Knowledge Suppression and Resilience under Censorship: Three-century Book Publications in China*

Ying Bai† Ruixue Jia‡ Jiaojiao Yang §

September 11, 2024

Abstract

This study investigates the short-, medium-, and long-term impacts of state censorship on knowledge production, focusing on the largest book banning in Chinese history, triggered by the creation of the *Siku Quanshu* (Complete Library in Four Sections) during 1772–1783. By analyzing publication data of over 161,000 books spanning from the 1660s to the 1940s, we find that categories subjected to more severe bans experienced a significant decline in publications in the decades following the bans (1780s to 1840s). However, as state control relaxed from the 1840s onwards, there was a marked resurgence in the publication of books in previously restricted categories. Further text analysis reveals notable spillover effects on less sensitive books in the same categories as those banned, indicating a chilling effect and associated self-censorship. We also document dynamic responses from publishers and authors, finding that the exit and entry of publishers help explain both the suppression and subsequent revival of knowledge production.

*We thank Matthew Gentzkow, Sergei Guriev, Shaoda Wang, Jaya Wen, David Yang, Li-an Zhou, Katia Zhuravskaya and seminar/conference participants at Monash and University of Chicago for their comments.

†Department of Economics, Chinese University of Hong Kong; ybai@cuhk.edu.hk

‡UCSD, CEPR and NBER; rxjia@ucsd.edu.

§Department of Economics, Chinese University of Hong Kong; jiaojiaoyang@link.cuhk.edu.hk

1 Introduction

Censorship, a practice with a centuries-long history, has manifested in various forms worldwide, from the suppression of ideas during medieval Europe's Inquisition to modern-day internet censorship in China. Its potential detrimental impact on knowledge production is a subject of considerable interest.[1] Essential questions to be addressed include: How does censorship influence knowledge creation? What role does self-censorship play? If the power of censors wanes, could there be a permanent loss of knowledge due to decreased interest and availability in censored subjects, or might there be a revival? This paper seeks to explore questions by analyzing the most extensive book banning in Chinese history and its effects on book production and content over short, medium, and long durations.

The focused book ban began with the establishment of *Siku Quanshu* (translated as *Complete Library in Four Sections*)—the largest book collection in Chinese history. Conducted during a time of economic growth, this initiative enabled the government to influence the knowledge that was accessible. The project, spanning from 1772 to 1783, cataloged more than 13,000 book titles. Among these, about 3,500 were deemed state-approved knowledge, with their complete texts included, while another roughly 3,000 were classified as banned books, representing forbidden knowledge. Although the books in the library and those banned represented only a minor fraction of the total books, they served as benchmarks for allowed versus prohibited knowledge, potentially affecting future knowledge creation. Analyzing the 50 book categories employed in the Chinese publication system,[2] our data show that the most censored categories (measured as the share of banned books in all collected books) were chronicle history, imperial decrees and memorials, military strategy, and various religions, which tended to include publications that concerned the legitimacy of the Qing state.

To study how censorship affects book production and content, we construct a comprehensive

[1] Existing research on censorship in economics and political science has focused on a ruler's decision of whether to control media and implement censorship (e.g., Besley and Prat, 2006; Gehlbach and Sonin, 2014; Shadmehr and Bernhardt, 2015) and what content is censored (e.g., King, Pan and Roberts, 2013). Despite the critical role of knowledge production, the literature remains limited, with exceptions on how religious censorship in Europe influenced publications of censored authors (Becker, Pino and Vidal-Robert, 2021; Blasutto and De la Croix, 2023) or firms with censored publications (Comino, Galasso and Graziano, 2024). More literature is discussed later.

[2] This bibliographic classification, which consists of four main sections and about 50 subcategories, was fully developed around 600 CE and was employed in subsequent dynasties (Zhou, 1996).

database to document book publications for nearly three centuries (1660s–1940s). Our data includes 161,007 titles across 50 publication categories,[3] along with information on their publishing time and authors. For approximately half of the books, we have additional information about their publishers. Using a standard difference-in-differences approach, we examine the effects of censorship by studying the dynamic patterns in book publications across categories with varying levels of censorship.

Our setting has characteristics that help shed light on the broader issue of censorship. First, the banned books acted as a guideline, but the actual enforcement of censorship was marked by ambiguity and uncertainty (see Section 2.3 for further details) – a common feature in many censorship practices.[4] This ambiguity can result in self-censorship, a widespread phenomenon that we can study thanks to the well-established book classification system. Second, our data spanning three centuries allows us to analyze how publications reacted under different political climates. Specifically, during the 1840s–1900s, China experienced many conflicts with Western countries and a major civil war, leading to diminished government control. The Qing dynasty's fall in 1911 accentuated these transformations. These dramatic political changes provide an opportunity to study the long-term responses of knowledge production after sustained repression, a topic that, despite its relevance, has rarely been examined empirically. Additionally, the textual information from book titles permits the use of text analysis methods to examine book contents, helping interpret the mechanisms behind the dynamic responses.

Our research includes three sets of analysis. The first set illustrates a fluctuating pattern in book production influenced by political changes. In the seventy years succeeding the bans, from the 1780s to the 1840s, we find a significant decline in publications for categories under stricter bans. Specifically, a one standard deviation increase in the share of banned books is associated with a 16% drop in the number of book titles. Post-1840s, however, a noteworthy resurgence was observed: previously suppressed categories returned to pre-censorship trends. This resurgence was also seen in the decades after the fall of the Qing dynasty (1910s–1940s). In contrast, no distinct trends were observed across book categories in the eleven decades before the organization of the *Complete Library*.

Heterogeneity analysis reveals that most of our findings are due to books published by private publishers, rather than state-published works (which did not recover until the collapse of the

[3]These titles consist of both first prints and reprints. We include both in our main analysis and also examine them separately.

[4]For instance, film censorship often comprises vague guidelines on forbidden subjects, yet films might still be censored without clear reasons.

dynasty). Regional heterogeneity shows that during the 1840s–1911, the revival was more evident in treaty ports, consistent with the fact that treaty ports were the first to lose state control.[5] We also examine the role of natural sciences by aligning historical and modern book classifications and find similar dynamics when considering the share of natural science books by category. Furthermore, we show that our results hold after accounting for author censorship (including associated Literary Inquisition cases) by category and potential measurement error in book publication data.

Our second analysis focuses on the content of book titles to determine which types were affected. We start with the keywords in the titles of the banned books and observe that the prevalence of these keywords decreased between 1773 and 1839, with a recovery occurring after the 1840s. This suggests that not only did the quantity of books in more censored categories recovered, but the topics previously banned did as well. Subsequently, using a method akin to Gentzkow and Shapiro (2010), we sort books into two categories: sensitive and less-sensitive, based on their resemblance to banned titles (forbidden knowledge) and *Complete Library* full-text titles (orthodox knowledge). We find that that both categories experienced cycles of decline and revival. Because the sensitive group constituted a much smaller fraction of the total books, the less-sensitive books explained the majority of the observed changes in publications during both phases of decline and revival. We interpret the significant changes in less-sensitive books as indicative of a chilling effect and associated self-censorship since these books were not directly banned. Furthermore, we study the impacts of both lost and surviving banned books and find that their dynamic effects on knowledge production are alike, reinforcing the importance of the chilling effect in our results (compared with the potential knowledge loss resulting from banned books).

Our final analysis examines publisher and author behaviors. Given that most publishers and authors did not remain active throughout the three-century study period, our findings are largely shaped by exit and entry dynamics. Empirically, we observe that the number of publishers in more censored categories dropped post-censorship and recovered after the 1840s. Moreover, incorporating the number of publishers into our analysis explains the publication patterns over time, suggesting their relevance. This conclusion regarding the role of publishers is reinforced using a Bartik method, where we predict the number of publishers in each category based on aggregate changes and the initial distribution across categories. In addition, by separating publishers and authors active in different periods, we find indicative evidence on authors' responses to censorship.

[5]Treaty ports refer to the cities that were forced to open to Western nations after the Opium War in 1840 and subsequent conflicts. These cities were among the first to experience significant growth and industrialization after the 1840s (Jia, 2014).

In sum, our study presents two key novel findings: (1) the revival of knowledge production after prolonged suppression, and (2) the role of self-censorship. These patterns introduce fresh evidence to the burgeoning economics literature, illustrating that censorship hinders knowledge creation under religious censorship in Europe (e.g., Becker, Pino and Vidal-Robert, 2021; Blasutto and De la Croix, 2023; Comino, Galasso and Graziano, 2024), impacts the supply of information in contemporary China (Qin, Strömberg and Wu, 2018), and may decrease the demand for forbidden information in well-constructed experimental settings (Chen and Yang, 2019). Previous research on religious censorship in Europe primarily examines censored authors and publishers. The absence of clearly-defined book categories makes it challenging to investigate the spillover effects on non-censored authors or publishers in historical Europe. The enduring bibliographic classification system in Chinese publications allows us to reveal the significant *chilling* effects on new publications and associated self-censorship. Our results show that despite censored books representing only a minor portion of total publications, the chilling effect can be considerable. Furthermore, the chilling effect fluctuates with the degree of state control, as evidenced by the resurgence of publications in more censored categories after the 1840s.

Clearly, the resuregence of book publications does not mean that censorship is trivial. Despite considerable population growth, the period from the 1770s to 1840s is viewed as a time of intellectual stagnation, with many scholars focusing on studying methods (known as *Kaoju* that emphasizes philology/textual research) rather than substaintial topics. Influential thinkers at the turn of the 19th and 20th centuries, such as Liang Qichao, speculated that censorship was a driving factor behind the cultural stagnation (Liang, 1921). Our findings thus provide empirical support for this broad conjecture about intellectual history.

The related literature also looks at historical instances of inquisition in censorship processes and identifies long-term detrimental effects on economic development, social capital (e.g., Xue, 2021; Drelichman, Vidal-Robert and Voth, 2021) and innovation (Wang, 2022). A challenge for such long-term analyses is the scarcity of data over short and medium timescales. By examining book production across both short and long durations, our study assesses the negative impacts of censorship on book production and suggests that societal actors – publishers, authors, and readers – can modify their behaviors in reaction to notable changes in the political climate, aligning with the significance of environmental change in influencing persistence versus change (Giuliano and Nunn, 2021).

Our findings also illustrate the role of publishers as key agents in implementing censorship. Investigating publisher behavior joins the aforementioned studies on book censorship in historical

Europe as well as research on the role of technology and industrial organization in knowledge creation (e.g., Dittmar, 2011; Dittmar and Seabold, 2019). Our contribution lies in documenting that the dynamic pattern in book production can be attributed to publishers' exits and entries.

Broadly speaking, our study adds to the growing literature that uses books as a metric of ideas (e.g., Alexopoulos, 2011; Abramitzky and Sin, 2014; Squicciarini and Voigtländer, 2015; Almelhem et al., 2023). Our results highlight both how political control shapes idea generation. In situations of censorship and government control, certain ideas are repressed. However, as state control begins to wane, these ideas can resurface, even after being suppressed for decades – a phenomenon that is not unique to the Chinese context.

2 Historical Background

2.1 The *Complete Library* and Associated Censorship

The *Complete Library in Four Sections* is regarded as one of the most extensive collections of Chinese literature. Initiated by the Qianlong Emperor during the Qing dynasty in 1772 and completed in 1783, a period characterized by notable economic prosperity, this project aimed to preserve and display the vast corpus of Chinese knowledge. Following the enduring bibliographic classification system, the library was categorized into four divisions: Classics (including Confucian works), Histories (covering various historical genres), Masters (covering fields such as agriculture, medicine, and religion), and Collections (featuring genres such as poetry and stories). The initiative involved over 3,800 scholars and resulted in a library cataloged over 13,000 book titles, accumulating over 79,000 volumes and roughly 997 million words in total. Among these book titles, approximately 3,500 were deemed state-approved knowledge, with their complete texts included (Encyclopedia of China, 2021).

The creation of the *Complete Library*, however, as also marked by significant censorship. Concerned about the Qing dynasty's legitimacy as a minority Manchu ruling over a predominantly Han population, the Qianlong Emperor sought to suppress dissent or criticism, resulting in the banning of books deemed politically sensitive or critical of Qing rule. Approximately 3,102 books were banned. The prohibition of these books was linked to around 100 cases of the Literary Inquisition, leading to the execution of those involved. Both officials and citizens were encouraged to identify and report contentious content, creating an atmosphere of fear and self-censorship among scholars. Although there are official records of banned books, historians suspect that these records

may have missing issues and have created overlapping yet slightly differing lists. We use both the official records and historians' work in our analysis.

The compilation process and associated censorship has captured the interest of many historians, with significant works (e.g., Guo, 1937; Guy, 1987; Huang, 1989). Based on historians' work, we outline the book collection and censorship process in five stages in Figure 1. As illustrated, local officials were tasked with collecting books from families and individuals. The collected books were then submitted to an editorial office established by the central government, which classified the books and determined which ones should be banned.

2.2 Characteristics of Censorship

Three prominent aspects of this censorship process are notable, which share similarities with censorship and policy implementation in contemporary China and other contexts.

The first feature is *delegation*. Although the central government ultimately decided on book bans, the task of gathering books was given to local officials. The quantity of books confiscated had an impact on the future promotions of these local officials (Zhang, 1997). As a result, local officials were incentivized to meticulously gather books, often recruiting citizens to assist in their searches. This collaboration between officials and citizens fostered an environment of fear.

The second feature is *ambiguity and uncertainty*. While citizens generally had an idea of which topics might be contentious based on past events, there was still some level of ambiguity and uncertainty. For instance, it was common for local authorities to search beyond the specified list, and the possibility of being reported by fellow citizens added to the uncertainty. Like in other settings, reporting by fellow citizens was often driven by personal conflicts (Wang, 2015).

The third feature is *low probability but severe punishment*. Although most banned books were confiscated and no penalties were enforced, the potential for severe punishment in certain cases, such as those witnessed during the Inquisition, was ominously present.

Appendix A.1 presents instances of banned books, along with the reasons for their prohibition. For example, *Chronicles of the Ming Dynasty* was banned for "containing unconventional accounts of the Ming dynasty's history." Another book on military conflicts, *Record of the Northern Expedition*, got banned because it documented the battles against the Qing troops at the end of the Ming Dynasty. We analyze the variation in content censorship in Section 3. Generally, most banned cases did not result in direct punishment of individuals. Nevertheless, when punishments did occur, they were exceedingly harsh, including the death penalty for the author, the editor, the

publisher, and their family members. In many scenarios, the authors of these banned books, such as Chen Jian (1497–1567) and Shen Deqian (1673–1769) were already deceased. However, their descendants, along with the descendants of the editors and publishers, faced repercussions. These aspects collectively suggest a chilling effect, which we explore using data.

2.3 The Publication Market

The book market in Qing China was both dynamic and complex. On the demand side, the long-standing Civil Service Examination system, which emphasized an in-depth understanding of the Confucian classics and other scholarly texts, played a pivotal role in cultivating a large and educated population. This cultivated intellectual environment led to a consistent and substantial demand for books across various social strata.

On the supply side, Qing China's publishing landscape differed from that of Europe, where movable-type printing had become the standard. Instead, Qing China predominantly relied on woodblock printing. This method, while more labor-intensive per print, had significantly lower initial costs compared to movable-type printing, which required substantial investment in machinery and typefaces. The accessibility of woodblock printing made it feasible for smaller, decentralized publishers to produce books, thereby fostering a more diverse and privately-controlled book market. In our data, this decentralized nature is reflected by the dominance of private publishers and their dispersed locations (see related analysis in Section 4.3).

Another factor that added to the decentralized nature of publishing was the choice of many individuals to publish their works and circulate them within a limited literate community (Zhang, 1989), seeking recognition or the intrinsic value of self-expression rather than profits. In our study, we do not distinguish between publications aimed at expression or profit, as they generally operated together.

The decentralized nature of the book market in Qing China presented considerable challenges for a thorough enforcement of state censorship. Despite official efforts to suppress certain works, many banned books managed to survive these censorship attempts (Brook, 1988), which is confirmed in our data. However, our primary interest is not in the banned books themselves, but in understanding how the imposition of book bans influenced the broader creation and dissemination of new knowledge.

3 Data and Measurement

China developed a bibliographic system during the Tang dynasty, known as the Four Section system, which was used in later dynasties. Each section contained 10-15 categories, resulting in a total of 50 categories. We employ variation in censorship across these 50 categories and conduct our main analysis at the category–year level. Our research employs a dozen data sources, with summary statistics presented in Table 1 and the data construction process detailed in Appendix A. Below, we outline the key variables in our study.

Book Production and Content To measure knowledge production, we obtain comprehensive records of Chinese books from the 26-volume *General Catalogue of Pre-modern Chinese Books* (General Catalogue Editorial Office, 2012). In our studied period, 1662-1949, the Catalogue included over 161,000 books, with information on their publication category, the author, the publisher, publication time, and whether they were reprints (see an example of the records in Appendix A.2).

There are two key measurement concerns in the book publication data. First of all, the *General Catalogue* was assembled from 1992 to 2009 and is based on books that have survived. If a surviving book has several editions with some earlier editions missing, this is recorded. We include these missing editions in our analysis. However, it could be suggested that books from more heavily censored categories are less likely to survive. We have three remarks on this interpretation. First, our dynamic results indicate a revival of publications in more censored categories after the 1840s, which eases the concern of survival bias. Second, we look at the relationship between censorship and the proportion of missing editions in each category (see Appendix A.3) and find no systematic relationship. This evidence is merely suggestive, as it is based on books that have survived but have missing editions. Finally, even if our findings are partially influenced by the possibility that more censored categories are less likely to diffuse and survive, censorship still incurs a cost in terms of access to knowledge. Therefore, we do not intend to dismiss the relevance of diffusion for our findings.

In addition, 24.5% of the books lacks specific publications years within the reigns. In Appendix A.4, we examine whether the missing rates are linked to censorship and find this is not the case, thus suggesting that the missing publication year data is not a result of censorship. We also consider the rate of missing years in each category in our analysis.

We are also interested in book contents and and how they react to censorship. To represent the

contents, we use text data from book titles, specifically focusing on keywords in the titles.

Category-level Censorship To measure the level of censorship, we collect the records of banned books from two sources: (i) the summary of banned books wrote by the editors of the *Complete Library* (Zhang, 1997); (ii) the catalogues of banned books compiled by historians (Chen, 1932; Sun, 1957; Yao, 1957). As these sources heavily overlap with each other, we use the union in our baseline and present additional results from each source.

We then calculate the share of banned books in each category among all collected books (i.e., banned books and *Complete Library* full-text and indexed books) as a measure for censorship. A complexity in this calculation is the uncatalogued status of the lost books by the *General Catalogue*. For these lost books, we code their categories based on their titles and summaries and validate our approach with historical research on banned books (see Appendix A.5 for the procedures). Moreover, we differentiate between the proportions of lost and surviving banned books in our later analysis.

Figure 2 shows the level of censorship in each category. The censorship degree has a mean of 12% and a standard deviation of 16%. The most censored categories are chronicle history, imperial decrees and memorials, military strategy, and various religions, which tended to include publications that concerned the legitimacy of the Qing state. In contrast, the least censored categories are Erya (a dictionary-like glossary), General Classics, Legalism, Medicine, and Family Genealogy.

There were two types of book bans: the majority (71.59%) were content-based, while a smaller fraction were author-based. In cases of author-based bans, all works by those authors were prohibited. We focus on content-based censorship, as it captures forbidden ideas. We also consider author-censorship in our analysis. As detailed in Appendix A.6, the pattern of author-censorship differs greatly from idea-censorship. The most frequently censored topics related to authors are Family Genealogy, Buddhism, Filial Piety, Collected words of multiple authors and Biographies, likely reflecting that one's collection of works and family history was censored. The correlation between content-based and author-based censorship is weak, with a correlational coefficient of -0.03.

Category-level Characteristics We collect more data and consider category-level characteristics that may affect our analysis, organized around market factors, political factors, and data issues. The main new data is whether the publishers were state or non-state, which we code based on information

on historical presses (Du, 2001; Du, 2009; Zhai, 2009). We explain the data construction process in Appendix A.7.

Markets factors include category size measured by the share of books in each category among all books, HHI of publisher concentration in each category, and reprinting share in each category. Political factors include the share of books printed by state publishers in each category and the level of author-base censorship in the same way as we construct the level of content-based censorship. These variable are all based on publication before 1662 (our start year). Additionally, we also consider missing-year rate in each category.

We check the correlations between censorship and these characteristics and find no strong relationship (Appendix Table A.8), indicating that size or market considerations are not key determinants of censorship. In our analysis, we allows time-varying impacts of these characteristics.

4 Decline and Revival in Book Production

4.1 Descriptive Evidence and Research Design

In Figure 3, we categorize books according to their level of censorship, dividing them into groups above and below the mean, and display their yearly trends over time. The impact of these bans appears evident: initially, the two categories were similar, but post-ban, a noticeable divergence emerged between the less and more censored books from the 1780s to the 1840s. This disparity, however, did not persist forever as the more heavily censored books eventually caught up after the 1840s.

Motivated by the descriptive trend in Figure 3, we use a standard difference-in-differences approach alongside an event-study method. The difference-in-differences specification is as follows:

$$\begin{aligned} ln\#Book_{c,t} = \beta_1 Censor_c * \mathbf{I}_t^{1773-1839} + \beta_2 Censor_c * \mathbf{I}_t^{1840-1911} + \\ \beta_3 Censor_c * \mathbf{I}_t^{1912-1949} + \alpha_c + \lambda_t + Section_s * t + \mathbf{X}_c * \mathbf{I}_t + \epsilon_{c,t}, \end{aligned} \tag{1}$$

where $\#Book_{c,t}$ refers to the number of books in category c and year t, and $Censor_c$ represents the level of censorship in category c. To address the observations of zeros, We start with a linear model using $ln(\#Bookproduction_{c,t} + 1)$ as the dependent variable, use a dummy indicator for books in category c and year t, and employ the Poisson to ensure our findings are not driven by any specific linear transformation (Chen and Roth, 2024).

To address category-level characteristics that exhibit little variation over time and factors that affect all categories over time, we incorporate category-level fixed effects (α_c) and year fixed effects (λ_t) as controls in our analysis. Furthermore, we gradually control for the category-level characteristics introduced above and their interaction with the period dummy ($\mathbf{X}_c$*$\mathbf{I}_t$), allowing the effects of these characteristics to vary across different periods. Additionally, we account for section-specific time trends to accommodate different trends for each section. Lastly, to account for potential correlation within categories, we cluster the standard errors at the category level.

Our assumption for estimating equation (1) is that categories with different level of censorship were on a similar trend before the compilation of the *Complete Library*. To check whether this assumption is reasonable, we use an event study strategy to estimate the impacts of censorship every five years, using 1772 as the reference year.

We anticipate $\beta_1 < 0$ because censorship probably caused the suppression of publications. However, we do not have priors on the signs of β_2 and β_3 and rely on data to reveal whether and when recovery took place.

4.2 Main Results

Our analysis reveals that categories subject to more stringent censorship experienced a significant decline in publication in the seven decades following the book bans, and followed by a notable revival afterwards. In Table 2, we present the results for different model specifications. The dependent variable in Columns (1)-(3) is under logarithmic transformation. In Column (1), we include only year and category fixed effects. In Column (2), we introduce section-specific time trends, and in Column (3), we incorporate category-level controls and their interactions with the period dummy. They imply that a one standard deviation increase in the level of censorship is associated with a reduction in publication by approximately 15.5% during 1773–1839. However, the coefficients are not different from zeros during 1840–1911 and 1911–1949, implying a recovery of book publications in the more-censored categories so that they return back to trends similar to those of the less-censored.

In column (4), we examine the extensive margin by using the dummy indicator, whether there is any book published in category i in year t, as the dependent variable. The dynamic pattern is similar to those in Columns (1)-(3).

In Column (5), we present the Poisson model, with the dependent variable being the number of books. The dynamic pattern is similar to that in Column (3). For simplicity, we focus on

Poisson estimates in most of our following analysis. It is worth noting the positive coefficient during 1911-1949 is sizable but the standard error is also large. Therefore, we refrain from making strong assertions about any clear overshooting.

To check the assumption of parallel trends in estimating equation (1), we present the five-year-by-five-year coefficients of the effects of censorship in Figure 4, using 1772 as the reference year. Prior to 1773, there is no discernible correlation between censorship and book publication, indicating that book categories did not exhibit divergent trends before the compilation of the Complete Library. However, following the compilation, censorship consistently exerted a negative impact on book publication that lasted several decades. However, after the 1840s, the negative effect of censorship diminished considerably, and book publication began to follow a trend similar to the pre-ban period.

4.3 Heterogeneous Patterns

We study four distinct types of heterogeneity to investigate differences in responses to censorship concerning publisher ownership, geographical locations, and the relevance for natural sciences.

Publisher Data and Ownership In our book publication records, 51% of them contain publisher information. The absence of publisher information often reflects that authors published books to distribute within a limited literary community (Zhang, 1989), while state-published books typically included publisher details. Before employing publisher data, we examine the proportion of books missing publisher information in each category and its relationship with censorship over time. As detailed in Appendix B.1, there is no evident connection, indicating that the absence of publisher information is not necessarily influenced by censorship.

We then separate books based on whether the books were published by the state and find that our baseline results are primarily influenced by the non-state publishers (which we call private publishers for simplicity). As indicated in Columns (1)-(2) of Table 3, state publishers were more responsive to censorship, and their publications in more censored categories did not rebound until the dynasty's fall. These results suggest that state publishers were slower to recover compared to private publishers.

First Prints vs. Reprints Our primary analysis includes both first prints and reprints from the *General Catalogue* data, with first prints making up 71% of the total publications.[6] As shown in Column (3) and (4) of Table 3, first prints are the main contributors to our results. On the other hand, reprints had a negligible impact.

We examine the responses to censorship in treaty ports and the rest of China by incorporating the interaction of censorship, treaty ports and periods. As reported in Column (6), we find evidence that the recovery in treaty ports during 1840s–1911 is more apparent than that in non-treaty ports. This pattern is consistent with the interpretation is that the loss of state control started in the treaty ports.

Importance of Natural Sciences To study whether natural sciences and social sciences responded to censorship differently, we calculate an index to access the proportion of natural sciences within each historical category. To do so, we match the 50 categories for pre-modern Chinese books with modern book categories and describe this procedure in Appendix B.2. The overall share of natural sciences in historical Chinese publications is relatively small, around 14%.

In Column (5), we incorporate the interaction between the level of censorship and the share of natural science as a control variable, while also controlling for the share of natural science itself. We find that our main findings on the decline and revival apply to all categories. If anything, categories that contained more natural sciences experienced a more pronounced revival after the demise of the dynasty.

Treaty Ports vs. Non-Treaty Ports The Opium War of 1840 marked a turning point in Chinese history. Since the 1840s, a series of war resulted in various "unequal treaties" that China was forced to sign with Western nations and was forced to open up to the rest of the world. These dramatic changes imply that the state could not control the society as in the past, especially in the treaty ports. Motivated by this institutional background, we construct a category-period-(publishing) prefecture panel data, which enables us to study regional variation (see Appendix A.7 for a map of the spatial distribution of publishers). The dispersed locations of publishers reflects the decentralized nature of publishing technology in our studied.

[6]Here, a first print refers to the initial appearance of a book in the *General Catalogue*. Some of these books may have been published earlier but did not survive.

4.4 Measurement Error

We are concerned with measurement error when using historical data on book publications. As discussed in Section 3, the absence of publication years and the likelihood of missing editions across different categories do not show a systematic relationship with censorship. Additionally, we contact a category-emperor analysis, which is not affected by missing publication years. Appendix Table B.3 shows that the effect of censorship, though less precisely estimated, aligns with the baseline findings.

Measurement error on the level of censorship is also likely. In Appendix B.4, we determine the level of censorship by counting banned books from various sources. Table B.4 displays the findings, showing that assessing the level of censorship from both official sources and historical research produces similar outcomes. In addition, we consider both lost and survived banned books when calculating the degree of censorship. When separating them, we find similar dynamic patterns, which are discussed later to help interpret our findings.

5 Book Contents and Self-Censorship

In this section, we analyze the content of publications and the crucial role of self-censorship (and its change post 1840s) in our findings. We perform text analysis on the titles of the publications. We also investigate the behaviors of publishers and authors to understand their responses.

5.1 Book Contents

Using the natural language processing (NLP) techniques, we analyze book contents. Our first analysis focuses on keywords in book titles. As previously noted, banned content represents a minor portion of the overall book contents. In particular, the banned keywords constitute only 1.6% on average of the total keywords from all publications.

In Columns (1)-(3) of Table 4, we examine how the same set of keywords changed along with censorship. Column (1) presents the decline and revival of all keywords. Column (2) focuses on the keywords of banned books and shows that these keywords declined during 1773–1839 and recovered after 1840s, indicating that the banned topics declined and recovered. Column (2) presents a similar dynamic pattern on other keywords, implying that the effects of censorship goes beyond banned topics per se. Due to the fact that banned books make up a minor proportion of total books, the responses of unbanned keywords were more pronounced and significantly contributed

to the bulk of our results.

Moreover, we categorize books' sensitivity based on their similarity to banned books. We first identify two word sets representing the forbidden knowledge (banned books) and orthodoxy knowledge (the *Complete Library* full-text books), and then convert words into vectors using a word embedding model. The vector difference($\overrightarrow{Keywords}_{\text{Bannedbooks}} - \overrightarrow{Keywords}_{\text{Completelibrary}}$) serves as the benchmark for comparison. We then split the title of each book into several keywords, and embed words into vectors. For each book, we calculate the average vector. Finally, we compute the Cosine similarity between the vector of each book title and the vector difference. We present a detailed description of the method in Appendix C.1.

Figure 5 plots the two word sets to illustrate the ideas. As shown, the most frequent keywords in forbidden knowledge often concerns history whereas the orthodoxy knowledge highlights classics. Figure 6 shows the distribution of similarity levels across various groups: Panel (a) illustrates this for banned books, Panel (b) for those in the *Complete Library* full-text books, and Panel (c) for all books. We use the average value from the banned books (-0.053) as the benchmark to define sensitive books, which corresponds to the 85th percentile in the distribution of all books.

In Columns (3)-(4) of Table 4, we observe that both sensitive and less sensitive books reacted to censorship: both categories declined between 1773 and 1839 and rebounded after the 1840s. Importantly, given that less sensitive books constitute 85% of the total, the decline and subsequent resurgence can largely be attributed to the spillover effects on this group.

Together, these observations underscore the significance of self-censorship: not only did books similar to those on the banned list become less prevalent between 1773 and 1839, but also numerous less sensitive books in the same categories that experienced increased censorship.

5.2 Chilling Effect vs. Knowledge Loss

Conceptually, banned books can influence knowledge production through two mechanisms: the chilling effect, where publishers and authors avoid certain subjects, and knowledge loss, where knowledge producers lose access to certain information and produce less as a result. Our research on the importance of spillover effects and the resurgence of publications in more censored categories post-1840s indicates that the chilling effect of banned books is more prevalent than the knowledge loss mechanism. Moreover, as stated earlier, banned topics (keywords) constituted only a small fraction of the overall topics (keywords).

To validate our interpretation, we leverage information on lost and surviving banned books.

Should knowledge loss be a significant factor, we would expect that the keywords in the lost banned books were less likely to recover. This is not the case in the data.

Among the 3,102 banned books, 1,783 have survived, 1,319 were lost. Using the same method in Section 5.1, we identify the keywords representing the forbidden knowledge using surviving banned books and lost banned books, respectively. The dependent variables of Columns (1) and (2) in Table 5 are the numbers of banned keywords, derived from surviving and lost banned books, and the dependent variables of Columns (3) and (4) are the numbers of unbanned keywords, which refer to the set of all keywords excluding those from banned books. As shown, we observe similar dynamics in decline and revival when using these two measures. Thus, we again infer that our findings are primarily influenced by a chilling effect, rather than a mechanism of knowledge loss.

5.3 Responses of Publishers and Authors

Publishers serve as a critical channel connecting knowledge producers and consumers. Here, we investigate their response to censorship by looking at their behaviors. As discussed above (see more details in Appendix B.1), we observe no correlation between censorship and the occurrence of missing publisher information across different categories over time. Therefore, our following analysis uses the number of publishers within each category based on books with publisher information.

Fall and Rise in the Number of Publishers Using the same specification as in equation (1), we examine the number of publishers across four periods (1662–1773, 1773–1839, 1840–1911, and 1912–1949) to study publishers' responses to censorship. Parallel to our main finding on the number of books, we observe a decline and resurgence in number of publishers. As shown in Columns (1)-(2) of Table Table 6, the number of publishers and authors venturing into sensitive fields significantly declined after 1773, and then experienced a rebound after 1840s. Based on the estimates, a one standard deviation increase in the level of censorship is associated with a 12% reduction in the number of publishers.

In Columns (3)-(4), we include the number of publishers in each period as control variables, and we observe that the effect of censorship on book publications can be fully absorbed by the role of publishers. This evidence, however, is merely suggestive as the number of publishers across categories is also influenced by censorship.

To formally test the relevance of publishers in explaining our baseline finding on book production, we take a Bartik approach to address the endogeneity. Specifically, we count the number of

new publishers entering the market in the periods 1773–1839, 1840–1911, and 1912–1949, and assign the new publishers to each category according to the proportion of books in each category at the end of the last period. Namely, we use the product of the initial proportion and the number of new publishers to generate a Bartik instrument. Column (5) shows that the Bartik instrument is strongly correlated with the number of publishers, and Column (6) presents the reduced-form results on how the instruments affects the number of books. In Column (7), once controlling for the number of publishers, we find that the effect of censorship did not significantly differ across periods, indicating that the number of publishers can explain our baseline finding on book production.

Exit, Entry, and Responses of Surviving Publishers To further understand the relevance of publisher exits, entries, and responses of surviving publishers (who might change categories due to censorship), we divide all publishers into two categories: surviving publishers and others. Surviving publishers are those that remained active for at least two periods. As shown in Column (1) of Table 7, more publishers exited heavily censored categories between 1773-1839, while more entered these categories during 1840-1911. Additionally, these trends are mostly driven by compositional changes (Column (3)). The actions of the surviving publishers (Column (2)) follow the same pattern of exits and entries, but since they represented a minor portion of the total publishers, it was the former that accounted for a significant part of the overall changes in the number of publishers across categories.

Taken together, we find that the changes in the number of publishers can explain our main findings on the decline and revival of book production. On an aggregate level, the margins of publisher entry and exit play a pivotal role in explaining the observed dynamics.

Suggestive Evidence on Authors Our findings on the significance of publishers do not disregard the reactions from authors. However, it is more challenging to examine authors' reactions since we can only observe those who have successfully published books. Censorship might result in the silence of authors who remain unobserved. Additionally, it is difficult to determine whether authors ceased writing on a subject or if they were unable to find publishers willing to publish their work.

As suggestive evidence to look at publishers and authors separately, we leverage the living era of 7,907 authors whose biographical data is available (see the Appendix C.2 for additional information on authors). We start by examining authors who died before 1773 to study how censorship impacted their publications across three periods: before 1773, from 1773 to 1839, and after 1840. As these authors were no longer alive, the changes in their publications reflect the actions of publishers. Table

8 shows the cross-category correlations between censorship and book publications. As illustrated in Column (1) of Table 8, there was a reduction in this correlation for publishers active from 1773 to 1839, compared to those active before 1772, suggesting that publishers became more inclined to respond to censorship by publishing fewer works in more censored categories. However, after the 1840s, this correlation became insignificant again, aligning with the recovery observed in our main analysis.

We then focus on the publisher who were active post 1840, when the effect of censorship disappeared. To explore how authors reacted to censorship, we analyze the publication of books by authors from different time periods: those who died before 1772 and those who lived between 1773 and 1839. As indicated in Row (3) of 8, there was no clear association between books published after 1840 by authors who passed away before 1772 and the level of censorship. This is reassuring as neither these authors nor the publishers were influenced by censorship in this context. However, for books published after 1840, a significant negative correlation persists between the level of censorship and books by authors who lived during 1773–1839. Given that censorship's impact had faded, this implies that authors from 1773 to 1839 also responded to censorship.

6 Conclusion

In this study, we examined the largest book ban in Chinese history to explore the impact of censorship on knowledge production over short-, medium-, and long-term periods. The nature of this censorship, characterized by ambiguous enforcement, echoes practices seen across various historical and contemporary contexts. By utilizing extensive publication data and a well-established book classification system in China, we were able to systematically assess how censorship shaped book production.

Our findings reveal that, despite seven decades of suppression, there was a remarkable revival in knowledge production following the loosening of state control. This resurgence highlights the resilience of knowledge and the adaptive capacity of society in response to shifting political climates. Importantly, this pattern is not unique to China; similar phenomena have been observed in other historical settings where censorship and repression were prevalent.[7] The three-century scope of our data provides a clear view of these dynamic changes over time.

The revival of suppressed knowledge, however, does not imply that censorship is not detrimental.

[7]For instance, the resurgence of Orthodoxy in post-Soviet Russia may follow a similar trend (e.g., Evans and Northmore-Ball, 2012), though the revival of religion is more complex due to potential state involvement.

On the contrary, the suppression and self-censorship observed during the short and medium terms following the initial ban likely contributed to a prolonged period of intellectual stagnation in China from the 1770s to the 1840s. This period stands in stark contrast to the contemporaneous developments in Europe, where the Industrial Revolution and the expansion of progressive ideas were flourishing (e.g., Mokyr, 2016; Almelhem et al., 2023). This divergence suggests that the suppression of knowledge in China may have had significant political and economic consequences during a critical period of global change.

Furthermore, our analysis indicates that the dynamic patterns in book production can be largely explained by the responses of publishers. This finding suggests that understanding the process of knowledge production requires not only a focus on authors and readers but also on the behaviors of intermediaries and platforms that play a crucial role in implementing censorship. This phenomenon remains relevant in today's digital age, where online platforms and intermediaries continue to influence the flow of information.

References

[1] **Abramitzky, Ran, and Isabelle Sin.** 2014. "Book Translations as Idea Flows: The Effects of the Collapse of Communism on the Diffusion of Knowledge." *Journal of the European Economic Association*, 12(6): 1453–1520.

[2] **Alexopoulos, Michelle.** 2011. "Read All about It!! What Happens Following a Technology Shock?" *American Economic Review*, 101(4): 1144–1179.

[3] **Almelhem, Ali, Murat Iyigun, Austin Kennedy, and Jared Rubin.** 2023. "Enlightenment Ideals and Belief in Progress in the Run-up to the Industrial Revolution: A Textual Analysis." *IZA Discussion Paper No. 16674*.

[4] **Becker, Sascha O, Francisco J Pino, and Jordi Vidal-Robert.** 2021. "Freedom of the Press? Catholic Censorship during the Counter-Reformation." *CEPR Discussion Paper No. DP16092*.

[5] **Besley, Timothy, and Andrea Prat.** 2006. "Handcuffs for the Grabbing Hand? Media Capture and Government Accountability." *American Economic Review*, 96(3): 720–736.

[6] **Blasutto, Fabio, and David De la Croix.** 2023. "Catholic Censorship and the Demise of Knowledge Production in Early Modern Italy." *Economic Journal*, 133(656): 2899–2924.

[7] **Brook, Timothy.** 1988. "Censorship in 18th Century China: A View from the Book Trade." *Canadian Journal of History*, 22(2): 177–96.

[8] **Chen, Jiafeng, and Jonathan Roth.** 2024. "Logs with Zeros? Some Problems and Solutions." *Quarterly Journal of Economics*, 139(2): 891–936.

[9] **Chen, Naiqian.** 1932. *Index of Banned Books Catalogue (Suoyingshi de jinshu zongmu).* Shenchu Tang Publishing House.

[10] **Chen, Yuyu, and David Y Yang.** 2019. "The Impact of Media Censorship: 1984 or Brave New World?"

American Economic Review, 109(6): 2294–2332.

[11] **Comino, Stefano, Alberto Galasso, and Clara Graziano.** 2024. "Censorship, Industry Structure, and Creativity: Evidence from the Catholic Inquisition in Renaissance Venice." *Journal of Law, Economics, and Organization*, ewae015.

[12] **Dittmar, Jeremiah, and Skipper Seabold.** 2019. "New Media and Competition: Printing and Europe's Transformation after Gutenberg." *CEP Discussion Papers No. DP1600*.

[13] **Dittmar, Jeremiah E.** 2011. "Information Technology and Economic Change: the Impact of the Printing Press." *Quarterly Journal of Economics*, 126(3): 1133–1172.

[14] **Drelichman, Mauricio, Jordi Vidal-Robert, and Hans-Joachim Voth.** 2021. "The Long-run Effects of Religious Persecution: Evidence from the Spanish Inquisition." *Proceedings of the National Academy of Sciences*, 118(33): e2022881118.

[1] **Du, Xinfu.** 2001. *A Comprehensive Survey of Book Publishing by Province and County in the Ming Dynasty (Quan Ming fensheng fenxian keshu kao).* Thread-Binding Book Bureau.

[2] **Du, Xinfu.** 2009. *A Comprehensive Survey of Book Publishing by Province and County in the Qing Dynasty (Quan Qing fensheng fuxian keshu kao).* Thread-Binding Book Bureau.

[17] **Encyclopedia of China, Editorial Board,** ed. 2021. *The Complete Library in Four Branches (Siku quanshu).* Encyclopedia of China Publishing House.

[18] **Evans, Geoffrey, and Ksenia Northmore-Ball.** 2012. "The Limits of Secularization? The Resurgence of Orthodoxy in post-Soviet Russia." *Journal for the Scientific Study of Religion*, 51(4): 795–808.

[19] **Gehlbach, Scott, and Konstantin Sonin.** 2014. "Government Control of the Media." *Journal of Public Economics*, 118: 163–171.

[20] **General Catalogue Editorial Office, The,** ed. 2012. *General Catalogue of Pre-modern Chinese Books (Zhongguo guji zongmu).* Chung Hwa Book Co. Ltd.

[21] **Gentzkow, Matthew, and Jesse M. Shapiro.** 2010. "What Drives Media Slant? Evidence from U.S. Daily Newspapers." *Econometrica*, 78(1): 35–71.

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

[23] **Guo, Bogong.** 1937. *Study on the Compilation and Editing of the Complete Library in Four Branches (Siku quanshu zuanxiu kao).* Shangwu Press.

[24] **Guy, R Kent.** 1987. *The Emperor's Four Treasuries: Scholars and the State in the Late Chien-lung Era.* Harvard University Asia Center.

[25] **Huang, Aiping.** 1989. *Research on the Compilation of the Complete Library (Siku Quanshu zuanxiu yanjiu).* Beijing: Renmin University Press.

[26] **Jia, Ruixue.** 2014. "The Legacies of Forced Freedom: China's Treaty Ports." *Review of Economics and Statistics*, 96(4): 596–608.

[27] **King, Gary, Jennifer Pan, and Margaret E Roberts.** 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review*, 107(2): 326–343.

[28] **Liang, Qichao.** 1921. *Intellectual Trends in the Qing Period (Qingdai xueshu gailun).* Shangwu Press (reprinted 1947).

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

[30] **Qin, Bei, David Strömberg, and Yanhui Wu.** 2018. "Media Bias in China." *American Economic Review*, 108(9): 2442–2476.

[31] **Shadmehr, Mehdi, and Dan Bernhardt.** 2015. "State Censorship." *American Economic Journal: Microeconomics*, 7(2): 280–307.

[32] **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.

[33] **Sun, Dianqi.** 1957. *Records of Banned Books in the Qing Dynasty (Qingdai jinshu zhijianlu).* The Commercial Press.

[34] **Wang, Fansen.** 2015. *The Capillary Action of Power (Quanli de maoxixueguan zuoyong).* Peking University Press.

[35] **Wang, Xuebo.** 2022. "Freedom of Speech, Spirit of Innovation, and Long-Term Economic Development: Evidence from the Qing Dynasty of China." *Available at SSRN 4011630*.

[36] **Xue, Melanie Meng.** 2021. "Autocratic Rule and Social Capital: Evidence from Imperial China." *Available at SSRN 2856803*.

[37] **Yao, Jinyuan.** 1957. *List of Banned Books in the Qing Dynasty (Qingdai jinhui shumu).* The Commercial Press.

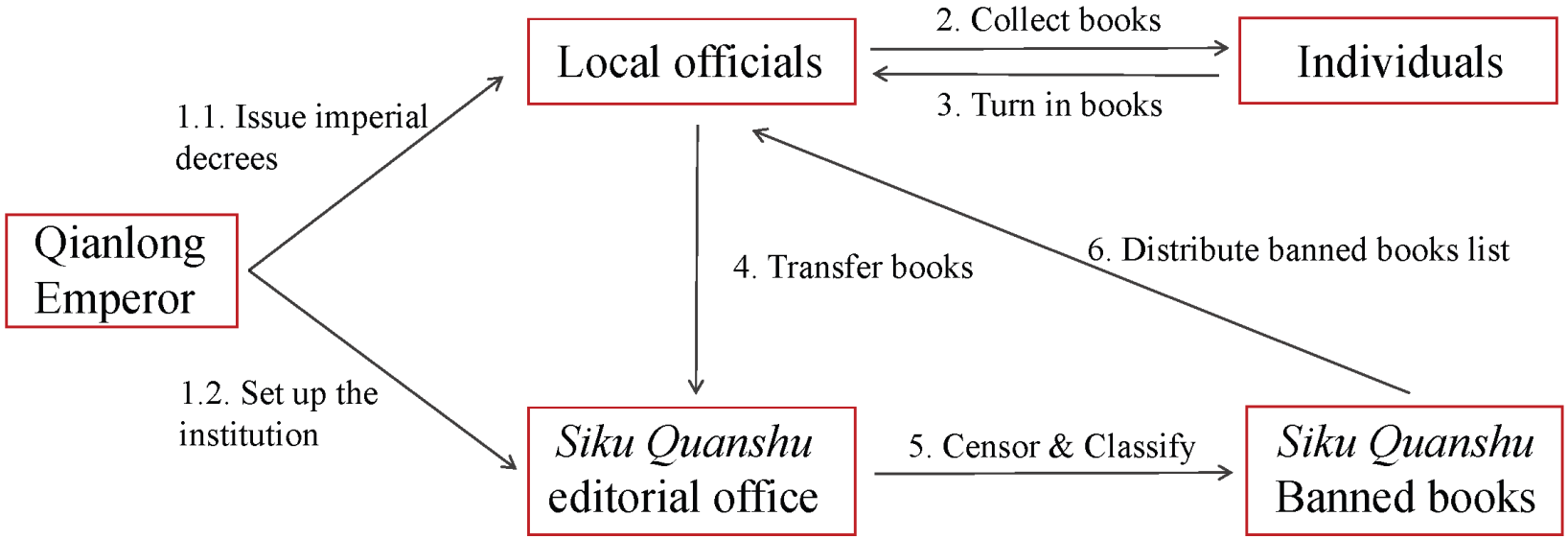
[7] **Zhai, Mianliang.** 2009. *A Dictionary of Historical Chinese Book Printing (Zhongguo guji banke cidian).* Suzhou University Press.

[39] **Zhang, Shucai,** ed. 1997. *The Archives of Compilation of the Complete Library (Zuanxiu Siku quanshu dang'an).* Shanghai Classics Publishing House.

[40] **Zhang, Xiumin.** 1989. *The History of Chinese Printing (Zhongguo yinshua shi).* Shanghai Renmin Press.

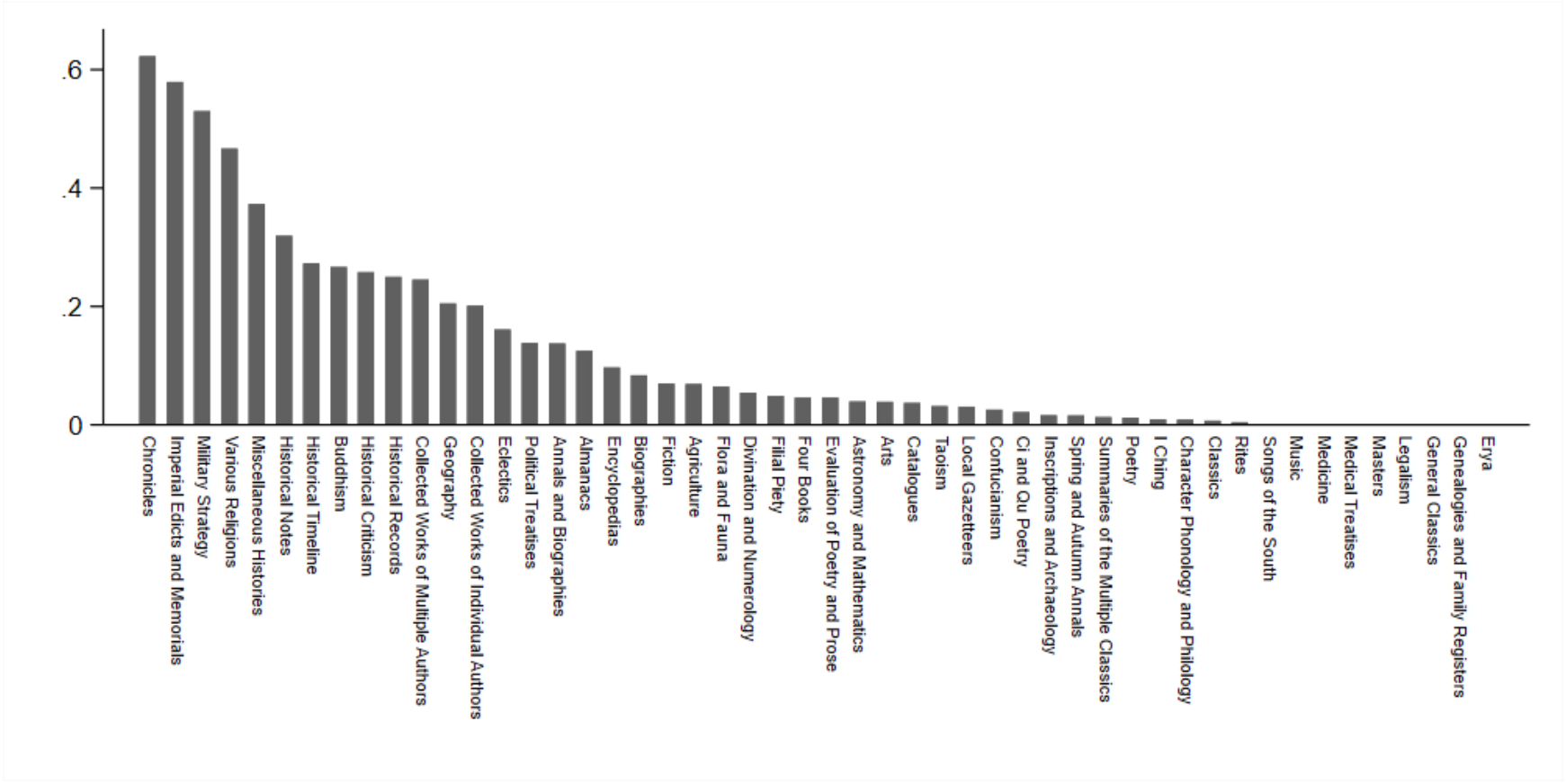
[41] **Zhou, Shaochuan.** 1996. *Studies on Catalogues of Historical Books (Guji mulu xue).* Zhongzhou Historical Books Publishing House.

Figure 1: The Process of Book Collection and Compilation



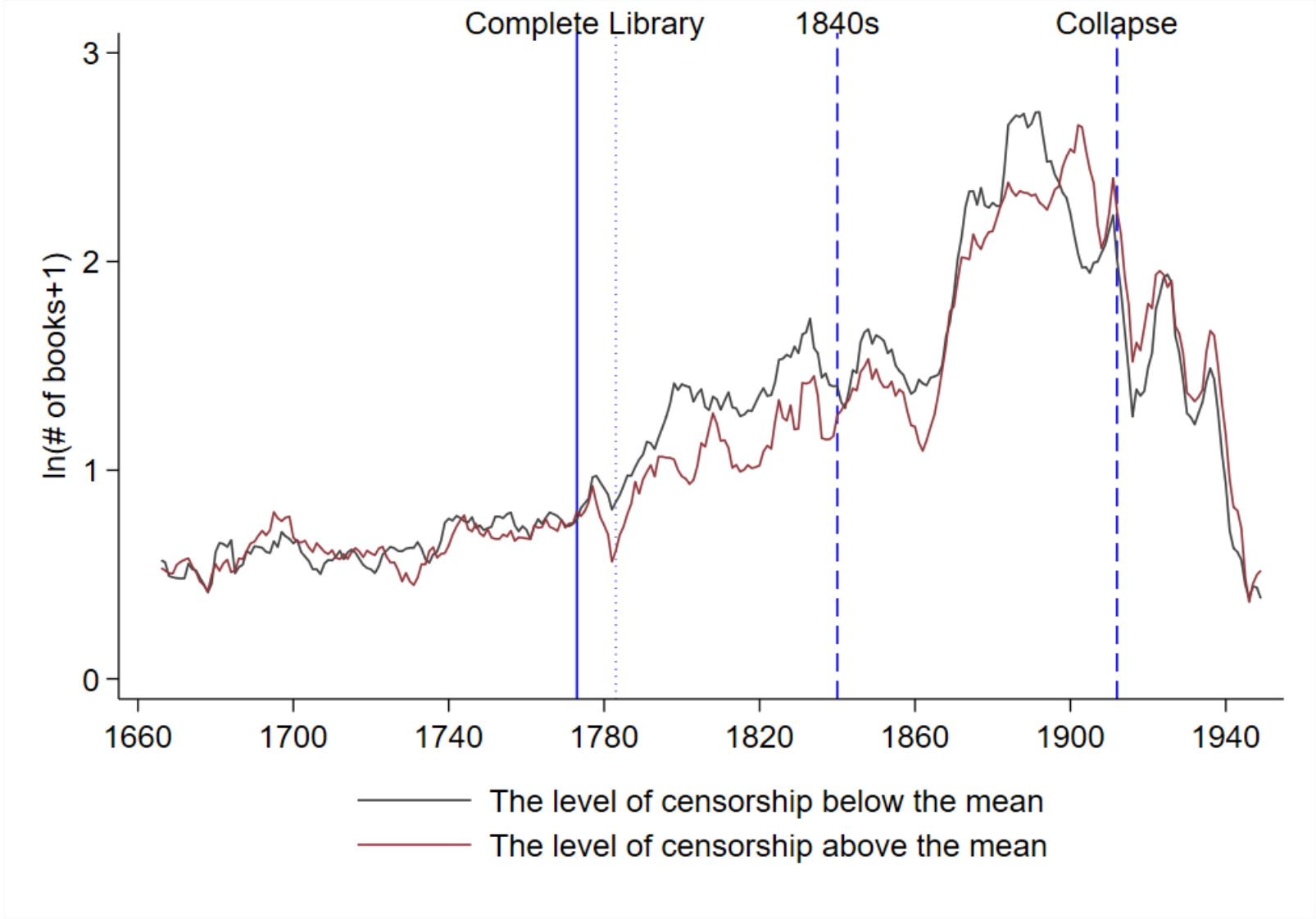
Note. This figure plots the main steps in book collection and censorship. The collection was delegated to local bureaucrats whereas the categorization and censorship decision were centralized.

Figure 2: The Level of Censorship: Based on Content



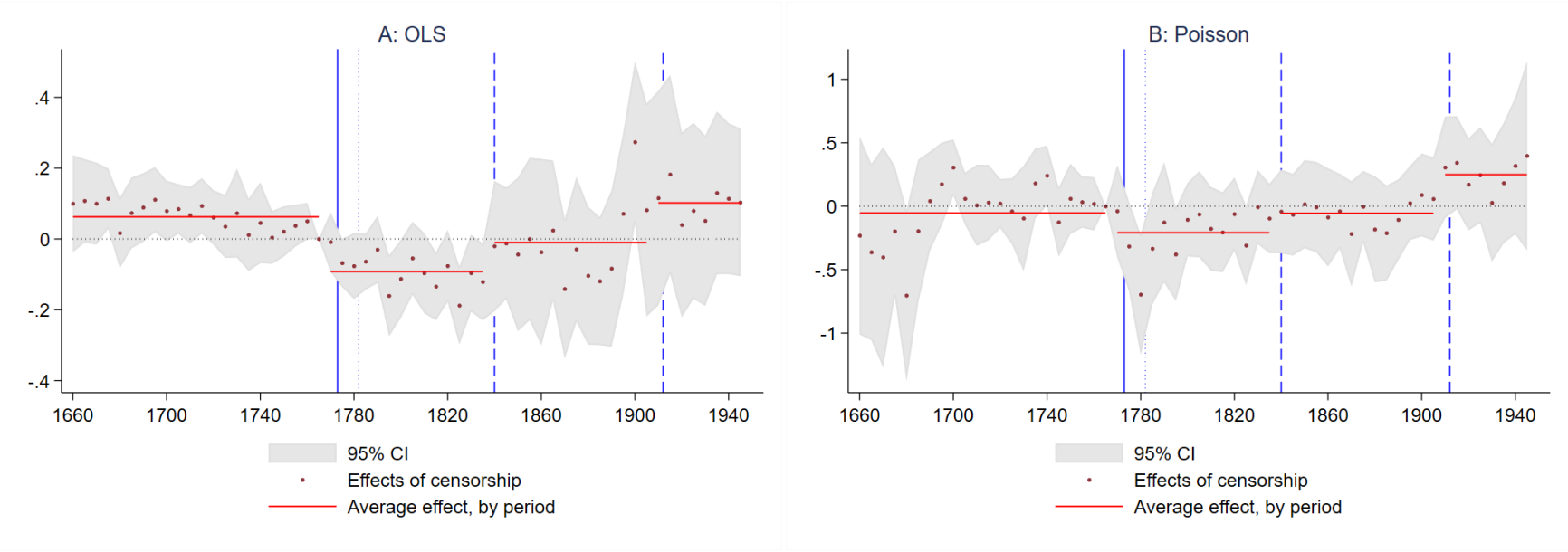
Note. This figure plots the level of censorship across 50 categories. For each category, the level is measured by the share of banned books in total collected book. The most censored categories are chronicle history, imperial decrees and memorials, military strategy, and various religions.

Figure 3: The Impact of Censorship on the # of Books: Descriptive Evidence



Note. This figure plots the yearly trends in book publication for two groups based on censorship degree. The red line represents trends for books whose level of censorship is above the mean, and the black line for those with censorship level below the mean.

Figure 4: The Impact of Censorship on #Books



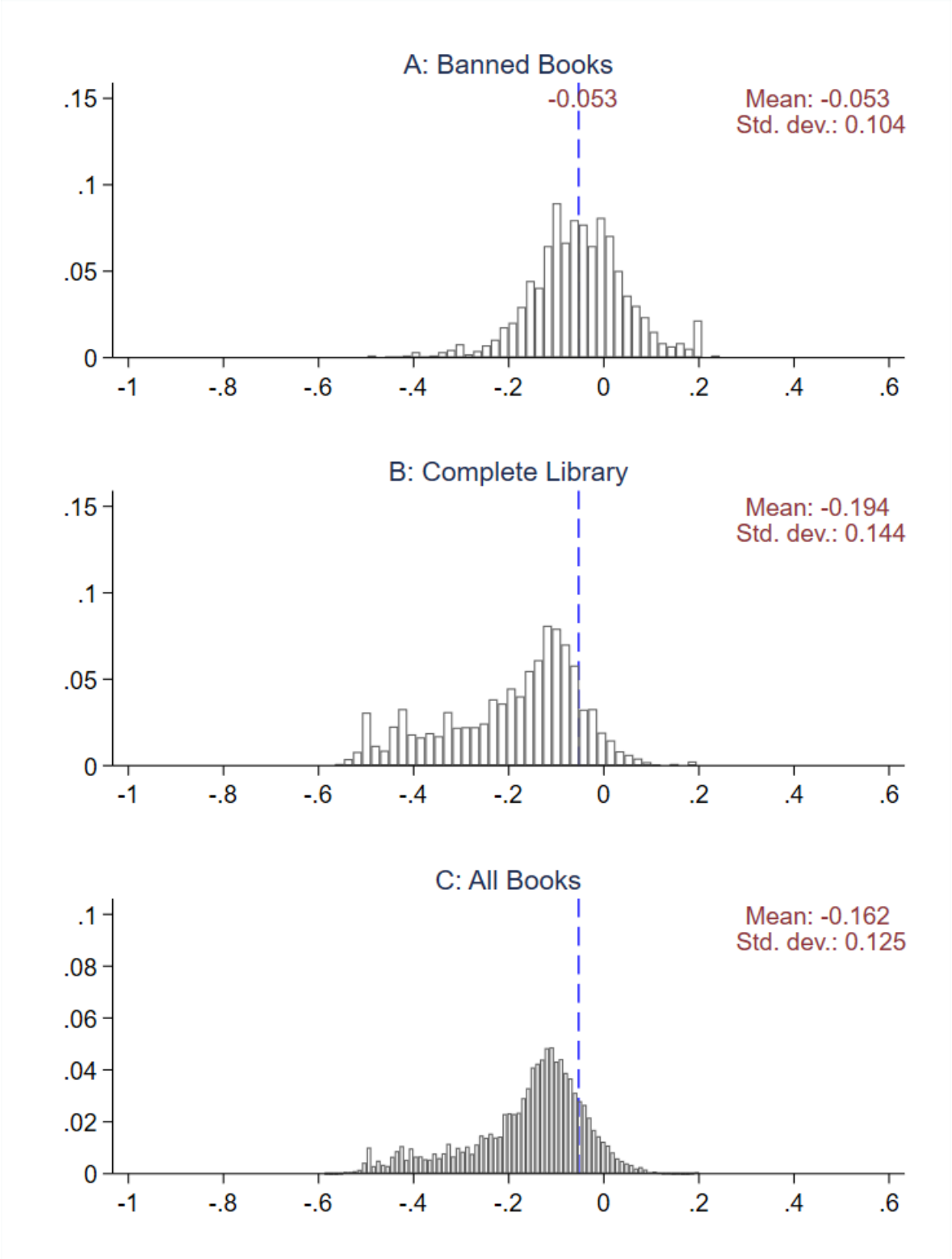
Note. These two figures plot the estimates of the effect of censorship on book publication by every five years, using 1772 as the reference year. The left panel displays estimates based on the OLS model, while the right panel shows estimates obtained from the Poisson model. The blue lines in each figure, situated from left to right, represent the years 1773 (the start of *Complete Library* project), 1782 (the completion of the *Complete Library*), 1840 (a turning point), and 1912 (a new regime), respectively.

Figure 5: Keywords in Banned Books and *Complete Library* Full-text Books



Note. The left panel displays the key words found in banned books, and the right panel displays the key words from the *Complete Library* full-text books. Font size indicates word frequency. Red colors indicate words used in the section of Histories, and blue colors indicate words used in the section of Classics.

Figure 6: The Level of Similarity to Banned Books across Groups



Note. The figure plots the level of similarity across various groups, Panel A on the banned books, Panel B on the *Complete Library* full-text book, and Panel C on all books. The blue line indicates the mean similarity derived from banned books. We use this value as the threshold to categorize all books into sensitive and less sensitive books.

Table 1: Summary Statistics

	Data Source	Obs.	Mean	SD	Min	Max
Baseline						
Books	A	14,400	7.719	20.652	0	467
ln (# books +1)	A	14,400	1.178	1.225	0	6.148
New books	A	14,400	5.475	16.583	0	275
Reprints	A	14,400	2.245	8.006	0	325
State-Published Books	A,G,H,J,K	14,400	0.493	2.905	0	93
Nonstate-Published Books	A,G,H,J,K	14,400	7.226	19.698	0	460
Content-based censorship	B,C,D,E,F	14,400	0.121	0.160	0	0.623
Author-based censorship	B,C,D,E,F	14,400	0.046	0.076	0	0.400
Category size	A	14,400	0.020	0.032	0.0003	0.213
State penetration	A,G,H,J,K	14,400	0.045	0.076	0	0.504
Share of private publishers	A,G,H,J,K	14,400	0.180	0.080	0.006	0.364
Market concentration	A,G,H,J,K	14,400	0.067	0.068	0.004	0.333
Reprinted level	A	14,400	1.551	0.310	1.003	2.802
Missing rate	A	14,400	0.654	0.117	0.148	0.868
Share of natural sciences	A,L	14,400	0.141	0.306	0	1
Treatyport	A,M	52,400	0.130	0.336	0	1
Textual Analysis						
Keywords	A	14,400	15.616	48.124	0	1014
Keywords from banned books	A,B,C,D,E,F	14,400	0.253	0.991	0	38
Keywords from unbanned books	A,B,C,D,E,F	14,400	15.363	47.900	0	1014
Sensitive books	A,B,C,D,E,F	14,400	0.977	3.763	0	176
Less-sensitive books	A,B,C,D,E,F	14,400	5.342	14.167	0	266
Publisher and Authors						
Publishers	A,G,H,J,K	14,400	149.650	355.200	0	3852
ln (# publishers +1)	A,G,H,J,K	14,400	4.014	1.415	0	8.257

Note. See details of our data construction process in Appendix A. Our data sources are: A. General Catalogue Editorial Office (2012). B. Zhang (1997). C. Ji (2000). D. Chen (1932). E. Sun (1957). F. Yao (1957). G. Yang (2014). H. Du (2001). J. Du (2009). K. Zhai (2009). L. Yang (2007). M. Yan (1955).

Table 2: The Impact of Censorship on #Books: DID Estimates

Dependent variable	ln(# Books+1) OLS			# Books>0 Logit	# Books Poisson
	(1)	(2)	(3)	(4)	(5)
Censorship (sd) × 1773-1839	-0.120***	-0.129***	-0.155***	-0.317***	-0.165**
	(0.041)	(0.037)	(0.041)	(0.097)	(0.082)
Censorship (sd) × 1840-1911	-0.012	-0.028	-0.066	0.041	-0.051
	(0.078)	(0.084)	(0.102)	(0.184)	(0.107)
Censorship (sd) × 1912-1949	0.110	0.088	0.043	0.325	0.218
	(0.095)	(0.106)	(0.113)	(0.239)	(0.147)
Category FE	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y
Section × time trend		Y	Y	Y	Y
Controls × period FE			Y	Y	Y
R-squared	0.704	0.708	0.730	0.426	0.750
Observations	14,400	14,400	14,400	13,776	14,400

Note. These estimates are obtained from our category-year data. Category refers to the 50 publishing categories and section refers to the four broad branches. Controls include category size, state penetration, share of private publishers, market concentration index, reprint level, missing rate, and the level of author-based censorship. Standard errors displayed in parentheses are clustered at the book category level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 3: The Impact of Censorship on #Book: Heterogeneous Patterns

Dependent variable	#Books, Poisson					
	Private publishers	State pub.	First prints	Reprints	Nat. sciences	Treaty ports
	(1)	(2)	(3)	(4)	(5)	(6)
Censorship (sd) × 1773-1839	-0.147*	-0.598***	-0.219***	0.065	-0.193**	-0.159*
	(0.081)	(0.136)	(0.085)	(0.113)	(0.095)	(0.089)
Censorship (sd) × 1840-1911	-0.029	-0.629***	-0.020	0.050	-0.053	-0.094
	(0.101)	(0.201)	(0.106)	(0.159)	(0.119)	(0.123)
Censorship (sd) × 1912-1949	0.219	-0.256	0.143	0.274*	0.080	0.111
	(0.144)	(0.164)	(0.160)	(0.163)	(0.153)	(0.083)
Censorship (sd) × nat. sci% × 1773-1839					-0.053	
					(0.302)	
Censorship (sd) × nat. sci% × 1840-1911					-0.412	
					(0.376)	
Censorship (sd) × nat. sci% × 1912-1949					1.193**	
					(0.589)	
Censorship (sd) × treaty port × 1773-1839						-0.103
						(0.076)
Censorship (sd) × treaty port × 1840-1911						0.219**
						(0.101)
Censorship (sd) × treaty ports × 1912-1949						0.087
						(0.095)
Category FE	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	
Section × time trend	Y	Y	Y	Y	Y	Y
Controls × period FE	Y	Y	Y	Y	Y	Y
Nat. sci share × period FE					Y	
Period FE						Y
Prefecture FE						Y
Treaty port× period FE						Y
Category FE × prefecture FE						Y
Pseudo R-squared	0.751	0.597	0.758	0.642	0.755	0.832
Observations	14,400	12,800	14,400	14,400	14,400	11,932

Note. The estimates are all from the Poisson model. Columns (1)-(5) are based on category-year data and Column (6) on category-prefecture-period data. Controls are the same as in Table 2. Standard errors displayed in parentheses are clustered at the book category level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 4: The Impact of Censorship on Book Contents: Textual Analysis

Dependent variable	# Keywords			# Books	
	All	Banned	Unbanned	Sensitive	Less-sensitive
Dependent var. mean	(15.616)	(0.253)	(15.363)	(0.977)	(5.342)
	(1)	(2)	(3)	(4)	(5)
Censorship (sd) × 1773-1839	-0.237***	-0.163**	-0.229***	-0.182**	-0.197**
	(0.079)	(0.080)	(0.082)	(0.079)	(0.096)
Censorship (sd) × 1840-1911	-0.093	0.024	-0.088	-0.054	-0.049
	(0.092)	(0.093)	(0.093)	(0.086)	(0.118)
Censorship (sd) × 1912-1949	0.175	0.123	0.188	0.145	0.195
	(0.148)	(0.156)	(0.148)	(0.140)	(0.140)
Category FE	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y
Section × time trend	Y	Y	Y	Y	Y
Controls × period FE	Y	Y	Y	Y	Y
Pseudo R-squared	0.771	0.427	0.773	0.662	0.689
Observations	14,400	13,064	14,400	13,248	14,400

Note. The estimates are all from the Poisson model. The dependent variables in column (1)-(3) are the number of key words. The dependent variables in column (4) and (5) are the number of books. We categorize all books into sensitive and less sensitive books based on the mean mean similarity derived from banned books(-0.053).Books with a similarity above this threshold are classified as sensitive, while those below are considered less sensitive. Controls are the same as in Table 2. Standard errors displayed in parentheses are clustered at the book category level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 5: The Impact of Censorship on # Keywords: Identify with Survived and Lost Banned Books

Dependent variable	# Banned keywords		# Unbanned keywords	
Identify keywords w. banned books	Surviving	Lost	Surviving	Lost
Dependent var. mean	(0.224)	(0.169)	(15.392)	(15.448)
	(1)	(2)	(3)	(4)
Censorship (sd) × 1773-1839	-0.185**	-0.143*	-0.228***	-0.233***
	(0.083)	(0.076)	(0.082)	(0.081)
Censorship (sd) × 1840-1911	-0.032	-0.040	-0.085	-0.086
	(0.076)	(0.090)	(0.095)	(0.092)
Censorship (sd) × 1912-1949	-0.095	0.142	0.193	0.190
	(0.137)	(0.122)	(0.148)	(0.148)
Category FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Section × time trend	Y	Y	Y	Y
Controls × period FE	Y	Y	Y	Y
Pseudo R-squared	0.493	0.341	0.774	0.772
Observations	12,510	12,276	14,400	14,400

Note. The estimates are all from the Poisson model. The dependent variables in Column (1) and (2) are the numbers of banned keywords, derived from survived banned books and lost banned books respectively. The dependent variables in Column (3) and (4) are the numbers of unbanned keywords, which refer to the set of all keywords excluding those from banned books. Controls are the same as in Table 2. Standard errors displayed in parentheses are clustered at the book category level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 6: Censorship and Responses of Publishers

Dependent variable	# Publishers	ln(# Publishers+1)	# Books	ln(# Books+1)	ln(# Publishers+1)	ln(# Books+1)	
	Poisson	OLS	Poisson	OLS	First Stage	Reduce Form	IV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
#NewPulishers$_p$× Share$_{c,p-1}$					1.275***	0.953***	
					(0.253)	(0.202)	
Censorship (sd) × 1773-1839	-0.206***	-0.168**	0.081	-0.063	-0.132*	-0.122***	-0.023
	(0.079)	(0.076)	(0.059)	(0.047)	(0.069)	(0.041)	(0.068)
Censorship (sd) × 1840-1911	-0.086	0.030	0.057	-0.077	0.090	-0.015	-0.082
	(0.093)	(0.097)	(0.079)	(0.093)	(0.072)	(0.101)	(0.094)
Censorship (sd) × 1912-1949	0.064	0.158	0.091	-0.039	0.155	0.042	-0.074
	(0.126)	(0.151)	(0.076)	(0.090)	(0.107)	(0.097)	(0.085)
ln (# Publishers+1)$_p$			0.898***	0.523***			0.747***
			(0.056)	(0.089)			(0.122)
Category FE	Y	Y	Y	Y	Y	Y	Y
Period FE	Y	Y					
Year FE			Y	Y	Y	Y	Y
Section × time trend	Y	Y	Y	Y	Y	Y	Y
Controls × period FE	Y	Y	Y	Y	Y	Y	Y
R-squared	0.948	0.900	0.770	0.757	0.932	0.736	0.094
Observations	200	200	14,400	14,400	14,400	14,400	14,400

Note. Columns (1)-(2) present how the number of publishers respond to censorship over time. Columns (3)-(7) show that the number of publishers can explain our baseline finding on the number of books. Controls are the same as in Table 2. Standard errors displayed in parentheses are clustered at the book category level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 7: Responses of Publishers: Exit, Entry, and Surviving Publishers

Dependent variable	Changes in ln # Publsihers		
	All	Surviving publishers	Others
	(1)	(2)	(3)
Censorship (sd) × 1773-1839	-0.167**	-0.083	-0.169**
	(0.076)	(0.085)	(0.081)
Censorship (sd) × 1840-1911	0.199*	0.130	0.209*
	(0.100)	(0.087)	(0.105)
Censorship (sd) × 1912-1949	0.129	0.063	0.152
	(0.134)	(0.068)	(0.159)
Section FE	Y	Y	Y
Period FE	Y	Y	Y
Controls × period FE	Y	Y	Y
R-squared	0.801	0.509	0.817
Observations	150	150	150

Note. The estimates are from the OLS model. Surviving publishers are those who spanned at least two periods. Controls are the same as in Table 2. Standard errors displayed in parentheses are clustered at the book category level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 8: Separating Publishers and Authors

				# Books, Poisson	
				Authors	
				Died before 1772	Lived dur. 1773-1839
				(1)	(2)
Publishers	Active before 1772	(1)	Censorship (sd)	-0.126	
				(-0.092)	
	Active during 1773-1839	(2)	Censorship (sd)	-0.313***	-0.526***
				(-0.084)	(-0.114)
	Active post 1840	(3)	Censorship (sd)	-0.122	-0.655***
				(-0.102)	(0.137)
	Observations			50	50

Note. This table shows the correlation across categories between censorship and publications for five groups. The estimates are otaineed from the Poisson model. These five groups are categorized based on the periods during which the publishers were active and the lifetimes of the authors. Column (1) fixes the authors who died before 1773 and indicates possible reactions from the publishers. Row (3) fixes the publishers active post-1840s and indicates possible reactions from the authors. Standard errors displayed in parentheses are clustered at the book category level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Online Appendix

Table of Contents

A Data Construction and Description **A-2**

A.1 Banned Books A-2

A.2 Book Publication Data A-2

A.3 Measurement Error: Book Survival A-3

A.4 Measurement Error: Missing Publication Year A-4

A.5 The Level of Censorship across Categories A-5

A.6 Author-based Censorship A-6

A.7 Publisher Ownership and Location A-7

A.8 Censorship and Other Category Characteristics A-9

B Decline and Revival in Book Production: Additional Results **A-10**

B.1 Missing Publisher Information A-10

B.2 Coding Natural Sciences A-11

B.3 Category-by-Emperor Analysis A-14

B.4 Different Sources of Banned Books A-15

C Book Contents and Self-Censorship: Additional Results **A-16**

C.1 Methodology of Textual Analysis A-16

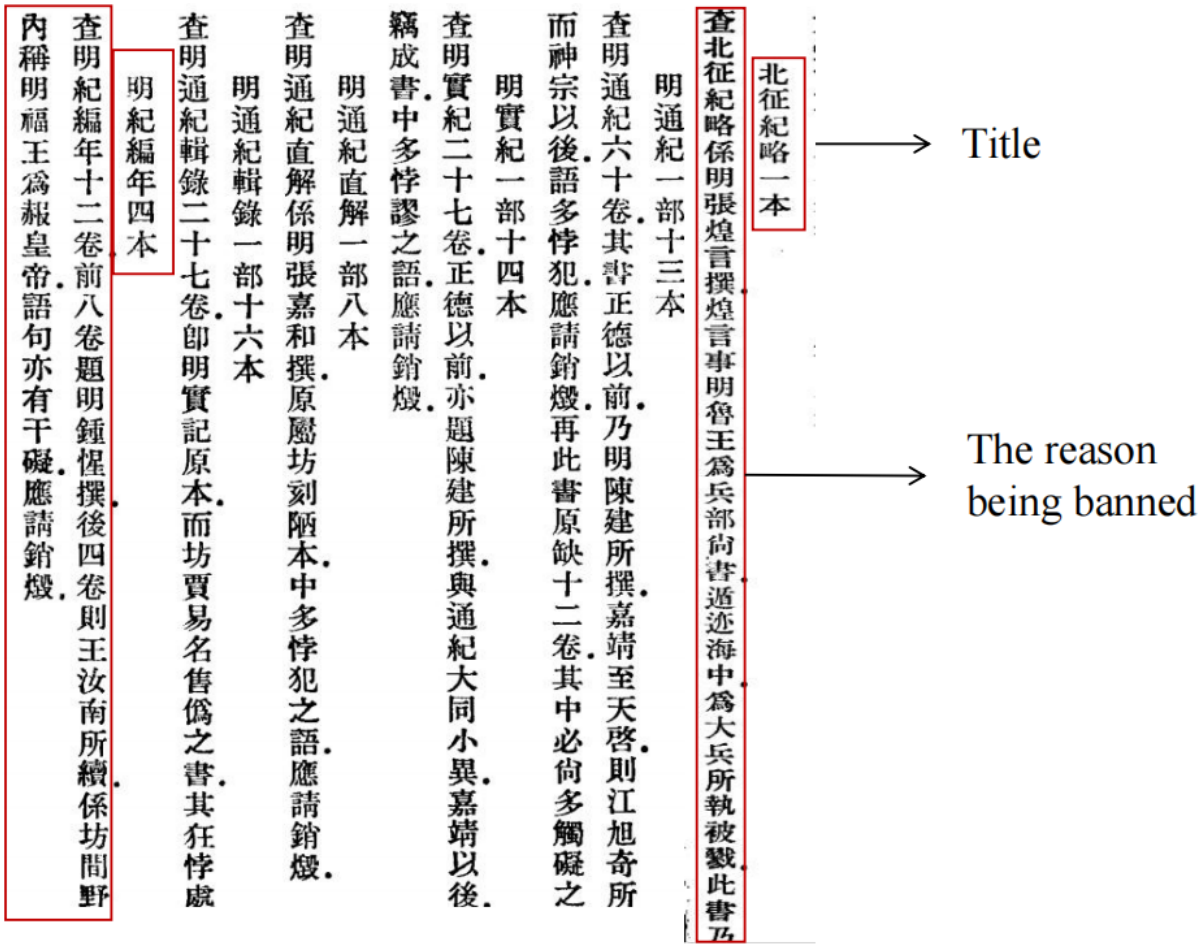
C.2 Authors' Information A-20

A Data Construction and Description

A.1 Banned Books

Figure A.1 displays an example of a banned book, detailing both its content and the reasons for its ban. In general, books are banned for three main reasons: (i) reflecting anti-Manchu or rebellion sentiment, or nostalgia for the Ming dynasty; (ii) recording of the conflicts between Manchus and Han in the late Ming dynasty; (iii) departing from the Confucianism classics and rebel against Confucian orthodoxy.

Figure A.1: Examples of Banned Books

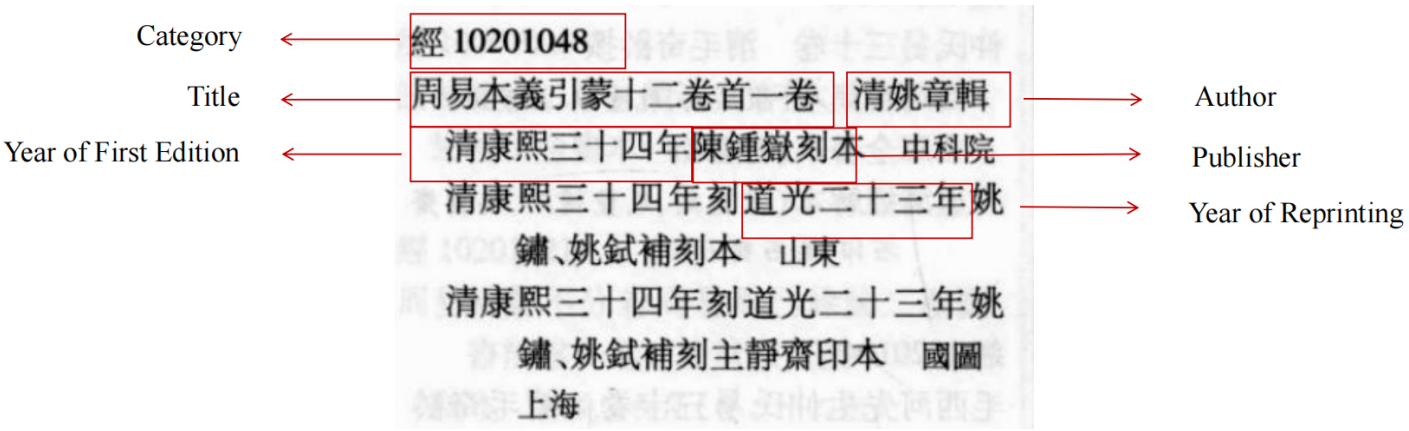


A.2 Book Publication Data

To measure knowledge production, we obtain book publication records from the *General Catalogue of Pre-modern Chinese Books*. This catalog contains records of over 180,000 books and nearly 270,000 editions. From the period spanning 1662 to 1949, it documents over 161,000 editions of

books. Figure A.2 presents an entry from the catalog. For each entry, the catalog details essential information including the publication category, the author, the publisher, publication time, whether they were reprints and year of reprinting. Based on the category and year of publication, we construct the variable to measure the annual book publications within each category. Additionally, we further refine these variables to track both first prints and reprints for each category and year in the data.

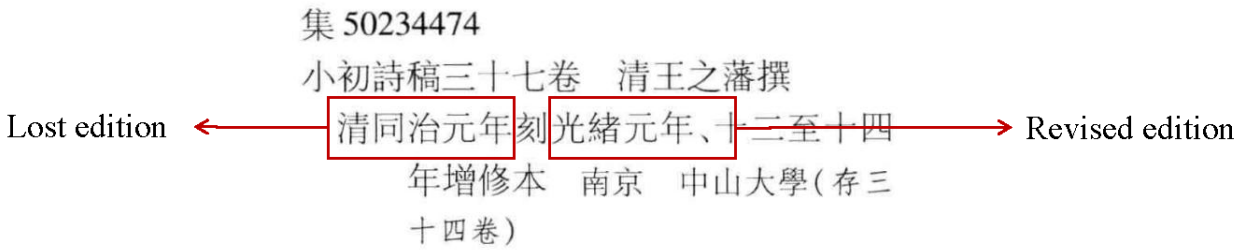
Figure A.2: Records on Book Publications



A.3 Measurement Error: Book Survival

Our data source for book publications is based on books that survived to the present. It also contains information on the lost editions of these books, which we include in our main analysis. This information provides an opportunity to examine whether book survival is highly correlated with censorship. Figure A.3 provides an example. It recorded that the edition of the book published in 1875 was revised from the edition printed in 1862, but the 1862 edition was lost.

Figure A.3: Records on Lost Editions



Based on these records, we calculate the share of lost editions for each category and examine its relationship with censorship. Table A.3 presents the result. As shown, there is no significant correlation between censorship and the share of lost editions.

Table A.3: Censorship and the Share of Lost Editions

Dependent variable	The share of lost editions					
	Books printed between 1773 and 1839			Books printed between 1662 and 1949		
	(1)	(2)	(3)	(4)	(5)	(6)
Censorship (sd)	-0.004	-0.006	-0.014	-0.003	-0.005	-0.010
	(0.008)	(0.009)	(0.013)	(0.006)	(0.008)	(0.012)
Category size		0.069*	0.078*		0.076*	0.083
		(0.039)	(0.043)		(0.038)	(0.052)
State penetration		-0.028	-0.028		-0.020	-0.020
		(0.020)	(0.018)		(0.018)	(0.018)
Private publisher		-0.064	-0.072		-0.053	-0.065
		(0.047)	(0.046)		(0.046)	(0.045)
Market Concentration		0.021	0.015		0.049	0.045
		(0.026)	(0.022)		(0.031)	(0.029)
Reprint level		-0.008	-0.006		-0.004	-0.002
		(0.006)	(0.004)		(0.006)	(0.004)
Author-based censorship		-0.020	-0.027		-0.022	-0.031
		(0.031)	(0.032)		(0.027)	(0.028)
Section FE			Y			Y
R-squared	0.004	0.272	0.359	0.003	0.212	0.335
Observations	50	50	50	50	50	50

Note. Standard errors displayed in parentheses are clustered at the book category level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

We should note that this exercise has an important limitation, as the data is based on surviving books and their lost editions. For our main analysis, the revival of book publications in more heavily censored categories post-1840s offers a valuable countermeasure to the issue of survival bias.

A.4 Measurement Error: Missing Publication Year

In the book publication records, some are missing the publication year information. We examine whether the missing probability correlates with the level of censorship and find it not to be the case. Table A.4 reports the results using the probability of book productions with missing publication year to the total book productions as the dependent variable. We find no significant correlation between the missing probability and the level of censorship. Besides, we find that the categories with higher state penetration have higher missing year probability. This is because books published

by the state publishers are usually large series of books. It is often only known during which emperor's reign they were published, without information on the year of publication.

Table A.4: Censorship and Missing Rate of Publication Years

Dependent variable	Missing rate		
	(1)	(2)	(3)
Censorship (sd)	0.140	0.126	0.050
	(0.111)	(0.080)	(0.117)
Category size		0.411	0.036
		(0.465)	(0.491)
State penetration		0.383*	0.405*
		(0.200)	(0.229)
Private publisher		0.764**	0.715**
		(0.285)	(0.311)
Market Concentration		-0.009	-0.076
		(0.230)	(0.258)
Reprint level		-0.054	-0.054
		(0.068)	(0.065)
Author-based censorship		-0.061	-0.073
		(0.220)	(0.231)
Section FE			Y
R-squared	0.033	0.302	0.369
Observations	50	50	50

Note. Standard errors displayed in parentheses are clustered at the book category level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

A.5 The Level of Censorship across Categories

Banned Book Records To measure the level of censorship, we collect records of banned books from multiple sources: (i) summaries of banned books from the editors of the Complete Library, and (ii) catalogs of banned books compiled by historians. After eliminating duplicate entries, we compile a comprehensive catalog of banned books, comprising 3,102 unique entries.

Matching with the *General Catalogue* Among the 3,102 banned books, 1,783 can be matched to the *General Catalogue*, which means that these books have survived to the present day and are clearly catelogued.

For the 1,319 lost books, we manually code their categories based on their titles and summaries. As noted earlier, for most of these banned books, the editors of the *Complete Library* wrote summaries for each book. The summaries were sufficiently detailed to determine the categories. We also validate our coding with historical research on banned books (6). Nevertheless, 272 of them do not contain detailed information, making it impossible to determine their categories. We do not consider these 272 banned books, i.e., 12% in our censorship measures. In addition, we separate the lost and surviving banned books and obtain two measures of censorship and use them in our analysis as well.

Measuring In our main analysis, we measure the level of censorship for each category by calculating the proportion of banned books relative to the total number of books collected in that category:

$$Censor_c = \frac{N_{banned}}{N_{banned} + N_{Siku}}. \tag{2}$$

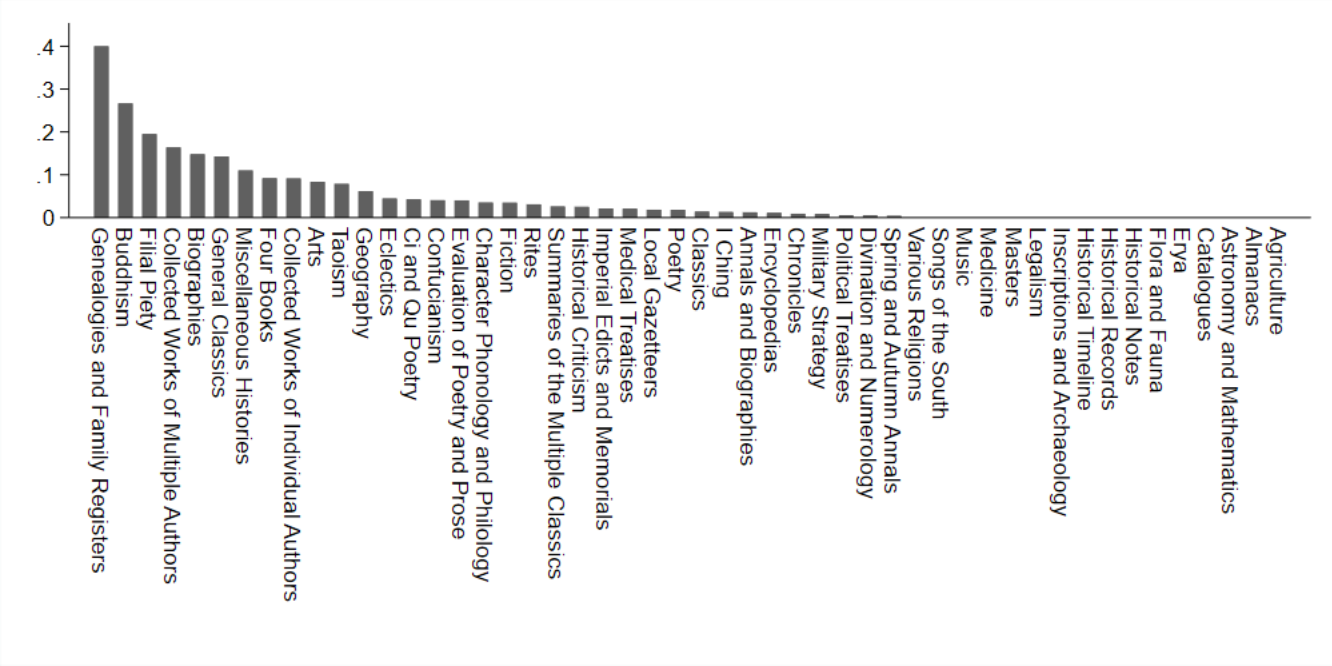
where N_{banned} denotes the number of banned books, and N_{Siku} denotes the number of books in the *Complete Library*, including both the full-text books and indexed books. We also separate lost and surviving banned books in our additional analysis.

A.6 Author-based Censorship

As discussed in the main text, 28% of the banned books were due to authors, encompassing those penned by banned authors or referencing their works. A well-known instance is Lu Liuliang, a prominent scholar from the late Ming dynasty who declined to serve in the Qing administration.

We determine the extent of author-based censorship using the same approach as for content-based censorship. Figure A.6 illustrates the level of author-based censorship across different categories. The trend diverges significantly from content-based censorship. For example, only one history-related category appears among the top 10 categories with the highest censorship levels. Most categories pertain to personal details, such as genealogies, biographies, and collected works, aiming to suppress the authors and their personal information.

Figure A.6: Author-based Censorship across Categories



A.7 Publisher Ownership and Location

The *General Catalogue* provides the basic information for each book, including the publishers. However, additional information on the publishers is incomplete and crude. For example, we sometimes do not know the ownership of publishers. To improve this information, we collect additional historical data on the printing market and publishers and integrate them with the *General Catalogue*. These historical data sources include (Du, 2001; Du, 2009; Zhai, 2009), which document the catalogue of publishers during the Ming and Qing dynasties, including the names of publishers, the ownership of publishers and the books they printed.

Figure A.7(I) provides an example of how to match the historical data on publishers to the *General Catalogue* data. Based on these sources, we construct a dataset of publishers, which includes the unique ID, ownership, the origin of the publisher, and the address of the publishers.

Figure A.7: I. Matching Publishers from the Two Historical Sources

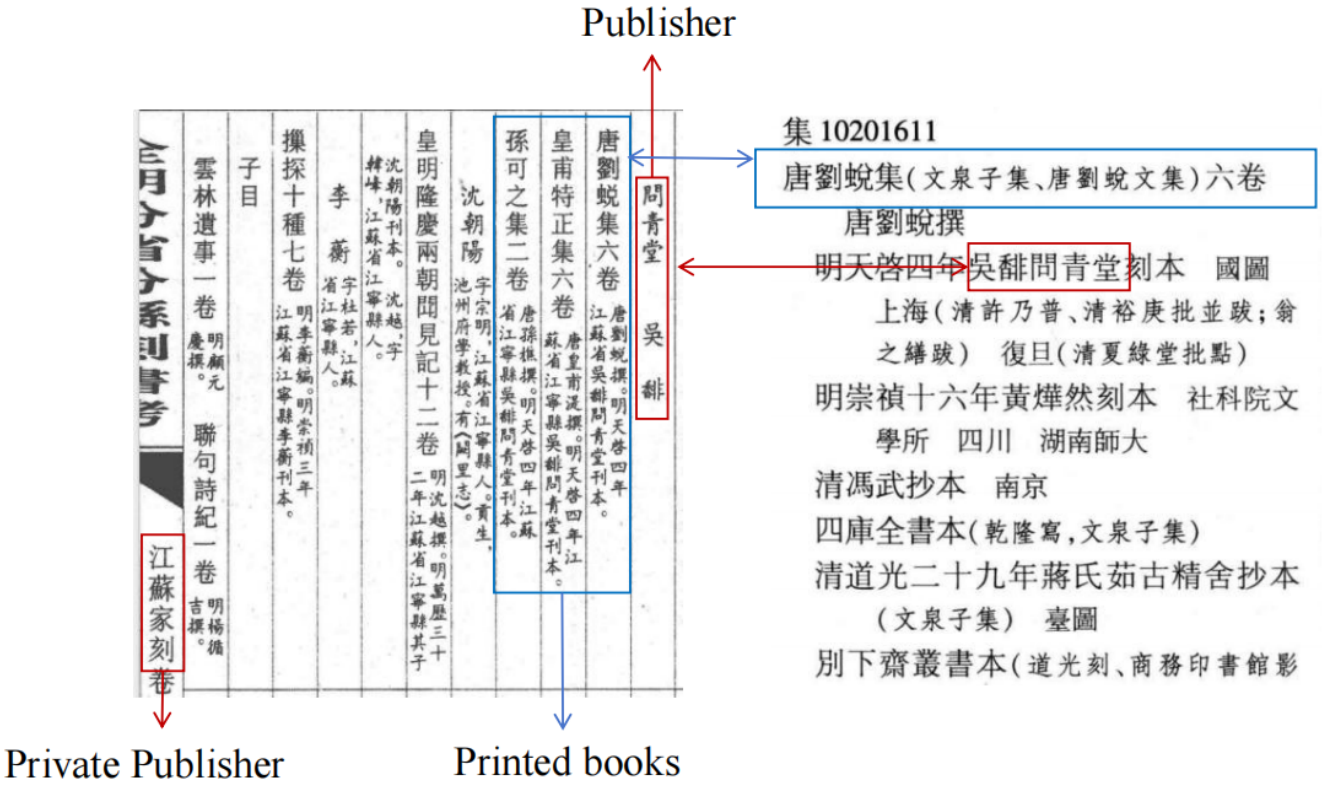
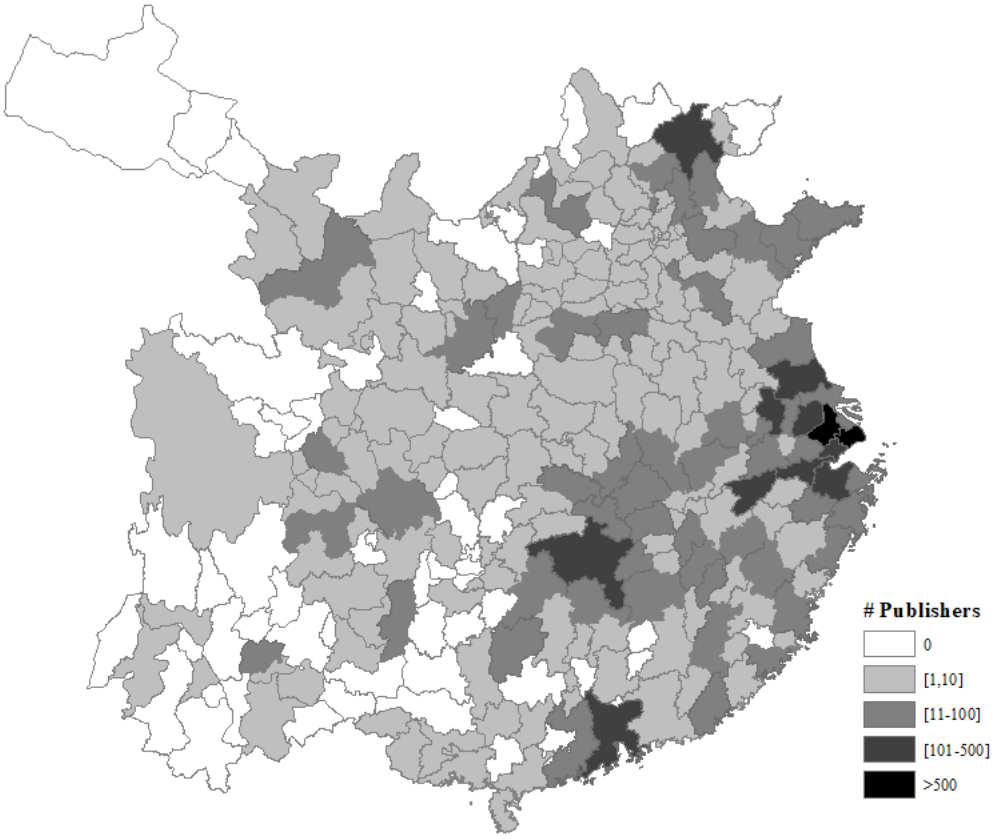


Figure A.7(II) illustrates the spatial distribution of publishers. The dispersed locations reflect a decentralized printing market in this era.

Figure A.7: II. Spatial Distribution of Publishers



A.8 Censorship and Other Category Characteristics

Table A.8 shows that there are no evident relationships between the degree of censorship and other category-level traits, indicating that factors such as size or market considerations do not significantly influence censorship.

Table A.8: Censorship and Other Category Characteristics

Dependent variable	Censorship (sd)		
	(1)	(2)	(3)
Category size	0.235	0.293	0.467
	(0.468)	(0.453)	(0.481)
State penetration	-0.191	-0.196	-0.169
	(0.166)	(0.170)	(0.178)
Private publishers	0.042	0.005	-0.087
	(0.426)	(0.470)	(0.495)
Market concentration	0.016	0.014	-0.199
	(0.450)	(0.451)	(0.446)
Reprinted level before Qing	-0.037	-0.043	0.006
	(0.072)	(0.076)	(0.062)
Missing rate	0.044	0.051	0.149
	(0.174)	(0.185)	(0.238)
Author-based censorship		-0.113	-0.132
		(0.330)	(0.375)
Section FE			Y
R-squared	0.021	0.023	0.249
Observations	50	50	50

Note. Standard errors displayed in parentheses are clustered at the book category level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

B Decline and Revival in Book Production: Additional Results

B.1 Missing Publisher Information

In the book publication records, nearly half lack publisher information. Table B.1 reports the results using the share of books without publisher information in category i during period t as the dependent variable. We find no significant correlation between the missing probability and the level of censorship.

Table B.1: Censorship and Missing Rate of Publishers

Dependent variable	Share of books w/o publisher info.		
	(1)	(2)	(3)
Censorship (sd) × 1773–1839	0.003	0.009	0.007
	(0.022)	(0.021)	(0.024)
Censorship (sd) × 1840–1911	-0.007	0.005	-0.002
	(0.024)	(0.021)	(0.020)
Censorship (sd) × 1912–1949	-0.043	-0.025	-0.027
	(0.027)	(0.023)	(0.027)
Category FE	Y	Y	Y
Period FE	Y	Y	Y
Section × time trend		Y	Y
Controls × period FE			Y
R-squared	0.727	0.768	0.806
Observations	200	200	200

Note. Standard errors displayed in parentheses are clustered at the book category level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

B.2 Coding Natural Sciences

Natural sciences are not clearly categorized in the historical publication classifications. We develop a method to match the historical classifications with the modern ones, using the following procedure.

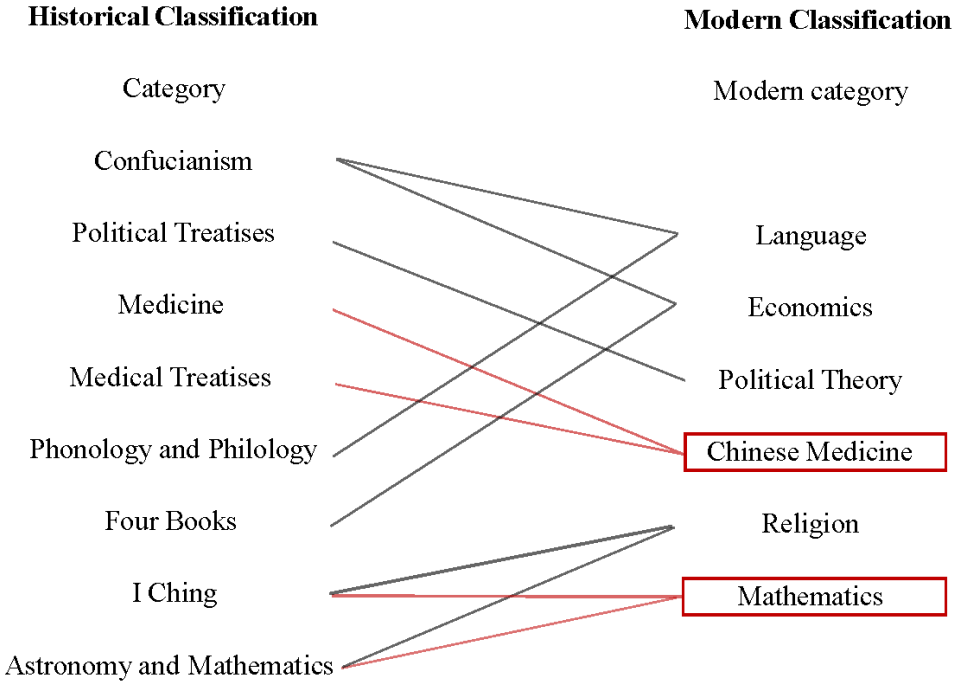
Nearly 4,000 historical books were republished after 1949, on which there exist modern classification. Thus, we know both the historical and modern classification for these republished books. Figure B.2(I) provides an example. *Exploiting the Works of Heaven* belongs to the category of political treatises by historical classification, while it belongs to natural science theory by modern classification.

Figure B.2: I. Republication of Historical Books in Modern Times: An Example



These republished books provide us with a channel to match historical classifications with modern ones. Figure B.2(II) shows an demonstration for matching.

Figure B.2: II. Matching Historical Classification with Modern Classification



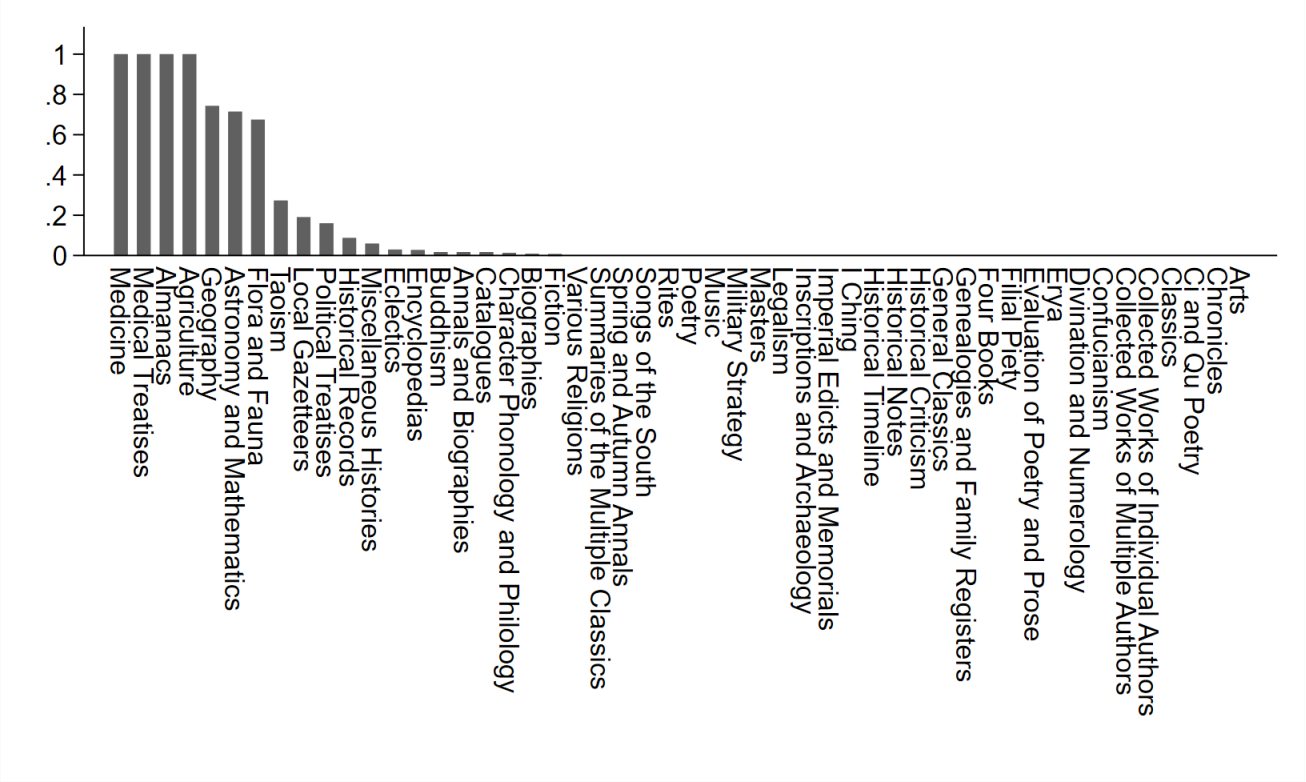
In each historical category, some can be matched to natural sciences and others can be matched to social sciences. We then calculate the share of natural sciences within each historical category as follows:

- Let N_i denote the number of republished books in an historical category i;
- In each historical category i, we count the number of books matched to different modern categories. Let N_{ij} denote the number of books matched to modern category j in historical category i;
- We then classify the modern categories into natural science and social science based on the modern classification. Let S_{in} denote the share of books matched to natural science in category i:

$$S_{in} = \frac{\sum N_{ij}}{N_i}, \text{ if } j \text{ belongs to natural sciences} \tag{3}$$

Figure B.2(III) illustrates the proportion of natural science within each historical category. In general, approximately 14% of the books are related to natural sciences, showing that they represented a minor portion of the total publications. The top 5 categories with the greatest share of natural sciences are: Medicine, Medical Treatises, Agriculture, Astronomy and Mathematics, and Geography.

Figure B.2: III. Share of Books in the Natural Science



Note. Figure B.2 plots the share of books belongs to the natural science in each historical category among the reprinted books after 1949.

B.3 Category-by-Emperor Analysis

Table B.3 reports the results using the number of books published during each emperor's reign in each category as the dependent variable. We find that the magnitudes of the censorship effects are similar to our baseline results. Due to a much smaller sample size, the estimates are less precisely estimated than our baseline.

Table B.3: Robustness Checks: Category-by-EmperorAnalysis

Dependent variable	ln(# Books+1)	# Books
	OLS	Poisson
	(1)	(4)
Censorship (sd) × 1773-1839	-0.130	-0.140*
	(0.086)	(0.082)
Censorship (sd) × 1840-1911	0.025	-0.065
	(0.119)	(0.112)
Censorship (sd) × 1912-1949	0.207	0.200
	(0.164)	(0.130)
Category FE	Y	Y
Emperor FE	Y	Y
Section × time trend	Y	Y
Controls × period FE	Y	Y
R-squared	0.918	0.957
Observations	600	600

Note. Standard errors displayed in parentheses are clustered at the book category level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

B.4 Different Sources of Banned Books

In our baseline analysis, we combine the records of two sources and delete the duplicate ones to get the number of banned books and calculate the level of censorship. Table B.4 shows the results using the level of censorship calculated based on the number of banned books from different sources.

Table B.4: Measurement of Censorship: Different Sources of Banned Books

Dependent variable	ln(# Books+1), OLS		# Books, Poisson	
	Official	Historians'	Official records	Historians'
	(1)	(2)	(3)	(4)
Censorship (sd) × 1773-1839	-0.142***	-0.168***	-0.134*	-0.157*
	(0.040)	(0.039)	(0.080)	(0.092)
Censorship (sd) × 1840-1911	-0.054	-0.125	-0.040	-0.093
	(0.106)	(0.081)	(0.123)	(0.098)
Censorship (sd) × 1912-1949	0.037	-0.026	0.203	0.190
	(0.113)	(0.085)	(0.149)	(0.138)
Category FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Section × time trend	Y	Y	Y	Y
Controls × period FE	Y	Y	Y	Y
R-squared	0.728	0.731	0.750	0.753
Observations	14,400	14,400	14,400	14,400

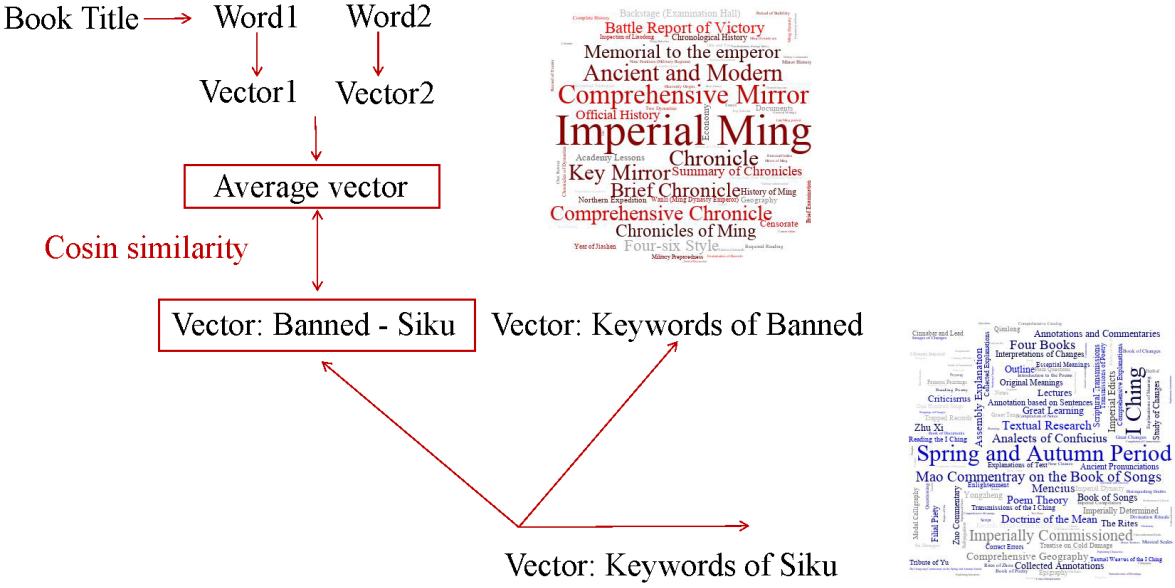
Note. Standard errors displayed in parentheses are clustered at the book category level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

C Book Contents and Self-Censorship: Additional Results

C.1 Methodology of Textual Analysis

In Section 5.1, we use the natural language processing(NLP) techniques to categorize books into two groups: sensitive and less-sensitive books, and examine the importance of self-censorship. This section outlines the textual analysis process, which is conducted in four steps summarized below. Figure C.1(I) illustrates these key steps.

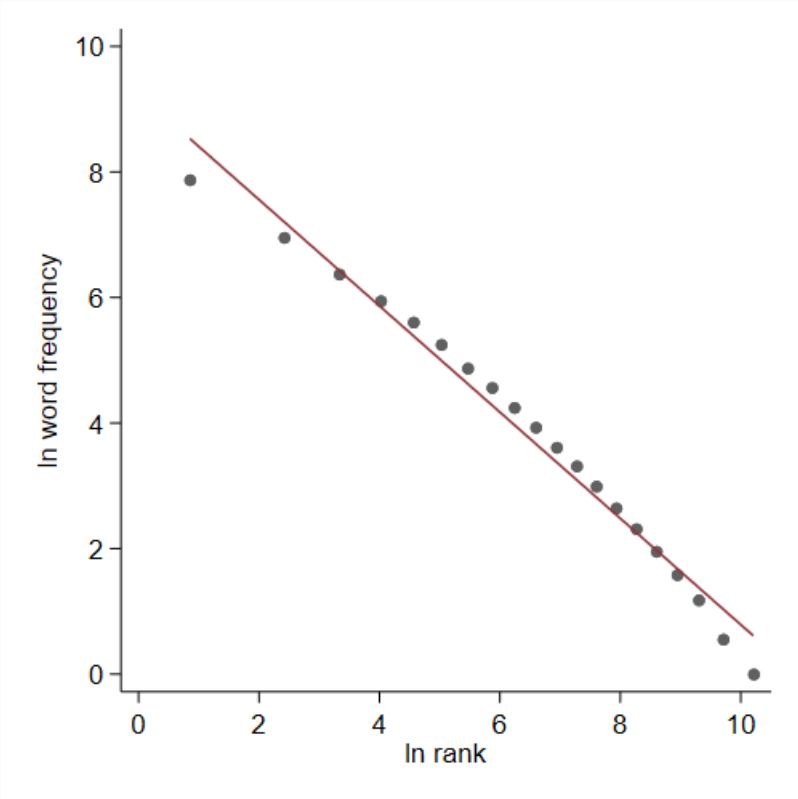
Figure C.1: I. Word Embedding and Similarity Analysis



Building Corpora A corpus is a set of texts or words that are used in natural language processing. It serves as a foundational database from which various linguistic patterns can be extracted and analyzed, including word frequency, collocations (how words are commonly paired together), syntactic structures, and more.

In this paper, we build our corpus using book titles. We first segment each book title into words according to their lexical meaning and then clean the data following conventional textual analysis procedures. Such procedures includes (i) removing non-textual elements such as numbers, years, and the names of their study rooms; (ii) eliminating words that are frequently occurring yet provide minimal unique information, such as collected works and collection of poems. Ultimately, we compile a set of words, which contains 34,053 unique words with a combined word frequency of 260,850. Figure C.1(II) displays the relationship between word frequency and its rank, demonstrating a perfect alignment with Zipf's Law. It is an empirical law that describes a specific pattern of distribution in natural languages, where a few words are used very frequently, while the majority are used rarely.

Figure C.1: II. Zipf's Law: Word Frequency and Rank



Finding Keywords After constructing the corpus, we document keywords that characterize the banned books and the *Complete Library* full-text books. The first group represents forbidden knowledge, while the second signifies orthodox knowledge.

For each word w, we calculate its frequency in the word sets of both the banned books and the *Complete Library* full-text books. Additionally, we compute the total frequencies of all other words (excluding w) in these two sets. Let f_b and f_s represent the total frequencies of word w in the word sets of banned books and the Complete Library, respectively. Similarly, let $f_{\sim b}$ and $f_{\sim s}$ denote the total occurrences of words other than w in the word sets of the banned books and the *Complete Library*. We use χ_w to denote Pearson statistic for each word in the word set of banned books.[8]

$$\chi_w^2 = \frac{(f_b f_{\sim s} - f_s f_{\sim b})^2}{(f_b + f_s)\,(f_b + f_{\sim b})\,(f_s + f_{\sim s})\,(f_{\sim b} + f_{\sim s})}. \tag{4}$$

By assuming that the counts are drawn from multinomial distributions. χ_w is a test statistic

[8]Simpler statistics, such as the ratio of word usage by the Siku Quanshu compared to the banned books, could bias selection towards words that are used infrequently by the Siku Quanshu and not at all by the banned books.

Table C.1: Top 20 Keywords from the Banned Books and the *Complete Library* Full-text Books

Complete Library	Banned Books
Spring and Autumn Period	Ming Dynasty
I Ching	Memorial to the emperor
Imperially Commissioned	Imperial Ming
Book of Documents	Comprehensive Mirror
Mao Comentary on the Book of Songs	Past and Present
Four Books	Chronicle
Textual Research	Key Mirror
Analects of Confucius	Comprehensive Chronicle
Mencius	Brief Chronicle
Poem Theory	Chronicles of Ming
Doctrine of the Mean	Four-six Style
Comprehensive Geography	Battle Report of Victory
Assembly Explanation	Summary of Chronicles
Rites	Official History
Great Learning	Backstage (Examination Hall)
Lectures	Economy
Outline	Academy Lessons
Annotations and Commentaries	Censorate
Imperial Edicts	Chronological History
Zhu Xi	History of Ming

whose null hypothesis is that the propensity to use word w is equal for the *Complete Library* and the banned books. Consequently, a significantly high value of this statistic suggests that the word is frequently used within a particular word set. The rationale behind this method is that if a word appears with unusually high frequency in one corpus compared to another, it likely plays a crucial role in characterizing that corpus.

To select the words for our analysis, we organize the statistic values for the words used in both the *Complete Library* and the banned books. We then choose the words that fall within the highest 5% of these values for each group. Ultimately, we identify 88 keywords for the banned books and 141 keywords for the *Complete Library*. Table C.1 lists the top 20 most frequently occurring words within the set of keywords. Notably, the keywords associated with the banned books predominantly relate to various history topics, while those for the *Complete Library* are primarily linked to Confucian classics.

Word Embedding For similarity analysis, we convert words into vectors using a word embedding model. Essentially, a word embedding model leverages the co-occurrence of words to develop representations within a relatively low-dimensional Euclidean space (Mikolov et al., 2013). The

basic intuition behind it is that words appearing in similar contexts are likely to have similar meanings. The process typically involves training a neural network model on a specific corpus with the goal of either predicting a word based on its surrounding context (Continuous Bag of Words, or CBOW) or predicting the surrounding context based on a word (Skip-Gram model). Through the optimization process in training, these models generate word embeddings, which encode semantic information such that words with similar meaning are positioned closely together in the vector space. Thus, each word is represented as a vector in a continuous space where semantically similar words located in proximity to one another.

Using the pretrained model provided by Mikolov et al. (2018), we represent each word by a 300-dimension vector. We then calculate the weighted vectors based on the word frequency. Ultimately, we derive two distinct vectors, one representing the banned books and the other representing the *Complete Library*:

$$\overrightarrow{Keywords}_{\text{Banned books}} = [0.035, 0.079, 0.231, -0.013, \ldots] \tag{5}$$

$$\overrightarrow{Keywords}_{\text{Complete library}} = [0.034, 0.071, 0.313, -0.066, \ldots] \tag{6}$$

Measuring Similarity Difference between the vectors representing the banned books and the *Complete Library* serves as the benchmark for similarity analysis. Let $\overrightarrow{Booktitle}_{\text{i}}$ denote the word representation of book i, the similarity of any book to the banned books can be computed as:

$$\text{sim}_i = \frac{\overrightarrow{Booktitle}_{\text{i}} \cdot \overrightarrow{Benchmark}}{\left\|\overrightarrow{Booktitle}_{\text{i}}\right\| \times \left\|\overrightarrow{Benchmark}\right\|} \tag{7}$$

C.2 Authors' Information

Our primary source of authors' information is the China Biographical Database Project (CBDB, 2019). Figure C.2 shows an instance from this database, including authors' names, birth and death years, and hometowns, along with their works. This aids us in identifying authors with identical names. Additionally, we cross-check author details using Wikipedia and Baidu Baike. In total, we have comprehensive information on 7,907 authors, out of which 5,805 lived after 1662. We use this subsample of authors and their living eras to distinguish the responses of publishers and authors in our analysis.

Figure C.2: Records on Authors' Information from China Biographical Database Project



References

[1] **Du, Xinfu.** 2001. *A Comprehensive Survey of Book Publishing by Province and County in the Ming Dynasty (Quan Ming fensheng fenxian keshu kao).* Thread-Binding Book Bureau.

[2] **Du, Xinfu.** 2009. *A Comprehensive Survey of Book Publishing by Province and County in the Qing Dynasty (Quan Qing fensheng fenxian keshu kao).* Thread-Binding Book Bureau.

[3] **Harvard University, Academia Sinica, Peking University.** 2019. *China Biographical Database (CBDB).* https://projects.iq.harvard.edu/cbdb.

[4] **Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin.** 2018. "Advances in Pre-Training Distributed Word Representations." *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).*

[5] **Mikolov, Tomás, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean.** 2013. "Distributed Representations of Words and Phrases and their Compositionality." *Proceedings of the 26th International Conference on Neural Information Processing Systems*, 2: 3111–3119.

[6] **Wang, bin.** 1999. *An Overview of Banned Books in the Qing Dynasty (Qingdai jinshu zongshu).* China Book Publishing House.

[7] **Zhai, Mianliang.** 2009. *A Dictionary of Historical Chinese Book Printing (Zhongguo guji banke cidian).* Suzhou University Press.