Population Census, State Legibility, and Politics of Data in China

Junchao Tang

Introduction

The state legibility of China, defined as its capacity to collect correct information from society, is a puzzle. On the one hand, the communist regime is constantly depicted as an omnipresent big brother surveilling every molecule of the society. On the other hand, ethnographies often suggest that it is common for local officials and citizens to deceive the Chinese state. For example, sources have shown that local officials hid, misrepresented, or otherwise failed to articulate information on the health of the local population which is thought to be linked to the 2020 COVID-19 outbreak in Wuhan city. How should we evaluate these paradoxical depictions of China's state legibility? This project is an attempt to resolve the paradox by imposing demographic data-checking techniques on available census data with a socio-political lens.

In fact, calling the narrative a paradox is a trap. First, such a narrative confuses the front line of data collection where information extraction happens with the backstage of data processing for bureaucratic information reporting. Unlike what the narrative suggests, state legibility is *manifold* but not *singular*. Second, the narrative is implicitly built upon the common belief that the Chinese state is effective in data collection while ineffective in ensuring undistorted data. It is this implicit assumption that makes the narrative a paradox, but a *paradox-on-paper*. While we have evidence in scholarly literature supporting the distortion in data reporting, however, empirical support for inaccurate data collection is rare if it even exists. The efficiency of the Chinese state in data collecting is thus *assumed*, not *ascertained*. In short, the Chinese state legibility is not a paradox; it is a *mess* plus a *myth*.

Theoretical Background

Unfortunately, an overview of recent scholarly discussions on state legibility will lead us to a similar trap involving both flaws. Political scientists have developed elaborate theories about how information can be distorted in an authoritarian context through top-down manipulation of data (Svolik 2012) or bottom-up distortion by local officials (Tsai 2008). Without a theorization about legibility formation through data collection, however, this picture of state legibility is incomplete and thus unsatisfactory, offsetting the intellectual merit of both traditions. After all, the bottom-up distortion makes a real difference in state legibility only if accurate data could otherwise be collected. Only when the state legibility has somewhat been established through data collection and reporting, and only when the state has a reflexive knowledge of its level of legibility, can the top-down manipulation becomes a reasonable performative tool for the state to attain symbolic power or legitimacy (Bourdieu 1994; Ding 2022). Unfortunately, much less has been formally theorized about legibility through data collection so far. except for Jerven (2013) who challenged the validity and reliability of economic statistics in Africa.

It is no surprise that the empirical examination of state legibility is unsatisfactory when the formulation of the legibility-formation process is incomplete. Ghosh (2020) and Travers (1982) discussed how the prevailing norm of typical sampling and enumeration prevented the Chinese

Communist Party from establishing an effective statistical system during Mao's era, but they did not give out a quantitative evaluation of the magnitude of inaccuracy. Though efforts have been made to detect the level of data irregularity, rarely can scholars say with confidence about the source of irregularity beyond suggestive evidence (Wallace 2016, 2022): Is the irregularity a product of inability, distortion, or manipulation? A sound evaluation of the data collection capacity is thus vital to a proper evaluation and decomposition of the multiple sources of state legibility: if we have an idea of the irregularity arising from data collection, then we have more confidence to say that the residuals in irregularity should be attributed to distortion or manipulation.

Therefore, there is a need to supplement the literature in a manner that addresses both points: a theoretical formulation of legibility formation through data collection, followed by an empirical quantitative evaluation of data collection capabilities. My general theoretical proposal will be that state legibility lives on two hands. On the left-hand side, there is unbiased data reporting through the bureaucracy. On the right-hand side, there is effective "ex-ante" data collection by the statistical bureaucrats. The left-hand legibility is a product of the despotic power of the state, while the right-hand legibility is a product of the infrastructural power of the state (Mann 1984, 2012). Through a quantitative analysis of existing population census data, I hope to provide a critical benchmark against which the existence and magnitude of state legibility from each hand can be evaluated.

Aims and Contributions

I aim to improve our intellectual understanding of state legibility and its crystallization in contemporary China. Drawing from the quality-checking techniques developed in demography (Lee and Zhang 2013), I plan to provide a quantitative evaluation of the two hands of Chinese state legibility. More specifically, I will advocate that two processes of state legibility formation should be distinguished: state legibility brought by its information collection infrastructure, followed by a modification to this state legibility brought by the distortion happening in the bureaucratic reporting system. I will then provide an empirical approximation of the infrastructure-based state legibility and its subnational variation based on a quantitative evaluation of the quality of age data in recent decennial censuses in China and explore exploratory factors contributing to the subnational variation. I will finally use the subnational variation in census data quality as a predictor of the irregularities in socioeconomic statistics like GDP to approximate the modification component brought by bureaucratic juking.

This project aims to provide one of the first quantitative evaluations of the magnitude and the subnational variation of Chinese state legibility and its various sources. The subnational-level analysis complements the national-level data-quality checking efforts made by demographers (Cai and Feng 2021). The findings of the project will have additional implications for the sociological understanding of population census as an information-extracting tool of the state and thus, a way of exerting its infrastructural power over the society (Mann 2012).

Research Design

I will seek a mixed pool of quantitative and qualitative evidence to estimate state legibility and to understand its formation process. I will first conduct a statistical analysis of population census data

for a description of the distribution of the Chinese state legibility. I will conduct interviews with leading demographers in China and read archival about the post-Mao re-establishment and operation of the statistics and census bureau to understand the data collection and legibility formation process.

Through statistical modeling of the age-heaping patterns in post-Mao decennial censuses, I will examine the regional variation in the accuracy of population censuses and treat it as a measure of local state legibility. Demographers have known well about the "ideal" characteristics of age distribution if the data quality is good. Compared to other socioeconomic indicators like Gross Domestic Product (GDP) or Total Fertility Rates (TFR), age is a less politically and socially sensitive statistic for the public and the CCP. Known characteristics for an ideal distribution and being mundane make the pattern of age distribution a particularly good candidate to minimize the possibility of bottom-up and top-down information distortion and capture the "pure" quality of information collection.

Technically, I will apply the data quality measure developed by Lee and Zhang (2017) on microdata of the population census grouped at the prefectural level, the lowest geographical unit provided in population censuses. Lee and Zhang's measure was based on the smoothness and heaping pattern of age distribution. I look forward to the opportunity of introducing the latest methodological development in quality-checking techniques of census data for an improved measure. Four to five waves of decennial census data will be used, of which the 1982, 1990, and 2000 censuses have been open to public access through the Integrated Public Use Microdata Series (IPUMS) project. Microdata for the 2010 population census is only available through physical access at Tsinghua University after approval. Depending on the availability of data at the time of research, I would like to add the 2020 population census to the analysis. I am also keen on learning the latest in data quality checking developed in recent years. After the measurement has been established, I will use state legibility as a dependent variable and explore what factors explain the variation in the accuracy of the population census.

Moreover, I will conduct interviews with leading Chinese demographers at Fudan, Peking, and Renmin University, and examine archives about the establishment and daily operation of statistical infrastructure in post-Mao China. I am particularly interested in whether the professionalization of the statistical bureaucracy(Ghosh 2020) and the extensive appearance of the Chinese Communist Party (Koss 2018) contribute to the state legibility of the Chinese state.

Preliminary Findings

I have conducted preliminary analysis using the 1982, 1990, and 2000 census data, and successfully made a map of the age heaping pattern of population censused in China at the prefecture level.

Figure 1 (attached in the end) Changing legibility of the Chinese government from 1982-2000. Deep areas represent higher age heaping and indicate weak state legibility in this area. Myers and Whipper indexes are calculated using the method by Lee and Zhang (2017). Based on the 1982, 1990 and 2000 Chinese Population Census data from IPUMS(2018).

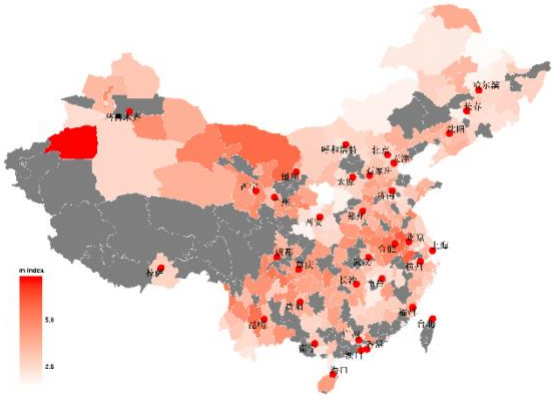
I constructed prefectural level Whipple and Myers indexes using the micro level census data from the Integrated Public Use Microdata Series (IPUMS). IPUMS provides academic access to the 1982, 1990 and 2000 waves of the census with the lack of the latest 2010 census data. I combine the computed indexes with GADM shape files of Chinese prefectural administrative boundaries and make the maps using the simple feature package of R.

On the maps, deep colors represent higher age heaping and suggest poorer knowledge of the real population age, while light reds indicate regions where the information gathering from the state apparatus is more successful. The grey areas represent where indexes are not available, mostly caused by changes in administrative division.

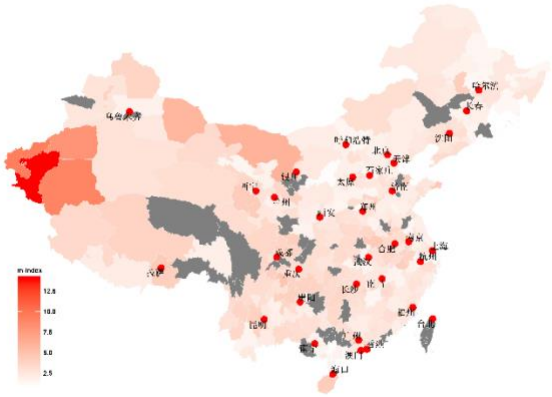
The maps show some interesting cross-sectional and longitudinal patterns. In general, the communist regime has good knowledge of their citizens with less than 5 Myers score, far smaller than the world average (8.21) reported by Lee and Zhang(2017). The informative agencies are more effective in the east coastal areas and in the north where the party organization is more developed, though it is somewhat surprising to observe some strong age continuity in the northeast. The weakest prefectures lie in the western boundaries in Xinjiang and Tibet, which is consistent with common expectations.

The most striking characteristic in a longitudinal gaze at the maps will be the large portion of missing data in the 1982 census. It displays that the administrative division of prefectures has experienced a huge transformation since the 1980s. Compared to 1990 and 2000, the prevailing deep reds in the 1982 map demonstrate the weakness of the state bureaucracy after the cultural revolution. The maps resonate with Vivian Shue's judgment that instead of the retreat of the state from the socioeconomic activities, China is rehabilitating its bureaucratic and party apparatus and strengthening its monitor and control of citizens' life in the past decades (Shue 1990). Decentralization reforms and flexibility for local experimentation should be understood as intentional strategies from Beijing to boost economic growth. Another striking discovery is the age heaping increased in south Xinjiang and Tibet from 1990 to 2000, contrary to the impression that the CCP achieved progress in governing Xinjiang and Tibet in the 1990s (Becquelin 2004).

Prefecture-level Myers Index in China: 1982

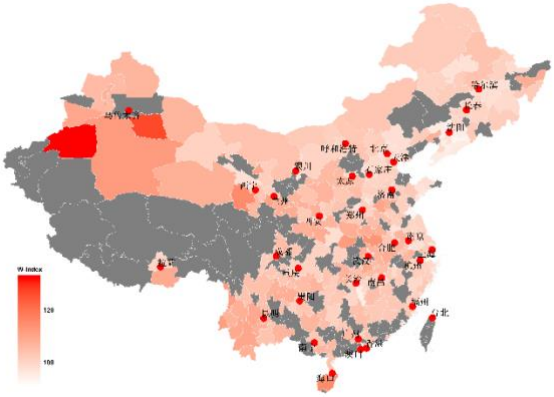


Prefecture-level Myers Index in China: 1990



Prefecture-level Myers Index in China: 2000

Prefecture-level Whipper Index in China: 1982



Prefecture-level Whipper Index in China: 1990

Prefecture-level Whipper Index in China: 2000

References

Becquelin, Nicolas. 2004. “Staged Development in Xinjiang.” *The China Quarterly* 178:358–78. doi: 10.1017/S0305741004000219.

Bourdieu, Pierre. 1994. “Rethinking the State: Genesis and Structure of the Bureaucratic Field.” *Sociological Theory* 12(1):1. doi: 10.2307/202032.

Cai, Yong, and Wang Feng. 2021. “The Social and Sociological Consequences of China’s One-Child Policy.” *Annual Review of Sociology* 20.

Ding, Iza. 2022. *The Performative State: Public Scrutiny and Environmental Governance in China*. Ithaca [New York]: Cornell University Press.

Ghosh, Arunabh. 2020. *Making It Count: Statistics and Statecraft in the Early People’s Republic of China*. Princeton: Princeton University Press.

Jerven, Morten. 2013. *Poor Numbers: How We Are Misled by African Development Statistics and What to Do about It*. Ithaca: Cornell University Press.

Koss, Daniel. 2018. *Where the Party Rules: The Rank and File of China’s Communist State*. 1st ed. Cambridge University Press.

Lee, Ching Kwan, and Yonghong Zhang. 2013. “The Power of Instability: Unraveling the Microfoundations of Bargained Authoritarianism in China.” *American Journal of Sociology* 118(6):1475–1508. doi: 10.1086/670802.

Mann, Michael. 1984. “The Autonomous Power of the State: Its Origins, Mechanisms and Results.” *European Journal of Sociology / Archives Européennes de Sociologie* 25(2):185–213. doi: 10.1017/S0003975600004239.

Mann, Michael. 2012. “A Theory of the Modern State.” Pp. 44–91 in *The Sources of Social Power, Volume 2: The Rise of Classes and Nation-states, 1760–1914*. Vol. 1. New York: Cambridge University Press.

Shue, Vivienne. 1990. *The Reach of the State: Sketches of the Chinese Body Politic*. Stanford University Press.

Svolik, Milan W. 2012. *The Politics of Authoritarian Rule*. Cambridge: Cambridge University Press.

Travers, S. Lee. 1982. “Bias in Chinese Economic Statistics: The Case of the Typical Example Investigation.” *The China Quarterly* 91:478–85. doi: 10.1017/S0305741000000679.

Tsai, Lily L. 2008. “Understanding the Falsification of Village Income Statistics.” *The China Quarterly* 196:805–26. doi: 10.1017/S0305741008001136.

Wallace, Jeremy L. 2016. “Juking the Stats? Authoritarian Information Problems in China.” *British Journal of Political Science* 46(1):11–29. doi: 10.1017/S0007123414000106.

Wallace, Jeremy L. 2022. *Seeking Truth and Hiding Facts: Information, Ideology, and Authoritarianism in China*. Oxford, New York: Oxford University Press.