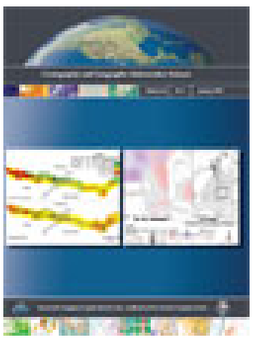

Taylor & Francis
Taylor & Francis Group

Cartography and Geographic Information Science

ISSN: 1523-0406 (Print) 1545-0465 (Online) Journal homepage: http://www.tandfonline.com/loi/tcag20

SinoGrids: a practice for open urban data in China

Yulun Zhou & Ying Long

To cite this article: Yulun Zhou & Ying Long (2016): SinoGrids: a practice for open urban data in China, Cartography and Geographic Information Science, DOI: 10.1080/15230406.2015.1129914

To link to this article: http://dx.doi.org/10.1080/15230406.2015.1129914

Published online: 13 Jan 2016.

CARTOGRAPHY AND GEOGRAPHIC INFORMATION SCIENCE, 2016
http://dx.doi.org/10.1080/15230406.2015.1129914

Taylor & Francis
Taylor & Francis Group

SPECIAL ISSUE ARTICLE

SinoGrids: a practice for open urban data in China

Yulun Zhou[a] and Ying Long[b,c]

[a]Department of Geography and Resource Management, The Chinese University of Hong Kong, Hong Kong, China; [b]School of Architecture, Tsinghua University, Beijing, China; [c]Beijing Institute of City Planning, Beijing, China

ABSTRACT

In the past decade, an explosion of data has taken place in Chinese cities due to widespread use of mobile Internet devices, Web 2.0 applications, and the development of the "Wired City." With advances in data storage and high-performance computing, big/open urban data have opened up important avenues for urban studies, planning practice, and commercial consultancy. Urban researchers and planners are eager to make use of these abundant, sophisticated, and dynamic data to deepen their understanding on urban form and functions. However, in practice, access to such urban data is limited in China due to institutional constraints on data distribution and data holders' hesitation to share data. And this hampers urban analytics. To draw reliable conclusions about the workings of complex urban systems, efficient and effective interoperation of multi-source urban datasets is needed. Also, dealing with the heterogeneity between datasets is an equally critical challenge, especially for urban planners and government officers. They would derive value from data analytics, but have little data processing experience. To address these issues, we initiated SinoGrids (Plan Xu Xiake), a crowdsourcing platform that standardizes (or "downscales") microscale urban data in China to facilitate its sharing and interoperation. To assess the performance evaluation of SinoGrids, we propose field-testing with actual urban data and their potential users. Digital desert, a son project of SinoGrids is also included.

ARTICLE HISTORY
Received 22 May 2015
Accepted 7 December 2015

KEYWORDS
Open data; crowdsourcing; urban analytics; citizen science; China

1. Introduction

Urbanization and industrialization are taking place worldwide at accelerating rates, prompting urban researchers, planners, and commercial consultants to keep up and deepen their understanding of urban form and functions. In China, public concerns on various urban challenges, for example, air pollution, have grown to high, even overwhelming levels (Enserink and Koppenjan 2007), strongly encouraging related urban studies. At the same time, an explosion of data has taken place in cities. Urban big/open data have opened up important opportunities for urban-related researchers to develop more sophisticated, large-scale, and dynamic analytic methods to understand urban issues. For example, urban data analytics have been proposed for house prices (Huang, Wu, and Barry 2010), mobile phone use (Tranos and Nijkamp 2015), and accessibility to health care (Aoun, Matsuda, and Sekiyama 2015). For purposes of urban and regional planning, using bus smart card data and points of interest in Beijing, Han, Yu, and Long (2015) discovered functional zones, and Long analyzed jobs–housing relationships (Long and Thill 2015) and profiled underprivileged residents (Long et al. 2014) based on open/big urban data.

All the aforementioned urban analyzing practices are empowered by new urban data. Unlike conventional urban data, new urban data are larger in size, finer-grained, more sophisticated, more dynamic, and closely involve urban residents' behaviors. In the past decade, the amount of new urban data has boomed, for reasons such as the following:

1. Development of "Wired Cities" with governments installing digital sensors everywhere in cities, for monitoring, managing, and regulating urban flows;
2. Rapid spread of mobile Internet technologies;
3. The popularity of social media and other Web 2.0 applications;
4. Accelerating development of data storage and distributed computing techniques (Kitchin 2014).

However, as a matter of fact, access to much of these urban data has been limited. Beyond accessibility, Gurstein (2011) noted there is a growing gap between enhancing citizens' "access" to data and enhancing

CONTACT Ying Long longying1980@gmail.com

their "usage" of it. When it comes to urban systems, effective data use is complex and always involves interactions between disciplines. This complexity gives abundant, high-quality, accessible, and usable data extra significance for researchers wishing to capture varied aspects of urban form and functions for urban and regional studies. For example, urban air pollution studies often model interactions between urban morphology, climatology, urban land use, etc., and thus, a wide range of urban data from various data sources are needed to draw reliable conclusions. In such a context, open urban data should not only be accessible but also usable, reusable, and redistributable if they are to benefit urban data applications (Wilbanks 2014).

The open data movement has been initiated with an overall intention to make local, regional, and national data, particularly publicly acquired data, available in a form that allows for direct manipulation, for example, cross-tabulation, visualization, etc. (Gurstein 2011). Typical successful practices include volunteered datasets, for example, Open Street Map (Haklay and Weber 2008), which provide information about urban form and function, with the providers themselves determining urban morphology through their own activity (Crooks et al. 2014). But currently, open data still face various critical issues with regard to data dispersion, heterogeneity, and provenance (Gurstein 2011; Overpeck et al. 2011; Reichman, Jones, and Schildhauer 2011).

China has been undergoing an unprecedented, vast, and rapid process of urbanization and industrialization (Bai, Shi, and Liu 2014). As a result, urban studies in China have even higher requirements for urban data in terms of spatial-temporal extent and scale, and data and metadata provisioning. The purpose of this paper, focusing on open data challenges for regional and urban studies in China, is twofold. First, it briefly reviews the status of open urban data in China and generally, together with related initial practices. Then, it proposes a new approach to a crowdsourcing platform for providing more sharable and interoperable microscale urban data. This approach, called SinoGrids (Plan Xu Xiake, in Chinese, online address: http://www.beijingcitylab.com/projects-1/14-sinogrids/), can empower urban and regional studies in China. It standardizes urban data using a uniform grid base so as to minimize conflict between original data holders and data users and, in so doing, bridge the gap between "access" and "effective use" of urban data. The Chinese name for SinoGrids, Xu Xiake Plan honors Xu Xiake, a famous Chinese geographer and travel writer of the Ming dynasty (1368–1644), who spent 30 years in traveling all around China and documented his travels. By naming SinoGrids after Xu Xiake, we hope to encourage sharing crowdsourced urban data all across China, regardless of location and platform. In the long run, we expect SinoGrids to help data holders in China develop the habit of data sharing and empowering each other.

2. Open urban data in the world and China

The definition of urban data we are using, as the context of discussion in this paper, is quite broad. Urban data refer to all datasets that can characterize and facilitate interpretation of aspects of urban form and functions (Crooks et al. 2014). Open urban data, subsequently, are urban data that are openly accessible and effectively usable to researchers, planners, commercial consultants, local residents, etc.

Traditional urban datasets are mainly based on government efforts, for example, national censuses and cadastral maps. Also included in urban data are administrative records, such as approvals of construction permits from planning departments, economic development reports from statistics departments, etc. As many of these datasets are based on regional statistics generated through sampling, urban studies based on them suffer from limitations of spatial-temporal scale and geographical coverage. Furthermore, those conventional government datasets that are stored as paper records can require huge digitization efforts to be integrated into detailed, large-scale, and comprehensive urban studies. However, recent open/big data efforts have partially changed the conventional situation.

Nowadays, a variety of portals for open/big urban data are coming online. Some of them are official data portals, enabled by recent open government initiatives that open previously nonaccessible data sources to the general public. Others are community-generated big data initiatives, collecting data from mobile phone activities, vehicle trajectories, public transit smart card data, business catalogs, and other smart city initiatives (Batty 2012).

2.1. Government-generated urban data

Government-run online data portals are mushrooming, indicating a rise in the official awareness and willingness of promoting social services and their transparency, and empowering urban studies by opening data. For example, Local Law 11 approved by New York City Council in 2012, which requires all agencies to open their data. The city-funded NYC OpenData

(https://nycopendata.socrata.com/) provided about 1300 datasets as of July 2013, and further plans to release 345 more datasets before 2018. The included datasets cover a range of aspects of urban issues, for example, public safety, city government, education, health care, and so on. All datasets are in machine-readable formats, paired with corresponding metadata. According to the US City Open Data Census (Open Knowledge 2014), dozens of American cities have created municipal data portals. In terms of data volume, New York City ranked first in 2014, followed by San Francisco (https://data.sfgov.org/), Los Angeles (https://data.lacity.org/), and Boston (https://data.cityofboston.gov/). At the federal level, the US General Services Administration has established an open data platform, *Data.gov*, providing over 150,000 online datasets gathered from hundreds of organizations including many Federal agencies. In Europe, the European Commission has been leading the project *INSPIRE* (http://inspire.ec.europa.eu/), which is intended to build up a European spatial data infrastructure, and has published spatial data from over 700 data communities. Detailed technical guidelines of data specifications for spatial elements, for example, addresses, coordinate reference systems, etc., together with specifications on different categories of data, have been proposed to make these datasets as manageable, usable, and interoperable as possible.

As China is a developing country, like many others, open data in China face pressure and limitations from tight regulations. Despite the current situation, the Chinese government has devoted efforts for easier data access and more relaxed data control. A national data portal (http://data.stats.gov.cn/) has been initiated by the National Bureau of Statistics of China, providing digitized census datasets and monthly, seasonal, and annual statistical reports, as well as some data visualization products. Beijing and Shanghai, the two metropolises in China, are the first to have established open urban data platforms (Beijing: http://www.bjdata.gov.cn/; Shanghai: http://www.datashanghai.gov.cn/), respectively, publishing over 400 and 209 urban datasets from various government departments. Wuhan is also expecting the opening of its "one-cloud, one-map, one-standard, one-model, one-stop" open data platform with 520 datasets, claiming that all government data in Wuhan will be accessible from this platform in the near future. Other cities are joining the open data movement, for example, Qingdao, Guiyang, Guangzhou, etc. The opening of local government data has become a trend in China (Guo 2014). Based on government data platforms introduced to date, the main motivations and intentions of the opening of government data are:

(1) Urban big/open data have been regarded as an effort that supports the human-oriented "new-type urbanization" in China.
(2) Urban big/open data are considered as a signature for the improvement of the accessibility and transparency of government, fulfilling the people's right to know. Open data movement also helps convey a clean, efficient, and open-minded image of the government.
(3) The governments are aware that, as the largest holder of public data resources, the best way to make the most out of these data is not by monopolizing data but by utilizing and sharing it.

China's government-driven open data platforms are comparable to foreign ones. For example, *Data.gov* and *INSPIRE*, both of them are government-driven open spatial data infrastructures with detailed data-sharing legislation, data specification, etc. As government-driven projects, the main data sources of *Data.gov* (USA) and *INSPIRE* (EU) are official agencies and data communities. China's national spatial data infrastructure (NSDI) has developed considerably since the 1980s, and China's NSDI has played significant roles in the nation's economic construction and social development (Chen and Chen 2003). But still, the openness of China's NSDI needs improvement, compared with the aforementioned foreign practices. According to administrative regulations on licenses for using national fundamental geographic data promulgated by State Bureau of Surveying and Mapping, in most cases, data from NSDI are freely accessible only to first-tier users, that is, central government agencies and provincial governments using it for macro decision-making and social welfare. Noncommercial organizations and individuals, as second-tier users, do not enjoy free access to the data from NSDI (Yang, Chen, and Wu 2001), indicating that the promotion of geosocial data sharing in China needs further efforts (Chen and Chen 2003). For current government-open portals in China, it is found that most datasets are still in a tabular format, which needs further preprocessing to use for data analytics. Most datasets are still based on traditional regional statistics, whose limited spatial and temporal scales constrain their applications in regional and urban studies. Also, the number of currently opened datasets is still limited and is not enough to fulfill urban analytics needs.

Government-funded research institutes have been taking a leading role in opening data in China. *The*

Geospatial Data Cloud (http://www.gscloud.cn/), established by the Chinese Academy of Sciences, is an open data platform for spatial data, for example, remote sensing images and retrieval products. Also, some preprocessing services, such as atmospheric correction, image gap-filling, etc., are provided online, in order to make the datasets more useful for application. Also, the Chinese Academy of Sciences *Data Center for Resources and Environmental Sciences* (http://www.resdc.cn/) effort is intended to empower studies on sustainable resource and environment in China. Its datasets, mostly about physical geography, focusing on vegetation, land, terrain, etc., are freely accessible for researchers (an official request letter is needed).

2.2. Community-generated urban data

Government-generated data developed for administrative purposes often fail to capture the characteristics of urban form and functions based on the public's perception (Crooks et al. 2014). In recent years, community-generated urban data, including social media, volunteered datasets, etc., have arisen and opened up new research opportunities in urban studies and planning (Batty 2013). One typical example is the *Open Street Map* (OSM, http://www.openstreetmap.org/), which was built by a community of mappers that contribute and maintain data about road networks. As a platform, *Open Street Map* has the following characteristics typical of community-generated open data initiatives:

1. Community driven. The open data portal is maintained by a community of mappers based on local knowledge, which also reflects a public perspective of urban space and activity.
2. Explicit. The data from *Open Street Map* are in geographical information system (GIS)-based format and are compatible with ESRI ArcMap, the most widely used platform for geographic information processing, making the data highly usable and interoperable.
3. Usable, reusable, and redistributable. Data from *Open Street Map* are freely usable as long as the user credits *Open Street Map* as its contributor. The data and results generated may be distributed under certain copyright and license terms.

In terms of coverage, Open Street Map has also covered many places in China, but due to certain limitations, data for middle- and small-sized Chinese cities tend to have lower precision and granularity. Despite that, it has still become one of the best ways of accessing basic urban geospatial infrastructure data in China.

Locally, leading map service providers, such as *Baidu Map* (http://developer.baidu.com/map/) and *Amap* (http://lbs.amap.com/), provide open data services through the application programming interface (API), empowering planners and commercial companies with cloud services. Big/open data initiatives in China have also been trying to face these challenges. *Datatang* (http://www.datatang.com/) was established as the first data trading platform in China to resolve conflicts between the original data holders and data users in a beneficial way for both. *Datatang* held over 44,000 datasets online as of May 2015, covering a wide range of functions, for example, semantic analysis, transportation, health care, etc. But most of its datasets are for sale, and only a small proportion is freely accessible and usable. Last but not least, *Beijing City Lab* (http://www.beijingcitylab.com/), a research community focused on urban topics in China, has opened 28 datasets characterizing urban China, all of which are freely accessible and in formats that support effective reuse. All the datasets are from open datasets online, supplemented by donations from researchers both in and out of the community.

Social media is a new kind of community-generated urban datasets but has already become a significant resource. Currently, the most popular microblogging service in the world is Twitter with over 284 million users as of December 2014. Twitter messages are mostly in English, and it is not freely accessible in Mainland China. In China, as alternatives, similar microblogging services are available, such as Sina Weibo, Tencent Weibo, etc. By September 2014, the number of monthly active users (MAU) of Sina Weibo reached 167 million, with an annual growth rate of 36%. Promoting the spirit of Web 2.0, which is to encourage user-generated content (UGC), microblogging has been an indispensable part of urban life and a major way of expressing personal feelings and opinions in China. Various studies have been proposed for analyzing social media data for urban activities based on Twitter. Tumasjan et al. (2010) tracked public opinion by monitoring political sentiment expressed via social media and predicted election results. Propagation patterns of news have been analyzed (Lerman and Ghosh 2010), and social media data have even been used to predict earthquakes (Sakaki, Okazaki, and Matsuo 2010) and stock market performance (Si et al. 2013). More and more studies have emerged using Weibo for detecting the pulse of city life in China.

2.3. Challenges in open urban data in China

Open data still face many challenges. Challenges for open data have been reviewed in the field of ecology (Reichman, Jones, and Schildhauer 2011) and climate (Overpeck et al. 2011). Among those noted, challenges of data dispersion, heterogeneity, and provenance also apply to urban open data (Liu et al. 2015).

Urban data are generated from various data sources, for example, government-generated data, volunteered datasets, UGC, etc. Data from different sources may have overlaps, and each characterizes a certain aspect of urban form and functions. Variations in spatial and temporal coverage and scale, data formats, etc. make data interoperation difficult. Also, for implicit data, such as paper-based maps and other documents, extra efforts of digitization and preprocessing are needed.

One more important challenge that is unique to open data in China—rapid and unprecedented changes in urban morphology and behaviors due to rapid urbanization and industrialization process in China—was mentioned by Liu et al. (2015). Characterizing frequently changing patterns imposes higher requirements on spatial-temporal scale and coverage of urban data to make datasets suitable to support more sophisticated urban studies.

In China, as government-dominated open data portals are at their early stage, urban data are mainly shared through community-driven efforts. However, data holders of community-driven open data often hesitate to share it. According to our survey, 79.75% of the surveyed individuals claim that they "often" find themselves lacking proper data. But only 22.14% of all individuals are willing to share data with other data users who might need it. The majority of people surveyed would selectively share datasets with their partners, colleagues, or students. Not surprisingly, 8.23% of the surveyed would rather not share data with others. The main concerns of data holders are losing research advantages, not receiving credit for sharing, and being limited by data distribution policies of the institution or data provider. Clearly, such hesitations of data holders have been a serious bottleneck for improving the data opening situation in China.

How to balance the benefits of the original data holders and that of data users remains a fundamental challenge for all open data projects, which constrains the motivation of data holders for opening community-driven data. Most data holders in China do not have the habit of sharing data.

3. SinoGrids: A crowdsourcing sharing platform for microscale basic urban data in China

Aiming to solve the aforementioned major challenges for community-based open urban data to empower urban studies in China, we initiated SinoGrids (Plan Xu Xiake), a crowdsourcing platform for encouraging sharing and interoperation of microscale urban data in standardized ways based on discrete gridding.

Discrete global grids have a lively research topic. A discrete global grid is a partitioning of the surface of the globe into approximately identical tiles (Sahr, White, and Kimerling 2003). Kimerling et al. (1999) compared the geometry properties of global grids according to a set of criteria, including equal area, no overlapping, consistent topology, regularity of shape, computational efficiency, etc. For SinoGrids, a square grid was selected to be the geometric basis of the grids, as it enables simple and efficient applications.

We rooted SinoGrids in a uniform 1-km grid system across the extent of China. In order to preserve the area of spatial units, we applied the Albers equal-area projection with standard parallels at 25° N and 47°N and the central meridian at 105°E. The projected extent of China was then discretized uniformly into 1-km squares. One kilometer is selected as the current scale for the grid, which is available for both regional analysis and internal urban studies. Other grid resolutions of data products will be added, providing users with other level of access according to their own contribution of data. The standardization (downscaling) process projects the original datasets onto a uniform grid. Guidelines and tools are provided without cost, which, other than being used for making datasets more sharable, could also be used for multisource data interoperation. There are many potential users, such as urban planners and government officers, who care about and understand cities, and who need the support of urban data analysis, but know little about data processing. The guidelines are written with such people in mind. By explaining the procedure in detail, SinoGrids can lower the technical requirement for data donators, making the platform open to a more general public.

Unlike many other open data platforms in China, as a crowdsourcing platform, the main data sources of SinoGrids are individual data holders. SinoGrids proposes a way to encourage data sharing and

improve the usability and interoperability of data. Besides, SinoGrids are mainly focusing on community-generated data, for example, social media records, instead of government-generated data, providing a public perspective for characterizing urban form and urban function instead of the administration perception that government-driven platforms normally provide.

In China, Data Center for Resources and Environmental Sciences (http://www.resdc.cn/Default.aspx) also provides 1-km gridded datasets. But those datasets are mainly for physical resources, for example, vegetation, land use, etc. On the social side, only gridded gross domestic product and population datasets are available. As community-driven data platforms, *Datatang* encourages data sharing by building up tunnels for data trading. SinoGrids is the first to focus on providing open gridded basic urban datasets on the social side to empower urban studies in China.

We designed a scheme for the platform to format every dataset characterizing different aspect of urban form or function as a single .dbf file. The users can download selected .dbf files corresponding to the attributes they want, and after joining with the constant 1-km grid, the dataset is ready to use. In this way, the users do not have to download datasets with all attributes or download a large grid every time, which dramatically improves the efficiency of data distribution. The independently provided grid is a guaranteed constant.

Currently, at the starting stage of SinoGrids, we have received various datasets from generous donators and intentions of cooperation from planning institutes and government departments. The available datasets mainly focus on the social aspect of urban behaviors, including social media datasets, road junctions, etc. Shortly, more datasets, such as the population grid and the public infrastructure grid, would be added. SinoGrids is a project initiated under Beijing City Lab, a virtual community for urban planners and researchers in China with over 40 research fellows, 42 junior research members, and over 8000 followers, all of whom are potential data donators and users for SinoGrids.

Interactive visualizations (Figures 1–3) are generated based on gridded data to illustrate the dataset for a general public with no technical background and no interest in data processing.

3.1. User evaluation of SinoGrids

For user evaluation of SinoGrids, we proposed a formal questionnaire survey in order to collect feedback and comments from previous and potential data donators and data users. The questionnaire mainly comes in two parts, with questions from both the perspectives of being a data holder and being a data user. There are 16 questions in total, containing both multiple-choice and subjective questions. For a certain individual, the questionnaire is able to self-adjust according to previous answered questions based on preset rules. For example, it will skip the next question if it can be inferred from previous answers. Some sample questions are listed below.

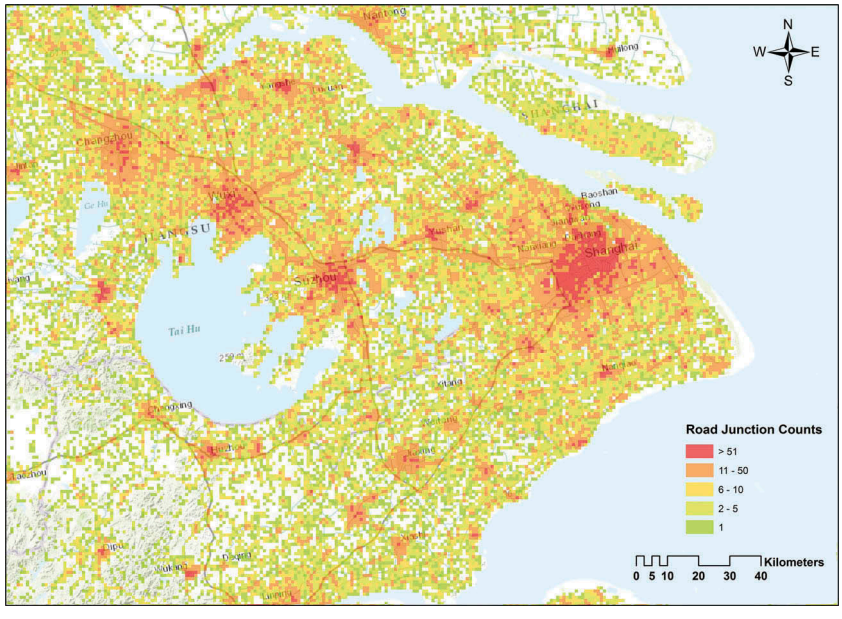

Figure 1. Map of road junction counts per square kilometer in Shanghai and its nearby area, based on SinoGrids.

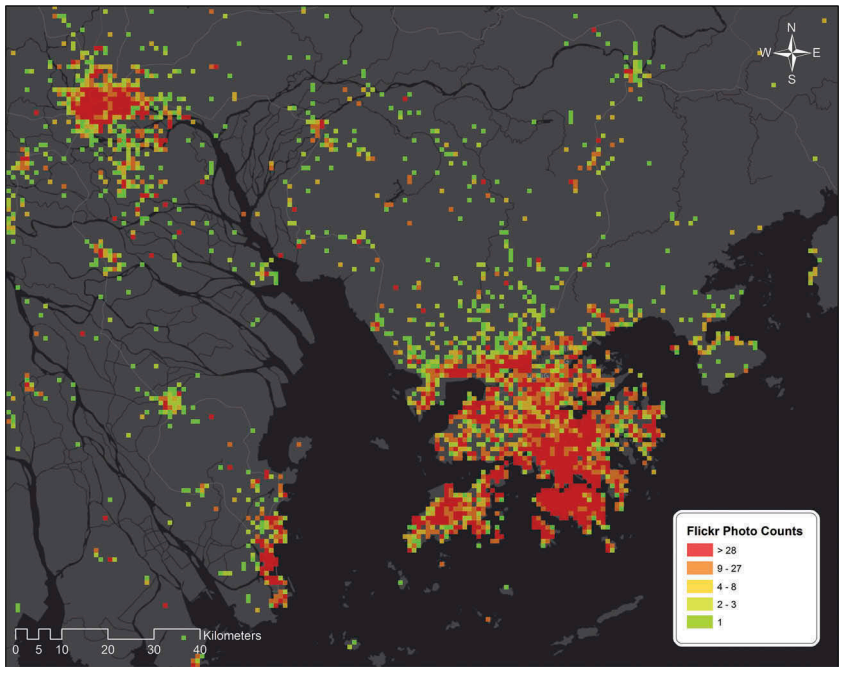

Figure 2. Map of Flickr photo counts per square kilometer in Hong Kong, Macau, Shenzhen, Guangzhou, and their nearby areas, based on SinoGrids.

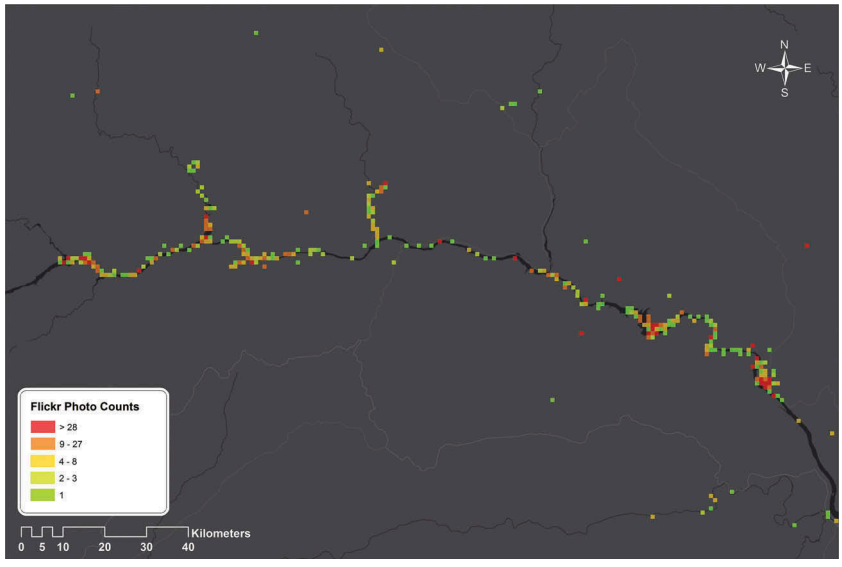

Figure 3. Map of Flickr photo counts per square kilometer along the Yangtze River and near the Three Georges Dam, based on SinoGrids.

Part 1: As a data holder,

2. What are the major concerns preventing you from sharing your urban data? (multiple choices)
   a. Potential competition.
   b. No real credit (e.g., citation), lack of motivation.
   c. Extra workload.
   d. Limitations by the primary data holder.
   e. Others (please specify):
6. Do you think the SinoGrids way reduces your concern while sharing your research data?
   a. Yes, dramatically.
   b. Yes, partially.
   c. Not exactly.
7. (If not) What are your remaining concerns?

Part 2: As a data user,

9. How often do you meet the dilemma of lacking proper urban data in realizing your ideas?
   a. Very often.
   b. Sometimes.
   c. Occasionally.
   d. Never.
10. What are the main resources of urban data in your current urban practice?
   a. Purchase.
   b. Project cooperation.
   c. Internet crawler.
   d. Open datasets or APIs.
   e. Data sharing.
13. What do you think is the proper form of spatial analysis unit for SinoGrids? (two choices at most)
   a. 100-m grid.
   b. 1-km grid.
   c. 5-km grid.
   d. 10-km grid.
   e. Irregular spatial units (e.g., blocks)
14. What do you think is the main bios of the SinoGrids way? What else do you think is effective in encouraging data sharing? Any other suggestions or opinions? (subjective question)

We promoted our questionnaire online using the most popular social media applications in China, Weibo and WeChat. The distribution of the questionnaire also took place simultaneously through Beijing City Lab, a virtual research community consisting of a wide range of urban-focused researchers, planners, commercial consultants, and news reporters. The online spread of the questionnaire was very rapid and effective. The questionnaire was able to reach large numbers of professionals by being forwarded for multiple times. In just 1 day, we got over 50 responses.

When the survey was over, a total of 158 effective questionnaires were collected and analyzed. As illustrated in Figure 4, responders mainly consist of urban researchers, students, urban planners, and others, including GIS specialists, commercial consultants, government officers, etc. As mentioned before, 79.75% of the surveyed individuals claim that they "often" find themselves lacking proper data for urban analyses.

The main concerns of the original data holders that prevent them from opening data are:

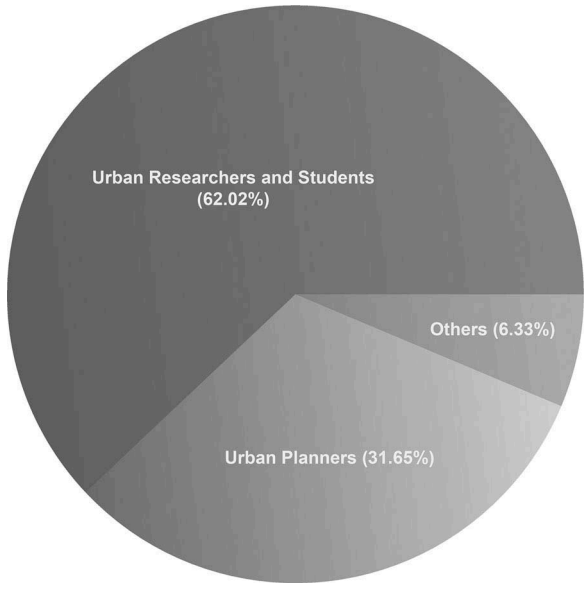

Figure 4. Summary of the professions of the surveyed individuals.

1. The possibility of losing research advantage (50%);
2. Constraints on data distribution proposed by the institute or original data generator (53%);
3. Extra workloads providing data would entail (35%);
4. Lack of real credit for data holders (54%).

SinoGrids is intended to help relieve the major concerns and promote urban data opening in a crowdsourcing way. The survey shows that 91.51% of the surveyed individuals would "have fewer concerns donating data" using the SinoGrids approach. The feedback shows that the downscaling process does contribute to solving the benefit conflict between data holders and users, making the original data holders more willing to share the data. In addition, data preparation process is simplified and smoothed, as detailed manual and GIS tools are provided, and it requires little data processing experience. The questionnaire survey also shows that 94.94% of the surveyed individuals think the SinoGrids approach could improve the efficiency and effectiveness of the interoperation of urban data from various sources.

Regarding data users, we surveyed the current data sources for urban analytics. Majority of the surveyed individuals (69.62%) obtained urban data from project collaboration, 50.63% used commercial data services, 37.34% got data through online data sprawl, and only 31.65% benefited from online open datasets.

Datasets from SinoGrids have been downloaded and effectively used by urban planners, researchers, and consultants. We interviewed several users who have used datasets from SinoGrids in their urban application. We received feedback such as "finally found free data for social media on SinoGrids, though downscaled, but good enough for my research" and "provides an effective way for data interoperation."

There are also debates on SinoGrids. Some claim that regular grids have the drawback that they do not match the geometry of real-world features. Thus, it is better to use units in irregular polygons, for example, administration blocks. However, as blocks vary in size, and China is a rapidly changing country, it is not possible to find a set of high-resolution blocks that is applicable everywhere and does not change with time. Regularly gridded datasets are more general, simpler, and more appropriate for data sharing. When needed, high-resolution, uniformly gridded data could also be further aggregated to irregular units, for example, district, block, etc.

3.2. Digital Desert: A son project of SinoGrids

What can we do with datasets in SinoGrids? One SinoGrids application is to study digital deserts, areas where social media data hardly cover. Study results based on social media data in these areas are less reliable. Urban data analysis does not necessarily provide equally valid results everywhere. The distribution of digital deserts, as illustrated in Figure 5, could improve the estimation of error for analysis results based on social media data.

In this evaluation of digital deserts, social media data from two different sources, Flickr and Weibo, are considered. Grid cells with a total number of social media data records less than a certain threshold value, in this case 6, are considered to be digital desserts, where social media data could hardly characterize urban behaviors due to the limited amount of data. The threshold is determined by local knowledge and judgment, as well as by various experiments.

The degree of digital desert percentage at provincial level (Figure 6) and municipal level (Figure 7) is calculated and visualized based on SinoGrids gridded social media datasets. A measure of digital desert can be defined as a percentage:

$$\text{Digital desert percentage} = \frac{\text{Total area of digital deserts in urban built-up area}}{\text{Total area of urban built-up area}}.$$

The rankings of 10 provinces with most/least digital deserts (Figures 8 and 9) and 15 cities with most digital deserts (Figure 10) are calculated and made open online.

At provincial level, this measure shows that more developed provinces (or SAR), Shanghai, Hong Kong, Sichuan, Zhejiang, and Guangdong, have the lowest digital desert percentage, while Heilongjiang, Xinjiang, Shandong, Inner Mongolia, and Jilin have the highest proportion of urban built-up land without noticeable social media coverage.

At the municipal level, there are 77 cities with the percentage of digital desert under 2%. Among these cities, 14 are directly governed cities, provincial capitals, and sub-provincial-level cities. Due to larger population and density, better economic development, better access to Internet, and fast-paced lifestyle, it is reasonable that more developed cities, like the aforementioned ones, have a relatively lower percentage of digital desert. The theory should also be applicable for most cities in the Beijing-Tianjin area, Yangtze River Delta, and Pearl River Delta, the three most productive and wealthy areas in China. But this

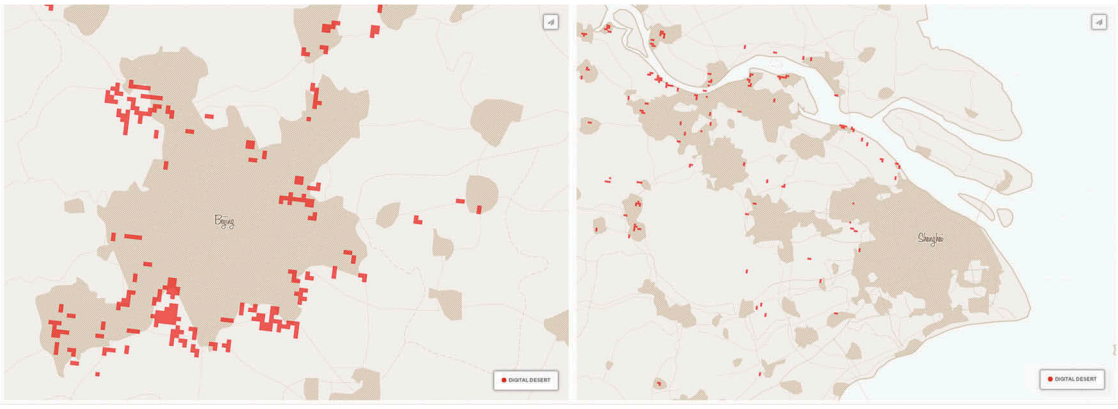

Figure 5. Online interactive mapping of digital deserts in and around Beijing (left) and Shanghai (right).

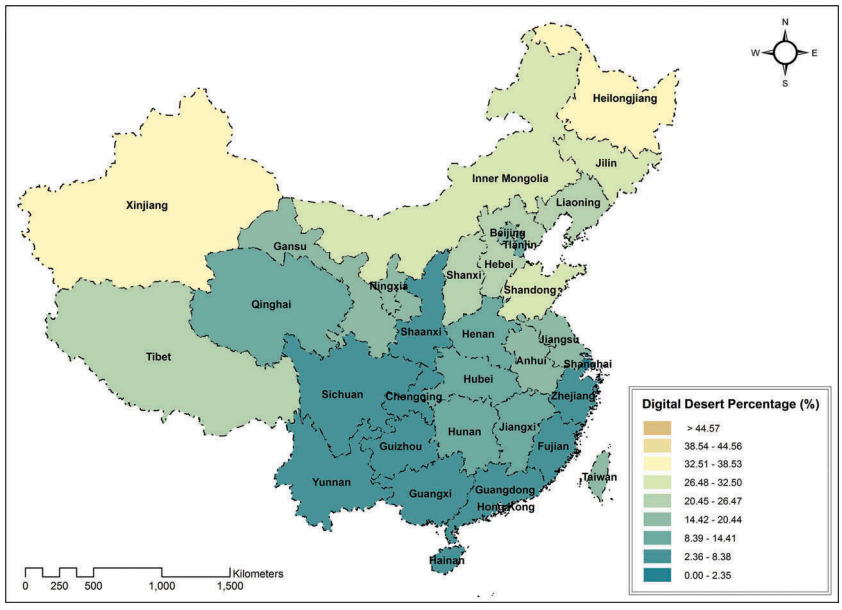

Figure 6. Map of provincial-level distribution of digital desert percentage.

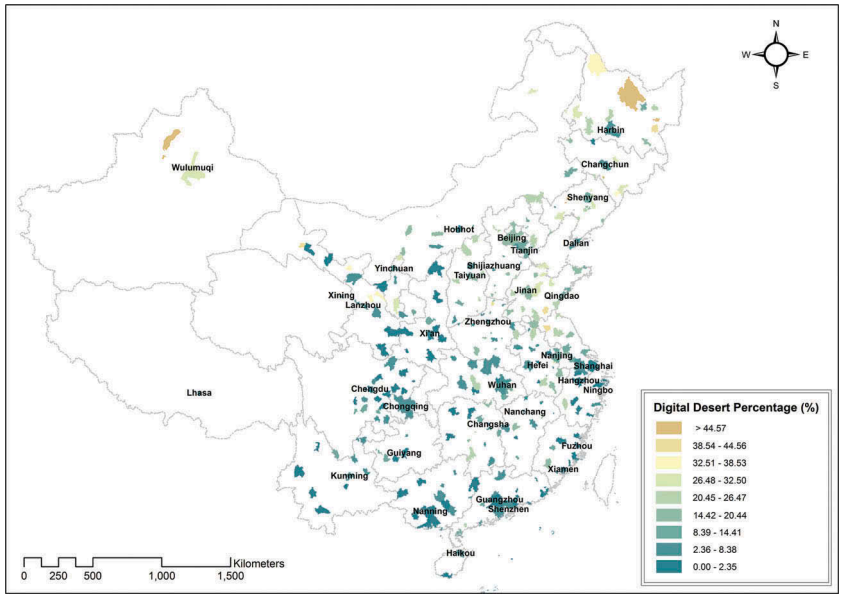

Figure 7. Map of municipal-level distribution of digital desert percentage.

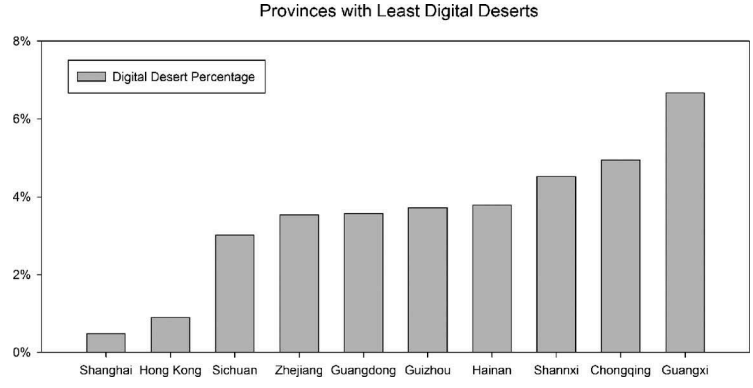

Figure 8. Ten provinces (or directly governed cities) with the lowest digital desert percentages in China.

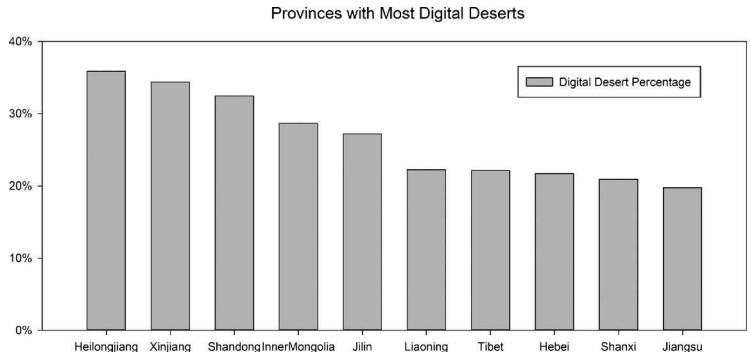

Figure 9. Ten provinces (or directly governed cities) with the highest digital desert percentages in China.

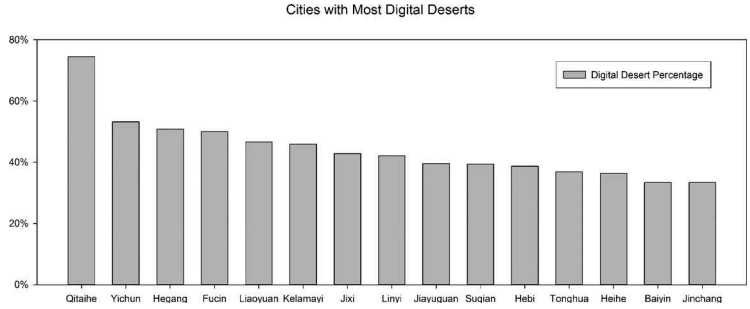

Figure 10. Fifteen cities with the highest digital desert percentages in China.

is not always true. There are also major cities with high digital desert rates, for example, Beijing (15.31%), capital of China. Beijing has the largest area of urban construction land in China, and it also has the most digital deserts in terms of net area, 385 km^2. Most of the digital deserts lie near the rural–urban fringe, and in between Beijing and its satellite towns. Having so many digital deserts within a metropolis is mainly due to the rapid urban development style. Cities in China, especially major cities, are growing fast with dramatically expanding urban boundaries. The newly constructed urban area needs time for people to move in and for the improvement of urban infrastructures, before it becomes a real urban area in terms of the intensity of human activity. Specifically for Beijing, the occurrence of massive digital deserts is also due to its dramatic urban sprawl. In contrast, 49 of the 77 cities with low percentage of digital desert (<2%; Table 1) are small cities with less than 30 km^2 of built-up land. The small area of urban land could be caused by slower growth in minor cities or natural physical constraints. Last but not least, among minor cities, tourist destinations, for example, Lijiang, Zhangjiajie, etc., have dense social media coverage because their population density is larger and people use social media more for sharing and memorizing while traveling.

Table 1. Examples of cities with low (<2%) percentage of digital desert.

City name	Percentage of digital desert (%)	Possible cause
Shanghai	0.39	Directly governed
Chengdu	1.57	Provincial capital
Guangzhou	1.52	Provincial capital
Ningbo	1.73	Sub-provincial
Xiamen	1.80	Sub-provincial
Foshan	1.93	Third largest city in Guangdong
Mianyang	0.00	Second largest city in Sichuan
Huzhou	0.00	Historical city
Sanya	0.00	Tourist destination, physical constraint
Lijiang	0.00	Tourist destination, minor city
Zhoushan	0.00	Minor city, physical constraint
Yan'an	0.00	Historical city, minor city

The nationwide distribution and ranking of digital desert percentages, together with the grid areas identified as digital deserts, illustrate how SinoGrids provides unique and interesting facts and points of reference for researchers, planners, consultants.

Shortly, more social media datasets, for example, Jiepang, are to be taken into account to make the calculations of digital deserts more reliable. The identification of digital deserts would be more valid as more social media datasets are donated, standardized, made public on SinoGrids, and integrated into digital desert detection.

4. Concluding remarks

In this paper, we discussed the current situation of open urban data in the whole world and in China, and especially listed the challenges that open urban data in China are facing, the three major ones being:

1. The balance between benefits to original data holders and data users.
2. The gap from accessible data to effectively usable data.
3. Many data holders in China fail to share data.

SinoGrids, as a crowdsourced sharing platform for microscale urban data in China, is aimed at solving the challenges mentioned above. The main approach is through standardized process for downscaling based on a discrete uniform grid. The benefits are obvious:

1. High-resolution data (e.g., geotagged social media points) from the data holders are downscaled onto a uniform grid, so that the ability to conduct further research is preserved for the data providers. Meanwhile, the data users also benefit from having more usable datasets to empower regional analysis and intracity studies.
2. The standardization (downscaling) process enables further data interoperation by normalizing data format and spatial analysis units. Guidelines and tools related to SinoGrids are freely provided. The guidelines are detailed, making SinoGrids more efficient and accessible to the general public, including urban planners and consultants with little data processing experience. Furthermore, interactive visualizations of gridded data convey the sense of urban data to the general public that has no capacity or interest in doing quantitative data analytics.

A human participated test is proposed for user performance evaluation of SinoGrids. The survey shows that lack of proper urban data to study has been a typical dilemma. But still, few people are now open to sharing data with others. Among the main concerns of data holders is the possibility of losing proprietary advantage. SinoGrids has been developed in part to relieve the concerns. The proposed survey shows that 91.51% of the surveyed individuals would "have fewer concerns donating data" via the SinoGrids way. The user evaluation also shows that 94.94% of the surveyed individuals think the SinoGrids way could improve the efficiency and effectiveness of the interoperation of urban data from various sources.

As a typical application of SinoGrids, the Digital Desert project was outlined for the purpose of identifying areas with little social media data coverage, using gridded social media datasets from SinoGrids. Rankings of 10 provinces with most/least digital deserts and 15 cities with most digital deserts, together with example cities with very few digital deserts, were generated. The nationwide distribution of digital desert percentage, along with the city rankings, based upon grids identified as digital deserts, provides unique and interesting facts to researchers, planners, consultants, and the general public. Further research could look into the underlying patterns of the generation and sprawling of digital deserts. The calculation of digital deserts would be more reliable as more social media datasets are donated, standardized, and made public on SinoGrids. The project of digital desert is also a solid example that proves SinoGrids empowers the production of valuable urban analysis results by opening and sharing urban datasets.

There are several avenues by which SinoGrids could be improved in the near future:

1. Donation bonus. Future plans for SinoGrids call for classifying the datasets into two types: 5 km gridded and 1 km gridded, and for every user, the availability of more precise datasets could be gained by contributing data that other data users could use. Everyone can be both a data provider and a data user. The more one contributes, the more one would gain. The credit system aims to encourage data sharing.
2. Data citation. For current open data portals, the common practice for data citation is by making data users cite the name of the data platform. As SinoGrids is a crowdsourcing platform, we plan to set a rule to require citation of the original data donor. By giving the credit to the contributor of the dataset, further data opening will be encouraged.
3. More datasets. The number of datasets is limited at this early stage of SinoGrids. With more datasets donated, standardized, and made public, SinoGrids would be more energetic, and be able to further strengthen interaction within the urban research, planning, and consulting community, by encouraging everyone to empower each other with open data sharing and, hopefully, forming a habit of data sharing.

Finally, SinoGrids is designed timely and accordingly based on the current data sharing context in China. However, in terms of data sharing, it has certain

limitations. For example, the downscaling method preserves geographical information of datasets reasonably, like Flickr or Weibo points. Although aggregated geographical information alone is of great value and could fulfill the needs of a wide range of regional and urban applications, the downscaling process loses spatial resolution. When it comes to data attribute information, some attributes, for example, passenger flow volume of bus stations, could be effectively summed up to the grid, but the downscaling process may not be ideal for preserving some other attributes. In the future, improvements or supplements should be made accordingly to make SinoGrids capable of leveraging more kinds of shared data as well as motivating increased urban data sharing.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

We would like to acknowledge the financial support of the National Natural Science Foundation of China (grant numbers 51408039, 51278526).

References

Aoun, N., H. Matsuda, and M. Sekiyama. 2015. "Geographical Accessibility to Healthcare and Malnutrition in Rwanda." *Social Science & Medicine* 130: 135–145. doi:10.1016/j.socscimed.2015.02.004.

Bai, X., P. Shi, and Y. Liu. 2014. "Society: Realizing China's Urban Dream." *Nature* 509 (7499): 158–160. doi:10.1038/509158a.

Batty, M. 2012. "Smart Cities, Big Data." *Environment and Planning B: Planning & Design* 39 (2): 191–193. doi:10.1068/b3902ed.

Batty, M. 2013. *The New Science of Cities*. Cambridge, UK: MIT Press.

Chen, J., and X. Chen. 2003. "Development of National Spatial Data Infrastructure (NSDI) in China: Progress and Applications." *Journal of Geospatial Engineering* 5 (2): 1–10.

Crooks, A., D. Pfoser, A. Jenkins, A. Croitoru, A. Stefanidis, D. Smith, S. Karagiorgou, A. Efentakis, and G. Lamprianidis. 2014. "Crowdsourcing Urban Form and Function." *International Journal of Geographical Information Science* 29 (5): 720–741. doi:10.1080/13658816.2014.977905.

Enserink, B., and J. Koppenjan. 2007. "Public Participation in China: Sustainable Urbanization and Governance." *Management of Environmental Quality: An International Journal* 18 (4): 459–474. doi:10.1108/14777830710753848.

Guo, H. 2014. "What Is the Role for Government, When Big Data Comes?" [In Chinese]. *Harvard Business Review China*. Accessed March 22 2015. http://www.hbrchina.org/2014-12-02/2623.html?mobile.

Gurstein, M. B. 2011. "Open Data: Empowering the Empowered or Effective Data Use for Everyone?" *First Monday* 16 (2). doi:10.5210/fm.v16i2.3316.

Haklay, M., and P. Weber. 2008. "Openstreetmap: User-Generated Street Maps." *IEEE Pervasive Computing* 7 (4): 12–18. doi:10.1109/MPRV.2008.80.

Han, H., X. Yu, and Y. Long. 2015. "Discovering Functional Zones Using Bus Smart Card Data and Points of Interest in Beijing." *arXiv preprint arXiv:1503.03131*.

Huang, B., B. Wu, and M. Barry. 2010. "Geographically and Temporally Weighted Regression for Modeling Spatio-Temporal Variation in House Prices." *International Journal of Geographical Information Science* 24 (3): 383–401. doi:10.1080/13658810802672469.

Kimerling, J. A., K. Sahr, D. White, and L. Song. 1999. "Comparing Geometrical Properties of Global Grids." *Cartography and Geographic Information Science* 26 (4): 271–288. doi:10.1559/152304099782294186.

Kitchin, R. 2014. "The Real-Time City? Big Data and Smart Urbanism." *GeoJournal* 79 (1): 1–14. doi:10.1007/s10708-013-9516-8.

Lerman, K., and R. Ghosh. 2010. "Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks." *ICWSM* 10: 90–97. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1509.

Liu, X., Y. Song, K. Wu, J. Wang, D. Li, and Y. Long. 2015. "Understanding Urban China with Open Data." *Cities* 47: 53–61. doi:10.1016/j.cities.2015.03.006.

Long, Y., X. Liu, H. Zhou, and Y. Gu. 2014. "Profiling Underprivileged Residents with Mid-Term Public Transit Smartcard Data of Beijing." *arXiv preprint arXiv:1409.5839*.

Long, Y., and J.-C. Thill. 2015. "Combining Smart Card Data and Household Travel Survey to Analyze Jobs–Housing Relationships in Beijing." *Computers, Environment and Urban Systems* 53: 19–35. doi:10.1016/j.compenvurbsys.2015.02.005.

Open Knowledge. 2014. U.S. City Open Data Census. Accessed March 20 2015. http://us-city.census.okfn.org/.

Overpeck, J. T., G. A. Meehl, S. Bony, and D. R. Easterling. 2011. "Climate Data Challenges in the 21st Century." *Science* 331 (6018): 700–702. doi:10.1126/science.1197869.

Reichman, O. J., M. B. Jones, and M. P. Schildhauer. 2011. "Challenges and Opportunities of Open Data in Ecology." *Science* 331 (6018): 703–705. doi:10.1126/science.1197962.

Sahr, K., D. White, and A. J. Kimerling. 2003. "Geodesic Discrete Global Grid Systems." *Cartography and Geographic Information Science* 30 (2): 121–134. doi:10.1559/152304003100011090.

Sakaki, T., M. Okazaki, and Y. Matsuo. 2010. "Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors." In *Proceedings of the 19th International Conference on World Wide Web*, 851–860. New York: ACM.

Si, J., A. Mukherjee, B. Liu, Q. Li, H. Li, and X. Deng. 2013. "Exploiting Topic based Twitter Sentiment for Stock Prediction." Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics, 24–29, Sofia, August 4–9.

Tranos, E., and P. Nijkamp. 2015. "Mobile Phone Usage in Complex Urban Systems: A Space–Time, Aggregated Human Activity Study." *Journal of Geographical Systems* 17 (2): 157–185. doi:10.1007/s10109-015-0211-9.

Tumasjan, A., T. O. Sprenger, P. G. Sandner, and I. M. Welpe. 2010. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." *ICWSM* 10: 178–185.

Wilbanks, J. 2014. *Portable Approaches to Informed Consent and Open Data. Privacy, Big Data, and the Public Good: Frameworks for Engagement*, 234. Cambridge: Cambridge University Press.

Yang, K., J. Chen, and W. Wu. 2001. "Recent Progress in China's NSDI Development." Proceedings of the 7th meeting of the Permanent Committee on GIS Infrastructure for Asia and the Pacific, Tsukuba, April 24–27.