Big spatial data for urban and environmental sustainability

Bo Huang and Jionghua Wang

Department of Geography and Resource Management, The Chinese University of Hong Kong, Hong Kong, China

ABSTRACT
Eighty percent of big data are associated with spatial information, and thus are Big Spatial Data (BSD). BSD provides new and great opportunities to rework problems in urban and environmental sustainability with advanced BSD analytics. To fully leverage the advantages of BSD, it is integrated with conventional data (e.g. remote sensing images) and improved methods are developed. This paper introduces four case studies: (1) Detection of polycentric urban structures; (2) Evaluation of urban vibrancy; (3) Estimation of population exposure to PM2.5; and (4) Urban land-use classification via deep learning. The results provide evidence that integrated methods can harness the advantages of both traditional data and BSD. Meanwhile, they can also improve the effectiveness of big data itself. Finally, this study makes three key recommendations for the development of BSD with regards to data fusion, data and predicting analytics, and theoretical modeling.

1. Introduction

The era of “Big Data” is coming and transforming our understanding of this world. Data whose size is exceeding the capacities of standard contemporary data management tools are considered “Big Data” (Batty 2013; Jo and Lee 2018). However, the rapid advances in technology blur the definition of “Big”. To further consolidate the definition of “Big Data”, some dimensions, like the Three V’s (volume, variety, and velocity) have been specified (Chen, Chiang, and Storey 2012; Laney 2001; Patgiri 2018), which dissociate the concept of “Big” from a fixed volume. Hereby, “Big data” refers to extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions. Based on these traits, emerging spatial data characterized by the three V’s which exceeds contemporary computing technologies counts as “big” (Shekhar et al. 2012). Big Spatial Data (BSD) in domains like satellite imagery, Internet of Things (IoT), Location-Based Services (LBS) and climate simulations suits the characteristics identified above well and spawns specialized systems, techniques and algorithms from its very beginning. The wave of utilizing BSD started even before the era of big data itself. This wave was predicted by Deren Li and Deyi Li in 1993 (Li, Wang, and Li 2015), and is continuing without a sign of ending.

New opportunities are provided by incorporating BSD and its analysis into solutions to challenging urban and environmental problems (Chen et al. 2018a; Song et al. 2018). With advances in data acquisition, new data sources (e.g. cellular signaling, social media, smart cards, review forums, video cameras, smart equipment, online taxi services, and satellite imagery) are emerging. These “new” data are adopted to boost applications like Volunteered Geographic Information (VGI) systems, demography studies, sharing economy apps, and recommendation systems. Reviewing these innovations, the benefits of BSD are threefold: (1) BSD can support spatial analysis of unstructured data in real-time. Maps can visualize unstructured data (e.g. E-mails, blogs, social media content, in-store sensor data, meteorological data, and driving times), which functions as raw data in location analysis for domains like retail, finance, and insurance. (2) BSD can integrate data from various sources, providing a more comprehensive picture. When doing so, a huge amount of data is pulled from different formats, devices or systems and given a geographic context to facilitate building a complete picture or analysis. The utilization of multiple data sources together can enhance the performance of BSD in terms of its coverage, spatio-temporal resolution and interpretation capability via techniques like image fusion (Ghasemian 2016). For instance, Chen, Huang, and Xu (2017) formulated a land cover classification framework with multi-source remotely sensed data. In their case, the land cover classification accuracy rate was improved to 92.31% with a 5% improvement compared to the conventional single data source approach. (3) BSD can tap huge datasets for policy measures. GIS tools for BSD processing will facilitate deep insights and predictive modeling for policymaking in health care (Song et al. 2018; Liu, Paciorek, and Koutrakis 2009), crime detection (Bogomolov et al.
2014), and disaster response (Voigt et al. 2007). Compared with conventional data sources (e.g. survey data), BSD is typically obtained at lower cost but has better coverage. Therefore, policies can be created based on more concrete and focused evidence provided by the emerging BSD. For example, Bengtsson et al. (2011) tracked population movements with a high spatiotemporal resolution by using Subscriber Identity Module (SIM) data during the Haiti 2010 earthquake and cholera outbreak. Their techniques can provide highly valuable information for developing rescue strategies.

There is a growing discussion that BSD leads to challenges as well as bringing opportunities (Sivarajah et al. 2017). Traditional solutions and theories thus now have to be reworked under a BSD context. Five challenges are highlighted in urban and environmental sustainability: (1) Huge data volumes; (2) Diversity: Different sources, structures, formats, scales and resolutions; (3) Representative bias; (4) Heterogeneity: Relationships that vary over space and time; (5) Deep patterns. In general, the challenges of big data in GIS today are less about the hardware and more about managing the huge volume of information and transforming it into value.

To advance BSD analytics, four researches are proposed to fuses traditional data sources (e.g. remote sensing data) with emerging BSD (e.g. social media check-ins, GPS position requests, and housing transactions). In this way, these two discrepant data sources (shown in Table 1) complement each other and thereby bring new insights to BSD analytics. In the following section, four applications (Cai, Huang, and Song 2017; Huang et al. 2019; Huang, Zhao, and Song 2018; Song et al. 2019, 2018) will be reviewed to share the experience in solving urban and environmental sustainability problems with emerging BSD and innovative methods. China is selected as the study area for these applications because China is one of the world’s fastest urbanizing countries and has immense amounts of available emerging BSD. These two points grant China great potential to advance the study of urban and environment sustainability with emerging BSD and through new analytics. The first and second applications focus on answering conventional problems (urban vibrancy and city polycentricity) with BSD directly. In these two studies, social media data is employed as the major data source for detecting urban activity. The third and fourth applications focus on generating high quality base data with BSD. Satellite images are incorporated with new analytics reworked for BSD and show significant improvements. Specifically, land-use and air pollution data are generated with improved spatiotemporal resolution and measurement accuracy.

The remaining parts of this paper are organized as follows: The methodology section introduces the four case studies regarding their background, method, data, and study area. Subsequently, the result section discusses the major findings from each study. The discussion section highlights the identified future directions and limitations of this study. Finally, a brief conclusion is provided.

2. Methodology

2.1. Detection of polycentric urban structures

BSD analytics is capable of integrating multiple data sources and harnessing them to rework conventional urban problems from new perspectives. Urban areas with a polycentric structure are emerging worldwide (Bai, Shi, and Liu 2014). Urban systems today are bigger but characterized by more independent settlements as well as more complex structures (Liu and Wang 2016). This shift has aroused widespread attention in studies on urban sprawl, population movement dynamics, and urban management. Hence, it is of great importance to delineate polycentric structures within cities. However, conventional detection methods for polycentric urban structures are limited and unable to leverage the advantages of BSD: Population census data and economic statistical data are used to detect urban sub-centers with techniques like Locally Weighted Regression (LWR) and semiparametric employment density function. This framework was firstly proposed by McMillen (2001) and later advanced by subsequent studies. However, these approaches have two disadvantages: (1) The Central Business Districts (CBDs) are selected subjectively; (2) The statistical data is categorized by administrative units and has relative low update frequencies. Concerning the first issue, the identification of sub-centers in the city is difficult for a user without detailed knowledge of the study area. The CBDs may have vague and constantly changing boundaries which are difficult to mark manually. Concerning the second

| Table 1. Comparison between remote sensing data and BSD. |
|----------------------------------------------------------|
| **Remote Sensing Data**                                      | **BSD**                                                                 |
| **Timeliness** | Time delay, thus hard to obtain real-time spatial data | Easy access to the real-time spatial data |
| **Data Type** | Continuous surface data (Raster) | Most of them are points and polylines Human daily activities and events Extremely big |
| **Direct Reflection** | Earth surface physical properties | Unstructured Data service providers every citizen |
| **Data Size** | Big | |
| **Data Structure** | Structured | |
| **Data Source** | Data service providers | |

2018
issue, population censuses and economic statistical data are usually updated every five or ten years, which limits the data’s ability to catch up with rapidly developing cities as well as their structure. For instance, the urban expansions happening between census dates and the current status quo are overlooked. Meanwhile, the identification of sub-centers, in terms of their boundary, is also limited by the spatial units of the administrative boundary. Consequently, some dense activity sites might be neglected when they are located within a large administrative region.

To address the abovementioned issues, a new detection method for polycentric urban structures is proposed in this study: BSD, nighttime light satellite images and social media check-in data are utilized to delineate urban structures with higher resolution in both spatial and temporal dimensions. The proposed method includes three steps: (1) development of observation units, (2) main center definition, and (3) sub-center definition, as shown in Figure 1. The proposed method is later applied to three megacities in China (Shanghai, Beijing, and Chongqing) for testing its effectiveness.

Three mega-cities (Shanghai, Beijing, and Chongqing) including their municipal (city-controlled) districts are selected as the study area. These three cities are selected due to their discrepant geographical characteristics and urban morphology patterns. Shanghai is located in the Yangtze River Delta. The city is subdivided by the Huangpu River into an old urban center and a modern downtown. Beijing is the capital of China and lies on the North China Plain. There are no natural conditions limiting urban expansion. The city has been developing in classic pie form with concentric ring roads. Chongqing is famous as a “mountain city”. The urban area is criss-crossed by mountains and strongly sloped.

![Figure 1. Workflow for detecting the polycentric urban structure with BSD.](image-url)
rivers. The city is thus fragmented into small pieces and has a complex urban structure.

Two kinds of BSD (nighttime light imagery and social media data) are adopted in this study to identify urban structures. Nighttime light imagery is used to characterize the urban texture as well as to observe statistical units anew. The sensor of this satellite imagery was launched in October 2011 and has a ground spatial resolution at 500 m (Miller et al. 2012). The social media data consists of Check-in data (from 25 April 2015, to 25 May 2016) collected from the Weibo Application Programing Interface (API) (http://open.weibo.com/wiki/API%E6%96%87%E6%A1%A3_V2/en) and is utilized to represent human activities. The Check-in events upload the users’ geo-locations as a geo-tag when the user creates a post with their social media account (Andreas and Haenlein 2009). Both datasets are later projected into a Universal Transverse Mercator (UTM) grid with a resolution of 500 m.

After the data progressing step, three steps are employed to identify polycentric urban structures. First, new observation units are built up. Human activity is not restricted by the administrative unit boundaries, which is typically too coarse. Therefore it is necessary to redevelop new observation units that can better suit the spatial characteristics of human activities. Then, a Multiagent Object-based Classification Framework (MAOCF) is introduced to build up these new observation units (Zhong, Zhao, and Zhang 2013). This approach is selected because it can control the procedure of object merging and leverage the advantages of contextual information. Three factors (scale, shape, and compactness) are set using the optimization method based on segmentation quality criteria (Chabrier et al. 2006). Here, both inter-segment homogeneity and intra-segment heterogeneity are taken into account (Espindola et al. 2006; Zhou et al. 2013). Second, the main centers are picked out by utilizing the Local Moran’s I (LMI) (Anselin 1995). Segments with a positive z-score (larger than 1.96) are picked out and merged if two segments are spatially adjacent. Then, the larger segment with more check-in records is identified as the main center area. Third, sub-centers are identified through Geographically Weighted Regression (GWR) (Fotheringham, Brunsdon, and Charlton 2003). Under normal circumstances, the closer an area is to the main city center, the more human activity (obtained through the square root of the check-in density for one segment) should be observed. However, the existence of a sub-center will have a similar effect as a main center and thus cause a spatially contiguous positive residual error which is revealed in the GWR model. Given this assumption, the relationship between human activity and distance to the main center centroid (check-in density weighted) is built up. By doing so, the contiguous tracts with positive residual error indicate the sub-centers in terms of their boundaries and locations.

2.2. Evaluation of urban vibrancy

In a previous study, polycentric urban structures of three cities in China were identified with high spatiotemporal resolution BSD. In this study, the association between urban structures and human activity, i.e. the urban vibrancy, will be further deciphered. Urban vibrancy is of great importance to sustainable urban development (Hall and Pfeiffer 2013). Evidence shows that vibrancy can be associated with social and economic sustainability (Brenner 2014; Brenner, Peck, and Theodore 2010), affirmative living conditions (Couitre 2013), promotion of human activity and interaction (Jacobs 1992), citizen subjective well-being (Pinquart and Sörensen 2000), innovation capability (Montgomery 1998) and urban resilience (Dale, Ling, and Newman 2010). However, defining the concept of urban vibrancy is challenging. Common approaches usually measure urban vibrancy with the following factors: (1) Urban Built Environment (UBE) (Gehl 2011; Jacobs 1992), (2) population density (Simmel 2012; Wirth 2011), (3) safety (Jacobs 1992), (4) social capital, and (5) cultural capital (Stern and Seifert 2010). Among them, social and cultural capital have rarely been discussed quantitatively due to data availability issues before the late 1990 s.

With advances in data acquisition techniques, the emerging BSD is now providing us with new opportunities to investigate the dynamics of urban vibrancy with higher spatiotemporal resolution. BSD (e.g. call detailed record data, Wi-Fi access point data, GPS data, check-in data, public transport smart card data, and mobile phone tracking data) typically come with massive sample sizes and fine spatiotemporal resolution. These features will benefit the analysis of urban vibrancy under a spatiotemporal heterogeneous context. However, most of the previous studies using BSD are limited by two issues: (1) single-source data sets are used to represent vibrancy; (2) the UBE and its connection with vibrancy are rarely quantitatively discussed in these studies. To address these issues, this approach thus has three objectives: (1) A comprehensive framework is built to evaluate and characterize urban vibrancy using multi-source BSD; (2) A hypothesis about the spatial dynamics of urban vibrancy is tested; (3) The association between urban vibrancy and UBEs is evaluated.

In this approach, Shanghai, a mega-city in China, is adopted as the study area. Shanghai is considered the pioneer in urbanization since the 1978 market-oriented reforms in China. With more than 24 million residents, this city is an ideal case for...
advancing the study of urban vibrancy from municipal (Montgomery 1998) or district (De Nadai et al. 2016) level to a finer spatial resolution. As shown in Table 2, three kinds of measurements are employed to capture vibrancy in urban areas: social activity intensity (SI), economic activity intensity (EI), and pedestrian density (PD). Meanwhile, UBE is measured by build density, the density of urban functions, mean building height, diversity of urban functions, diversity of building age, mean house price, mean building age, and density of road junctions. All data are assimilated into a grid with a 1 km² spatial resolution (Zhou and Long 2016).

The proposed urban vibrancy evaluation framework (shown in Figure 2) includes two steps: (1) Assessment of the Comprehensive Urban Vibrancy Index (CUVI); (2) Characterization of urban vibrancy. In the first step, a Factor Analysis (FA) is employed to extract urban vibrancy. By using the common factor with the largest eigenvalue, the CUVI is calculated for each grid pixel. This CUVI thus extracts information from multiple indicators of vibrancy and outperforms a single facet approach. In the second step, the CUVI is characterized by UBE indicators with a 10-fold linear regression.

2.3. Estimation of population exposure to PM$_{2.5}$

As evidenced in the previous study, BSD can measure human activity in urban areas with both higher spatiotemporal resolution and lower cost than conventional approaches. Such emerging high-resolution human activity measures are also a critical component when assessing the health impact of air pollution caused by fine particulate matter. PM$_{2.5}$ (fine particulate matter with aerodynamic diameters of less than

| Attributes                  | Indicator                        | Data source            | Year |
|-----------------------------|----------------------------------|------------------------|------|
| Social activity intensity   | Sina Weibo data                  |                        | 2016 |
| Economic activity intensity | Dianping life services reviews   |                        | 2016 |
| Pedestrian density          | Mobile phone GPS positioning     |                        | 2016 |
| UBE: Density                | Building density                 | tianditu.com           | 2016 |
| UBE: Density                | Density of urban functions       | map.baidu.com          | 2016 |
| UBE: Intensity              | Mean building height             | lianjia.com            | 2016 |
| UBE: Diversity              | Diversity of urban functions     | map.baidu.com          | 2016 |
| UBE: Diversity              | Diversity of building height     | lianjia.com            | 2016 |
| UBE: Diversity              | Diversity of building age        | lianjia.com            | 2016 |
| UBE: Diversity              | Diversity of house prices        | lianjia.com            | 2016 |
| UBE: Quality                | Mean house price                 | lianjia.com            | 2016 |
| UBE: Aging buildings        | Mean building age                | lianjia.com            | 2016 |
| UBE: Walkability            | Density of road junctions        | lianjia.com            | 2016 |

Figure 2. The workflow of urban vibrancy assessment.
2.5 mm) is a kind of fine particulate matter which can cause severe damage to human health (Anenberg et al. 2010; Cohen et al. 2001). This problem has led to increasing attention worldwide. China is particularly strongly affected, with an annual mean PM$_{2.5}$ concentration of 61 $\mu$g/m$^3$ (Fang, Wang, and Xu 2016). This number is six times higher than the air quality standard (10 $\mu$g/m$^3$) recommended by the World Health Organization (WHO). 1.2 million premature deaths per year from 1990 to 2010 are associated with PM$_{2.5}$ pollution (Peng et al. 2016). It is thus of great importance to estimate the population exposure to this pollutant with high spatiotemporal accuracy. This may be useful for related environment and epidemiologic studies (Chen et al. 2018a).

Conventionally, station-based data are widely used in estimating surface PM$_{2.5}$ yielding high update frequencies but sparse spatial distribution (Kwan 2016; Nyhan et al. 2016). However, gridded high spatial resolution PM$_{2.5}$ data (e.g. 1 km$^2$ spatial resolution concentration maps) can better suit the purpose of analysis. To this end, satellite-station-hybrid models are proposed to improve ground-level PM$_{2.5}$ estimations. This way, a large number of approaches leverage the satellite-derived Aerosol Optical Depth (AOD) to invert the spatial distribution of PM$_{2.5}$ (Lv et al. 2017; Wang and Christopher 2003). This inversion utilizes the relationship between station-based data (e.g. PM$_{2.5}$) and satellite-based data (e.g. AOD) regressed at the station locations to estimate the station-based data at locations without station observations. However, this relationship is spatiotemporally heterogeneous (Huang, Wu, and Barry 2010). To solve this problem, local (spatial) weighted regression techniques are widely used to build up the relationship between station-based data (e.g. PM$_{2.5}$) and satellite-based data under heterogeneous assumptions (He and Huang 2018b; Ma et al. 2015; Xie et al. 2015). However, most current PM$_{2.5}$ exposure studies use only station-based data and thus lead to uncertainty in measuring the exposure of the population. The uncertainty is mainly caused by coarse administration units (e.g. census tracts) and low update frequencies (e.g. 5–10 years population survey data). Some approaches, like questionnaire surveys (Cohen et al. 2001) and GPS-based wearable air monitors (Wang, Kwan, and Chai 2018), have been proposed to provide exposure measures of individuals with high spatiotemporal accuracy. However, the sample sizes of this kind of approach are typically limited due to privacy issues and costs. The spatiotemporal accuracy, in terms of the distribution of the population and the pollutants, is still limited.

To address these issues, this approach thus proposes a new PM$_{2.5}$ exposure and health risk assessment method using data on dynamic population distribution and PM$_{2.5}$ concentration generated by the satellite-station-hybrid method. The dynamic population distribution is generated with Location-Based Service (LBS) data (i.e. geo-tagged posters) from Weibo (http://weibo.com). The PM$_{2.5}$ concentration is acquired using the satellite-station-hybrid method: A Geographically and Temporally Weighted Regression (GTWR) algorithm (He and Huang 2018b; Huang, Wu, and Barry 2010) is adopted to model the relationship between the Moderate Resolution Imaging Spectroradiometer (MODIS) 3-km AOD product and the ground-level PM$_{2.5}$ measurements.

In this study, thirteen cities (Beijing, Tianjin, Shijiazhuang, Tangshan, Qinhuangdao, Handan, Xingtai, Baoding, Zhangjiakou, Chengde, Cangzhou, Langfang, and Hengshui) in the Beijing-Tianjin-Hebei (BTH) region in China are selected as the study area due to serious air pollution problems in the corresponding regions. With a total of 111 million people living in the BTH region, the annual mean PM$_{2.5}$ concentration exceeded 90 $\mu$g/m$^3$ in 2014 (He and Huang 2018b), which is significantly higher than the average level in China (61 $\mu$g/m$^3$) (Fang, Wang, and Xu 2016) and the air quality standard recommended by the WHO (10 $\mu$g/m$^3$).

The workflow (Figure 3) of the proposed method includes four sections: (1) High spatiotemporal resolution dynamic population estimation; (2) High spatiotemporal resolution PM$_{2.5}$ concentration estimation; (3) Exposure assessment, and (4) Health impact assessment. Specifically, four kinds of datasets (dynamic population, demographic data, PM$_{2.5}$ concentration, and mortality data) are used in this approach, as shown in Table 3. First, the dynamic population (monthly population density map) is estimated. In this step, the Weibo records are aggregated into a pre-generated grid with each grid pixel representing the number of Weibo records for a given month. Next, the number of records in each pixel is converted to a percentage by dividing it by the total number of records in the pixel’s belonging city. The total population of a city is then distributed into this grid according to this generated percentage. Second, a high spatiotemporal resolution PM$_{2.5}$ dataset is generated. To do so, an AOD dataset with high spatial coverage is created by fusing the 3-km AOD product’s newly released 3-km Dark Target and the 10-km Deep Blue AOD product (Wei et al. 2019). Following this, the relationship between AOD and PM$_{2.5}$ is built with GTWR at 384 ground stations. The built GTWR model includes a dependent variable (observed PM$_{2.5}$ concentration at ground stations) and a set of explanatory variables (i.e. AOD, relative humidity, temperature, wind speed, and normalized difference vegetation index) extracted at each spatiotemporal position. The built spatiotemporal heterogeneous model is then applied to all spatiotemporal positions (the BTH region) along with the explanatory variables (He and Huang 2018b, 2018a). By doing so, a high spatiotemporal resolution PM$_{2.5}$
dataset is generated by applying the built relationship to all spatial and temporal points. The result shows the generated dataset can capture more than 80% of the variation of daily PM$_{2.5}$ in the corresponding area. Third, a population weighted scheme is applied to generate the PM$_{2.5}$ exposure assessment. Finally, the health impact of PM$_{2.5}$ is assessed with a concentration-response (C-R) function (Anenberg et al. 2010). Meanwhile, the exposure assessments are cross-compared under different model assumptions (Table 4) by which the impact of uncertainty caused by using different models can be fully revealed. Overall, Model 1 and 2 use low resolution (interpolated) PM$_{2.5}$ concentration. Model 1 and 3 use low resolution (county-level demographic) population data. Model 4 and 5 use both high resolution satellite-station-based PM$_{2.5}$ concentration data and population data. Specifically, Model 4 uses Landscan population data, and Model 5 (the proposed framework in the paper) uses improved LBS-based dynamic population data. This way, Model 5 can better fit the ground truth with high spatiotemporal resolution population and PM$_{2.5}$ concentration data. Based on the above points, the Model 5, which uses more accurate data (population distribution and pollutant concentration), is assumed to better evaluate the health impact. Therefore, the health impacts of other models are compared with the Model 5 estimates accordingly. If using the results of Model 5 as baseline, Model 1 and 2 overestimate premature deaths by 3309 and 4841 respectively, while Model 3 and 4 underestimate premature deaths by 6017 and 1881.

### 2.4 Urban land-use classification via deep learning

BSD can not only improve existing algorithms using conventional analytic frameworks as an additional data source, but can also be suitable for creating
analysis methods with more complex model structures (e.g., learning based models with thousands of parameters). This advantage is important when a study is targeting an urban area with high complexity, since in such scenarios traditional models tend to oversimplify the problems and provide limited interpretations (Albert et al. 2019; Şalap-Ayça et al. 2018). Urban land-use mapping yields critical basic data for environmental monitoring, urban planning, public health and other related urban studies concerns (Cao et al. 2011; Chen et al. 2018b; Huang 2017; Liu et al. 2015). However, even in this era of big data, traditional methods like filed surveys and aerial photos are still adopted as the major acquisition method of this kind of data. With the advance in satellite imagery techniques, emerging high-resolution remote sensing images with distinct spatial, temporal, spectral, angular and radiometric characteristics are becoming available. However, these advancements of satellite imagery haven’t been fully utilized for generating a high-quality land-use product, because algorithms that are compatible with these super high resolution multi-spectral images are still underdeveloped. Specifically, the high-quality images render one land parcel as a set of pixels with different “sub” land-use types. For instance, pixels representing trees, buildings, and water bodies are found in one residential area. The advancement of this data source has proceeded beyond the level of conventional analytics which use low-resolution satellite images, yielding a mix of objects in one pixel (i.e. the mixed pixel problem) (Anenberg et al. 2010).

As far as conventional algorithms are concerned, land-use classification methods can be categorized as three types: per-pixel (Wu et al. 2009; Zhao, Zhong, and Zhang 2016), object-based (Aksoy et al. 2005; Blaschke et al. 2014; Voltersen et al. 2014), and per-field (Hu and Wang 2013; Wu et al. 2009). In all of them, only shallow architectures are extracted from the images. Features are generated, selected or described manually. Deep architectures and high-level features are thus not well utilized in solving classification problems. To address this issue, the use of Deep Convolutional Neural Networks (DCNN) was proposed, which display significantly higher accuracy at land-use classification. By extracting deep architectures achieved with multiple nonlinear transformations (LeCun, Bengio, and Hinton 2015), high-level abstract features are generated from the original High Spatial Resolution (HSR) images (Jia, Liu, and Sun 2015). However, DCNN require a large number of training samples. To solve this problem, two improvements are made to the original DCNN. First is the transfer DCNN (Castelluccio et al. 2015; Hu et al. 2015), which decreases the required training sample size by embedding a pre-trained Convolutional Neural Network (CNN) into the new classification task. Second are small DCNNs that prevent overfitting by only using a few layers (Zhang, Du, and Zhang 2015). As far as the status quo is concerned, there are still three areas that can be improved in land-use classification. First, transfer DCNN only utilize gray or RGB channel information even for an HSR multispectral image. Second, small DCNNs can use all channels in a multispectral image but cannot take advantages of deep architectures. Third, conventional land-use classification methods evenly split HSR images with a uniform decomposition method. Land-use patterns are thus chopped. Consequently, the classified land-use mapping may lead to “false blocks”. To solve the abovementioned problems, a Semi-Transfer Deep Convolutional Neural Network (STDCNN) is proposed. The proposed network combines a transfer DCNN with a small DCNN. By doing so, advantages acquired from both the channels in HSR multispectral images and the high-level abstractions achieved through deep architecture can be well leveraged. This network includes three parts: (1) a transfer DCNN using the AlexNet Model; (2) a small DCNN with a few layers, and (3) a full connected layer and a softmax layer. Subsequently, a skeleton-based decomposition method is employed to fix the “false block” issue in the land-use mapping process.

To validate the effectiveness of the proposed model, two areas in Hong Kong and one area in Shenzhen were selected as the study area as shown in Table 5. The images were subtracted as 256 × 256 samples and later classified as different land-use types (Table 5). Both the images for Hong Kong and the ones for Shenzhen were taken in 2015.

The proposed land-use classification framework includes three parts, as shown in Figure 4 (a-c): (1) Data preprocessing; (2) STDCNN training and classification; (3) Land-use mapping.

In the first step, the HSR multispectral images and vector data (street blocks and roads) are projected into a UTM/WGS84 geo-referenced coordinate system. Following this, a skeleton-based decomposition (Voltersen et al. 2014; Zhang, Du, and Zhang 2015) is applied to decompose the HSR images into a set of training samples (images with a size equal to 256 × 256).

In the second step, the training samples are handled separately by a transfer DCNN and a small DCNN as shown in Figure 4 (c). Their results are later joined by a fully connected layer (FL) and a softmax layer. The transfer DCNN uses only the information from the RGB channels. Two fully connected layers (TL1 and TL2) are employed to convert the export of AlexNet (1 × 1 × 4096) to a (1 × 1 × 512) vector. The small DCNN uses the information from all channels. Three convolution layers (SL1, SL2, and SL3) and one fully connected layer (SL4) are employed. For each of the
convolution layers, a pooling layer is attached. This STDCNN network is trained through Back-Propagation (BP) learning (Rumelhart, Hinton, and Williams 1988). Finally, the network will export a vector representing the confidence level of L (eleven in this study) types land-use for a given image sample (process unit).

In the third step, a street block (mapping unit) is classified as one land-use type according to the achieved confidence level in the former steps. Specifically, the confidence levels of the multiple process units are weighted by the number of their overlapping pixels to one mapping unit and summed up. A maximum confidential rule is later applied to decide the final land-use type of this mapping unit.

This approach (STDCNN) has three improvements. First, the STDCNN can utilize both RGB channel information with deep structures (AlexNet) as well as information from multispectral images. Second, the STDCNN leverages pre-trained parameters from AlexNet which significantly speeds up the convergence of the whole model. Third, a skeleton-based decomposition is employed to enhance the mapping

| Area | I | II | III |
|------------------|------------------|------------------|------------------|
| Location         | Kowloon Peninsula and the northern edge of Hong Kong Island | Shatin in the New Territories | Nansha area      |
| Region           | Hong Kong        | Hong Kong        | Shenzhen         |
| Data Type        | WorldView-3      | WorldView-3      | WorldView-2      |
| Spatial resolution | 1.24 m          | 1.24 m          | 0.5 m            |
| Image size       | 9,472*9,728      | 4,352*4,352      | 14,546*17,361    |
| Spectral         | Coastline, blue, green, yellow, red, red edge, near infrared 1 and near infrared 2 (400–1040 nm) | Coastline, blue, green, yellow, red, red edge, near infrared 1 and near infrared 2 (400–1040 nm) | Blue, green, red, and near infrared |
| Land-use type     | Commercial, institutional, port, dense residential, sparse residential, woodland, water, open space, vacant, industrial and container terminal classes | Commercial, institutional, port, dense residential, sparse residential, woodland, water, open space, vacant, industrial and container terminal classes | Commercial, institutional, dense residential, sparse residential, woodland, open space, vacant, industrial, container terminal, and road classes |

Figure 4. STDCNN based land-use classification framework.
process of land-use types. By doing so, the “false blocks” problem can be solved.

3. Results

3.1. Detection of polycentric urban structures

In the descriptive analysis, as shown in Figure 5, the nighttime light imagery provides better spatial coverage, while the check-in data can better identify human activity in places such as harbors, which might be mistakenly identified as city centers if relying only on nighttime light data. In this way, conventional remote sensing data and BSD combined can provide us with a more comprehensive picture of polycentric urban structures.

Conducting a preliminary analysis following the proposed three steps detection framework, 927, 866, and 558 segments are generated for three study areas (Shanghai, Beijing, and Chongqing) respectively. The segments are denser and more fragmented in plains and places with more human activity. Conversely, mountainous areas with little human activity are represented in larger segments. For delineating main and sub-centers, the LMI approach outperforms the conventional threshold one in identifying main centers. However, the situation is more complex when comparing the results of sub-centers identification between the OLS and GWR-based approaches. The detected centers differ considerably, more sub-centers being detected in places distant from the main center with the OLS approach. Meanwhile, the OLS analysis identifies some detected sub-centers as part of the main center. This may occur because the OLS has difficulties handling spatial heterogeneity in a local model, and consequently causes more outliers than the GWR result.

Given these preliminary results, two model evaluation steps are performed to further validate the effectiveness of the proposed model. First, the detection results

![Figure 5. Night time light data and Check-in data.](image)
(threshold-based, LMI cluster/OLS based, and LMI cluster/GWR based) are compared amongst each other according to their correspondence to human daily life. For this, human daily life is represented by Google Maps POI (https://developers.google.com/places webservice intro). The LMI cluster/GWR achieves the highest Kappa coefficient in 16 (out of 17) types of POI points, meaning this method can well identify the places with concentrated human activity. Second, the detected main and sub-centers are compared with the Master plan of the three cities respectively, shown in Table 6. The main center and all sub-centers mentioned in the Master plan are well recognized by this approach. Meanwhile, some emerging centers are also detected, showing the advantages of motoring urban development with BSD.

3.2. Evaluation of urban vibrancy

Urban vibrancy shows a distinctive variation across different surface attributes in Shanghai. The CUVI, as a linear combination of SI, EI, and PD, accounts for 75.56% of the commonalities among surface attributes. The distribution of the CUVI is heavy-tailed, which indicates a significant agglomeration in terms of urban vibrancy in Shanghai. Meanwhile, by mapping the CUVI in Shanghai, its polycentric structure can be easily identified, as shown in Figure 6. Urban vibrancy (measured with the CUVI) is characterized by UBEs using 10-fold OLS. The result shows that UBEs can well explain the variation in the CUVI (with a training $R^2$ equal to 0.764, and a testing $R^2$ equal to 0.760). In this sense, UBEs are significantly associated with urban vibrancy. Further, among all UBE indicators, building density and density of urban functions have the strongest connection with urban vibrancy.

3.3. Estimation of population exposure to PM$_{2.5}$

This approach proposed a robust PM$_{2.5}$ exposure assessment model using satellite-station-hybrid based PM$_{2.5}$ measurements and LBS-based dynamic population data. The monthly exposure levels, as well as the health outcomes, were evaluated in thirteen cities in 2015. According to the results from this study, about half of the population was exposed to PM$_{2.5}$ levels higher than 80 µg/m$^3$ in 2015. The situation is worst in December, 138,150, 80,945, and 18,752 premature deaths can be attributed to multi-cause, cardiovascular, and respiratory diseases respectively due to PM$_{2.5}$ exposure estimated by the relative risk model (He and Huang 2018a; Anenberg et al. 2010). In the comparison of models, the dynamic population distribution and the high spatiotemporal PM$_{2.5}$ estimations significantly affect the assessment result. Conventional methods, like using station-interpolated PM$_{2.5}$ concentration and pixel-based population data, yield measurements with substantial discrepancies to the proposed method.

3.4. Urban land-use classification via deep learning

Two evaluation criteria (Kappa coefficients and overall accuracy) are introduced to, using the same datasets (Hong Kong and Shenzhen), compare the proposed STDCNN to four other classification methods: spatial pyramid co-occurrence image classification (Yang and Newsam 2011); scene classification based on a latent Dirichlet allocation (Zhao, Zhong, and Zhang 2013); a transfer DCNN; and a small DCNN. Overall, the STDCNN achieves the highest performance in both Kappa coefficients (Hong Kong: 0.903; Shenzhen: 0.780) and overall accuracy (Hong Kong: 91.25%; Shenzhen: 80.00%). Further, the F-test results show that the improvement of the STDCNN over all other methods are significant at the 0.05 confidence level except for the transfer DCNN in Hong Kong (p-value: 0.0801) and the small DCNN in Shenzhen (p-value: 0.0736).

Meanwhile, the STDCNN achieves better results in identifying land-use types like dense residential, institutional, open space, and road areas by comparing the confusion matrices across all models. From a training perspective, the STDCNN testing curve stays above all other methods. The STDCNN and the transfer DCNN with pre-trained parameters can converge faster than the small DCNN. From a mapping perspective, there are two major advantages achieved by the STDCNN. First, skeleton-based decomposition can well preserve land-use patterns which are falsely split by conventional uniform decomposition. Second, the constitution of both major and minor land-use types are identified for each street block by using skeleton-based decomposition. Here, the confidence level of each

| Table 6. Detection accuracy assessment. |
|----------------------------------------|
| **City** | **Main center** | **Emerging center** | **Accuracy** |
|---------|-----------------|---------------------|--------------|
| Beijing | Dongcheng, Xicheng, Chaoyang, Fengtai, Shijingshan, Haidian, Shunyi, Pingguo, Huairou, Changping, Yizheng, Daxing, Fangshan, Mentougou, Tongzhou | Tongzhou Airport, Huanggezhuang | 88.2% |
| Shanghai | Huangpu, Xuhui, Changning, Jingan, Putuo, Hongkou, Yangpu | Haidian, Anting, Caolu | 83.3% |
| Chongqing | Baoshan, Jiading, Songjiang, Jinshan, Minhang, Nanqiao, Huinan, Nanhui, Airport, Yuzhong, Dadukou, Jiangbei, Shapingba, Jiulongpo, Nanan, Yubei, Banan, Beibei | Hechuan, Nanchuan, Fuling, Changshou, Longxing | 86.7% |

**Note:** Accuracy is calculated using Kappa coefficients at the 0.05 confidence level except for the transfer DCNN in Hong Kong (p-value: 0.0801) and the small DCNN in Shenzhen (p-value: 0.0736).
land-use type can be used to mimic the percentage of corresponding land-use type in one given mapping unit (street block). Compared with the conventional pixel-based land-use mapping approach, skeleton-based decomposition can generate land-use maps better matching local land-use units.

Monitoring land-use changes is of great importance in urban planning. Conventional survey approaches are accurate but severely limited by spatial coverage, data update frequency (temporal resolution), and cost issues. This approach introduces an alternative way to generate land-use maps with both high spatiotemporal resolution and low costs. Deep learning techniques are employed and show a strong capability to understand complex urban patterns by modeling highly abstract features with deep structures and multispectral images. Given that complexity is a common issue in urban systems, findings from this study also imply a great potential to model urban systems with deep structures.

4. Discussion

This paper briefly discussed emerging BSD and its related analytics with focus on the opportunities and challenges it brings. Four related studies were described. The first and second studies exemplify how to incorporate BSD to understand the urban in finer spatiotemporal scales (with higher spatiotemporal resolution). The third and fourth studies show new ways to incorporate BSD to generate critical base data (PM$_{2.5}$ concentration and land-use maps) for urban and environment studies. The four reviewed studies evidence that BSD can significantly improve documentation methods, which can greatly facilitate developing solutions in addressing both urban and environmental problems. From an analytics perspective, new frameworks can be designed or reworked based on a previous “small data” version. In this process, some future directions emerge as follows:

Figure 6. CUVI in Shanghai.
First, data fusion (data integration). Objects are observed by multiple sensors, but typically, for each of the sensors, the information captured about the observed objects is typically partial or inconsistent in terms of the quality, coverage, accuracy, and timeliness (Dong and Srivastava 2013). The information captured by each single sensor individually may thus be missing or incomplete for a certain analysis purpose. This missing data issue may be common in a BSD context, where the raw BSD may fail to match highly diverse analytic purposes. Thus, a data fusion process is necessary for obtaining the full picture. The output, the fused dataset, typically has better (spatial, temporal, spectral, and radiometric) resolution and coverage. For instance, pedestrians’ location information can be observed by LBS position requests, roadside cameras, social media posts and population survey data. The population in one region can be exactly estimated by population census data, but with low spatial (administrative unit based) and temporal (5 or 10 years) resolution. However, high spatiotemporal resolution dynamic population data can be created by disaggregating the census population of one administrative unit to gridded pixel level with LBS position requests or social media post count information.

Second, data and predictive analytics. Conventional analytic frameworks (e.g. linear regression and autoregressive integrated moving average models) have limited capacity to fully leverage the massive amount of information from BSD. The underlying model of BSD is usually of high complexity (heterogeneity and nonlinearity). Meanwhile, conventional algorithms might be computationally inefficient under a BSD context. Therefore, advanced data and predictive analytics for urban and environmental applications should be developed. The analytics should be capable of capturing non-linear and heterogeneous characteristics with high computational efficiency.

Third, theoretical models for BSD. To deal with unstructured data and rapidly emerging problems, “black-box” machine learning methods (e.g. deep learning) are widely adopted. However, theoretical models are typically missing in such approaches. Most of the problems are solved at the application level but not at the theoretical level. Hence, solid theoretical models might be necessary for a deep understanding of “black-box” models. Specifically, emerging techniques, like explainable artificial intelligence (Gunning 2017) and interpretable machine learning (Samek, Wiegand, and Müller 2017), could be of great importance to improve the BSD analytics.

Despite the advantages of using BSD, there are still two limitations found in the reviewed studies. The first issue is scale dependence. BSD analytics include multiple (spatial and temporal) scale data sources with more complex and discrepant data acquisition procedures than traditional ones (Renslow 2012; Hariri, Fredericks, and Bowers 2019). Therefore, abundant (dis)aggregation measures are applied to handle BSD with varying resolution and spatiotemporal units, which potentially causes modifiable areal unit problems (Openshaw and Taylor 1979). Compared with frameworks using only traditional data sources with relatively simple data acquisition and preprocessing processes, it is more challenging to generalize BSD approaches to other scales. The second issue is the interpretation. BSD causes more problems in representativeness, coverage bias, metrology, and data quality (Cox, Kartsonaki, and Keogh 2018). The complex structure of the data also demands more non-linear and “black-box” learning models, which are less interpretable (Beam and Kohane 2018). Consequently, it is difficult to interpret BSD approaches with conventional statistical measures, especially when ground truth data is missing. For instance, the ground truth of the dynamic population investigated in the third study is theoretically unavailable.

5. Conclusion

BSD has significantly changed urban and environmental sustainability studies today. Emerging data sources advanced conventional studies (e.g. polycentric urban structures and PM2.5 exposure) with better spatiotemporal resolution and lower cost. Meanwhile, new analytics (e.g. deep learning methods) suit BSD are also developed. Despite limitations such as scale dependence and interpretation ability, BSD has shown great potential in the reviewed studies. Future directions, such as data fusion, data and predictive analytics, and theoretical models for BSD are also identified in this study.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Bo Huang is a Professor with the Department of Geography and Resource Management, The Chinese University of Hong Kong, where he is also the Associate Director of Institute of Space and Earth Information Science and the Director of MSc Program in Geoinformation Science. Prior to this, he held faculty positions at University of Calgary (Geomatics Engineering), Canada and National University of Singapore (Transportation). His research interests are broad, covering most aspects of Geographical Information Science (GIScience), specifically: satellite image fusion for environmental monitoring, spatial/spatio-temporal statistics for land cover/land use change modeling, spatial optimization for sustainable urban and land use planning, Intelligent Transportation Systems (ITS), and web/wireless GIS for location-based services.
Bo Huang is currently a postdoctoral fellow in the Department of Geography and Resource Management, The Chinese University of Hong Kong. He received his Ph.D. degree from the CUHK (HK, China). His research interests include spatial social/temporal modeling and land-use optimization.

**ORCID**

Bo Huang [http://orcid.org/0000-0002-5063-3522](http://orcid.org/0000-0002-5063-3522)

Jionghua Wang [http://orcid.org/0000-0001-8265-3985](http://orcid.org/0000-0001-8265-3985)

**References**

Aksoy, S., K. Koperski, C. Tuske, G. Marchisio, and J. C. Tilton. 2005. “Learning Bayesian Classifiers for Scene Classification with a Visual Grammar.” *IEEE Transactions on Geoscience and Remote Sensing* 43 (3): 581–589. doi:10.1109/TGRS.2004.839547.

Albert, A., J. Kaur, E. Strano, and M. Gonzalez. 2019. “Spatial Sensitivity Analysis for Urban Land Use Prediction with Physics-constrained Conditional Generative Adversarial Networks.” *arXiv Preprint* arXiv:1907.09543.

Andreas, M. K., and M. Haenlein. 2009. “The Challenges and Opportunities of Social Media.” *ESCP Europe* 79: 59–68.

Anenberg, S. C., L. W. Horowitz, D. Q. Tong, and J. J. West. 2010. “An Estimate of the Global Burden of Anthropogenic Ozone and Fine Particulate Matter on Premature Human Mortality Using Atmospheric Modeling.” *Environmental Health Perspectives* 118 (9): 1189–1195. doi:10.1289/ehp.0901220.

Anselin, L. 1995. “Local Indicators of Spatial Association—LISA.” *Geographical Analysis* 27 (2): 93–115. doi:10.1111/j.1538-4632.1995.tb00338.x.

Bai, X., P. Shi, and Y. Liu. 2014. “Society: Realizing China’s Urban Dream.” *Nature News* 509 (7499): 158. doi:10.1038/509158a.

Batty, M. 2013. “Big Data, Smart Cities and City Planning.” *Dialogues in Human Geography* 3 (3): 274–279. doi:10.1177/2043820613513390.

Beam, A. L., and I. S. Kohane. 2018. “Big Data and Machine Learning in Health Care.” *Jama* 319 (13): 1317–1318. doi:10.1001/jama.2017.18391.

Bengtsson, L., X. Lu, A. Thorson, R. Garfield, J. Von Schreeb, and P. W. Gething. 2011. “Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-earthquake Geospatial Study in Haiti.” *PLoS Medicine* 8 (8): e1001083. doi:10.1371/journal.pmed.1001083.

Blaschke, T., G. J. Hay, M. Kelly, S. Lang, P. Hofmann, E. Addink, R. Q. Feitosa, F. Van der Meer, H. Van der Werff, and F. Van Coillie. 2014. “Geographic Object-based Image Analysis—towards a New Paradigm.” *ISPRS Journal of Photogrammetry and Remote Sensing* 87: 180–191. doi:10.1016/j.isprsjprs.2013.09.014.

Bogomolov, A., B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland. 2014. “Once upon a Crime: Towards Crime Prediction from Demographics and Mobile Data.” In *Proceedings of the 16th international conference on multimodal interaction*, 427–434, ACM, Istanbul, Turkey.

Brenner, N., ed. 2014. *Implosions/Explosions: Towards a Study of Planetary Urbanization*. Berlin: Jovis.

Brenner, N., J. Peck, and N. Theodore. 2010. “Variegated Neoliberalization: Geographies, Modalities, Pathways.” *Global Networks* 10 (2): 182–222. doi:10.1111/j.1471-0374.2009.00277.x.

Cai, J., B. Huang, and Y. Song. 2017. “Using Multi-source Geospatial Big Data to Identify the Structure of Polycentric Cities.” *Remote Sensing of Environment* 202: 210–221. doi:10.1016/j.rse.2017.06.039.

Cao, K., M. Bath, B. Huang, Y. Liu, L. Yu, and J. Chen. 2011. “Spatial Multi-objective Land Use Optimization: Extensions to the Non-dominated Sorting Genetic algorithm-II.” *International Journal of Geographical Information Science* 25 (12): 1949–1969. doi:10.1080/11773489.2011.570269.

Castelluccio, M., G. Poggi, C. Sansone, and L. Verdoliva. 2015. “Land Use Classification in Remote Sensing Images by Convolutional Neural Networks.” *arXiv Preprint* arXiv:1508.00092.

Chabrier, S., B. Emile, C. Rosenberger, and H. Laurent. 2006. “Unsupervised Performance Evaluation of Image Segmentation.” *EURASIP Journal on Applied Signal Processing* (2006): 217.

Chen, B., B. Huang, and B. Xu. 2017. “Multi-source Remotely Sensed Data Fusion for Improving Land Cover Classification.” *ISPRS Journal of Photogrammetry and Remote Sensing* 124: 27–39. doi:10.1016/j.isprsjprs.2016.12.008.

Chen, B., Y. Song, M.-P. Kwan, B. Huang, and B. Xu. 2018b. “How Do People in Different Places Experience Different Levels of Air Pollution? Using Worldwide Chinese as a Lens.” *Environmental Pollution* 238: 874–883. doi:10.1016/j.envpol.2018.02.093.

Chen, B., Y. Song, T. Jiang, Z. Chen, B. Huang, and B. Xu. 2018a. “Real-time Estimation of Population Exposure to PM2.5 Using Mobile-and Station-based Big Data.” *International Journal of Environmental Research and Public Health* 15 (4): 573. doi:10.3390/ijerph15040573.

Chen, H., R. H. Chiang, and V. C. Storey. 2012. “Business Intelligence and Analytics: From Big Data to Big Impact.” *MIIS Quarterly* 36 (4): 1165. doi:10.2307/41703503.

Cohen, A., H. Anderson, B. Ostro, K. Pandey, M. Krzyzanowski, N. Kunzli, K. Gutschmidt, C. Pope, I. Romieu, and J. Samet. 2001. “Urban Air Pollution: Comparative Quantification of Health Risks.” *World Health Organization*.

Couture, V. 2013. “Valuing the Consumption Benefits of Urban Density,” *University of California, Berkeley. Processed*.

Cox, D., C. Kartsonaki, and R. H. Keogh. 2018. “Big Data: Some Statistical Issues.” *Statistics & Probability Letters* 136: 111–115. doi:10.1016/j.spl.2018.02.015.

Dale, A., C. Ling, and L. Newman. 2010. “Community Vitality: The Role of Community-level Resilience Adaptation and Innovation in Sustainable Development.” *Sustainability* 2 (1): 215–231. doi:10.3390/su2010215.

De Nadai, M., J. Staiano, R. Larcher, N. Sebe, D. Quercia, and B. Lepri. 2016. “The Death and Life of Great Italian Cities: A Mobile Phone Data Perspective.” In *Proceedings of the 25th international conference on world wide web*, 413–423, International World Wide Web Conferences Steering Committee, Montreal, Canada, April 11 - 15.

Dong, X. L., and D. Srivastava. 2015. “Some Statistical Issues.” *Statistics & Probability Letters* 124: 1245–1248, IEEE, Brisbane, Australia, April 8–12.

Espindola, G., G. Câmara, I. Reis, L. Bins, and A. Monteiro. 2006. “Parameter Selection for Region-
