Applied Geography 75 (2016) 200–212

Contents lists available at ScienceDirect

Applied Geography

journal homepage: www.elsevier.com/locate/apgeog

Mapping the fine-scale spatial pattern of housing rent in the metropolitan area by using online rental listings and ensemble learning

Yimin Chen [a, b], Xiaoping Liu [a, b, *], Xia Li [a, b], Yilun Liu [c], Xiaocong Xu [a, b]

[a] School of Geography and Planning, Sun Yat-sen University, Guangzhou, China
[b] Guangdong Key Laboratory for Urbanization and Geo-simulation, Sun Yat-sen University, Guangzhou, China
[c] College of Natural Resources and Environment, South China Agricultural University, Guangzhou, China

ARTICLE INFO

Article history:
Received 11 March 2016
Received in revised form
3 August 2016
Accepted 13 August 2016

Keywords:
Housing rent mapping
Online rental listings
Anjuke
Ensemble learning

ABSTRACTS

The accurate mapping of housing rent is crucial to the understanding of residential dynamics. In this study, we proposed the use of online rental listings as a new reliable data source for mapping housing rent. With the collected individual rental information from an online platform, we attempted to produce the fine-scale spatial pattern of housing rent in the metropolitan area of Guangzhou, China, at the neighborhood committee (NC) level. This involves the task of estimating the housing rent for areas with no observation data of housing rent. To this end, we evaluated six numeric prediction methods of machine learning. We further enhanced their performance through ensemble learning, an approach which can form new classifiers with even better performance than any of the individual constituent classifiers. We implemented ensemble learning through ways of bagging and stacking, and selected the most accurate ensemble classifier to produce the spatial pattern of housing rent at the NC-level. In the resulting housing rent pattern, we identified a distance decay relationship between the housing rent and the distance from the city center. The data sources and the ensemble learning platform in this application of housing rent mapping are generally open access. Therefore, the proposed approach in this study can provide useful hints for housing rent mapping in other geographical areas. Our mapping results can also be integrated with additional information to support the studies of urban residential problems in China.

1. Introduction

Rent is always a crucial variable to explain urban phenomena in either a theoretical or an empirical manner. In urban economic theory, the core concept of bid rent is used to represent how much money the competing land users are willing to pay for a specific land unit with a certain distance to the central business district (CBD) (Ahlfeldt, 2011; Y. M.; Chen, Li, Wang, & Liu, 2012). The bid rent in this theory can also be viewed as the trade-off between accessibility and commuting cost. In the realistic real estate markets, however, rent is more frequent to be considered as the indicator of housing cost for residents, or in turn the economic return for residential investors. In the famous Smith's rent gap theory to explain the process of gentrification (Lopez-Morales, 2011), two different types of rents are defined, i.e. the current rent of a property and its expected rent after rehabilitation. The disparity between the current rent and the expected rent is regarded as the primary motivation of local investors to renovate the properties in the declining urban areas. Such local actions then collectively form the process of gentrification at the macro scale.

The accurate mapping of housing rent is essential to many urban research. Firstly, the mapping of housing rent and related variables, such as housing price and land price, characterizes the spatial distributions of real property values, and is fundamental to the monitoring and evaluation of local residential markets. By using data of this kind the price-to-rent ratio, for example, can be computed for detecting the inflation of housing markets and its spatial distributions (Frappa & Mésonnier, 2010). Secondly, the spatial delineation of housing rent is enables the analysis of how structural/neighborhood characteristics affect real property values (Ahlfeldt, 2011), and in turn explain the determinants of renting

* Corresponding author. School of Geography and Planning, Sun Yat-sen University, 135 West Xingang RD., Guangzhou 510275, China.
E-mail address: liuxp3@mail.sysu.edu.cn (X. Liu).

http://dx.doi.org/10.1016/j.apgeog.2016.08.011
0143-6228/

households' residential behaviors (W. Wu, Zhang, & Dong, 2013). The knowledge obtained from this kind of research is particularly useful for the design of new urban development (Waltert & Schläpfer, 2010) as well as the renewal of declining areas in the city (Lopez-Morales, 2011). Thirdly, the housing rent maps explicitly express the spatial variations of housing cost, a variable of which is important for research of households' housing/non-housing consumptions and associated policy makings towards the low-income group. As reported by Davis and Ortalo-Magné (2011), for the renting households in the U.S., the share of housing expenditure to total income is remarkably stable across regions and over time, but varies significantly among different income groups. For the low-income group, besides the private rental market (Kemp, 2011), governments in many countries have also launched a variety of projects to supply affordable housing, such as the public rental housing (Delang & Lung, 2010). In this respect, a clear spatial pattern of housing rent can provide valuable information of private market influences, and also crucial references for determining the residential subsidies at a reasonable level.

Despite the importance discussed above, the accurate mapping of housing rent patterns at the fine-scale still remains challenging. This is primarily because the contemporary mainstream data sources of housing rent are aggregate data of official statistics, population census and surveys, which fail to provide necessary micro-level attributes for the fine-scale mapping of housing rent. Rondinelli and Veronese (2011) mentioned the scantiness of housing rent data for empirical analysis in Europe. In their research, they assembled the rent data of Italy from multiple sources, including national statistics, household surveys and price surveys by the national associations of estate agents. Partridge, Rickman, Ali, and Olfert (2010) acquired the rental information from the population census to assess the impacts of proximity to urban consumer amenities and production on housing costs growth. Ahlfeldt (2011) made use of the residential property transaction samples in Berlin to establish a hedonic model with structural and neighborhood characteristics of properties. Lopez-Morales (2011) discussed the gentrification led by social dispossession in Santiago de Chile by using the housing data acquired from Santiago Property Market Bulletin. Yi and Huang (2014) also stressed the lack of housing data in China. With the latest population census data, they revealed the housing inequality across population groups and different provinces.

Overall, these data sources have two major limitations. Firstly, most often the productions of these kinds of data are costly and labor-intensive, and their update cycles could be as long as five or more years (e.g. population census). Therefore, it would be rather difficult to obtain the timely spatial dynamics of housing rent/price by solely using these data sources. Secondly, the official statistics are usually aggregate data (Rondinelli & Veronese, 2011) in which fine-scale attributes are lacking. To overcome these limitations, we proposed the use of online rental listings as a new reliable source of housing data. The advance of the Internet and related technology has brought more and more convenience for making real property transactions or rental housing (Hogan & Berry, 2011; Rae, 2015). Such online services not only smooth the procedures of housing consumption, but also can offer a wealth of fine-scale housing information for research. For example, Hogan and Berry (2011) assessed the racial discrimination in the online rental housing market of Toronto. Hanson and Hawley (2011) also adopted a similar Internet-based approach to examine the discrimination issues in the housing market of U.S. cities. Rae (2015) explored the geography of local housing submarket through a dataset of online housing search in U.K. However, the application of online housing information for geographical research is still in its infancy. Recently Batty et al. (2012) included the modeling of housing markets in the proposed seven future research areas of smart cities based on the new forms of data sources (e.g. Internet-based data). Arribas-Bel (2014) also emphasized the utility of various kinds of online individual data for the better understanding of cities.

In this study, we present the mapping of the fine-scale housing rent in the metropolitan area of Guangzhou by using the online rental listings collected from Anjuke (http://guangzhou.anjuke.com/), a domestic real estate platform in China. China is now experiencing fast urbanization (Liu et al., 2010, 2014) and radical changes in housing consumption (Yi & Huang, 2014). According to the recent national statistics, housing cost, which can be measured by or transformed from housing rent, becomes the second largest sector (22.5%) of the urban households' expenditure (China Statistical Bureau, 2015). To obtain a clear spatial pattern of housing rent in the metropolitan area, we carried out the analysis at the neighborhood committee (NC) level, which is the most basic level of administrative divisions in Chinese cities (A brief introduction of the administrative divisions is provided in Section 2). It is expected that our study can be helpful to support many other studies related to urban residential dynamics in China, such as housing affordability (J. Chen, Hao, & Stephens, 2010), urban poverty (He, Wu, Webster, & Liu, 2010) and residential segregation (Z. Li & Wu, 2008).

Our application of housing rent mapping also involves the estimation of housing rent for areas with no observation data of housing rent. The estimation of housing price/rent is not a new problem, and a variety of methods has been proposed in previous literature relevant to this topic. Traditionally, the hedonic price regression models are often used to estimate the housing price/rent (Ahlfeldt, 2011). Recently, Waltert and Schläpfer (2010) reviewed the hedonic price literature (46 studies) and reported the effects of landscape amenities on housing price. However, the spatial effects, which are referred to as the spatial autocorrelation and spatial heterogeneity, may not be well addressed by the traditional hedonic regression models (Dubé & Legros, 2014). The family of spatial econometric models explicitly account these effects and hence achieves better accuracy in the estimation of housing price/rent (Anselin & Le Gallo, 2006). The main types of spatial econometric models include the spatial lag model and the spatial autoregressive error model (Osland, 2010). These models have been increasingly applied in the empirical housing research during the past two decades. For example, Yu, Wei, and Wu (2007) employed geographically weighted regression (GWR) model to examine the spatial effects for the modeling of housing prices in Milwaukee. Bitter, Mulligan, and Dall'erba (2007) also applied GWR in their empirical research to deal with the spatial heterogeneity in housing markets. Dubé and Legros (2014) addressed the latent bias in the spatial autoregressive model when housing data is pooled over time. Besides these models, the method of spatial interpolation is also capable of estimating housing price/rent, although in this method only the factor of geographical distance is taken into account (S. Hu, Cheng, Wang, & Xie, 2012).

From the perspectives of accurate prediction, however, alternative methods from the field of machine learning present two appealing advantages over traditional statistical methods. Firstly, statistical methods usually require assumptions of the distribution of data, whereas machine learning methods are more flexible with no requirement of data distribution; secondly, machine learning methods can capture the higher-order interactions between data and hence have better prediction ability than traditional statistical models (Jerez et al., 2010). Moreover, the individual machine learning methods/classifiers can be further combined to form a new classifier that has even better performance than any of the individual constituent method/classifier. This is the so-called ensemble learning approach (X. Li, Liu, & Yu, 2014; Zhou, 2012).

Ensemble learning has been successfully applied in many fields, including real estate appraisal (Graczyk, Lasota, Trawiński, & Trawiński, 2010), but remains relatively new to human geography. Nevertheless, it is worth exploring the ability of ensemble learning to solve geographic problems. Therefore, we adopted the ensemble learning approach to develop the housing rent prediction model. There are several types of ensemble learning, such as bagging and stacking (Graczyk et al., 2010). As a result, we carried out multiple experiments to compare the outcomes of bagging and stacking, and chose the most accurate ensemble classifier for the housing rent prediction (See Section 3 for more details). Finally, we discussed the implications obtained from the results of housing rent mapping.

2. Study area and data

The study area is located in Guangzhou, China (Fig. 1). Guangzhou is the third biggest city in China, with a population of more than 13 million and an urbanization rate of 86% in 2014. For a better understanding of our analysis, it is worth mentioning the administrative divisions of Guangzhou, which consist of four levels. The top level is the whole city as a unit of China's prefectural-level divisions. The second level is the county-level divisions, which refers to the 11 districts in Guangzhou (Fig. 1). The third one is the township-level divisions, i.e. the sub-district units or 'Jiedaos' in urban areas or towns in the urban fringe and the rural areas. The forth level includes the most basic units of administrative divisions: the neighborhood committees (NC) in urban areas and the village committees (VC) in rural areas. This is the level in which we mapped the housing rent pattern. Fig. 2 takes Tianhe District as an example to illustrate the hierarchy of administrative divisions below the township-level. Moreover, the Guangzhou Land Resource and House Management Bureau has ranked the residential zones in 12 classes, with 1 indicating the highest class and 12 the lowest. Here we only included NCs within the residential zones from class 1 to class 9 and dropped the rest of them, because the residential zones below class 9 are basically in the rural areas where the rental market is distinctive from the urban areas. As a result, a total of 1996 NCs in the urban areas were selected for subsequent analysis.

We derived the observation data of housing rent from Anjuke, a popular online platform that publishes real estate information of homes/apartments for sale or rental. Anjuke has covered 67 cities in China and is already an influential online real estate agency in

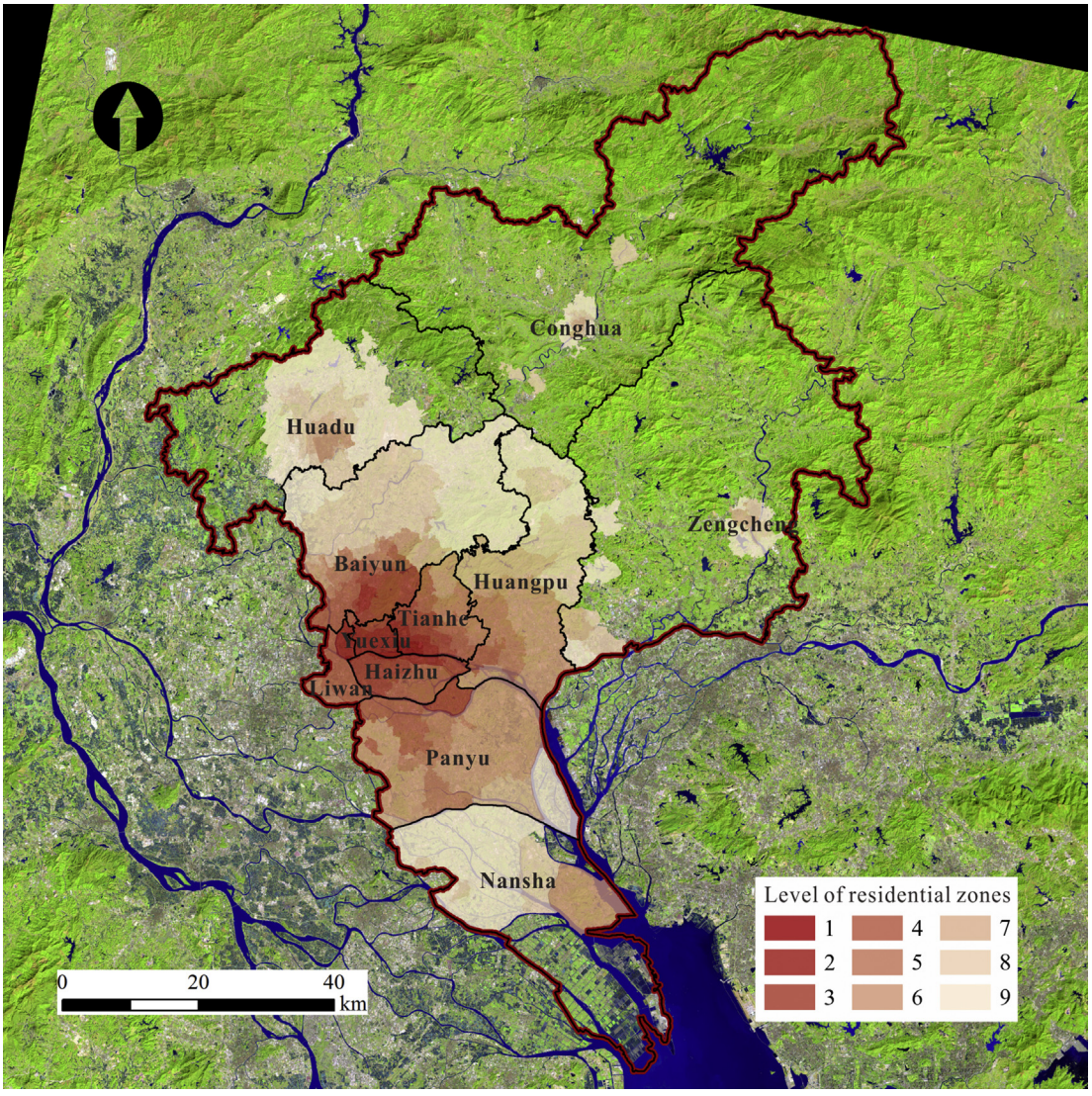

Fig. 1. Study area: Guangzhou City (23°08′N, 113°16′E). The rank of residential zones ranges from 1 to 9, with 1 being the highest class residential zone and 9 being the lowest one.

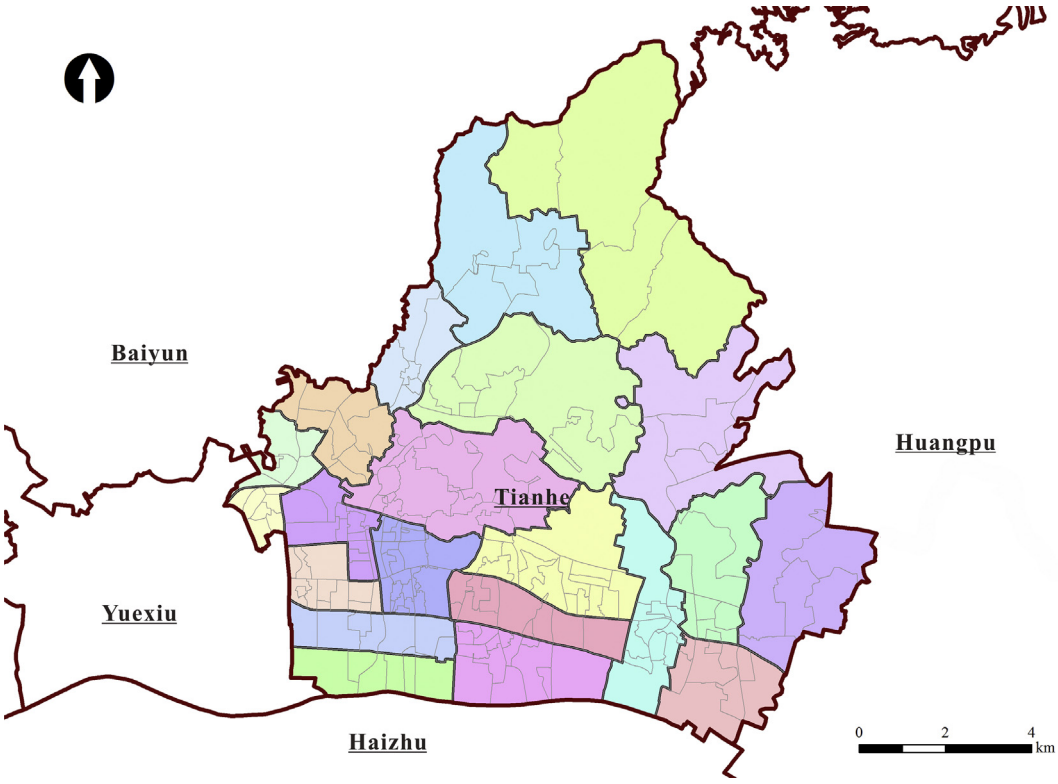

Fig. 2. Tianhe District as an example to illustrate the hierarchy of administrative divisions in Guangzhou. The bold red lines are the boundaries of county-level divisions (i.e. the districts). The bold grey lines are the boundaries of township-level divisions (i.e. 'Jiedao' in the city). The light grey lines are the boundaries of the lowest level divisions, the neighborhood committees. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Guangzhou. We created a program of web crawlers to collect the rental listings in 2015, and organized them as records with the attributes of 'identity number', 'title', 'floor area', 'number of rooms', 'rent', 'residential quarter name' and so on. We then added a new attribute of 'unit rent' into the records (i.e. the 'rent' divided by the 'floor area') so that it would be convenient to exclude the false records, such as those with an unreasonable unit rent of over 1000 yuan/m^2. There were 327,767 records in total after the removal of false records. The calculated unit rents indeed is the standardized housing rents and also an indicator to reflect the individual differences among apartments. This is quite straightforward: if two apartments have the same size, normally the one has higher (unit) rent would in turn offer better living conditions (e.g. good views or better layouts), and vice versa; however, if their sizes are different, their rents cannot be compared directly until they are transformed into unit rents (i.e. after standardization), and again the one has higher unit rent would in turn offer better living conditions (perhaps more rooms, desirable orientation or better decoration) than the other, and vice versa. In this sense, we believe that the pre-processing of converting the housing rent into unit rents is reasonable for the collected records. However, we identified one important omission in the collected records, i.e. the apartments' precise locations. Moreover, the room number or building number of the apartments are also unknown. The only useful attribute for identifying the apartments' locations is 'residential quarter name'. Thus, we designed another program to search the locations of all the residential quarters through Baidu Maps (http://lbsyun.baidu.com/index.php?title=webapi/guide/webservice-placeapi).

The above procedures of location determination would raise another two side-effects. The first one is that records of rental apartments in the same residential quarter, although quite likely to be in the separate buildings, will point to the same locations. As a result, we merged these records into a new single one according to the 'residential quarter name', and assigned the average unit rent of the original records to their corresponding residential quarters. This resulted in 3522 residential quarter records with observed locations and unit rents (Fig. 3(a) and (b)). We also compared the resulted mean unit rents of five representative residential quarters (Fig. 3(c)) with those acquired from another two important online real estate agencies in Guangzhou, including Soufun and Centaline Property Agency (CPAL). As shown by Table 1, no substantial differences are observed among these platforms, indicating that our data from Anjuke are reliable.

The second side effect is the mismatch between the residential quarters as areal objects in the real-world and their simplified representation as individual points in our analysis. To alleviate the uncertainty caused by this problem, we aggregated the residential quarter records into the NC-level according to their locations, and used their mean unit rent to represent the mean unit rent of the NCs (Fig. 3(b) and (c)). Even so, still 49% of the selected NCs' mean unit rent are unknown (i.e. no records contained) (Fig. 3(c)), and indeed they will be estimated with the available observed mean unit rent and a set of spatial variables, including the nighttime lights.

The nighttime lights data we used is the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) monthly cloud free composites from NOAA/NGDC (http://www.ngdc.noaa.gov/eog/viirs/download_monthly.html). Compared with the conventional nighttime lights data of DMSP-OLS, the VIIRS data has a finer spatial resolution (roughly 500 m) and solves the problem of saturation in the densely urbanized areas. However, the monthly composites are not filtered to exclude lights from aurora, fires and other temporal lights. Meanwhile, these composites also contain background noises. To reduce the influences caused by these uncertainties, we

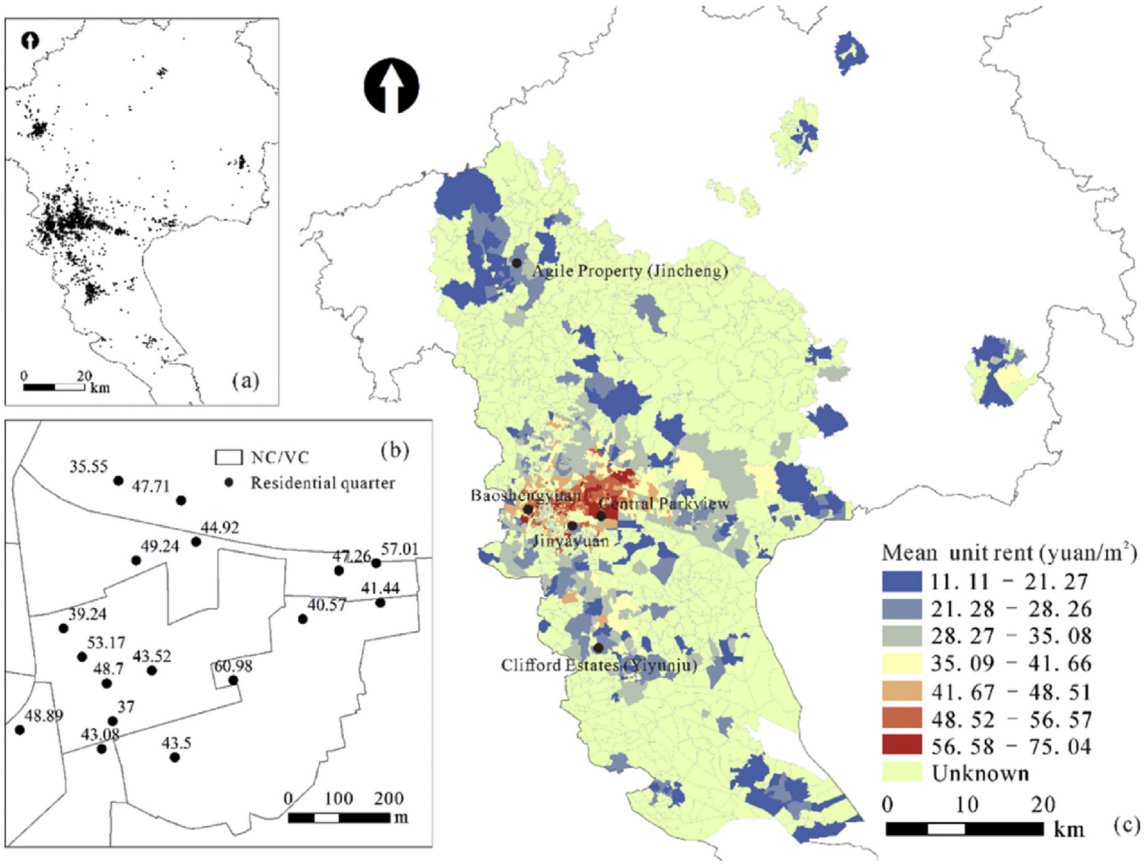

Fig. 3. The acquired housing rent data from Anjuke. (a) The locations of the acquired residential quarters; (b) Multiple residential quarters (with their unit rent on the top) exist within a single NC; (c) NC with their mean unit rents calculated based on the observations (i.e. the acquired Anjuke data), or labeled 'Unknown' if observations are missing.

Table 1
Comparison of mean unit rent and its standard deviation of five representative residential quarters (yuan/m^2) collected from different real-estate online platforms.

	Central parkview		Clifford estates (Yiyunju)		Jinyayuan		Baoshengyuan		Agile property (Jincheng)	
	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
Anjuke	97.24	21.93	28.33	2.88	42.21	5.06	56.50	11.04	21.16	2.83
Soufun	96.46	18.96	27.62	3.75	44.75	7.36	56.72	7.15	20.36	2.97
CPAL	97.33	15.05	25.60	3.95	46.66	5.87	56.32	8.43	20.65	4.06

The locations of these residential quarters are shown in Fig. 3(c). Std. = standard deviation; CPAL = Centaline Property Agency Limited; Anjuke: http://guangzhou.anjuke.com/ ; Soufun: http://gz.fang.com/; CPAL: http://gz.centanet.com/rental/gz/.

first overlapped the 12 scenes of monthly composites from January to December, 2015, and generated a new image in which the median values of the overlapping images were retrieved and assigned to each pixel. Additionally, pixels with negative values were regarded as background noises and excluded from the median image. The third step of data processing is the removal of outlier pixels. We adopt the method similar to that in Shi et al. (2014), which assumes that the lights in the core of the city should be the highest, while pixels exceeding that are considered as the outliers.

Fig. 4 demonstrates the extent of the city core, in which the highest pixel value is considered as the threshold for removing the outliers. By traversing the light values in the city core, we identified the max value of 57.59 in the median image. The outliers, which are defined as the pixels with the values exceeding 57.59, can then be easily detected by segmenting the image using this value. As shown by Fig. 4 (the white boxes), these outlier pixels primarily exist in two exurban areas, i.e. the Baiyun International Airport (19 outlier pixels) and the Nansha Xingang Terminal (27 outlier pixels). This is consistent with the findings in other applications of VIIRS image, which report the observed extremely high pixel values in the huge urban objects, such as airport and sea port (Guo, Lu, Wu, & Zhang, 2015; Ma, Zhou, Pei, Haynie, & Fan, 2014). By using this threshold method, we identified and removed a total of 49 outlier pixels. Because the outlier pixels in the Nansha Xingang Terminal are outside of the metropolitan area (Fig. 1), the actual number of outlier pixels is 22, a small proportion of pixels in the whole image (22/11681 ≈ 0.19%). For each NC, the mean nighttime lights were calculated so as to generate a variable comparable to the observed mean unit rent in the NC-level. Fig. 4(b) and (c) show the correlations between the mean nighttime lights (before and after outlier removal, respectively) and the observed mean unit rent. It is evident that the correlation is moderately improved after the removal of outliers while the other observations maintaining unchanged.

Besides the nighttime lights, we also prepared an additional set of spatial variables to further enhance the performance of our

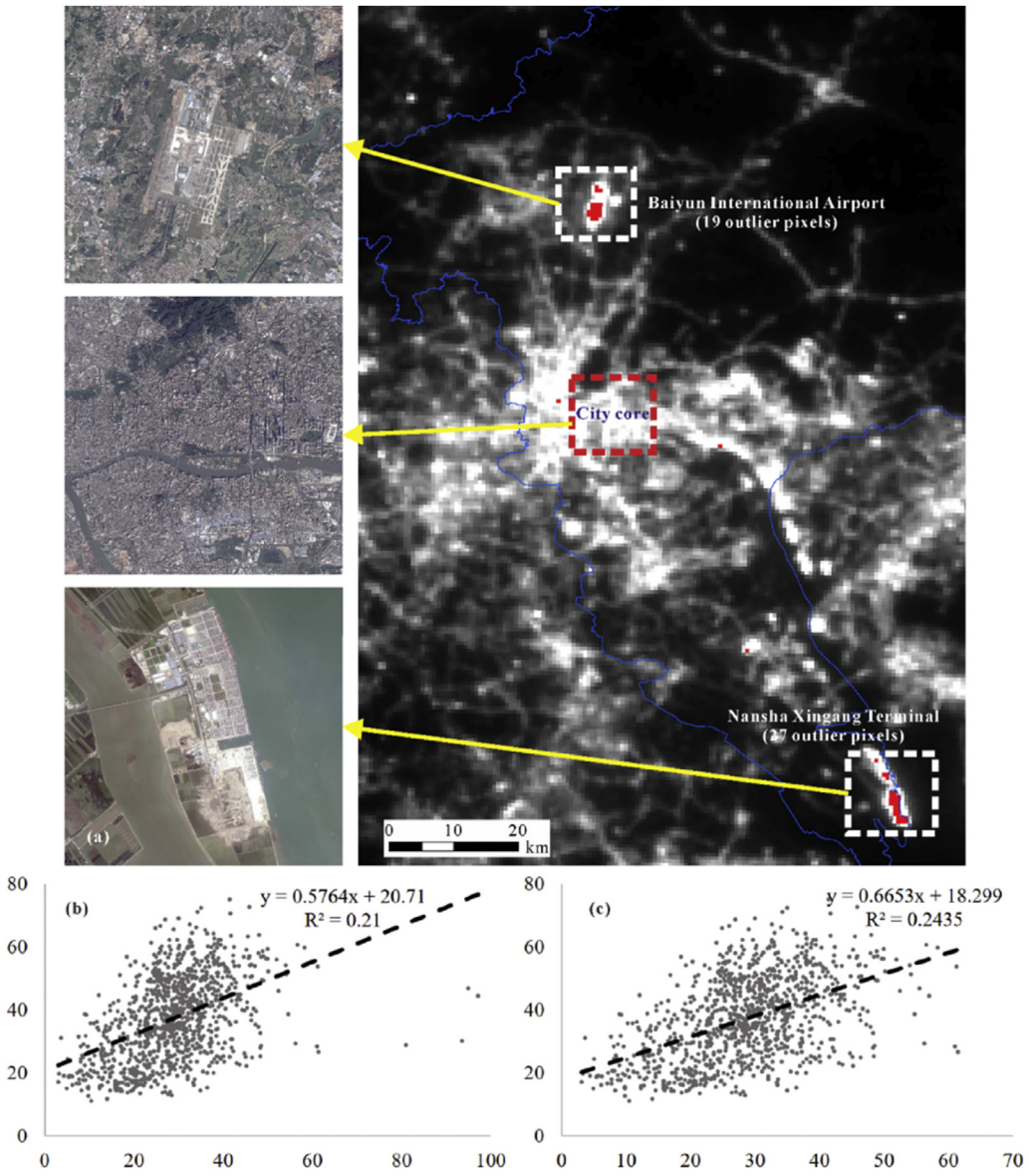

Fig. 4. (a) The spatial distributions of the outlier pixels (in red) in the nighttime light image. The images in the left column are captured from Google Earth (2015/9/17-2015/12/19). (b) and (c) are the correlations between the mean nighttime lights (before and after outlier removal, respectively) and the mean unit rent. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

model. These variables were selected according to the findings of previous residential research. Wu et al. (2013) identified several important residential determinants, including the environmental conditions, traffic systems, job opportunities, education and health care. In our work, we used the NC-level mean Normalized Difference Vegetation Index (NDVI), which was derived from the Landsat image (122/044; 2015.10.08), to represent the influence of environmental conditions (E_{NDVI}) on housing rent. The other variables were generated by using the POIs data. A POI is a point location with the attributes of its name, address and category. Thus, the data set of POIs can provide useful socioeconomic information for important locations. We chose six categories of POIs data, including education (e.g. elementary schools, middle schools or high schools), higher education (e.g. universities), enterprises, commercial buildings, hospitals and metro stations. Based on the POIs data, we created features to represent the supplies of facilities/services and potential amenities to the neighboring residential quarters. We followed the approach adopted by Hu, Yang, Li, and Gong (2016), in which the mean kernel densities of POIs is used as the features of parcel objects (polygons) for land-use classification. The residential analysis conducted by Wu et al. (2013) also employs the average density at the sub-district ('Jiedao') level to represent the facility abundance, such as school and public transportation. Therefore, we generated the kernel density for each of the selected POIs types, and aggregated into the NC-level by calculating their mean values. In the calculation of kernel density, the 'Silverman's rule-of-thumb' (Silverman, 1986) was used to automatically determined the bandwidth.

Intuitively, one can use other simpler methods instead of the kernel density calculation, such as counting the number of facility in a NC or within a buffering distance. However, these counting methods could raise bias, as illustrated by Fig. 5. Assume that NC-A has one facility (i), whereas B and C have zero. It seems that A is the best because it has the highest number of facility. However, it is

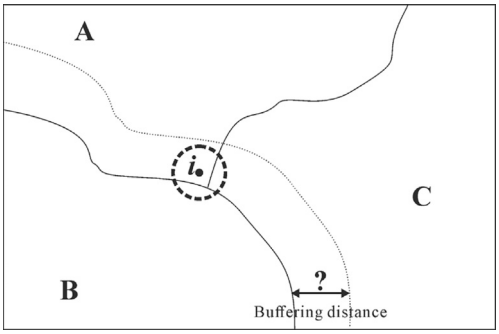

Fig. 5. A special example to illustrate the weakness of the simple counting methods.

more than likely that the influence of facility *i* is beyond the boundaries of A. If facility *i* represents an enterprise located in A (Fig. 5), for example, this enterprise can also offer job opportunities to the neighboring NCs of B and C. In other words, the accessibility of B and C to the work place is greater than zero, even though B and C do not have any enterprise of their own. In this sense, facility *i* can have an impact on the housing rent of the residential quarters nearby, including those located in B and C. However, such an impact cannot be revealed by the simple counting method, because it implicitly assumes that facilities only provide services to residents within the NC boundaries. Alternatively, one can generate the buffers of NCs with a distance weight to give more relevance to nearer facilities, and then aggregate the weighted number of facility count. This is another feasible approach to create POIs features for NCs besides the averaging kernel density we used. Both of these two approaches actually share the same idea to measure the facility abundance at a location: calculating the total quantity of facility within the neighborhood (e.g. the buffering area) weighted by the (linear or non-linear) distance function. However, the operations of these two approaches are different: the buffer approach directly applies the calculation in the NC-level, whereas the averaging kernel density approach obtain the pixel-level (local) facility abundance first and average it at the NC-level later. That is, the averaging kernel density approach indeed accounts the local heterogeneity of facility distribution in the calculation of NC-level facility abundance. Due to this characteristic and also the successful applications in previous research (T. Hu et al., 2016; W. Wu et al., 2013), we used the averaging kernel density approach to obtain the NC-level abundance of the selected POIs types.

3. Methodology: ensemble learning

Ensemble learning is a collection of methods that trains multiple classifiers/algorithms and combines their results to improve the accuracy of classification or numeric prediction (Zhou, 2012). Empirically, an ensemble classifier could perform better than a single method for most of time, unless the individual classifiers cannot provide sufficient diversity of generalization patterns (Graczyk et al., 2010). In other words, the individual classifiers should be accurate, and at the same time they should make errors on different instances (Windeatt & Ardeshir, 2004).

There are several ways to ensemble multiple classifiers for numeric prediction, such as training a single type of classifier/algorithm by using different subsets of training data, or combining different classifiers that have been already trained (Zhou, 2012). The former ensemble approach is called bagging, which is short for bootstrap aggregating. Bagging achieved the condition of diversity by using bootstrap subsets randomly drawn (with replacement) from the whole training data. Specifically, considering a full training data set with *n* instances, a sample of them will be drawn through bootstrap sampling. This sampling procedure can be applied *m* times to generate *m* sets of training samples, in which some original instances may exist in two or more training sample sets because of the replacement scheme. Next, the individual classifier is trained differently by using each of the drawn training sample sets and eventually form *m* trained classifiers that may give different predictions for the same input instance. To aggregate the different predictions, the bagging method adopts the strategy of voting for categorical classification problems or averaging for numeric prediction problems. That is, in our case, after the training procedure the bagging method averages the predicted rents from *m* trained classifiers as the final outputs. The bagging approach is particularly useful for unstable classifiers that are highly sensitive to even a small change of training conditions. This approach also avoids the overfitting problem of unstable classifiers. Therefore, bagging is usually applied to improve the performance of algorithms with tree structures, although also suitable for other types of classifiers.

Another typical approach of ensemble learning is stacking, which improves the prediction accuracy by combining multiple classifiers of different types. In stacking, the involved individual classifiers (i.e. base classifiers) are at the level-0, while the meta-classifier to combine the individual classifiers is at the level-1. It should be noted that the meta-classifier can also be one of the base classifiers (Graczyk et al., 2010). The first step in stacking is to train the level-0 classifiers separately by using the given training data set. Then, a new data set is generated, in which the outputs of the level-0 classifiers are regarded as features while the original true classes/values are still treated as the true classes/values. Next, this new data set is used to train the level-1 classifier for learning a combination of predictions from level-0 classifiers and hence achieving the improved prediction accuracy. More details of bagging and stacking (including the pseudo codes) can be found in (Zhou, 2012).

In this study, we selected six individual classifiers that have been frequently used to solve numeric prediction problems. These classifiers include Gaussian process regression (GPR), *k*-nearest neighbor algorithm (*k*-NN), backpropagation neural networks (BP-NN), radial basis function neural network (RBF-NN), fast decision-tree (FDT) and support vector regression (SVR). These methods have their own advantages and limitations. For example, *k*-NN as an instance-based learning algorithm has the promising ability to learn the complex interactions by using simple procedures of local approximation, but is rather sensitive to the configuration of parameter *k* and the choice of distance measures (X. Wu et al., 2008). Therefore, it is expected that the ensemble of these classifiers can reach an improved prediction accuracy. The open-source machine learning platform WEKA (Waikato Environment for Knowledge Analysis) (Srivastava, 2014) provides all of these classifiers and also the ensemble methods of bagging and stacking for numeric prediction. We utilized WEKA to carry out the experiments due to its convenience of data processing, implementation and visualization. Specifically, each selected classifier was run and validated separately to identify their individual performance in terms of mean absolute error (MAE) and its percentage (%MAE), and root mean squared error (RMSE) and its percentage (%RMSE):

$$\mathrm{MAE} = \frac{1}{n}\sum |r_{i,o} - r_{i,p}| \tag{1}$$

$$\%\mathrm{MAE} = \frac{1}{n}\sum \frac{|r_{i,o} - r_{i,p}|}{r_{i,o}} \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum (r_{i,o} - r_{i,p})^2} \quad (3)$$

$$\%RMSE = \frac{\sqrt{\frac{1}{n}\sum (r_{i,o} - r_{i,p})^2}}{\overline{r_o}} \quad (4)$$

where $r_{i,\,o}$ and $r_{i,\,p}$ are the observed and predicted mean unit rent for ith NC; n is the total number of NC; $\overline{r}_o$ is the observed mean unit rent. The performance of the individual classifiers was evaluated based on the approach of 10-fold cross validation, a standard way of validation in machine learning. In the 10-fold cross validation, training dataset is equally divided into 10 sets, with 9 of them being the training data and the remaining set being the test data. Then repeat this procedure by 10 times, with different test sets in each time, and a mean accuracy/error is calculated as the final outputs of validation for the model. After the evaluation of the individual classifiers, the bagging approach was applied to each of them. The stacking approach, however, can build several new classifiers based on the possible combinations of the six individual classifiers. Therefore, we ranked the individual classifiers according to their performances and added them one by one into the stacking classifier, with the two most accurate classifiers to form Stacking #1 and so forth. Finally, the individual classifiers, the bagging classifiers and five stacking classifiers were compared to identify the best classifier for the estimation of unknown mean unit rent.

4. Results and discussions

4.1. Implementation and results

It has long been recognized that methods in machine learning outperform traditional statistical methods, such as linear regression, in addressing the complexity of realistic datasets (Jerez et al., 2010). Nevertheless, before the implementation of ensemble learning, we carried out the linear regression analysis to explicitly obtain the general relationships between the input variables and the NC-level mean unit rent. As expected, the estimated coefficients demonstrate the significant positive correlations between the input variables and the mean unit rent (Table 2). The results indicate that the nighttime lights can partly explain the variation of mean unit rent. This is not surprised since nighttime lights have been widely applied to the estimation of social-economic characteristics (Ma et al., 2014) that are fundamental to the formation of housing rent patterns. In addition, the positive coefficients of the kernel densities of the selected POIs confirm that the abundance of facility/service supplies and local amenities can provide added values to residential quarters. This is in line with the common experiences, such as that residential quarters close to colleges (e.g. k_{HEdu}) or working places (e.g. k_{Enter}) usually have higher level of housing rents (Gibbons & Machin, 2008), but in turn they provide social-cultural amenities and also better accessibility from the perspective of daily commuting. The assessed errors of this regression model are shown in Table 2 (%MAE = 19.63% and %RMSE = 21.61%).

Table 2
Linear regression coefficients of the explanatory variables (Number of observations = 1028; F-statistics = 218.43; R^2 = 0.63).

	NTL	E_{NDVI}	k_{Edu}	k_{HEdu}	k_{Enter}
Coefficient	0.10**	16.18*	1.72*	1.18***	11.59***
t-statistics	2.76	2.03	2.04	4.21	9.10
	k_{Com}	k_{Hosp}	k_{Metro}	Constant	
Coefficient	0.39***	3.71*	3.19***	16.83***	
t-statistics	1.92	1.99	10.40	11.8	
MAE = 6.28, %MAE = 19.63%, RMSE = 8.05, %RMSE = 21.61%					

Note: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

The performance of the six individual classifiers was also evaluated before the implementation of ensemble (Table 3 and Fig. 6). The parameter configuration files of these classifiers in WEKA format are provided in Appendix A. The results indicate that their performance is generally good, with the MAE ranging from 5.29 to 5.68 (%MAE = [16.40–17.66%]) and the RMSE from 6.98 to 7.43 (%RMSE = [18.75–19.96%]). Among these classifiers, SVR, k-NN and GPR have similar prediction errors of %MAE ≈ 16.69% and %RMSE ≈ 18.82%, which are lower than those for BP-NN, RBF-NN and FDT (%MAE ≈ 17.62% and %RMSE ≈ 19.53%). Due to the insensitivity to the number of dimensions, SVR has been reported to be the most robust and accurate method than others in machine learning (X. Wu et al., 2008). Previous studies also revealed that k-NN as a kind of lazy learning method also has satisfactory performance if the training datasets are reliable (Kuramochi & Karypis, 2005). Thus, it is not surprising that SVR and k-NN are among the best individual classifiers in our preliminary experiments.

The ensemble approach of bagging was applied to each of the individual classifiers with the default parameter settings of WEKA for bagging (Appendix A), in which the sampling procedures are applied 10 times. The individual classifiers after bagging are denoted as B_GPR, B_k-NN, B_BP-NN, B_RBF-NN, B_FDT and B_SVR, respectively. The results shown in Table 4 reflect that the effectiveness of bagging varies from one individual classifier to another. B_RBF-NN (MAE = 5.55) and B_FDT (MAE = 5.49) gain the largest improvements in the prediction errors by comparing with their performance before bagging (5.68 and 5.62, respectively). B_BP-NN also achieved better outcomes than BP-NN, although the improvement is quite trivial (with reduced %MAE from 17.58% to 17.52%). These experimental results, which are consistent to the empirical findings in previous studies (Kim & Kang, 2010), suggest that the bagging approach is particularly effective for improving classifiers with neural networks or tree structures. However, compared with the performance of SVR, k-NN and GPR, their bagging forms (i.e. B_SVR, B_GPR and B_k-NN) cannot achieve any improvement but have even larger errors. Nevertheless, these three bagging classifiers still outperform B_RBF-NN, B_FDT and B_BP-NN. In summary, the experiments with our dataset of mean unit rent have demonstrated that the bagging approach can narrow down the gap of prediction errors between the 'good' classifiers and the 'bad' ones, but eventually it doesn't improve too much the performance of the 'good' classifiers. Therefore, we resorted to another ensemble approach of stacking (please see Appendix A for the parameter configurations of stacking).

Because we have six individual classifiers, there are five combinations of them to form the stacking classifiers. Therefore, we first ranked the individual classifiers according to their MAE in ascending order, i.e. SVR < GPR < k-NN < BP-NN < FDT < RBF-NN. Then we grouped them into five combinations:

Stacking #1: SVR + GPR;
Stacking #2: SVR + GPR + k-NN;
Stacking #3: SVR + GPR + k-NN + BP-NN;

Table 3
Prediction errors of the individual classifiers.

	GPR	k-NN	BP-NN	RBF-NN	FDT	SVR
MAE	5.33	5.35	5.56	5.68	5.62	5.29
%MAE	16.76%	16.91%	17.58%	17.66%	17.62%	16.40%
RMSE	6.98	6.99	7.12	7.43	7.26	7.04
%RMSE	18.75%	18.78%	19.12%	19.96%	19.50%	18.92%

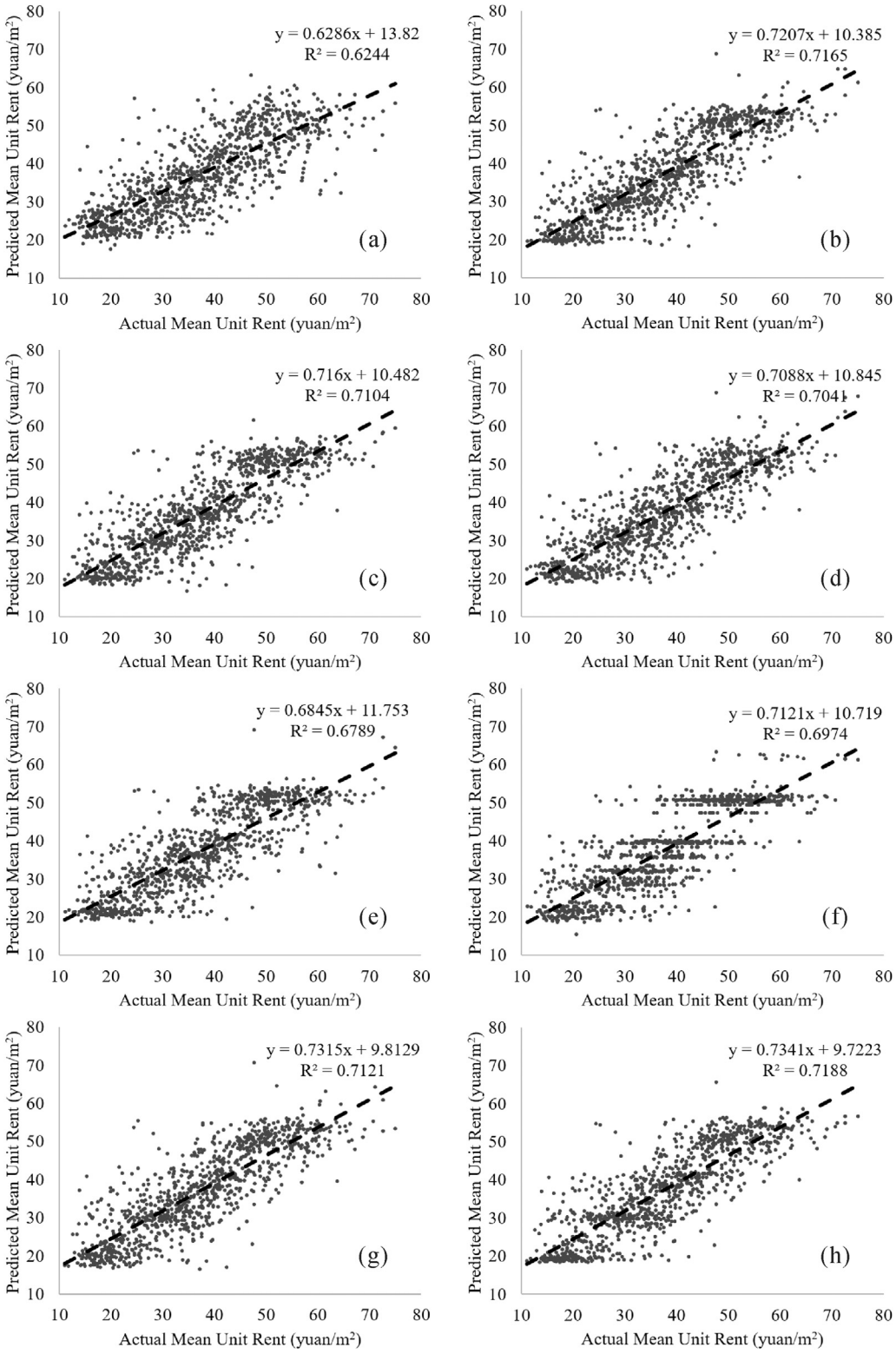

Fig. 6. Actual mean unit rent vs. predicted mean unit rent by individual classifiers of (a) Linear regression, (b) GPR, (c) *k*-NN, (d) BP-NN, (e) RBF-NN, (f) FDT, (g) SVR and (h) Stacking #2.

Stacking #4: SVR + GPR + *k*-NN + BP-NN + FDT; and
Stacking #5: SVR + GPR + *k*-NN + BP-NN + FDT + RBF-NN.

We also conducted multiple experiments to identify the best meta-classifier, and eventually we found that SVR as the meta-classifier ensured the most accurate results. Therefore, SVR was used as the meta-classifier for Stacking #1 to #5. The prediction errors of the five stacking classifiers are shown in Table 5. It can be found that by combing SVR and GPR (Stacking #1), the predictions errors (MAE = 5.26 and RMSE = 6.99) have already reduced to be less than those for either of the two input classifiers (Table 3). With

Table 4
Prediction errors of the bagging classifiers.

	B_GPR	B_*k*-NN	B_BP-NN	B_RBF-NN	B_FDT	B_SVR
MAE	5.33	5.40	5.56	5.55	5.49	5.31
%MAE	16.77%	16.92%	17.52%	17.25%	17.23%	16.51%
RMSE	6.98	7.06	7.13	7.27	7.11	7.03
%RMSE	18.76%	18.97%	19.54%	19.54%	19.11%	18.87%

Table 5
Prediction errors of the stacking classifiers (meta-classifier = SVR).

	Stacking #1	Stacking #2	Stacking #3	Stacking #4	Stacking #5
MAE	5.26	5.25	5.27	5.29	5.27
%MAE	16.32%	16.32%	16.32%	16.38%	16.34%
RMSE	6.99	6.96	6.96	6.98	6.96
%RMSE	18.77%	18.69%	18.70%	18.76%	18.70%

an extra classifier added, i.e. the *k*-NN, the MAE of Stacking #2 decreases into 5.25. However, the performance could not be improved but become worse after including BP-NN and FDT into the stacking classifier (Stacking #3 and #4). The prediction errors of Stacking #5 (MAE = 5.27 and RMSE = 6.96), which is the ensemble of all individual classifiers, are approximately equal to those of Stacking #3 but higher than the errors of Stacking #1. Overall, Stacking #2 has the best performance among all stacking classifiers. The ensemble approach of stacking also outperforms all individual classifiers and the bagging classifiers in terms of the prediction errors (Tables 3 and 4).

We employed the Wilcoxon signed ranks test (Demšar, 2006) to further assess the significance of the differences between Stacking #2, which is selected as the control classifier (i.e. the best one), and the remaining ensemble (bagging and stacking) classifiers. The Wilcoxon signed ranks test is a non-parametric test that ranks the performance differences between two classifiers. Given N data sets, let d_i be the performance difference (e.g. MAE) between two classifiers on data set i. All of the differences are ranked by their absolute values, in which a lower rank is assigned to the smaller absolute difference. Average ranks are assigned if equal absolute values exist. These ranks are summed up separately according to the signs of the differences:

$$R^+ = \sum_{d_i>0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i) \quad (5)$$

$$R^- = \sum_{d_i<0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i) \quad (6)$$

where R^+ and R^- represent the sum of ranks for non-negative and non-positive differences, respectively. Let $T = \min(R^+, R^-)$. With a given confidence level (e.g. $\alpha = 0.05$), if T is smaller than or equal the critical value, the null hypothesis that two classifiers perform equally well can be rejected. The critical values for different values of N and confidence levels are provided in (Demšar, 2006). We ran each of the ensemble classifiers five times ($N = 5$) on the housing rent data set, and derived the differences in terms of MAE between Stacking #2 and the remaining classifiers. The results and their ranks are shown in Table 6. For Stacking #2, R^+ is always less than R^- (i.e. $T = R^+$), and also satisfies the condition of equal or less than the critical value of 5 ($N = 5$, $\alpha = 0.05$; see (Demšar, 2006)). Therefore, it is evident that Stacking #2 significantly outperforms all of the other ensemble classifiers compared, and can be used to estimate the unknown mean unit rent.

Fig. 7(a) shows the composite of the original and the estimated mean unit rent at the NC-level. Based on these results, the spatial characteristics of housing rent in metropolitan Guangzhou can be revealed. By visually inspecting Fig. 7(a), it is evident that the mean unit rent is the highest in the central area of the city and gradually declines toward the outskirts. To quantify this pattern, we used a power function to estimate the relationship between distance from the city center and the NC-level mean unit rent. In this estimation, for NCs that already have observations of housing rent data, we directly used them as the inputs. For the NCs with unknown rents, we used the estimated rents as the inputs. Therefore, rather than a limited number of samples, all NCs were accounted in the estimation. As illustrated by Fig. 7(c), the estimated exponent is −0.327, which indicates a strong distance decay effect. This is in line with our own experiences about this city and the findings in previous research as well. That is, the city has formed a matured core (i.e. the CBD) at the macro scale (Shin, 2014). In this sense, the housing rent map we produced is reliable.

Table 6
MAE differences between Stacking #2 (the control classifier) and the remaining ensemble classifiers (the ranks are shown in the parentheses).

N	B_GPR	B_*k*-NN	B_BP-NN	B_RBF-NN	B_FDT
1	−0.0665 (3)	−0.1351 (3)	−0.3000 (2)	−0.2867 (4)	−0.2266 (3)
2	−0.0655 (2)	−0.1246 (2)	−0.2951 (1)	−0.2804 (2)	−0.2344 (4)
3	−0.1664 (5)	−0.1825 (5)	−0.3152 (4)	−0.3055 (5)	−0.2641 (5)
4	−0.0483 (1)	−0.1405 (4)	−0.3197 (5)	−0.2346 (1)	−0.2172 (2)
5	−0.0742 (4)	−0.1194 (1)	−0.3034 (3)	−0.2853 (3)	−0.2064 (1)
R^+	0	0	0	0	0
R^-	15	15	15	15	15
N	B_SVR	Stacking #1	Stacking #3	Stacking #4	Stacking #5
1	−0.0412 (3)	0.0102 (2)	−0.0020 (1)	−0.0269 (2)	−0.0068 (1)
2	−0.0380 (2)	−0.0112 (3)	−0.0514 (3)	−0.0682 (3)	0.0122 (3)
3	−0.1323 (5)	−0.0629 (5)	−0.0810 (4)	−0.0981 (5)	−0.1112 (5)
4	−0.0344 (1)	−0.0255 (4)	−0.0865 (5)	−0.0929 (4)	−0.0706 (4)
5	−0.0602 (4)	0.0096 (1)	−0.0184 (2)	−0.0105 (1)	0.0074 (2)
R^+	0	3	0	0	5
R^-	15	12	15	15	10

4.2. Discussions

There are several implications that can be drawn from the results discussed above. Firstly, our experiments have demonstrated the usefulness of the online rental listings and other auxiliary data (e.g. nighttime lights and POIs) for mapping housing rent at a relatively fine-scale (i.e. the NC-level). All of these data sets share two promising advantages: (1) they are generally open access, including the housing rent data and POIs, and hence one can collect these data for related analysis by implementing the web-crawlers; (2) the online rental listings are usually updated at much finer temporal interval than conventional statistical housing data. While the explanatory variables of nighttime lights and POIs are also regularly updated, one can easily capture the spatial-temporal variations of housing rent annually, monthly or even weekly. Although in this study only the year 2015 data are assembled to produce the spatial pattern of housing rent, the time-series mapping of housing rent, which enables the long-term monitoring and analysis of local residential markets, can be anticipated if the continuous collection of corresponding data sets is achieved.

Secondly, we demonstrate the utility of ensemble learning methods in the production of spatial housing data. The results reflect that ensemble learning can effectively reduce the overall prediction errors. However, the actual performance of ensemble learning could be significantly affected by the original individual classifier. In this study, the selected individual classifiers are the most well-known ones, which have already been applied in a vast

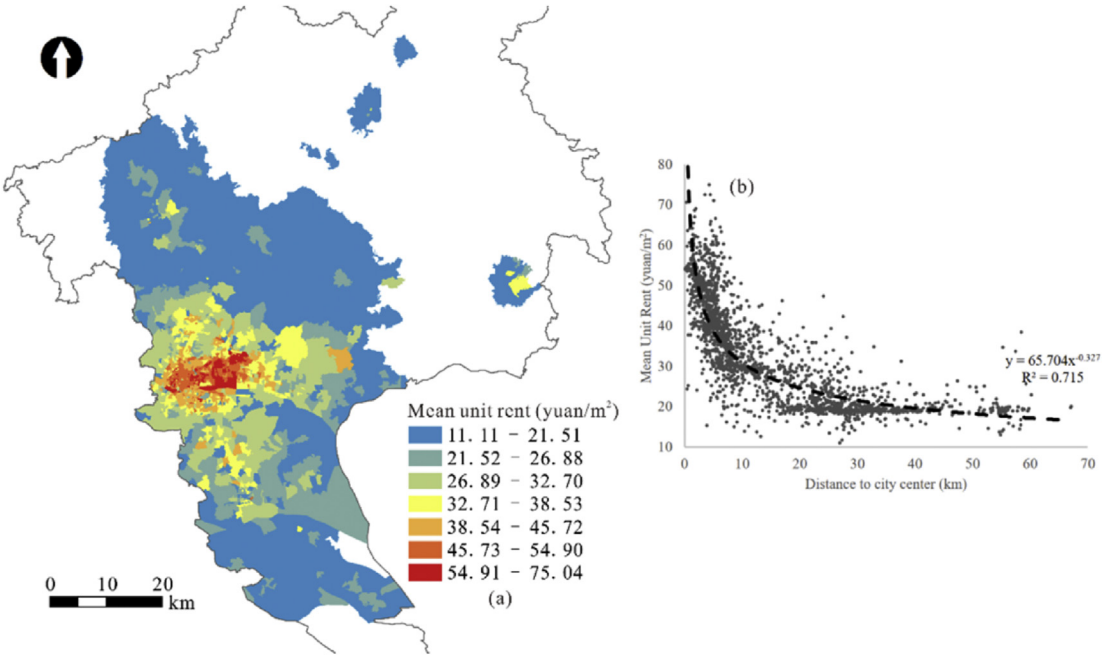

Fig. 7. (a) The estimated mean unit rent at the NC-level. (b) Distance decay in the NC-level mean unit rents.

number of situations (Graczyk et al., 2010; Srivastava, 2014). Therefore, the effectiveness of ensemble learning can be examined more clearly for these representative individual classifiers. Consistent with the findings in previous research (Wang, Hao, Ma, & Jiang, 2011), the ensemble learning approach of bagging in our case works better for algorithms based on neural networks and tree structures. Even so, the predictions made by the bagging neural networks and tree structure based algorithms are still less accurate than those from support vector regression. Another important lesson learned from the experiments is that ensemble learning by combining some classifiers might work better than combining them all. This is evident in the stacking experiments, in which the best model is the combination of SVR, GPR and *k*-NN (Stacking #2) rather than all of the six classifiers (Stacking #3, #4 and #5). Therefore, despite the better performance of the stacking approach, one should be cautious to bring in the individual classifiers instead of pooling all the classifiers at hand.

Besides the housing research, machine learning methods and their ensemble forms also have promising potential to solve other problems in social sciences and human geography, such as land use legacies (Tayyebi, Pijanowski, & Pekin, 2015), public health research (Grubesic, Miller, & Murray, 2014) and demographic mapping (Grekousis & Thomas, 2012). Indeed, machine learning methods are particularly useful for social sciences and human geography, in which complex relationships exist and cannot be easily captured by traditional analytical methods. Machine learning methods are not only applicable in cases with available complete data sets, but also adaptive to problems with incomplete data sets. Traditionally researchers adopt the statistical techniques of small area estimation and missing value imputation methods to address the missing data problems that frequently occur in demographic research. However, recent studies of Jerez et al. (2010) and Nelwamondo, Golding, and Marwala (2013) consistently report the better performance of machine learning methods (e.g. *k*-NN, neural network and decision trees) over the traditional statistical techniques for missing data imputation. However, despite the advantage, interpreting the results of machine learning methods might be difficult (e.g. the neural network as 'black box'), because these methods make predictions/classifications purely based on the interactions of data (Jerez et al., 2010). Therefore, it should be cautious to apply machine learning methods if the objective is to mimic the mechanism and causality of social phenomena.

Thirdly, in a broader context, our results can contribute to the understanding of urban residential dynamics, especially for the Chinese cities. For instance, the recent rush-up in Chinese cities' housing prices has raised the concern of 'housing bubbles', which refers to the excessive deviation of housing price from the fundamental housing values (Hou, 2010). In this sense, the resulting spatial distributions of housing rent in our work can be served as an input variable to calculate the price-to-rent ratio, which is frequently used along with other factors (e.g. income and financial conditions) to measure the likelihood of 'housing bubbles'. Moreover, as the generated mean unit rent are at the NC-level, i.e. the very basic level of administrative divisions, they are compatible with other socioeconomic statistics in the mainstream data sources (e.g. population census). This advantage can facilitate the research of social problems related to residential conditions, including housing affordability (J. Chen et al., 2010), urban poverty (He et al., 2010) and residential segregation (Z. Li & Wu, 2008). Finally, housing rent is also an important variable to measure the land value for residential uses. While a typical mode of urban expansion in contemporary China is the rapid development of real estates (S. Hu et al., 2012), the mapping of housing rent can provide valuable information to indicate the orientation as well as the potential return of future growth.

5. Conclusions

In this study, we have explored the utility of online rental listings for mapping the spatial pattern of housing rent in the metropolitan Guangzhou. It is expected that our estimation of the NC-level housing rent pattern can be integrated with other information to provide better understanding of urban residential dynamics and related problems. To this end, we established a prediction model based on ensemble learning using the input variables of VIIRS nighttime lights, NDVI and several types of POIs. The

experimental results indicate that the ensemble approach of bagging can effectively reduce the prediction errors of classifiers with neural networks or tree structures (e.g. RBF-NN and FDT), whereas the stacking approach produces ensemble classifiers with much better performance than the bagging classifiers. As a result, we chose Stacking #2, the classifier with the lowest errors (% MAE = 16.32% and %RMSE = 18.69%), to estimate the mean unit rent at the NC-level. Based on the mapping results, we identified a significant distance decay relationship in the NC-level mean unit rents with respect to the city center. This reveals that the city has a matured core (i.e. the CBD) at the macro scale. Although the established ensemble model can effectively improve the prediction accuracy of housing rent, the performance can be further refined by introducing variables of social characteristics, such as household incomes and financial conditions. Moreover, the presented method can also be applied in a finer scale if extra spatial data are available. We have explained in Section 2 that the choice of the NC-level for mapping the housing rent is a compromising decision due to the absence of the spatial footprints of residential quarters. Therefore, we will seek to acquire more useful data in the future so that the spatial pattern of urban housing rent can be delineated with an even finer resolution.

Acknowledgements

We thank the anonymous reviewers for their valuable comments and suggestions. This research was supported by the Key National Natural Science Foundation of China (Grant No. 41531176), the National Natural Science Foundation of China (Grant No. 41601420, 41371376, 41301452, and 41401432), the Guangdong Natural Science Foundation (Grant No. 2015A030310288), and the Fundamental Research Funds for the Central Universities (16lgpy03).

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.apgeog.2016.08.011.

References

Ahlfeldt, G. (2011). If Alonso was right: Modeling accessibility and explaining the residential land gradient. *Journal of Regional Science, 51*(2), 318–338.
Anselin, L., & Le Gallo, J. (2006). Interpolation of air quality measures in hedonic house price models: Spatial aspects. *Spatial Economic Analysis, 1*(1), 31–52.
Arribas-Bel, D. (2014). Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography, 49*, 45–53.
Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., et al. (2012). Smart cities of the future. *The European Physical Journal Special Topics, 214*(1), 481–518.
Bitter, C., Mulligan, G. F., & Dall'erba, S. (2007). Incorporating spatial variation in housing attribute prices: A comparison of geographically weighted regression and the spatial expansion method. *Journal of geographical systems, 9*(1), 7–27.
Chen, J., Hao, Q., & Stephens, M. (2010). Assessing housing affordability in post-reform China: A case study of Shanghai. *Housing Studies, 25*(6), 877–901.
Chen, Y. M., Li, X., Wang, S. J., & Liu, X. P. (2012). Defining agents' behaviour based on urban economic theory to simulate complex urban residential dynamics. *International Journal of Geographical Information Science, 26*(7), 1155–1172.
China Statistical Bureau. (2015). *China statistical yearbook 2015*. Beijing: China Statistical Bureau.
Davis, M. A., & Ortalo-Magné, F. (2011). Household expenditures, wages, rents. *Review of Economic Dynamics, 14*(2), 248–261.
Delang, C. O., & Lung, H. C. (2010). Public housing and poverty concentration in urban neighbourhoods: The case of Hong Kong in the 1990s. *Urban Studies, 47*(7), 1391–1413.
Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research, 7*, 1–30.
Dubé, J., & Legros, D. (2014). Spatial econometrics and the hedonic pricing model: What about the temporal dimension? *Journal of Property Research, 31*(4), 333–359.
Frappa, S., & Mésonnier, J.-S. (2010). The housing price boom of the late 1990s: Did inflation targeting matter? *Journal of Financial Stability, 6*(4), 243–254.
Gibbons, S., & Machin, S. (2008). Valuing school quality, better transport, and lower crime: Evidence from house prices. *Oxford Review of Economic Policy, 24*(1), 99–119.
Graczyk, M., Lasota, T., Trawiński, B., & Trawiński, K. (2010). Comparison of bagging, boosting and stacking ensembles applied to real estate appraisal. In *Intelligent information and database systems* (pp. 340–350). Springer.
Grekousis, G., & Thomas, H. (2012). Comparison of two fuzzy algorithms in geo-demographic segmentation analysis: The Fuzzy C-Means and Gustafson–Kessel methods. *Applied Geography, 34*, 125–136.
Grubesic, T. H., Miller, J. A., & Murray, A. T. (2014). Geospatial and geodemographic insights for diabetes in the United States. *Applied Geography, 55*, 117–126.
Guo, W., Lu, D., Wu, Y., & Zhang, J. (2015). Mapping impervious Surface distribution with integration of SNNP VIIRS-DNB and MODIS NDVI Data. *Remote Sensing, 7*(9), 12459–12477.
Hanson, A., & Hawley, Z. (2011). Do landlords discriminate in the rental housing market? Evidence from an internet field experiment in US cities. *Journal of Urban Economics, 70*(2), 99–114.
He, S., Wu, F., Webster, C., & Liu, Y. (2010). Poverty concentration and determinants in China's urban low-income neighbourhoods and social groups. *International Journal of Urban and Regional Research, 34*(2), 328–349.
Hogan, B., & Berry, B. (2011). Racial and ethnic biases in rental housing: An audit study of online apartment listings. *City & Community, 10*(4), 351–372.
Hou, Y. (2010). Housing price bubbles in Beijing and Shanghai? A multi-indicator analysis. *International Journal of Housing Markets and Analysis, 3*(1), 17–37.
Hu, S., Cheng, Q., Wang, L., & Xie, S. (2012). Multifractal characterization of urban residential land price in space and time. *Applied Geography, 34*, 161–170.
Hu, T., Yang, J., Li, X., & Gong, P. (2016). Mapping urban land use by using Landsat images and open social data. *Remote Sensing, 8*(2), 151.
Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., et al. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine, 50*(2), 105–115.
Kemp, P. A. (2011). Low-income tenants in the private rental housing market. *Housing Studies, 26*(7–8), 1019–1034.
Kim, M.-J., & Kang, D.-K. (2010). Ensemble with neural networks for bankruptcy prediction. *Expert Systems with Applications, 37*(4), 3373–3379.
Kuramochi, M., & Karypis, G. (2005). Gene classification using expression profiles: A feasibility study. *International Journal on Artificial Intelligence Tools, 14*(04), 641–660.
Li, X., Liu, X., & Yu, L. (2014). Aggregative model-based classifier ensemble for improving land-use/cover classification of Landsat TM Images. *International Journal of Remote Sensing, 35*(4), 1481–1495.
Liu, X., Li, X., Chen, Y., Tan, Z., Li, S., & Ai, B. (2010). A new landscape index for quantifying urban expansion using multi-temporal remotely sensed data. *Landscape Ecology, 25*(5), 671–682.
Liu, X. P., Ma, L., Li, X., Ai, B., Li, S. Y., & He, Z. J. (2014). Simulating urban growth by integrating landscape expansion index (LEI) and cellular automata. *International Journal of Geographical Information Science, 28*(1), 148–163.
Li, Z., & Wu, F. (2008). Tenure-based residential segregation in post-reform Chinese cities: A case study of Shanghai. *Transactions of the Institute of British Geographers, 33*(3), 404–419.
Lopez-Morales, E. (2011). Gentrification by Ground Rent Dispossession: The Shadows Cast by Large-Scale Urban Renewal in Santiago de Chile. *International Journal of Urban and Regional Research, 35*(2), 330–357.
Ma, T., Zhou, C., Pei, T., Haynie, S., & Fan, J. (2014). Responses of Suomi-NPP VIIRS-derived nighttime lights to socioeconomic activity in China's cities. *Remote Sensing Letters, 5*(2), 165–174.
Nelwamondo, F. V., Golding, D., & Marwala, T. (2013). A dynamic programming approach to missing data estimation using neural networks. *Information Sciences, 237*, 49–58.
Osland, L. (2010). An application of spatial econometrics in relation to hedonic house price modeling. *Journal of Real Estate Research, 32*(3), 289–320.
Partridge, M. D., Rickman, D. S., Ali, K., & Olfert, M. R. (2010). Recent spatial growth dynamics in wages and housing costs: Proximity to urban production externalities and consumer amenities. *Regional Science and Urban Economics, 40*(6), 440–452.
Rae, A. (2015). Online housing search and the geography of submarkets. *Housing Studies, 30*(3), 453–472.
Rondinelli, C., & Veronese, G. (2011). Housing rent dynamics in Italy. *Economic Modelling, 28*(1), 540–548.
Shin, H. B. (2014). Urban spatial restructuring, event-led development and scalar politics. *Urban Studies, 51*(14), 2961–2978.
Shi, K., Yu, B., Huang, Y., Hu, Y., Yin, B., Chen, Z., et al. (2014). Evaluating the ability of NPP-VIIRS nighttime light data to estimate the gross domestic product and the electric power consumption of China at multiple scales: A comparison with DMSP-OLS data. *Remote Sensing, 6*(2), 1705–1724.
Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (vol. 26). CRC press.
Srivastava, S. (2014). Weka: A tool for data preprocessing, classification, ensemble, clustering and association rule mining. *International Journal of Computer Applications, 88*(10).
Tayyebi, A., Pijanowski, B. C., & Pekin, B. K. (2015). Land use legacies of the Ohio River Basin: Using a spatially explicit land use change model to assess past and future impacts on aquatic resources. *Applied Geography, 57*, 100–111.
Waltert, F., & Schläpfer, F. (2010). Landscape amenities and local development: A review of migration, regional economic and hedonic pricing studies. *Ecological*

Economics, 70(2), 141–152.
Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications, 38*(1), 223–230.
Windeatt, T., & Ardeshir, G. (2004). Decision tree simplification for classifier ensembles. *International Journal of Pattern Recognition and Artificial Intelligence, 18*(05), 749–776.
Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., et al. (2008). Top 10 algorithms in data mining. *Knowledge and information systems, 14*(1), 1–37.
Wu, W., Zhang, W., & Dong, G. (2013). Determinant of residential location choice in a transitional housing market: Evidence based on micro survey from Beijing. *Habitat International, 39*, 16–24.
Yi, C., & Huang, Y. (2014). Housing consumption and housing inequality in Chinese cities during the first decade of the twenty-first century. *Housing Studies, 29*(2), 291–311.
Yu, D., Wei, Y. D., & Wu, C. (2007). Modeling spatial dimensions of housing prices in Milwaukee, WI. *Environment and Planning B: Planning and Design, 34*(6), 1085–1102.
Zhou, Z. (2012). *Ensemble methods: Foundations and algorithms*. Boca Raton: CRC Press.