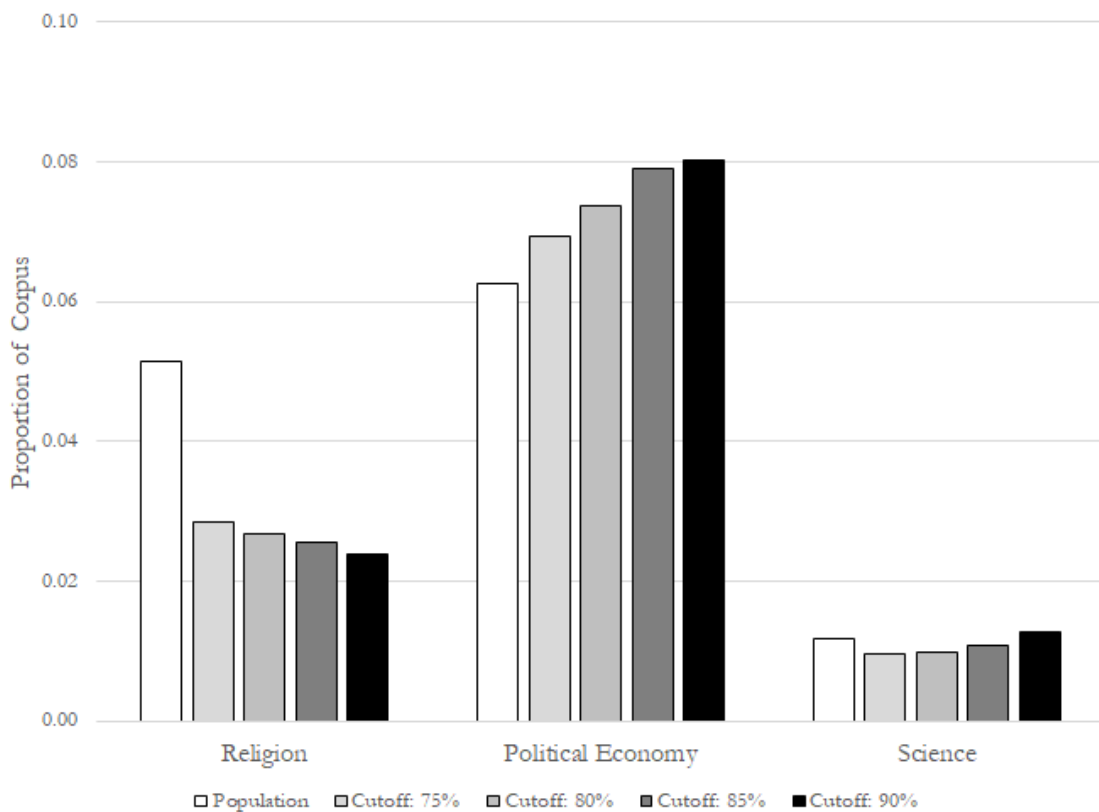
With this dataframe constructed, we can create an intersection and set difference by partitioning the data based on some real number threshold value. For our case we utilize 75%, 80%, 85%, and 90%. That is, a book is included in the intersection of the two data sets if their match is greater than or 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%).

⁴²Since 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 D.1 reports the results.

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

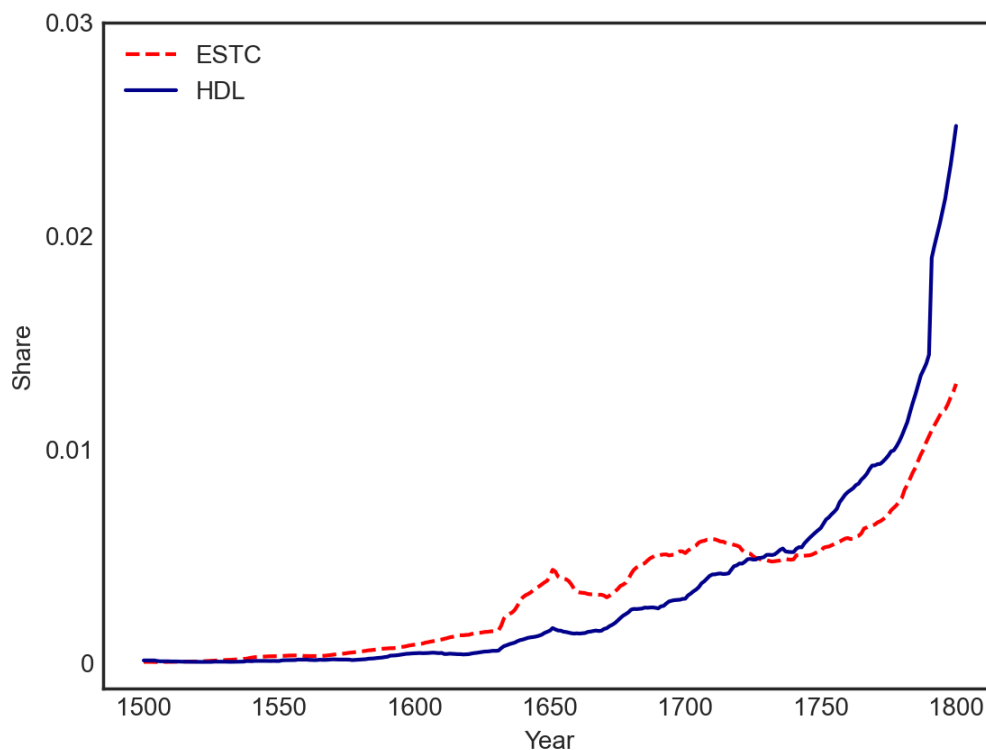


There are several features of this figure that are of interest for the present exercise. First, it appears that the HDL data is *not biased with respect to science*. This is reassuring, as works of science are the primary focus of the paper. However, the HDL data do appear to under-represent religion. 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. 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 D.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 D.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 HDL 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.

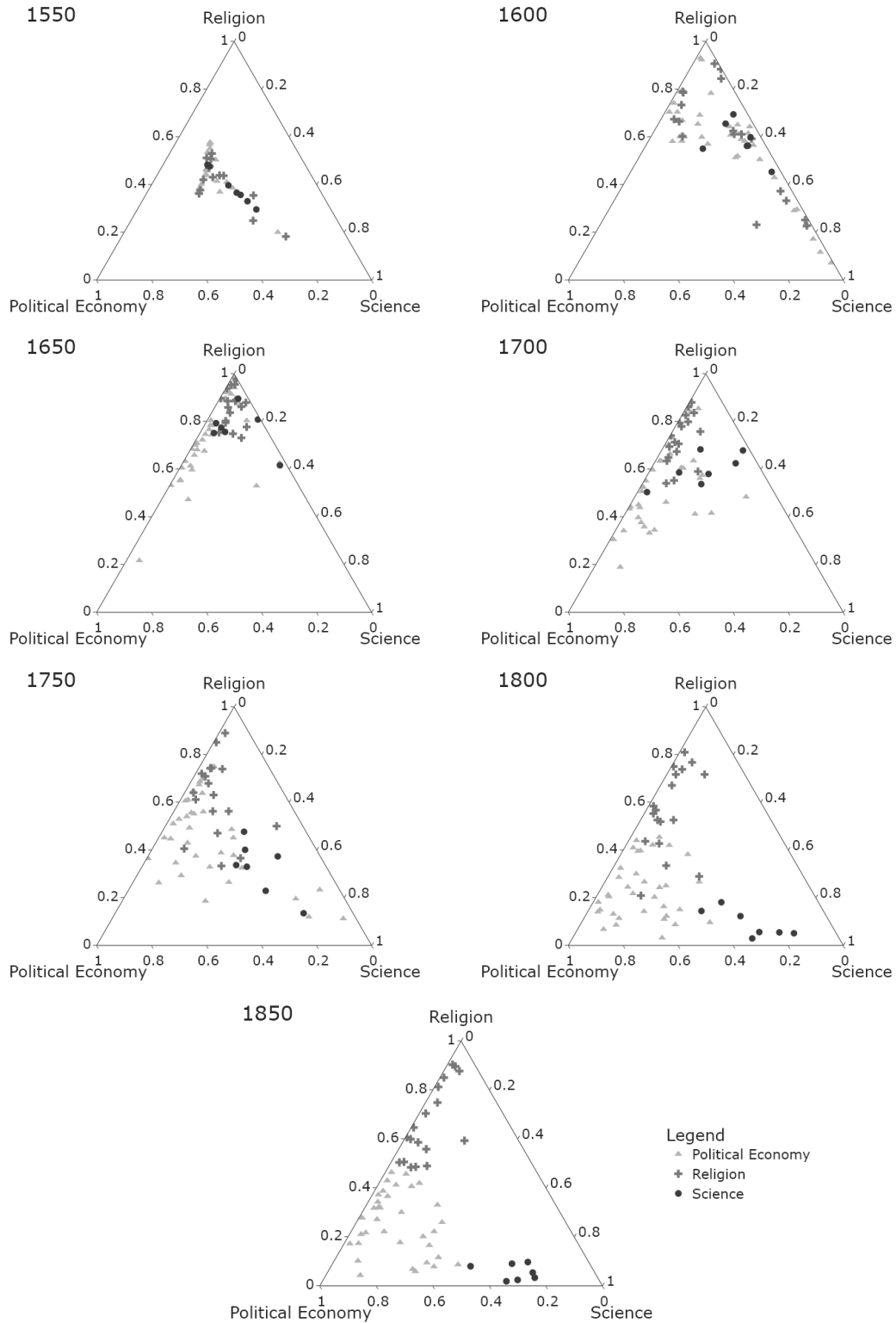
Finally, the exercise described above cannot account for books that are no longer in existence. This is a potential source of bias, particularly if those books were widely read and contributed to the type of language people used at the time. We believe this to be unlikely, however. The heroic efforts 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.”

E Analysis with Unbinned Data

In the analysis presented in the body of the paper, topics are placed in moving 20-year bins. We do this in part due to the low number of volumes in early years, as well as to reduce measurement error in the classification by year. One issue, however, is that binning the data may obscure the *timing* of changes in the sentiment in the corpus. We address this issue in this appendix by presenting the main results when data are unbinned. Below, we re-make Figures 2, 3, 6, 7, 8, 9, and 10 using the unbinned data.

From Figures E.1 and E.3, it appears that some volumes did share the languages of science and religion early in the period (around 1600). However, by the early 18th century there appears to be a clear distinction between the languages of science and religion, similar to what we found in the primary analysis. Figures E.3 and especially E.4 indicate that by far the most progress-oriented volumes were found at the science-political economy nexus, and (from Figure E.4) the rise in progress-oriented language seems particularly strong in the late 17th through mid-18th centuries, much like we found in the primary analysis. Finally, Figure E.7 reveals that it is those volumes using language at the nexus of science and political economy *that also have high industry scores* that are the most progress-oriented, especially in the 18th and 19th centuries. This is also similar to what we found in the primary analysis.

Figure E.1: Topics by Category, 1550–1850, Unbinned data



Note: Categorization into “Science”, “Political Economy” or “Religion” based on topics’ placement in 1850.

Figure E.2: Relationship between Categories over time, within volumes, unbinned

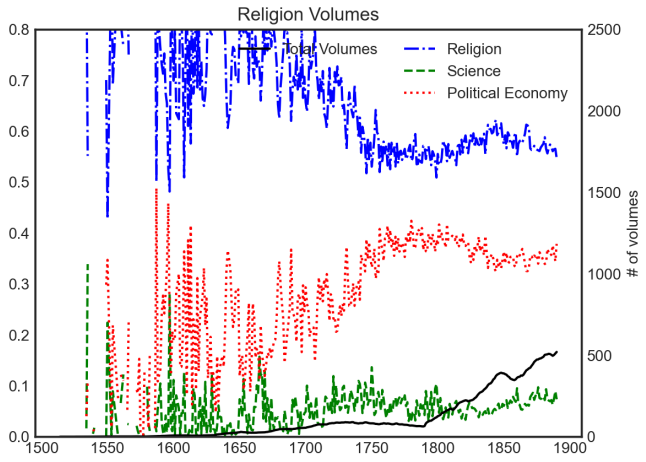
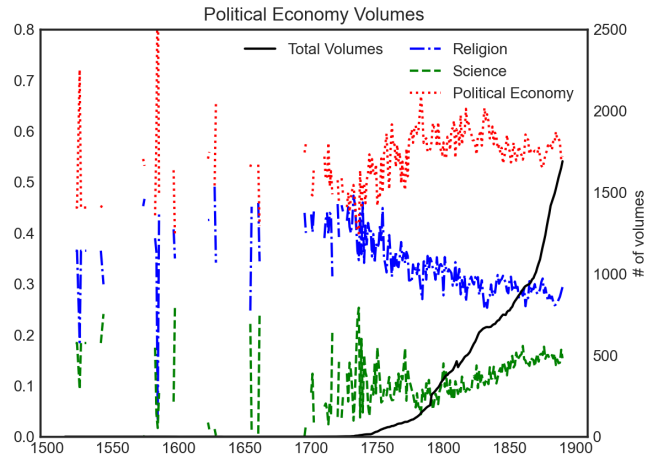
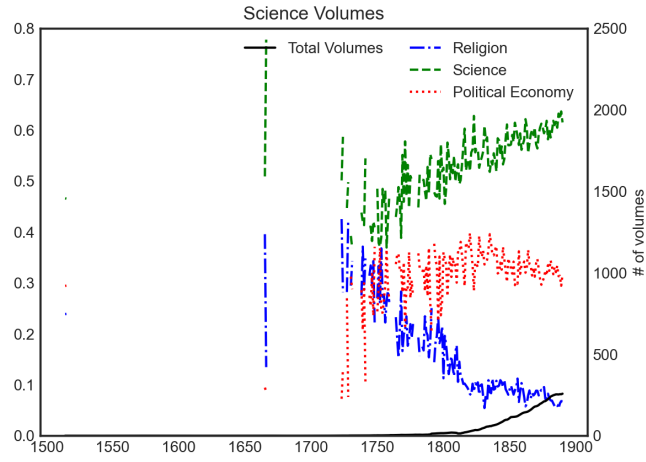
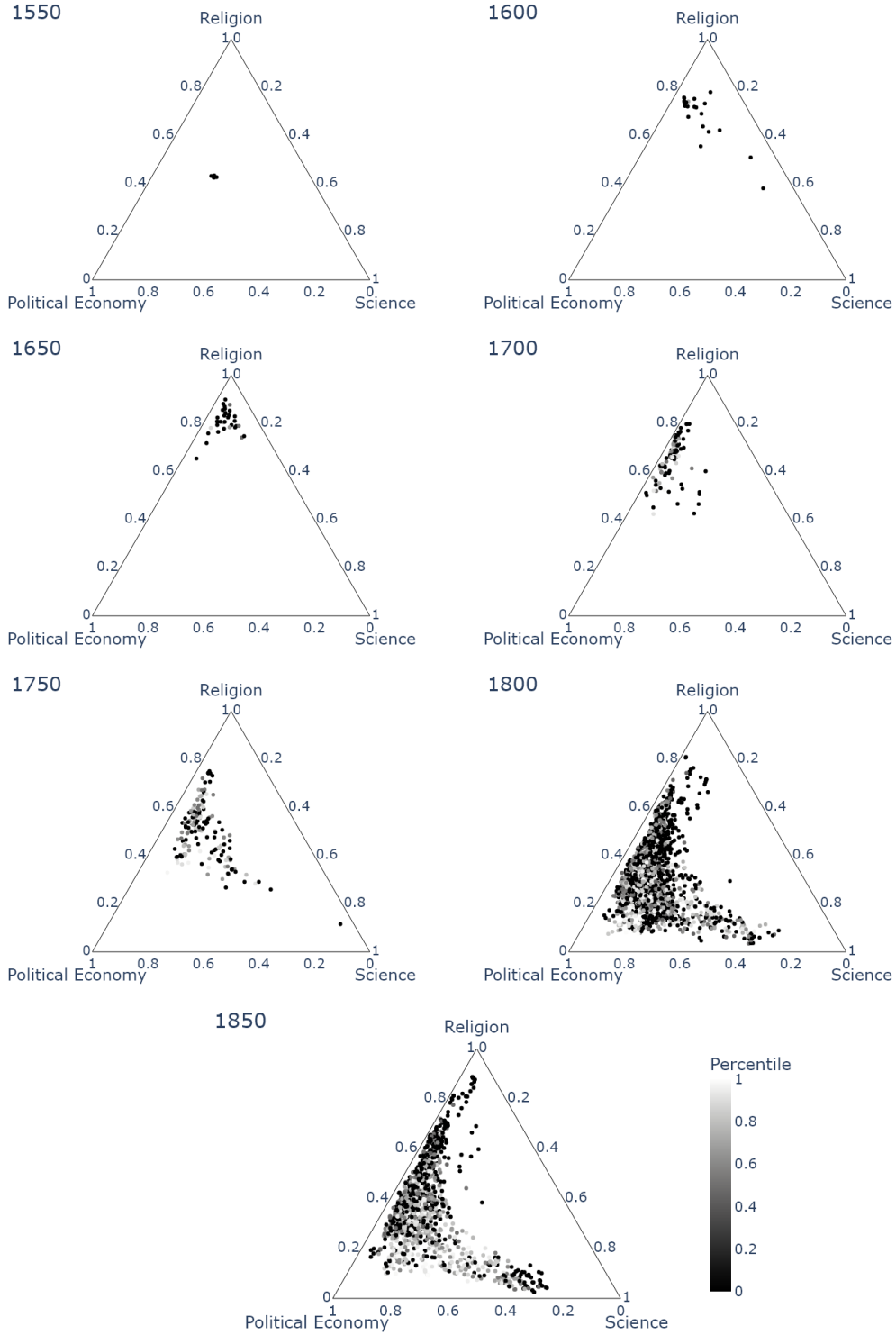


Figure E.3: Progress Sentiment, 1550–1850, unbinned



Note: Each dot represents a volume. The shade of each dot represents the sentiment of that volume, with lighter shades representing more progressive sentiment.

Figure E.4: Marginal Effects and Predicted Values, Progress Sentiment Regressions, un-binned

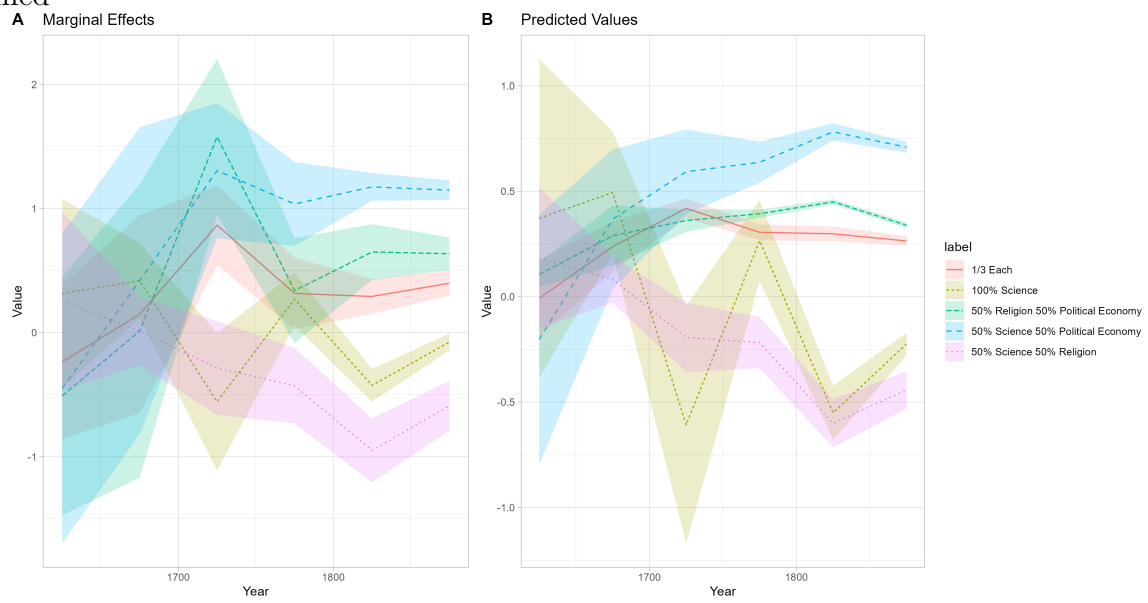
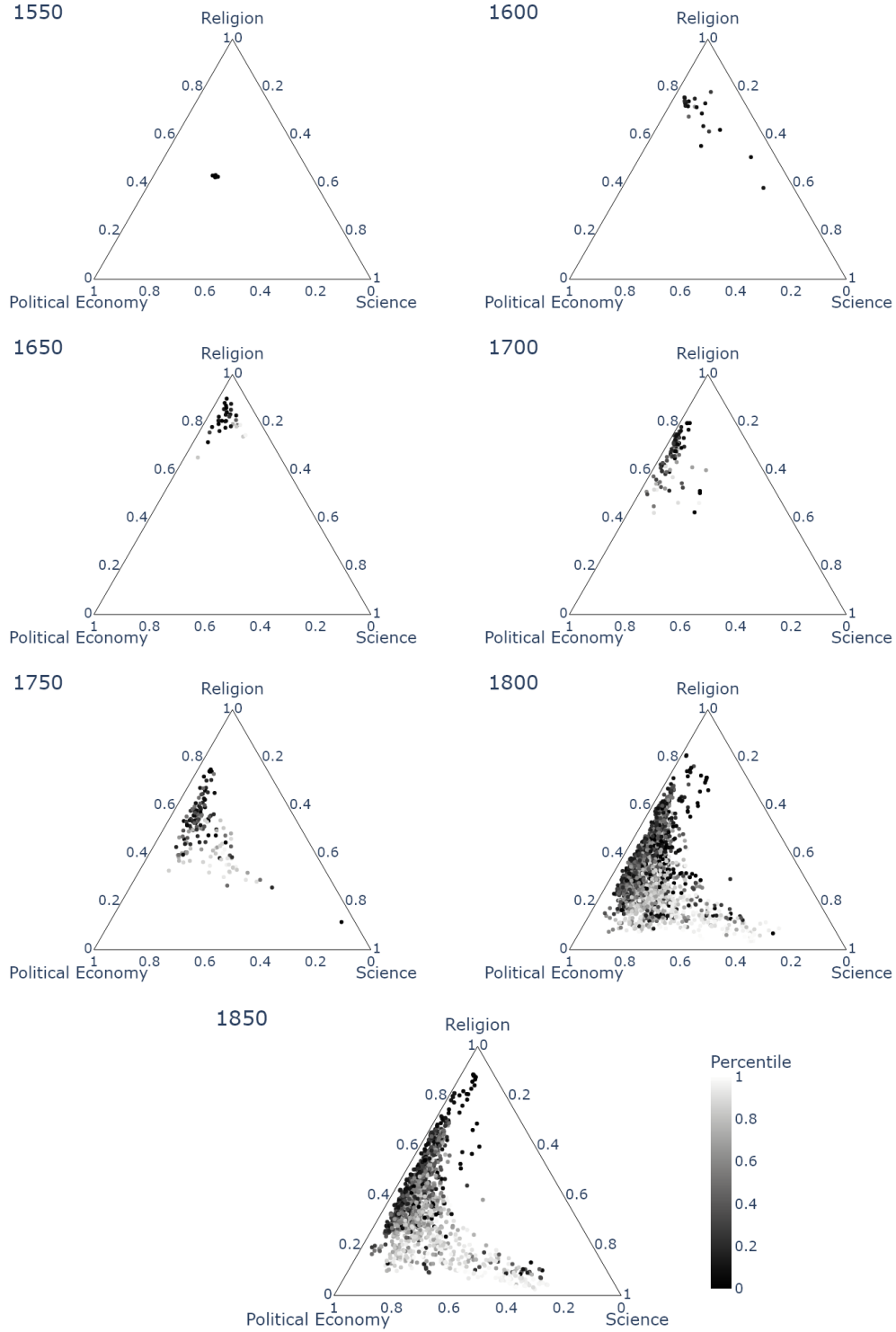
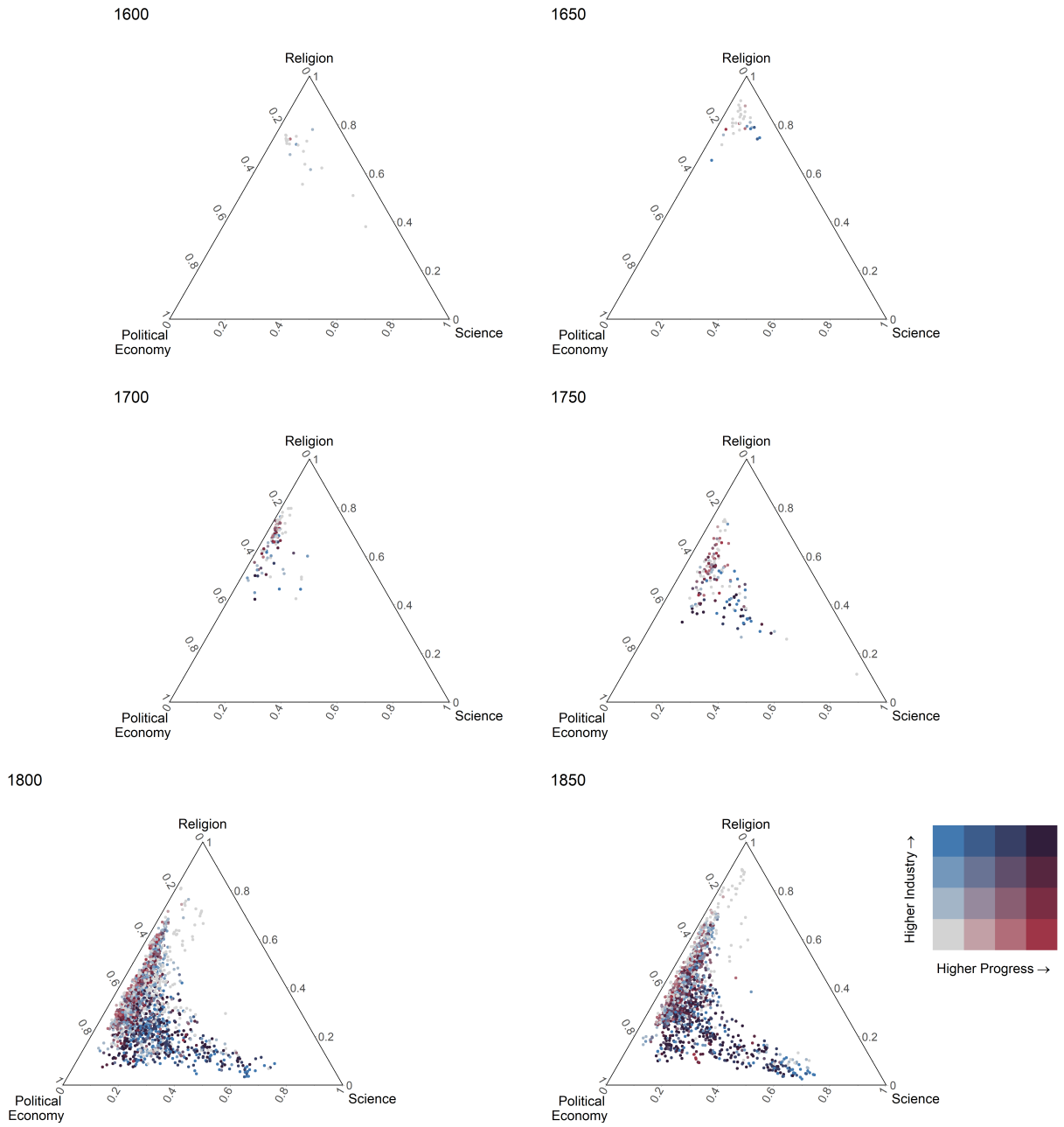


Figure E.5: Industry Sentiment, 1550–1850, unbinned



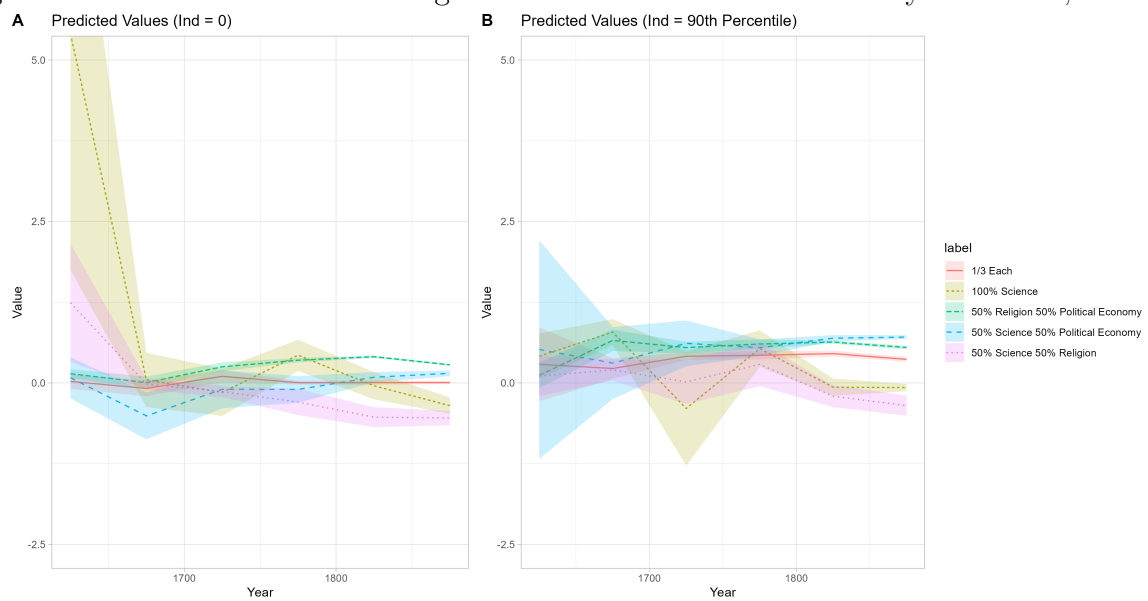
Note: Each dot represents a volume. The shade of each dot represents the sentiment of that volume, with lighter shades representing more industrial sentiment.

Figure E.6: Progress and Industry Sentiment, 1600–1850, unbinned



Note: Each dot represents a volume. For each year. The shade of each dot represents the sentiment of that volume along both the industry and progress axes.

Figure E.7: Predicted Values of Progress Sentiment at 0 and 90 Industry Percentile, unbinned



F Placebo Test: “Optimistic” Sentiment and the Language of Science

It is possible that our analysis 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 F.1. These words are used to calculate sentiment in the same manner as we calculated progress-oriented sentiment (i.e., using equation (5)).

Table F.1: Optimism Dictionary Word Lists

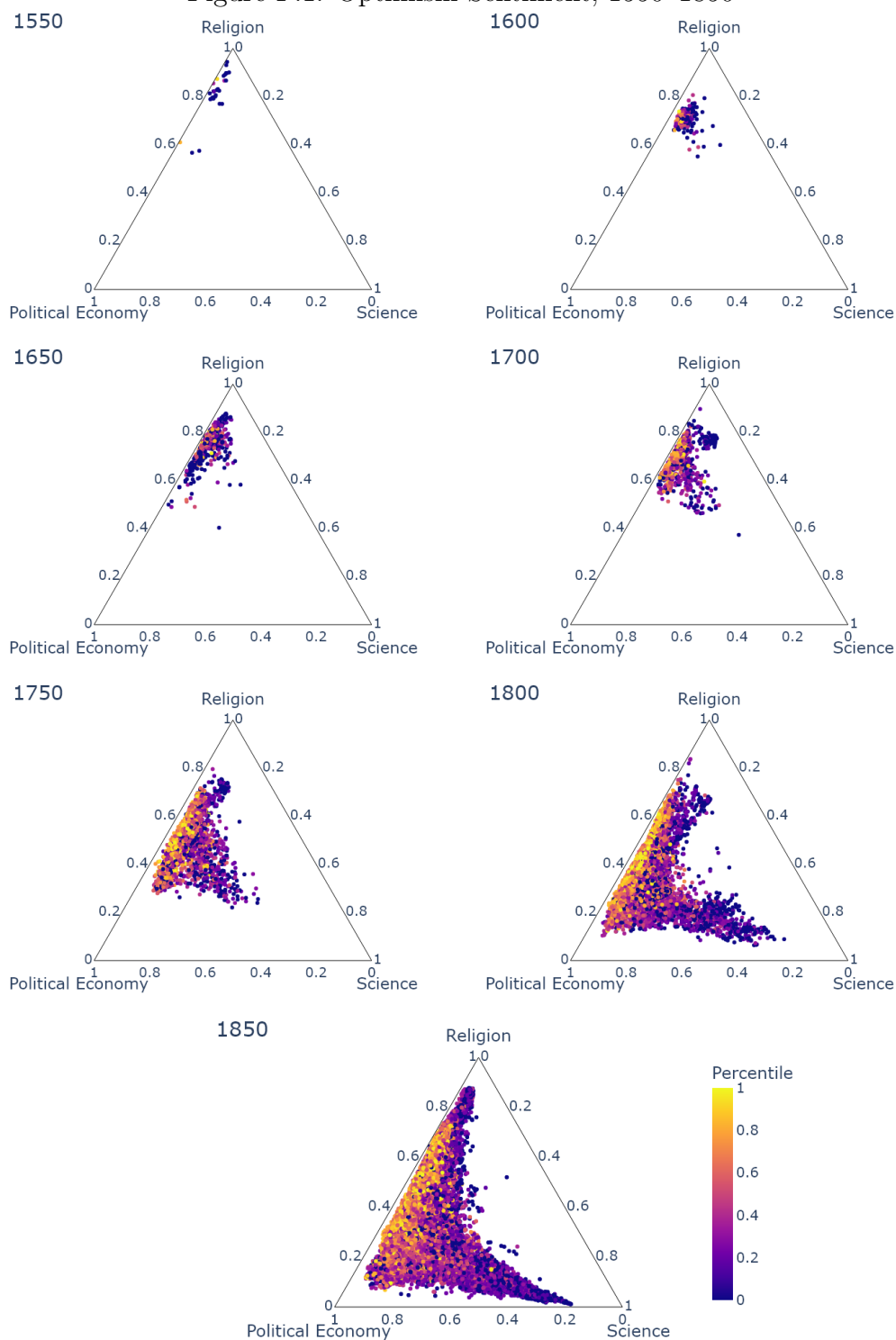
optimistic	optimism	anticipation
assurance	assured	calmness
cheer	cheerful	cheerfulness
cheering	confidence	confident
easiness	elation	encouraged
encouragement	enthusiasm	exhilaration
expectant	happiness	happy
hopeful	hopefulness	hoping
idealism	idealistic	merry
promising	rosy	sanguine
sanguineness	sureness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

Each volume is assigned an optimism sentiment score. Figure F.1 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 6. Those results indicated

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.

Figure F.1: 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.