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

Ying Bai† Ruixue Jia‡ Jiaojiao Yang §

March 18, 2025

Abstract

This study examines the short-, medium-, and long-term impacts of state censorship on knowledge production, focusing on the largest book-banning campaign in Chinese history, initiated during the compilation of the *Siku Quanshu* (Complete Library in Four Sections) between 1772 and 1783. Analyzing publication data from over 161,000 books spanning the 1660s to the 1940s, we find that categories subjected to stricter censorship experienced significant declines in publication during the seven decades following the bans (1770s–1830s). However, beginning in the 1840s, political upheavals and the erosion of state control triggered a resurgence of publications in previously restricted categories. This pattern of suppression and recovery extends to ideas, as evidenced by text analysis of book titles. Further analysis highlights the central role of chilling effects in driving these dynamics, with the exit and entry of publishers playing a crucial role in both the suppression and subsequent revival of knowledge production.

*We thank Scott Gehlbach, Matthew Gentzkow, Sergei Guriev, Jennifer Pan, Harry Pickard, Tuan Twee Sng, Konstantin Sonin, David Stromberg, Shaoda Wang, Jaya Wen, David Yang, Li-an Zhou, Katia Zhuravskaya and conference and seminar participants at the AEA-ASREC meeting, Arthur Lewis Lab Conference, Monash, National University of Singapore, Oxford, Singapore Management University, 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 the *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 books. Among these, about 10,000 were deemed state-approved knowledge, with their complete texts included or titles indexed, while another roughly 3,000 were classified as banned books, representing forbidden knowledge. Although the books banned represented only a minor fraction of the total books, they served as a benchmark for 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 database to document book publications for nearly three centuries (1660s–1940s). Our data includes

[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 categories, was fully developed around 600 CE and was employed in subsequent dynasties (Zhou, 1996).

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 and contents 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 significant spillover effects and potential self-censorship, a widespread phenomenon that we can explore 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 the forced opening of parts of the country and diminished government control. Many scholars consider the 1840s as the beginning of "modern China" (e.g., Fairbank and Goldman, 2006). Furthermore, the Qing dynasty's fall in 1911 accentuated ongoing 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 seven decades succeeding the bans, from the 1770s to the 1830s, we find a significant decline in publications for categories under stricter bans. Specifically, a one standard deviation increase in the share of banned books in a category is associated with an 18% 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*. To clarify, the lack of pre-trends does not imply the Qing

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

[4] For example, film censorship often includes vague guidelines on prohibited topics, yet films may still be censored without explicit reasons. Similarly, lawyers and journalists frequently navigate unclear censorship boundaries in authoritarian regimes (e.g., Stern and Hassid, 2012).

state before the *Complete Library* compilation was devoid of censorship, but rather highlights the impacts of systematic idea-based censorship (see more discussion on the background in Section 2.1).

To better understand the revival around the 1840s, we examine regional variations and find that it first emerged in treaty ports and prefectures along the coast or the Yangtze River. This pattern is consistent with the fact that these areas were among the earliest to experience a decline in state control.[5] Multiple important changes, including the forced opening of parts of the country and the weakening of state capacity due to wars, likely contributed to this decline in state control.

Further heterogeneity and comparison analyses reveal that our findings are primarily driven by books published by non-state publishers rather than state-published works. We assess the role of the natural sciences by aligning historical and modern book classifications and find similar dynamics within natural science categories, even though natural sciences comprised only 11% of total publications. We also consider author-based censorship in our analysis and address potential measurement error in publication data.

Our second analysis examines the content of book titles to identify which types were affected, which reveals the importance of chilling effects for our main finding. Using a method similar to Gentzkow and Shapiro (2010), we extract keywords from book titles and analyze the numbers of keywords and *unique* keywords as the dependent variables to approximate book content. The results indicate a decline in the prevalence of these keywords from the 1770s to the 1830s, with a resurgence after the 1840s. This pattern suggests that, alongside the increased volume of books in heavily censored categories, previously restricted topics also began to reappear.

When we separate keywords into banned and unbanned categories, we find both responded dynamically to censorship pressures. However, unbanned keywords played a particularly important role in shaping our overall findings, as they made up the majority of publications. When we categorize keywords as either new or pre-existing based on their initial appearance, we observe that new keywords also experienced a cycle of decline and revival, implying spillover effects on the production of new topics from censorship.

Next, we categorize books into "sensitive" and "less-sensitive" groups based on their similarity to banned titles and *Complete Library* full-text titles. Both groups show cycles of decline and recovery. Since sensitive books constituted a smaller share of total publications, changes in less-

[5]Treaty ports refer to cities that were forced to open to Western nations following the Opium War of 1840 and subsequent conflicts. These cities typically held extraterritorial status and were among the first to undergo significant growth and industrialization after the 1840s (e.g., Jia, 2014).

sensitive books accounted for the majority of observed trends across both the decline and revival phases. We interpret these shifts in less-sensitive books as evidence of a chilling effect and associated self-censorship, as they were not directly banned. Additionally, we evaluate the impact of both lost and surviving banned books, finding that both had similar dynamic effects on knowledge production. Once again, this finding highlights the central role of chilling effects in our results, beyond any potential knowledge loss due to 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.

Our study contributes to the growing economics literature on censorship by demonstrating its broad and lasting effects on knowledge creation. Previous research has shown that censorship hinders intellectual development in Europe under religious censorship (e.g., Becker, Pino and Vidal-Robert, 2021; Blasutto and De la Croix, 2023; Comino, Galasso and Graziano, 2024), distorts the information supply in modern China (Qin, Strömberg and Wu, 2018), shapes content production in U.S. cinema (Tan and Wang, 2024), and diminishes demand for restricted information in controlled experimental settings (Chen and Yang, 2019). Different from studies of religious censorship in Europe that primarily focus on the direct impacts on censored authors and publishers, we leverage the long-standing bibliographic classification system in Chinese publications to examine category-level censorship (a proxy for topics and ideas) and uncover significant chilling effects on new publications.

Our findings reveal that even though censored books account for only a small fraction of total publications, the chilling effects on knowledge production can be substantial. Moreover, these effects are dynamic, fluctuating with the degree of state control, evidenced by a resurgence of publications in previously censored categories after the 1840s. This dynamic interplay between repression and resilience in book production represents a novel contribution to the literature.

Importantly, the resurgence of publications after the 1840s should not be interpreted as evidence

that censorship's detrimental impact is negligible. From the 1770s to the 1830s – a period often characterized as one of intellectual stagnation – China experienced significant population growth but saw limited innovation and cultural development. Influential thinkers such as Liang Qichao later speculated that censorship played a pivotal role in this stagnation (Liang, 1921). This intellectual stagnation contrasts sharply with the contemporaneous flourishing of progressive ideas and technological advancements in Europe during the Industrial Revolution (e.g., Mokyr, 2016; Almelhem et al., 2023). The suppression of knowledge in China during this critical period may have had important political and economic consequences, hindering the country's ability to adapt to and engage with the transformative global changes of the era.

The related literature also looks at historical instances of literary inquisitions[6] in censorship processes and identifies long-term detrimental effects on economic development through channels such as social and human capital (e.g., Xue, 2021; Drelichman, Vidal-Robert and Voth, 2021; Wang, 2022; Dewitte, Drago, Galbiati and Zanella, 2024). A typical challenge in such long-term analyses is the scarcity of data across short, medium, and long timescales. By analyzing book production across both short- and long-term periods, our study not only confirms the negative effects over several decades but also suggests that societal actors – publishers, authors, and readers – can modify their behaviors in reaction to important 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; Chaney, 2016; Biasi and Moser, 2021; Almelhem et al., 2023; Duan and Zhang, 2024). Our results highlight 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.

[6] In our context, literary inquisitions represent a small subsample of book ban cases, characterized by the imposition of severe punishments on those implicated. For more information, refer to Section 2.2.

2 Historical Background

This section describes the creation of the *Complete Library* and its related censorship. We discuss the characteristics of this censorship and outline the publication landscape of our study period.

2.1 The *Complete Library* and Related Censorship

The *Complete Library in Four Sections* is regarded as one of the most extensive collections of Chinese literature. The four sections represent China's book classification system, which became stable during the Tang dynasty (618–907 CE) and used in subsequent dynasties: Classics (Confucian works), Histories (historical genres), Masters (non-Confucian schools, science, technology, and religion), and Collections (literature). Each section was further divided into 10–15 categories.

The *Complete Library* project was 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. 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 included with their complete texts. In addition, about 6,700 additional books were indexed and included in the library (Encyclopedia of China, 2021).

The creation of the *Complete Library*, however, as also marked by systematic censorship. Concerned about the Qing dynasty's legitimacy as a minority Manchu ruling over a predominantly Han population, the Qing emperors always sought to suppress dissent or criticism. Prior to the *Complete Library* campaign, censorship was relatively sporadic and typically author specific, but the creation of the library led to systematic banning of certain books considered politically sensitive or critical of Qing authority, and facilitated idea-based censorship. Approximately 3,102 books were banned. These bans occasionally resulted in Literary Inquisition cases, leading to the execution of those involved. There were about 100 such cases during the Qianlong Era. Both government officials and the public were encouraged to report controversial content, creating an atmosphere of fear and self-censorship among academics. Although official records of banned books exist, historians suspect these may be incomplete, leading to slight discrepancies in different lists. In our study, we draw upon official lists as well as historical research.

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 could have 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 *probabilistic 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 Literary Inquisitions, 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 played a pivotal role in fostering a large, educated population. Confucian values also emphasized expression, often through writing, as a means of achieving lasting significance, which necessitated extensive reading.[7] This cultivated intellectual environment generated 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 data and map in Appendix A.7).

Another factor that added to the decentralized nature of publishing was the choice of many individuals to publish their works and circulate them within their literate communities (Zhang, 1989), guided by Confucian values of expression and seeking recognition rather than profit. 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. Our main focus, however, is not on the banned books, which constituted a minor fraction of the total publications, but on examining how the implementation of book bans impacted the broader process of creating and spreading new knowledge.

3 Data and Measurement

As mentioned above, China's book classification system has a long history, beginning in the Han dynasty (202BC–220 CE) and stabilizing during the Tang dynasty (618–907 CE) with the

[7] A classical Confucian ideal of personal and moral excellence emphasizes virtue (*li de*), action (*li gong*), and expression (*li yan*), known as three ways of achieving lasting significance (*san bu xiu*).

"four-section system". Each section was divided into 10–15 categories, totaling 50. This system remained stable over time, even as the number of books grew during the Ming and Qing dynasties, and was used by both state and private collections. For our study, it is useful that classifications were fixed, not influenced by book production or censorship. Moreover, book classification was common knowledge for knowledge producers. Since the Han dynasty, numerous bibliographies were compiled –official, private, and specialized – making it easy for authors and publishers to access classification information (Zhang, 1989).

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 Catalog of Pre-modern Chinese Books* (General Catalog Editorial Office, 2012). In our studied period, 1662–1949, the Catalog included over 161,000 books, with information on their publication category, the author, publication time, and whether they were reprints (see an example of the records in Appendix A.2). About 51% of the books also include information the publishers. Despite relatively rich information, there are two key measurement concerns in the book publication data we need to address.

First of all, the *General Catalog* was assembled from 1992 to 2009 and is based on books that have survived. It is well-known that many books were destroyed and lost in wars and conflicts. If a surviving book has several editions with some earlier editions missing, this is recorded. However, it could be suggested that books from more heavily censored categories are less likely to survive. We address this interpretation with three points. First, our dynamic results show a revival of publications in the more censored categories after the 1840s, which alleviates concerns about survival bias. Second, we examine the relationship between censorship and the proportion of missing editions by category (see Appendix A.3) and find no systematic relationship. While this evidence is only suggestive – relying as it does on surviving books with missing editions – it indicates that censorship alone does not account for these gaps. Finally, even if the decline observed from the 1770s to 1830s is partially influenced by the possibility that more censored categories were less likely to be disseminated and preserved during periods of suppression, censorship still imposes a cost on access to knowledge. While diffusion limitations may contribute to the observed decline, they do not fully explain the subsequent revival.

Second, 28.1% of the books lacks specific publication 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.[8] 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 written by the editors of the *Complete Library* (Zhang, 1997); (ii) the Catalogs 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, defined as $Censor_c = \frac{N_{banned}}{N_{banned}+N_{Siku}}$, to measure category-level censorship. Here, N_{Siku} includes both full-text and indexed books in the *Complete Library*. A complexity in this calculation is the uncataloged status of the lost books by the *General Catalog*. 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 separate lost from 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 include Erya (a dictionary-like glossary), genealogies and family registers, general classics, legalism, masters, medical treatises, and medicine.

There were two types of book bans: the majority (72%) were content-based, while a smaller fraction were author-based, typically due to accusations of anti-government activities against certain authors. 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,

[8]The likelihood of missing publication years decreases over time (defined according to emperors), but it does not show a correlation with censorship.

buddhism, filial piety, collected works 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.

Information on Publishers As noted earlier, 51% of the books in our dataset include information about their publishers. We assess whether the absence of publisher data affects our findings when using this information.

To classify publishers as either state or non-state entities, we use historical press records (Du, 2001; Du, 2009; Zhai, 2009). A detailed explanation of the data construction process is provided in Appendix A.7. Our study period covers a total of 18,080 publishers.

The publisher data reveals two interrelated patterns that align with the discussion in Section 2.3 about the decentralized nature of the publishing industry. First, the industry was geographically dispersed across the nation, as illustrated in the spatial distribution map of publishers in Figure 5. Second, many books were produced by small-scale private publishers. In our dataset, 30% of publishers had published more than one book, and the relationship between the number of books published and publisher rank follows a power-law distribution (see the figures in Appendix A.7): a small number of publishers produced a large volume of books, while many publishers issued only one book.

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. Market factors include category size measured by the share of books in each category among all books and its squared term (to allow for a flexible size effect), HHI of publisher concentration in each category, and reprinting share in each category (measured by the ratio of the number of editions to the number of books). Political factors include the share of books printed by state publishers in each category and the level of author-based censorship in the same way as we construct the level of content-based censorship. Additionally, we consider the missing-year rate in each category. These variables are all based on publications before 1662, the start year of our main analysis.

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 allow for 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 the level of censorship in their categories, 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 1830s. This disparity, however, did not persist forever as the more heavily censored books eventually caught up in latter decades.

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:

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

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 such observations of zeros,[9] we start with a linear model using $ln(\#Bookproduction_{c,t} + 1)$ as the dependent variable, and then employ the Negative Binomial and Poisson models 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 and their interaction with the period dummy ($\mathbf{X}_c$*$\mathbf{I}_t$), allowing the effects of these characteristics to vary across different periods. The pre-1662 characteristics variables refer to category size and its square term, HHI index, reprinting share, state publisher share, degree of author censorship, and the probability of missing years.

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.

We anticipate that $\beta_1 < 0$ 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

[9] In our category-year data, 36% of the $Book_{c,t}$ observations are zeros.

when recovery took place.

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 1765–72 as the reference period. The event-study specification is as follows:

$$\ln \#Book_{c,t} = \sum_{k \neq 0} \beta_k \, Censor_c \times \mathbf{D}_{t=0 \pm k} + \alpha_c + \lambda_t + \text{Section}_s \times t + \mathbf{X}_c \times \mathbf{I}_t + \epsilon_{c,t}, \tag{2}$$

where $\mathbf{D}_{t=0 \pm k}$ represents dummy variables for each 5-year interval, where k denotes the number of years relative to 1765–72, the reference period. The rest of the terms are the same as in equation (1).

4.2 Main Results

Baseline Estimates 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 dummies. They imply that a one standard deviation increase in the level of censorship is associated with a reduction in publication by approximately 15% 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 present Negative Binomial estimates, and in Column (5), Poisson estimates, with the dependent variable being the number of books. The dynamic pattern mirrors that of Column (3). These estimates suggest that a one standard deviation increase in censorship levels is associated with an approximate 18% decline in publication between 1773 and 1839, an effect that dissipates after the 1840s. For simplicity, we focus on Poisson estimates in most of our subsequent analysis. Additionally, the positive coefficient for the period 1911–1949 is sizable in both Negative Binomial and Poisson estimates, but the large standard error tempers any strong conclusions regarding possible overshooting effects.

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 1765–72 as the reference period. Prior to 1773, there was 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 publications in more-censored categories, and this effect 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.

Regional Patterns To better understand the revival that began in the 1840s, we examine regional variations in dynamic responses. The Opium War of 1839–1842 marked a pivotal turning point in Chinese history, often regarded as the beginning of modern China (Fairbank and Goldman, 2006). In the aftermath of a variety of conflicts, a series of unequal treaties forced China to open its coastal and Yangtze river regions to Western nations through the treaty port system. This era of external pressures was further compounded by the Taiping Rebellion (1851–1864), a devastating civil war that further weakened the state. These profound upheavals suggest that the state's ability to exert control over society significantly diminished during this period, beginning with treaty port regions and likely extending to other areas.

Motivated by this institutional background, we construct a category-period-(publishing) prefecture panel dataset, which allows us to study regional variations. In our book publication records, 51% of the books contain publisher information. The absence of publisher information often reflects that authors published books to distribute within their literary communities (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. Thus, we rely on available publisher information.

Figure 5 illustrates the locations of treaty ports alongside the spatial distribution of publishers. As depicted, the majority of treaty ports were situated along the coast or the Yangtze River.

At the more granular category-period-prefecture level, however, over 90% of book observations are zeros, so we focus on a binary variable indicating whether books in a given category were published within a specific period and prefecture, examining its association with censorship. Due to regional variation in book presence, we report coefficients relative to the sample mean for

each group, allowing interpretation as changes from the mean, similar to our baseline Poisson regressions.

As shown in Columns (1) and (2) of Table 3, we find evidence that the likelihood of books being published in more heavily censored categories declined in both treaty ports and non-treaty ports between 1773 and 1839. The estimated declines are approximately 13% for both groups. However, the recovery in treaty ports from 1840 to 1911 is more pronounced, while non-treaty ports continued to experience a decline during this period. This pattern suggests that the weakening of state control began in treaty ports.

To account for potential spatial spillover effects, we also analyze prefectures along the coast or the Yangtze River (where treaty ports were typically located) and compare them with those in inland China. We again observe a more pronounced recovery in coastal and Yangtze River areas during 1840–1911, aligning with the effects of the country's opening and the initial loosening of state control in these regions.

It is important to clarify that these findings do not imply a single political event behind the revival of restricted categories in the 1840s. Instead, we interpret the revival as the result of multiple significant political upheavals that collectively weakened state control.

4.3 Heterogeneities and Comparisons

We explore several dimensions of heterogeneity to examine variations in responses to censorship, focusing on differences between state and private publishers, first prints versus reprints, and natural sciences compared to other fields. Our main findings are summarized below, with detailed results available in the appendix.

State Publisher vs. Non-state Publisher We separate books based on whether the books were published by the state and find that our baseline results are primarily influenced by non-state publishers.[10] As indicated in Columns (1)-(2) of Appendix Table B.2, 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.

[10]Here, We include the missing publishers in the non-state group, as state publishers are rarely absent from the publication data. Excluding these missing publishers does not affect our results.

First Prints vs. Reprints Our primary analysis includes both first prints and reprints from the *General Catalog* data, with first prints making up 71% of the total publications.[11] As shown in Column (3) and (4) of Appendix Table B.2, first prints are the main contributors to our results. On the other hand, reprints had a negligible impact. This finding indicates that censorship hampers the creation of new knowledge, which is fundamental to drive scientific advancements and fuel economic growth (Mokyr, 2016).

Natural Sciences vs. the Rest To investigate if censorship affected the natural and social sciences differently, we first compute an index to gauge the proportion of natural sciences within various historical categories. We achieve this by aligning the 50 pre-modern Chinese book categories with contemporary classifications, using information from reprinted historical books for which we know both historical and modern classifications. This matching process is detailed in Appendix B.3. Notably, the share of natural sciences in historical Chinese publications is small, at approximately 11%.

We then divide the categories into two groups based on their varying shares of natural sciences: share of natural sciences above zero; share of natural sciences above the mean; the share of natural sciences being 100%. In Appendix Table B.3, we incorporate the interaction between the level of censorship and the dummy indicating natural science as a control variable. We find that our main findings on the decline and revival apply to both categories. However, given that only a small portion of historical books pertain to natural sciences in our setting, our findings primarily reflect knowledge about facts, history, and politics, which may influence but do not necessarily directly encompass science and technology.

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 conduct a category-emperor analysis, which is not affected by missing publication years. Appendix Table B.4 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.5, we determine the

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

level of censorship by counting banned books from various sources. Table B.5 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 Chilling Effects

In this section, we analyze the content of publications and the crucial role of chilling effects (and their changes post the 1840s) for 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

Keyword and Similarity Analysis Using the natural language processing (NLP) techniques, we analyze book contents. Our first analysis focuses on keywords in book titles. We focus on two sets of keywords – 1,714 unique keywords in banned books and 34,053 in all books – and examine how the same set changed along with censorship (see Appendix C.1 for the procedure of identifying keywords following Gentzkow and Shapiro (2010)). As previously noted, banned content represents a minor portion of the overall book contents. Specially, banned keywords account for only 2.3% of the total keyword frequency and 2.5% of all unique keywords across publications.

We begin by examining keyword counts in Columns (1)-(3) of Table 4, where each keyword may appear multiple times. Column (1) captures the overall decline and resurgence of all keywords. Column (2) isolates keywords from banned books, revealing a decline from the 1770s to the 1830s, followed by a recovery after the 1840s, suggesting that topics associated with banned materials followed similar cycles. Column (3) presents a comparable pattern among other keywords, indicating that censorship effects extended beyond explicitly banned topics.

Next, we examine the count of keywords in Columns (4)-(6), removing the influence of keyword frequency. These results confirm that the observed decline and recovery apply. However, based on this measure, those associated with banned topics remained suppressed from the 1840s through 1911. The overall trend is largely driven by unbanned keywords, as banned ones represent only a small portion of total publications.

Since yearly differences between unique and total keyword counts are small, we perform a category-by-period analysis in Appendix C.2. On average, unique keywords constitute less than one-third of total keywords at this aggregate level, and similar patterns emerge, showing a decline and subsequent revival in unique keywords.

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 6 plots the two word sets to illustrate the ideas. As shown, the most frequent keywords in banned books often concerns history whereas the *Complete Library* full-text books highlight classics. Figure 7 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 (-0.038) as the benchmark to define sensitive books, which corresponds to approximately the 85th percentile in the distribution of all books.

In Columns (7)-(8) of Table 4, we observe that both sensitive and less sensitive books reacted to censorship: both categories declined between the 1770s and the 1830s and rebounded after the 1840s. Importantly, given that less sensitive books constitute around 85% of the total, the decline and subsequent resurgence can largely be attributed to the spillover effects on this group.

New Keywords vs. Pre-existing Keywords We also examine the timing of keywords to assess how censorship influenced the emergence of new ideas or topics. For each keyword, we identify the initial year of its appearance in the dataset. In this first year, the keyword is classified as "new," while in all subsequent years, it is categorized as "pre-existing."

As shown in Table 5, both new and pre-existing keywords exhibit patterns of decline and recovery, observable at the yearly level (Columns (1)-(2)) and in aggregated periods (Columns (3)-(4)). The effect on new keywords highlights a spillover effect of censorship, constraining the generation of new topics. However, much like other measures, this restrictive effect on new keywords also dissipates after the 1840s.

Together, these observations underscore the significance of spillover and potential self-censorship: not only did topics explicitly banned or similar to those on the banned list become less prevalent between the 1770s and 1830s, but numerous unbanned and less-sensitive books in related categories also declined as censorship intensified. Additionally, censorship hindered the creation of new topics. Nevertheless, these suppression effects shifted dramatically around the 1840s, suggesting societal responses to major political changes.

5.2 Chilling Effects vs. Knowledge Loss

Banned books can impact knowledge production through two primary mechanisms: the chilling effect, wherein publishers and authors preemptively avoid certain subjects, and knowledge loss, whereby knowledge producers lose access to specific information and, consequently, generate less output. Our research, highlighting the importance of spillover effects and the resurgence of publications in more censored categories post-1840s, indicates that the chilling effect of censorship outweighs the knowledge loss mechanism. As noted earlier, banned topics (keywords) constituted only a small fraction of overall topics.

To further substantiate this interpretation, we use data on both lost and surviving banned books to construct two measures of censorship: $\frac{N_{banned,lost}}{N_{banned,lost}+N_{Siku}}$ and $\frac{N_{banned,surviving}}{N_{banned,surviving}+N_{Siku}}$. If knowledge loss were a significant factor, we would expect the censorship measure based on lost banned books to correlate with a permanent absence of specific keywords. Empirically, however, we observe similar dynamic responses across keywords for both measures, supporting the predominance of the chilling effect.

Of the 3,102 banned books, 1,783 have survived, and 1,319 were lost. We calculate two measures of censorship within each category based on the surviving and lost banned books and analyze how they influenced book contents (i.e., keyword appearances) over time. Table 6 presents the results for censorship measured by the percentage of surviving and lost banned books in each category. Columns (1), (3), and (5) show results for censorship measured by the percentage of surviving banned books, while Columns (2), (4), and (6) display results for lost banned books. In both cases, we find a decline and subsequent recovery pattern consistent with our baseline findings.

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 7, the number of publishers venturing into more-censored fields significantly declined after 1773, and then experienced a rebound after the 1840s. Based on the estimates, a one standard deviation increase in the level of censorship is associated with a 22% 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 8, 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 lifetimes of 5,805 authors whose biographical details are accessible through the China Biographical Database Project (CBDB, 2019). Additional details on these authors are provided in Appendix C.3. 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 9 shows the cross-category correlations between censorship and book publications. As illustrated in Column (1) of Table 9, 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 publishers who were active post the 1840s, 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 Table 9, there was no clear association between books published after the 1840s 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 the 1840s, 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 examine 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 constructing extensive publication data and leveraging a well-established book classification system in China, we are 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.[12] 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 negate the detrimental effects of censorship. On the contrary, the suppression and self-censorship that occurred in the short and medium terms following the initial ban likely contributed to a prolonged period of intellectual stagnation in China between the 1770s and 1830s – a critical time when Europe was undergoing the transformative changes of the Industrial Revolution.

Furthermore, we document responses from both publishers and authors and find 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.

An important limitation of our study is the inability to fully disentangle supply-side from demand-side factors in explaining the dynamic patterns we observe. While our empirical analyses focus on the knowledge producers, whose behaviors we can partially observe, the results also suggest a revival in demand after the 1840s. This renewed demand likely contributed to the

[12] 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.

consistent eagerness of publishers to release books in restricted areas, despite several decades of suppression. Thus, our findings imply a resurgence in demand for knowledge within these more restricted categories. Ultimately, the dynamic pattern we observe reflects an equilibrium between supply and demand.

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] **Biasi, Barbara, and Petra Moser.** 2021. "Effects of Copyrights on Science." *American Economic Journal: Microeconomics*, 13(4): 218-260.

[7] **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.

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

[9] **Chaney, Eric.** 2016. "Religion and the Rise and Fall of Islamic Science." Working paper.

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

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

[12] **Chen, Yuyu, and David Y Yang.** 2019. "The Impact of Media Censorship: 1984 or Brave New World?" *American Economic Review*, 109(6): 2294–2332.

[13] **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.

[14] **Dewitte, Edgard, Francesco Drago, Roberto Galbiati, and Giulio Zanella.** 2024. "Science under inquisition: The allocation of talent in early modern Europe." *CEPR Discussion Paper No. 17644.*.

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

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

[17] **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.

[18] **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.

[19] **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.

[20] **Duan, Li, and Xiaoming Zhang.** 2024. "Awakening Latent Human Capital: The Opening-Up and Entrepreneurship in 19th-Century China." *HKU Working Paper.*

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

[22] **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.

[23] **Fairbank, John King, and Merle Goldman.** 2006. *China: A New History, Second Enlarged Edition.* Harvard University Press.

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

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

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

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

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

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

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

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

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

[33] **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.

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

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

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

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

[38] **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.

[39] **Stern, Rachel E., and Jonathan Hassid** 2012. "Amplifying silence: uncertainty and control parables in contemporary China." *Comparative Political Studies*, 10 (2012): 1230-1254.

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

[41] **Tan, Hui Ren, and Tianyi Wang.** 2024. "McCarthyism, Media, and Political Repression: Evidence from Hollywood." No. w32682. National Bureau of Economic Research.

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

[43] **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*.

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

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

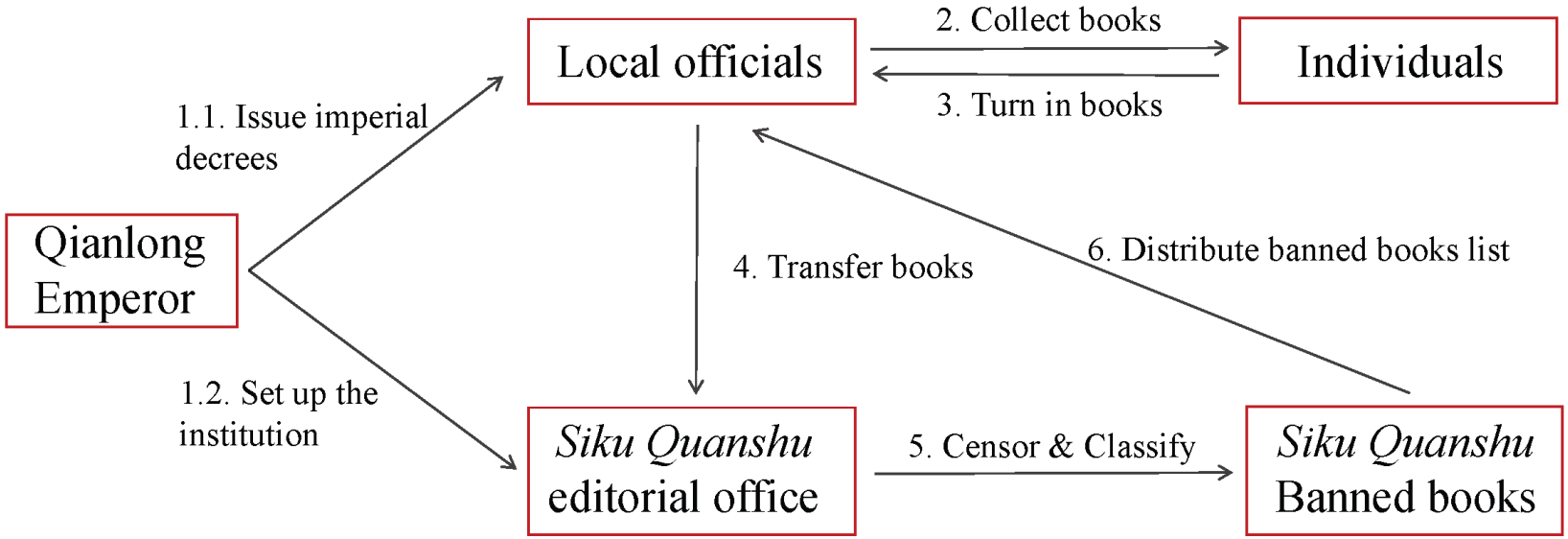
[46] **Zhai, Mianliang.** 2009. *A Dictionary of Historical Chinese Book Printing (Zhongguo guji banke cidian).* Suzhou University Press.

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

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

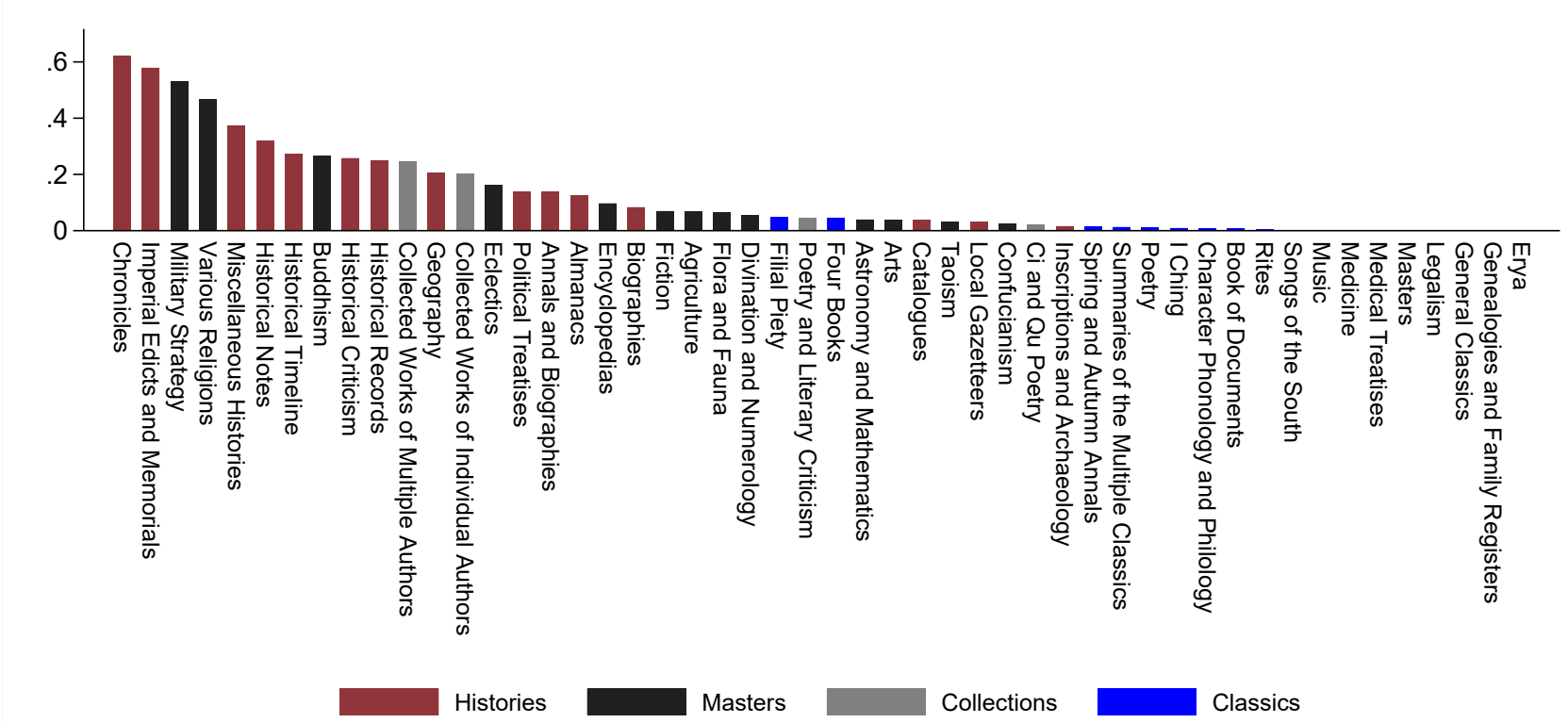
[49] **Zhou, Shaochuan.** 1996. *Studies on Catalogs of Historical Books (Guji mulu xue).* Zhongzhou Historical Books Publishing House.

Figure 1: The Process of Book Collection and Compilation



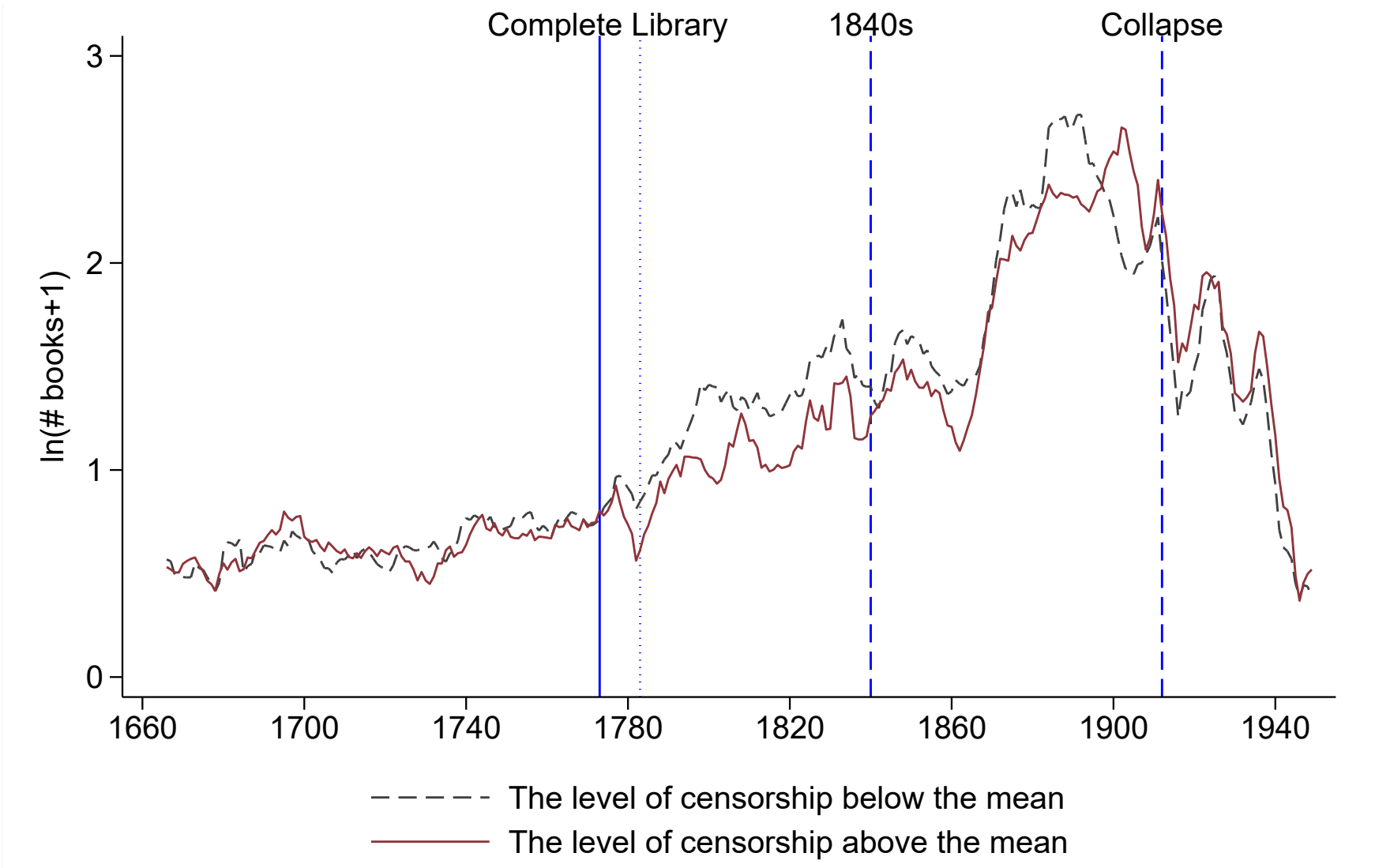
Note. This figure plots the main steps in book collection and censorship. The collection was delegated to local bureaucrats whereas the categorization and censorship decision were centralized.

Figure 2: The Level of Censorship: Based on Content



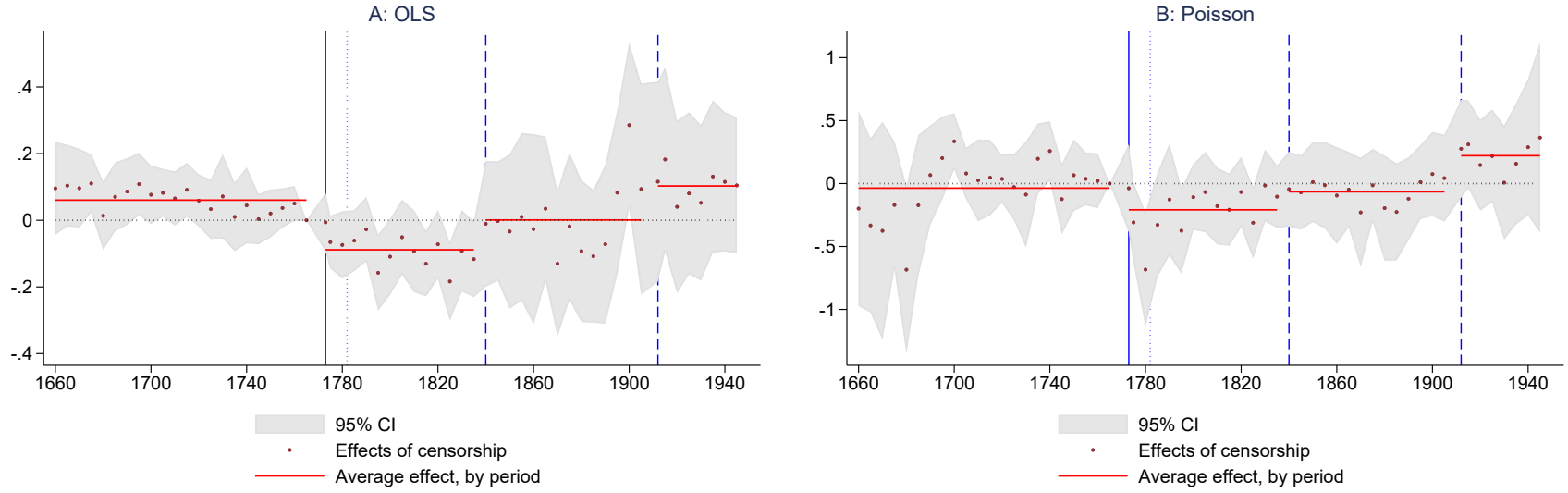
Note. This figure plots the level of censorship across 50 categories. For each category, the level is measured by the share of banned books in total collected book. The most censored categories are chronicle history, imperial decrees and memorials, military strategy, and various religions.

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



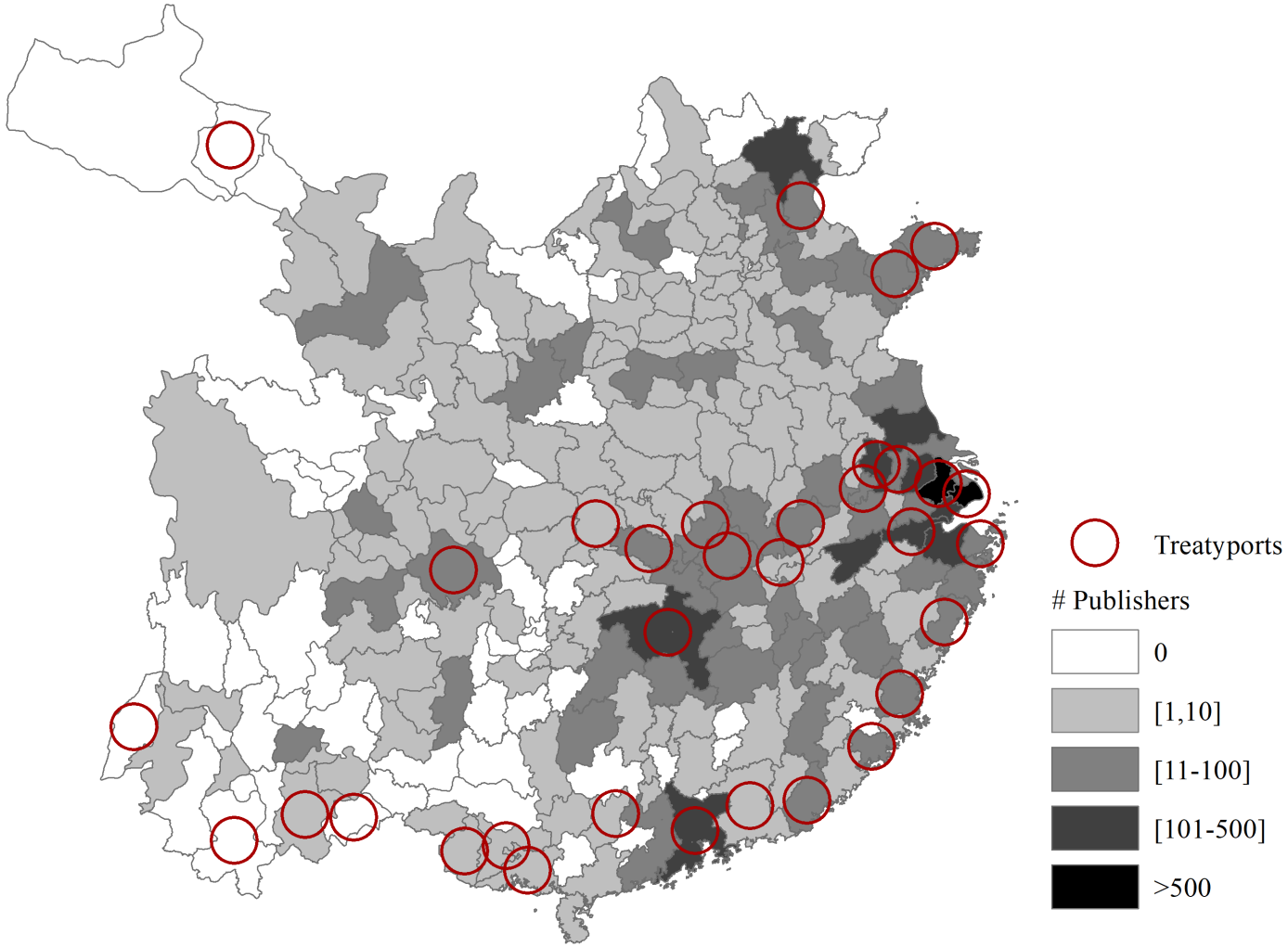
Note. This figure plots the yearly trends in book publication for two groups based on censorship degree. The red line represents trends for books in categories whose level of censorship is above the mean, and the black line for those in categories with censorship level below the mean.

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 1765–72 as the reference period. 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 in Chinese history), and 1912 (a new regime), respectively.

Figure 5: Treaty Ports and Spatial Distribution of Publishers



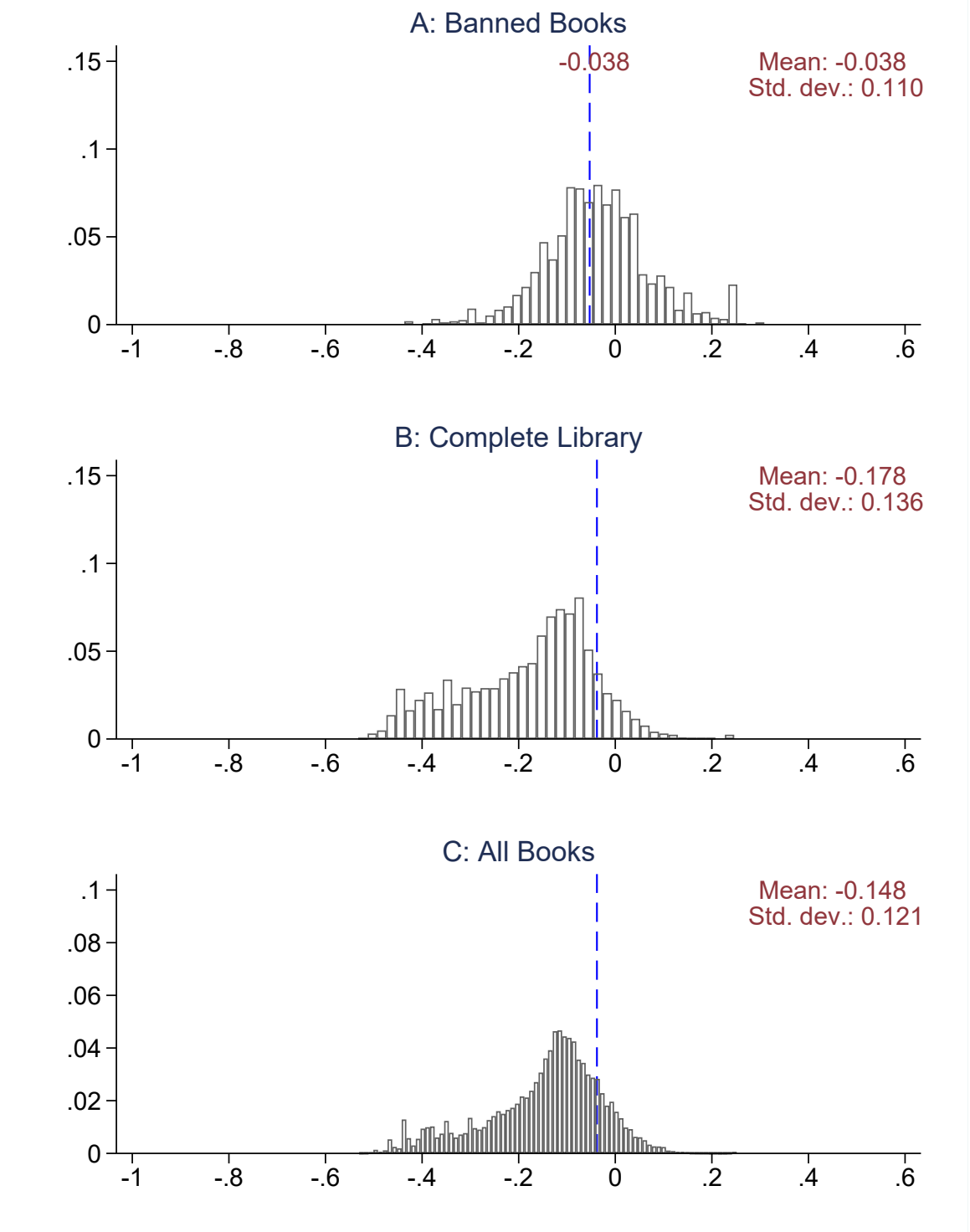
Note. This map plots the number of publishers across prefectures. The red circles mark the locations of treaty ports.

Figure 6: Keywords in Banned Books and *Complete Library* Full-text Books



Note. The left panel displays the keywords found in banned books, and the right panel displays the keywords 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 7: 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
Pre-1662 category size	A	14,400	0.020	0.032	0.0003	0.213
Pre-1662 state penetration	A,G,H,J,K	14,400	0.164	0.192	0	0.859
Pre-1662 market concentration	A,G,H,J,K	14,400	0.067	0.068	0.004	0.333
Pre-1662 reprinted level	A	14,400	1.551	0.310	1.003	2.802
Pre-1662 missing rate	A	14,400	0.654	0.117	0.148	0.868
Treaty port	A,M	52,400	0.130	0.336	0	1
Coastal or along the Yangtze River	A,M	52,400	0.233	0.423	0	1
Share of natural sciences	A,L	14,400	0.106	0.273	0	1
Textual Analysis						
Total keywords	A	14,400	15.525	48.023	0	1014
Unique keywords	A	14,400	10.691	25.659	0	458
Total keywords from banned books	A,B,C,D,E,F	14,400	0.369	1.332	0	49
Unique keywords from banned books	A,B,C,D,E,F	14,400	0.277	0.777	0	13
Total keywords from unbanned books	A,B,C,D,E,F	14,400	15.156	47.661	0	1013
Unique keywords from unbanned books	A,B,C,D,E,F	14,400	10.414	25.336	0	457
Sensitive books	A,B,C,D,E,F	14,400	0.965	3.680	0	162
Less-sensitive books	A,B,C,D,E,F	14,400	5.310	14.224	0	265
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 Catalog 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)			# Books	
	OLS			Negative Binomial	Poisson
	(1)	(2)	(3)	(4)	(5)
Censorship (sd) × 1773-1839	-0.120***	-0.129***	-0.150***	-0.181**	-0.182***
	(0.041)	(0.037)	(0.041)	(0.074)	(0.068)
Censorship (sd) × 1840-1911	-0.012	-0.028	-0.053	0.008	-0.078
	(0.078)	(0.084)	(0.113)	(0.126)	(0.113)
Censorship (sd) × 1912-1949	0.110	0.088	0.046	0.112	0.174
	(0.095)	(0.106)	(0.111)	(0.163)	(0.151)
Dependent variable mean	1.178	1.178	1.178	7.719	7.719
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	0.734	0.239	0.750
Observations	14,400	14,400	14,400	14,400	14,400

Note. The dataset comprises 50 categories × 288 years. Section denotes the four overarching sections that encompass these 50 categories. Controls include pre-1662 category size and its square term, HHI index, reprinting share, state publisher share, degree of author censorship, and the probability of missing years. 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 the Presence of Certain Books: Regional Patterns

Dependent variable	Whether there is a book published (coefficients relative to sample mean)			
	Treaty Ports	Non Treaty Ports	Coastal or along the Yangtze River	Inland
	(1)	(2)	(3)	(4)
Censorship (sd) × 1773-1839	-0.132**	-0.131***	-0.112**	-0.156***
	(0.057)	(0.046)	(0.049)	(0.056)
Censorship (sd) × 1840-1911	-0.064	-0.253*	-0.074	-0.308*
	(0.130)	(0.129)	(0.105)	(0.154)
Censorship (sd) × 1912-1949	-0.018	-0.004	-0.013	-0.006
	(0.123)	(0.082)	(0.104)	(0.093)
Dependent variable mean	0.163	0.036	0.127	0.030
Category FE	Y	Y	Y	Y
Period FE	Y	Y	Y	Y
Prefecture FE	Y	Y	Y	Y
Section × time trend	Y	Y	Y	Y
Controls × period FE	Y	Y	Y	Y
Prefecture FE × period FE	Y	Y	Y	Y
Prefecture × Category FE	Y	Y	Y	Y
R-squared	0.681	0.561	0.665	0.540
Observations	6,800	45,600	12,200	40,200

Note. The dataset comprises 50 categories × 262 prefectures × 4 periods. The estimates are all obtained from the OLS model, and the coefficients are relative to the sample mean. Controls include pre-1662 category size and its square term, HHI index, reprinting share, state publisher share, degree of author censorship, and the probability of missing years. Standard errors displayed in parentheses are clustered at the book category and prefecture level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

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

Dependent variable	# Total Keywords			# Unique Keywords			# Books	
	All	Banned	Unbanned	All	Banned	Unbanned	Sensitive	Less-sensitive
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Censorship (sd) × 1773-1839	-0.234***	-0.252***	-0.228***	-0.198***	-0.231***	-0.190***	-0.232***	-0.202***
	(0.073)	(0.077)	(0.077)	(0.069)	(0.076)	(0.073)	(0.083)	(0.073)
Censorship (sd) × 1840-1911	-0.085	-0.121	-0.075	-0.017	-0.169*	-0.000	-0.137	-0.046
	(0.102)	(0.091)	(0.105)	(0.094)	(0.102)	(0.093)	(0.098)	(0.124)
Censorship (sd) × 1912-1949	0.174	-0.047	0.195	0.210	-0.036	0.235	0.183	0.165
	(0.160)	(0.138)	(0.161)	(0.159)	(0.127)	(0.160)	(0.159)	(0.136)
Dependent variable mean	15.525	0.369	15.156	10.691	0.277	10.414	0.965	5.310
Category FE	Y	Y	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y	Y	Y
Section × time trend	Y	Y	Y	Y	Y	Y	Y	Y
Controls × period FE	Y	Y	Y	Y	Y	Y	Y	Y
Pseudo R-squared	0.771	0.445	0.774	0.732	0.390	0.734	0.665	0.692
Observations	14,400	13,965	14,400	14,400	13,965	14,400	13,248	14,400

Note. The dataset comprises 50 categories × 288 years. The estimates are all obtained from the Poisson model. In Columns (1) through (3), the dependent variable is the count of keywords appearing in book titles published across various time periods. In Columns (4) through (6), the dependent variable is the count of unique keywords appearing in book titles published across various time periods. In Columns (7) and (8), the dependent variable represents the number of books. We categorize all books into sensitive and less sensitive books based on the mean similarity derived from banned books. Controls include pre-1662 category size and its square term, HHI index, reprinting share, state publisher share, degree of author censorship, and the probability of missing years. Singletons are excluded from the number of observations. 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 Book Contents: New and Pre-existing Keywords

Dependent variable	# Total keywords			
	New	Pre-existing	New	Pre-existing
	(1)	(2)	(3)	(4)
Censorship (sd) × 1773-1839	-0.263***	-0.224***	-0.231***	-0.179**
	(0.081)	(0.076)	(0.080)	(0.079)
Censorship (sd) × 1840-1911	0.064	-0.090	0.092	-0.023
	(0.124)	(0.100)	(0.124)	(0.103)
Censorship (sd) × 1912-1949	0.262	0.175	0.275	0.236
	(0.174)	(0.164)	(0.176)	(0.169)
Dependent variable mean	1.743	13.782	125.505	992.270
Category FE	Y	Y	Y	Y
Year FE	Y	Y		
Period FE			Y	Y
Section × time trend	Y	Y	Y	Y
Controls × period FE	Y	Y	Y	Y
Pseudo R-squared	0.519	0.776	0.906	0.960
Observations	14,400	14,400	200	200

Note. The dataset of column (1) and (2) comprises 50 categories × 288 years, and the dataset of column (3) and (4) comprises 50 categories × 4 periods. The estimates are all obtained from the Poisson model. For each word, we determine the initial year of its appearance in the dataset. During this first year, the word is classified as "new." In all following years, it is categorized as a "pre-exiting" word. In Columns (1) and (3), the dependent variable is the count of new keywords appearing in book titles published across various time periods. In Columns (2) and (4), the dependent variable represents the number of pre-existing keywords. Controls include pre-1662 category size and its square term, HHI index, reprinting share, state publisher share, degree of author censorship, and the probability of missing years. Singletons are excluded from the number of observations. Standard errors displayed in parentheses are clustered at the book category level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 6: Chilling Effect vs. Knowledge Loss: Evidence from Survived and Lost Banned Books

Dependent variable	# All keywords		# Banned keywords		# Unbanned keywords	
Censorship measured by	Surviving	Lost	Surviving	Lost	Surviving	Lost
	(1)	(2)	(3)	(4)	(5)	(6)
Censorship (sd) × 1773-1839	-0.217***	-0.280***	-0.269***	-0.175**	-0.208***	-0.286***
	(0.071)	(0.097)	(0.082)	(0.072)	(0.074)	(0.105)
Censorship (sd) × 1840-1911	-0.091	-0.082	-0.097	-0.114	-0.082	-0.069
	(0.104)	(0.109)	(0.104)	(0.071)	(0.107)	(0.115)
Censorship (sd) × 1912-1949	0.249	0.071	-0.001	-0.080	0.267	0.097
	(0.163)	(0.151)	(0.166)	(0.093)	(0.165)	(0.157)
Dependent variable mean	15.525	15.525	0.369	0.369	15.156	15.156
Category FE	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y
Section × time trend	Y	Y	Y	Y	Y	Y
Controls × period FE	Y	Y	Y	Y	Y	Y
Pseudo R-squared	0.769	0.773	0.446	0.445	0.773	0.777
Observations	14,400	14,400	13,965	13,965	14,400	14,400

Note. The dataset comprises 50 categories × 288 years. The estimates are all obtained from the Poisson model. In Columns (1), (3), and (5), the level of censorship is determined by the proportion of surviving banned books relative to all collected books. In Columns (2), (4) and (6), the level of censorship is calculated by the share of lost banned books among all collected books. Controls include pre-1662 category size and its square term, HHI index, reprinting share, state publisher share, degree of author censorship, and the probability of missing years. Singletons are excluded from the number of observations. Standard errors displayed in parentheses are clustered at the book category level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 7: Censorship and Responses of Publishers

Dependent variable	# Publishers Poisson	ln(# Publishers+1) OLS	# Books Poisson	ln(# Books+1) OLS	ln(# Publishers+1) First Stage	ln(# Books+1) Reduce Form	ln(# Books+1) IV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
#NewPulishers$_p$× Share$_{c,p}$					0.772**	1.093***	
					(0.295)	(0.232)	
Censorship (sd) × 1773-1839	-0.222***	-0.175**	0.080	-0.051	-0.151**	-0.103***	0.111
	(0.062)	(0.075)	(0.063)	(0.049)	(0.072)	(0.038)	(0.140)
Censorship (sd) × 1840-1911	-0.104	0.026	0.046	-0.062	0.060	0.009	-0.076
	(0.086)	(0.101)	(0.087)	(0.096)	(0.082)	(0.110)	(0.106)
Censorship (sd) × 1912-1949	0.009	0.142	0.109	-0.028	0.165	0.083	-0.151
	(0.121)	(0.165)	(0.089)	(0.081)	(0.137)	(0.113)	(0.106)
ln (# Publishers+1)			0.916***	0.536***			1.416***
			(0.058)	(0.078)			(0.386)
Dependent variable. mean	149.650	4.014	7.719	1.178	4.014	1.178	1.178
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	0.899	0.769	0.764	0.925	0.740	
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, based on 50 categories × 4 periods. Columns (3)-(7) show that the number of publishers can explain our baseline finding on the number of books, where the dataset comprises 50 categories × 288 years. Columns (5)-(7) employ a Bartik approach instrumenting the number of publishers in each category with aggregate changes and initial distribution. Controls include pre-1662 category size and its square term, HHI index, reprinting share, state publisher share, degree of author censorship, and the probability of missing years. Standard errors displayed in parentheses are clustered at the book category level. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

Table 8: 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	-0.174**	-0.085	-0.176**
	(0.075)	(0.085)	(0.079)
Censorship (sd) × 1840-1911	0.201*	0.129	0.211*
	(0.106)	(0.084)	(0.116)
Censorship (sd) × 1912-1949	0.117	0.050	0.140
	(0.140)	(0.066)	(0.162)
Dependent variable mean	0.064	0.074	0.027
Section FE	Y	Y	Y
Period FE	Y	Y	Y
Controls × period FE	Y	Y	Y
R-squared	0.801	0.532	0.816
Observations	150	150	150

Note. The estimates are obtained from the OLS model based on changes in logged number of publishers across 50 categories × 4 periods. Surviving publishers are those who spanned at least two periods. Controls include pre-1662 category size and its square term, HHI index, reprinting share, state publisher share, degree of author censorship, and the probability of missing years. Standard errors displayed in parentheses are clustered at the book category level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 9: Separating Responses from Publishers and Authors

				# Books, Poisson	
				Authors	
				Passed away before 1772	Lived during 1773-1839
				Column (1)	Column (2)
	Active before 1772	Row (1)	censorship(sd)	-0.026	
				(0.076)	
Publishers	Active during 1773-1839	Row (2)	censorship(sd)	-0.187**	-0.405***
				(0.086)	(0.088)
	Active post 1840	Row (3)	censorship(sd)	-0.047	-0.521***
				(0.108)	(0.114)
	Observations			50	50

Note. This table shows the correlation across 50 categories between censorship and publications for five groups, with the same controls in our earlier analyses: pre-1662 category size and its square term, HHI index, reprinting share, state publisher share, degree of author censorship, and the probability of missing years. The estimates are obtained 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-7

B Decline and Revival in Book Production: Additional Results **A-9**

B.1 Missing Publisher Information A-9

B.2 Heterogeneities: Publisher Ownership and Reprints A-10

B.3 Considering Natural Sciences A-11

B.4 Category-by-Emperor Analysis A-14

B.5 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 Unique Keywords: Category-by-Period Analysis A-20

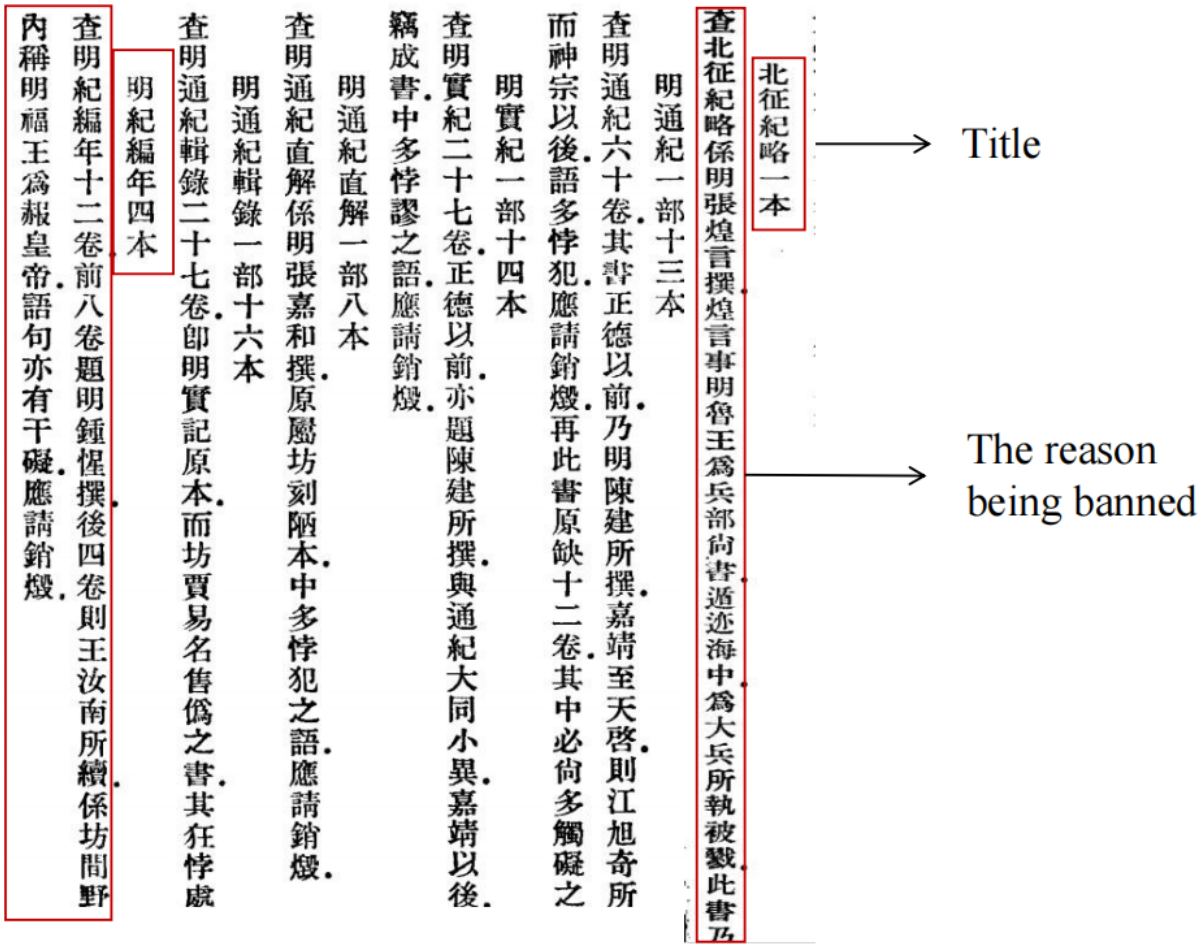
C.3 Authors' Information A-20

A Data Construction and Description

A.1 Banned Books

Figure A.1 presents instances of banned book records, outlining the reasons for their prohibition. In the two boxed examples, *Record of the Northern Expedition* was banned for detailing conflicts with Qing forces at the end of the Ming Dynasty. *Chronicles of the Ming Dynasty* was banned for presenting non-official accounts of Ming dynasty history. Generally, books are banned for three primary reasons: (i) expressing anti-Manchu sentiments or nostalgia for previous dynasties; (ii) documenting conflicts between Manchus and Han; (iii) deviating from the Confucian classics.

Figure A.1: Examples of Banned Books

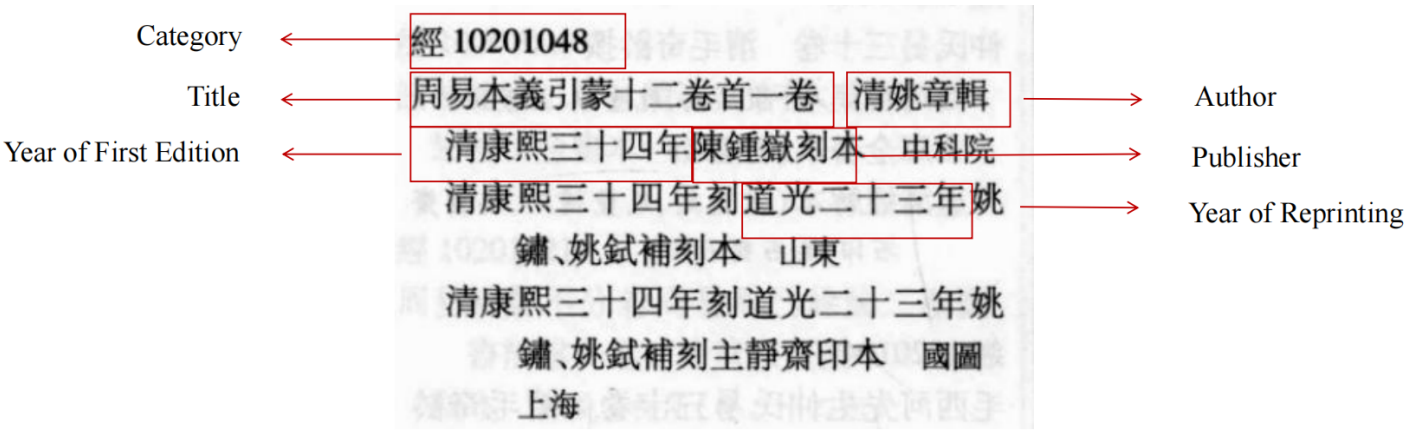


A.2 Book Publication Data

To assess knowledge production, we employ book publication records from the *General Catalog of Pre-modern Chinese Books*. From 1662 to 1949, the Catalog lists over 161,000 book editions. An

example Catalog entry is shown in Figure A.2. Each entry includes vital details such as publication category, author, publication date, reprint status, and reprint year. Moreover, publisher information is available for 51% of the books. Using the publication category and year, we construct variables to represent annual book publications per category. Additionally, these variables are refined to differentiate between first prints and reprints across each category and year in the dataset.

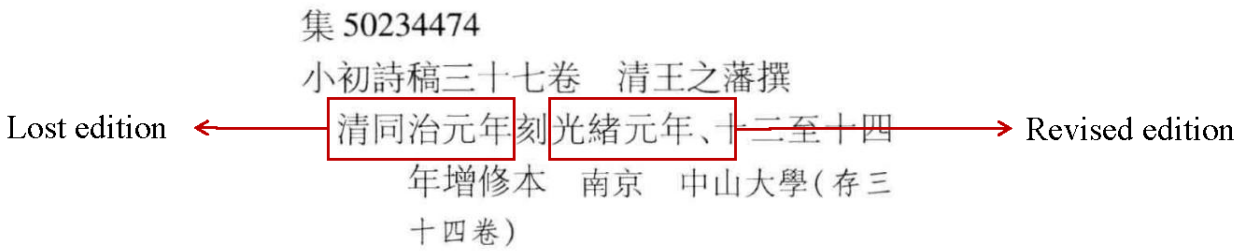
Figure A.2: Records on Book Publications



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 notes 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, whether we consider books published during the suppression decades (Columns (1)-(3)) or all books in our data (Columns (4)-(6)).

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	-0.009
	(0.008)	(0.008)	(0.013)	(0.006)	(0.007)	(0.012)
Category size		0.030	0.067		0.068	0.149
		(0.154)	(0.190)		(0.139)	(0.177)
Category size squared		0.008	-0.068		-0.116	-0.429
		(0.708)	(0.795)		(0.644)	(0.732)
State penetration		-0.003	-0.004		-0.003	-0.004
		(0.008)	(0.008)		(0.006)	(0.006)
Market Concentration		-0.003	-0.009		0.029	0.026
		(0.015)	(0.016)		(0.023)	(0.024)
Reprint level		-0.010	-0.008		-0.005	-0.003
		(0.008)	(0.006)		(0.008)	(0.006)
Author-based censorship		-0.001	-0.004		-0.007	-0.013
		(0.025)	(0.027)		(0.019)	(0.021)
Section FE			Y			Y
R-squared	0.004	0.107	0.166	0.003	0.085	0.169
Observations	50	50	50	50	50	50

Note. The dataset comprises characteristics across 50 categories. 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)	0.064	0.069	0.004
	(0.094)	(0.091)	(0.117)
Category size		2.561	2.814
		(1.628)	(1.938)
Category size squared		-8.947	-10.670
		(7.314)	(7.950)
State penetration		0.217**	0.226**
		(0.095)	(0.096)
Market Concentration		0.559*	0.485
		(0.328)	(0.339)
Reprint level		-0.025	-0.008
		(0.058)	(0.044)
Author-based censorship		-0.321	-0.369
		(0.210)	(0.240)
Section FE			Y
R-squared	0.008	0.246	0.316
Observations	50	50	50

Note. The dataset comprises characteristics across 50 categories. 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 Catalog* Among the 3,102 banned books, 1,783 can be matched to the *General Catalog*, which means that these books have survived to the present day and are clearly catalogued.

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 3. 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., 9% 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:

$$Censor_c = \frac{N_{banned}}{N_{banned} + N_{Siku}}. \tag{3}$$

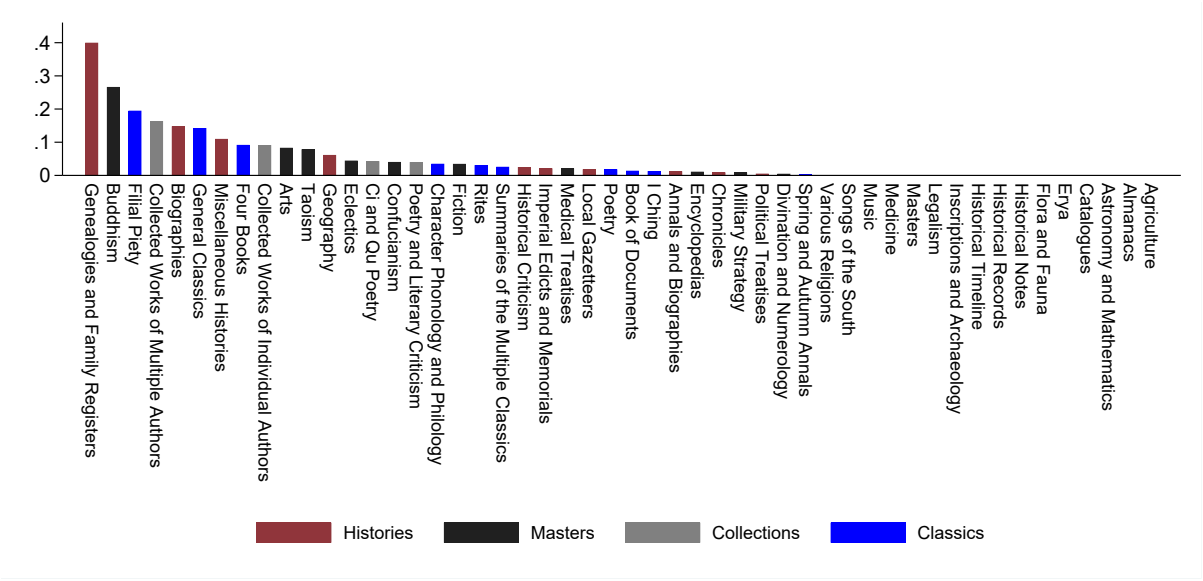
where N_{banned} denotes the number of banned books, and N_{Siku} denotes the number of books in the *Complete Library*, including both the full-text books and indexed books. We also separate lost and surviving banned books in our additional analysis.

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 methodology 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 Catalog* provides the basic information on the names of 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 gather additional historical data on the printing market and publishers, incorporating them into the *General Catalog*. These historical data sources (Du, 2001; Du, 2009; Zhai, 2009) provide information on publishers during the Ming and Qing dynasties, detailing the names of publishers, their ownership, and the books they produced.

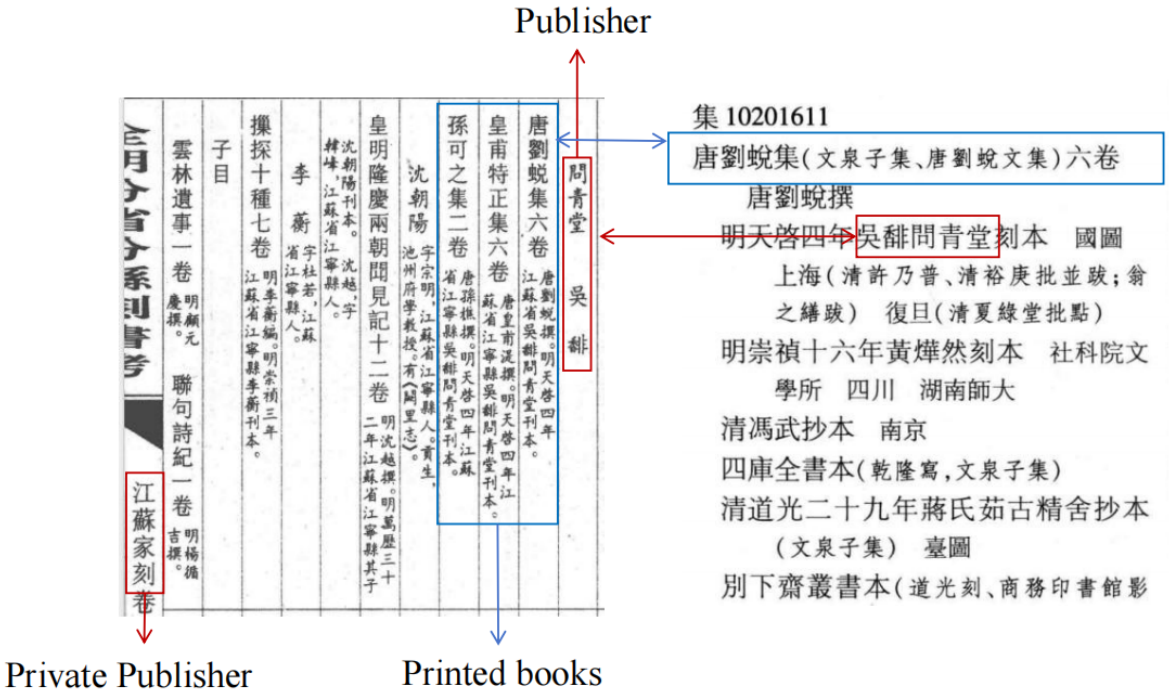
Figure A.7(I) provides an example of how to match the historical data on publishers to the *General Catalog* data. Based on these sources, we construct a dataset of publishers, which includes the unique ID, ownership, the origin of the publisher, and the address of the publishers.

Figure 5 illustrates the spatial distribution of publishers. The dispersed locations reflect a decentralized printing market in this era. Figure A.7(II) illustrates the relationship between the logged number of books published by publishers and their ranks, revealing a power-law distribution.

A.8 Censorship and Other Category Characteristics

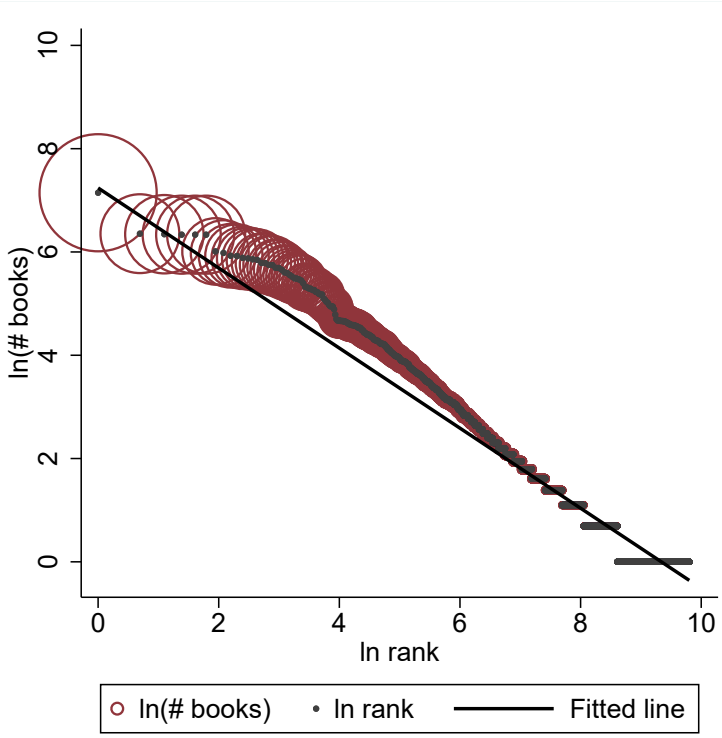
Table A.8 shows that there are no evident relationships between the degree of censorship and other category-level traits, indicating that factors such as size or market considerations do not significantly influence censorship.

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



Note. The left panel comes from data on publishers and the right panel comes from the *General Catalog*.

Figure A.7: II. Publications-Rank Relationship



Note. This figure plots the number of books published by publishers in relation to their ranks.

Table A.8: Censorship and Other Category Characteristics

Dependent variable	Censorship (sd)		
	(1)	(2)	(3)
Category size	-0.499	-0.356	-0.847
	(1.615)	(1.511)	(2.014)
Category size squared	3.700	3.134	6.217
	(6.544)	(6.167)	(8.463)
State penetration	-0.052	-0.050	-0.037
	(0.073)	(0.077)	(0.064)
Market concentration	-0.014	-0.019	-0.283
	(0.397)	(0.404)	(0.389)
Reprinted level	-0.044	-0.050	-0.002
	(0.071)	(0.076)	(0.062)
Missing rate	0.082	0.080	0.146
	(0.147)	(0.150)	(0.184)
Author-based censorship		-0.089	-0.056
		(0.285)	(0.354)
Section FE			Y
R-squared	0.019	0.020	0.251
Observations	50	50	50

Note. The dataset comprises characteristics across 50 categories. 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 without publisher		
	(1)	(2)	(3)
Censorship (sd) × 1773-1839	0.003	0.009	0.009
	(0.022)	(0.021)	(0.024)
Censorship (sd) × 1840-1911	-0.007	0.005	-0.001
	(0.024)	(0.021)	(0.022)
Censorship (sd) × 1912-1949	-0.043	-0.025	-0.027
	(0.027)	(0.023)	(0.028)
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	0.820
Observations	200	200	200

Note. The dataset comprises characteristics across 50 categories × 4 periods. Controls include pre-1662 category size and its square term, HHI index, reprinting share, state publisher share, degree of author censorship, and the probability of missing years. Standard errors displayed in parentheses are clustered at the book category level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

B.2 Heterogeneities: Publisher Ownership and Reprints

Table B.2: The Impact of Censorship on #Books: Heterogeneous Patterns

Dependent variable	# Books, Poisson			
	Non-state publishers	State publishers	First prints	Reprints
	(1)	(2)	(3)	(4)
Censorship (sd) × 1773-1839	-0.164**	-0.552***	-0.242***	0.032
	(0.067)	(0.110)	(0.061)	(0.132)
Censorship (sd) × 1840-1911	-0.053	-0.591***	-0.048	-0.030
	(0.115)	(0.206)	(0.117)	(0.165)
Censorship (sd) × 1912-1949	0.168	-0.225	0.090	0.187
	(0.150)	(0.189)	(0.176)	(0.160)
Dependent variable mean	7.226	0.493	5.475	2.245
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	0.750	0.597	0.757	0.641
Observations	14,400	12,800	14,400	14,400

Note. The dataset comprises 50 categories × 288 years. The estimates are all obtained from the Poisson model. Controls include pre-1662 category size and its square term, HHI index, reprinting share, state publisher share, degree of author censorship, and the probability of missing years. Standard errors displayed in parentheses are clustered at the book category level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Figure B.3: I. Republication of Historical Books in Modern Times: An Example



B.3 Considering 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.

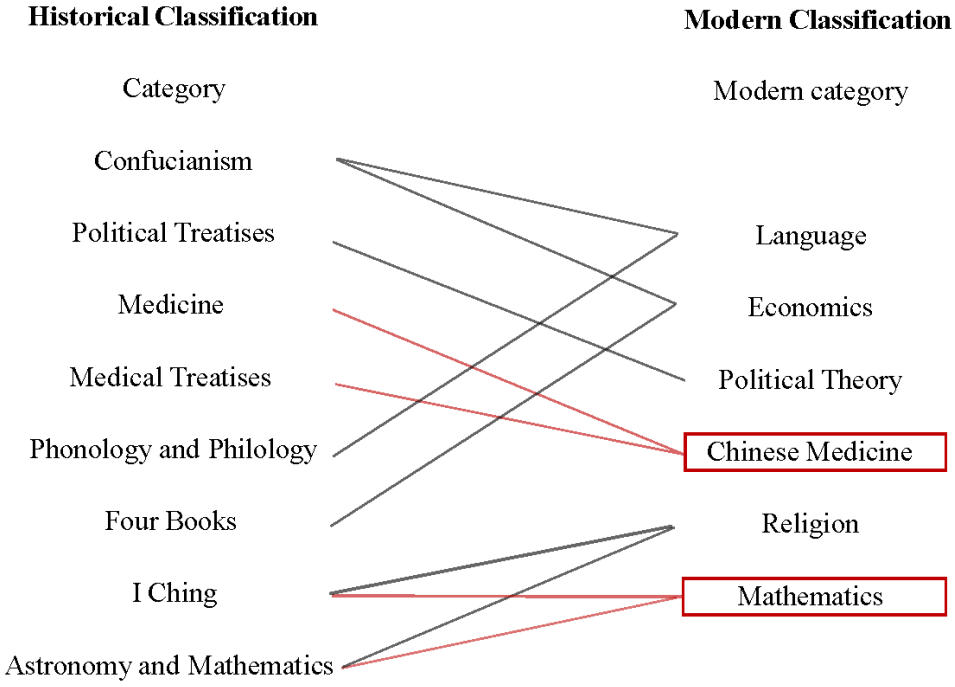
Nearly 4,000 historical books were republished after 1949, on which there exist modern classification. Thus, we know both the historical and modern classification for these republished books. Figure B.3(I) provides an example. *Exploiting the Works of Heaven* belongs to the category of political treatises by historical classification, while it belongs to natural science theory by modern classification.

These republished books provide us with a pathway to match historical classifications with modern ones. Figure B.3(II) shows an demonstration for matching.

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

Figure B.3: II. Matching Historical Classification with Modern Classification



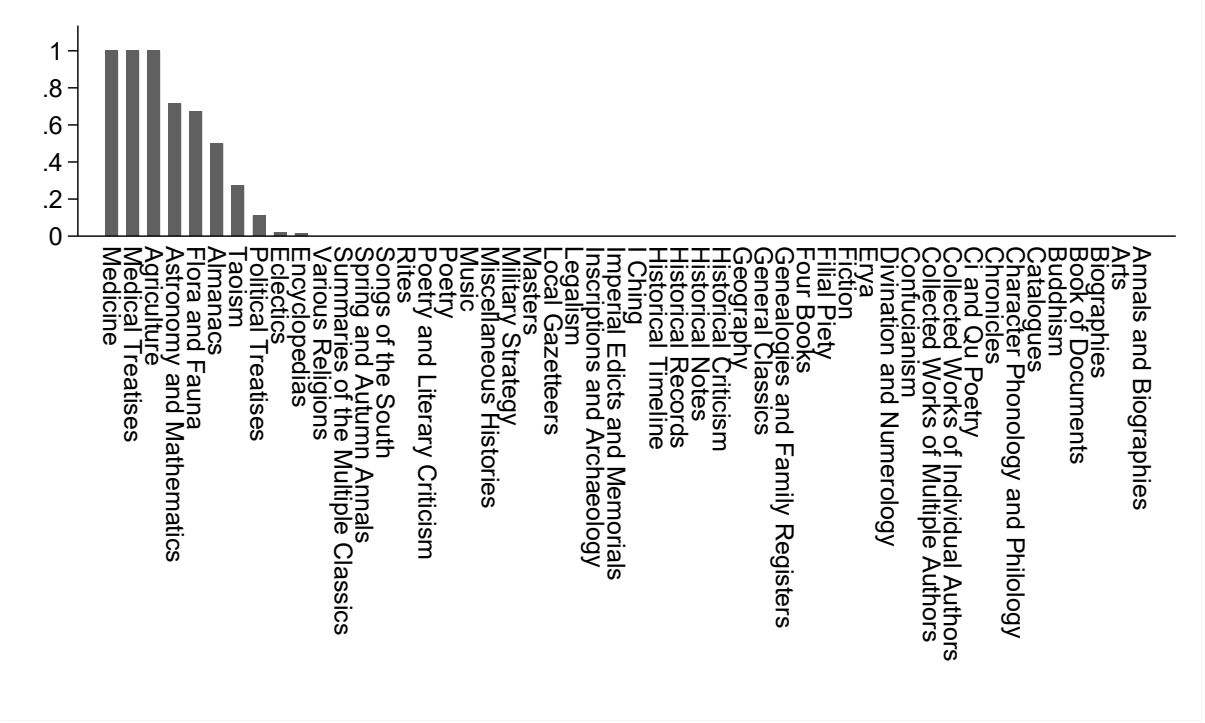
categories. Let N_{ij} denote the number of books matched to modern category j in historical category i;

- 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:

$$S_{in} = \frac{\sum N_{ij}}{N_i}, \text{ if } j \text{ belongs to natural sciences} \tag{4}$$

Figure B.3(III) illustrates the proportion of natural science within each historical category. In general, approximately 11% of the books are related to natural sciences, showing that they represented a minor portion of the total publications. The top 3 categories with the greatest share of natural sciences are: Medicine, Medical Treatises and Agriculture.

Figure B.3: III. Share of Books in the Natural Science



Note. Figure B.3 plots the share of books belongs to the natural science in each historical category among the reprinted books after 1949.

We then divide all categories into two groups based on their share of natural sciences. For robustness, we employ two definitions, where a category is classified as natural sciences if its share of natural sciences is (1) above the mean, or (2) equal to one. We interact these dummy variables of natural sciences with the level of censorship, and the results are presented in Table B.3. As shown, the decline and revival in the natural sciences are similar to that in the rest. Because natural sciences only made up a small portion of total publications, the estimates are noisier if we only consider natural sciences.

Table B.3: The Impact of Censorship on # Books: Natural Sciences

Dependent variable	# Books, Poisson			
	Share > mean		Share == 1	
	(1)	(2)	(3)	(4)
Censorship (sd) × 1773-1839	-0.201***	-0.197***	-0.208***	-0.208***
	(0.070)	(0.074)	(0.072)	(0.072)
Censorship (sd) × 1840-1911	-0.085	-0.077	-0.084	-0.085
	(0.117)	(0.122)	(0.121)	(0.122)
Censorship (sd) × 1912-1949	0.073	0.073	0.080	0.080
	(0.162)	(0.168)	(0.148)	(0.149)
Censorship (sd) × natural sciences × 1773-1839		-0.123		-0.162
		(0.264)		(0.507)
Censorship (sd) × natural sciences × 1840-1911		-0.190		0.226
		(0.306)		(0.696)
Censorship (sd) × natural sciences × 1912-1949		0.536		0.800
		(0.506)		(0.861)
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
Natural sciences × period FE	Y	Y	Y	Y
Pseudo R-squared	0.756	0.756	0.754	0.754
Observations	14,400	14,400	14,400	14,400

Note. The dataset comprises 50 categories × 288 years. The estimates are all obtained from the Poisson model. Controls include pre-1662 category size and its square term, HHI index, reprinting share, state publisher share, degree of author censorship, and the probability of missing years. Standard errors displayed in parentheses are clustered at the book category level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

B.4 Category-by-Emperor Analysis

Table B.4 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.4: Robustness Checks: Category-by-EmperorAnalysis

Dependent variable	ln(# Books+1)	# Books
	OLS	Poisson
	(1)	(2)
Censorship (sd) × 1773-1839	-0.141*	-0.168**
	(0.080)	(0.074)
Censorship (sd) × 1840-1911	0.016	-0.079
	(0.110)	(0.120)
Censorship (sd) × 1912-1949	0.186	0.166
	(0.149)	(0.142)
Category FE	Y	Y
Emperor FE	Y	Y
Section × time trend	Y	Y
Controls × period FE	Y	Y
R-squared	0.736	0.733
Observations	600	600

Note. The dataset comprises 50 categories × 12 emperors. Controls include pre-1662 category size and its square term, HHI index, reprinting share, state publisher share, degree of author censorship, and the probability of missing years. Standard errors displayed in parentheses are clustered at the book category level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

B.5 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.5 shows the results using the level of censorship calculated based on the number of banned books from different sources.

Table B.5: Measurement of Censorship: Different Sources of Banned Books

Dependent variable	ln(# Books+1), OLS		# Books, Poisson	
	Official records	Historians'	Official records	Historians'
	(1)	(2)	(3)	(4)
Censorship (sd) × 1773-1839	-0.134***	-0.156***	-0.171**	-0.161**
	(0.038)	(0.043)	(0.069)	(0.081)
Censorship (sd) × 1840-1911	-0.036	-0.124	-0.089	-0.122
	(0.118)	(0.082)	(0.128)	(0.086)
Censorship (sd) × 1912-1949	0.040	-0.023	0.135	0.154
	(0.112)	(0.084)	(0.150)	(0.150)
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
Pesudo R-squared	0.733	0.736	0.749	0.753
Observations	14,400	14,400	14,400	14,400

Note. The dataset comprises 50 categories × 288 years. Controls include pre-1662 category size and its square term, HHI index, reprinting share, state publisher share, degree of author censorship, and the probability of missing years. 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

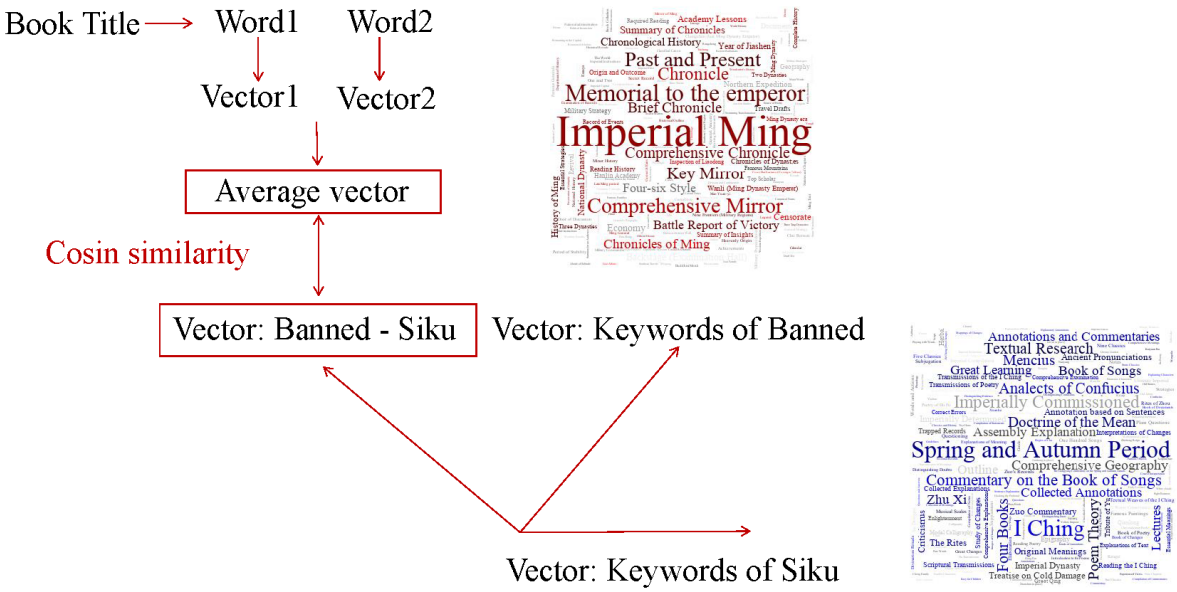
C.1 Methodology of Textual Analysis

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.

Building Corpora A corpus is a collection of texts or words 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 include (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

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



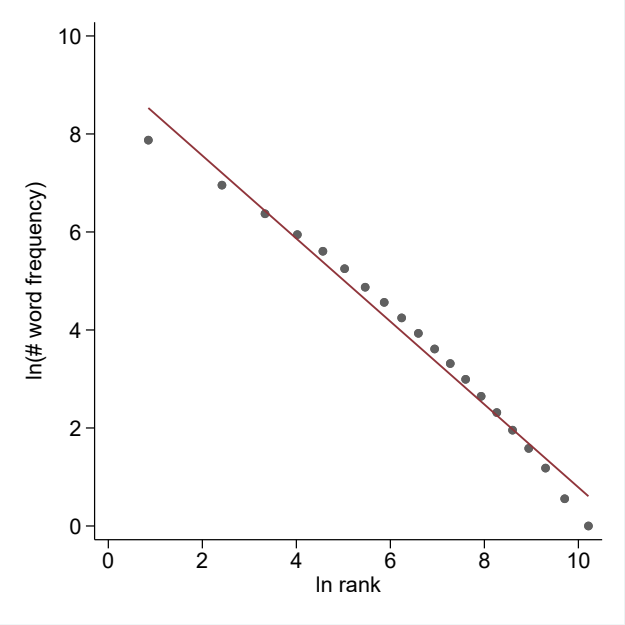
minimal unique information, such as collected works and collections of poems.

Ultimately, we compile a set of words, which contains 29,541 unique words with a total word frequency of 223,555. Figure C.1(II) displays the relationship between word frequency and its rank, demonstrating a strong 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.

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

Figure C.1: II. Word Frequency and Rank



books:[13]

$$\chi_w^2 = \frac{(f_b f_{\sim s} - f_s f_{\sim b})^2}{(f_b + f_s)\,(f_b + f_{\sim b})\,(f_s + f_{\sim s})\,(f_{\sim b} + f_{\sim s})}. \tag{5}$$

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

In our dataset, there are 2,864 unique (two-character) keywords found in the titles of books within the *Complete Library*, while the set of banned book titles includes 1,714 unique keywords. Of these, 353 keywords are shared between the *Complete Library* and the banned titles. We calculate the Pearson statistic values and choose the words that fall within the highest 10% of the values for each group. Ultimately, we identify 190 keywords for the banned books and 588 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.

[13]Simpler statistics, such as the ratio of word usage by the *Complete Library* compared to the banned books, could bias selection towards words that are used infrequently by the *Complete Library* 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
Commentary on the Book of Songs	Past and Present
Four Books	Chronicle
Textual Research	Key Mirror
Analects of Confucius	Comprehensive Chronicle
Poem Theory	Brief Chronicle
Doctrine of the Mean	Chronicles of Ming
Mencius	Four-six Style
Comprehensive Geography	Battle Report of Victory
Rites of Zhou	Summary of Chronicles
Imperial Edicts	Official History
Assembly Explanation	Backstage (Examination Hall)
Great Learning	Economy
Illustrated Text	Censorate
Lectures	Academy Lessons
Annotations and Commentaries	Chronological History
Outline	History of Ming

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 create representations in a relatively low-dimensional Euclidean space (Mikolov et al., 2013). The basic intuition 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). 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 are located in proximity to one another.

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

$$\overrightarrow{Keywords}_{\text{Banned books}} = [0.035, 0.079, 0.231, -0.013, \ldots] \tag{6}$$

$$\overrightarrow{Keywords}_{\text{Complete library}} = [0.034, 0.071, 0.313, -0.066, \ldots] \tag{7}$$

Measuring Similarity The 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 follows:

$$\text{sim}_i = \frac{\overrightarrow{Booktitle}_{\text{i}} \cdot \overrightarrow{Benchmark}}{\left\|\overrightarrow{Booktitle}_{\text{i}}\right\| \times \left\|\overrightarrow{Benchmark}\right\|} \tag{8}$$

C.2 Unique Keywords: Category-by-Period Analysis

Table C.2: The Impact of Censorship on Book Contents: Category-by-Period Analysis

Dependent variable	# Total keywords			# Unique keywords		
	All	Banned	Unbanned	All	Banned	Unbanned
	(1)	(2)	(3)	(4)	(5)	(6)
Censorship (sd) × 1773-1839	-0.192**	-0.231***	-0.186**	-0.136**	-0.211***	-0.129**
	(0.076)	(0.078)	(0.080)	(0.058)	(0.061)	(0.062)
Censorship (sd) × 1840-1911	-0.025	-0.095	-0.015	0.071	-0.156*	0.087
	(0.104)	(0.094)	(0.107)	(0.092)	(0.094)	(0.090)
Censorship (sd) × 1912-1949	0.229	-0.039	0.251	0.211	-0.018	0.226
	(0.164)	(0.144)	(0.164)	(0.140)	(0.102)	(0.142)
Dependent variable mean	1117.775	26.565	1091.210	340.185	6.225	333.960
Category FE	Y	Y	Y	Y	Y	Y
period FE	Y	Y	Y	Y	Y	Y
Section × time trend	Y	Y	Y	Y	Y	Y
Controls × period FE	Y	Y	Y	Y	Y	Y
Pseudo R-squared	0.958	0.885	0.958	0.931	0.660	0.932
Observations	200	196	200	200	196	200

*Note.*The dataset comprises 50 categories × 4 periods. The estimates are all obtained from the Poisson model. In Columns (1) through (3), the dependent variable is the count of a specific set of keywords appearing in book titles published across various time periods. In Columns (4) through (6), the dependent variable is the count of unique keywords appearing in book titles published across various time periods. Controls include pre-1662 category size and its square term, HHI index, reprinting share, state publisher share, degree of author censorship, and the probability of missing years. Singletons are excluded from the number of observations. Standard errors displayed in parentheses are clustered at the book category level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

C.3 Authors' Information

Our primary source of author information is the China Biographical Database Project (CBDB, 2019). Figure C.3 shows an instance from this database, including authors' names, birth and death

years, and hometowns, along with their works. This helps us identify authors with identical names. Additionally, we cross-check author details using Wikipedia and Baidu Baike.

Figure C.3: Records on Authors' Information from China Biographical Database Project



In total, we have comprehensive information on 7,907 authors, 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.

References

[1] **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).*

[2] **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.

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