SCIENTIFIC DATA

Check for updates

OPEN

DATA DESCRIPTOR

A gridded establishment dataset as a proxy for economic activity in China

Lei Dong [1,2,3 ✉], Xiaohui Yuan[3], Meng Li[3], Carlo Ratti[2] & Yu Liu [1 ✉]

Measuring the geographical distribution of economic activity plays a key role in scientific research and policymaking. However, previous studies and data on economic activity either have a coarse spatial resolution or cover a limited time span, and the high-resolution characteristics of socioeconomic dynamics are largely unknown. Here, we construct a dataset on the economic activity of mainland China, the gridded establishment dataset (GED), which measures the volume of establishments at a 0.01° latitude by 0.01° longitude scale. Specifically, our dataset captures the geographically based opening and closing of approximately 25.5 million firms that registered in mainland China over the period 2005–2015. The characteristics of fine granularity and long-term observability give the GED a high application value. The dataset not only allows us to quantify the spatiotemporal patterns of the establishments, urban vibrancy, and socioeconomic activity, but also helps us uncover the fundamental principles underlying the dynamics of industrial and economic development.

Background & Summary

An uneven geographical distribution of economic activity has a crucial impact on society[1]. It is a driving force behind a variety of social phenomena, such as population migration, industrial upgrading, and segregation. Detailed geographic data on economic activity, despite their importance, are not readily available[2]. Statistical departments in a few developed countries provide aggregated establishment data to measure economic activity geographically. The Census Bureau of the United States, for example, publishes annual statistics for businesses with paid employees. The data are available by ZIP code or at coarser administrative levels[3]. Similarly, Statistics Sweden complies a gridded Swedish socioeconomic dataset that covers all individuals and all establishments in Sweden[4]. However, such data are not available in China or most other countries in the world. In China, the main sources of socioeconomic data are statistical yearbooks, which only provide aggregated figures at the city level and cannot be used to analyze socioeconomic activity inside cities. Meanwhile, the comparability of the statistics from yearbooks across regions has been questioned by researchers due to inconsistent statistical standards in different regions[5].

Because of the limitations of official sources, researchers often seek to use alternative data as proxies for existing measures of economic output[6]. Some 'big data' sources – e.g., nighttime lights[7,8], streetview imagery[9–11], mobile phone data[12,13], social media data[14], and restaurant data[15,16] – have been examined by investigating their relationships with traditional indicators of socioeconomic activity (e.g., population, GDP, and income). However, the time span covered by most big data sources is very limited; therefore, the relevant research mainly deploys cross-sectional analysis, and it is difficult to track temporal changes of a small area. Importantly, due to potential biases in the spatiotemporal coverage of big data, even with long-term observations, constructing a consistent and reliable measure is still an enormous challenge for scientists[17,18]. For instance, the number of users of different web platforms – the main sources for big data – are changing dynamically. It is difficult to distinguish whether the changes reflected in the data are caused by changes in the number of users of the platform itself or changes in the real world. Additionally, many of these big data sources may not be publicly available or may face privacy issues.

Under the above background, the nighttime light (NTL) dataset, a publicly available source of data covering most of the world land areas since 1992, becomes the most widely used dataset for socioeconomic

[1]Institute of Remote Sensing and Geographical Information Systems, School of Earth and Space Sciences, Peking University, Beijing, 100871, China. [2]Senseable City Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA. [3]QuantUrban Inc., Beijing, 100176, China. ✉e-mail: arch.dongl@gmail.com; liuyu@urban.pku.edu.cn
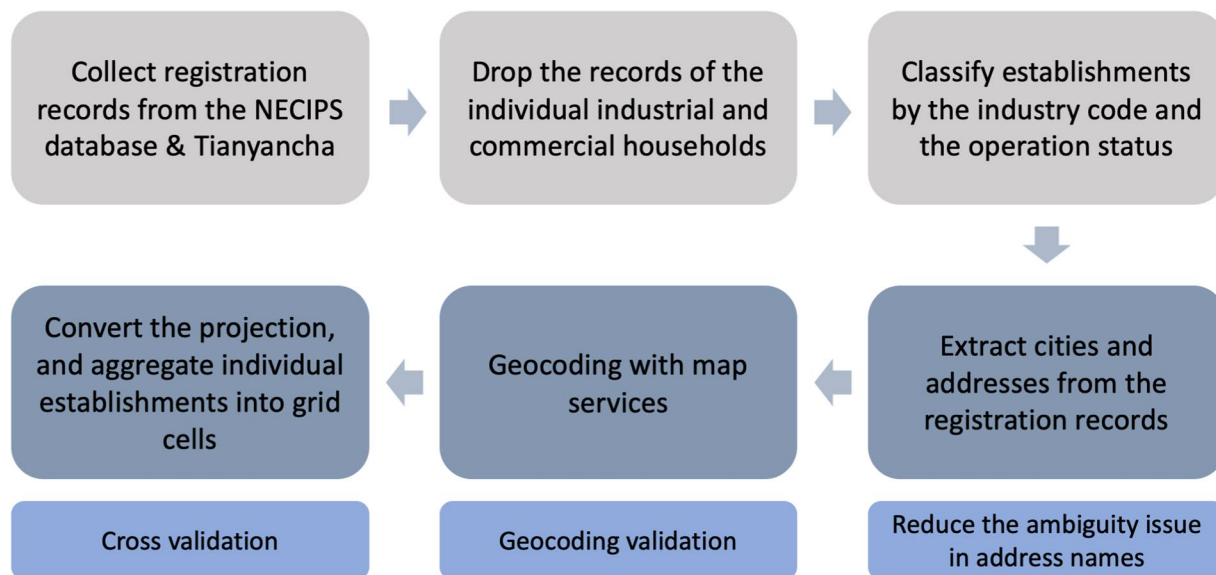
Fig. 1 Schematic overview of the method used to generate the gridded establishment dataset (GED).

Industry	Subgroup
Primary	Agriculture; forestry; husbandry; fishery
Secondary	Mining; manufacturing; construction
	Production and supply of electricity, water and gas
Tertiary	Transport, storage and post services
	Accommodation and catering services
	Leasing and business services; education; finance
	Scientific research and technical services
	Wholesale and retail; real estate
	Culture, sports and entertainment
	Health and social work
	Resident services, maintenance and other service
	Public management, social security, and social organization
	Information transmission, software, and IT services
	Water conservancy, environment, and public facilities management

Table 1. Industry classification.

activity measurement[4,19–21]. However, there are still some limitations in the NTL data. First, it mainly reflects electricity-powered illumination at night[4], which does not necessarily match real socioeconomic activity. Second, because of the saturation effect, NTL can easily reach the upper limit of luminance value within cities, making it difficult to capture the variation of economic activity inside cities. Third, for each grid cell, there is only the brightness value provided by NTL; however, development modes of two regions with the same luminance value (e.g., a manufacturing park and a high-tech park) may be very different. NTL cannot capture this potentially heterogeneous development. A comprehensive comparison between NTL and micro-level socioeconomic variables can be found in ref. [4]. In addition to NTL data, mobile phone data and social media data have also been used in recent studies to investigate economic activity[14,22,23]. Although generally regarded as effective data sources, three main reasons limit their wide application in the estimation of economic activity. First, access to such data is limited. Due to privacy issues, operators or web platforms are reluctant to release their data. Second, mobile phone data or social media data mainly measure users' online social activity, which is not directly related to economic activity. Third, there are often multiple platforms in a region, so that the information reflected in mobile phone data or social media data usually represent the users covered by a particular platform, i.e., the data may not be representative[24,25].

In this paper, using approximately 25.5 million firm registration records, we construct a geographically based dataset – the gridded establishment dataset (GED) – on the economic activity of mainland China. To the best of our knowledge, this is the first geocoded establishment dataset covering mainland China. An important advantage of this dataset is that it simultaneously provides high spatial resolution (0.01° latitude by 0.01° longitude, approximately 1.1 km × 1.1 km at the equator) and long-duration observations (2005–2015, 11 years). For each
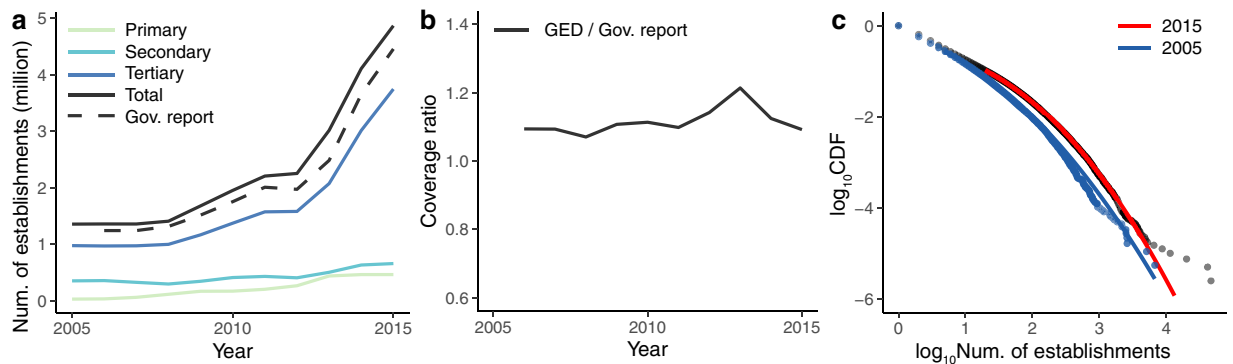
Fig. 2 Basic statistics of the GED. (**a**) Number of establishments in different sectors each year. (**b**) The coverage ratio of the GED. (**c**) The cumulative distribution function (CDF) of establishments at the grid cell level. The log-normal fitting lines are estimated by the *poweRlaw* package in R[45].

grid cell, we also provide the number of establishments by industry, which allows us to analyze the heterogeneous development trajectories of a region. The workflow to construct the dataset is shown in Fig. 1 and detailed in the Methods section.

To validate this dataset, we analyze the accuracy of the geocoding process and compare the GED with several socioeconomic indicators and NTL data at the city level. The results show that the GED can effectively reflect socioeconomic activity and achieves a better performance than that of the NTL. The characteristics of fine granularity and long-term observability give this dataset high application values. It not only can allow researchers and policymakers to quantify the spatiotemporal patterns of economic activity but also can help us uncover the fundamental principles underlying the dynamics of industrial development. Additionally, having a greater availability of these granular data can advance the potential of machine learning models for urban studies. For example, the GED can serve as tags for satellite/streetview imagery to build machine learning models to predict socioeconomic dynamics[26].

Although the GED is a major step towards measuring the geographical distribution of economic activity in China, we should note that it also has several limitations. First, we only have the registered addresses of firms, which may not be the same as the operational addresses. Especially for some industrial parks, due to some location-based policies, e.g. tax subsidies, some firms may register within these industrial parks but operate elsewhere. Second, because of the change in place names, there may be errors in the geocoding process for historical addresses. Therefore, we limit the starting date of the dataset to 2005, when Amap and Baidu, the location-based service providers we use to geocode the address, began to provide online map services in China. Third, business size (e.g., number of employees or revenue-related information) is undocumented in the original data; thus, we can only use establishment counts to represent economic activity.

Methods

Data source and preprocessing. Each newly established firm in mainland China needs to register at the local Administration for Industry and Commerce (AIC) by providing detailed information, such as firm name, address, industry classification, and stockholders. After approval, the relevant information is publicized on the Internet, and everyone can retrieve these data from an online system named the National Enterprise Credit Information Publicity System (NECIPS, http://www.gsxt.gov.cn/index.html). Our original data are collected from this system via web scraping. Specifically, we collect the names of firm that were established between 2005 and 2015 from Tianyancha[27], a website that provides firm-related information in China. By querying firm names, we then obtain the detail page of each firm. Finally, we collect 83.7 million registration records.

To reduce potential noise in the original data, we perform preprocessing as follows: 1) We drop the record of "individual industrial and commercial household" (57.8 million records), a business type whose scale is generally small – fewer than seven employees by law. Another reason for this filter is that the registered addresses of "individual industrial and commercial households" are often at village or township level, and there is no clear street address to perform geocoding. 2) According to the operational status in the year (2018) of data collection, we classify firms into three categories: *existing*, *cancelled*, and *other*, see the preprocessing code for details. Here, *existing* means that the firm has been operating from the time of registration until (at least) 2018; *cancelled* means that the firm was closed at a certain time between the registration year and 2018; *other* represents the remaining status, such as unknown status and moving out. Note that this classification does not affect the number of establishments in the statistics every year, but only adds a dimension to evaluate the operation status of establishments. 3) According to the industry classification, we further label establishments with tags of primary, secondary, and tertiary sectors (Table 1). After data preprocessing, there are a total of 25,545,850 establishments left.

Figure 2a show the number of newly established firms by industry each year. To demonstrate the representativeness of this dataset, we further collect the official statistics based on a report published by the State Administration for Industry and Commerce[28]. Specifically, according to the report, the number of newly established firms in 2015 was about 12,200 per day, or 4.45 million per year (the dashed line in Fig. 2a). In our data, there were 4.85 million establishments in 2015, accounting for 109.2% of the government released figures
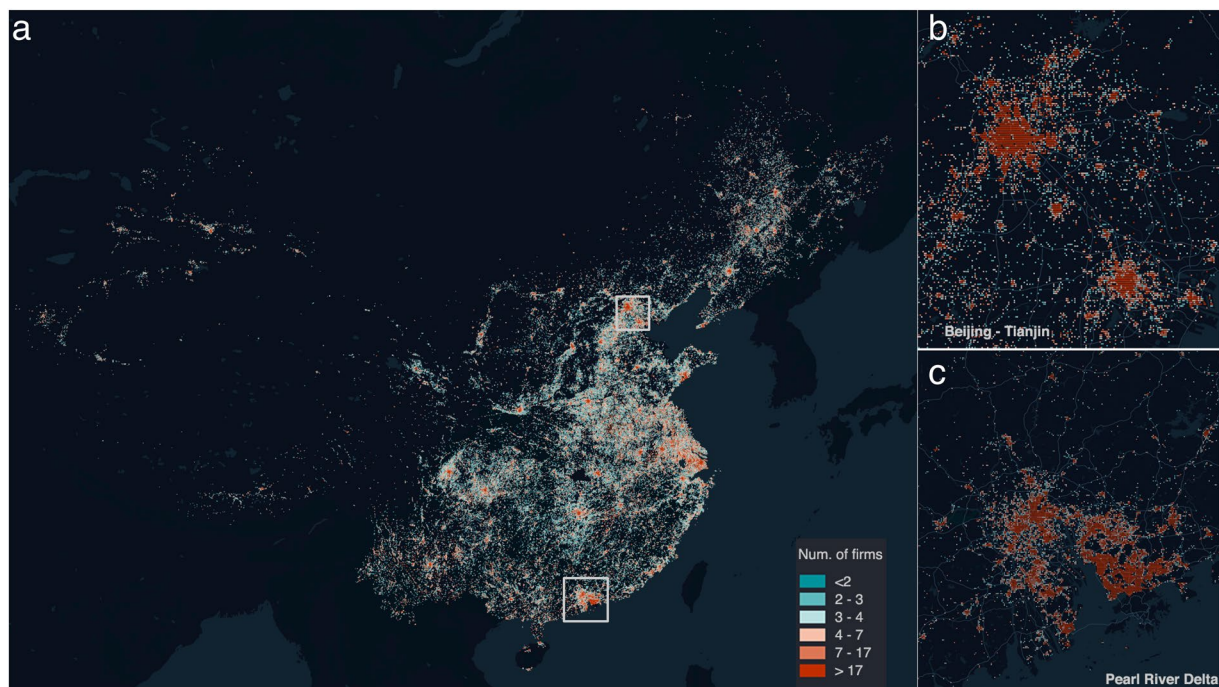
Fig. 3 Geographical distribution of establishments (2015). (**a**) All of mainland China. (**b**) Beijing-Tianjian-Hebei areas. (**c**) Pearl River Delta. The South China Sea Islands are not fully shown on the map. (Visualization tool: Kepter.gl.).

(Fig. 2b). The coverage ratio greater than 100% is probably because the government has subtracted the firms that were cancelled during the registration year when counting the yearly establishments. Overall, our results are identical with those of official statistics (Fig. 2b).

Geocoding. Geocoding is the key step for building a geolocated dataset[29]. To obtain high-quality geocoding results, it is important to know the city to which the firm address belongs. The city and address can then be passed into the online map application program interface (API) to ensure that the search results belong to the input city, which can significantly reduce the problem of ambiguity in place names. In other words, cities may have streets or buildings with the same name, and only querying the street name may return multiple coordinates, for example, one in city *A* and one in city *B*, making it difficult to match the address with the right coordinate. We notice that a considerable proportion of the city data is missing or inaccurate in the NECIPS database. To solve the missing value issue and obtain the coordinates of each firm, we have taken the following steps:

- 1) We build a city list containing the full name and short name of all prefecture-level cities (including *dijishi*, *diqu*, *meng*, and *autonomous prefecture*) and counties (including *shixiaqu*, *xian*, *qi*, and *autonomous county*) in mainland China, see ref. [30] for details on the administrative divisions of China.
- 2) We inspect the city list, if a city name (full name) appears in the address of a registration record, the firm is then classified into this city. For example, the address 'Beijing City, Haidian District, Zhongguancun Street' belongs to Beijing City.
- 3) For cities where the full name cannot be found, we check whether the short name of the city appears in the address. For example, 'Shanghai, Pudong District, and Haiyang Building' belong to Shanghai City because 'Shanghai' is a short name of 'Shanghai City.'
- 4) For cities that can not be matched to any cities, we then inspect the county/district names. For example, the address 'Haidian District, Zhongguancun Street' belongs to Beijing City, as 'Haidian' is a district of Beijing.
- 5) If the address field does not contain any description of the city/county/district name, we try to extract the city information from other fields such as the registration authority. In the end, the corresponding cities or countries could not be found for approximately 3% of the total records. The step of extracting city names can be processed automatically by some Python packages, e.g., the 'cpca' package[31].
- 6) After having the city and address information, we geocode them with AMap API[32], and obtain the coordinate information (longitude, latitude). Because the registered addresses will be checked by AIC, it is unlikely that there would be typo issues. AMap also automatically handles typo issues according to the API document[32].
- 7) Finally, we convert all geocoded coordinates from the GCJ02 projection, a special projection system in China[33], into the world geodetic system 1984 (WGS84) projection, and then aggregate the coordinates according to a 0.01° latitude by 0.01° longitude grid.

Variable name	Description
idx	Index of the grid cell
lon	Longitude of the center of the grid cell (WGS84 projection)
lat	Latitude of the the center of the grid cell (WGS84 projection)
date	Year of the establish record
establish_total	Total number of establishments within the cell
- establish_prim	- Number of newly established firms in primary industry
- establish_seco	- Number of newly established firms in secondary industry
- establish_tert	- Number of newly established firms in tertiary industry
exist_total	Number of existing firms
- exist_prim	- Number of existing firms in primary industry
- exist_seco	- Number of existing firms in secondary industry
- exist_tert	- Number of existing firms in tertiary industry
cancel_total	Number of cancelled firms
- cancel_prim	- Number of cancelled firms in primary industry
- cancel_seco	- Number of cancelled firms in secondary industry
- cancel_tert	- Number of cancelled firms in tertiary industry
other_total	Number of establishments with unknown operation status
county	Name of the county-level division where the grid cell is located
city	Name of the prefecture city where the grid cell is located
province	Name of the province where the grid cell is located

Table 2. List of variables.

As shown in Fig. 2b, we find that the distributions of the grid-cell values are well fitted by a log-normal function, which is in line with most socioeconomic characteristics[34]. Meanwhile, we plot the map of newly established firms in 2015 (Fig. 3). As seen from the map, our dataset clearly shows the geographic pattern of economic activity in China – the vast majority of economic activity is concentrated in the eastern coastal areas, especially the three megacity clusters: Beijing-Tianjin-Hebei, the Yangtze River Delta, and the Pearl River Delta. Unlike the NTL data, the establishment dataset does not have the saturation effect, so it can easily capture variations within cities (Fig. 3b,c).

Data Records

Data are available at the figshare repository[35]. The repository contains four data files, and the variable descriptions of the main file ('GED_2005–2015_v2.csv') are reported in Table 2. Each row of the file represents the record of one grid cell, including the centroid longitude and latitude, year of the record, total number of establishments, number of establishments in different operational statuses, and number of establishments in different industries (Table 2). It should be noted that for a cancelled firm, the date variable here represents the year when the firm was registered, not the year when it was cancelled.

To reduce the file size, for a given year if a grid cell does not have any registration record, this grid cell is not included in the dataset. For ease of use, we also link the coordinates of cells with the county, city and province information. The basic statistical summary at the province level is shown in Table 3. Guangdong, Jiangsu, Shangdong, Zhejiang and Shanghai rank among the top five in terms of number of establishments. The reason why Shandong ranks high is because of its large population, and the other four provinces are recognized as the regions with the strongest entrepreneurial spirit in China.

We also provide three-year – 2005, 2010, and 2015 – unaggregated data to facilitate researchers to conduct research on different scales. For the data description of the unaggregated data, see the 'read-me' file under the figshare repository path.

Technical Validation

To validate the dataset, we analyze the accuracy of geocoding and the correlation between our dataset and other socioeconomic indicators.

Validation of geocoding. According to the city size and geographical location, we select three representative cities and randomly sample 10,000 records for analysis. The three cities are Beijing (large city), Changsha (medium-sized city), and Anshan (small city). For geocoding accuracy, we use two methods to evaluate. The first is the spatial resolution returned by the Amap API. This is the geographical level that the map can match according to the input address. The finer the geographical resolution, the higher the accuracy may be. As shown in Table 4, the finest match is to the street number, followed by the point of interest (POI). Any address that can match these two levels is generally of high accuracy, and we find that more addresses are matched at these two levels in large cities than in medium and small cities: 88% of Beijing compared with 77% of Changsha and 48% of Anshan. Moreover, if we consider two more levels, to the level of villages and towns, then the geocoding accuracy of these three cities is 96%, 95% and 84%, respectively.

Province	N_{total}	Min	Max	Mean	SD
Anhui	812,773	1	1,275	6.73	20.23
Beijing	1,356,979	1	4,758	30.48	104.11
Chongqing	651,720	1	1,409	9.30	34.55
Fujian	809,078	1	2,526	8.97	34.67
Gansu	304,597	1	1,718	5.63	19.94
Guangdong	3,004,904	1	41,232	18.45	134.45
Guangxi	629,717	1	1,642	8.39	30.40
Guizhou	551,186	1	2,000	7.28	31.97
Hainan	184,424	1	1,714	11.90	49.48
Hebei	1,017,968	1	938	5.27	16.36
Heilongjiang	460,459	1	782	6.28	17.07
Henan	1,101,598	1	1,172	5.72	20.22
Hubei	952,163	1	1,261	7.65	27.60
Hunan	635,266	1	860	5.54	19.51
Innter Mongolia	410,483	1	474	6.31	15.15
Jiangsu	2,635,874	1	46,388	10.49	104.50
Jiangxi	547,476	1	950	6.25	17.78
Jilin	422,966	1	11,412	7.11	62.65
Liaoning	726,469	1	685	7.15	20.37
Ningxia	127,697	1	643	7.58	21.55
Qinghai	83,245	1	450	5.28	15.18
Shaanxi	79,774	1	992	2.46	11.70
Shandong	1,840,092	1	1,238	6.21	19.90
Shanghai	1,607,515	1	6,724	40.72	142.14
Shanxi	515,799	1	819	4.29	13.51
Sichuan	1,044,063	1	3,383	7.06	36.84
Tianjin	392,245	1	3,500	12.97	45.23
Tibet	37,346	1	489	5.29	16.63
Xinjiang	306,840	1	650	7.01	20.04
Yunnan	519,001	1	1,394	7.43	26.84
Zhejiang	1,734,792	1	2,109	8.46	25.01

Table 3. Statistical summary of the establishments (2005–2015). N_{total} is the total number of establishments within the province. *Min*, *Max*, *Mean*, and *SD* (standard deviation) are the statistical indicators at the cell level, i.e., each cell is one observation.

Spatial resolution	Beijing	Changsha	Anshan
Street number	7,108	4,673	3,475
POI	1,665	3,005	1,276
Village	537	1,387	2,500
Town	313	431	1,128
Road	249	347	962
Other	128	157	659

Table 4. Geocoding resolution.

Another verification method is to use the 'confidence indicator' returned by Baidu Maps. (Amap and Google Maps do not provide this measure, but considering Amap and Baidu Maps provide similar services in China, therefore we use the 'confidence indicator' obtained from Baidu Maps as a benchmark to evaluate the quality of the address attribute of firms.) According to the document of Baidu Map[36], geocoding confidence ≥ 50 means the absolute geocoding error is less than 1000 m. We use Baidu Maps to re-geocode these sampled addresses, the results show that the proportions of coordinate errors less than 1000 m (i.e. confidence ≥ 50) in the three cities are 93.1%, 92.2% and 74.1%, respectively, implying that the accuracy of the geocoding services is significantly higher in large cities than in small cities.

Validation with city socioeconomic indicators.   To further justify the validity of the data, we aggregate the number of newly established firms at the city level, and compare it with the city's socioeconomic indicators: GDP, fiscal revenue, and employment. These indicators are collected from the city statistical yearbook, and some cities are excluded due to a lack of statistical data from the yearbook. We present the results of two fitting models
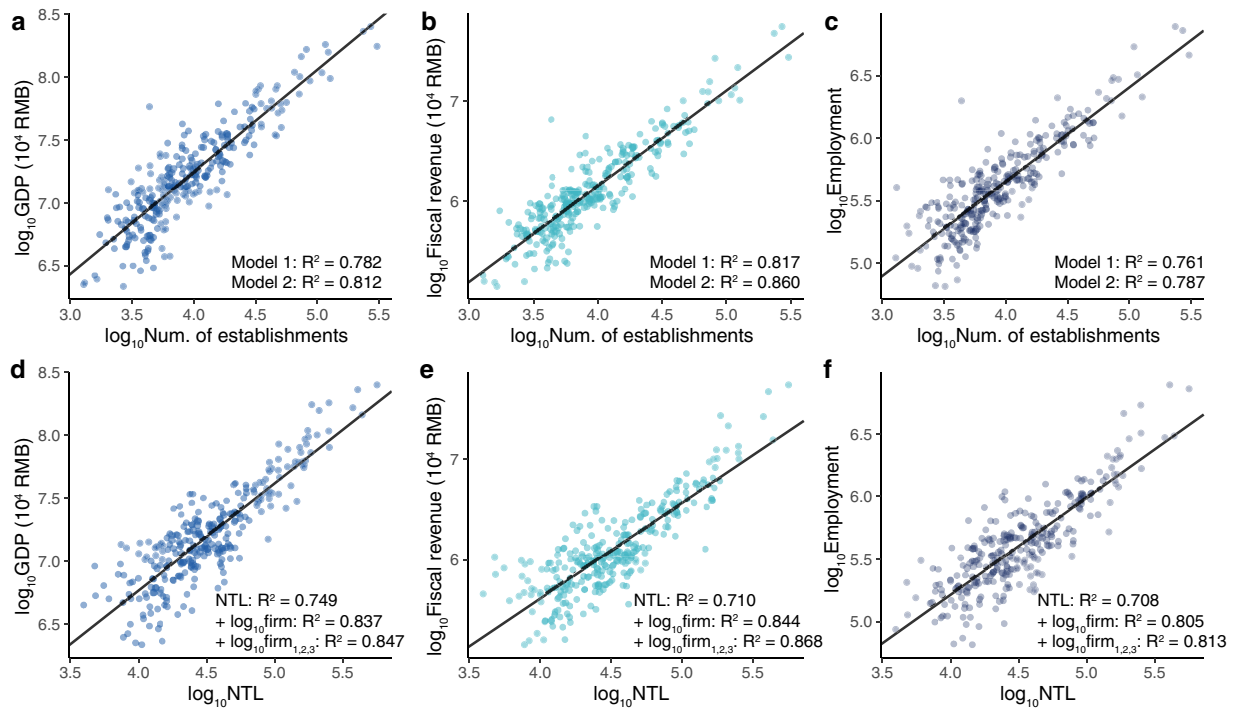
Fig. 4 Validation with socioeconomic indicators at city level. (**a**–**c**) Correlation between number of establishments and (**a**) GDP, (**b**) fiscal revenue, and (**c**) employment. (**d**–**f**) Similar to (**a-c**) but replacing establishments with the nighttime light data. We also add $\log_{10}$ NTL in Eqs. (1) and (2), and present the regression results in (**d**–**f**).

in Fig. 4. One is a linear regression between the total number of establishments and the variables of interest y, denoted as Model 1:

$$\log_{10}(y) = \beta_0 + \beta_1 \log_{10}(\text{Est}) + \in \tag{1}$$

The second model decomposes the number of establishments into three sectors, i.e., the primary, secondary, and tertiary sectors. Model 2 is as follows:

$$\log_{10}(y) = \beta_0 + \beta_1 \log_{10}(\text{Est}_1) + \beta_2 \log_{10}(\text{Est}_2) + \beta_3 \log_{10}(\text{Est}_3) + \in \tag{2}$$

To make a comparison with the nighttime light data, we also collect the Visible Infrared Imaging Radiometer Suite (VIIRS) NTL data for 2015 (available through www.ngdc.noaa.gov/eog/viirs). We use the annual 'vcm-orm-ntl' average radiance layer, which has undergone an outlier removal process (exclusion of data impacted by stray light, lightning, cloud cover, etc.). Similar to the establishment data, we aggregate the values of nighttime light data at the city level. To make the results comparable, we used the 2015 data of each dataset for comparison.

Figure 4 shows the fitting results (adjusted R^2) of GDP, fiscal revenue, and employment at the city level. In general, both our data and the nighttime light data perform well in estimating these socioeconomic variables, with adjusted R^2 values over 0.70 in all cases. However, the accuracy of the establishment data is better than that of the the nighttime light data in predicting all three variables (Model 1), especially for fiscal revenue (0.817 vs. 0.710) and employment (0.761 vs. 0.708), and decomposing establishments into three sectors, i.e., Model 2, further improves the adjusted R^2 (Fig. 4a–c). Additionally, we add $\log_{10}$ NTL in Eqs.(1) and (2) to test how much additional variance can be explained by the combined model (Fig. 4d–f). Results demonstrate that the combined version of Eq. (2) achieves the highest predicting accuracy, implying that the establishment data and NTL data respectively capture some elements of the socioeconomic indicators, thus they can be jointly used in practice.

Finally, we validate our dataset by linking it with a granular population dataset to verify the scaling law of cities. As demonstrated in numerous studies[37,38], urban socioeconomic activity Y scales with population P in a simple power-law manner: $Y \sim P^\alpha$, where α is called the scaling exponent (or the elasticity in the term of economics). To test the power-law relationship within cities, we collect a mobile phone estimated population dataset from ref. [39]. We match the population with the establishment data under the same spatial resolution (0.01° latitude by 0.01° longitude, see Fig. 5), and then run a simple regression to estimate the scaling coefficient α of Beijing. According to the assumption of the scaling theory, we constrain the analysis within the urbanized area by setting a population density threshold of 1,000 people/km², and drop cells with population density lower than this threshold. Result shows that the power-law relationship between population and establishments holds well within cities, and the scaling exponent $\alpha = 1.16$ is greater than 1, which is very close to the theoretical numbers derived from different models[40].
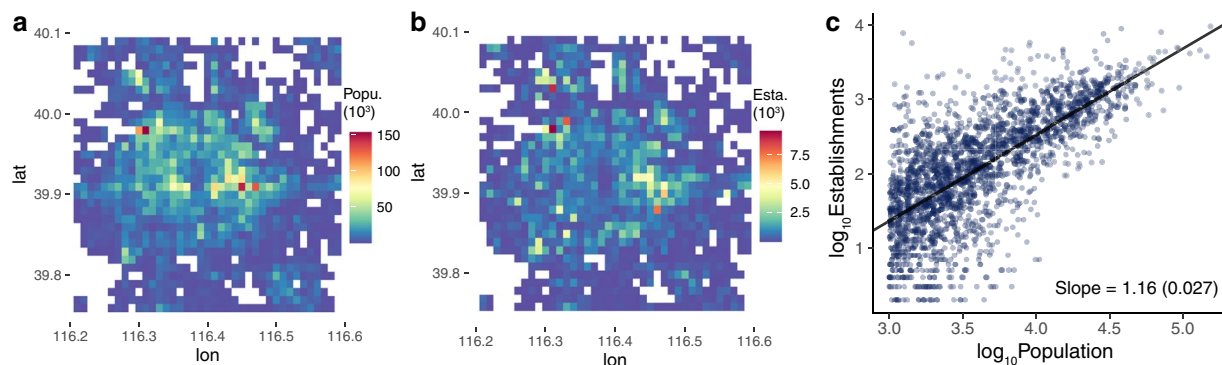
Fig. 5 Validation with urban scaling theory. (**a,b**) Geographical distributions of (**a**) population and (**b**) establishments. (**c**) Power-law relationship between population and establishments. The exponent $\alpha = 1.16$ is greater than 1.

Usage Notes

The GED can be used in geographic information systems (GIS), such as ArcGIS and QGIS, as well as statistical software, such as Python, R, MATLAB, and Stata. In GIS software, the dataset can be imported as a vector layer or a raster layer. To match with other geographical datasets, such as industrial park boundaries[41] or water/air quality monitoring records[42], users can apply the spatial join function in GIS software by joining attributes from GED to another data based on the spatial relationship[43]. For example, we can draw a buffer for each industrial park, and then extract the grid cells and the corresponding establishments that fall within this buffer. If the resolution of the GED is inconsistent with other data sources in use, some resampling methods used in remote sensing areas could also be applied to this dataset[44].

Additionally, the county/city/province name provided in this dataset is based on the administrative boundary of 2019. Therefore, if the GED is matched with other statistical data by couty/city/province name, the impact of changes in the name (and the corresponding spatial range) should be considered. It's also notable that the grid cell index of GED is under WGS84 coordinates. Thus, the spherical area of grid cells in different regions is different (approximately 1.1 km × 1.1 km at the equator and 0.85 km × 1.1 km at Beijing).

Finally, as mentioned in the introduction section, due to some location-based policies, there will be a large number of registered enterprises in some specific areas, such as the Shanghai Free Trade Zone. The data points on the far right of Fig. 2c indicate that there are more than 10,000 establishments in each of these grid cells. Such spatial 'outliers' would have a certain impact on some micro-scale analyses. To address this issue, we suggest taking an upper limit on the value in each cell or using a logarithmic transformation of the data when performing micro-scale analyses.

Code availability

The preprocess script, validation dataset and the R code that performs the statistical analysis are available through https://github.com/quanturban/firm.

References

1. Storper, M. Why do regions develop and change? The challenge for geography and economics. *J. Econ. Geogr.* **11**, 333–346 (2011).
2. Nordhaus, W. D. Geography and macroeconomics: New data and new findings. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 3510–3517 (2006).
3. Census Bureau. County business patterns. https://www.census.gov/programs-surveys/cbp.html (2018).
4. Mellander, C., Lobo, J., Stolarick, K. & Matheson, Z. Night-time light data: A good proxy measure for economic activity? *PLoS One* **10**, e0139779 (2015).
5. Rawski, T. G. What is happening to China's GDP statistics? *China Econ. Rev.* **12**, 347–354 (2001).
6. Einav, L. & Levin, J. Economics in the age of big data. *Science* **346**, 1243089 (2014).
7. Chen, X. & Nordhaus, W. D. Using luminosity data as a proxy for economic statistics. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 8589–8594 (2011).
8. Henderson, J. V., Storeygard, A. & Weil, D. N. Measuring economic growth from outer space. *Am. Econ. Rev.* **102**, 994–1028 (2012).
9. Naik, N., Kominers, S. D., Raskar, R., Glaeser, E. L. & Hidalgo, C. A. Computer vision uncovers predictors of physical urban change. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 7571–7576 (2017).
10. Glaeser, E. L., Kominers, S. D., Luca, M. & Naik, N. Big data and big cities: The promises and limitations of improved measures of urban life. *Econ. Inq.* **56**, 114–137 (2018).
11. Ma, R., Wang, W., Zhang, F., Shim, K. & Ratti, C. Typeface reveals spatial economical patterns. *Sci. Rep.* **9**, 15946 (2019).
12. Blumenstock, J., Cadamuro, G. & On, R. Predicting poverty and wealth from mobile phone metadata. *Science* **350**, 1073–1076 (2015).
13. Dong, L. *et al.* Measuring economic activity in China with mobile big data. *EPJ Data Sci.* **6**, 29 (2017).
14. Llorente, A., Garcia-Herranz, M., Cebrian, M. & Moro, E. Social media fingerprints of unemployment. *PLoS One* **10**, e0128692 (2015).
15. Glaeser, E. L., Kim, H. & Luca, M. Nowcasting the local economy: Using yelp data to measure economic activity. Working Paper 24010, National Bureau of Economic Research https://doi.org/10.3386/w24010 (2017).
16. Dong, L., Ratti, C. & Zheng, S. Predicting neighborhoods' socioeconomic attributes using restaurant data. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 15447–15452 (2019).

17. Lazer, D., Kennedy, R., King, G. & Vespignani, A. The parable of Google flu: Traps in big data analysis. *Science* **343**, 1203–1205 (2014).
18. Panczak, R., Charles-Edwards, E. & Corcoran, J. Estimating temporary populations: A systematic review of the empirical literature. *Palgrave Commun.* **6**, 1–10 (2020).
19. NOAA. Version 4 DMSP-OLS nighttime lights time series. https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html (2019).
20. Gaughan, A. E. *et al.* Spatiotemporal patterns of population in mainland China, 1990 to 2010. *Sci. Data* **3**, 160005 (2016).
21. Donaldson, D. & Storeygard, A. The view from above: Applications of satellite data in economics. *J. Econ. Perspect.* **30**, 171–198 (2016).
22. Toole, J. L. *et al.* Tracking employment shocks using mobile phone data. *J. Roy. Soc. Interface* **12**, 20150185 (2015).
23. Almaatouq, A., Prieto-Castrillo, F. & Pentland, A. Mobile communication signatures of unemployment. In *International Conference on Social Informatics*, 407–418 (Springer, 2016).
24. e Silva, F. B. *et al.* Uncovering temporal changes in Europe's population density patterns using a data fusion approach. *Nat. Commun.* **11**, 1–11 (2020).
25. Grantz, K. H. *et al.* The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. *Nat. Commun.* **11**, 1–8 (2020).
26. Jean, N. *et al.* Combining satellite imagery and machine learning to predict poverty. *Science* **353**, 790–794 (2016).
27. Tianyancha. Business directory. https://top.tianyancha.com/companies/ (2019).
28. The State Administration for Industry and Commerce. Analysis on the development of national enterprises since the 18th CPC National Congress. http://www.gov.cn/zhuanti/2017-10/27/content_5234848.htm (2017).
29. De Rassenfosse, G., Kozak, J. & Seliger, F. Geocoding of worldwide patent data. *Sci. Data* **6**, 1–15 (2019).
30. The Central People's Government of the People's Republic of China. Administrative divisions of the People's Republic of China. http://www.gov.cn/test/2005-06/15/content_18253.htm (2019).
31. DQinYuan. Chinese province city area mapper. https://github.com/DQinYuan/chinese_province_city_area_mapper (2019).
32. Amap Open Platform. Geocoding. https://lbs.amap.com/api/webservice/guide/api/georegeo (2019).
33. wandergis. Coordinate transform. https://github.com/wandergis/coordtransform (2019).
34. Limpert, E., Stahel, W. A. & Abbt, M. Log-normal distributions across the sciences: Keys and clues. *BioScience* **51**, 341–352 (2001).
35. Dong, L., Yuan, X., Li, M., Ratti, C. & Liu, Y. A gridded establishment dataset as a proxy for economic activity in China. *figshare* https://doi.org/10.6084/m9.figshare.c.5000618 (2021).
36. Baidu Maps. Geocoding. https://lbsyun.baidu.com/index.php?title=webapi/guide/webservice-geocoding (2019).
37. Bettencourt, L. M., Lobo, J., Helbing, D., Kühnert, C. & West, G. B. Growth, innovation, scaling, and the pace of life in cities. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7301–7306 (2007).
38. Li, R. *et al.* Simple spatial scaling rules behind complex cities. *Nat. Commun.* **8**, 1841 (2017).
39. Dong, L., Huang, Z., Zhang, J. & Liu, Y. Understanding the mesoscopic scaling patterns within cities. *Sci. Rep.* **10**, 21201 (2020).
40. Bettencourt, L. M. The origins of scaling in cities. *Science* **340**, 1438–1441 (2013).
41. Zheng, S., Sun, W., Wu, J. & Kahn, M. E. The birth of edge cities in China: Measuring the effects of industrial parks policy. *J. Urban Econ.* **100**, 80–103 (2017).
42. He, G., Wang, S. & Zhang, B. Watering down environmental regulation in China. *Q. J. Econ.* **135**, 2135–2185 (2020).
43. ArcGIS. Spatial join. https://desktop.arcgis.com/en/arcmap/10.3/tools/analysis-toolbox/spatial-join.htm (2020).
44. Lyons, M. B., Keith, D. A., Phinn, S. R., Mason, T. J. & Elith, J. A comparison of resampling methods for remote sensing classification and accuracy assessment. *Remote Sens. Environ.* **208**, 145–153 (2018).
45. Gillespie, C. Fitting heavy tailed distributions: The powerlaw package. *J. Stat. Softw.* **64** (2015).

Acknowledgements

Author contributions

L.D., X.Y., C.R. and Y.L. designed research; L.D. performed research; L.D. and M.L. analyzed data; L.D. wrote the paper. All authors reviewed the paper.

Competing interests

L.D. and X.Y. are co-founders of QuantUrban, a company that uses big data and machine learning to measure socioeconomic dynamics of cities.

Additional information

Correspondence and requests for materials should be addressed to L.D. or Y.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.