The Reach of the State
Online Appendix

Comparative Political Studies

Charles Chang
Duke Kunshan University
charles.c.chang@dukekunshan.edu.cn

Yuhua Wang
Harvard University
yuhuawang@fas.harvard.edu

Last updated July 5, 2023

DATA COLLECTION PROCESS

We collected the POI data from Amap in 2018. The Amap Place Nearby (*zhoubian jiansuo*) Application Programming Interface (API) permits us to use a programming interface to easily access Amap's POI database that stores all such data. The API requires input parameters such as geographical coordinates, distance, and POI types. Once these parameters are set, the API retrieves the relevant list of POIs from Amap's database.

We wrote Python script to retrieve POIs within a radius of a large number of search location points. To retrieve the POIs efficiently, we set our script on a High-Performance Cluster (HPC) and managed the retrieved POIs in a PostgreSQL database. The HPC and database allow our script to run continuously and help efficiently remove any duplicates. To systematically retrieve all POIs in China, we designed a fishnet of 388,201 location search points; each pair is systematically set 0.05 degrees apart. At different places in China, 0.05 degrees indicate slightly different distances on the ground, approximately 3,000–5,000 meters. However, because Amap's API allows us to retrieve POIs within a maximum 10-kilometer radius, this setting enables us to retrieve all POIs in any area of China with a sufficient spatial overlap (see Appendix Figure A1-1).

Amap is not the only mapping service in China. To select a POI data source for our study, we also collected POIs from other major mapping service companies, including Baidu, Tencent, Google, and OpenStreetMap, in a sample urban area of 10,000 square kilometers in China. Although most mapping companies provide data access through web portals or APIs, some limit the number of daily API requests.[1] Moreover, most mapping companies only provide access to the latest version of their POIs, and their data quality varies by their service areas geographically. We compare their services in Appendix Table A1-5.

We collected the POIs from these sources by navigating the same fishnet of search locations. We then compared the completeness, locational precision, and categorical information of their POIs. To do so, we randomly placed 100 points in the sample area. We then manually identified the place, consulting Google Earth images and Tencent Street View panorama. Because the locations from Google Earth images are very precise (Potere 2008), we can use it to identify the precise geolocations of any point. At the identified geolocations, we then used Tencent Street View panorama to identify the name of the place.[2] For example, if the point is located at a Bank of China (BoC) ATM, we labelled its POI as BoC ATM. We then searched the nearby BoC ATM in Amap, Baidu, Tencent, Google, and OpenStreetMap according to its geographic location and estimate the distance between its location on Google Earth images and its POI location. Appendix Table A1-1 reports the average locational error. We also compared the number of POIs in the sample area. Different mapping companies have different norms of POI digitization. For example, Amap often delineates separate gates and buildings in a residential compound, while Tencent only provides a centralized location. In conclusion, we find that while Amap, Baidu, and Tencent all provide excellent mapping services in China, Amap has the most complete database of POIs (Appendix Table A1-1).

[1] Mapping companies often limit API use to protect their data. However, for scholars who only collect POIs on government agencies, the API limitation does not pose a serious threat to data collection.

[2] Tencent Street View panorama uses a China-specific coordinate system. However, the geographical coordinates identified from Google Earth can be easily converted to this system (Chang 2020).

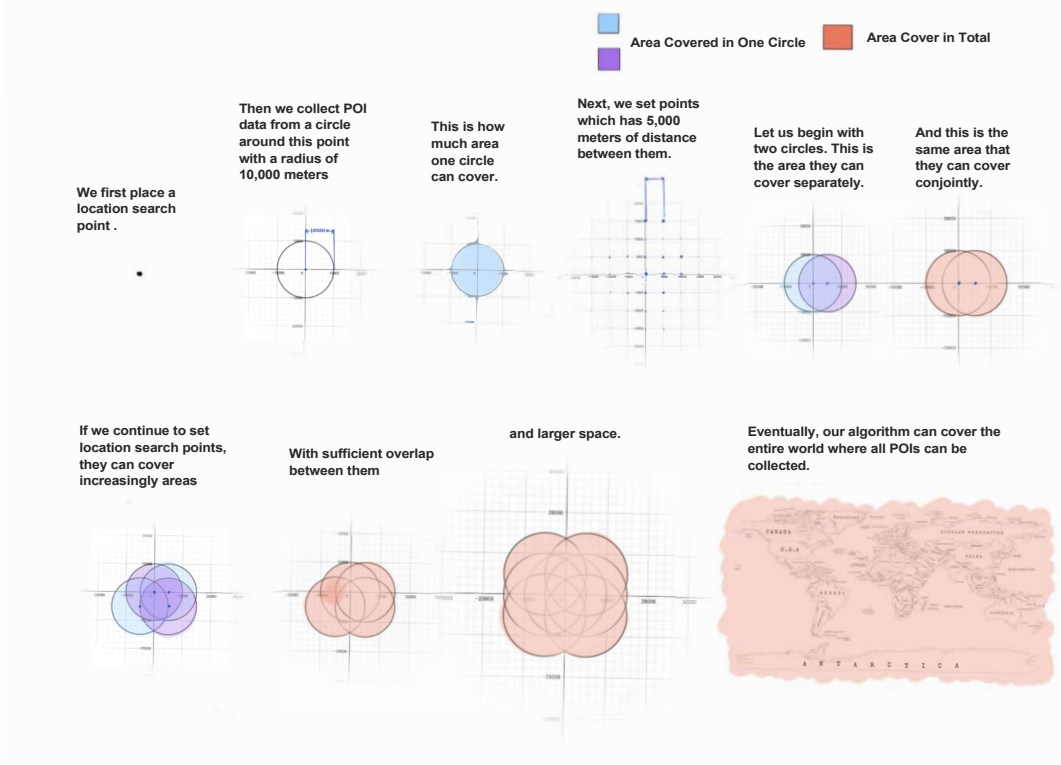


Figure A1-1: POI Data Collection Process

Table A1-1: Coverage and Location Accuracy of Different Location-Based Services

Source	**Tencent**	**Google**	**Amap**	**Baidu**	**OpenStreetMap**
Total number of POIs	192,468	70,428	472,290	101,584	2,049
Geographical error (meter)	5	394	20	3	11
Metadata	Name, address, land use type	Name, land use type	Name, address, land use type	Name, address, land use type, consumers' review	Land use type

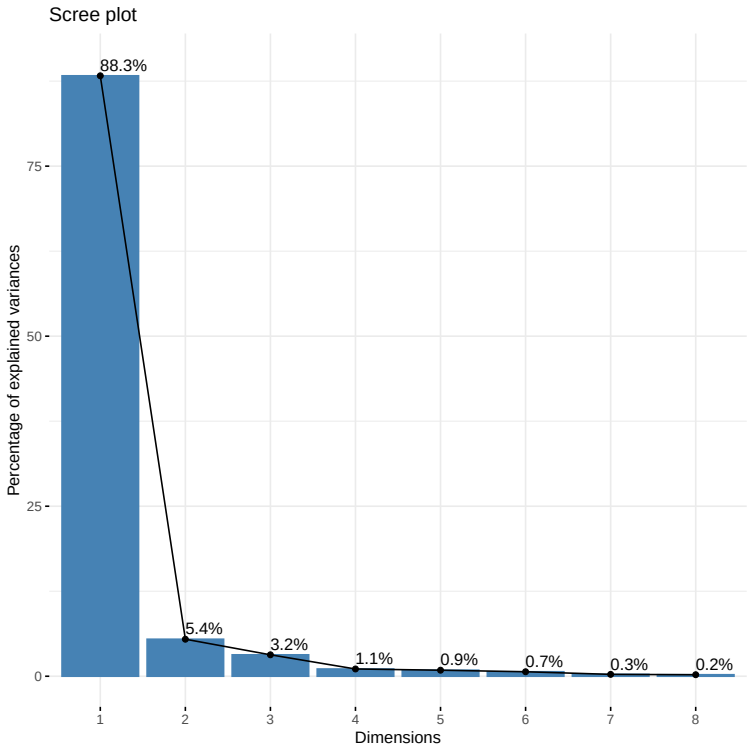
Source: Chang (2020).

Table A1-2: Amap Codes and Categories of State Agencies

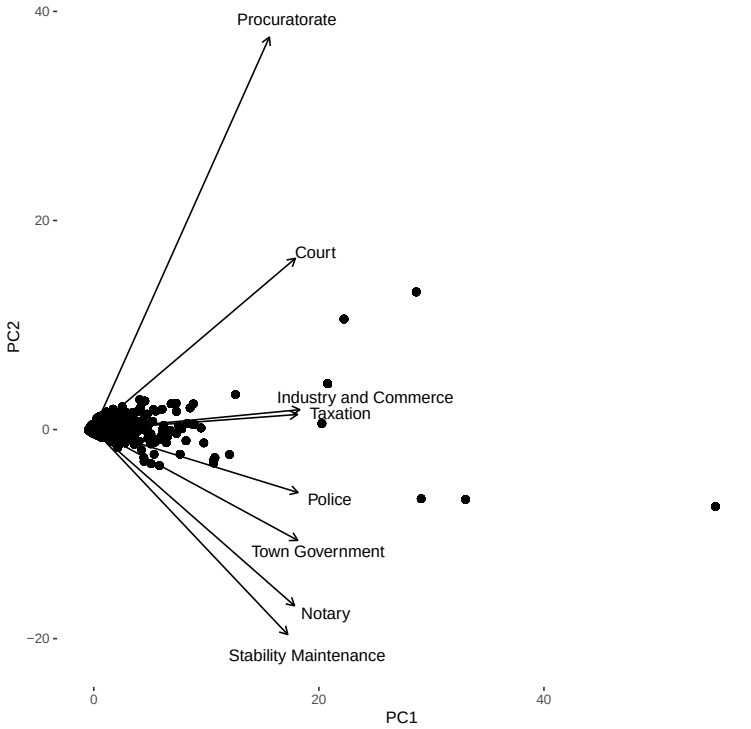
Type	Category
130101	National-Level Government Administration
130102	Provincial-Level Government Administration
130103	Prefectural-Level Government Administration
130104	District- and County-Level Government Administration
130105	Town-Level Government Administration
130106	Below Town Level Government Administration
130501	Police
130502	Procuratorate
130503	Court
130505	Notary
130506	Stability Maintenance
130701	Industry and Commerce
130702	State Tax Authority
130703	Local Tax Authority

PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA is a dimension reduction procedure that transforms correlated variables into a smaller set of important composite variables – principal components (PCs) – that explain the most variability in the original data, plus errors. We first aggregate the data to the township level – the lowest administrative level in China – and calculate the total number of each category of state agency in every township (or street in cities). If the state locates its agencies randomly, then the first PC would not explain much more of the variance in the data than the subsequent PCs. Yet this is not what we observe in the data. Appendix Figure A1-2 (panel (a)) is a scree plot that displays the eigenvalue of each PC, which corresponds to the amount of variance each PC explains in the data. We see that the first PC explains 88.3% of the total variance. Then there is an observable "elbow" from the first to the second PC, indicating a significant reduction in the explanatory power of the latter. Panel (b) maps the loadings of each variable on the first two PCs. The loadings represent the correlations between each of the variables and the estimated components. A component may therefore be substantively interpreted by examining the variables with which it is most highly correlated. Most agencies are correlated with the first PC, while procuratorate and court – both legal institutions – are also correlated with the second PC.



(a) Scree Plot



(b) Biplot of the First Two Principal Components

Figure A1-2: Principal Component Analysis Results

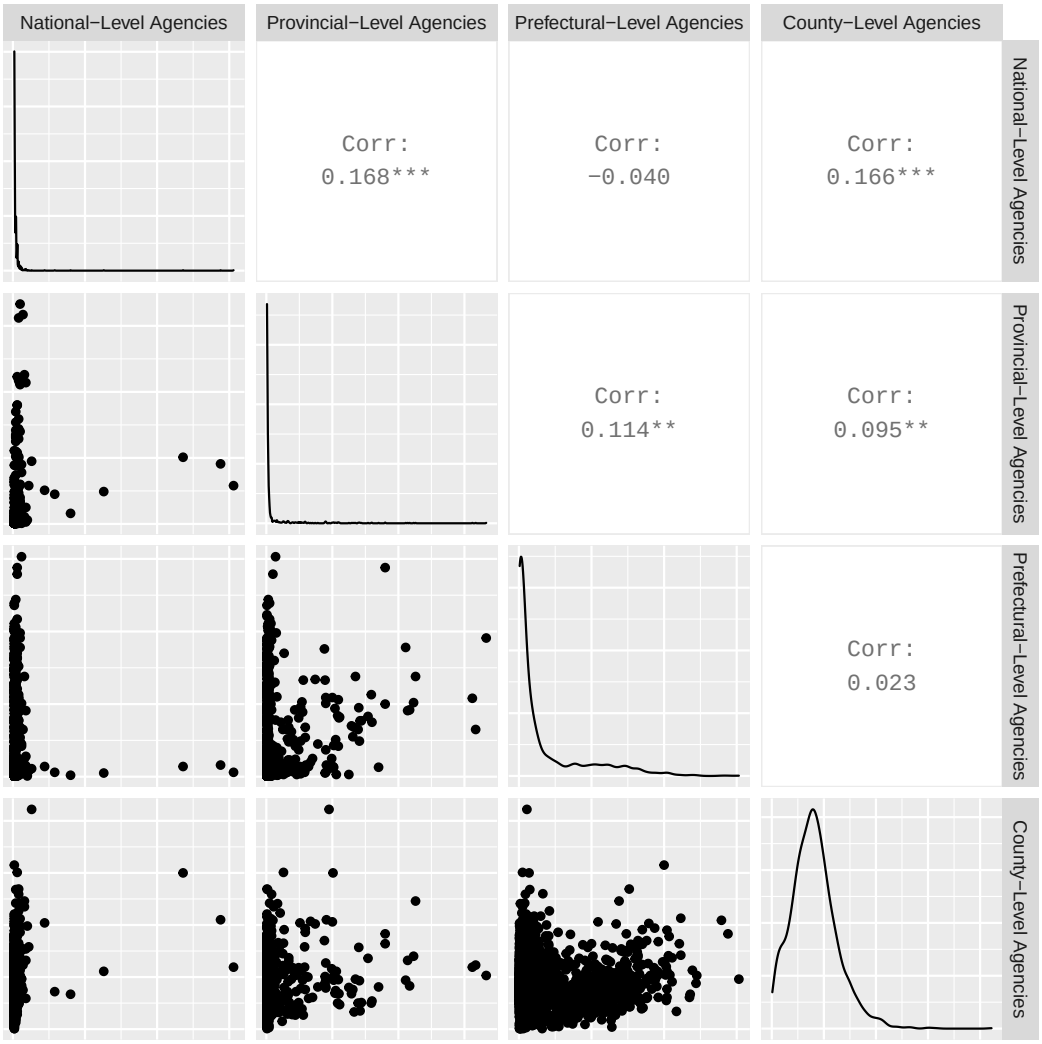


Figure A1-3: Correlations between State Agencies at Different Levels

Notes: The variables are 1) Number of national-level agencies in each county, 2) Number of provincial-level agencies in each county, 3) Number of prefectural-level agencies in each county, and 4) Number of county-level agencies in each county.

MAPPING STATE AGENCIES

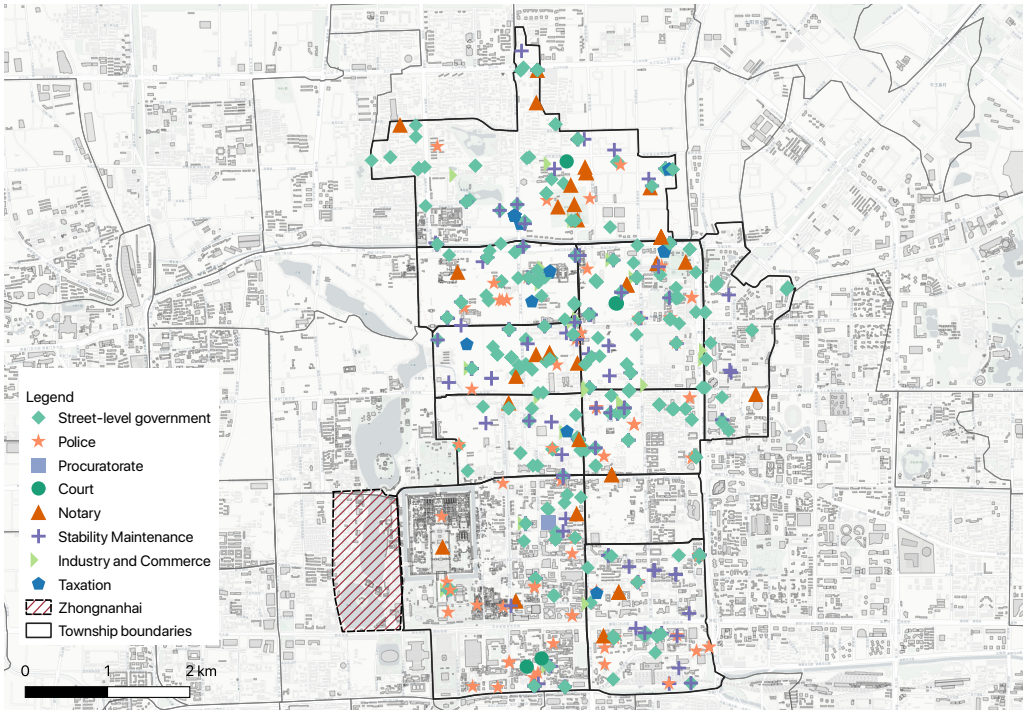


Figure A1-4: Locations of State Agencies in Dongcheng District (Beijing)

STATE AGENCIES AND OFFICIAL STATISTICS

Appendix Figure A1-5 compares counties that have missing data on four important indicators – population, GDP, tax revenue, and fiscal spending – with those that released such data.[3] The counties that did not disclose these statistics had more state agencies on a per capita basis than those that disclosed this information.[4]

There are two possible interpretations regarding why local governments with more state agencies do not disclose important demographic and economic indicators. First, the counties with more per capita state agencies might govern a bigger area, so it is difficult to collect social and economic statistics. Second, the counties with more per capita state agencies might be politically important, so the local governments want to obfuscate these indicators from observers to avoid public scrutiny. We argue that the latter is more plausible because Chinese local officials need to report social and economic statistics to the higher-ups, which use these indicators (e.g., fiscal and economic conditions) to determine local officials' career advancement (Lü and Landry 2014; Landry, Lü, and Duan 2018). This suggests that these local governments have these statistics, but they decide not to make them public. While a definitive answer would require further research, this interpretation is consistent with the argument that some authoritarian governments, despite a higher density of state agencies, might choose not to disclose important economic and fiscal indicators due to transparency or career concerns (Hollyer, Rosendorff, and Vreeland 2015; Wallace 2016).

[3] We rely on statistical yearbooks published by the Chinese government at various levels to measure whether a county has released such data.

[4] We use the 2010 census data, which is more complete, to calculate per capita state agencies.

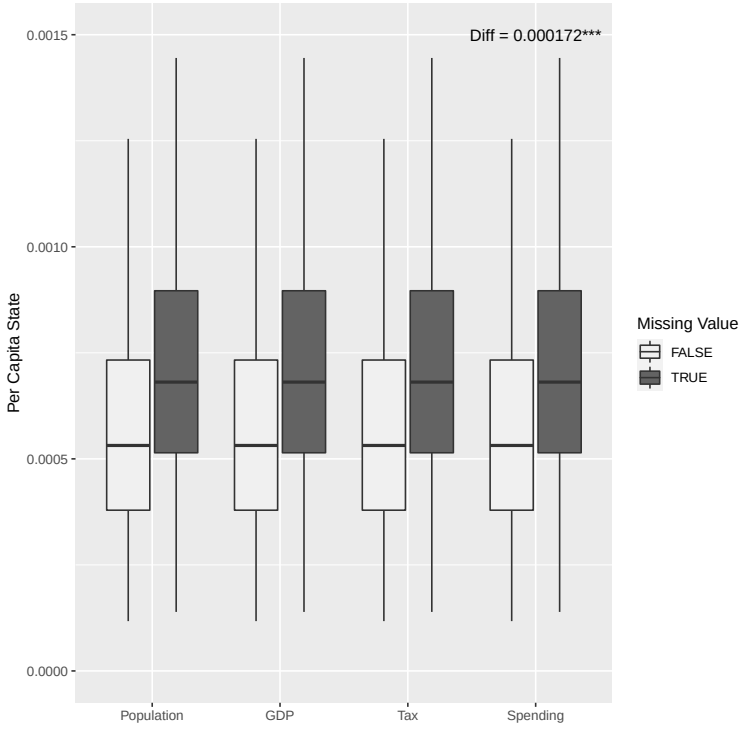


Figure A1-5: State Agencies and Missingness in Official Statistics

Table A1-3: Coercion and Protest: DID Estimates with County-Level Data

Dependent variable:	N of Protests
Year2017×Treatment	-1.384***
	(0.323)
Year2015×Treatment	-0.301
	(0.242)
Year2015	-0.350***
	(0.123)
Year2017	-2.295***
	(0.145)
County FE	Yes
Outcome mean	2.672
Outcome std.dev.	5.142
Observations	7,434
R^2	0.709

Notes: The unit of analysis is county-year. The data is a panel data of Chinese counties in 2013, 2015, and 2017. Treatment is defined as having at least one new stability maintenance agency in 2016. Standard errors are robust, clustered at the county level. P values based on two-tailed tests, * p<0.1, **p<0.05, ***p<0.01

Table A1-4: State Agencies and Protest: DID Estimates with County-Level Data

Dependent variable:	N of Protests			
Year2017×Administration	-0.401			
	(0.292)			
Year2015×Administration	-0.165			
	(0.250)			
Year2017×Court		-0.286		
		(0.304)		
Year2015×Court		0.053		
		(0.225)		
Year2017×Industry and Commerce			-0.336	
			(0.288)	
Year2015×Industry and Commerce			0.042	
			(0.221)	
Year2017×Taxation				-0.329
				(0.281)
Year2015×Taxation				0.010
				(0.218)
Year2015	-0.340	-0.495***	-0.486***	-0.463***
	(0.216)	(0.181)	(0.171)	(0.146)
Year2017	-2.504***	-2.591***	-2.570***	-2.634***
	(0.236)	(0.253)	(0.226)	(0.191)
County FE	Yes	Yes	Yes	Yes
Outcome mean	2.672	2.672	2.672	2.672
Outcome std.dev.	5.142	5.142	5.142	5.142
Observations	7,434	7,434	7,434	7,434
R^2	0.706	0.706	0.706	0.706

Notes: The unit of analysis is county-year. The data is a panel data of Chinese counties in 2013, 2015, and 2017. Treatment is defined as having at least one new state agency (administration, court, industry and commerce, or taxation) in 2016. Standard errors are robust, clustered at the county level. P values based on two-tailed tests, * p<0.1, **p<0.05, ***p<0.01

Our measure can be used to conduct research on state infrastructure in a large number of countries in which official statistics have been opaque or difficult to obtain. In this section, we demonstrate how to apply our measure beyond China using a cross-national database that we constructed using POIs from OpenStreetMap in 2021.

While Amap provides the most competitive location-based services in China, Google Maps and OpenStreetMap are the two most popular mapping platforms in the world. Google Maps is currently the most frequently used location-based service in the world, with over 1 billion monthly active users.[5] OpenStreetMap is a collaborative project that created a free editable geographic database that covers more than 200 countries.[6] It relies on volunteer contributors around the world to collect geodata. Many early contributors were cyclists who surveyed with (and for) other cyclists.[7] Others are geographic information system professionals who contribute data with Esri tools.[8] By 2013, OpenStreetMap had reached 1 million registered contributors worldwide.[9]

We used data from OpenStreetMap rather than Google Maps for two reasons. First, OpenStreetMap offers website download and is accessible on Amazon Web Services and Google Bigquery, while Google Maps requires API access; this allows us to download the data (which is logistically simpler than accessing the data via an API) in a variety of formats and store them as ZIP files to reduce their file sizes.[10] The second reason is that OpenStreetMap is completely free; Google Maps charges a fee for additional information about each POI. Appendix Table A1-5 compares multiple mapping platforms. Although our data demonstration is based on OpenStreetMap, we provide Python code on Github, which researchers can consult if they need to collect POI data from Google Maps.[11]

Appendix Figure A1-6 (panel (a)) illustrates the number of state agencies across the 216 countries for which OpenStreetMap data are available. Appendix Figure A1-6 (panel (b)) shows the number of state agencies per million people in these countries. The figure demonstrates that scale economies manifest at the cross-national level as well: more populous countries, such as China and India, need fewer state agencies per capita than less populated countries, such as the USA, Canada, and Russia. Appendix Figure A1-8 shows a cross-national scatter plot between per capita state agencies (log) and population (log), which follows a power law as well (the estimated slope is -0.3, which indicates that a 1% increase in population (log) is associated with a 0.7% increase in the number of state agencies (log)).

---

[5] https://sites.google.com/a/pressatgoogle.com/google-maps-for-iphone/google-maps-metrics (accessed October 28, 2021).

[6] https://www.openstreetmap.us/ (accessed October 28, 2021).

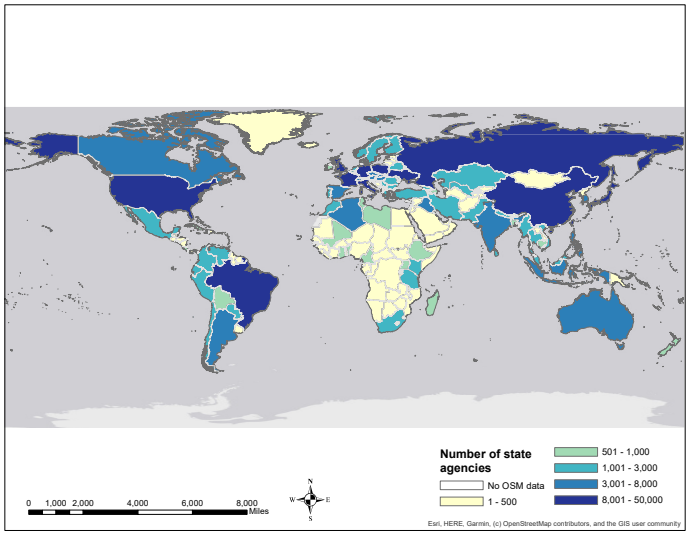
[7] http://www.opencyclemap.org/docs/ (accessed October 28, 2021).

[8] https://www.esri.com/news/releases/10_3qtr/openstreetmap.html (accessed October 28, 2021).

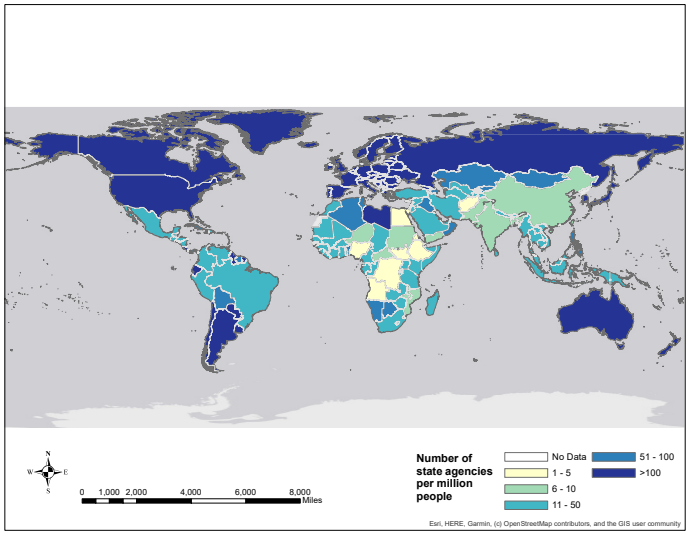
[9] https://blog.openstreetmap.org/2013/01/06/1-million-openstreetmappers/ (accessed October 28, 2021).

[10] OpenStreetMap is also the only platform that provides archived data back to 2005, so researchers can construct a panel dataset of more than 200 countries over 15 years, which allows them to study more countries than what has been possible using existing data sources.

[11] https://github.com/placeasmedia/stateagencies/tree/main/data_collections/google_poi_collection.



(a) Number of State Agencies across Countries



(b) Number of State Agencies Per Million People across Countries

Figure A1-6: Number of State Agencies around the World

We validate this cross-national dataset by examining the correlations between per capita state agencies (log) and a group of variables, including Polity score, per capita GDP (log), per capita government spending (log), and per capita government employees (log).

The correlation plots, displayed in Appendix Figure A1-7, reveal three important findings. First, more democratic countries, proxied by Polity scores (Marshall, Gurr, and Jaggers 2014), have more state agencies per capita. There are two possible interpretations. The first is that democracies generally have larger governments since they can rely on consent to extract more revenues and build more government agencies (North and Weingast 1989) or because they need

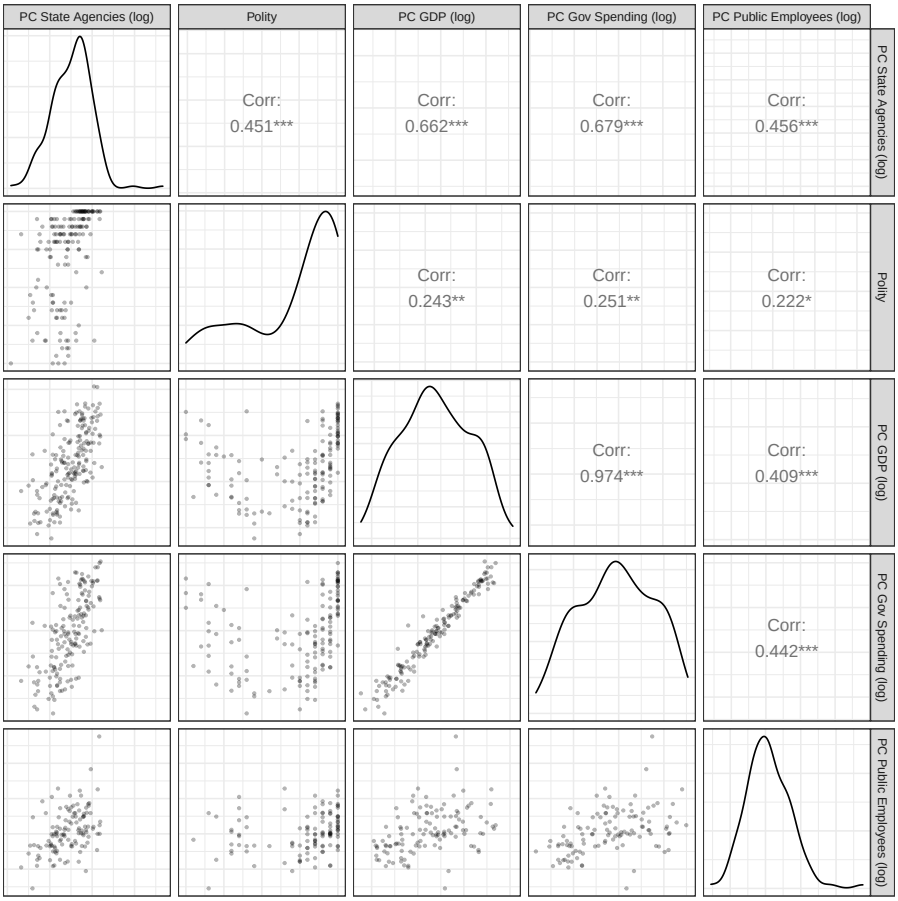


Figure A1-7: Correlation Plots of Cross-National Indicators

a bigger government to respond to redistributive demands (Meltzer and Richard 1981). Alternatively, it may be the case that it is more accessible for volunteer contributors in democracies, which are more transparent and open (Hollyer, Rosendorff, and Vreeland 2015), to collect data on government agencies.

The second main finding is that wealthier countries, measured by per capita GDP, have more state agencies per capita. This might simply be a corollary of the first finding: wealthier countries tend to be democracies, which have larger governments. In addition, as Besley and Persson (2009) show, state capacity and economic development are highly correlated because a strong state provides public goods and protects property rights, thus promoting economic development, which in turn increases government revenues.

Thirdly, our measure of per capita state agencies is significantly correlated with two other measures scholars often use to measure government size: per capita government spending and

per capita public employees.[12] But our measure has a clear advantage: while our data cover 216 countries, the World Bank data on government spending include only 159 countries, and data on public employees for only 119 countries.

We hypothesize that many countries did not disclose their data to the World Bank *not* because they did not have the capacity to collect it, but because they intentionally obfuscated these statistics from the public. As Malesky, Schuler, and Tran (2012) and Hollyer, Rosendorff, and Vreeland (2015) argue, transparency would generate unwelcome public scrutiny and, in authoritarian regimes, trigger mass protests. To corroborate this idea, we examined whether authoritarian regimes are more likely than democracies to hide data from the public, controlling for each country's level of state infrastructure. Appendix Table A1-6 reports the results. The dependent variable is an indicator that equals 1 if the country did not disclose its data on public employees to the World Bank. The independent variable is the country's Polity score. The regression controls for per capita GDP (log), number of state agencies per million (log) from OpenStreetMap, and population (log). The estimates are consistent with our claim that, holding constant the number of state agencies per million (log), more democratic countries are more likely to disclose their data to the World Bank.

[12] We obtained both variables from the World Bank. See https://data.worldbank.org/indicator/NE.CON.GOVT.ZS and https://govdata360.worldbank.org/indicators/haa733075?country=BRA&indicator=42305&viz=line_chart&years=2000,2017 (both accessed October 28, 2021).

Table A1-5: Comparison of Major Location-Based Services

Source	Accessibility	Availability	Spatial Coverage	Temporal Coverage	Update Frequency
OpenStreetMap	Website download; also accessible on Amazon Web Services (AWS) and Google Bigquery	Free	More than 200 countries	2005 – present	Weekly
Amap	Gaode Nearby API	Free for 30,000 API requests per day (with a registered account) or request by approval with an enterprise account	Mostly China	Present only	Daily
Baidu	Baidu Nearby API	Free for 2,000 API requests per day	Mostly China	Present only	Three months
Google	Google Place Search API	Free for basic information of POI or pay for additional information	More than 200 countries	Present only	Daily
Tencent	Tencent WebService Place API	Free for 10,000 API requests per day	Mostly China	Present only	Three months
SafeGraph	Downloadable through ArcGIS Marketplace; also accessible on AWS	Basic information free for educators and researchers or pay for additional information for non-researchers	USA, UK, and Canada	2018 – present	Monthly

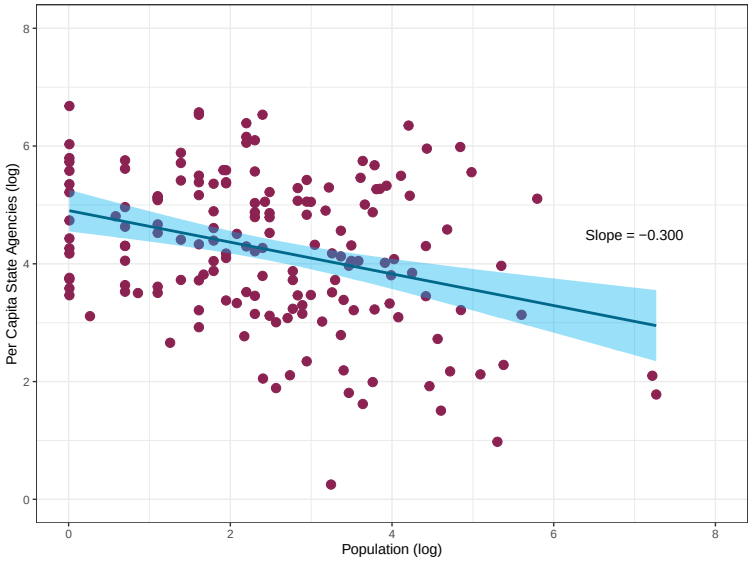


Figure A1-8: Economies of Scale in State Administration across Countries

Table A1-6: Patterns of Missingness in World Bank Data: OLS Estimates

Dependent variable:	Missingness in Public Employee Data (World Bank)
Polity score	-0.031***
	(0.006)
Per capita GDP (log)	0.076**
	(0.031)
N of state agencies per million (log)	0.051
	(0.035)
Population (log)	-0.007
	(0.023)
R^2	0.212
Observations	151

Notes: The data is a cross section of countries. Dependent variable is an indicator that equals 1 if the country did not disclose its number of public employees to the World Bank. Standard errors are robust. P values based on two-tailed tests, * p<0.1, **p<0.05, ***p<0.01

Lastly, we also examine how our measure is associated with existing measures of state capacity. For our measure – OSM – we use the number of state agencies per million people from OpenStreetMap in 2021. The existing measures of state capacity include: 1) Hanson/Sigman, which is a latent variable of different dimensions of state capacity developed by Hanson and Sigman (2021); 2) State Antiquity, which is an index estimated by Bockstette, Chanda, and Putterman (2002); 3) WGI, which is an indicator measuring government effectiveness included in the Worldwide Governance Indicators; 4) Fragile State, an index developed by Rice and Patrick (2008); 5) Tax/GDP ratio, which is the share of tax revenue in GDP (Besley and Persson 2009); 6) Myers Index (log), an index developed by Lee and Zhang (2017) by examining national census age heaping. For measures (1)-(5), we take their 2006-2015 averages. For the Myers Index, we take the latest year in which the index was available. Hanson/Sigman, State Antiquity, WGI, and Tax/GDP are positive measures of state capacity: the higher the index, the stronger the state. Fragile State and Myers Index are negative measures: the higher the index, the weaker the state.

Appendix Figure A1-9 shows the correlation plot of these measures. All correlation coefficients are statistically significant and have the right signs. OSM is positively correlated with Hanson/Sigman, State Antiquity, WGI, and Tax/GDP and negatively correlated with Fragile State and Myers Index. Our measure is particularly correlated well with the more comprehensive indexes, such as Hanson/Sigman, Fragile State, and Myers Index, indicating the multidimensional nature of our data.

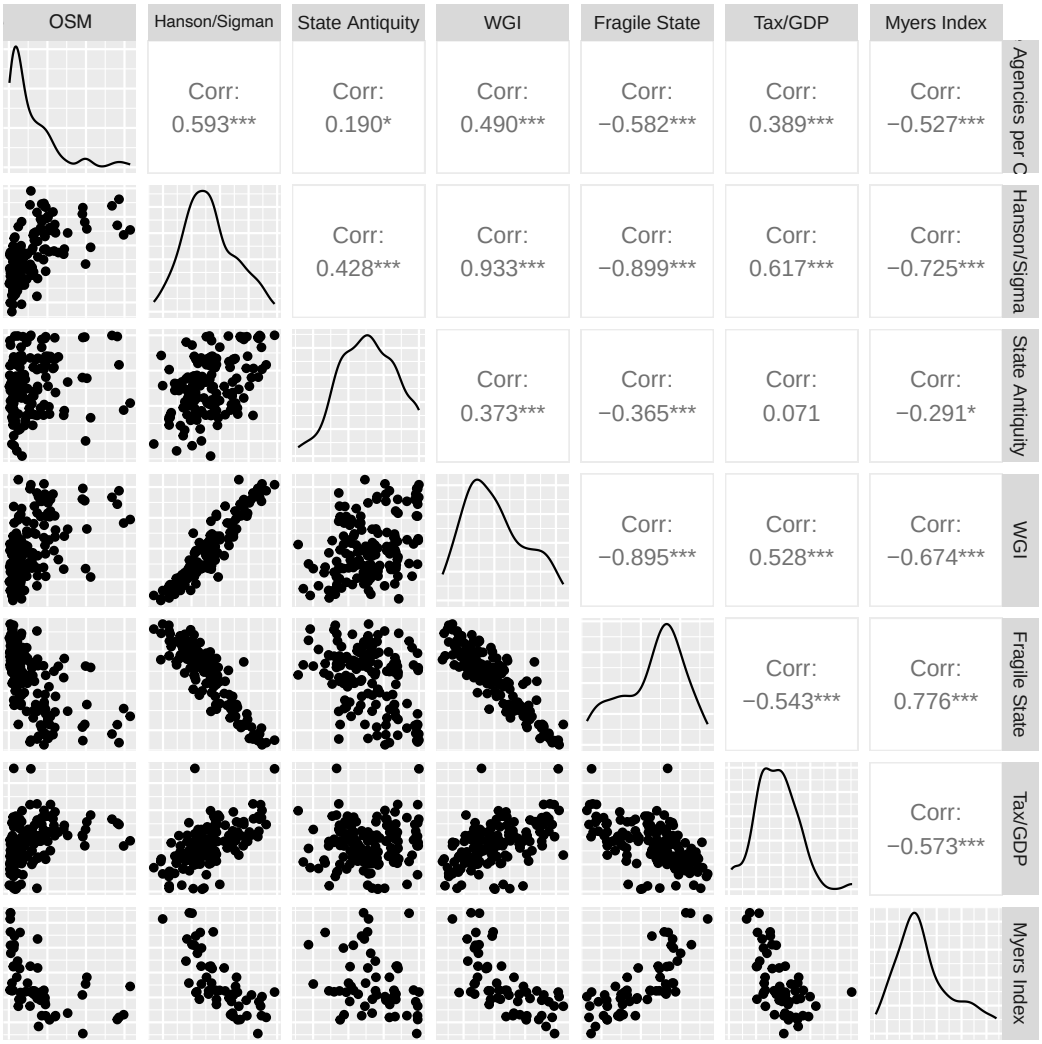


Figure A1-9: Correlation Plot of State Infrastructure and State Capacity Measures

References for the Online Appendix:

Besley, Timothy, and Torsten Persson. 2009. "The Origins of State Capacity: Property Rights, Taxation, and Politics." *American Economic Review* 99 (4): 1218–44.

Bockstette, Valerie, Areendam Chanda, and Louis Putterman. 2002. "States and Markets: The Advantage of an Early Start." *Journal of Economic Growth* 7 (4): 347–69.

Chang, Charles. 2020. "A Data Driven Approach to Study the Social and Political Statuses of Urban Communities in Kunming." *Journal of Chinese History* 4(2): 461–481.

Hanson, Jonathan K, and Rachel Sigman. 2021. "Leviathan's Latent Dimensions: Measuring State Capacity for Comparative Political Research." *Journal of Politics* 83 (4): 1495–1510.

Hollyer, James R, B Peter Rosendorff, and James Raymond Vreeland. 2015. "Transparency, Protest, and Autocratic Instability." *American Political Science Review* 109 (4): 764–784.

Landry, Pierre F, Xiaobo Lü, and Haiyan Duan. 2018. "Does Performance Matter? Evaluating Political Selection along the Chinese Administrative Ladder." *Comparative Political Studies* 51 (8): 1074–1105.

Lee, Melissa M., and Nan Zhang. 2017. "Legibility and the Informational Foundations of State Capacity." *Journal of Politics* 79 (1): 118–32.

Lü, Xiaobo, and Pierre F Landry. 2014. "Show Me the Money: Interjurisdiction Political Competition and Fiscal Extraction in China." *American Political Science Review* 108 (03): 706– 722.

Malesky, Edmund, Paul Schuler, and Anh Tran. 2012. "The Adverse Effects of Sunshine: A Field Experiment on Legislative Transparency in an Authoritarian Assembly." *American Political Science Review* 106(4): 762–786.

Marshall, Monty G, Ted Robert Gurr, and Keith Jaggers. 2014. "Polity IV Project: Political Regime Characteristics and Transitions, 1800–2013." *Center for Systemic Peace*.

Meltzer, Allan H, and Scott F Richard. 1981. "A Rational Theory of the Size of Government." *Journal of Political Economy* 89 (5): 914–927.

North, Douglass C., and Barry R. Weingast. 1989. "Constitutions and Commitment: The Evolution of Institutions Governing Public Choice in Seventeenth-Century England." *The Journal of Economic History* 49:803–832.

Potere, David. 2008. "Horizontal Positional Accuracy of Google Earth's High-Resolution Imagery Archive." *Sensors* 8(12): 7973–7981.

Rice, Susan E., and Stewart Patrick. 2008. "Index of State Weakness in the Developing World." *Brookings Institution, Washington, DC.*