Manuscript

B Urban Analytics and City Science

EPB: Urban Analytics and City Science
2023, Vol. 0(0) 1–21

DOI: 10.1177/23998083231158370
journals.sagepub.com/home/epb
SAGE

Quantifying the environmental characteristics influencing the attractiveness of commercial agglomerations with big geo-data

Zhou Huang and Ganmin Yin
Institute of Remote Sensing and Geographical Information Systems, Peking University, China
Beijing Key Lab of Spatial Information Integration & Its Applications, Peking University, China

Xia Peng
Tourism College, Beijing Union University, China

Xiao Zhou and Quanhua Dong
Institute of Remote Sensing and Geographical Information Systems, Peking University, China
Beijing Key Lab of Spatial Information Integration & Its Applications, Peking University, China

Abstract

Understanding the attractiveness of commercial agglomerations contributes to urban planning. Existing studies focus less on commercial agglomerations, and most directly use environmental supply factors to characterize attractiveness. This study measures attractiveness from the perspective of human demand. Specifically, we build a novel bipartite graph based on big geo-data of human mobility, using node centralities (degree, betweenness, and pagerank) to measure attractiveness. Next, we summarize multisource environmental features such as Point-of-Interests (POIs), land cover, transportation, and population, and use them as inputs to accurately predict attractiveness based on random forest. Finally, the spatial heterogeneity of the effects of these environmental variables on attractiveness is analyzed by multiscale geographically weighted regression. The results of the Beijing case show that: (1) All three centralities show a trend that the urban center is higher than the surrounding area, and betweenness is more reasonable. (2) Random forest can accurately predict attractiveness, with R^2 for degree, betweenness, and pagerank at 0.903, 0.846, and 0.760, respectively. (3) The number of shopping POIs, the length of main roads, and the number of bus stops positively affect attractiveness, while the effects of greening ratio and population density are bidirectional. As for the service scope, about 70% of commercial agglomerations have an average service radius of less than 15 km, which is significantly correlated with the Voronoi

Corresponding author:
Xia Peng, Tourism College, Beijing Union University, Beijing 100101, China.
Email: ivy_px@163.com

diagram. Our results can inspire understanding the human–environment relationship and guide urban policymakers in business planning.

Keywords
Urban planning, commercial agglomeration attractiveness, bipartite graph centrality, supply and demand, human–environment relationship, big geo-data

Introduction

As an important element of the built environment, a shop refers to a functional place that sells goods or services to provide residents with shopping, dining, entertainment, and other living needs. This supply–demand interplay is the embodiment of the coupling relationship between human and the built-up environment. In many cases, the shops in the city do not exist alone, but show a clustering effect, forming a series of commercial agglomerations, thereby enhancing the city's economic vitality (Bettencourt et al., 2007; Gomez-Lievano et al., 2016). Compared to the concept of the trading area (Huff, 1964), the commercial agglomeration in this study focuses only on its infrastructure part and does not consider the scope of services. Commercial agglomeration is a geographical object on a mesoscopic scale between a shop and a city (Hall, 2014; Dong et al., 2020), the latter two are more studied, and the former is less involved (Ballantyne et al., 2022; Seong et al., 2022; Yang et al., 2019), which is still a relatively research gap.

Understanding the attractiveness of commercial agglomerations can help us better carry out business planning and urban design. So far, much work has studied attractiveness at the city and shop levels, with relatively little consideration given to the scale of commercial agglomerations (Estiri et al., 2021; Kunc et al., 2016; Romao et al., 2018; Sobolevsky et al., 2015). In addition, there are two main ways to portray business attractiveness in the existing studies: (1) *interviews and questionnaires* (González-Hernández and Orozco-Gómez, 2012; Wong et al., 2001), which require surveyed customers to have strong imaginary ability and high shopping involvement, and have disadvantages like subjectivity, labor-intensive and coarse-resolution; (2) *the combination of environmental variables* (such as commodity diversity, environmental comfort, transportation convenience, population congestion) as input to calculate the probability that a customer from an origin will travel to different destinations. They can be roughly divided into three categories (Kunc et al., 2016; Teller and Reutterer, 2008): spatial interaction models, such as the Huff model (Huff, 1964; Liang et al., 2020) and the competing destinations models (De Mello-Sampayo, 2017; Fotheringham, 1983; Ishikawa, 1987), random utility models (Grashuis et al., 2020; Oppewal et al., 1997), and multiplicative competitive interaction (MCI) models (Cooper and Nakanishi, 1989; Wu et al., 2019). These environmental factors do reflect attractiveness to some extent, that is, there is an "environment-attractiveness" mapping relationship, but there is little work to deeply explore more details of the mapping relationship, such as "to what extent" and "how to reflect."

With the development of positioning technology, location-aware devices, and mobile internet, big geo-data containing individual movement information is constantly generated and stored. These big geo-data include mobile phone positioning data, smart card data, taxi trajectory data, etc (Bao et al., 2021; Wang et al., 2022). Compared with traditional questionnaires, big geo-data has the advantages of wide-coverage, low-cost, high-resolution and has been widely used in human mobility, traffic optimization, urban planning (Chen et al., 2018; Liu et al., 2015; Wang et al., 2020, 2022; Yin et al., 2022), which also provides us with a new perspective for studying the attractiveness of commercial agglomerations. The most intuitive way to measure the attractiveness of a thing is to see how people react to it. For example, the number of people who buy can represent the

attractiveness of goods, because people do "vote with their feet." Big geo-data can objectively and quantitatively reflect human demand (Giglio et al., 2019; Liu et al., 2015; Wang et al., 2021). Compared with the above environmental supply, it is a more reasonable way to characterize attractiveness, and studying this supply–demand relationship is also helpful in understanding the human–environment relationship.

Batty (2013) systematically explores the application of bipartite graphs in urban studies, which are widely used in business analysis, transportation modeling, and urban management (Chakraborty et al., 2019; Eubank et al., 2004; Von Ferber et al., 2009). Bipartite graphs can model two types of entities and their relationships, such as user-item, road-intersection, and people-place (Asratian et al., 1998). Considering the characteristics of the Origin-Destination (OD) dataset used, if the source grids (Origin) analogous users, the commercial agglomerations (Destination) analogous items, with the OD flow between them as the links, we can establish a "source grid-commercial agglomeration" bipartite graph, which can be regarded as a user-item graph at the aggregate level or a larger scale. This enlightens us to construct a bipartite graph based on the human demand inferred from the OD flow, and then we can use the indicators of the node importance in the graph to portray the attractiveness of commercial agglomerations.

After characterizing attractiveness, as mentioned earlier, two questions remain to be answered, one is the extent to which environmental variables reflect attractiveness, and the other is how these variables affect attractiveness. As data-driven and black-box models, machine learning models represented by random forests and neural networks have a powerful ability to model irregular patterns (Breiman, 2001; Drucker et al., 1996; Jordan and Mitchell, 2015). It can help us better predict attractiveness based on environmental variables to answer the first question. And its practical significance is to help us estimate the popularity of new commercial centers among residents before designing. In addition, due to spatial heterogeneity, the relationship between these variables and attractiveness is usually not spatially stationary. Geographically weighted regression (GWR) allows its regression coefficients to vary with spatial location, so it can effectively discover the spatial variation (Brunsdon et al., 1996) and answer the second question. It has been widely adopted in the driving factors analysis of carbon emission (Zhou et al., 2022), public transport usage (Chiou et al., 2015), and housing price (Kang et al., 2021). This spatial analysis can guide us in the targeted allocation of urban infrastructure to improve business attractiveness, such as knowing which areas depend more on point-of-interests (POIs) and which areas show more urgent transportation needs.

Therefore, this paper takes the commercial agglomeration as the research object, from the side of the human demand rather than the environmental supply, to more objectively and intuitively characterize the attractiveness, and further analyze its influencing factors. Specifically, based on the mobile phone positioning data, we extract the travel flow from the source grids to the commercial agglomerations, and construct a bipartite graph to calculate the centrality of the nodes as the measure of attractiveness. Then, after summarizing various environmental variables (including: POIs, land cover, transportation, population density), we select the random forest model to predict the attractiveness of commercial agglomerations, and analyze the importance of each variable. Finally, we further adopt the multiscale geographically weighted regression to study the spatial heterogeneity of the relationship between environmental variables and attractiveness.

The main contributions of this research are the following three points:

- We measure the attractiveness of the commercial agglomeration from the human demand side rather the environmental supply, using the node centralities in the bipartite graph.
- We use the random forest model to achieve accurate predictions of attractiveness, with environmental features like POIs, land cover, transportation and population density as input.
- We adopt the multiscale geographically weighted regression to analyze the spatial non-stationarity of the effects of some typical environmental features on attractiveness.

The rest of the paper is organized as follows. Section "Study area and dataset" provides an overview of the study area and datasets. Section "Methodology" describes the roadmap of the study: feature engineering, bipartite graph modeling, attractiveness prediction, and spatial heterogeneity analysis. Section "Results" and "Discussions" presents the interesting findings and important results of our research. Section "Conclusions" summarizes the main conclusions and improvement directions of the study.

Study area and dataset

Study area

The study area is Beijing, the capital of China, which is also a commercially developed and densely populated economic center in China, as shown in Figure S1 in the Supplementary Material. The commercial agglomeration data is obtained from an open-access dataset (Peng et al., 2022), which contains the built-up footprints of 249 commercial agglomerations in Beijing, 2019. The dataset includes information such as ID, Chinese and English names, and geometry in the format of Esri shapefile and the coordinate system of WGS 84. Figure 1(a) shows the study area and the footprints of some typical commercial agglomerations in Beijing.

Dataset

OD flow data is provided by China Unicom Smart Steps Company (http://www.smartsteps.com/), one of China's largest population big data service providers. The dataset is extracted from China Unicom's mobile phone signaling data in May 2019, with a spatial resolution of 250-m. According to the user stay patterns within 1 month, three stay types (residence, work and visit) can be identified, and we have eliminated the human flow whose stay type is residence or work. Further, we aggregate all the flow visiting the commercial agglomerations and count the average number of visitors per day as the intensity of the flow. Finally, we extract 123,769 OD flows and the distribution of the flow intensity is shown in Figure 1(b).

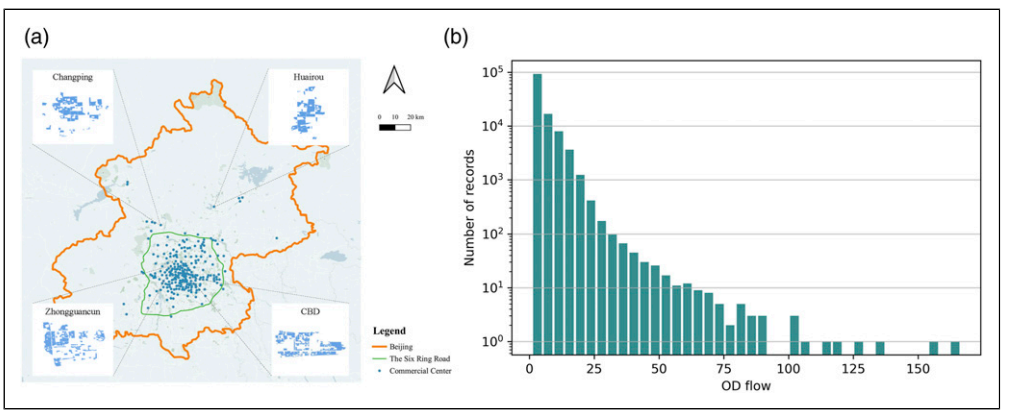

Figure 1. Study area and dataset: (a) Study area and the footprints of some commercial agglomerations. (b) The distribution of the human flow intensity.

Methodology

The research framework of the paper is shown in Figure 2. First is the *feature engineering*, which integrates features like POIs, land cover, transportation, and population for each commercial agglomeration to reflect its environmental supply level. Then is the *bipartite graph modeling*, which extracts the OD flow from source grids to commercial agglomerations based on the mobile phone location data, to construct the bipartite graph, and calculate the centralities of the nodes like degree,

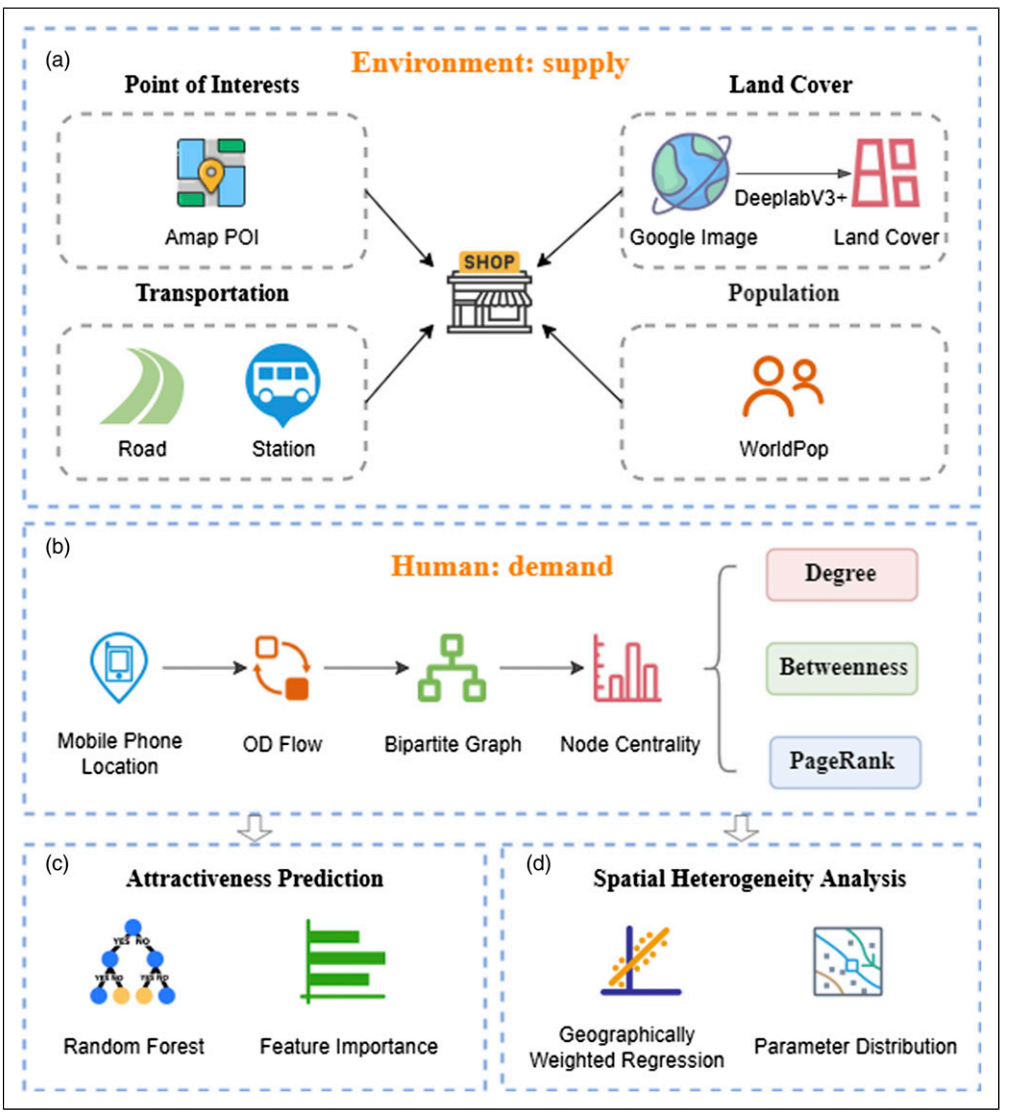

Figure 2. The framework of the study: (a) Feature engineering: integrate environmental factors like POIs, land cover, transportation and population. (b) Bipartite graph modeling: calculate node centrality to measure attractiveness based on human flow. (c) Attractiveness prediction: achieve accurate predictions with environmental factors as inputs and attractiveness as outputs based on random forest. (d) Spatial heterogeneity analysis: analyze spatial variation in the effects of various environmental factors on attractiveness based on the multiscale geographically weighted regression.

betweenness, and pagerank as the measures of attractiveness from the perspective of human demand. Next is the *attractiveness prediction*, which uses the environmental factors of the commercial agglomeration as the independent variables and the attractiveness as the target, realizes a highly accurate prediction based on random forest, and calculates the importance of each feature. The final step is the *spatial heterogeneity analysis*, where we further adopt the multiscale geographically weighted regression to analyze spatial variation in the effects of various environmental factors on attractiveness, taking into account spatial non-stationarity.

Feature engineering

Based on the existing studies, we decide to use data such as POIs, land cover, transportation, and population to characterize the service diversity, environmental comfort, transportation convenience, and population density, thereby reflecting the environmental supply side of the commercial agglomerations.

(1) POIs. We obtain the complete POI data of Beijing through the Amap API (https://lbs.amap.com/api/webservice/summary/), which has information like location and category. We make an intersection between the POIs and the footprints of the commercial agglomerations to associate them. Due to a large number of POI categories, the top 10 categories of POIs related to commercial attributes are kept, as shown in Table S1 in the Supplementary Material, and the number of each category is used as the POI feature of the commercial agglomerations.
(2) Land cover. This physical feature is based on remote sensing imagery through semantic segmentation. Remote sensing imagery is provided by Google Earth in 2018 with a spatial resolution of 1 m. The model is the DeepLabV3 + which adopts a pre-trained Xception on the ImageNet-1k dataset as the backbone (Chen et al., 2018a; Deng et al., 2009). We extract four types of land cover, as shown in Table S2 in the Supplementary Material, and calculate the proportion of the area of each type in each commercial agglomeration as the land cover feature.
(3) Transportation. We use the road network to reflect the convenience of private transportation and the bus/subway stations to measure the accessibility of public transportation. Road network data comes from OpenStreetMap. (https://www.openstreetmap.org/). We choose three categories: residential (street or road generally used for local traffic within settlement), main (highways with higher levels than residential), and others (other roads in addition to the above two categories). We count the length of three types of roads in each commercial agglomeration as the private transportation feature. Bus and subway station data also comes from the Amap API, for each commercial agglomeration, we count the number of bus and subway stations in its 1 km buffer as the public transportation feature.
(4) Population density. On the one hand, population density reflects economic dynamism, on the other hand, may bring congestion, that is, there is a dual effect, so it is necessary to consider the impact of population on attractiveness. The population data comes from WorldPop (https://www.worldpop.org/) for China in 2020, with a resolution of 100-m. We count the average population density of each commercial agglomeration as the population feature.

Bipartite graph modeling

Big geo-data such as mobile phone location data records a wealth of individual movement information, so we can easily extract the flow of people who have reached the commercial agglomerations and their source grids. We use the bipartite graph as an effective tool to model this visiting relationship.

The bipartite graph, also called the bigraph or two-mode graph, can model two types of entities and their relationships. A graph $G = (U, V, E)$ is bipartite if its nodes can be divided into two disjoint subsets U, V, such that each edge in $E \subseteq U \times V$ connects a node in U to one in V. Figure S2(a) in the Supplementary

Material is an example that can represent the relationship between user-item, road-intersection, and people-place. The bipartite graph can also be described by an adjacency matrix W of size $|U| \times |V|$, which can be 0-1 or weighted, w_{ij} represents the weight of the edge between node i in U and node j in V.

In this study, if we treat the commercial agglomerations and source grids as two node sets, and the human flow between them as the edges, then we can build a weighted bipartite graph of "commercial agglomeration-source grid," as shown in Figure S2(b) in the Supplementary Material. A key issue here is how to get the weights of the edges based on the flows. From the perspective of people-place interaction, a highly attractive location tends to attract a large number of residents, and residents tend to travel long distances to visit the location (Wang et al., 2021; Arbués et al., 2016). This means that we need to consider both the flow volume and the flow length to define the edge weights (Wang et al., 2021), as shown below

$$w_{ij} = \frac{vol_{ij} \cdot len_{ij}}{\sum_j vol_{ij} \cdot \sum_j len_{ij}}, 0 \le i \le |U| - 1, 0 \le j \le |V| - 1 \quad (1)$$

where vol_{ij} and len_{ij} denote the volume and length of the flow between the commercial agglomeration i and source grid j, respectively. To avoid the effect of excessive outliers, the weights are normalized by the total volume and length of all the flows that reach the commercial agglomeration.

After building the bipartite graph, we use the centrality scores of the nodes in the graph to measure the attractiveness of commercial agglomerations. Many studies use centralities such as degree, betweenness, closeness, and pagerank (Koschützki et al., 2005) to measure the importance of a location in a city, based on street networks (Crucitti et al., 2006), bus routes (Soh et al., 2010), taxi trajectories (Huang et al., 2015), and human flows (Cheng et al., 2013). But these studies are all about one-mode graphs, and little attention is paid to the commercial agglomeration as the research unit.

We choose to measure the attractiveness with the three centralities of degree, betweenness, and pagerank, and the descriptions are shown in Table 1. The input of the calculation is a bipartite graph, where the commercial agglomerations and source grids are the node sets, the human flow is the edge set, and the output is the centrality score of each commercial agglomeration. The degree and betweenness in the bipartite graph are calculated through the NetworkX package in Python. And the pagerank is implemented through the BiRank package in Python, which is also based on the iterative propagation of the nodes' scores according to the edge weights, which is an extension of pagerank for the bipartite graph (He et al., 2016). After calculating the centrality scores, we use the maximum normalization to uniformly scale the scores to within 0–100 to make them comparable

Table 1. The definitions of three centralities: degree, betweenness, and pagerank.

Centrality	Description	Formula
Degree	The sum of the weights of the edges that connect the node i	$C_D(i) = \sum_j w_{ij}, i \in U, j \in V$
Betweenness	The sum of the fraction of all-pairs shortest paths that pass through the node i	$C_B(i) = \sum_{s \neq i \neq t \in (U \cup V)} \frac{\sigma_{st}(i)}{\sigma_{st}}$, where σ_{st} is the total number of shortest paths from node s to node t, $\sigma_{st}(i)$ is the number of these paths that pass through the node i
PageRank	The likelihood that a surfer randomly walks to the node i based on the weights of the edges. A node should be ranked high if connected to many high-ranked nodes	$C^U_{PR}(i) = \alpha \sum_j \frac{w_{ij}}{\sqrt{d_i}\sqrt{d_j}} C^V_{PR}(j) + \theta$, $C^V_{PR}(j) = \beta \sum_i \frac{w_{ij}}{\sqrt{d_i}\sqrt{d_j}} C^U_{PR}(i) + \gamma$, where d_i, d_j are the weighted degrees of node i and j for normalization, α, β are the damping factors, θ, γ are the initialization variables (He et al., 2016)

$$C_{norm} = 100 \times \frac{C}{C_{max}} \tag{2}$$

Attractiveness prediction

So far, we have summarized multisource environmental factors to reflect its supply capacity, and measured the attractiveness based on human demand. It is necessary to accurately predict the attractiveness based on these environmental factors, which will help with urban planning, for example, provide a theoretical basis for policymakers to estimate the popularity among the residents before designing new commercial agglomerations.

Machine learning models have proven to be powerful in modeling irregular and nonlinear patterns, among which linear regression (LR), support vector machine (SVM), and random forest (RF) are the most widely used classical models (Jordan and Mitchell, 2015). LR assumes a linear relationship between the independent and dependent variables, and generally fits the model parameters by ordinary least squares. Because its model formulas are always simple and the linear assumptions are too strong, it performs well when modeling linear relationships, but not in other situations. SVM enables classification by mapping sample points into high-dimensional space and separating them with maximum margins using hyperplanes (Drucker et al., 1996). Therefore, SVM is generally adopted for classification tasks (especially binary classification), and does not stand out in regression tasks. RF is an ensemble learning model based on decision trees, which avoids overfitting and improves model performance by building multiple trees and combining the results (Breiman, 2001). In many classification and regression tasks, RF usually works better than other single-learner models, such as LR and SVM. Therefore, we choose three classical models of LR, SVM, and RF, with features such as POI, land cover, transportation, and population as inputs, and three centrality scores as outputs to compare their performance. We divide the data into training set and test set in a 7:3 ratio, performing a 10-fold cross-validation on the training set to avoid overfitting. In addition, analyzing the importance of each feature to the model can also provide valuable suggestions for policy development. In this study, the machine learning models are all implemented with default parameters based on the Sklearn package in Python, and it has been experimentally found that it can achieve or approach optimal performance.

Spatial heterogeneity analysis

In addition to achieving an accurate prediction of attractiveness, it is also necessary to study the spatial variation in the impact of these environmental factors on attractiveness due to spatial heterogeneity. It will not only help to guide the configuration of urban infrastructure, but also deepen understanding of the human–environment relationship. Compared to the global ordinary least squares (OLS), the geographically weighted regression (GWR) is a local regression model whose parameters vary with spatial location. So it can effectively model spatial non-stationarity (Brunsdon et al., 1996), and has been used to analyze the driving factors of carbon emission (Zhou et al., 2022), public transportation usage (Chiou et al., 2015), housing price (Kang et al., 2021). Its formula is as follows (Brunsdon et al., 1996)

$$y_i = \beta_0(u_i, v_i) + \sum_j \beta_j(u_i, v_i) x_{ij} + \epsilon_i \tag{3}$$

where i denotes the i -th commercial agglomeration, (u_i, v_i) is the geographical coordinates of its geometric center. y_i represents the dependent variable, and x_{ij} represents the j -th independent

variable. $\beta_0(u_i, v_i)$ is the intercept term, $\beta_j(u_i, v_i)$ is the local regression coefficient, and $\in_i$ is the random error term. As can be seen from the formula, the regression coefficients vary with location, so GWR can effectively model spatial non-stationarity. Although GWR is efficient, it implicitly assumes that each independent variable has the same spatial scale, which may not be true in some cases (Mansour et al., 2021; Zhou et al., 2023). Later, a multiscale geographically weighted regression (MGWR) is proposed to allow the relationship between the independent and dependent variables to vary at different spatial scales (Fotheringham et al., 2017). In MGWR, the coefficients $\beta_{bwj}(u_i, v_i)$ of different independent variables have different bandwidths to describe different spatial scales, as shown in the following formula (Fotheringham et al., 2017)

$$y_i = \beta_{bw0}(u_i, v_i) + \sum_j \beta_{bwj}(u_i, v_i) x_{ij} + \epsilon_i \tag{4}$$

In this study, due to the many factors, we select the explanatory variables by category: (1) For POIs, we choose SHP because it is the most numerous and most relevant to the commercial attributes. (2) For land cover, we choose the green ratio because it represents natural environmental comfort to some extent. (3) For transportation, we choose the main road length to reflect the private transport convenience when traveling across regions, and choose the number of bus stops to indicate the public transport accessibility. 4) For population, we choose population density. The variance inflation factors (VIFs) of the above variables are tested to be far less than 10, so multicollinearity does not pose a problem here (Wheeler and Tiefelsdorf, 2005).

We select the adaptive bisquare kernel function to generate the spatial weights matrix, while the criterion for optimal bandwidth is the corrected Akaike Information Criterion (AICc). The coefficient of determination (R^2) and the Akaike Information Criterion (AIC) are adopted as goodness-of-fit metrics compared to the OLS model. This study implements the MGWR and OLS through the MGWR package from PySAL (https://github.com/pysal/mgwr).

Results

Bigraph centrality

After building the bipartite graph, we calculate the centrality score of each commercial agglomeration based on Table 1, and its boxplot is shown in Figure S3 in the Supplementary Material. We can see that most of the scores of degree and betweenness are at a low level, and the median is less than 20% of the maximum, reflecting the great inequality of the attractiveness level of the commercial agglomeration. In pagerank, this difference is much smaller, and the minimum is more than 40% of the maximum, because in the formula in Table 1, we use the degrees of nodes to normalize, thus suppressing the contribution of high-degree nodes to some extent (He et al., 2016).

We further draw the spatial distribution maps of the three centralities, as shown in Figure 3. First, we can find that the three centralities all roughly show the trend of high values in urban centers and low values in the surrounding areas. The distribution of the high-value regions of degree is relatively scattered, not only in urban areas within the 6th Ring Road, but also in the center of some suburbs, such as Yanqing and Changping in the northwest, Huairou, Miyun, Pinggu, and Shunyi in the northeast. Betweenness presents a similar spatial layout with degree, the general high-value area is more concentrated, but the above suburban centers with high-degree value, in addition to Changping and Shunyi, have a lower betweenness value. Finally, there is pagerank, whose distribution of high-value areas is the most concentrated, almost all located within the 5th Ring Road, and has the most apparent difference in values from the inside to the outside of the city.

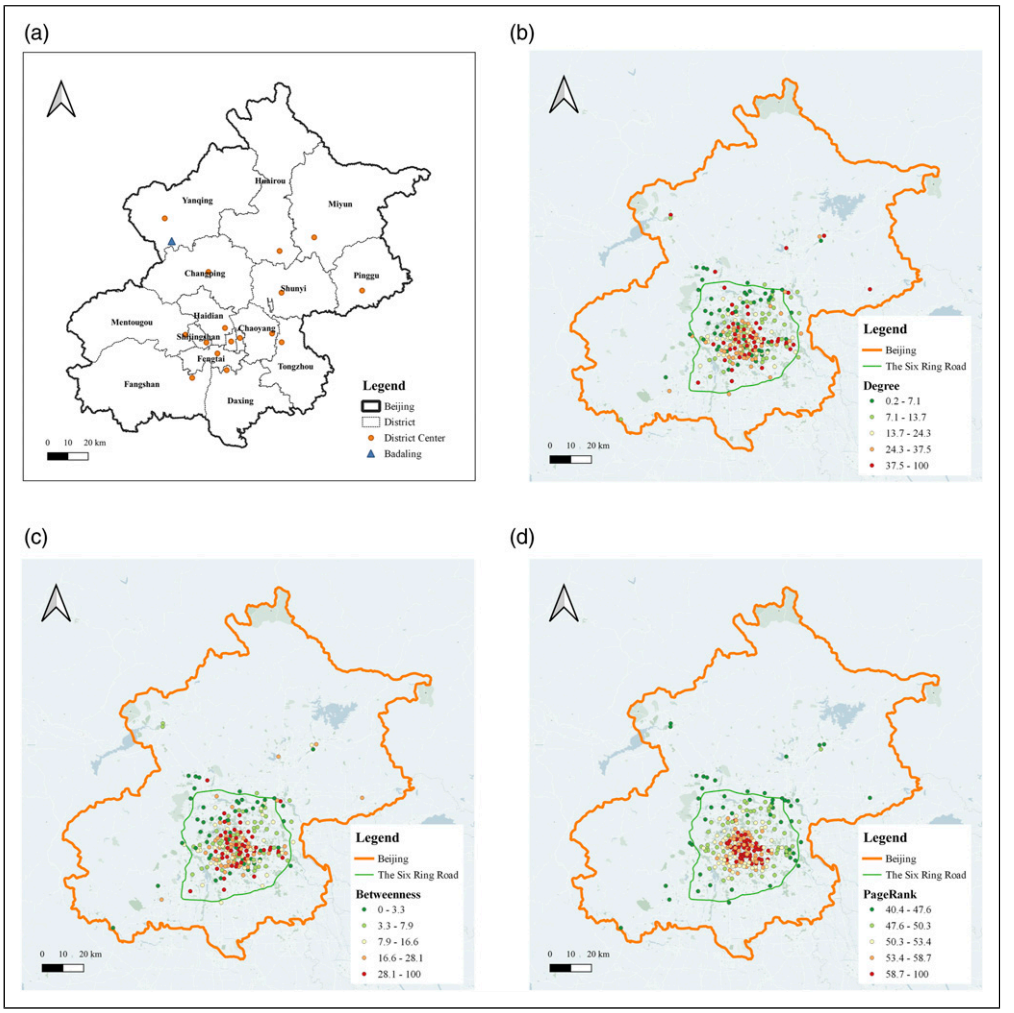

Figure 3. Spatial distribution of three centralities: degree, betweenness, and pagerank.

We select the top 10 commercial agglomerations on the three indicators for further comparison, as shown in Table 2. First of all, the top 10 of degree, we find that they can be roughly divided into two types: one is the center of the surrounding suburbs, such as Changping Center (Changping District), Yuegezhuang (Pinggu District), Baihe East (Miyun District), Liangxiang (Fangshan District), Huangcun (Daxing District); the other is the economically developed and densely populated areas in the city center, such as CBD, Shuangjing, Qingnian Road (the core commercial belt in Chaoyang District), Zhongguancun (high-tech industrial center in Haidian District), Tiantongyuan (a famous commuter town in Changping District that carries the function of residence). The reason for this may be that degree only considers the connection nature of the node to its first-order neighbors (Koschützki et al., 2005), and is easily affected by long-distance flows, as shown in equation (1). At the same time, suburban commercial agglomerations are outstanding in degree centrality because they have less competition and can attract distant customers.

Then there is betweenness, except for Changping Center, the other commercial agglomerations are located in the urban center. They are economically developed and densely populated, such as

Table 2. Top 10 commercial agglomerations based on three centralities: degree, betweenness, and pagerank.

Rank	Degree	Score	Betweenness	Score	PageRank	Score
1	Changping Center	100	CBD	100	CBD	100
2	CBD	88.87	Zhongguancun	79.95	Hongmiao	99.63
3	Yuegezhuang	88.54	Shuangjing	78.89	Tuanjie Lake	96.31
4	Baihe East	83.82	Huilongguan	78.89	Yansha	92.69
5	Shuangjing	81.12	Wangjing	74.96	Chaoyang Park	86.19
6	Liangxiang	81.03	Qingnian road	67.43	Xizhimen	85.83
7	Huangcun	78.90	Jiuxianqiao	65.24	Guangqumen	80.32
8	Zhongguancun	77.01	Changping Center	64.25	Donghuashi	79.81
9	Huilongguan	74.85	Sanyuanqiao	63.86	Balizhuang	77.78
10	Qingnian Road	69.99	Tiantongyuan	62.12	Shilibao	73.39

CBD, Shuangjing, Qingnian Road (the core business districts along Chaoyang Road and Jingtong Expressway within the 4th Ring Road in the west of Chaoyang District); Wangjing, Jiuxianqiao, Sanyuanqiao (emerging new towns along Airport Expressway within the 5th Ring Road in the north of Chaoyang District); Zhongguancun (high-tech industry center in Haidian District); Huilongguan and Tiantongyuan (the two famous commuter towns in Beijing). It can be seen that the attractive commercial agglomerations identified by betweenness are more in line with reality and our cognition, which may be because betweenness focuses on the bridge-like nodes between groups and groups in the network. And those nodes are the hub nodes in the network (Barthelemy, 2004), and the above commercial agglomerations exactly play such a role in the urban system.

Finally, pagerank, we find that in addition to the CBD and Xizhimen, the other commercial agglomerations are generally very small, with an average area of 0.461 km^2 and an average POI count of 781.375, which is lower than the average value of 0.733 km^2 and 1181.755 of all commercial agglomerations. However, CBD has an area of 3.195 km^2 and 10,547 POIs, and Xizhimen has an area of 2.404 km^2 and 3293 POIs. Further, we find that the average distance from those small-size commercial agglomerations to the CBD is only 3.18 km, geographically highly close. Considering that the pagerank algorithm has a "stronger the stronger" Matthew effect (Page et al., 1999; Merton, 1968), nodes connected to high-rank nodes also have a high-rank trend, as shown in Table 1, which makes commercial agglomerations close to the CBD also have high pagerank scores.

Attractiveness prediction results

Taking the features of POIs, land cover, transportation, and population as inputs and the three centrality scores as outputs, we compare the prediction effects of three models: linear regression, support vector machine, and random forest, as shown in Table 3. Based on the results, linear regression outperforms support vector machine a lot, probably due to the relatively strong linear relationship between inputs and outputs. As a classical ensemble learning method, random forest performs much better than the other two models, up to 0.903, 0.846, and 0.760 on R^2 of degree, betweenness, and pagerank, respectively. Therefore, we choose the random forest model for subsequent analysis. Figure 4(a) shows the scatter plots between the true and predicted values of the random forest model. We find a high degree of match between the two values, indicating that random forest can better model the complex nonlinear relationships between environmental factors and attractiveness, resulting in high prediction accuracy.

Table 3. Comparison of the prediction effects of different models on three centralities: degree, betweenness, and pagerank.

Centrality	Model	R^2
Degree	SVM	0.512
	LR	0.750
	RF	0.903
Betweenness	SVM	0.483
	LR	0.672
	RF	0.846
PageRank	SVM	0.461
	LR	0.654
	RF	0.760

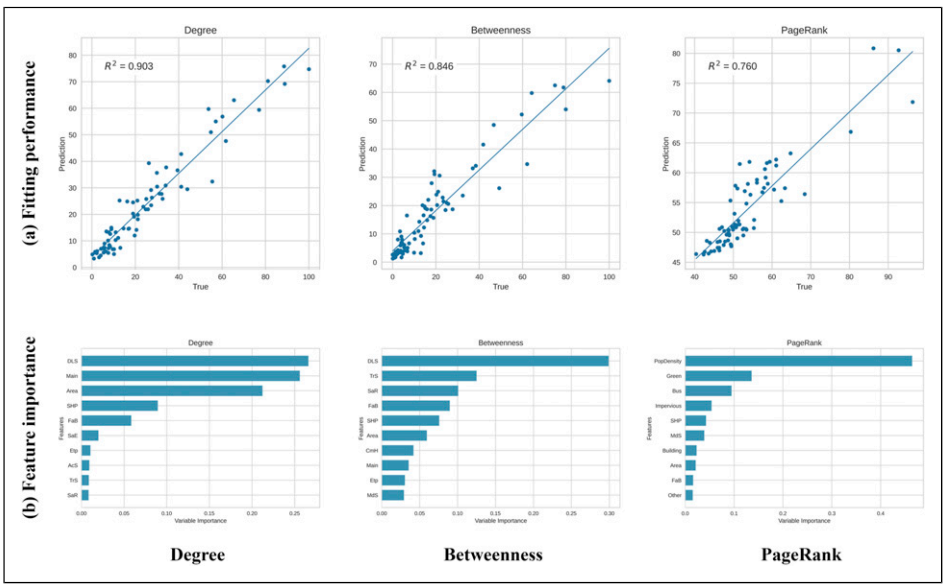

Figure 4. The prediction results of three centralities using random forest: (a) fitting performance, (b) feature importance.

We further calculate the importance of each variable for the random forest model, and Figure 4(b) shows the top 10 variables of three indicators based on their contribution to the model. The first is degree, which can be seen that the number of POIs in categories such as daily life service (DLS), shopping (SHP), food and beverages (FaB), has a significant influence on the model. Besides, the length of the main roads (Main), representing the convenience of private transport, contributes a lot to the model, and the area representing the size of the commercial agglomeration also plays an important role. Then there is betweenness, the variable that has the greatest impact on the model is POIs, of which the contributions of daily life service (DLS), transportation service (TrS), sports and recreation (SaR), food and beverages (FaB) and shopping (SHP) are particularly prominent. It also confirms the phenomenon above that the high-value areas of betweenness are mostly distributed in the city's economically developed and densely populated core areas. Finally, there is pagerank, where we find that population density is the most important feature, and it is consistent with the

previous finding that many of the high-value areas of pagerank occur near the CBD with extremely high population density. Besides, the main influencing factors of pagerank are relatively mixed. In addition to population density, POIs (i.e., the number of shopping (SHP), medical service (MdS), and food and beverages (FaB)), land cover (i.e., the area ratio of green, impervious and buildings), transportation (i.e., the number of bus stops (Bus) and the length of other roads (Other)) are all in the top 10 important variables of pagerank. In general, the number of POIs representing the service diversity is the most dominant factor of commercial agglomeration attractiveness, which reflects its functional attributes (i.e., providing goods or services to residents), transportation and population density are also important variables, but land cover, although to some extent represents environmental comfort, the impact is relatively not noticeable.

Spatial heterogeneity analysis results

We compare MGWR to OLS to explore spatial variations in the relationship between the explanatory variables and attractiveness. Table 4 shows that for the three centralities, MGWR is about 15%–45% better than OLS in terms of R^2 and AIC. Specifically, compared to OLS, MGWR increases by 15.9% on R^2 and decreases by 18.4% on AIC for degree, 28.9% and 28.1% for betweenness, and 36.8% and 44.6% for pagerank, respectively. This suggests a large spatial heterogeneity in the study area, which allows MGWR as a local model to achieve better results in fitting the data.

As mentioned in Section "Bi-graph centrality", the calculated attractiveness of betweenness is more in line with reality and our cognition, so we only choose betweenness for subsequent analyses for simplification, and the explanatory variables are shown in Section "Spatial heterogeneity analysis". Table S3 in the Supplementary Material shows the statistical results of the MGWR parameter estimates, including statistics such as mean, standard deviation, minimum, median, and maximum. First of all, we can find that the mean values of the parameters of all explanatory variables are greater than 0, which indicates that No.Shopping POIs, Green Ratio, Main Road Length, No. Bus Stops and Population Density generally have a positive impact on attractiveness, that is, increasing the values of these factors will bring about an overall increase in attractiveness. Second, the regression coefficients span a wide range in standard deviation, with a minimum of 0.005 (No. Shopping POIs) and a maximum of 0.074 (No. Bus Stops). This large difference tells us that there is a significant spatial variation in the effects of variables in the study area, and also proves the need to use local regression models. Finally, we find that Green Ratio and Pop Density have a bidirectional effect on attractiveness, such as the minimum and maximum of the regression coefficient of Green Ratio are −0.031 and 0.014, while −0.021 and 0.102 for Pop Density, respectively. This suggests that an increase in Green Ratio and Pop Density in some areas will lead to an increase in attractiveness, while in others, it may be the opposite, which is something that global models cannot reveal.

To more intuitively show the spatial heterogeneity of the relationship between the explanatory variables and attractiveness, we generate the spatial distribution maps of their regression

Table 4. Comparison between MGWR and OLS on three centralities: degree, betweenness, and pagerank.

	MGWR		OLS	
Centrality	R^2	AIC	R^2	AIC
Degree	0.749	268.151	0.646	328.626
Betweenness	0.754	264.771	0.585	368.293
PageRank	0.781	222.707	0.571	402.044

coefficients. Each commercial agglomeration is represented by its Thiessen polygon for visualization, as shown in Figure 5. First, in the city center, the local R^2 is a little bit lower than the surrounding areas (about 0.675 vs 0.800). It indicates that the city center is more structurally complex, and more implicit factors need to be considered to achieve better predictions. The remaining subgraphs in Figure 5 are the distribution of the regression coefficients of the selected explanatory variables in the study area. The coefficient of **No. Shopping POIs** roughly shows a trend that the city center is lower than surrounding areas, especially Yanqing and Changping in the northwest show high values. It may be due to the limited level of economic development of the suburbs, and the number of POIs representing the service level to some extent is insufficient. Therefore, for the suburbs, we can increase the number of shopping POIs to increase the attractiveness of the commercial agglomeration by a larger margin. **Green Ratio** shows a bidirectional effect on attractiveness, showing a positive effect in the city center and the southern area, because these areas are mostly urban built-up areas and green space resources are scarce, so increasing the green ratio of the commercial agglomerations in the city center will help improve the attractiveness. In the north, the suburbs show a negative effect, that is, the higher the green ratio, the more remote the location, which brings less attractiveness. For **Main Road Length**, the stronger effect occurs in the northeast, northwest and southwest suburbs, while the effect in the urban center is lower. It may be due to the more developed road network in urban centers and the more inadequate coverage in the suburbs, which results in a "basin-like" phenomenon. For **No. Bus Stops**, it can be seen that the overall spatial variation in the regression coefficients is very obvious. The effects are lower in the northeast and southern suburbs, while high-value areas appear from the city center to the northwest. The reason may be that many suburban residents choose to travel by private car, motorcycle, bicycle, and walking, and rely less on public transportation compared to the residents in the city center. In addition, for the high-value areas in the northwest, Yanqing District has a huge number of tourists brought by "5A" scenic spots such as the Badaling Great Wall (Figure S1 in the Supplementary Material), and Changping District has a large number of commuters living in the two

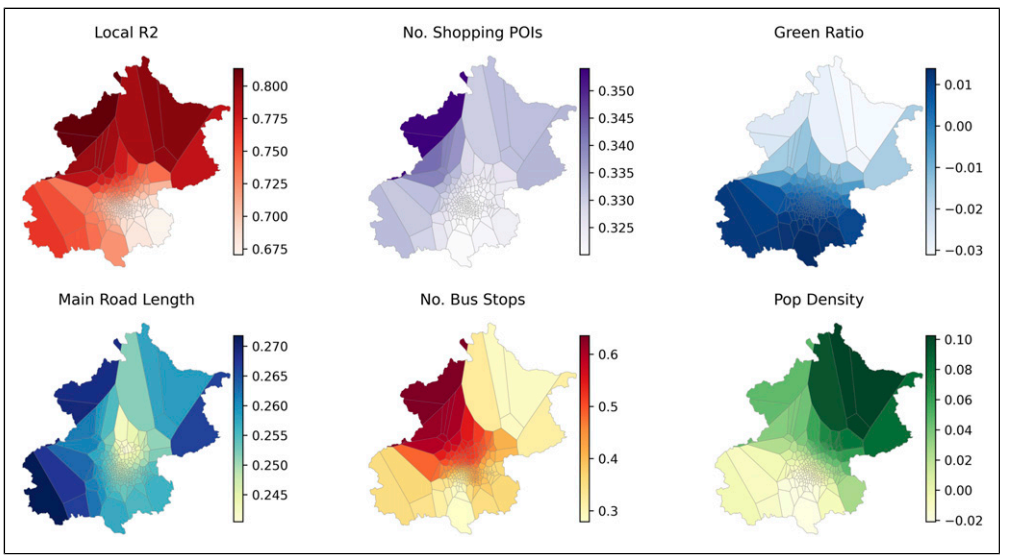

Figure 5. Spatial distribution of the MGWR coefficients.

large communities of Tiantongyuan and Huilongguan, so there exists a more urgent demand for public transportation resources. Finally, there is **Pop Density**, its effect on the attractiveness of commercial agglomerations is also bidirectional. In the city center, population density negatively correlates with attractiveness, which can be attributed to the congestion caused by overpopulation. And in the northern suburbs, increased population density means an increase in labor and customers, thus leading to a rise in the supply and demand for services and products, which positively impacts on the attractiveness of commercial agglomerations.

Discussions

As described in Section "Introduction", the commercial agglomeration of this study is focused only on its physical infrastructure, which is reflected in the open-access datasets used. However, the scope of services is also a crucial element in business analytics, such as the trading area defined by Huff (Huff, 1964). Therefore, the service scope of commercial agglomerations is further explored based on human mobility big data. For each commercial agglomeration, we make a weighted average of the length of all the flows arriving at it by the volume, denoted as the *average radius*, as shown in equation S(1) in the Supplementary Material. This metric reflects the actual scope of services for each commercial agglomeration, and we further have some interesting findings.

Finding a grading effect of service radius

We calculate the average radius of 249 commercial agglomerations and divide them into three groups based on Jenks natural breaks (Jenks, 1967). The average radius shows an obvious grading effect at the breakpoints of 15.3 km and 33.6 km, as shown in Figure 6(a). According to statistics, there are 176 commercial agglomerations in the [0: 15.3 km) interval, accounting for 70.7%, 65 in the [15.3 km: 33.6 km) interval, accounting for 26.1%, and 8 in the [33.6 km) interval, accounting for 3.2%. Clearly, the average service radius of most commercial agglomerations is within 15 km, which takes about 30 min by car in the city.

The above three intervals are denoted from small to large as Level 1, 2, and 3, and their spatial distribution is shown in Figure 6(b). Obviously, the distribution of the levels is regular, showing a concentric circle phenomenon that gradually increases from the inside to the outside: Level 1 is basically all located within the 6th Ring Road and is concentrated in the city center, Level 2 is

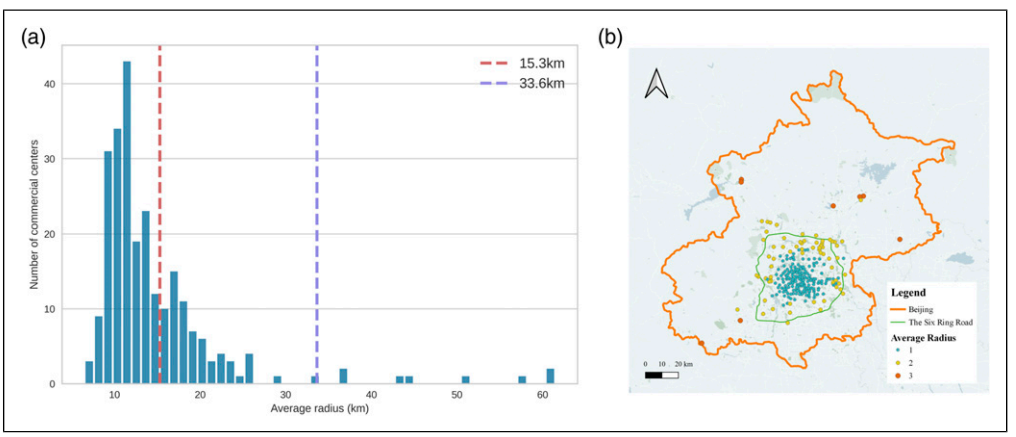

Figure 6. The average radius of each commercial agglomeration: (a) Grading effect. (b) Spatial distribution.

distributed near the 6th Ring Road, and Level 3 is mainly found in the centers of several suburbs. It is reasonable because in urban centers, there are always a large number of commercial agglomerations to choose from, and residents do not have to travel long distances to meet their needs, while in the suburbs, the opposite is true, residents sometimes have to go far to get the goods or services they want. But this finding, while intuitive (as we have explained), is inconsistent with the Central Place Theory (Berry and Garrison, 1958; Getis and Getis, 1966), in which it is pointed out that higher-order central places often have a larger range of services.

Explanation on the grading effect

As we all know, Voronoi diagram is a classical spatial partitioning method based on the spatial layout of a given set of sample points (Aurenhammer, 1991). Each sample point has a corresponding Voronoi cell (or Thiessen polygon), within which each location is closer to that point than to any other. In urban geography, it is often used in catchment analysis such as stores, hospitals, and cell phone base stations (Boots and South, 1997; Guruprasad, 2011; Rezende et al., 2000). Therefore, we hope to explore whether the Voronoi diagram could be used to explain the "strange" phenomenon of the service radius mentioned in Section "Finding a grading effect of service radius", and whether there is an intrinsic correlation between them. Within the boundary of Beijing, the Voronoi diagram is generated by QGIS, where commercial agglomerations are represented by their geometric centers. The corresponding Thiessen polygon for each commercial agglomeration can represent its theoretical service area. For each polygon, its area and perimeter are calculated, and the area-perimeter ratio can be regarded as a representation of the polygon size, which is recorded as the *average width*, as shown in equation S(2) in the Supplementary Material.

After calculating each commercial agglomeration's average radius and width, we make its scatter plot and calculate the Pearson correlation coefficient between the two values, up to 0.796 ($p < 0.001$), as shown in Figure S4 in the Supplementary Material. This indicates a high correlation between the average radius of the commercial agglomerations and the competitive relationship brought about by their spatial layout, which is well in line with the Voronoi theory (Brassel and Reif, 1979). This significant correlation also reminds us that traditional spatial analysis methods still have great application value in urban planning today.

Conclusions

In summary, using the tool of the bipartite graph, this study adopts the centrality scores (degree, betweenness, and pagerank) to measure the attractiveness of commercial agglomerations from the human demand side rather than the environmental supply side. Then, random forest is used to predict attractiveness accurately using environmental variables such as POIs, land cover, transportation, and population as inputs. And the spatial heterogeneity of the effects of these variables on attractiveness is analyzed based on multiscale geographically weighted regression. Each of the three centralities has its characteristics, and the attractiveness measured by betweenness is more in line with reality and our cognition. Based on the summarized environmental features, random forest can accurately predict attractiveness, with R^2 of 0.903, 0.846, and 0.760 for the above centralities. The estimated parameters of MGWR show that there is indeed a large spatial heterogeneity, among which the number of shopping POIs, the length of main roads, and the number of bus stops positively affect attractiveness, while the impact of greening ratio and population density is bidirectional. As for the service scope, we observe a grading effect in the average radius of commercial agglomerations, about 70% is less than 15 km. Moreover, we further find that this grading effect significantly correlates with the Voronoi diagram, which can be explained to some extent.

This study still has deficiencies and needs to be improved in the future. For example, more characteristics need to be considered, such as average price, customer praise, parking convenience, and visual appearance, which will impact attractiveness. Questionnaires or interviews can be implemented to verify the calculated attractiveness's rationality further. In addition, considering the characteristics of different centralities, an effective combination of multiple centrality scores to form a more reasonable measure of attractiveness is also a good direction for future work.

Acknowledgements

We thank the editors and reviewers for their suggestions to improve the quality of this paper.

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

We acknowledge the financial support from the National Natural Science Foundation of China (42271471, 41830645), the International Research Center of Big Data for Sustainable Development Goals (CBAS2022GSP06), the State Key Laboratory of Resources and Environmental Information System, and the Academic Research Projects of Beijing Union University (ZK70202002).

ORCID iDs

Zhou Huang https://orcid.org/0000-0002-1255-1913
Ganmin Yin https://orcid.org/0000-0002-4712-9077

Supplemental Material

Supplemental Material for this article is available online.

References

Arbues P, Banos JF, Mayor M, et al. (2016) Determinants of ground transport modal choice in long-distance trips in spain. *Transportation Research Part A: Policy and Practice* 84: 131–143.

Asratian AS, Denley TM and Haggkvist R (1998) *Bipartite Graphs and their Applications*. Cambridge, UK: Cambridge University Press, volume 131.

Aurenhammer F (1991) Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Computing Surveys (CSUR)* 23(3): 345–405.

Ballantyne P, Singleton A, Dolega L, et al. (2022) A framework for delineating the scale, extent and characteristics of american retail centre agglomerations. *Environment and Planning B: Urban Analytics and City Science* 49(3): 1112–1128.

Bao Y, Huang Z, Li L, et al. (2021) A BiLSTM-CNN model for predicting users' next locations based on geotagged social media. *International Journal of Geographical Information Science* 35(4): 639–660.

Barthelemy M (2004) Betweenness centrality in large complex networks. *The European Physical Journal B - Condensed Matter* 38(2): 163–168.

Batty M (2013) *The New Science of Cities*. Cambridge, Massachusetts, US: MIT press.

Berry BJL and Garrison WL (1958) A note on central place theory and the range of a good. *Economic Geography* 34(4): 304–311.

Bettencourt LMA, Lobo J, Helbing D, et al. (2007) Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences of the United States of America* 104(17): 7301–7306.

Boots B and South R (1997) Modeling retail trade areas using higher-order, multiplicatively weighted voronoi diagrams. *Journal of Retailing* 73(4): 519–536.

Brassel KE and Reif D (2010) A procedure to generate thiessen polygons. *Geographical Analysis* 11(3): 289–303.

Breiman L (2001) Random forests. *Machine Learning* 45(1): 5–32.

Brunsdon C, Fotheringham AS and Charlton ME (2010) Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis* 28(4): 281–298.

Chakraborty A, Krichene H, Inoue H, et al. (2019) Exponential random graph models for the japanese bipartite network of banks and firms. *Journal of Computational Social Science* 2(1): 3–13.

Chen LC, Zhu Y, Papandreou G, et al. (2018a) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), 06 October 2018, Springer, pp. 801–818.

Chen W, Huang H, Dong J, et al. (2018b) Social functional mapping of urban green space using remote sensing and social sensing data. *ISPRS Journal of Photogrammetry and Remote Sensing* 146: 436–452.

Cheng YY, Lee RKW, Lim EP, et al. (2013) Delayflow centrality for identifying critical nodes in transportation networks. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, New York, NY, USA, Association for Computing Machinery, 25 August 2013, pp. 1462–1463.

Chiou YC, Jou RC and Yang CH (2015) Factors affecting public transportation usage rate: Geographically weighted regression. *Transportation Research Part A: Policy and Practice* 78: 161–177.

Cooper LG and Nakanishi M (1989) *Market-share Analysis: Evaluating Competitive Marketing Effectiveness*. New York, US: Springer Science & Business Media, volume 1.

Crucitti P, Latora V and Porta S (2006) Centrality measures in spatial networks of urban streets. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* 73(3): 036125.

de Mello-Sampayo F (2017) Competing-destinations gravity model applied to trade in intermediate goods. *Applied Economics Letters* 24(19): 1378–1384.

Deng J, Dong W, Socher R, et al. (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20-25 June 2009, IEEE, pp. 248–255.

Dong L, Huang Z, Zhang J, et al. (2020) Understanding the mesoscopic scaling patterns within cities. *Scientific Reports* 10(1): 21201–21211.

Drucker H, Burges CJ, Kaufman L, et al. (1996) Support vector regression machines. *Advances in Neural Information Processing Systems* 9: 155–161.

Estiri M, Heidary Dahooie J, Hosseini F, et al. (2021) Proposing a new model for shopping centre attractiveness assessment by a combination of structural equation modelling (sem) and additive ratio assessment (aras). *Current Issues in Tourism* 24(11): 1542–1560.

Eubank S, Guclu H, Kumar VSA, et al. (2004) Modelling disease outbreaks in realistic urban social networks. *Nature* 429(6988): 180–184.

Fotheringham AS (1983) A new set of spatial-interaction models: The theory of competing destinations. *Environment and Planning A* 15(1): 15–36.

Fotheringham AS, Yang W and Kang W (2017) Multiscale geographically weighted regression (mgwr). *Annals of the American Association of Geographers* 107(6): 1247–1265.

Getis A and Getis J (1966) Christaller's central place theory. *Journal of Geography* 65(5): 220–226.

Giglio S, Bertacchini F, Bilotta E, et al. (2019) Using social media to identify tourism attractiveness in six italian cities. *Tourism Management* 72: 306–312.

Gomez-Lievano A, Patterson-Lomba O and Hausmann R (2016) Explaining the prevalence, scaling and variance of urban phenomena. *Nature Human Behaviour* 1(1): 0012–0016.

González-Hernández EM and Orozco-Gómez M (2012) A segmentation study of Mexican consumers based on shopping centre attractiveness. *International Journal of Retail & Distribution Management* 40: 759–777.

Grashuis J, Skevas T and Segovia MS (2020) Grocery shopping preferences during the covid-19 pandemic. *Sustainability* 12(13): 5369.

Guruprasad K (2011) Generalized voronoi partition: A new tool for optimal placement of base stations. In: 2011 Fifth IEEE International Conference on Advanced Telecommunication Systems and Networks (ANTS), 18-21 December 2011, Bangalore, India, IEEE, pp. 1–3.

Hall P (2014) *Cities of Tomorrow: An Intellectual History of Urban Planning and Design since 1880*. Oxford, UK: John Wiley & Sons.

He X, Gao M, Kan MY, et al. (2017) Birank: Towards ranking on bipartite graphs. *IEEE Transactions on Knowledge and Data Engineering* 29(1): 57–71.

Huang X, Zhao Y, Yang J, et al. (2016) Trajgraph: A graph-based visual analytics approach to studying urban network centralities using taxi trajectory data. *IEEE Transactions on Visualization and Computer Graphics* 22(1): 160–169.

Huff DL (1964) Defining and estimating a trading area. *Journal of Marketing* 28(3): 34–38.

Ishikawa Y (1987) An empirical study of the competing destinations model using japanese interaction data. *Environment and Planning A: Economy and Space* 19(10): 1359–1373.

Jenks GF (1967) The data model concept in statistical mapping. *International Yearbook of Cartography* 7: 186–190.

Jordan MI and Mitchell TM (2015) Machine learning: Trends, perspectives, and prospects. *Science* 349(6245): 255–260.

Kang Y, Zhang F, Peng W, et al. (2021) Understanding house price appreciation using multi-source big geo-data and machine learning. *Land Use Policy* 111: 104919.

Koschutzki D, Lehmann KA, Peeters Let al. (2005) *Centrality Indices Network analysis*. New York, US: Springer, 16–61.

Kunc J, Križan F, Bilková K, et al. (2016) Are there differences in the attractiveness of shopping centres? experiences from the czech and slovak republics. *Moravian Geographical Reports* 24(1): 27–41.

Liang Y, Gao S, Cai Y, et al. (2020) Calibrating the dynamic huff model for business analysis using location big data. *Transactions in GIS* 24(3): 681–703.

Liu Y, Liu X, Gao S, et al. (2015) Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers* 105(3): 512–530.

Mansour S, Al Kindi A, Al-Said A, et al. (2021) Sociodemographic determinants of covid-19 incidence rates in oman: Geospatial modelling using multiscale geographically weighted regression (mgwr). *Sustainable Cities and Society* 65: 102627.

Merton RK (1968) The matthew effect in science: The reward and communication systems of science are considered. *Science* 159(3810): 56–63.

Oppewal H, Timmermans HJP and Louviere JJ (1997) Modelling the effects of shopping centre size and store variety on consumer choice behaviour. *Environment and Planning A: Economy and Space* 29(6): 1073–1090.

Page L, Brin S, Motwani R, et al. (1999) *The Pagerank Citation Ranking: Bringing Order to the Web*: Stanford InfoLab. Technical report.

Peng X, Yin G and Huang Z (2022) *Dataset of the Footprints of Commercial Agglomerations in Beijing*, 2019. London, UK: Figshare. DOI: 10.6084/m9.figshare.19656957.v1.

Rezende FAVS, Almeida RM and Nobre FF (2000) Diagramas de Voronoi para a definição de áreas de abrangência de hospitais públicos no Município do Rio de Janeiro. *Cadernos de Saude Publica* 16(2): 467–475.

Romao J, Kourtit K, Neuts B, et al. (2018) The smart city as a common place for tourists and residents: A structural analysis of the determinants of urban attractiveness. *Cities* 78: 67–75.

Seong EY, Lim Y and Choi CG (2022) Why are convenience stores clustered? the reasons behind the clustering of similar shops and the effect of increased competition. *Environment and Planning B: Urban Analytics and City Science* 49(3): 834–846.

Sobolevsky S, Bojic I, Belyi A, et al. (2015) Scaling of city attractiveness for foreign visitors through big data of human economical and social media activity. *IEEE 2015 International Congress on Big Data*. IEEE, pp. 600–607.

Soh H, Lim S, Zhang T, et al. (2010) Weighted complex network analysis of travel routes on the singapore public transportation system. *Physica A: Statistical Mechanics and Its Applications* 389(24): 5852–5863.

Teller C and Reutterer T (2008) The evolving concept of retail attractiveness: what makes retail agglomerations attractive when customers shop at them? *Journal of Retailing and Consumer Services* 15(3): 127–143.

Von Ferber C, Holovatch T, Holovatch Y, et al. (2009) Public transport networks: empirical analysis and modeling. *The European Physical Journal B* 68(2): 261–275.

Wang X, Chen J, Pei T, et al. (2021) I-index for quantifying an urban location's irreplaceability. *Computers, Environment and Urban Systems* 90: 101711.

Wang Y, Huang Z, Yin G, et al. (2022) Applying ollivier-ricci curvature to indicate the mismatch of travel demand and supply in urban transit network. *International Journal of Applied Earth Observation and Geoinformation* 106: 102666.

Wang H, Huang Z, Zhou X, et al. (2022) DouFu: A Double Fusion Joint Learning Method for Driving Trajectory Representation. *Knowledge-Based Systems* 258: 110035.

Wang Y, Zhu D, Yin G, et al. (2020) A unified spatial multigraph analysis for public transport performance. *Scientific Reports* 10(1): 9573.

Wheeler D and Tiefelsdorf M (2005) Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems* 7(2): 161–187.

Wong GKM, Lu Y and Yuan LL (2001) Scattr: An instrument for measuring shopping centre attractiveness. *International Journal of Retail & Distribution Management* 29: 76–86.

Wu SS, Kuang H and Lo SM (2019) Modeling shopping center location choice: Shopper preference–based competitive location model. *Journal of Urban Planning and Development* 145(1): 04018047.

Yang T, Pan H, Hewings G, et al. (2019) Understanding urban sub-centers with heterogeneity in agglomeration economies—where do emerging commercial establishments locate? *Cities* 86: 25–36.

Yin G, Huang Z, Bao Y, et al. (2022) *Convgcn-rf: A hybrid Learning Model for Commuting Flow Prediction Considering Geographical Semantics and Neighborhood Effects*. GeoInformatica 1–21. pages.

Zhou X, Dong Q, Huang Z, et al. (2023) The spatially varying effects of built environment characteristics on the integrated usage of dockless bike-sharing and public transport. *Sustainable Cities and Society* 89: 104348.

Zhou X, Wang H, Huang Z, et al. (2022) Identifying spatiotemporal characteristics and driving factors for road traffic Co2 emissions. *Science of The Total Environment* 834: 155270.

Zhou Huang is Associate Professor of GIScience at the Institute of Remote Sensing and Geographical Information Systems, Peking University. He received the BSc degree in GIS and the Ph.D. degree in Cartography and GIS from Peking University, China, in 2004 and 2009, respectively. Currently, his main research interests include big geo-data, high-performance geocomputation, distributed geographic information processing, spatial data mining, and spatial database. He has published more than 60 academic papers in international journals or conferences. In addition, Zhou Huang serves as Deputy Director of Institute of Remote Sensing and GIS, Peking University, and Deputy Director of Engineering Research Center of Earth Observation and Navigation, Ministry of Education, China. He was also selected for the Youth Talent Innovation Plan in Remote Sensing Science and Technology in 2015, funded by the Ministry of Science and Technology of China.

Ganmin Yin received the B.S. degree from Peking University in 2020. He is currently pursuing the Ph.D. degree in GIScience with the Institute of Remote Sensing and Geographical Information Systems, Peking University. His research interests include human mobility, transportation, urban data mining, social sensing and GeoAI.

Xia Peng received the B.S. degree in geographical information system from the China University of Geosciences, Wuhan, China, in 2004, the M.S. degree in cartography and geographical information system from Peking University, Beijing, China, in 2007, and the Ph.D. degree in urban and rural planning from Tsinghua University, Beijing, in 2013. She is currently an Associate Professor with the Tourism College, Beijing Union University, Beijing. Her major research interests include spatio-temporal data mining, GIS, and tourism decision support systems.

Xiao Zhou received the Ph.D. degree from the School of Geography and Ocean Science at Nanjing University in 2021. He is currently a post-doctor with the Institute of Remote Sensing and Geographical Information Systems, Peking University. His research interests include environmental analysis and sustainable development.

Quanhua Dong received the Ph.D. degree from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing at Wuhan University in 2021. She is currently a post-doctor with the Institute of Remote Sensing and Geographical Information Systems, Peking University. Her research interests include three-dimensional spatial analysis and urban system.