Big Data for urban studies: opportunities and challenges: a comparative perspective

Jianquan Cheng*, Nicholas Gould
School of Science and the Environment
Liangxiu Han
School of Computing, Mathematics and Digital Technology
Manchester Metropolitan University (MMU)
Manchester, UK
J.Cheng@mmu.ac.uk; N.Gould@mmu.ac.uk;
L.Han@mmu.ac.uk

Cheng Jin
School of Geographical Sciences
Nanjing Normal University, Nanjing, China
jincheng@njnu.edu.cn

***Abstract*— A city is a complex system with complicated interactions between transport, land use, environment and population at a variety of scales. Understanding these interactions is the prerequisite for predicting urban changes and supporting sustainable urban development planning. In this paper, we divide the evolution of urban data into four stages. Then the opportunities of big data (new stage) for urban studies application are explored and followed by case studies from both the UK and China. The main challenges are evaluated and the solutions to which are further discussed and compared between both countries.**

***Keywords*- Urban Big Data, Urban System, Urban Data Evolution, Interactions, Analytics, Model and Users.**

I. INTRODUCTION

A city is a complex system with complicated interactions between transport, land use, environment and population at a variety of scales from global, national, urban, and neighborhood [1]. Sustainable urban development needs scientific evidence for understanding the system and predicting its dynamics from short to long term. Since the first occurrence of computers in 1950s, urban studies have been being massively stimulated by the exciting progressions in data availability, urban modelling and planning support system. Nowadays, the emergence of big data (theories, methods and techniques) has brought great opportunities for urban studies. However, the current use of big data for urban studies is still at its early stage due to the nature of city complexity (e.g. the integration of social, technical, political, cultural aspects). In this paper, we will discuss the challenges of big data for urban studies through comparing China with the UK. In the next section, we will review the evolution of data sets for urban studies. The third section will explore the opportunities of big data for urban studies in general, followed by case studies from both countries. The fourth section will be focused on analyzing and discussing the challenges and solutions in China and UK.

II. EVOLUTION OF URBAN DATA

In addition to political and socio-economic dimensions, spatial and environmental dimension is very crucial for exploring urban systems and particularly supporting urban planning. Since the first GIS in 1968, spatial data has evolved significantly due to rapid advances in senor and computer technologies. The evolution of urban data can be specifically divided into four stages in terms of resolution, scale and :1 (1990-2000) satellite images dominated spatial data; 2 (2000-2010) census surveys based socio-economic data; 3 (2010-2015) disaggregate and locational data and 4 (2015-present) "big data" (e.g. movement, qualitative) though the exact year of each stage may vary with country (e.g. the North and South).

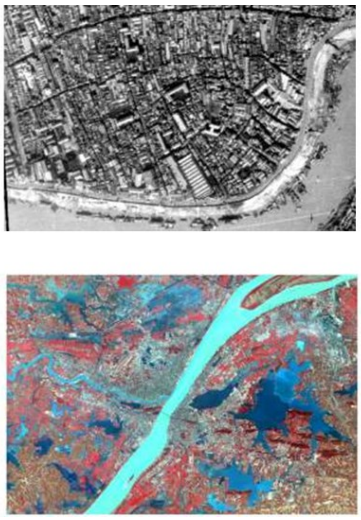

Fig 1. Aerial photograph in 1965 (inner city) & SPOT imagery in 1986

At stage one, urban data was dominated by primary spatial data captured by remote sensing including satellite imagery and aerial photography. For example, in the study of monitoring temporal urban growth of Wuhan city [2], the aerial photographs of 1955 and 1965, and SPOT images of 1986 and 2000 (Fig 1) were deployed for producing land cover maps in different periods. These temporal sets of spatial data enabled us to map (Fig 2) and model the spatial and temporal patterns and processes of urban development [3-5]. However, the spatial and temporal resolutions of these

imageries were not high enough in that period even for creating urban land use maps.

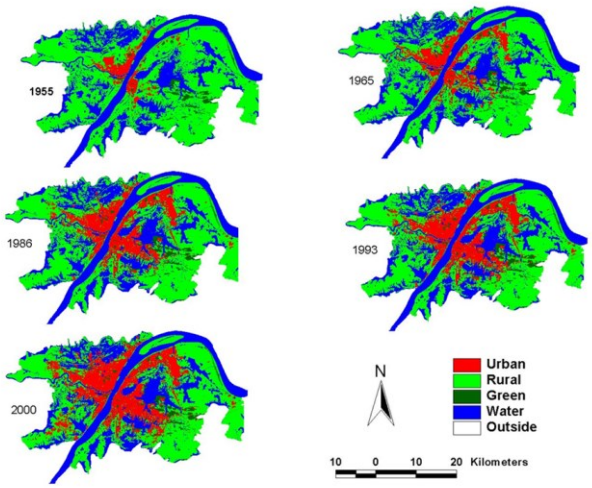

Fig 2. Monitoring urban growth of Wuhan city

At stage two, with the regular and legal implementation of large-scale national census, economic and other surveys, socio-economic and particularly demographic data sets at selected levels of spatial statistical units (including district, sub-district, community or residential committee in China (Fig.3) and district, ward, output area in the UK) have become available.

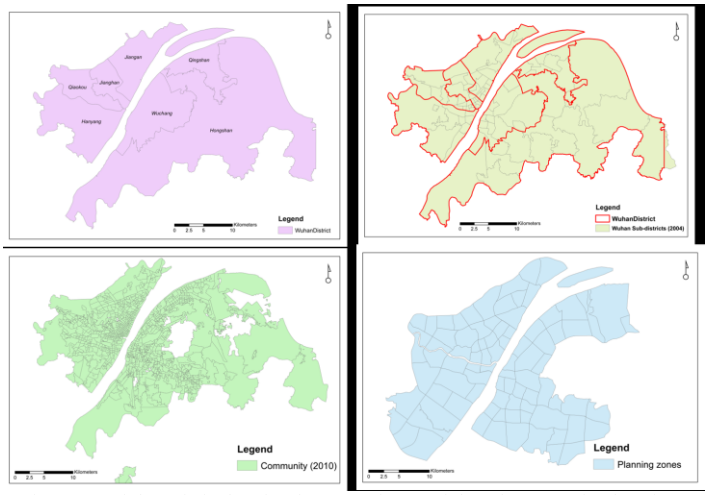

Fig 3. Spatial statistical units for reporting social and economic data in Wuhan, China (district, sub-district, community and planning zone levels)

For urban studies, these survey data sets were accessible to academia at sub-district level before 2010 in China. The availability of these non-spatial data sets has enabled the integration of spatial and social-economic data into urban studies. For example, population data can be dis-aggregated from sub-district level to pixel level (e.g. 30x30 m^2) using Dasymetric method and land use data from satellite imagery [2] (Fig.4). The detailed census data at urban community level can be used for measuring residential segregation in inner city (an example of Nanjing City in Fig 5) [6].

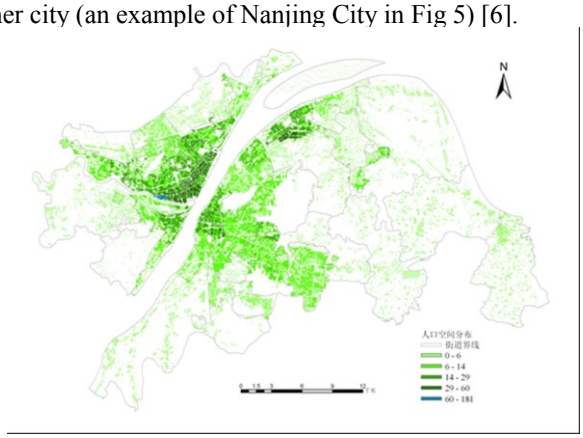

Fig 4. Population density of Wuhan City in 2000 at pixel level

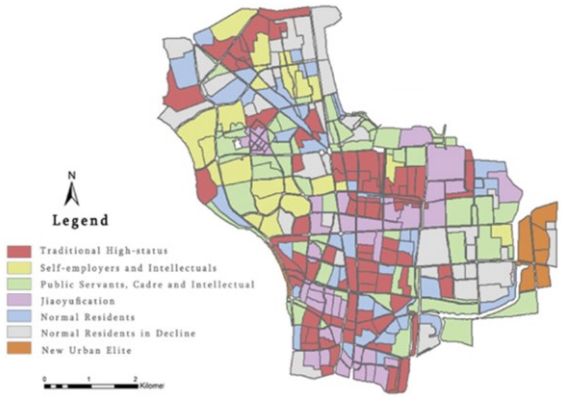

Fig 5. Residential segregation using 2000 census data at urban community level

However, these units in the case of China are too coarse to be useful for spatial-social analysis as discussed by Cheng et al. [7]. Particularly, the census data at the finest spatial statistical unit – urban community (or residential committee) has become unavailable through official channels recently.

At stage three, with the increasingly wide use of GPS techniques by the public, location data has been extensively deployed for urban studies, such as location of bus stops, location of healthcare facilities (Fig 6). These data sets are mostly point data, which can be georeferenced easily using GPS-embedded smart phones or other equivalent devices. However, due to unavailability of attribute data (e.g. number of people) at individual level, such as building level, it is not sufficiently accurate to measure accessibility to public transport and healthcare facility in Fig 6. As a consequence, it is imperative to demand data about the location of people and their dynamics during 24 hours, for example, where do they live, work, drive, shop and visit? In summary, the spatial and particularly socio-economic data are not only unavailable but also with low resolutions in China. Comparatively, in the UK, the spatial data sets (e.g. building footprint and height, which is called DSM (Digital Surface Model), integrated transport network) have been accessible

freely to academia. The socio-economic data sets (e.g. population census) available are georeferenced from district level down to output area level (Fig 7). The average size of output area is about 15 hectare, which is much smaller than the unit of urban community in China (average size is around 71 hectare). However, both countries share the unavailability of the high-resolution data sets, regarding population and their movement at individual level and qualitative data, at this stage.

At stage four, the new era of big data is providing massive opportunity for urban studies, which will be discussed in details in the next section.

Fig 6. Location of bus stops and healthcare facilities in Wuhan (2014)

Fig 7. Spatial statistical units in the UK

III. URBAN BIG DATA

A. *Opportunities*

Although there is no standard definition of big data, it is commonly recognized that big data has several properties, such as 3V (*high Volume, high Variety and high Velocity*). Volume refers to the amount of data, variety to the number of types of data and velocity to the speed of data processing. *High Volume* means that the data generated by machines, networks and human interaction on systems are recorded as petabytes or even exabytes, instead of gigabytes. *High Velocity* indicates that data are generated through a streaming process in real time and in a continuous fashion, rather than batch processing. High variety denotes that the data collected from a variety of sources have hundreds of formats, structured and unstructured.

"Big Data" is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it [8]. Big data has been growing exponentially, motivated by the technical advances in GPS, social media (e.g. Facebook, Twitter, Instagram, Google) and smart devices (e.g. sensors, smart phone and meters).

With big data, it allows for analyzing and modelling the increasingly complicated interactions in dynamic urban systems, particularly in China, where rapid urbanization has been transforming the cities at an unprecedented pace, physically and socially. There have been increasing challenges for local urban planning to achieve sustainable, resilient and smart development (theories/methods/practice) by human-centered approach.

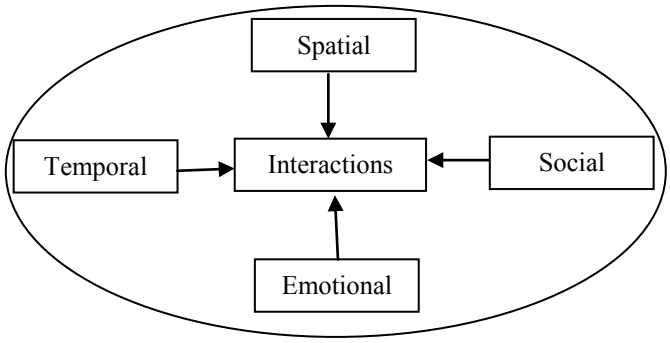

Fig 8. Interactions in urban system

For instance, with big data, we can analyze and model not only the spatial and temporal interactions (such as traditional transport and land use interaction) that dominated the previous stages, but also the social and emotional interactions (Fig 8). Social media (e.g. Facebook and Twitter) provides a platform for social interaction between all age groups at every moment. Emotional interaction means people' feeling when interacting with the

environment. For example, when people walk in a street, the same level of light and noise may give different pedestrians different feelings and accordingly results in different actions.

B. Case studies (China and UK)

In China, there is an open tourism-oriented country-wide web service called "Where to go?" (www.qunar.com), which provides an interactive social media platform for tourists to upload and exchange their trip related information (e.g. story and photo for destinations and routes) across China. This kind of data with *high variety* could be recorded before or after a trip. Such data, though in the format of text or photo, have high spatial and temporal resolution (e.g. the spots visited and on which day), which enables to analyse and model the temporal patterns of tourist flow network. In our case study of Nanjing, there are totally 1424 valid samples of tourist, who visited the destination city of Nanjing in 2012, involving 45 scenic spots. With these data sets, we can model and compare the distance decay effects between one-day, two-day and three-day tours (Table 1), which indicates the scale-dependent property of tourists flow (Fig. 10) [9].

Table 1 Distance decay effects between one-day, two-day & three-day tours

	Function	Regression Equation	*Adjusted* R^2
One-day tour	Revised power function	$T_{ij} = 0.1307 \times C_{out}^{0.424} C_{inj}^{0.433} / d_{ij}^{0.248}$	0.724
Two-day tour	Revised power function	$T_{ij} = 02047 \times C_{out}^{0.390} C_{inj}^{0.387} / d_{ij}^{0.370}$	0.662
Three-day and above tour	Revised Power function	$T_{ij} = 0.3848 \times C_{out}^{0.308} C_{inj}^{0.330} / d_{ij}^{0.529}$	0.633

Fig 9. Tourist flow network of three-type tours in Nanjing city

Transaction data, as a typical example of big data particularly with *high volume* and *high velocity*, has been increasingly becoming accessible in many countries including China. Rapid urbanization in this country has been significantly promoted by the provision of a high-quality motorway network. The flows of vehicles between counties within a province and their spatial variations indicate the structure and quality of the transport system. In the case study of Jiangsu province, which has the highest density of motorway in China, there are more than 334 fee-collecting stations on the motorway network within the province, which record each vehicle passing through it. This big data enables to analyze the patterns of the flows and variations between the counties. In 2014, there have been a total of 234,535,769 records collected, which can be split into many matrices (334×334) for temporal analysis. Some preliminary results presented in Fig 10 indicate the spatial concentration of flows, distribution of flows between 59 counties and the centrality degree of transport system within each county across the province.

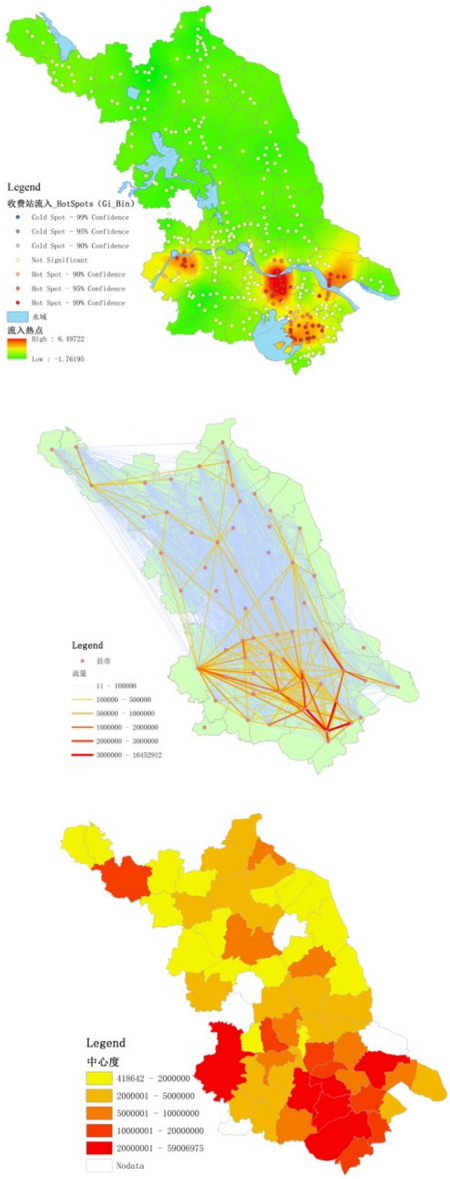

Fig 10. Flows of vehicles between stations and counties

Big data has been increasingly available in the UK through research collaborations between institutes and industrial partners. For example, based on collaboration with Transport for Greater Manchester, the real-time historical traffic monitoring data has been deployed for detecting congestion and measuring resilience of transport system. In this case study, the big data consists of journey time data on links collected from passive Bluetooth sensors and traffic

volume data from permanent induction loops (Fig.11). For both data sources, the data with *high volume* and *high velocity* was aggregated into 10-minute time slots. In this study, relative congestion, or nonrecurring congestion, is defined by the standard deviation from the mean journey time on links on "normal" days. It is focused on the day of 13th January 2016 when there was a major incident on the M62 motorway and a football match at the stadium in the evening. Such big data enabled the detection of spatial and temporal clusters (Fig 12) and animation of the congestions within the day (http://tinyurl.com/mcrcongestion).

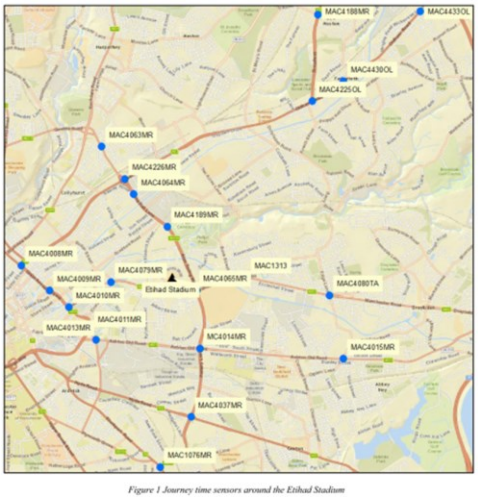

Figure 1 Journey time sensors around the Etihad Stadium

Fig 11. Journey time sensors around the Etihad Stadium

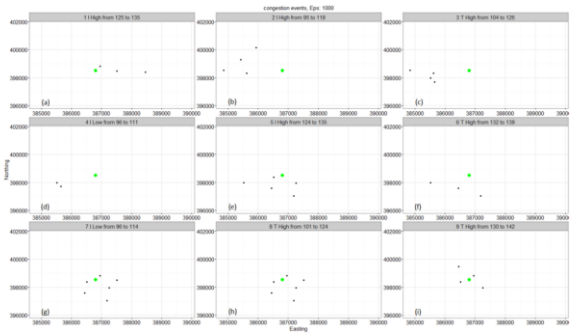

Fig 12. Spatial and temporal clustering around the stadium

IV. CHALLENGES

It is well recognized that big data has offered opportunity for integrating data with increasing resolutions and coverages at multiple levels; developing multi-scale and diverse-purpose system modelling with more variables (indicators) and larger-size samples, and particularly understanding people' movement, behavior, preferences and opinions with dynamic and qualitative data. However, there are many challenges for big data into urban studies, from the perspectives of data, analysis, modelling, system and users.

First of all, in the aspect of data, the quality of data needs to be further improved, particularly the qualitative data (e.g. emotion, perception and comments) from social media may have noise. There is limited number of sensors in Fig 11. An urgent demand is an integrated data infrastructure that is able to combine aggregate, disaggregate and particularly individual data sets at a variety of spatial and temporal scales. Another challenge is the need to integrate heterogeneous data sources - the *Variety* of big data. Considering, for example, the two types of traffic sensor, discussed earlier (Fig 11), although both react to the passing of vehicles, they provide significantly different information; the Bluetooth sensors provide a measure of *journey time*, which is directly related to congestion, whereas the induction loops provide a measure of the *volume* of traffic, which is indicator of *potential* congestion.

Secondly, in the aspect of analytics, it is very demanding to visualize big data rapidly, interactively and interpretably. The 3D geo-visualization of people's movement and activity based on time geography is an appropriate attempt but it can hardly handle large-volume samples. The spatio-temporal analysis is subject to MAUP and MTUP [10]. Photo and text analytics need to be integrated into spatial and image analytics, which is particularly instrumental for modelling the complicated interactions in the urban system, such as sentiment analysis. There is lack of photo analysis in the case study of Fig 9 and insufficient spatio-temporal analysis of flows in Fig 10.

Thirdly, with the aspect of modelling, big data provides full and large size sample, which facilitates the calibration and validation of systematic models and easily achieves high model accuracy. However, it does not mean they are the best models for the target as the causality or correlation discovered from the data might be not true in reality. Compared to this so-called pseudo-model, big data also creates 'Big Model', which means plenty of parameters and statistics are produced. A good example is geographically weighted model (GWR), in which a map can be created for each independent variable, and its standard error and t-statistics [11]. However, big model has created challenges for interpreting these outcomes, some of which might be conflicting and controversial.

Fourthly, the storage, analysis and output of urban big data require appropriate computer architecture and software engineering strategy for coupling different programs. The mainstream GIS software -ArcGIS, has become increasingly a large and heavy system, which can hardly handle big spatial data in terms of storage, computation and visualization. Some open source software such as QGIS could provide a flexible environment for processing and analyzing big spatial data through an open structure and the wealthy community of developers and users. In addition to this, parallel /high performance computing are sought-after solutions for addressing big data processing and analytics challenges in urban studies. In the case study of workshop based planning support [12], it has been proved that the parallel computation approach enables the fast building of planning scenarios within an acceptable time for workshop based participation when dealing with big data with *high volume*.

Last but not least, regarding the users related to big data, there are not only the issues of data ownership, privacy, and copyright but also of inequity. The inequity is indicated by data production and data access. In data production, there is significant disparity between urban and rural, large and small city, as the latter may have poorer access to devices, which can be used to produce data. In particular, those people with digital disability, literacy, and poverty may not be able to disseminate their own produced data. In data access, there is disparity between the central, provincial and local-level institutions across China. For example, researchers in Beijing have higher opportunity of accessing big data with national coverage than other cities as they may have had collaborations and networks with national organizations. There is also significant disparity between staff members at a variety of grades. Professors have better access to big data than others as they have had extensive network or industrial partnership with data producers or providers. In some sense, UK shares the similar situation in these inequities as China.

V. Discussion and Conclusion

It can be concluded that big data, as a new stage of urban data evolution, has offered massive opportunities for urban studies particularly in the aspects of understanding the complicated spatial, temporal, social and emotional interactions due to high availability of movement, qualitative and emotional data sets [13]. However, the successful application of big data still need to solve many challenges in data, analytics, model, system and users, most of which are shared between the North (e.g. UK) and South (China). For example, to respond to the challenges of big data research, UK has launched a campaign of open data in 2010, which aims to open up the large amounts of data to the public without fee but with license permission (https://data.gov.uk/). A good example is the crime data geo-referenced at street level (https://data.police.uk/), which is updated on a monthly basis and has become accessible freely online since December 2010. At national level, several universities (e.g. UCL, Leeds, and Liverpool) have established world-leading research centres on big data with different foci (e.g. transport, crime, retail and health). National research council such as ESRC has regularly provided funding to support the studies of big data (e.g. secondary data analysis initiatives).

Comparatively, China still has a long way to open urban data to the public. Urban data sets are provided through collaboration/agreement with governments or purchased from markets (e.g. taxi company) though this has been the case in the UK as well.

Looking back the four stages of urban data evolution, China has poorer data infrastructure than the UK, not only at aggregate levels but also in the new era of big data. Both data production and access have shown spatial and social inequities in China and the UK.

References

[1] M. Batty, “Cities and Complexity: Understanding Cities with Cellular Automata, Agent-Based Models, and Fractals”, MIT Press, 2007.

[2] J. Cheng, “Modelling Spatial and Temporal Urban Growth”, ITC: Enschede, 2003.

[3] J. Cheng and I. Masser, "Urban Growth Pattern Modelling, a Case Study of Wuhan, P.R.China." Landscape and Urban Planning, 62(4), 2003a, pp.199-217.

[4] J. Cheng and I. Masser, "Modelling Urban Growth Patterns: a Multi-scale Perspective." Environment and Planning A, 35(4), 2003b, pp.679-704.

[5] J. Cheng, and I. Masser, "Understanding Spatial and Temporal Processes of Urban Growth: Cellular Automata Modelling." Environment and Planning B: Planning and Design, 31(2), 2004, pp.167-194.

[6] Q. Wu, J. Cheng, C. Guo, D.J. Hammel, X.Wu, "Socio-spatial Differentiation and Residential Segregation in the Chinese City based on the 2000 Community-level Census Data: A Case Study of Inner City of Nanjing", Cities: International Journal of Urban Policy and Planning, 39, 2014, pp.109-119.

[7] J. Cheng, J. Turkstra, M. Peng, N. Du & P. Ho, "Urban Land Administration and Planning in China: Opportunities and Constraints of Spatial Data Models." Land Use Policy , 23 (4), 2006, pp.:604-616.

[8] M. A. Beyer and D. Laney. The Importance of Big Data: A Definition. Stamford, CT: Gartner, 2012.

[9] C. Jin, J. Cheng, J. Xu, “Using User-generated Data to Analyze and Model Tourist Flows between the Scenic Spots within Nanjing City, China”, Journal of Travel Research (in revision).

[10] T. Cheng, & M. Adepeju, “Modifiable Temporal Unit Problem (MTUP) and Its Effect on Space-time Cluster Detection. PLoS One, 9(6), doi:10.1371/journal.pone.0100465.

[11] A. S. Fotheringham, C. Brunsdon, M. Charlton, “Geographically Weighted Regression: The Analysis of Spatially Varying Relationships”, John Wiley & Sons: London, 2003.

[12] J. Tu, J. Cheng, L. Han, “Big Data Computation for Workshop-based Planning Support”, IEEE Xplore (2015 IEEE International Conference on), 2015, pp.1510 - 1514 (DOI:10.1109/CIT/IUCC/DASC/PICOM.2015.226) .

[13] M. Batty, “Big data, Smart Cities and City Planning”, Dialogues in Human Geography, vol.3(3), 2013, pp.274–279.