# BIG/OPEN DATA IN CHINESE URBAN STUDIES AND PLANNING
## A REVIEW

YING LONG · LUN LIU

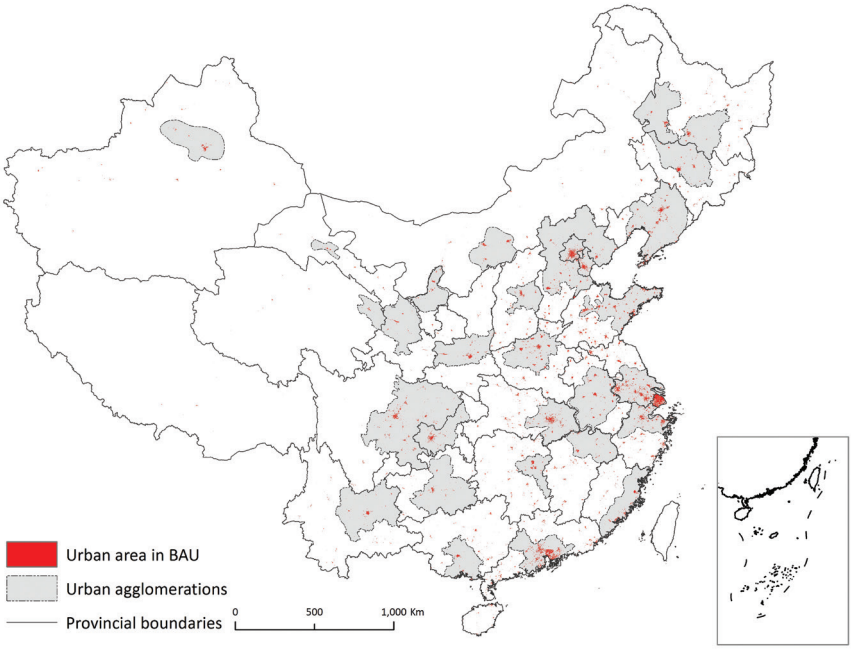
Senior lady on her smartphone

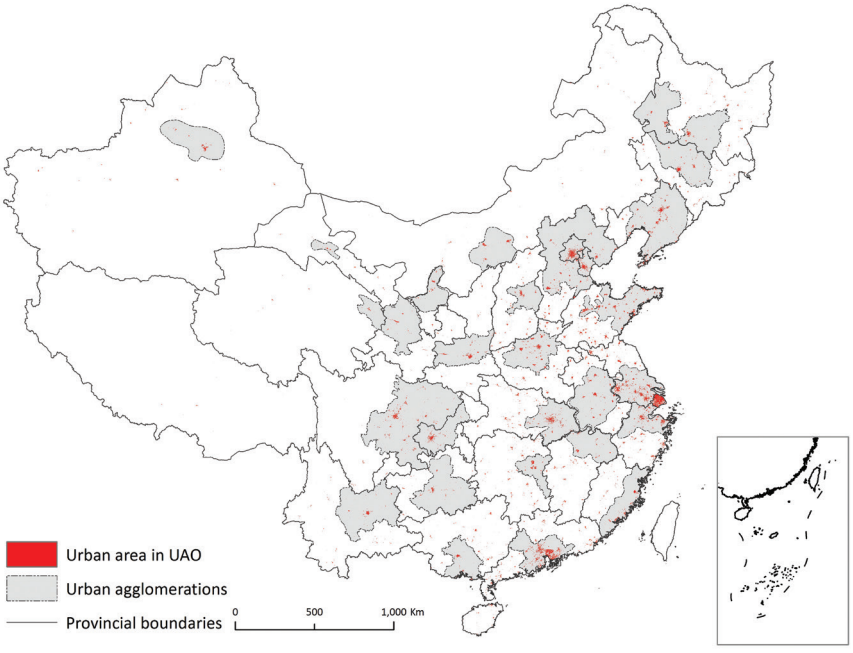
## BACKGROUND

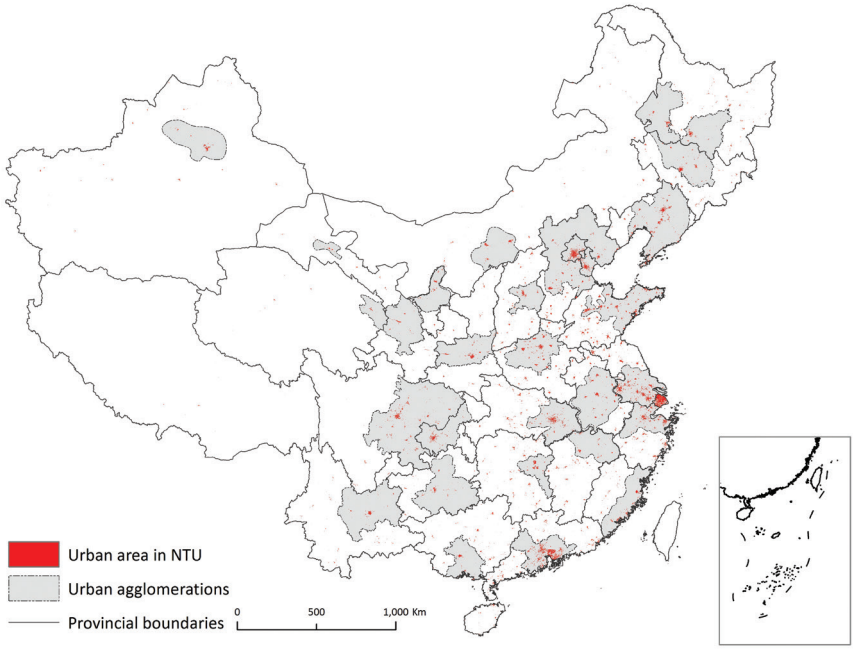
*China is the most rapidly urbanizing country in the world. Chinese cities have accommodated more than 700 million people and are attracting another 30 million annually. Consequently, China's urbanization has attracted extensive attention from academia, government, industry and the general public.*

*Despite this attention, gaining a solid and unbiased understanding of China's urbanization has been difficult due to a lack of information. In the past, urban data were mainly produced and maintained by government departments and these reports were usually not accessible to the public or to academic researchers. Now things have changed a lot due to both the government's awareness of the importance of openly providing data, and the emergence of new data sources to related the development of information and communication technologies (ICT). The collection of ICT sources, such as cell phone call records, big/open data collections has been identified as an important complement to conventional survey data and data collected by various administrative departments. These new data sets facilitate the understanding of both urban form and functions (Jia and Jiang, 2010; Goetz and Zipf, 2012; Crooks et al., 2014). As a result, urban big/open data, which a thought to correspond well to the human-oriented "New Type Urbanization" in China, have become a hotspot of Chinese urban research. Furthermore, big/open urban data have been recognized as a viable and cost-efficient option for collecting urban feature as the data infrastructure in developing countries is quite insufficient comparing with developed countries. The application of big/open data has opened up important development opportunities for urban studies, planning practice and commercial consultancy in China.*

*Before further discussion, we would like to provide a definition of big/open urban data to be used in this report. The types of data should meet two criteria: first, it would characterize certain aspects of urban form and functions (Crooks et al., 2014); and, second, it is openly accessible to the public. In terms of data sources, there are generally three overlapping though different sources of such data. The first are official data portals, enabled by the recent open government initiatives which grant public access to previously non-accessible data sources. The second are ICT-based big data initiatives, generating data from mobile phone activities, vehicle trajectories, public transit smart card data, business catalogues, as well as other smart city programs (Batty, 2012). Such data enables researchers to capture urban dynamics at very fine spatiotemporal scales and therefore gauge urban dynamics at finer spatial and temporal scales (Kitchin, 2014; Yue et al., 2014). The third source is Volunteer Geographic Information (VGI) and Crowdsourcing (Goodchild 2007; Crooks et al., 2014), which allows the general public to contribute to the urban data pool in a 'bottom-up' approach. Examples of such data*


Urban area in BAU
Urban agglomerations
Provincial boundaries
0
500
1,000 Km


Urban area in UAO
Urban agglomerations
Provincial boundaries
0
500
1,000 Km




The simulated results in the BAU scenario for typical cities

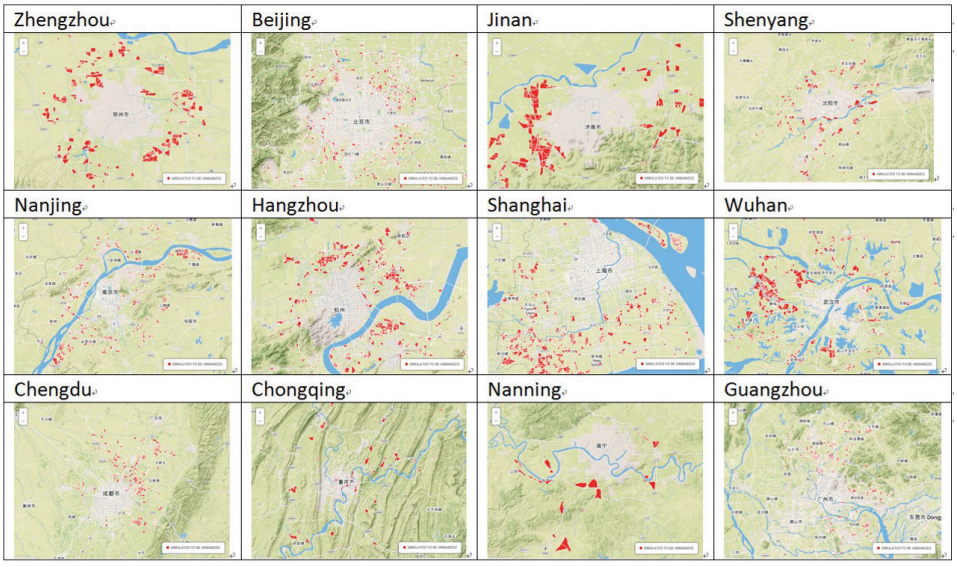



Figure 1: Urban growth simulation of all Chinese cities under various scenarios

*type include collaborative VGI mapping platforms, such as OpenStreetMap (OSM), and geo-tagged social media applications, such as Foursquare, Twitter, and Flickr (Liu et al., 2015).*

*Despite these positive changes, there are also several challenges related to the use of big/open data applications in urban studies and planning in China. First, most of data sets were collected for a given single city which makes them insufficient to formulate knowledge on the universal laws of Chinese cities. Second, visualization has been highlighted for research using instant or short-term big/open data, but very few studies are for understanding cities using big/open data accumulated for a long time, like two years. Third, some studies aggregate spatial units into grids and data are then reported using these grids. This practice leads to the loss of the granularity. Lastly, most studies have not been benefitted by including crowdsourcing in the designing and execution of academic studies nor have these data sources been used to validate results.*

*In response to these challenges, we set up the Beijing City Lab (BCL; http://www.beijingcitylab.org), an online research network to produce and store data about Chinese cities. The Beijing City Lab (BCL) is a virtual research community dedicated to studying, but not limited to, China's capital Beijing. The Lab focuses on employing interdisciplinary methods to quantify urban dynamics, generating new insights for urban planning and governance, and ultimately to the discovery of the science of cities required for sustainable urban development. The lab's current mix of planners, architects, geographers, economists, and policy analysts lends unique research strength. Through the endeavor of the core research team led by Dr. Ying Long, BCL has been developing rapidly and drawing a lot of attention from the urban planning community both in China and overseas.*

*BCL lays much emphasis on urban modeling and quantitative urban research at multiple scales. Moreover, the research conducted by BCL focuses on the living quality of human settlement in China and aims to provide both comprehensive measurement and the monitoring of China's urban development. Our research is expected to support related policy decision making. Furthermore, BCL also works on spreading messages on China's quantitative urban research in the international research community. It has now become one of the major gateway for foreign colleagues to learn about the latest progress in urban research in China.*

*Besides publishing Chinese scholarly works and data in international journals and platforms, BCL also brings in the words of foreign scholars. For instance, an interview about the past and prospect of urban modelling with Prof. Michael Batty, the director of Centre of Advanced Spatial Analysis (University College London), was conducted by BCL in 2013. Another event was the interview with Sir Peter Hall to discuss China's New Type Urbanization.*

*Through the works of BCL we have identified four major transformations of urban studies in China under the above mentioned technological and institutional background, which are transformation in spatial scale, in temporal scale, in granularity, and in methodology. The transformations are further illustrated below.*

## MAJOR TRANSFORMATIONS OF URBAN STUDIES IN CHINA

### Transformation in Spatial Scale – The Mega-Model

Existing Chinese urban and regional research can be generally categorized by the scale of study. The first type is in-depth research on a *single city*, for instance, the study of poverty in Guangzhou City (Yuan et al., 2008) or the study of the distribution of public facilities in Beijing. The second type is analysis on the *regional scale*, which covers several provinces or the entire country and uses province or county as the basic unit of analysis. An example of this type of research would be a na-

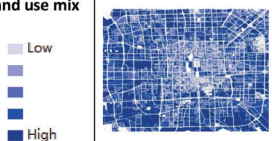
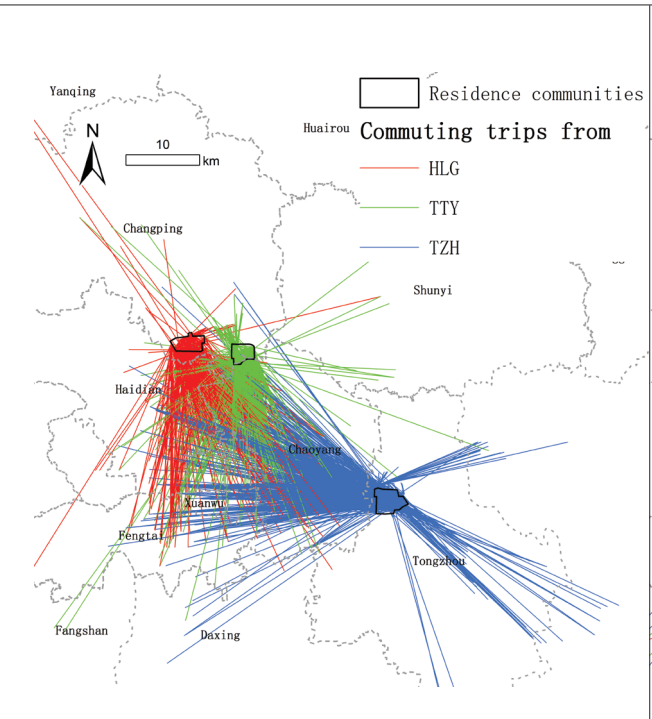
| | Beijing | Nanjing | Changsha | Weifang | Gongzhuling |
|---|---|---|---|---|---|
| City level | 1 MD | 2 SPC | 3 OPCC | 4 PLC | 5 CLC |
| Density |  | | | | |
| Function |  | | | | |
| Land use mix |  | | | | |



Figure 2: Derived land use map of 297 cities using MVP-CA



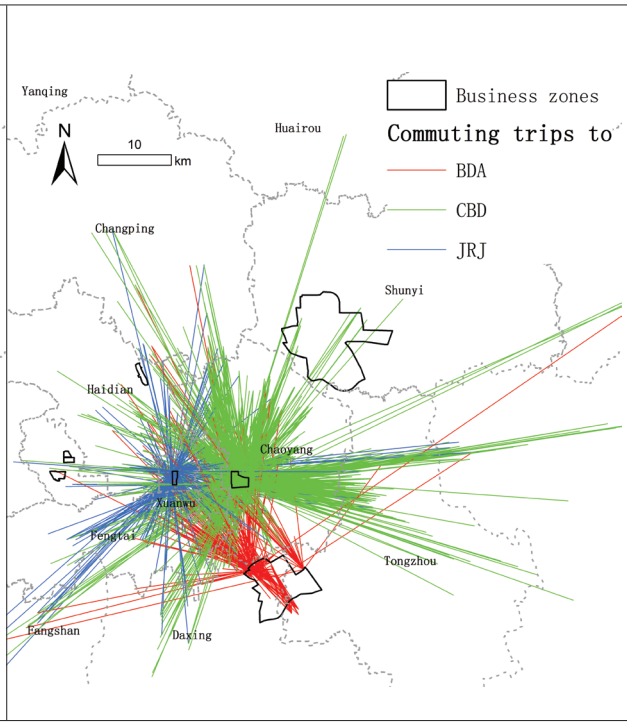




Figure 3: Commute pattern in typical areas

tional macroeconomic study. Most existing research is not able to achieve both regional scale coverage and high spatial resolution. In other words, the wide coverage of the study area is usually achieved by sacrificing details, while in-depth studies usually cover a relatively much smaller area.

To resolve this problem, we have developed a methodology capable of maintaining both a fine resolution and the capability to conducting research at the national or regional scale (Long & Shen, 2014). Termed the mega-model (or big model), this method is an effective tool for quantitative research, driven by the availability of big/open data and implemented through straightforward modelling approach (Long et al., 2014b). The resulting capability presents a new paradigm of urban and regional study (Hunt et al, 2005; Wegener, 2004; He et al., 2012). The subjects of mega-modeling are usually urban systems containing several cities, but by using this modeling method we can examine both the development of individual cities within the larger study area as well as the networking among cities.

Here we present two applications of the mega-model methodology. The first is the MVP-CA (Mega-Vector-Parcels Cellular Automata) Model which has produced detailed estimates of urban growth of all Chinese cities at the parcel scale. The full Model simulates the growth of all 654 cities in China in the next five years under various developmental scenarios (Long et al., 2014).

The second application of MVP-CA generated land use maps for 297 cities using street maps and POI (Point of Interest) data. This project was undertaken to overcome the unavailability of open land use data in China. It aims to provide free and open land use data for researchers. The layouts, land use functions, urban boundaries, densities, and degrees of land use mix are all identified as a result of the modeling (Long & Liu, 2013; Long & Shen, 2014; Long & Liu, 2015).

In our methodology, OpenStreetMap data are used to identify and delineate parcel geometries, while Points of Interest are gathered to infer land use intensity, function, and mixing at the parcel-level. To be more specific, five steps are involved. First, parcel boundaries are delineated with OSM. Second, land use density is calculated as the ratio between the counts of POIs in/close to a parcel to the parcel area and then standardized to a range between 0 and 1. Third, urban parcels are identified from all generated parcels with a vector-based constrained cellular automata (CA) model. Fourth, urban function for individual parcels is identified by examining dominant POI types within the parcels, which refers to the POI type accounting for more than 50% of all POIs within the parcel. The last step, the results are validated against both conventional manually collected parcel data and Ordnance Survey data.

### Transformation in Temporal Scale

Another breakthrough is the dynamic analysis of urban development. The data sources of conventional urban study and planning are mainly governmental statistic annals and self-conducted surveys, both of which are cross-sectional data at a single time-point. Moreover, due to the limited sampling technique, the spatial coverage of data is also limited. On the contrary, big data, such as bus/metro card records and taxi GPS traces, are able to reflect the dynamics of the urban system in minutes and seconds, which are obviously advantageous in consistency, wide coverage, and comprehensiveness. By combining these big data sets the limits of conventional urban study and research are solved (Long et al., 2012; Bagchi & White, 2004; Joh & Hwang, 2010; Jang, 2010; Roth et al., 2011; Zhou et al., 2007; Dong et al., 2009; Peng et al., 2007; Yang et al., 2009).

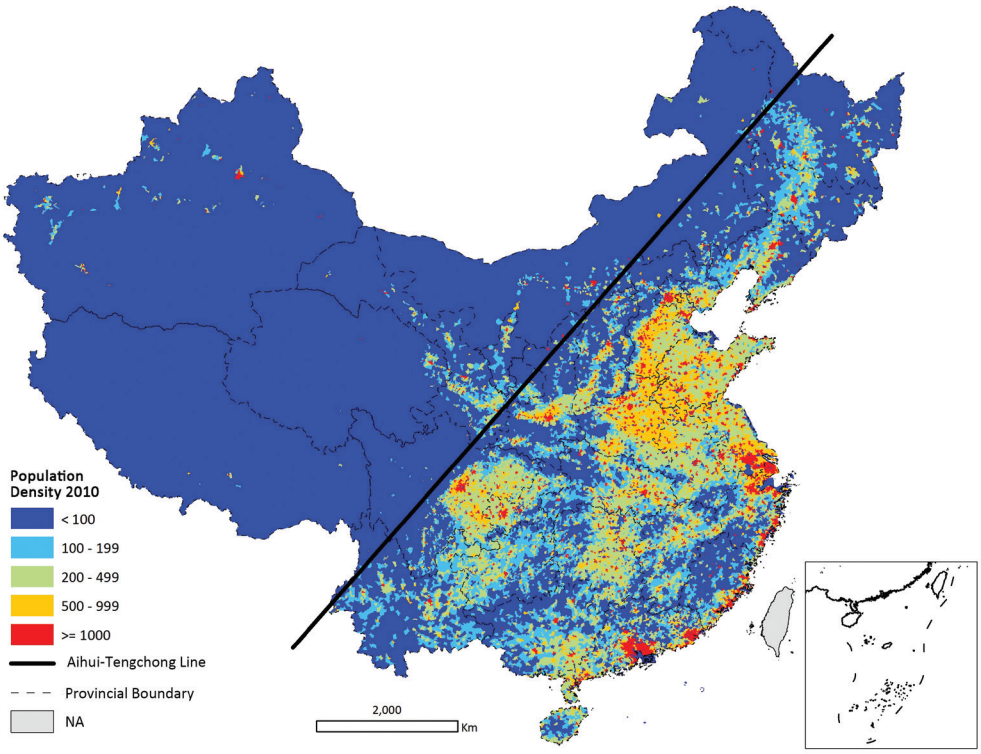



Figure 4: Distribution of population density at sub-district level in 2010

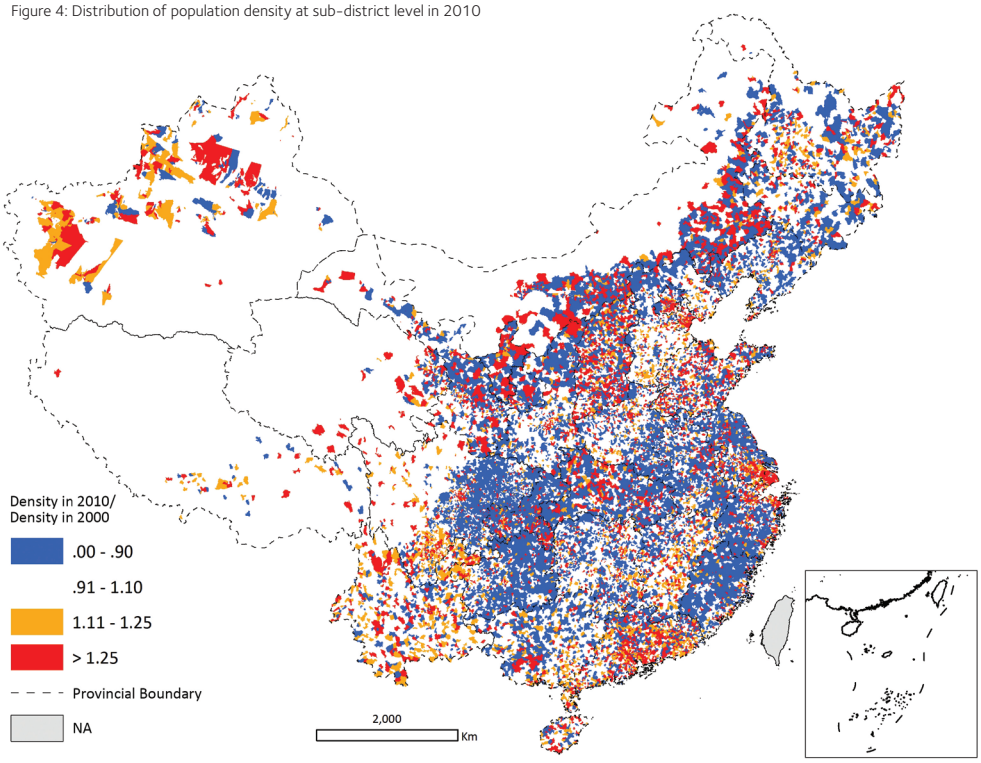



Figure 5: Population density ratio between 2000 and 2010

After being accumulated for a certain period, big data also can reflect the long-term changes and trends of urban development and life style over time. For instance, Long (2012)'s research on a week's bus card record of Beijing residents involves more than 10 million card holders and more than 100 million records, with which the commute pattern and the urban structure of Beijing are identified. The research indicates that more than 95% of full-time jobs are longer than six hours a day and 99.5% of people start their daily travel from their own homes. It also demonstrated that the influential area of the CBD is much larger than either the Shangdi technological cluster (a high tech boom town part of Beijing) or the "Financial Street" in Beijing. A further comparison of the records in 2008 and 2014 shows that the total bus trips are reducing and being replaced by metro rail trips. Moreover, having identified lower income people, from related socio-economic surveys, we found that they usually spend more time on bus and that they are more transient than expected. For example, 80% of these people move their homes within six years and 87% change jobs, which means that they live quite an unstable life. It therefore reminds us to give more consideration towards this group of people in "urban village" regeneration, public housing design, and urban design.

## Transformation in Granularity

Previously, lots of conventional urban planning and policy making was "building-oriented", which led to various social, economic and environmental problems caused by over-development. In response to this situation, many Chinese cities have put forward plans of smart growth and replaced the large expansion projects with small-scale urban regeneration and redevelopment projects. In such circumstance, planning techniques that function at large spatial scale are less useful, which gives rise to the need to develop planning information at finer scales. This new type of urbanization is defined to be human-oriented urbanization, which lays much emphasis on the human scale and the granularity of research.

An example of research with higher granularity is our study of the dynamics of nationwide population density at the town/sub-district scale. Our study reports that one third of the country's land is sparsely populated, due to the aggregation towards big cities and city centers. Besides the well-known phenomenon of decaying village (Liu et al., 2009), we also found a trend of declining populations in 180 of the 654 Chinese cities. This finding is informative for planning practice in China, which has always assumed that population growth would continue. With the recognition of this declining trend, the goal of planning in those 180 cities should no longer be land expansion, but the enhancement of residents' living quality. Another related issue is the so-called "ghost city" effect in several Chinese cities where large newly constructed areas remain vacant as a result of over-construction. Such places can be identified by evaluating the intensity of internet activity on Baidu map (Chinese version of Google map) and Weibo (Chinese twitter). The housing vacancy rate can be thus calculated for all cities with big data, from which the influencing factors and certain rules of urban development can be derived for use in policy making.

Another study reliant on big data found that 50% of all developments in Beijing are informal or illegal, with no planning permission; while, on the other hand, 95% of people's activity and mobility is still within the planning boundary. It indicates that the planning control is quite effective in the social sense, despite of its ineffectiveness in the physical sense.

The third example is a nationwide sub-district-scale analysis on the exposure to $PM_{2.5}$ pollution[1]. We relate satellite-based Aerosol Optical Depth (AOD) retrievals to ground-

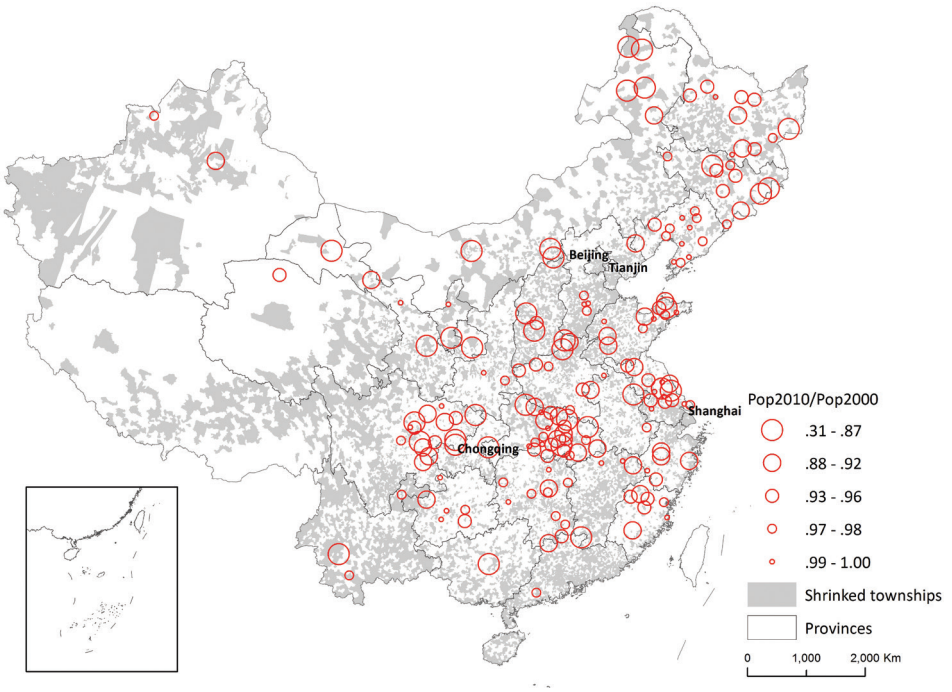



Figure 6: Ratio of shrinking of Chinese cities between 2000 and 2010

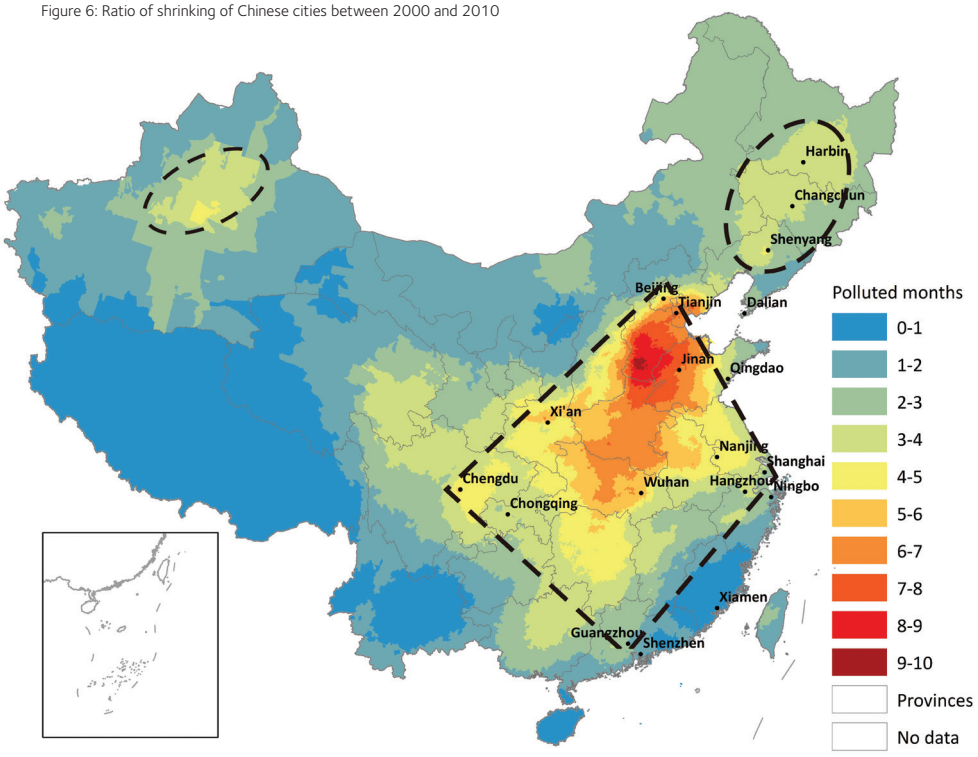



Figure 7: Urban environment at fine spatial scales: The number of polluted months in a year for each Chinese sub-district violating national PM2.5 standard.

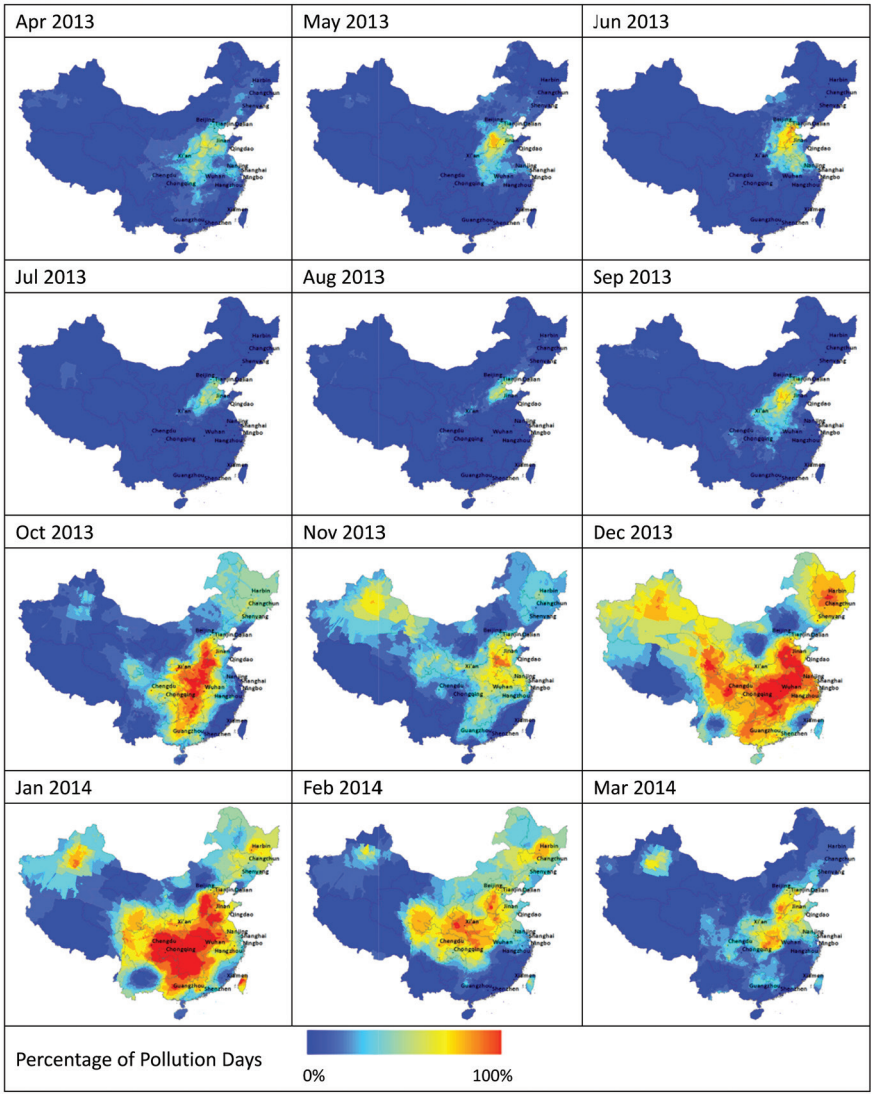



Figure 8: Exposed days in each month for each sub-district

based $PM_{2.5}$ observations. We use the sub-district level population data to estimate and map the potential population exposure to $PM_{2.5}$ pollution in China at the sub-district level, the smallest administrative unit with public demographic information. During April 08, 2013 and April 07, 2014, China's population-weighted annual average $PM_{2.5}$ concentration was nearly 7 times the annual average level suggested by the World Health Organization (WHO). About 1,322 million people, or 98.6% of the total population, were exposed to $PM_{2.5}$[1] at levels above WHO's daily guideline for longer than half a year.

### Transformation in Methodology

Similar to the trend of crowdsourcing in data collection (VGI), there is also a trend to use crowdsourcing to conduct urban research. For instance, it is almost impossible for a single research group to conduct detailed field work to measure socio-economic and spatial transformation changes in all 180 shrinking Chinese cities. Therefore, we propose the use of crowdsourcing as a new research paradigm. BCL has proposed several crowd-sourcing projects including the field survey in shrinking cities and counties, and the verification of the urban growth simulation by MVP-CA.

## CONCLUDING REMARKS AND DISCUSSION

In this short report, we present several major changes and challenges in assembling big/open urban datasets for Chinese cities, and showcase our attempt to apply big/open urban data to understand China's urbanization. This new data environment has been drawing more and more attention from both researchers and planners, since it enables detailed observation of individuals' activities in the urban space. These detailed data could be applied to provide helpful information for the decisions on heated topics such as urban regeneration, shrinking city, public participation, etc., as well as provide new developing opportunities for urban study, planning and design, construction, and commercial consultancy, which correspond to the human-oriented development strategy to the central government's New Type Urbanization policy.

Several Chinese urban research and planning institutes have started to conduct quantitative urban research using this new information, in addition to work underway at Beijing City Lab. Recognizing that the new data are able to cover large geographic area in fine resolution, we proposed the mega-model, a new regional and urban research paradigm. Meanwhile, we have also identified four transformations in quantitative urban research, namely transformation in spatial scale, in temporal scale, in granularity, and in methodology, which are all centered on the improving people's quality of life. However, there are also several issues that we need to pay attention to.

### Dealing With Data Bias

This issue has been repeatedly discussed since the emergence of big/open data. For instance, the studies on urban residents' happiness using geotagged Weibo are suffering from data bias on several aspects, including the duplicity of Weibo senders, the limitations of natural language processing technology, the representativeness of Weibo senders, and the black box of Weibo API, all of which bring doubts about the reliability of such Weibo-based studies. There are a few strategies to tackle this problem. The first strategy is to make use of the data bias. For instance, recognizing that low-income people are more likely to travel more frequently by bus, we studied the travel behavior and the change of residence and work locations of low-income people from 2008 to 2010 with smart card data (SCD). The second strategy is to study the behavior of special

groups, such as our study on the travel behavior of university students and four extreme social-economic groups in Beijing. The third strategy is to combine these big data with other data types to improve the stability of research results. For example, we combined SCD, travel survey data, social website check-in data, and taxi GPS data in our study Beijing which found that more than 95% people conduct their daily life within the planning boundary. The last strategy is to use more than one dataset to complement each other, thus depicting the whole urban system.

## Short-Term Data Visualization vs. Long-Term Data Exploration

Most data used in current research are collected in less than one week instead of years. Moreover, some research is merely data visualization. Comparing with the new data, conventional data, such as yearbook data, can reflect the transformation of the urban system over the years. However, the situation would change a lot with the accumulation of new data, which could lead to quite different research results. For example, one-day records of credit cards can be applied to identify the patterns of consumption, one-month's records might help identify the influence of festivals, while a year's records can further manifest the impacts of technological progress on consumption. Such a research trajectory is reflected in our research with SCD from 2008 to 2014.

## Current Situation Analysis vs. Future Planning Support

Up to now, there is more existing research aiming at analyzing the current situation of urban systems than evaluating their future development. This situation needs to be changed. In order to provide effective guidance to urban planning and design with the new data and research methods, we have proposed a new methodology named Data Augmented Design (DAD).

DAD is a planning and design method based on quantitative urban analysis, which provides whole-process tools for field survey, information processing, design, and short-term and long-term evaluation. DAD aims at enhancing the scientific base of design, to guide the creativity of planners and designers. To be more specific, DAD is a new design method that emphasizes the inspiration power of quantitative analysis. We expect DAD to reduce the working load of designers and thus let them focus on creative instead of repetitive work, and at the same time improve the measurability of design. Moreover, DAD is simple and straightforward, which makes it convenient to be generalized but also sensitive to the specialty of each project.

**Acknowledgement:** *We thank Dr. Xingjian Liu for his comments on the earlier manuscript of this report. Our thanks also go to all BCL members with whom we conducted academic studies summarized in the report.*

ENDONOTE

1 $PM_{2.5}$ refers to fine particulate matter in the air with diameters less than 2.5 micrometers. These fine particles are believed to present the greatest health risk.

REFERENCES


Bagchi M., & White P. (2004). What role for smart-card data from bus systems? *Municipal Engineer*, 157(1): 39-46.

Batty, M. (2012). Smart cities, big data. *Environment and Planning B*, *39*(2), 191.

Crooks, A., Pfoser, D., Jenkins, A., Croitoru, A., Stefanidis, A., Smith, D., & Karagiorgou, S. (2014). Crowdsourcing urban form and function. *International Journal of Geographical Information Science*, In press

Dong, X., Yu, Z., & Fu, W. (2009). Data Processing and Analyzing System for Bus IC Card Based on GIS. *Geospatial Information*, 7(5): 124-126.

He, L., Song, Y., & Dai, S. (2012). Research on the Paradigm of Urban Planning Responding to Uncertainty. *City Planning Review*. (07):15-22.

Hunt, J. D., Kriger, D. S., & Miller, E.J. (2005). Current operational urban land-use–transport modelling frameworks: A review. Transport Reviews, 25(3):329-376.

Jang W. (2010). Travel time and transfer analysis using transit smart card data: *Transportation Research Record*, 2144: 142149,

Jia, T., & Jiang, B. (2010). Measuring urban sprawl based on massive street nodes and the novel concept of natural cities. *arXiv preprint arXiv:1010.0541*.

Joh C.H., & Hwang C. A. (2010). Time-geographic analysis of trip trajectories and land use characteristics in Seoul metropolitan area by using multidimensional sequence alignment and spatial analysis, AAG Annual Meeting, Washington, DC.

Kitchin, R. (2014). The real-time city? Big data and smart urbanism. *GeoJournal*, *79*(1), 1-14.

Liu, L., Long, Y., & Batty, M. A Retrospect and Prospect of Urban Models: Reflections after Interviewing Mike Batty. *City Planning Review*. 38(8): 63-70.

Liu, X., Song, Y., Wu, K., Wang, J., Li, D., & Long, Y. (2015). Understanding urban China with open data. *Cities*. In press (doi: 10.1016/j.cities.2015.03.006)

Liu, Y., Liu, Y., & Zhai, R. (2009). Geographical Research and Optimizing Practice of Rural
Hollowing in China. *Acta Geographica Sinica*. 64(10):1193-1202.

Long, Y., Gu, Y., & Han, H. (2012). Spatiotemporal heterogeneity of urban planning implementation effectiveness: Evidence from five master plans of Beijing. *Landscape and Urban Planning*, 108: 103-111.

Long, Y., Liu, X. (2013). Featured graphic: How mixed is Beijing, China? A visual exploration of mixed land use. *Environment and Planning A, 45*, 2797–2798.

Long Y., & Liu X. (2015). Automated identification and characterization of parcels (AICP) with OpenStreetMap and Points of Interest. *Environment and Planning B*, Forthcoming.

Long, Y., Wu, K., Mao, Q. (2014a). Simulating parcel-level urban expansion for all Chinese cities. *arXiv preprint* arXiv:1402.3718.

Long, Y., Wu, K., Wang, J., & Shen, Z. (2014b). Big models: From Beijing to the whole China. *arXiv preprint* arXiv:1406.6417.

Long, Y., & Shen, Y. (2014). Mapping parcel-level urban areas for a large geographical area. *arXiv preprint* arXiv, 1403.5864.

Peng, H., Zhao, Q., & Zhao, S. (2007). Transfer Matrix Construction Method Based on Bus IC Card Data Processing. *Computer and Communications*. (4): 32-34.

Roth C., Kang S.M., & Batty M. (2011). Structure of urban movements: Polycentric activity and entangled hierarchical flows. *PLoS ONE*, 6(1): e15923.

Wegener, M. (2004). Overview of land-use transport models. *Handbook of Transport Geography and Spatial Systems*, 5:127-146.

Yuan, Y., Xu, X., & Xue, D. (2008). Spatial Types and Differentiation Mechanism of New Urban Poverty ofGuangzhou City in Transitional China. *Geographical Research*. 27(3):672-682.

Yue, Y., T. Lan, A. G. O. Yeh and Q.-Q. Li (2014). "Zooming into individuals to understand the collective: A review of trajectory-based travel behaviour studies." *Travel Behaviour and Society* 1(2): 69-78.

Zhou, T., Zhai, C., & Gao, Z. (2007). Approaching Bus OD Matrices Based on Data Reduced from Bus IC Cards. *Urban Transport of China*. 5(3): 48-52.