*Article*

B Planning and Design

Environment and Planning B:
Planning and Design
2016, Vol. 43(2) 341–360

DOI: 10.1177/0265813515604767
epb.sagepub.com
SAGE

# Automated identification and characterization of parcels with OpenStreetMap and points of interest

**Xingjian Liu**
The University of Hong Kong, Hong Kong

**Ying Long**
Tsinghua University and Beijing Institute of City Planning, China

**Abstract**
Against the paucity of information on urban parcels in China, we propose a method to automatically identify and characterize parcels using OpenStreetMap (OSM) and points of interest (POI) data. Parcels are the basic spatial units for fine-scale urban modeling, urban studies, and spatial planning. Conventional methods for identification and characterization of parcels rely on remote sensing and field surveys, which are labor intensive and resource consuming. Poorly developed digital infrastructure, limited resources, and institutional barriers have all hampered the gathering and application of parcel data in China. Against this backdrop, we employ OSM road networks to identify parcel geometries and POI data to infer parcel characteristics. A vector-based cellular automata model is adopted to select urban parcels. The method is applied to the entire state of China and identifies 82 645 urban parcels in 297 cities. Notwithstanding all the caveats of open and/or crowd-sourced data, our approach can produce a reasonably good approximation of parcels identified using conventional methods, thus it has the potential to become a useful tool.

**Keywords**
OpenStreetMap (OSM), points of interest (POI), land parcel, automatic generation, urban planning

## Introduction

Land-parcel data are one of the cornerstones of contemporary urban planning (Cheng et al., 2006).[1] Parcels are the basic spatial units of urban models; for example, the latest urban simulation models are oftentimes vector based and capture parcel-level dynamics (Pinto, 2012; Stevens and Dragicevic, 2007). More importantly, normative planning and

**Corresponding author:**
Ying Long, Beijing Institute of City Planning, No. 60, South Lishi Rd, Beijing 100045, China.
Email: longying1980@gmail.com

policies are performed using parcels, ranging from the formulation of comprehensive plans, to strategic plan implementation, and to policy evaluation (Alberti et al., 2007; Frank et al., 2006; Jabareen, 2006).

Whereas parcel data for the developed world are generated by robust digital infrastructure and supplemented by open data initiatives, for example, OpenStreetMap (OSM), researchers still lament the difficulty of attaining parcel data for developing countries. For example, the best available parcel map for China's capital, Beijing — supposedly one of the most technologically advanced and rapidly developing cities in the erstwhile Third World — is dated 2010 (Beijing Institute of City Planning, 2010). In addition, collecting parcel data in medium-sized and small- sized cities in China is constrained by poorly developed digital infrastructure. It goes without saying that complete parcel-level features (eg, land-use type, urban functions, and development density) often are hard to be found. In addition to hard infrastructures, soft institutions have also created barriers for Chinese urban planners' access to parcel maps. For instance, our interviews with fifty-seven planning professionals[2] suggest that access to existing parcel maps held by local planning bureaus and institutes is restrained, as parcel maps are tagged as confidential within the current Chinese planning institutions. In summary, parcel data for the developing world are oftentimes outdated, limited in geographical scope, and do not contain rich thematic information other than basic parcel geometry. As parcel data are at the center stage of urban planning (Cheng et al., 2006), the lack of parcel data would constrain our ability to trace urban changes at high spatial resolution, hinder the formulation and implementation of detailed urban plans, and restrain the possibility of adopting contemporary parcel-based urban management.

Built on manual interpretation of remote sensing images and field surveys, conventional ways of generating parcel data are time consuming, expensive, and labor intensive (Erickson et al., 2013). Thus, many developing countries do not have the necessary capital and resources to produce parcel data in the conventional fashion. Overcoming such data paucity seems to be a high priority for urban planning in developing countries.

Against this backdrop, we propose a method for automated identification and characterization of parcels (AICP), based on ubiquitous OSM and points-of-interest (POI) data. The proposed method could (1) provide quick and robust delineation of land parcels, and (2) generate a variety of parcel-level attributes, allowing the examination of urban functions, development density, and mixed land uses. We illustrate the usefulness of our method in a case study with 297 Chinese cities. In addition, the framework could work well with conventional data sources (eg, survey data), when these are available. In the next section we review the progress in obtaining parcel-level geometry and features, followed by an elaboration of methods and the case study. We conclude with a discussion of the strength and limitations, as well as future applications of our method.

## Identification and characterization of parcels

Parcel boundaries and their features are conventionally identified through manual interpretation of topographic maps, building plans, field surveys, and high-resolution remote sensing images (Cheng et al., 2006). Such manual operations could suffer from a series of problems. Firstly, manual operations are resource intensive and time consuming. For example, it would take an experienced operator 3 to 5 hours to identify and infer land use for 35 to 50 urban parcels covering the area of one square kilometer. Secondly, manual operations often result in an inconsistent dataset, as data quality largely depends on the experience and technical proficiency of individual practitioners. Thirdly, conventional

methods are less suitable for routine updates and longitudinal comparison. This becomes more problematic in the face of volatile urban development (eg, gentrification and urban sprawl) in many developing countries. Still, even though manual operations can identify parcel geometries, the generated parcels usually lack parcel-level information, such as density and land-use mix. As a case in point, data about parcel density for Beijing are limited to the area within the sixth ring road (approximately 13.8% of the Beijing Metropolitan Area).

Recent attempts have been made to automatically identify parcel geometries. For example, Yuan et al. (2012) proposed a raster-based approach for parcel delineation based on taxi trajectories and POIs. However, they omitted road space in the delineation of parcels, and the raster-based nature of the method generates computational burden, thus limiting the method's applicability to large geographical areas. Meanwhile, Aliaga et al. (2008) advanced an algorithm for interactively synthesizing parcel layouts for to-be-developed areas based on the structure of real-world urban areas. Their study however does not account for parcel characteristics, and performs parcel subdivision within predefined blocks, instead of identifying blocks from the data.

In light of this situation, OSM has been proposed as a promising candidate for quick and robust delineation of parcels and other urban features (Haklay and Weber, 2008; Over et al., 2010; Ramm, 2010). As one of the most successful volunteered GIS projects, OSM provides street-network data for a wide array of cities (Goodchild, 2007; Sui, 2008). Jokar Arsanjani et al. (2013a) predicted that the data coverage and quality of OSM will continue to be improved in the coming years. More specifically, the quality of OSM data in well-mapped and oftentimes large cities is on a par with that of topographic maps (Girres and Touya, 2010; Haklay, 2010; Over et al., 2010). The growth of OSM in developing countries has been encouraging, as the volume of OSM data in China has experienced a nine-fold increase during 2007–2013 (Figure 1).

Several preliminary studies suggest that OSM road networks are useful in identifying urban structures. For example, Hagenauer and Helbich (2012) extracted urban built-up areas from OSM, and Jiang and Liu (2012) identified natural grouping of city blocks

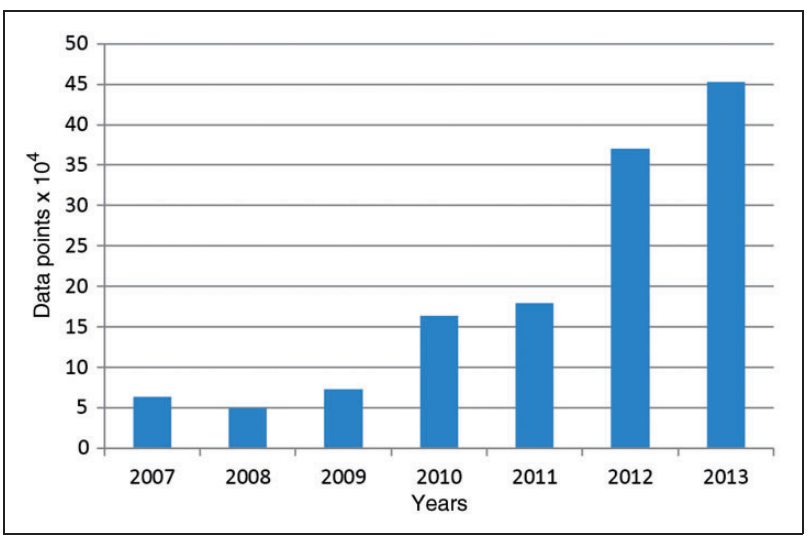


**Figure 1.** The increasing data volume in the OpenStreetMap (OSM)-China data (Unit: data points).

based on OSM data. Existing analyses using OSM focus more on deriving universal laws and social physics (Jiang and Liu, 2013) rather than producing data products for urban planning and studies. In a similar vein, Jokar Arsanjani et al. (2013b) identified land-use patterns for central Vienna, Austria (roughly 32 $km^2$) using OSM. Whereas Jokar Arsanjani et al. (2013b) introduced a volunteered geographic information-based method to generate land-use patterns, their approach focuses on developed countries with high-accuracy OSM data, uses predefined boundaries for parcel subdivision, and could be extended to generate additional parcel features.

As a corollary, it is understandable that researchers show hesitation in applying OSM-China data, as the data quality is not always clear and more metadata may be needed. In light of this situation, attempts have been made to assess the quality of Chinese OSM data (Zheng and Zheng, 2014). Preliminary results suggest that inter alia (1) OSM data are more dense and applicable in major cities, although the overall data coverage may be poor; (2) data completeness remains the major issue (Goodchild, 2008); (3) OSM data in China have been continuously improved. In line with observations, OSM data are gradually being used to understand Chinese cities (see, for example, Leitte et al., 2012; Liu et al., 2012; Zhang et al., 2013).

In addition to parcel geometries, planning practices also require parcel features such as urban functions and development density. There is a rich literature on inferring land use from remote sensing images (Herold et al., 2002; Kressler et al., 2001). However, as discussed previously, remote sensing images are not suitable for large-scale parcel-level analysis, due to inter alia data availability and computational burden. Although some automatic or semiautomatic techniques have been developed to address urban land-use classification (Herold et al., 2002; Pacifici et al., 2009), it is still difficult to identify certain landuse types, such as high-density residential areas and commercial areas, from remote sensing images. More importantly, remote-sensing-based methods often treat parcels as having homogenous land-use types, thus not allowing for mixed land use. More recently, researchers have inferred human use of urban space from human mobility data, such as smart card records (Long et al., 2013), mobile phone data (Soto and Frias-Martinez, 2011; Toole et al., 2012), and taxi trajectories (Liu et al., 2012b; Yuan et al., 2012). Nevertheless, human mobility data are often proprietary and involve privacy issues (Beresford and Stajano, 2003). Such limited data access greatly undermines the wide applicability of human-mobility-based methods.

To this end, we propose POI data as an alternative data source for characterizing parcels. The strength of POI data includes (1) containing subparcel- level business information, which could serve as proxies for land use and urban functions; (2) being available from online mapping and cataloguing service providers; (3) having a nearly global coverage; and (4) having high spatial (eg, geocoded business locations) and temporal (eg, routinely updated by service providers) resolution. In addition, while some POI data are proprietary and require small access fees, many are freely available from online business catalogues. With all these advantages, POI data seem to have great potential for characterizing parcel features.

Therefore, we propose a preliminary automatic process to supplement existing approaches to the identification and characterization of fine-scale urban land parcels. OSM data are used to identify and delineate parcel geometries, while POIs are gathered to infer land-use intensity, function, and mixing at the parcel level. With all possible caveats in mind (which will be elaborated in the concluding section), we emphasize that our empirical framework (1) is automatic and extensible, allowing for the incorporation of various data sources (eg, taxi trajectories, mobile phone data, as well as conventional

survey data); (2) produces not only parcel geometry and land-use types but also useful parcel-level information such as land-use mix; (3) is applicable to large geographic areas, while most previous studies are limited to small areas; and (4) enables routine updates and free distribution of urban parcel data for China. While enjoying the ubiquitous availability and other blessings of OSM data, in the application of our method it is necessay to be aware of the caveats of crowd-sourced and/or open data (Elwood et al., 2012; Neis et al., 2012; Sui et al., 2013).

## Data and methods

### *Data*

*Administrative boundaries of Chinese cities.* Our analysis covers a total of 654 Chinese cities (Figure 2),[3] ranging across five administrative levels: municipalities directly under the entral government (MD, 4 cities), subprovincial cities (SPC, 15), other provincial capital cities (OPCC, 16), prefecture-level cities (PLC, 251), and county-level cities (CLC, 368) [MOHURD (2013); see Ma (2005) for a more detailed discussion on the Chinese administrative system). As a city *proper* in China contains both rural and urban land uses, we narrow our analytical scope onto legally defined urban land within each city proper and use administrative boundaries of urban lands to carve out OSM and POI data layers. In addition to administrative boeundaries, we also gather information about the total built-up area of individual cities in 2012 (MOHURD, 2013), which will be used in the urban parcel identification process.

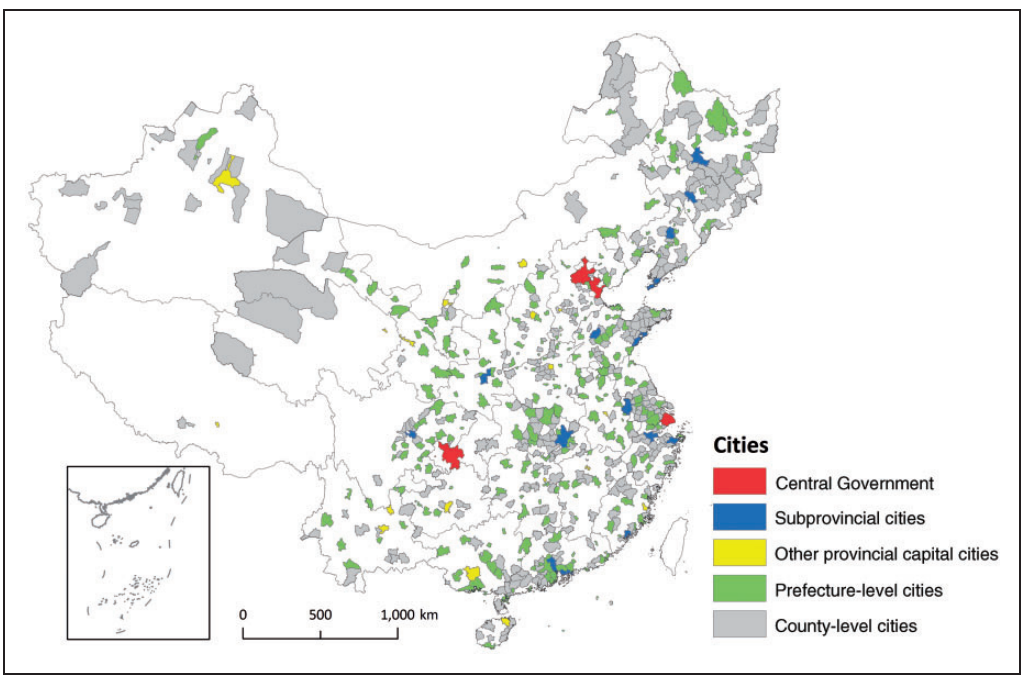


**Figure 2.** Administrative boundaries of Chinese cities.

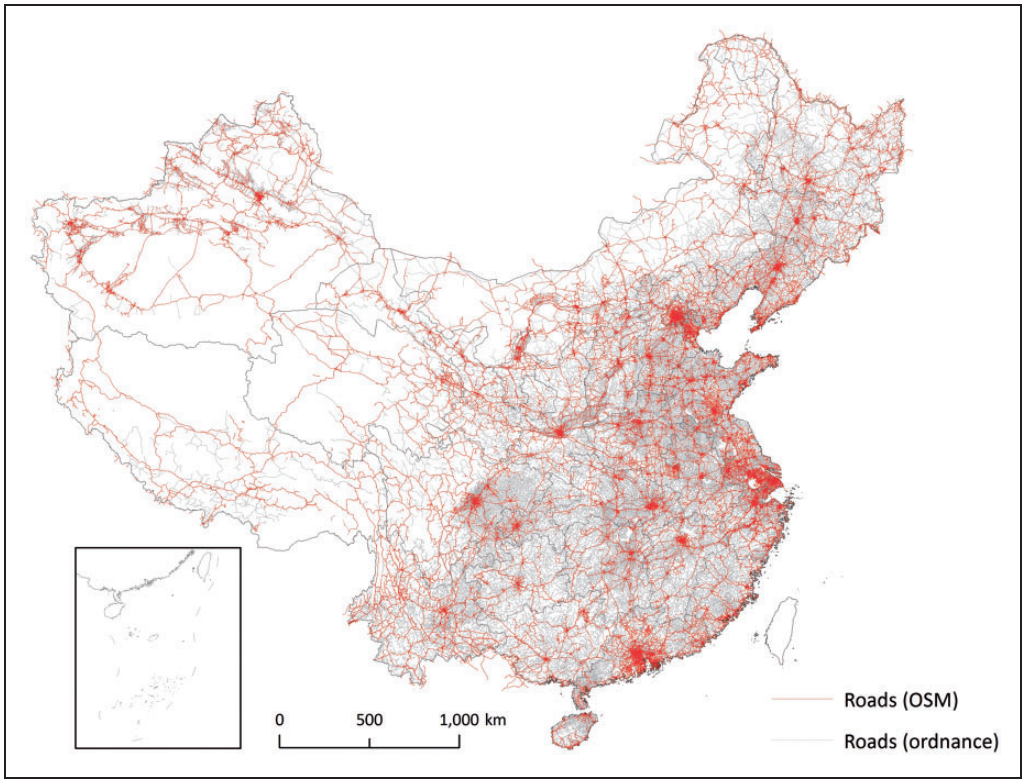


**Figure 3.** A comparison of roads in the OpenStreetMap (OSM) and survey map.[4]

*OSM in China.* OSM road networks for China were downloaded on 5 October 2013. We also amassed the Chinese equivalent of Ordnance Survey data (hence the use of 'ORDNANCE' as the abbreviation) at the end of 2011 with detailed road networks in order to verify results produced by the OSM data. The OSM dataset contains 481 647 road segments (8.0% of that of the survey map) of 825 382 kilometers (31.5% of the survey map). Furthermore, road networks in OSM and the survey map were overlaid for a visual inspection of data quality (Figure 3).[4] The preliminary completeness check is consistent with previous findings on OSM data quality in China (Zheng and Zheng, 2014). In spite of capturing a portion of the survey map, OSM data cover most urban areas in China, especially large cities (Figure 3), and are potentially useful for identifying urban land parcels. The implication of OSM data quality will be elaborated in subsequent sections.

*POIs.* A total of 5 281 382 geotagged POIs were synthesized from an online Chinese business catalogue synthesized by Sina Weibo. The initial twenty POI types were aggregated into eight more general categories (Table 1): commercial sites account for most POIs, followed by office building/space, transportation facilities, and government buildings. POI labeled as 'others' are used in estimating land-use density, but removed from land–use-mix analyses as POI of this type are not well classified and would conduce to the estimation of the degree of mixed land use. We manually checked randomly sampled POI data points to gauge and ensure the overall data quality. Our empirical framework is extensible in the sense that POI

**Table 1.** Points-of-interest types and aggregated information.

| Type | Abbreviation | Counts |
|---|---|---|
| Commercial sites | COM | 2 573 862 |
| Office building/space | OBS | 677 056 |
| Transport facilities | TRA | 561 236 |
| Others | OTH | 534 357 |
| Government | GOV | 468 794 |
| Education | EDU | 285 438 |
| Residence communities | RES | 167 598 |
| Green space | GRE | 13 041 |

counts can be replaced by other human activity measurements, ranging from the more conventional landuse cover derived from remote-sensing images to online check-in service data (eg, Foursquares).

*Other data.* DMSP/OLS [1 km spatial resolution; Yang et al. (2013)] and GLOBCOVER [300 m spatial resolution; Bontemps (2009)] remote sensing images were obtained for model validation, as we compare parcels identified by our empirical framework with those generated from remote sensing images. In addition, for benchmarking purposes, manually generated parcel data for Beijing were gathered from Beijing Institute of City Planning (BICP). In our analyses, we bear in mind that the spatial resolutions of parcels generated by different methods may differ.

## Methods

*Delineating parcel boundaries.* The working definition of a parcel in our framework is a continuously built-up area bounded by roads. Identifying land parcels and delineating road space are therefore *dual* problems. In other words, our approach begins with the delineation of road space, and individual parcels are formed as polygons bounded by roads.

The delineation of road space and parcels is performed as follows: (1) All OSM road data are merged as line features into a single data layer; (2) individual road segments are trimmed with a threshold of 200 m to remove hanging segments; (3) individual road segments are then extended at both ends for 20 m to connect nearby but topologically separated lines; (4) road space is generated as buffer zones around road networks. A varying threshold ranging between 2 m and 30 m is adopted for different road types, considering both surface conditions and different levels of roads (eg, national highways and local streets); (5) parcels are delineated as the space left when road space is removed; and (6) a final step involves overlaying parcel polygons with administrative boundaries to determine which individual parcels belong to which cities. Parameters used in these steps are determined pragmatically and based on practitioners' experiences, while bearing in mind the topological errors in OSM data.

*Calculating density for all parcels.* We define land-use density as the ratio between the counts of POIs in or close to a parcel to the parcel area. We further standardized

the density to range from 0 to 1 for better intercity and intracity density comparison using equation (1):

$$d = \frac{\log d_{\text{raw}}}{\log d_{\text{max}}} \tag{1}$$

where $d$ is standardized density, $d_{\text{raw}}$ and $d_{\text{max}}$ correspond to the density of individual parcels and the nation-wide maximum density value.[5] As mentioned previously, other measures (eg, online checkins and floor-area ratio) can substitute POIs and approximate the intensity of human activities.

*Selecting urban parcels.* The next step selects urban parcels from all generated parcels. The total urban land of individual cities was gathered from MOHURD (2013). We employ a vector-based constrained cellular automata (CA) model to identify urban parcels in individual cities[6] (Zhang and Long, 2013). More specifically, we use the CA model to predict individual parcels' possibility of being urban, and the total urban land is used as constraints for the aggregated amount of urban parcels.

In the CA model, each parcel is regarded as a cell, and the cell status was 0 (urban) or 1 (nonurban). Essentially, the CA model simulates the urban development. At the onset of the simulation, all cells are set to be rural. At each step during the simulation, whether a parcel is converted to 'urban', that is, the probability of being urban, depends on two factors (Li and Yeh, 2002): firstly, the proportion of neighboring parcels that are urban; secondly, individual parcels' attributes such as size, compactness, and the POI density. In our empirical operationalization, the neighborhood of a parcel includes all parcels within a 500 m radius. These three attributes are combined using a logit function to influence individual parcels' probability of being urban (Wu, 2002). We then multiply the two factors (neighborhood and parcel attributes) to determine whether the final probability is greater than a predefined threshold. In other words, a parcel surrounded by many urban parcels and with certain attributes would have more chance of being identified as an urban parcel in the simulation. Figure 4 provides a visual illustration of our CA model, where the final probability of being selected as an urban parcel for parcels A, B, and C is 0.6 ($0.8\times6/8$), 0.3 ($0.6\times4/8$), and 0.225 ($0.9\times2/8$), respectively. With a threshold of 0.5, the only parcel that would be selected as urban in our simulation is parcel A. The algorithm stops when the total area of urban parcels reaches the total urban land area.

In order to determine the parameters in the aforementioned logit function, we perform a logistic regression based on manually prepared parcels for the city of Beijing (12 183 $km^2$; Yanqing and Miyun counties in the Beijing Metropolitan Area are not included). Each parcel is treated as a sample in the logistic regression: there are 125 401 samples in total, 57 817 of which are urban. The observed status of being urban or nonurban is regressed on individual parcels' size, compactness, and POI density. The overall precision of logistic regression is 74.2%, and all parameters are statistically significant. The derived parameters, which capture the relative importance of different attributes, are incorporated into the constrained CA models for all cities in China.[7] Our constrained CA model is then validated with data for Beijing. The overall accuracy of 78.6% in terms of parcel counts implies the applicability of our CA model in identifying urban parcels from all parcels generated in a city.

*Inferring dominant urban function and land use mix for selected urban parcels.* Urban function for individual parcels is identified by examining dominant POI types within the parcels, defined

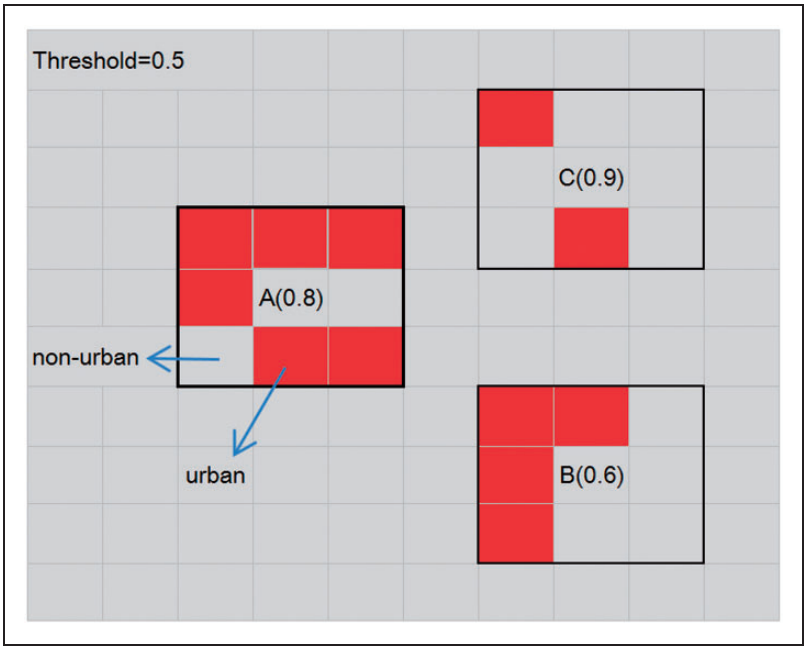


**Figure 4.** Examples of identifying urban parcels using constrained cellular automata. The parcels under investigations are labelled as A, B, and C, and the rectangles in black refer to individual parcels' neighborhoods. The numbers in parentheses denote individual parcels' probability of being urban based on their attributes.

as the POI type that has accounted for more than 50% of all POIs within the parcel. For example, if 31 out of 60 POI within a parcel are labeled as 'office building/space', the urban function for that parcel will be assigned as 'office'. Note that not all parcels would have a dominant urban function.

As a supplement measurement for the dominant function, we computed a mixed index to denote the degree of mixed land use (Frank et al., 2004). The mixed index ($M$) of a land parcel is calculated using equation (2):

$$M = -\sum_{i=1}^{n} (p_i \times \ln p_i) \tag{2}$$

where $n$ denotes the number of POI types, and $p_i$ is the proportion of POI type $i$ among all POIs in the parcel. This index has been used previously to understand evolving travel mode choice and public health outcomes, as well to study changing senses of community (Manaugh and Kreider, 2013).

*Validations.* Our parcel identification and characterization results are validated at two spatial levels. At the first, a finer spatial scale (ie, parcel level), we compare the geometry and attributes of urban parcels generated by our program with those identified manually in the conventional approach. Due to data availability, this fine-scale comparison is only performed for the city of Beijing (ie, the aforementioned BICP data). Since urban parcels for Beijing were collected in 2010 with a total urban area of 1677.5 $km^2$, we reran the

constrained CA model for Beijing using this total number and regenerated urban parcels for Beijing.[8]

In order to (partially) remedy the limited availability of manually collected parcel data, we perform a second set of validations at the aggregated level (ie, regional level). We compare the overall distribution of urban parcels identified from OSM with survey data. To ensure the comparability of urban parcels from both approaches, we use road networks from survey data in place of OSM roads and rerun our program to identify urban parcels. The survey dataset reports the actual roads, thus, according to our working definition of parcels, parcels generated from the survey dataset should in theory correspond more closely to real-world parcels. In other words, parcels generated based on the survey dataset are used to benchmark the validity of the OSM-based product. While this first aggregated analysis focuses on OSM's data quality, a second analysis at the aggregated level evaluates the algorithm itself. This second aggregated analysis is conducted by comparing survey-based urban parcels with 'urban patches' in the remote sensing products (GLOCOVER and DMSP/OLS).

## Results

### *Parcel characteristics*

We ran the proposed constrained CA model for all 654 cities. Our method generates exceedingly large parcels (ie, individual parcels that would exceed the total urban area constraints) for cities with limited OSM data. We adopt a pragmatic threshold of ten parcels and deem the 297 cities with ten or more urban parcels as 'successfully' processed by our algorithm (Figure 5). Due to its sheer size — roughly the same size as Austria — Chongqing was the only MD-level city absent from this group of successfully processed cities. All SPC cities, as well as half of the medium-to-small cities at the PLC and CLC levels resulted in more than ten urban parcels.

A total of 232 145 parcels were identified for these 297 cities (Figure 5), and 82 645 out of all generated parcels are labeled as 'urban' (total urban area 25 905 $km^2$). The average number of urban parcels for MD, SPC, OPCC, PLC, and CLC cities is 1411, 407, 199, 79, and 26, respectively. As discussed previously, cities with a larger population and higher administrative rank (eg, Beijing, the national capital; Nanjing, a provincial capital; and Qingdao, a subprovincial level city) tend to have more detailed OSM road networks and consequently a greater number of parcels.

For all urban parcels, we calculate (1) land-use density; (2) urban function; and (3) degree of land-use mix. Figure 6 illustrates the results for five representative cities. Density among parcels within a city or across cities could be compared in terms of inferred and standardized density attributes. Urban function and measurements of land-use mix point to substantive mixing of land use. More specifically, 58 915 (71.3%) out of the 82 645 urban parcels have 'dominant' urban functions (Figure 6), including 12 448 residential parcels, 11 353 commercial parcels, 9797 Office building/space parcels, and 3301 government parcels. Moreover, the average land-mix degree for all urban parcels in the 297 cities is approximately 0.66 (with a maximum of 1). The generated parcel data are distributed freely online at http://www.beijingcitylab.com.

### *Parcel validation*

For validation at the parcel level, Table 2 is a summary of the comparison of parcels generated by our approach and those contained in the BICP Beijing parcel data.

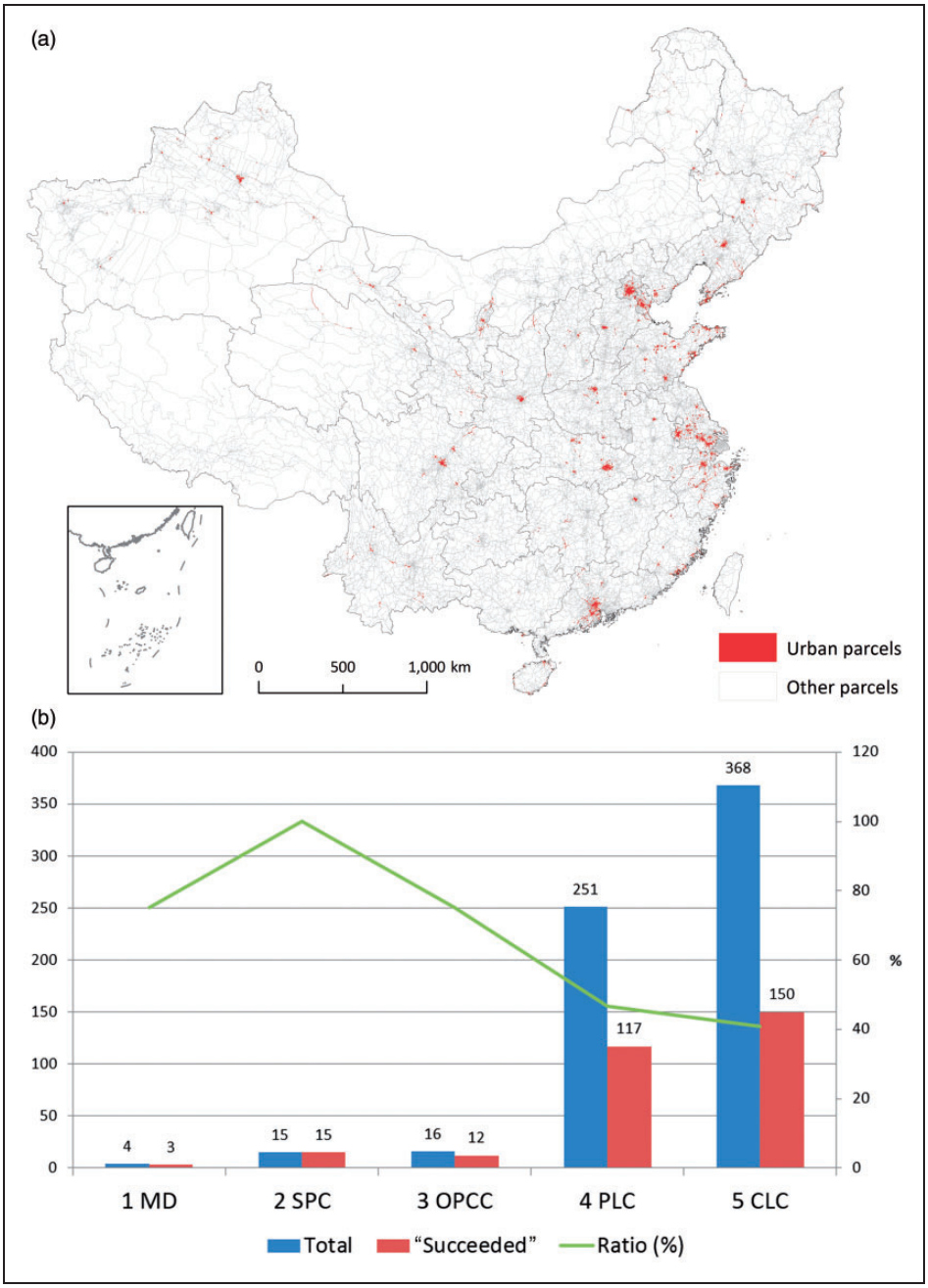


**Figure 5.** All generated parcels and urban parcels in China: (a) spatial distribution; (b) the profile of'successfully processed' cities.

Results given in Table 3 suggest that the OSM-based approach generally produces larger parcels, due to the lack of information about tertiary and more detailed roads in the OSM dataset.[9] Nevertheless, the overlapped area between urban parcels in the OSM and BICP data accounts for 71.2% of all OSM-based urban parcels, suggesting that both datasets

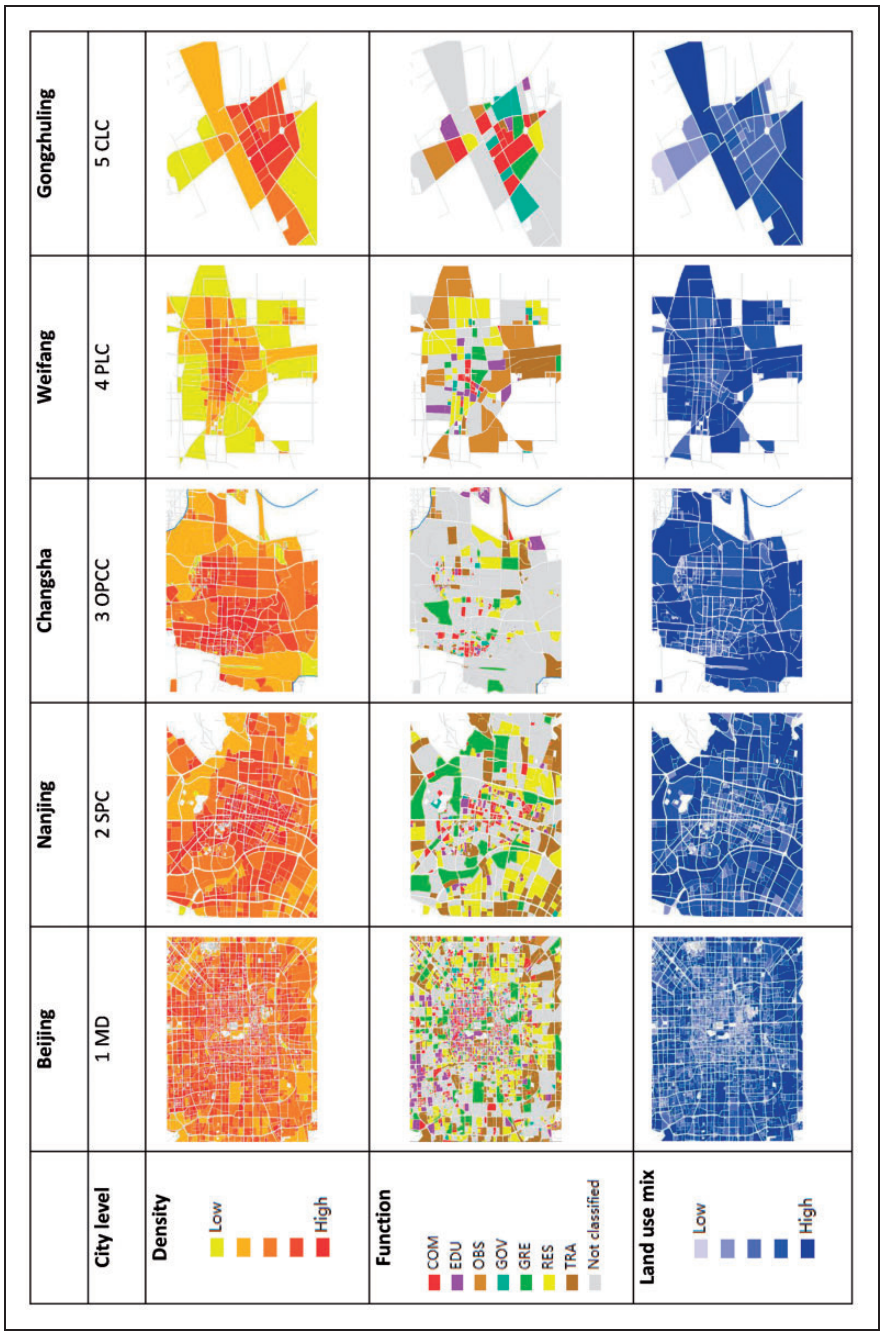


**Figure 6.** The generated parcels and their attributes in typical cities of China.

**Table 2.** Comparison of selected OpenStreetMap (OSM) and Beijing Institute of City Planning (BICP) parcels in Beijing (R = ring road).

| Parcels | Parcel count | Average size (ha) | Overlapped with BICP | Spatial distribution (in terms of area, $km^2$) | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | within R2 | R2–R3 | R3–R4 | R4–R5 | R5–R6 | beyond R6 |
| OSM | 7130 | 17.2 | 1194.2 $km^2$ (71.2%) | 42.5 | 74.0 | 113.4 | 263.5 | 666.5 | 519.9 |
| BICP | 57818 | 2.9 | – | 48.6 | 69.7 | 99.8 | 229.5 | 687.9 | 544.4 |
| OSM/BICP | 0.12 | 5.93 | – | 0.87 | 1.06 | 1.14 | 1.15 | 0.97 | 0.95 |

**Table 3.** The comparison of OpenStreetMap (OSM) and Survey (ORDNANCE) urban parcels for 297 cities.

| Data | Urban area ($km^2$) | Parcel count | Average parcel/patch size (ha) | Intersected with survey-based results ($km^2$) |
|---|---|---|---|---|
| OSM | 25905 | 82645 | 31.3 | 15053 |
| Survey | 25670 | 260098 | 10.0 | – |

largely capture the same geographic distribution of land-use activities. In addition, we decompose the city of Beijing into subregions bounded by major ring roads, and calculate the proportion of parcels falling into individual subregions. The proportion of parcels falling into subregions between ring roads is consistent across both datasets. We also compare the size distribution of parcels, with both approaches showing lognormal distribution with similar mean values.

Furthermore, density and urban functions of OSM-based urban parcels in Beijing are compared with other data sources. With the same OSM-generated parcel boundaries, we calculate development density for individual parcels (a total of 7130 parcels)[10] based on (1) building information such as floor area gathered from BICP for 2008 and (2) POI data, as building information is the common data for inferring development density. The Pearson correlation coefficient between development densities calculated in two different ways is 0.858, suggesting that POI data could be used as a proxy for urban density. As POI types and land-use types in BICP data were not totally aligned with each other, we limited our comparison to OSM parcels with a dominating residential function and residential parcels in BICP. We overlaid residential parcels in OSM and BICP, and the overlapping area is 211.5 $km^2$ (56.3% of the total 375.6$km^2$ OSM-based residential parcels). In other words, the parcel-level validation suggests that, despite only using data from online sources, our OSM-based approach often produces reasonably good approximations for data produced by the conventional manual method.

Validation at the aggregated regional level is performed by comparing urban parcels generated by OSM and survey (ORDNANCE) data in 297 cities where substantive numbers of urban parcels are identified (Table 3). Urban parcels based on survey data are generated and selected using the same parcel generation and selection methods as applied to the OSM data. From Table 4, it appears that the OSM-based approach tends to generate parcels of larger size, again due to the relative sparseness of road networks in the OSM data. The match degree between urban land by OSM and survey data is 58.1%, calculated as the

**Table 4.** The comparison of urban parcels/patches in various data for 627 cities.

| Data | Year | Spatial resolution | Urban area ($km^2$) | Parcel/patch count | Average parcel/patch size (ha) | Intersected with survey data ($km^2$) |
|---|---|---|---|---|---|---|
| Survey | 2011 | | 39 746 | 3 50 102 | 13.0 | |
| DMSP/OLS | 2008 | 300 m | 44 720 | 1293 | 3458.6 | 21 553 |
| GLOBCOVER | 2009 | 1 km | 39 389 | 12 515 | 314.7 | 15 206 |

ratio of the area of overlapping urban parcels to the area of all OSM-based urban parcels. When we disaggregate the overlapping results for each city level, the ratio for MC, SPC, and OPCC is around 70% and the ratio for FLC and CLC is around 45%. This, following the comparison of road networks in both datasets in Figure 2, further confirms the data completeness of OSM in big cities was much better than that in small cities in China.

Additionally, as the errors in OSM-generated parcels may come from (1) errors in the raw OSM data and (2) errors in our empirical framework, we attempt to single out pitfalls in our empirical framework. In this regard, we cross-reference survey-based parcels with remote-sensing-based parcels. On the one hand, we expected survey-based parcels would be less plagued by data issues that linger over the applicability of OSM, and thus reflect the effectiveness of our algorithms. On the other hand, we deem urban areas identified from remote sensing images as the 'ground truth'.

More specifically, we compare the urban parcels based on survey data with the 300 m-resolution urban area of China in GLOBCOVER (Bontemps et al., 2009), as well as 1 km-resolution urban area of China from DMSP/OLS in 2008 (Yang et al., 2013). We quantify the overlapping of survey-based urban parcels and 'urban patches' identified in the remote sensing products. There were 21 553 $km^2$ urban lands derived based on the survey data (ORDNANCE) (54.2%) overlapping those of DMSP/OLS. This overlapping percentage rises to 60% if we assume road spaces that are not accounted for in survey-based parcels were covered by DMSP/OLS. We also found the comparison results between survey-based and GLOBCOVER were robust, though less promising than those between survey-based and DMSP/OLS. The overlapped area of the two datasets was 19 501 $km^2$, 49.5% of urban area in GLOBCOVER and 43.6% in DMSP/OLS. Considering the sheer size of our study area and limited data sources, we believe that these percentages are reasonably good.

We note that these percentages of overlapping are achieved even though remote sensing images only capture the ground truth at a rather coarse scale. For example, results in Table 4 suggest that an average urban parcel was around 300 m × 400 m, much smaller than the average size of an 'urban patch' in GLOBCOVER or DMSP/OLS. In addition, the time lag between survey-based and DMSP/OLS, to a certain extent, underestimated the ratio of overlapping. Still, the percentage of overlapping might be hampered by inconsistency between the spatial resolutions of the two datasets. In the meantime, we are striving to collect fine-scale urban land-use data for the whole of China, which would enable more detailed validations of our method.

## Conclusions

Aiming at resolving the paucity of parcel data for cities of the developing world, our study proposes a novel and extensible empirical framework for the automatic identification and characterization of parcels using OSM and POI data. Our analysis represents a preliminary

attempt to use volunteered GIS data to identify and characterize urban parcels in China. Empirical results suggest that OSM and POI could help to produce a reasonably good approximation of parcels identified by conventional methods, thus making our approach a useful supplement. The bottom line, however, is that we may argue for more ways of identifying parcels, not just newer, faster, or more efficient ways. In fact, as we have enumerated in the review section, parcels have been conventionally produced in a number of different ways. Thus, our method provides more opportunities and alternative methods to characterize parcels and generate insights.

More specifically, our contribution lies in the following aspects. Firstly, we propose a straightforward approach to delineating parcels, identifying urban parcels, and characterizing parcel features, using ubiquitously available OSM data. Secondly, we employ a novel approach that incorporates a vector-based CA model into the identification of urban parcels. Thirdly, our approach has been applied to hundreds of Chinese cities, and could possibly be extended to generate parcel data for other areas that lack of conventional data sources.

The final product of our project is a dataset containing fine-scale urban parcels with detailed features for 297 Chinese cities. This dataset can be applied for, but is not limited to, the following purposes. Firstly, the dataset which can be updated periodically, can provide parcel maps for urban planning and studies in places where digital infrastructure development is weak. For example, official parcel data for Beijing are generally updated every three years but our approach would enable updating on a yearly basis so as to capture rapid growth in Chinese cities. Secondly, the dataset can serve as the base for emerging vector-based urban modeling, for example, vector-based CA models and agent-based models (Jjumba and Dragicevic, 2012; Stevens and Dragicevic, 2007). Urban parcels generated by our approach would enable us to establish large-scale urban expansion models for large geographic areas (eg, an entire nation) at parcel level. Such urban expansion models would open up new avenues for fine-scale regional growth management that are technically impossible without parcel data. Our attempt to establish such a parcel-level national-scale urban-expansion model will be reported in a related paper. Thirdly, parcel attributes such as urban functions and land-use intensity provide useful measurements for urban analysts to examine inter alia quality of life, urban growth, and land-use changes (Frank et al., 2010). As a sign of the usefulness of our project, though our dataset has been released for a very brief period of time, many planning projects and urban analyses have reportedly explored and used our parcel data; our parcel dataset has been downloaded more than 1500 times and we received over 100 comments in the first week of its release. In the past, planning professionals in China have had less access to land-use data at such fine spatial scale. Fourthly, the generated parcels could be used as spatial units for consolidating other spatially referenced data: for example, geotagged photographs, transportation smart card records, taxi trajectories, and mobile phone traces. The estimation of urban function, density, and land-use mix would be improved by integrating different data sources.

As discussed previously, we aware that our method is likely to be susceptible to the caveats of crowd-sourced data (Sui et al., 2013). For example, we note that, in addition to data completeness, other issues such as data accuracy, data 'vandalism', temporal inconsistency between data uploaded at different times, and flexible taxonomy in metadata may also affect OSM data quality. All these caveats need to be handled with caution when open data such as OSM data are applied (Haklay et al., 2010; Neis et al., 2012). Because the general limitations and setbacks of using open and crowd-sourced data to study urban dynamics have been detailed elsewhere (Elwood et al., 2012; Liu et al., 2014; Sui et al., 2013; Sun et al., 2013), we conclude by noting limitations and possible future research

avenues that are specific to our empirical framework. A first limitation of our approach is that OSM road networks are relatively sparse in many cities (especially those at the lower end of the administrative hierarchy) and lead to unrealistic, large urban parcels. This deficiency is likely to be alleviated by the ever-increasing coverage and quality of OSM data in China (Figure 1). Techniques for parcel subdivision would be an alternative means of generating more realistic urban parcels in small cities in China (Aliaga et al., 2008). A second limitation is related to the use of POI for estimating land-use density. Our current approach focuses on the quantity rather than quality of individual POI (eg, a large department store and a small convenience store are treated equally). Possible improvements include the incorporation of online check-in data (eg, Jiepang and SinaWeibo — Chinese equivalents of Twitter and Foursquares, respectively), taxi trajectories, and transportation smart card records to supplement inferring land-use intensity. Thirdly, the current fine-scale validation is limited to the city of Beijing, and our method needs to be validated and refined with real-world parcel data in more cities. As already mentioned, more efforts need to be put into data and model validation. Lastly, the CA model can be improved by incorporating more constraints, such as accessibilities to main roads and city centers, as well as exclusive development zones.

*** *** ***

As the generated parcel data are freely available online,[11] the online data distribution and visualization offer a new avenue to validate and improve our methods: crowd-sourced validations (Fritz et al., 2012). More specifically, individual users, with their local knowledge and experience, can identify and report geometric and/or thematic errors of parcels, which in turn can be incorporated and used to finetune our CA model.

## Acknowledgements

The two authors contributed equally to this article, thus the order of author list is alphabetical. The authors are very grateful to the Editor and anonymous reviewers for their insights and thorough critiques. All errors remain the authors' own. Ying Long would like to acknowledge the financial support of the National Natural Science Foundation of China (Grant No. 51408039).

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## Notes

1. Parcels in China roughly correspond to 'blocks' in the US context. In this paper, we use the term 'parcel' to be consistent with other literature on Chinese cities.
2. We have interviewed fifty-seven planning professionals in China (twenty-three of these are affiliated to research institutes and universities; twenty-one to foreign planning or architect firms such as AEOCOM and Atkins; and thirteen with domestic planning institutes and firms). Professionals often rely on manual digitalization of land-use maps (often in raster formats as

vector data files would not be released). This process is extremely time consuming and often produces land-parcel maps of less than desirable quality.
3. Sansha in Hainan, Beitun in Xinjiang, and Taiwan were not included in our analysis due to the lack of availability of OSM and POI data.
4. Roads in the survey map were partially covered by roads in OSM. According to our careful check, almost all roads in OSM were also roads in the survey map.
5. The unit is the POI count per $km^2$. For parcels with no POIs, we assume a minimum density of 1 POI per $km^2$.
6. Each city has its own constrained CA model for identifying urban parcels.
7. We admit the heterogeneity of weights in various cities, however we do not have existing parcels for other cities at the time of this research.
8. The urban area of the city of Beijing was 1445.0 $km^2$ in 2012 (MOHURD, 2013), which was less than that of the urban parcels prepared by BICP (1677.5 $km^2$). Such inconsistency between official yearbooks (ie, MOHURD reports) and geospatial data (ie, BICP data) in China is not rare.
9. Parcels derived from survey data in Beijing were similar with those by planners in BICP in terms of parcel size.
10. Density for BICP parcels is calculated based on floor space rather than POI. Floor space information was limited to the parcels within the sixth ring road of Beijing, and this information was used in the comparison.
11. Beijing City Lab, Data15, http://www.beijingcitylab.com/data-released-1/data1-20/.

## References

Alberti M, Booth D, Hill K, Coburn B, et al. (2007) The impact of urban patterns on aquatic ecosystems: an empirical analysis in Puget lowland sub-basins. *Landscape and Urban Planning* 80: 345–361.

Aliaga DG, Vanegas CA and Beneš B (2008) Interactive example-based urban layout synthesis. *ACM Transactions on Graphics (TOG) ACM* 27: 5–160.

Beijing Institute of City Planning (2010) Existing land use map of Beijing. Internal Working Report, Beijing Institute of City Planning, Beijing.

Beresford AR and Stajano F (2003) Location privacy in pervasive computing. *Pervasive Computing IEEE* 2: 46–55.

Bontemps S, Defourny P, Van Bogaert E, et al.(2009)GLOBCOVER: Products Description and Validation Report. http://due.esrin.esa.int/files/GLOBCOVER2009_Validation_Report_2.2.pdf.

Cheng J, Turkstra J, Peng M, et al. (2006) Urban land administration and planning in China: opportunities and constraints of spatial data models. *Land Use Policy* 604–616.

Elwood S, Goodchild M and Sui DZ (2012) Researching volunteered geographic information: spatial data, geographic research, and new social practice. *Annals of the Association of American Geographers* 571–590.

Erickson A, Rogers L, Hurvit P, et al. (2013) Challenges and solutions for a regional land use change analysis. Proceedings of ESRI, http://proceedings.esri.com/library/userconf/proc06/papers/papers/pap_1472.pdf.

Frank LD, Andresen MA and Schmid TL (2004) Obesity relationships with community design, physical activity, and time spent in cars. *American Journal of Preventive Medicine* 27(2): 87–96.

Frank LD, Sallis JF, Conway TL, et al. (2006) Many pathways from land use to health: associations between neighborhood walkability and active transportation, body mass index, and air quality. *Journal of the American Planning Association* 72: 75–87.

Frank LD, Sallis JF, Saelens BE, et al. (2010) The development of a walkability index: application to the Neighborhood Quality of Life Study. *British Journal of Sports Medicine* 44: 924–933.

Fritz S, McCallum I, Schill C, et al. (2012) Geo-Wiki: an online platform for improving global land cover. *Environmental Modelling and Software* 31: 110–123.

Girres JF and Touya G (2010) Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS* 14: 435–459.

Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. *GeoJournal* 69: 211–221.

Goodchild MF (2008) Spatial accuracy 2.0. In: Zhang J-X and Goodchild MF (eds) *Spatial Uncertainty, Proceedings of the Eighth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences Volume 1*. Liverpool: World Academic Union, pp. 1–7.

Hagenauer J and Helbich M (2012) Mining urban land-use patterns from volunteered geographic information by means of genetic algorithms and artificial neural networks. *International Journal of Geographical Information Science* 26: 963–982.

Haklay M (2010) How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design* 37: 682–703.

Haklay M and Weber P (2008) OpenStreetMap: user-generated street maps. *Pervasive Computing, IEEE* 7(4): 12–18.

Haklay M, Basiouka S, Antoniou V, et al. (2010) How many volunteers does it take to map an area well? The validity of Linus' Law to Volunteered Geographic Information. *The Cartographic Journal* 47: 315–322.

Herold M, Scepan J and Clarke KC (2002) The use of remote sensing and landscape metrics to describe structures and changes in urban land uses. *Environment and Planning A* 34: 1443–1458.

Jabareen YR (2006) Sustainable urban forms their typologies, models, and concepts. *Journal of Planning Education and Research* 26: 38–52.

Jiang B and Liu X (2012) Scaling of geographic space from the perspective of city and field blocks and using volunteered geographic information. *International Journal of Geographical Information Science* 26: 215–229.

Jiang B, Liu X and Jia T (2013) Scaling of geographic space as a universal rule for map generalization. *Annals of the Association of American Geographers* 103: 844–855.

Jjumba A and Dragićević S (2012) High resolution urban land-use change modeling: agent iCity approach. *Applied Spatial Analysis and Policy* 5: 291–315.

Jokar Arsanjani J, Helbich M, Bakillah M, et al. (2013) The emergence and evolution of OpenStreetMap: a cellular automata approach. *International Journal of Digital Earth* 8: 76–90. (1–30 (accepted).

Jokar Arsanjani J, Helbich M, Bakillah M, et al. (2013) Toward mapping land-use patterns from volunteered geographic information. *International Journal of Geographical Information Science* 27: 2264–2278.

Kressler FP, Bauer TB and Steinnocher KT (2001) Object-oriented per-parcel land use classification of very high resolution images. In: Remote Sensing and Data Fusion over Urban Areas, IEEE/ISPRS Joint Workshop, pp. 164–167.

Leitte AM, Schlink U, Herbarth O, et al. (2012) Associations between size-segregated particle number concentrations and respiratory mortality in Beijing, China. *International Journal of Environmental Health Research* 22: 119–133.

Li X and Yeh AGO (2002) Neural-network-based cellular automata for simulating multiple land use changes using GIS. *International Journal of Geographical Information Science* 16: 323–343.

Liu X, Biagioni J, Eriksson J, et al. (2012a) Mining large-scale, sparse GPS traces for map inference: comparison of approaches., In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 669–677.

Liu Y, Wang F, Xiao Y, et al. (2012b) Urban land uses and traffic 'source-sink areas': evidence from GPS-enabled taxi data in Shanghai. *Landscape and Urban Planning* 106(1): 73–87.

Liu Y, Sui Z, Kang C, et al. (2014) Uncovering patterns of inter-urban trips and spatial interactions from check-in data. *PLOS ONE* 10: 137.

Long Y, Han H Y and Yu X (2015) Discovering functional zones using bus smart card data and points of interest in Beijing. arXiv:1503.03131.

Ma L J (2005) Urban administrative restructuring, changing scale relations and local economic development in China. *Political Geography* 24: 477–497.

Manaugh K and Kreider T (2013) What is mixed use? Presenting an interaction method for measuring land use mix. *Journal of Transport and Land Use* 6: 63–72.

MOHURD (2013) *Chinese City Construction Statistics Yearbook 2012*. Ministry of Housing and Urban-rural Development of the People's Republic of China, Beijing: China Planning Press.

Neis P, Goet M and Zipf A (2012) Towards automatic vandalism detection in OpenStreetMap. *ISPRS International Journal of Geo-information* 1: 315–332.

Over M, Schilling A, Neubauer S, et al. (2010) Generating web-based 3D city models from OpenStreetMap: the current situation in Germany. *Computers, Environment and Urban Systems* 34: 496–507.

Pacifici F, Chini M and Emery WJ (2009) A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification. *Remote Sensing of Environment* 113: 1276–1292.

Pinto NN (2012) A cellular automata model based on irregular cells: application to small urban areas. *Environment and Planning B: Planning and Design* 37: 1095–1114.

Ramm F, Topf J and Chilton S (2010) *OpenStreetMap: Using and Enhancing the Free Map of the World*. Cambridge: UIT Cambridge.

Soto V and Frías-Martínez E (2011) Automated land use identification using cell-phone records. Proceedings of the 3rd ACM international workshop on MobiArch ACM, pp. 17–22.

Stevens D and Dragicevic S (2007) A GIS-based irregular cellular automata model of land-use change. *Environment and Planning B: Planning and Design* 34: 708–724.

Sui D Z (2008) The wikification of GIS and its consequences: or Angelina Jolie's new tattoo and the future of GIS. *Computers, Environment and Urban Systems* 32: 1–5.

Sui D Z, Goodchild M and Elwood S (2013) Volunteered geographic information, the exaflood, and the growing digital divide. In: Sui D et al. (eds) Crowdsourcing Geographic Knowledge Netherlands: Springer. pp. 1–12.

Sun L, Axhausen K W, Lee D H, et al. (2013) Understanding metropolitan patterns of daily encounters. *Proceedings of the National Academy of Sciences* 110: 13774–13779.

Toole J L, Ulm M, González M C et al. (2012) Inferring land use from mobile phone activity. Proceedings of the ACM SIGKDD International Workshop on Urban Computing ACM, pp.1–8.

Wu F (2002) Calibration of stochastic cellular automata: the application to rural-urban land conversions. *International Journal of Geographical Information Science* 16: 795–818.

Yang Y, He C, Zhang Q, et al. (2013) Timely and accurate national-scale mapping of urban land in China using Defense Meteorological Satellite Program's Operational Linescan System nighttime stable light data. *Journal of Applied Remote Sensing* 7(1): 073535.

Yuan J, Zheng Y and Xie X (2012) Discovering regions of different functions in a city using human mobility and POIs. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ACM, pp. 186–194.

Zhang L, Yang W, Wang J, et al. (2013) Large-scale agent-based transport simulation in Shanghai, China. *Transportation Research Record: Journal of the Transportation Research Board number* 2399: 34–43.

Zhang Y P and Long Y (2013) Urban growth simulation using V-BUDEM: a vector-based Beijing urban development model.In: The conference of spatial planning and sustainable development, Beijing.

Zheng S and Zheng J (2014) Assessing the completeness and positional accuracy of OpenStreetMap in China. In: Bandrova T, et al. (eds) *Thematic Cartography for the Society, Lecture Notes in Geoinformation and Cartography*. Cham, Switzerland: Springer International, pp. 171–189.

**Xingjian Liu**, is an assistant professor in the Department of Urban Planning and Design, the University of Hong Kong. He received degrees from Wuhan (China), Texas State (USA), Cambridge (UK). His main areas of research interest include urban China, globalization and urban changes, as well as urban planning and big data. He is a research fellow of

the Globalization and World Cities research network and associate director of the Beijing City Lab

**Ying Long**, a senior engineer in Beijing Institute of City Planning, is an inter-disciplinary scholar with a global vision and substantive planning experiences in China. His research focuses on urban planning, quantitative urban studies, and applied urban modeling. Familiar with planning practices in China and versed in the international literature, Dr. Long's academic studies creatively integrates international methods and experiences with local planning practices. Dr. Long is also the founder of Beijing City Lab (BCL), an open research network for quantitative urban studies.