IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 14, 2020, accepted December 2, 2020, date of publication December 9, 2020,
date of current version December 21, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3043317*

# Spatial Accuracy Evaluation for Mobile Phone Location Data With Consideration of Geographical Context

**XIAOQING SONG**[1,2,3,4], **YI LONG**[1,2,3], **LING ZHANG**[1,2,3], **DAVID G. ROSSITER**[3,5], **FENGYUAN LIU**[1,2,3], **AND WEI JIANG**[1,2,3,4]

[1] Key Laboratory of Virtual Geographic Environment, Ministry of Education, Nanjing Normal University, Nanjing 210023, China
[2] State Key Laboratory Cultivation Base of Geographical Environment Evolution, Nanjing 210023, China
[3] Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China
[4] School of Geography and Tourism, Anhui Normal University, Wuhu 241000, China
[5] Section of Soil and Crop Sciences, School of Integrated Plant Sciences, College of Agriculture and Life Sciences, Cornell University, Ithaca, NY 14853, USA

Corresponding author: Yi Long (yilong.njnu@hotmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 41571382, in part by the National Key Research and Development Program of China under Grant 2017YFB0503500, and in part by the State Key Program of National Natural Science Foundation of China under Grant 41930104.

**ABSTRACT** In recent years, mobile phone location (MPL) data have been widely used to determine the spatial trajectories of users. Although this massive amount of MPL data can provide insight into human movement, definite conclusions cannot be drawn because of positioning bias: the locations of MPL data are usually not the phone users' actual locations. In recent years, the spatial accuracy of MPL data has been increasingly evaluated. Such efforts have led to many insights regarding the quality and applicability of MPL data. Despite these achievements, to the best of our knowledge, no studies have quantitatively assessed the spatial accuracy of MPL data by considering geographical influencing factors. In this study, we built a linear evaluation model based on geographical weighted regression (GWR) and a nonlinear evaluation model based on a random forest (RF) to quantify the relationship between geographical factors and the positioning bias of MPL data. Nanjing city in China is used as the test case. The results show that both the GWR model and RF model have good stability. However, the RF model's overall prediction performance is much better than that of the GWR model. The RF model can estimate the spatial accuracy of the MPL data within narrow margins of error. The importance ranking of geographical variables shows that the population density, elevation and building density are the three most important factors, while the normalised difference water index (NDWI) and distance to the nearest cell tower (DNCT) are the least important variables. The RF model constructed in this study can be used to evaluate the spatial accuracy of MPL data and simulate the spatial distribution of the positioning bias of the MPL data covering the study area.

**INDEX TERMS** Mobile phone location data, positioning bias, geographical factors, spatial accuracy evaluation.

## I. INTRODUCTION

Recently, mobile phone location (MPL) data have become a crucial data source in various research areas, such as public health [1], human mobility patterns [2], and urban and transportation planning [3], [4]. Notably, MPL data can be used to determine the spatial trajectories of users and play a crucial role in revealing the dynamic pulse of a city. However, such data cannot provide sufficiently accurate space-time information to arrive at definite conclusions about human movement [5], [6]. A key characteristic of MPL data is that the locations are documented at the level of cell towers. These locations, which are usually represented as geographical coordinates of the cell towers, do not necessarily reflect the actual locations of the phone users [7], [8]. Obviously, the spatial accuracy of the MPL data directly affects the validity of human mobility research.

Evaluating spatial accuracy has always been an essential task in mobile positioning, and it is also the basis and premise of data applications. Many studies have focused on the spatial

The associate editor coordinating the review of this manuscript and approving it for publication was Feng Xia.

accuracy of MPL data [9], [10] and its influential factors [11], [12]. These studies have enhanced our understanding of the quality of MPL data. Existing research on the factors influencing the accuracy has two perspectives: a communication perspective and a geographical perspective. From the communication perspective, the equipment conditions of a cell tower, such as the carrier frequency and antenna height [13], influence the spatial accuracy of mobile positioning. From the geographical perspective, the complex channel environment (i.e., the geographical environment) is the main factor that influences the spatial accuracy of MPL data. Some researchers have qualitatively stated that the spatial accuracy of MPL data is affected by certain geographical factors, such as relief, buildings, vegetation, etc. [14]–[16]. However, to the best of our knowledge, no studies have quantitatively assessed the spatial accuracy of MPL data by considering the geographical influencing factors.

The objectives of this research are (1) to identify key geographical factors that significantly affect the spatial accuracy of MPL data, (2) to construct quantitative evaluation models of positioning bias based on these key geographical factors, and (3) to map the predicted spatial distribution of the positioning bias of MPL data over an area. In this paper, we describe our methods for evaluating the positioning bias of MPL data based on MPL data, GPS data and geographical data, and we discuss our validation case study experiments. We recruited forty volunteer college students to collect their GPS data and MPL data over the same time period. After filtering the GPS data, we calculated the positioning bias of the MPL data. The predictive variables were collected and pre-processed, and after correlation and multicollinearity tests, we identified seven predictive variables exhibiting significant influences. A linear evaluation model based on geographical weighted regression (GWR) and a nonlinear evaluation model based on random forest (RF) were used to construct quantitative relationships between the seven predictive variables and the positioning bias. We simulated the spatial distribution of the positioning bias of the MPL data for a study area (nine administrative districts of Nanjing, China) based on the two models. Our case studies showed that both models exhibited good stability, and the overall prediction performance of the RF model was much better than that of the GWR model. These results demonstrated that the RF model can be useful for spatial accuracy evaluation of MPL data.

The remainder of this article is organised as follows. The next section discusses the existing research related to this study. Section III describes the study area and data. The methods used to evaluate the positioning bias of MPL data are described in Section IV. Detailed results of the experiments conducted in this work are discussed in Section V. Section VI provides a summary.

## II. RELATED WORK

The goal of this work is to quantitatively assess the spatial accuracy of MPL data from a geographical perspective. The existing research related to the spatial accuracy of MPL data can be divided into three groups from the phenomenon to the mechanism: (1) the uncertainty of spatiotemporal analysis, (2) the accuracy of MPL data and (3) factors influencing the spatial accuracy. In this section, we review the related research on these three groups.

The first group of studies related to the spatial accuracy of MPL data focuses on the uncertainty of spatiotemporal analysis. Uncertainty is related to several concepts, such as accuracy, precision, consistency, etc. [17]. Many critical concerns have been raised about how uncertainties could influence results [8], [18] and the risk level in decision-making processes [19]. The uncertainties embedded in MPL data with respect to their spatiotemporal granularity should be examined. From the spatial perspective, the positional accuracy of phone records relies on the size of the tower coverage area [20]. Zhao *et al.* [21] compared three frequently used mobility indicators derived from call detail records (CDRs) with a dataset that contains both CDRs and actively generated logs. They found that CDRs tend to underestimate the total travel distance and the movement entropy while providing a reasonable estimate of the gyration radius. Yin *et al.* [22] took a mobile signalling dataset of 24-hour user tracking as a benchmark to evaluate the bias in the population distribution derived from CDRs. They found that the median relative errors were 25∼30% during hours when humans were active. From the perspective of temporal granularity, the interval between two phone communication activities is usually longer than two hours, especially with CDR data, which leads to a great deal of uncertainty in human mobility analysis. Zhao *et al.* [23] evaluated the effect of temporal sampling intervals (TSIs) on typical human mobility indicators obtained from MPL data. They showed that coarser TSIs tend to underestimate the four selected indicators (movement entropy, radius of gyration, eccentricity, and daily travel frequency) to different degrees. The temporal granularity of CDR data can be improved with data completion. Some modelling techniques have been proposed to predict missing locations in subscribers' trajectories. Hoteit *et al.* [24], [25] filled the spatiotemporal gaps in CDRs and examined the quality of filled data in the presence of ground-truth GPS data.

The second group of research explores the accuracy of MPL data. The accuracy evaluation of MPL data is the foundation of data applications. Ahas *et al.* [26] estimated the positioning bias using the differences between the GSM-measured location and GPS-measured location, i.e., the actual location of a mobile phone with high accuracy. They found that 52% of positioning points were accurate within 400 metres in urban areas and 50% were accurate within 2600 metres in rural areas. In addition, Ahas *et al.* [27] evaluated passive mobile positioning data for tourism surveys and found that the data can be collected for larger spatial units. They also found that the spatial and temporal precision of MPL data is higher than traditional tourism statistics. Isaacman *et al.* [28] and Becker *et al.* [10] quantified the differences between CDRs and the actual locations logged by volunteers. Their results showed that the median differences between daily ranges computed from CDRs and those derived from the ground-truth logs are less than 1.5 miles. Research on the spatial distribution characteristics of the positioning bias of MPL data has generally involved simply dividing a study area into urban and rural areas. Pospíšilová and Novák [9] calculated the accuracy of positioning based on one of the primary Czech mobile network operator's data and their distribution. They found that the average accuracy in Prague is 1 km. The precision within Czechia can be described as relatively high in regional centres (1.3 km), moderately high (3.5 km) in cities with over 10,000 inhabitants, and low in the countryside (6 km). Trevisani *et al.* [29] found that the average accuracy is approximately 800 metres for U.S. data and decreases to 480 metres for Italian data. In the Italian case, the average accuracy increases from urban (0.48 km) to suburban (0.75 km) and then to highway (1 km) scenarios. In the U.S. case, the average accuracy increases from suburban (0.49 km) to urban (0.79 km) and then to highway (2.91 km) scenarios.

The third group of studies identifies the sources and factors influencing positioning biases. This research is mainly divided into two aspects: the communication perspective and the geographical perspective. In the communication perspective, many studies have shown that the equipment conditions of a cell tower influence the spatial accuracy of mobile positioning. The Okumura-Hata model [13] indicated that the coverage of a cell tower is related to the carrier frequency and antenna height. Zhong and Xiao [11] found that the coverage of a cell tower can be controlled by adjusting the downdip antenna angle. From the geographical perspective, many studies have indicated that the complex channel environment (i.e., the geographical environment) is the main factor that influences the spatial accuracy of MPL data. Many works have shown that the error sources in wireless location systems include non-line-of-sight (NLOS) propagation and other factors [12]. The complexity of the channel environment is the ultimate cause of NLOS propagation. Many models have been proposed to describe the path loss of wireless signals in the channel environment. For example, Edwards and Durkin [30] proposed the Durkin model to predict large-scale wireless signal path loss. Okumura and Ohronofi [31] put forward the Okumura model to predict wireless signals, and it has been the most widely used model for predicting urban signals. In addition, the geographical environment affects the site selection of cell towers [32], thus affecting the distribution density of cell towers and indirectly affecting the spatial accuracy of CMPL data. Some researchers have qualitatively stated that the spatial accuracy of MPL data is affected by certain geographical factors, such as relief, buildings and vegetation, etc. [14]–[16]. Although these studies qualitatively stated that some geographical factors affect the spatial accuracy of MPL data, quantitative evaluations of the positioning bias of MPL data from the geographical perspective remain scarce. Therefore, for the first time, we build quantitative relationship models between the positioning bias of MPL data and certain geographical factors to evaluate the spatial accuracy of MPL data.

## III. STUDY AREA AND DATA

### A. STUDY AREA

The study area in this research covers nine administrative districts of Nanjing, China (Figure 1). This city covers an area of 6,587 km$^2$ and has eleven administrative districts. Among them, Xuanwu, Qinhuai, Gulou, Jianye, and Yuhuatai contain the most concentrated urban areas. The other six administrative districts are in rural areas, although urbanisation occurs in parts of these districts. The data collection of this study is mainly carried out in the nine administrative districts shown in Figure 1; thus, only these administrative districts are used to represent the study area.

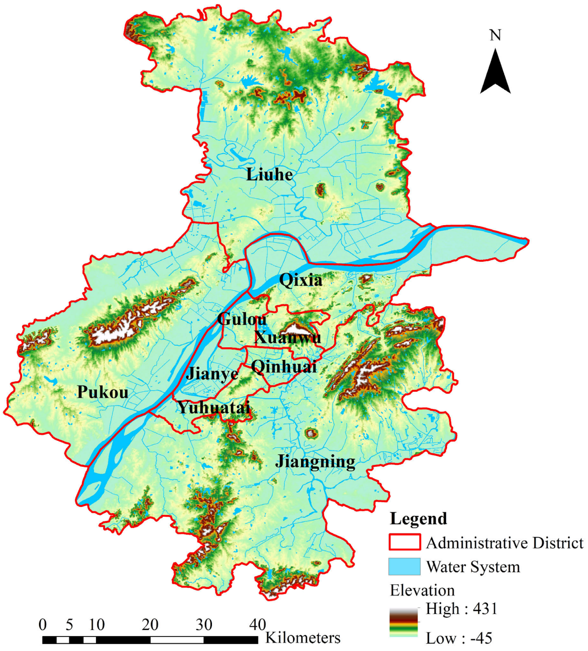


**FIGURE 1. The administrative districts and specific geographical environment details of the study area.**

As one of eastern China's financial centres, Nanjing is a megalopolis with a population of more than 8.5 million people. Among them, the urban population accounts for 83.2%. Nanjing has a complex and diverse geographical environment [33]. The terrain is mainly composed of hillocks (53% of the area); plains, depressions, rivers, and lakes (39.2%); low mountains (3.5%); and hills (4.3%). Nanjing is one of the eight backbone network nodes of China Telecom. Nanjing's unique socioeconomic and geographical status makes it an ideal area for studying the quantitative relationship between. positioning biases and the positioning environment.

### B. DATA COLLECTION AND PRE-PROCESSING

Two datasets were used in this study. The first dataset is the spatiotemporal location dataset that includes two types: MPL data and GPS data. We used this dataset to evaluate the spatial accuracy of MPL data quantitatively. The second dataset contains the predictive variables that affect the positioning accuracy of MPL data.

### 1) SPATIOTEMPORAL LOCATION DATASET

As shown in Figure 2, MPL data are a byproduct of cell tower positioning. This positioning technology obtains the spatiotemporal location information of subscribers (cell phones) through the network of mobile operators. Accordingly, MPL data are stored in the database server of mobile operators. The MPL dataset used in our study is from a major operator in Nanjing city. It contains both passive and active location data. Passive MPL data are those where data are generated only when particular types of human activities occur, e.g., placing or receiving a call, sending or receiving a short message, and connecting to the Internet. The typical passive MPL data are CDRs. Active MPL data are collected using mobile tracking, which does not depend on human use of the phone. They are generated by communication systems that periodically actively update or regularly update the location of mobile phones. Periodic updates are triggered by tower pinging if a subscriber has been 'silent' (i.e., no human use events detected) for a certain time period. Regular updates are triggered by moving from cell tower's service area to that of another tower [21]. Each record in this MPL dataset comprises an anonymous user ID, recording date, recording time, and the coordinates (longitude/latitude in the WGS84 coordinate reference system) of the cell tower that handled the mobile transaction. GPS is a high-precision radio navigation positioning system based on man-made satellites. Modern smartphones usually contain a GPS receiver, making it very convenient for us to collect and store GPS data. In this study, the cell phones used to collect MPL data were also used as the GPS receivers and GPS data memory (Figure 2). Each record in the GPS dataset includes the device ID, recording date, recording time, positioning accuracy $\alpha$, and coordinates (longitude/latitude in the WGS84 coordinate reference system) of the cell phone. Compared with MPL data, GPS data usually have higher spatial accuracy and smaller time intervals between adjacent anchor points. In our study, the location obtained using GPS is taken as the actual location of a mobile phone.

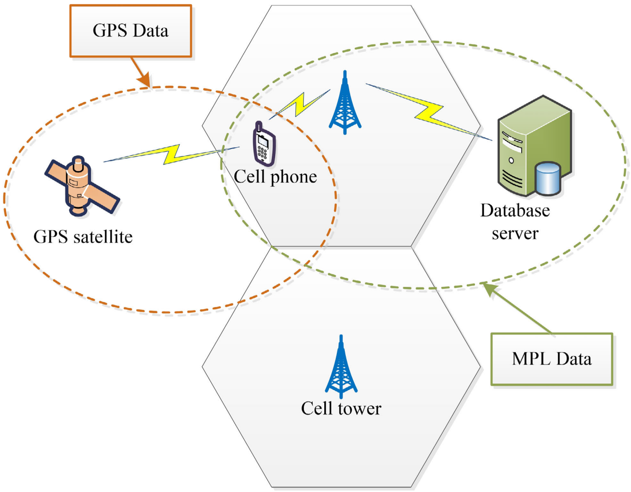


**FIGURE 2. Collection and storage of MPL data and GPS data.**

In the spatiotemporal location data collection phase, we recruited forty volunteer college students to collect the data. We provided each volunteer with a mobile phone, and each phone was equipped with a mobile card from the aforementioned Nanjing operator. We also used the same brand and model of cell phones to collect data and reduce potential biases. The volunteers collected GPS data with these mobile phone GPSs and permitted us to inspect their MPL data during the data collection period. Specifically, a volunteer's MPL data and GPS data were collected by the same phone in the same period of time (Figure 2). The GPS dataset was collected by these mobile phone GPSs at 5-second intervals. Also, to ensure that the samples were representative of the target population (i.e., all of the geographical environments in Nanjing), we constructed twenty itineraries so that volunteers could plan routes covering different geographical contexts, including various elevations, tall building densities, park areas, lakes, etc.

Four steps were applied in the data pre-processing phase. We first converted the latitude and longitude coordinates of the MPL dataset and the GPS dataset into projected coordinates (Beijing 1954 3 Degree GK CM 117E projected coordinate system). Then, we filtered the GPS data based on the positioning accuracy $\alpha$ of the GPS data. We kept only those GPS records with a positioning accuracy of fewer than 3 metres. Furthermore, each anchor point in the MPL dataset was matched to the anchor point in the GPS dataset based on the same timestamps of the two datasets. The positioning bias $y$ of each pair of points was calculated using Equation (1):

$$y = \sqrt{\left(H_{gps} - H_{mpl}\right)^2 + \left(V_{gps} - V_{mpl}\right)^2} \tag{1}$$

where $y$ is the positioning bias of the MPL data; $H_{gps}$ and $V_{gps}$ represent the horizontal position and vertical position, respectively, in the projected coordinates of a GPS anchor point; and $H_{mpl}$ and $V_{mpl}$ represent the horizontal position and vertical position, respectively, in the projected coordinates of an MPL anchor point. These point pairs constituted a set of samples $D$ in this study. At this stage, each sample in $D$ included the GPS anchor point coordinates, the positioning accuracy $\alpha$ of the GPS anchor point, the MPL anchor point coordinates, positioning time, and positioning bias $y$. Finally, we filtered the samples $D$ based on the value of $\alpha$ divided by $y$. Since the GPS anchor point was taken as the actual location of a volunteer in this study, the positioning accuracy $\alpha$ of the GPS anchor point will also affect the positioning bias $y$. We only kept samples in $D$ with a ratio ($\alpha/y$) of less than 0. 1 [17] to reduce this effect.

After the above data pre-processing phase, we obtained a dataset $D$ of 6,112 samples for modelling. The mean positioning bias $y$ of the 6,112 samples was 2,130.80 metres (minimum 39.73 m, maximum 7,264.98 m, and standard deviation $\sigma = 1,711.61$ metres). Figure 3 shows the comparative statistics.

**TABLE 1. Predictive variables, data sources, and pre-processing.**

| Category | Variable | Variable description | Data source | Pre-processing |
|---|---|---|---|---|
| Topography | Elevation | Vertical distance above or below sea level at a location on the ground. | 90 m Digital Elevation Database from NASA CGIAR-CSI Geoportal, SRTM | Resampling to a 200 m grid |
| | Slope | Degree of the steepness of surface units, expressed as the ratio of the vertical height to the horizontal distance of the surface. | Derived from the elevation data | Applying the 'slope' function in ArcGIS to generate a 200 m grid |
| | Aspect | Azimuth of the projection of the normal of the slope surface on the horizontal plane. | Derived from the elevation data | Applying the 'aspect' function in ArcGIS to generate a 200 m grid |
| Land use | Building density | Density of buildings within a spatial unit | Derived from building POI data (Crawled from Gaode Map) | Applying the 'point density' function in ArcGIS to generate a 200 m grid |
| Land cover | DNV | Distance to the nearest vegetation. | Derived from vegetation data (Crawled from Gaode Map) | Applying the 'near' function in ArcGIS to generate a 200 m grid |
| | DNWB | Distance to the nearest water body. | Derived from water data (Crawled from Gaode Map) | Applying the 'near' function in ArcGIS to generate a 200 m grid |
| | NDWI | Normalised Difference Water Index, an index reflecting water information in remote sensing images | Derived from remote sensing image data (Landsat8 30 m satellite digital products) | Applying the 'band math' function in ENVI to generate a 30 m grid of NDWI, then resampling to a 200 m grid. |
| | NDVI | Normalised Difference Vegetation Index, an index reflecting vegetation coverage | Derived from remote sensing image data (Landsat8 30 m satellite digital products) | Applying the 'band math' function in ENVI to generate a 30 m grid of NDVI, then resampling to a 200 m grid. |
| Population | Population density | The population density within a spatial unit | LandScan 2020 Global Population Database | Clipping the grids to the study area boundary and resampling to a 200 m grid |
| Cell tower | Cell tower density | The density of cell towers within a spatial unit | Mobile operator | Applying the 'point density' function in ArcGIS to generate a 200 m grid |
| | DNCT | Distance to the nearest cell tower. | Mobile operator | Applying the 'near' function in ArcGIS to generate a 200 m grid |

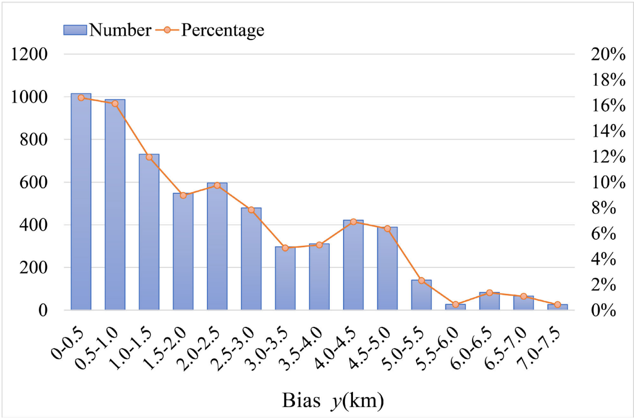


**FIGURE 3. Statistical distribution of the positioning bias.**

### 2) PREDICTIVE VARIABLES DATASET

The geographical coordinates of the cell tower serving the mobile phone at a given time are used as an approximation of the subscriber position [7]. Therefore, the spatial accuracy of MPL data is closely related to the coverage (i.e., service area) of the cell towers. As described in Section III, the complex channel environment (i.e., the geographical environment) is the main factor influencing the spatial accuracy of MPL data. On the one hand, the geographical environment affects the site selection of cell towers, thus affecting the distribution density of cell towers and indirectly affecting the spatial accuracy of MPL data. On the other hand, the geographical environment affects the propagation loss of wireless signals and indirectly affects the coverage of cell towers and the positioning accuracy of MPL data.

Gridded maps of spatial covariates were collated to describe topography, land use, land cover, population and cell tower (Table 1). Topography strongly impact the site selection of cell towers [34], which indirectly affects the spatial accuracy of MPL data. We used elevation, slope and aspect to describe the topography of the volunteer's geographical environment. Buildings and vegetation can cause NLOS propagation of wireless signals, resulting in mobile positioning biases [35], [36]. Building density and distance to the nearest vegetation (DNV) were used to depict the distribution of buildings around the volunteers. Water bodies usually refract or reflect wireless signals, causing a loss in wireless signal propagation [37]. Distance to the nearest water body (DNWB) was used to quantify the impact of water bodies on the positioning accuracy. The impact of vegetation and water bodies was also described using remotely sensed imagery. The normalised difference vegetation index (NDVI) [38] and the modified normalised difference water index (NDWI) [39] were calculated from a Landsat8 remote sensing image from July 2020 with a spatial resolution of 30 m. In some cases, a cell phone is not connected to the nearest cell tower due to load balance issues [40]. Therefore, the distribution of the population on the ground also affects the accuracy of MPL data. The population density is an important index of the population distribution. As mentioned above, the spatial accuracy of MPL data is closely related to the coverage of the cell towers. The cell tower density and the distance to the nearest cell tower (DNCT) were used to describe the distribution of the cell towers around the volunteers. In addition, considering the average spatial distance between cell towers in the study area is 197.3 m, gridded maps of the predictive variables are prepared at 200 m by 200 m resolution to describe these factors (Table 1).

All predictive variables were assigned to GPS anchor points, the actual locations of the volunteers from the GPS records. After data preparation, as stated above, we obtained a set of samples $D$ for modelling. Each sample in $D$ includes a positioning bias $y$ (calculated by Equation (1)) and a $p$-dimensional vector of the predictive variables $X$, $X = \{x_1,\ldots,x_p\}$. The form of the dataset $D$ is $D = \{(X_1,y_1),\ldots,(X_n,y_n)\}$, where $n$ is 6,112

## IV. METHODOLOGIES

In this section, we construct a linear and a nonlinear evaluation model for the spatial accuracy of MPL data. The GWR approach is mainly used to build the linear evaluation model. To further improve the prediction accuracy of the positioning bias of MPL data, a nonlinear evaluation model based on an RF is constructed. The analysis results are shown in the next section to demonstrate the impact of these methods and the choice of the key parameters on the prediction accuracy.

### A. LINEAR EVALUATION MODEL

Regression analysis is often used to quantify some geographical reasoning. The traditional linear regression model is shown as:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon_i \quad (i = 1, 2 \ldots n) \tag{2}$$

where $p$ is the number of predictive variables; $n$ is the number of observations (samples); $y_i$ is the dependent variable of the $i$th sample point; $x_j$ is the $j$th independent variable of the $i$th sample point; $\varepsilon_i$ is the residual of the $i$th sample; $\beta_0$ is the intercept, and $\beta_j$ is the regression coefficient of the $j$th independent variable.

The traditional linear regression model is based on the ordinary least square (OLS) method to estimate the parameters. OLS obtains model parameters (the intercepts and regression coefficients) by reducing the difference between the real and predicted values of dependent variables. Specifically, when the sum of squares of residuals is the smallest, the model parameters are optimal. The traditional linear regression model implements the average or global estimation of the parameters, ignoring spatial heterogeneity and nonstationarity. In fact, the influence of a predictive variable on the positioning bias $y$ may be different at different spatial locations. Therefore, we built a linear evaluation model for the spatial accuracy of MPL data based on the GWR approach. The GWR is an extension of the traditional linear regression model, which inserts the geographical position of the sample into the regression parameters [41]. A general version of the model can be expressed as:

$$y_i = \beta_0(u_i, v_i) + \sum_{j=1}^{p} \beta_j(u_i, v_i) x_j + \varepsilon_i \quad (i = 1, 2 \ldots n) \tag{3}$$

where $(u_i,v_i)$ represents the spatial coordinates of the $i$th sample point; $\beta_0(u_i,v_i)$ is the intercept, and $\beta_j(u_i,v_i)$ is the regression coefficient of the $j$th independent variable of the $i$th sample point. In essence, GWR measures the inherent relationships around each sample $i$, where weighted least squares estimate each set of regression coefficients. The matrix expression for this estimation is

$$\beta'(u_i, v_i) = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i) y \tag{4}$$

where $X$ is independent variable matrix; $y$ is the dependent variable vector; $\beta'(u_i,v_i)$ is the estimate of the location-specific regression coefficients, and $W(u_i, v_i)$ is a diagonal matrix denoting the geographical weighting of each sample point $i$. The kernel density function determines the weight assigned to neighbouring units. Several kernel functions can be used for the weighting scheme. The Gaussian kernel is specified in this study, and its usual continuous form can be defined as

$$W(u_i, v_i) = \exp[-\frac{1}{2}(\frac{d_{ij}}{b})^2] \tag{5}$$

where $d_{ij}$ is the distance between sample point $i$ and observation point $j$ and $b$ is the kernel bandwidth. The bandwidth is a key control parameter of the GWR model. The regression results are sensitive to the choice of neighbours, and the selected bandwidth may significantly impact the coefficient estimation [42]. Cross-validation ($CV$) is generally used to determine the optimal number of nearest neighbours. It takes the form of

$$CV = \sum_{i=1}^{n} {}^{[y_i - y'_{\neq i}(\beta)]^2} \tag{6}$$

where $y'_{\neq i}(\beta)$ is the predicted value of $y_i$ with the observations for sample point $i$ omitted from the calibration process. Here, we are trying to find $b$ that minimises $CV$. This in effect minimizes the sum of squared errors at all sample points and arrives at an optimal bandwidth. In this study, GWR was applied to build the linear evaluation model and identify key predictive variables that significantly affect the spatial accuracy of MPL data.

### B. NONLINEAR EVALUATION MODEL

To further improve the prediction accuracy of the positioning bias of MPL data, a machine learning model, i.e., RF, is used to build a nonlinear evaluation model. An RF is an ensemble of many decision trees with some randomness in selecting predictors at each split and in the training cases for each tree [43]. This model usually has a high prediction accuracy, high efficiency, and low probability of overfitting to training data [44]. As shown in Figure 4, the model is an ensemble of $K$ regression trees $\{T_1,\ldots,T_K\}$. In the training phase, random sampling with replacement is executed, generating $K$ datasets $\{D_1,\ldots,D_K\}$. The sample number of each dataset is equal to the number of samples in $D$. As described in Section III-B above, we obtain a dataset $D$ of 6,112 samples for modelling after data pre-processing. In our study, the number

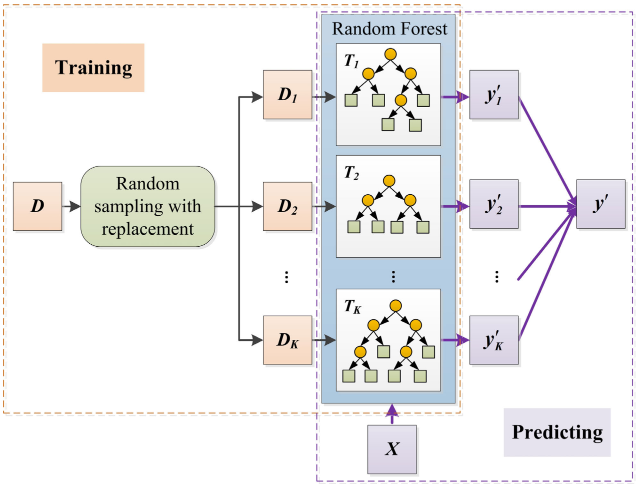


**FIGURE 4.** A nonlinear evaluation model based on a random forest.

of samples of each dataset in $\{D_1,\ldots,D_K\}$ is 6,112. Each dataset is used to train and generate a regression tree. Further, $K$ regression trees are assembled into a random forest, i.e., the nonlinear evaluation model. In the predicting phase, a $p$-dimensional vector $X$ of predictive variables of a sample is put into the model. This model produces $K$ outputs $\{y'_1 = T_1(X),\ldots,y'_K = T_K(X)\}$, where $y'_k$, $k = 1,\ldots, K$, is the prediction for a sample by the $k$th tree. A prediction is made by averaging the outputs of the ensemble, $y'$.

Several indicators are used to assess the model performance in this study; the mean square error (MSE), pseudo R-squared (RSQ), goodness of fit ($R^2$) and root mean square error (RMSE) are computed as shown in Equation (7), Equation (8), Equation (9) and Equation (10) respectively:

$$\mathrm{MSE} = n^{-1}\sum_{i=1}^{n}\left(y'_i - y_i\right)^2 \tag{7}$$

$$\mathrm{RSQ} = \left(n - \sum_{i=1}^{n}\left(y'_i - \bar{y}\right)^2\right)\Big/\sum_{i=1}^{n}\left(y_i - \bar{y}\right) \tag{8}$$

$$\mathrm{R}^2 = \sum_{i=1}^{n}\left(y'_i - \bar{y}\right)^2\Big/\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 \tag{9}$$

$$\mathrm{RMSE} = \sqrt{n^{-1}\sum_{i=1}^{n}\left(y'_i - y_i\right)^2} \tag{10}$$

where $n$ is the number of samples; $y_i$ is the actual value of the positioning bias of the $i$th sample, and $\bar{y}$ is the mean of the actual costs of the positioning bias of $n$ samples. Additionally, $y'_i$ is the predicted value of the positioning bias of the $i$th sample using the model. The RF performance is affected by two crucial parameters: '*ntree*' and '*mtry*'. The '*ntree*' variable is the number of regression trees in the RF. The value of *ntree* is set according to the quantitative relationship among the MSE, RSQ, and *ntree*. The minimum value of *ntree* that results in stable MSE and RSQ values is the appropriate value of *ntree*. The '*mtry*' variable is the number of predictive variables sampled for splitting at each node of the regression trees in the RF. We use RF regression implemented with a MATLAB package [45] to create the nonlinear evaluation model. The default of *mtry* in this implementation is the square root of the number of independent variables. To obtain the optimum *mtry* value, model tuning is performed (Figure 5). Specifically, the *mtry* value that produces the lowest RMSE is optimum [46].

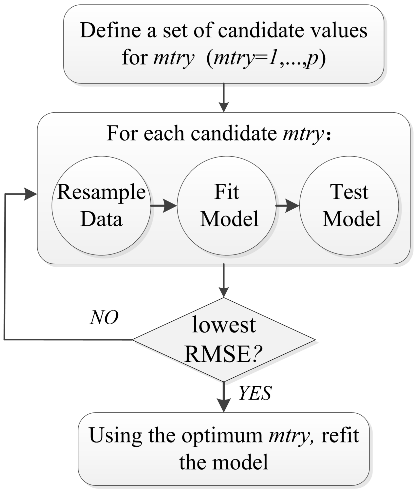

Define a set of candidate values for *mtry* (*mtry=1,…,p*)
For each candidate *mtry*:
Resample Data
Fit Model
Test Model
NO
lowest RMSE?
YES
Using the optimum *mtry,* refit the model

**FIGURE 5.** Flow chart of model tuning for the optimum mtry value.

An RF, as an ensemble of decision trees, inherits the ability to measure the importance of each predictive variable, i.e., how often it is used in the model and how much it contributes to reducing the residual sum of squares. In our study, the importance of the predictive variables is measured by the mean decrease in accuracy. The idea is to permute the values of each feature randomly and measure the effect of this change on the accuracy rate of the tree-based model. For unimportant variables, shuffling the assignment of their values to sample points has little effect on the accuracy rate of the model, but for important variables, the permutation will reduce the accuracy rate. The greater the influence on the accuracy rate of the model is, the more critical the feature.

## V. RESULTS AND DISCUSSION

### A. THE SPATIAL DISTRIBUTION OF POSITIONING BIAS

We first analyse the spatial distribution of the samples and the geographical environment of the significant positioning bias samples. Figure 6(a) shows the spatial distribution of the positioning bias of a volunteer. The trajectory derived from the MPL data is quite different from the actual route (i.e., the route recorded by GPS) of the volunteer (Figure 6(b)). In the central part of this trajectory, the volunteer is crossing the Xuanwu Lake park on foot via a path connecting several islands, three of which have cell towers. However, at times, the phone connects to cell towers on the lakeshore, far from the user. Even in the earlier and later parts of the trajectory, there are only a few cell towers along the volunteer's route. It should be noted that the distance differences in the 5-second

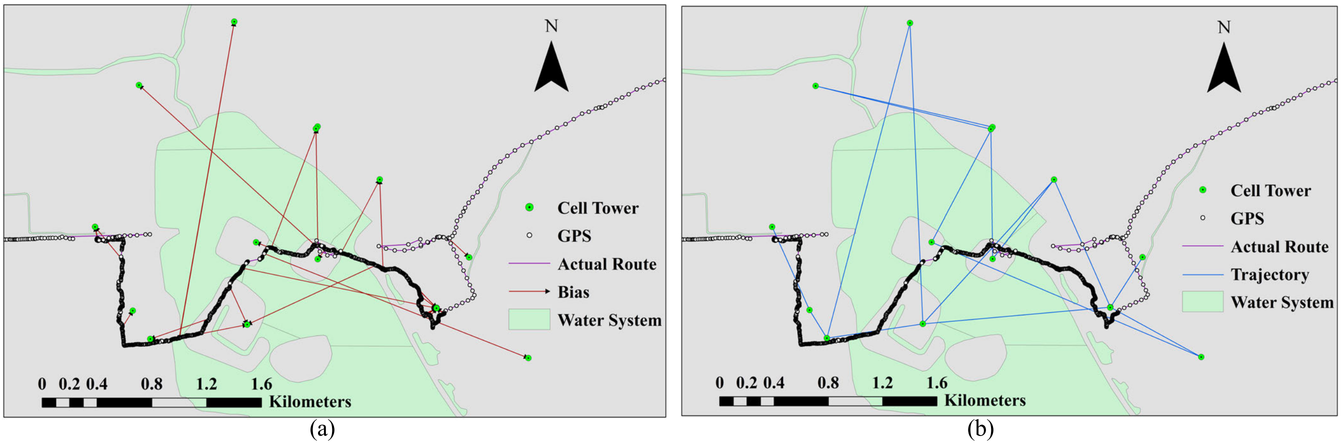


**FIGURE 6. (a) Differences in locations given by mobile positioning and GPS positioning, in the example of GPS positioning with 5-second intervals and mobile positioning with 10-minute intervals in Nanjing, China. (b) The trajectory derived from the MPL data (blue line) and the actual route obtained from the GPS data (purple line) of a volunteer on the ground.**

intervals of the GPS are from the volunteer travelling via different transportation modes.

As stated in Section III-B, we obtained a dataset *D* of 6,112 samples. Figure 7 shows the spatial distribution of the positioning biases of the 6,112 samples located in the study area. Of these samples, 3,004 are located in urbanised districts, and 3,108 are located in predominantly rural districts. We find that the spatial distribution of the positioning bias is related to the geographical environment. Specifically, the biases of samples in areas with a mountain or a lake are usually larger than those in areas with a high density of buildings. As shown in Figure 8, the positioning biases of samples in the Lao Mountain and Purple Mountain scenic area are very high, with a maximum of 7,264.98 m. Similarly, the biases of examples around the Xuanwu Lake park and Yangtze River are also high. These characteristics further indicate that the spatial accuracy of MPL data is greatly affected by certain geographical factors, and there is a relationship between these factors and the positioning bias. Therefore, we proceeded to quantify this relationship.

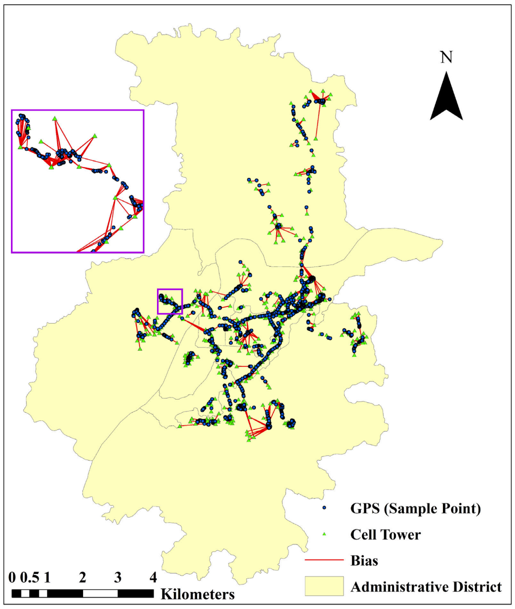


**FIGURE 7. Spatial distribution of the positioning bias of all samples.**

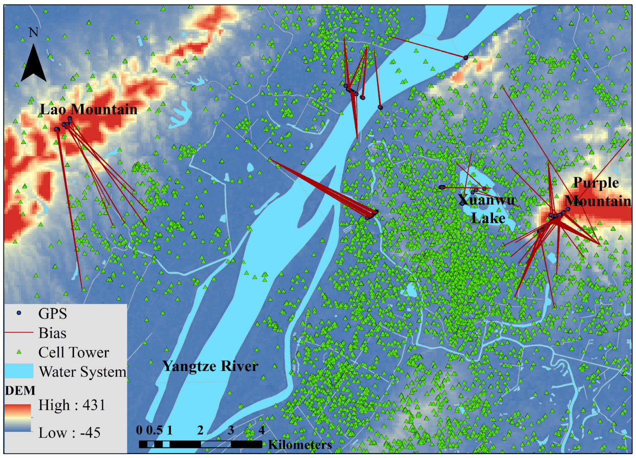


**FIGURE 8. The geographical environment of the significant positioning bias samples.**

## B. CORRELATION AND MULTICOLLINEARITY TESTS

Before building the models described in Section IV, we test the multicollinearity between the predictive variables and the correlation between the positioning bias and the predictive variables. In our study, the SPSS tool is used to diagnose multicollinearity between the predictive variables. The variance inflation factor (VIF) is often used to detect multicollinearity between independent variables. Generally, it is considered that VIF values greater than 10 are often taken as a signal that the data have collinearity problems [47]. As shown in Table 2, all VIF values are less than 10, and most of the predictive variables are significant factors except slope,

**TABLE 2. Multicollinearity between the predictive variables.**

| Predictive variable | VIF | Tolerance | t | Sig. |
|---|---|---|---|---|
| Elevation | 3.093 | 0.323 | 23.204 | 0.000* |
| Slope | 2.288 | 0.437 | -0.159 | 0.873 |
| Aspect | 1.153 | 0.868 | 13.759 | 0.000* |
| Building density | 3.597 | 0.278 | -4.615 | 0.000* |
| DNV | 1.119 | 0.894 | -2.965 | 0.003* |
| DNWB | 1.939 | 0.516 | -0.566 | 0.572 |
| NDWI | 3.266 | 0.306 | 4.317 | 0.000* |
| NDVI | 3.301 | 0.303 | -0.866 | 0.387 |
| Population density | 1.582 | 0.632 | -9.965 | 0.000* |
| Cell tower density | 3.890 | 0.287 | 4.557 | 0.000* |
| DNCT | 1.429 | 0.700 | -6.632 | 0.000* |

*. p ($p < 0.05$)

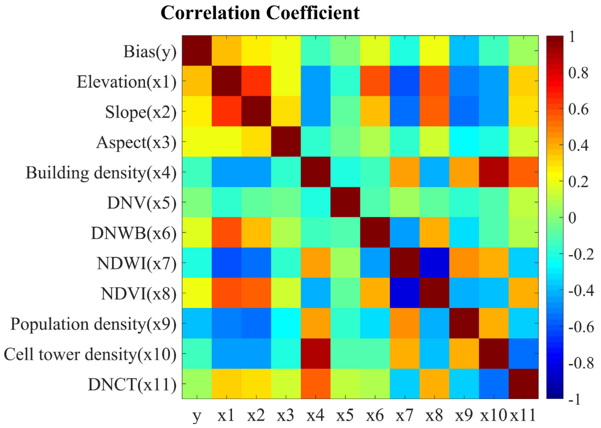


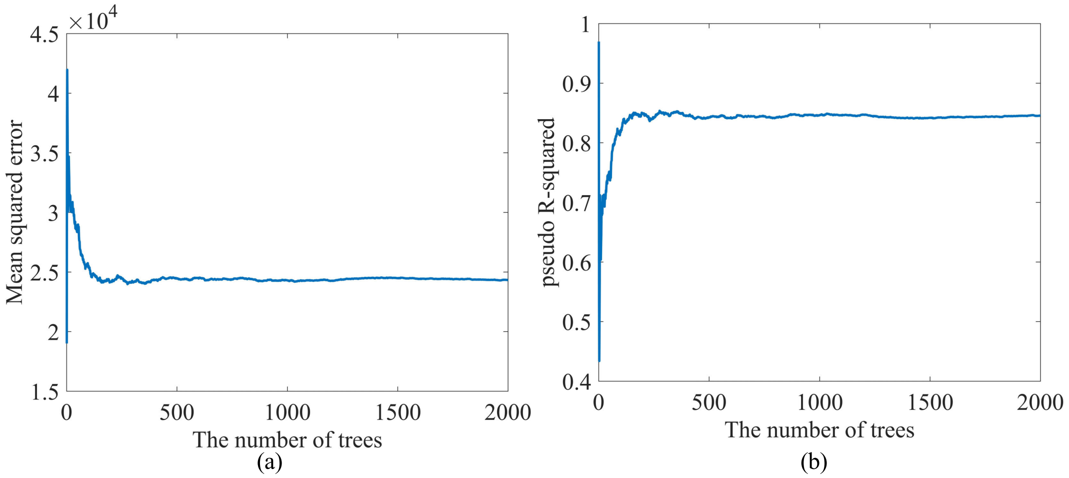
**FIGURE 9. Correlation between the positioning bias y and predictive variables.**

DNWB and NDVI. The VIF values results show that the multicollinearity of the predictive variables has little impact on the regression analysis.

We further investigate the correlation between the predictive variables and the correlation between the positioning bias and the predictive variables. The nonparametric Spearman (rank) correlation coefficient is used to test the correlation. The result shows that all Spearman correlation coefficient values are significant at the 0.01 level. Figure 9 shows that elevation, slope, aspect, DNWB, NDVI, and DNCT are positively correlated with the positioning bias *y*. Accordingly, the building density, DNV, NDWI, population density and cell tower density are negatively correlated with the bias *y*. In addition, there is a strong correlation between the NDWI and NDVI, with a correlation coefficient of -0.82. Similarly, the building density is strongly correlated with the cell tower density, with a correlation coefficient of 0.898. Combined with Table 2, the NDVI (nonsignificant factor) and cell tower density (with the highest VIF value) are removed from the predictive variables dataset to further reduce the impact of multicollinearity on the regression. Furthermore, two other nonsignificant factors (i.e., slope and DNWB) are eliminated, and seven key geographical factors, namely, elevation, aspect, building density, DNV, NDWI, population density and DNCT remain for modelling.

### C. EVALUATION MODEL PERFORMANCE: GWR METHOD VS. RF APPROACH

We compare the two methods using the indicators proposed in Section IV-B. As described in Section IV-A, a Gaussian kernel is specified in the linear evaluation model (i.e., the GWR model). *CV* is used to determine the optimal kernel bandwidth, which is 287.32. For the nonlinear evaluation model (i.e., the RF model), there are two important parameters, *ntree* and *mtry*. Figure 10(a) shows the relationship between the MSE and *ntree* of the model, and Figure 10(b) shows the relationship between the RSQ and *ntree*. The values of both the MSE and RSQ fluctuate rapidly as the number of trees increases, but after about *ntree* = 500, these values change slowly until *ntree* = 1,000. At that point, these metrics tend



(a)

(b)

**FIGURE 10. (a) Relationship between the MSE and ntree. (b) Relationship between the RSQ and ntree.**

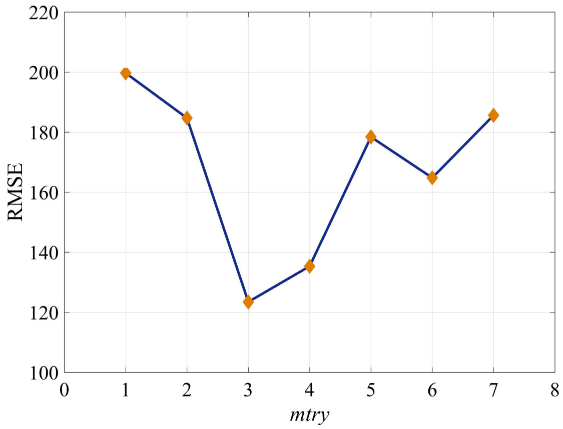


**FIGURE 11.** Relationship between the RMSE and mtry.

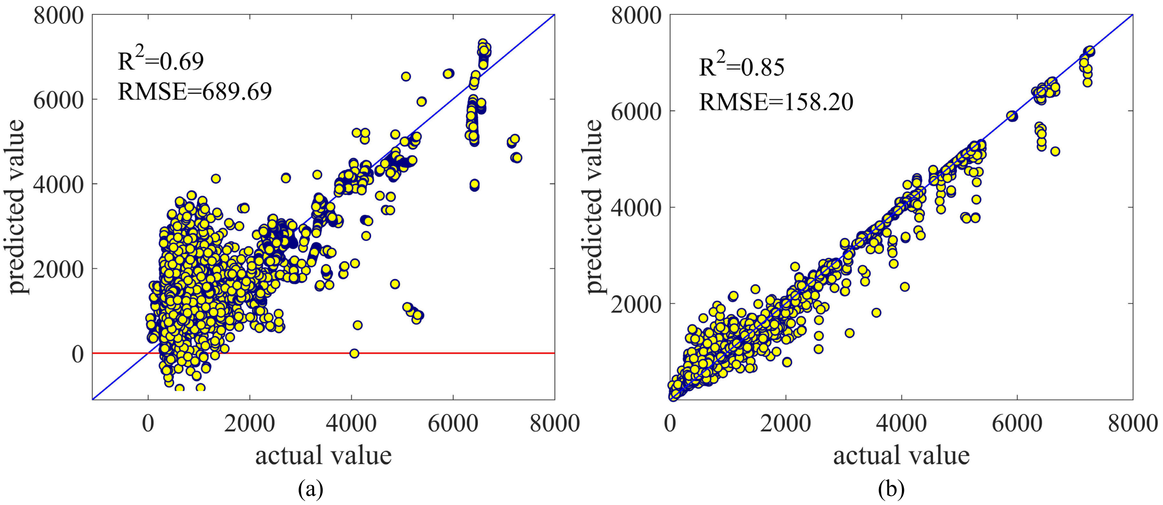
to be stable. Therefore, we set the value of *ntree* to 1000 to stabilise the overall error rate of the model while ensuring reasonable model efficiency. Figure 11 shows the relationship between the RMSE and *mtry* of the model. We obtain the lowest RMSE for *mtry* = 3. This result implies that only a few predictors dominate the others; therefore, if more predictors are tried, they will eliminate the effect of fewer significant predictors, resulting in a model de-emphasises the most important bias sources.

Ten-fold *CV* is used to evaluate the prediction accuracy of the two models. Specifically, the samples are randomly divided into 10 groups, among which 9 groups are taken as the training data and 1 group is taken as the testing data in turn to test the prediction accuracy of the models. Aggregating the results of 10 sets of testing data, Figure 12(a) and Figure 12(b) show scatter diagrams between the predicted and actual values of the GWR model and RF model, respectively. The $R^2$ and RMSE of GWR model are 0.69 and 689.69, respectively. The goodness of fit of the model is good; however, the RMSE is large. As shown in Figure 12(a), there are some large residuals in the predicted results. In particular, the predictions have many values less than zero (below the red line), which is impossible. By comparison, the overall performance of the RF model is better than that of the GWR model. Model evaluation with 10 sets of testing data indicate strong model performance. The $R^2$ and RMSE of the RF model are 0.85 and 158.2, respectively. As shown in Figure 12(b), the residuals of the RF model prediction results are small, and there are no obvious impossible results (e.g., less than zero).

Distributions of $R^2$ and RMSE of the two models' ten-fold *CV* results are shown in Figure 13(a) and Figure 13(b). Figure 13(a) shows that the $R^2$ of the ten-fold *CV* of the GWR model fluctuates between 0.65 and 0.71, and the $R^2$ of the ten-fold *CV* of the RF model fluctuates between 0.81 and 0.88. Figure 13(b) shows that the RMSE of the ten-fold *CV* of the GWR model fluctuates between 636.04 and 695.73, and the RMSE of the ten-fold *CV* of the RF model fluctuates between 136.33 and 203.14. The results show that both models have good stability, and the RF model's overall prediction performance is much better than that of the GWR model.

### D. INFLUENCE EVALUATION OF THE PREDICTIVE VARIABLES

This section further investigates the impact of the predictive variables on the GWR model and RF model. Figure 14 shows the local coefficients of the Gaussian-weighted GWR. Among them, the constant term and NDWI coefficient have a large range. The constant (i.e., intercept) varies from −7,702.8 to 42,438.1, with an average of 4,867.4. The NDWI coefficient varies from −20,169.1 to 6,983.8, with an average of −555. The building density and elevation coefficients also fluctuate greatly. The building density coefficient varies from −557.9 to 3,396.5, with an average of 93.2. The maximum value of the elevation coefficient is 296, the minimum value is −160, and the average value is 4.6. By comparison, the coefficients of the remaining



(a)

(b)

**FIGURE 12.** Comparisons of the actual bias and the model-predicted bias based on (a) the GWR model and (b) the RF model.

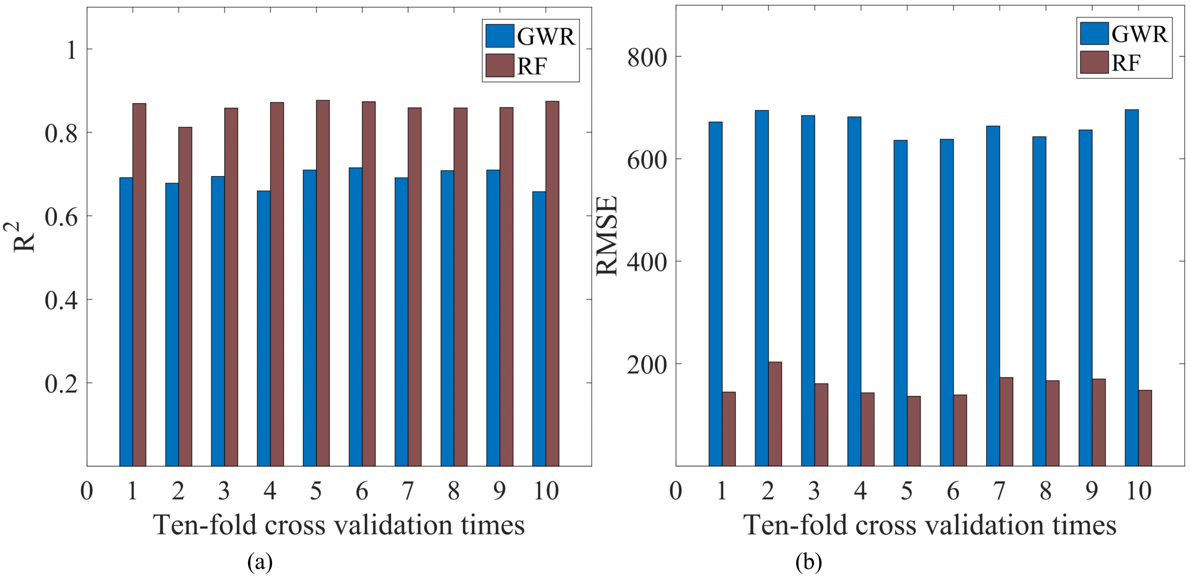


**FIGURE 13.** Distributions of (a) $R^2$ and (b) RMSE of the two models' ten-fold CV.

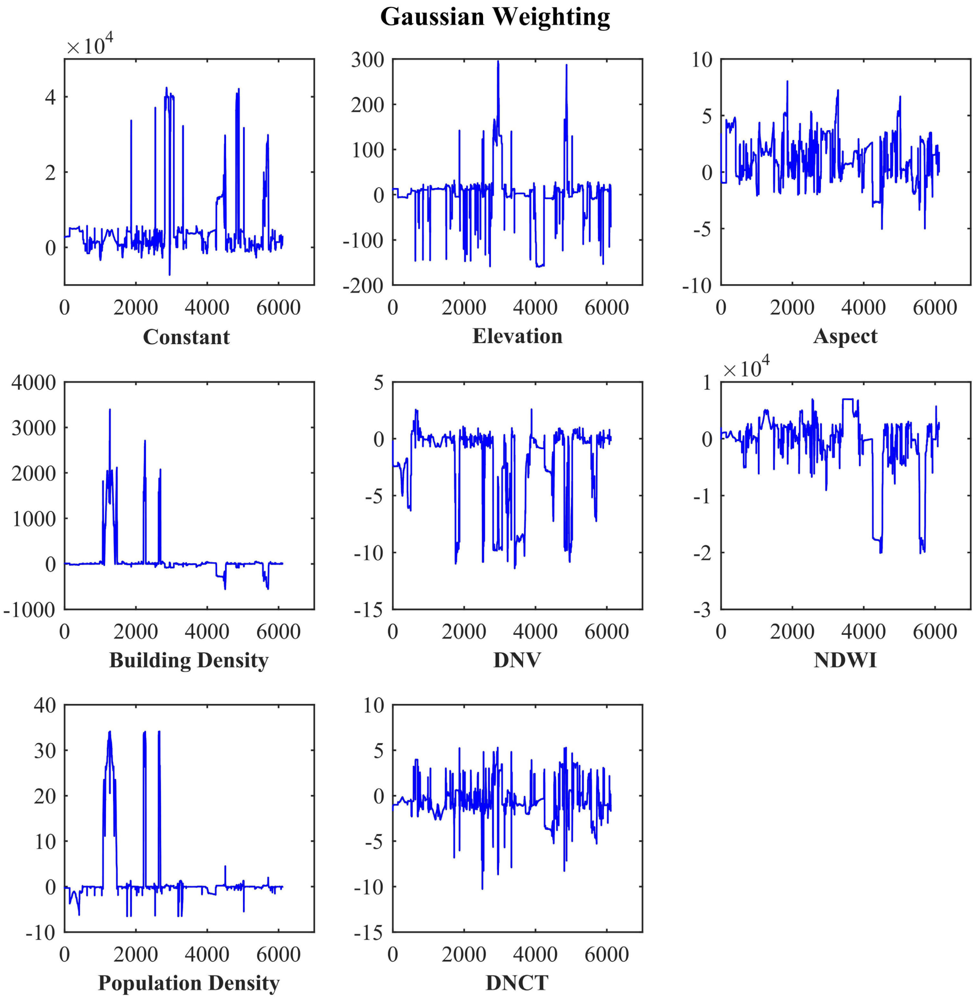


**FIGURE 14.** GWR estimates for Gaussian weighting.

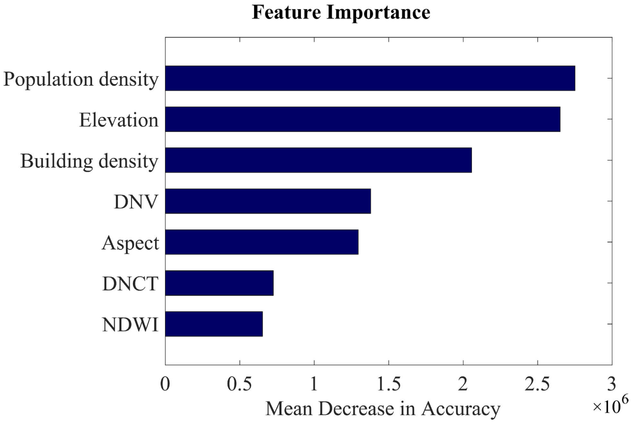


**FIGURE 15.** Influence evaluation of the predictive variables in the RF model.

predictive variables exhibit small ranges. The ranges of the population density, DNCT, DNV and aspect coefficients are 40.7, 15.6, 14 and 13.1, respectively, and the averages are 1.8, −0.5, −2.1 and 1.2, respectively.

Figure 15 shows the importance of the predictive variables in the RF model measured by the mean decrease in accuracy. The population density, elevation and building density are the three most important factors, while the NDWI and DNCT are the least important variables. This result implies that the population density, elevation and building density play a significant role in splitting each node of the regression trees in the RF model. In contrast, the NDWI and DNCT are rarely considered in this process.

### *E. THE PREDICTED SURFACE OF POSITIONING BIAS*

We simulate the spatial distribution of the positioning bias of the MPL data covering the study area based on the two models. The 7 key geographical variables from the study area are collected and pre-processed into a gridded map at 200 m by 200 m resolution. The entire study area is divided into grids of 200 m by 200 m, and the values of all geographical variables are assigned to each grid. The gridded dataset of 7 key geographical variables is input into the GWR model and RF model, and the predicted values of all grid cells are produced. After visualisation, the predicted surface maps of the positioning bias across the study area are generated (Figure 16). The predicted positioning bias varies dramatically across the study area and exhibits several scattered clusters. The predicted positioning bias is further divided into five somewhat arbitrary levels that loosely correspond to the accuracy levels needed for human geography studies. Figure 16(a) shows the spatial distribution of the predictive bias of the GWR model. The very low (0-500 m) and low (500-1,000 m) levels are mainly scattered in the eastern, western and southwestern parts of the study area. The moderate (1,000-1,500 m), high (1,500-2,000 m), and very high (>2,000 m) levels account for most of the area. These areas are mainly located in the northern, most of the southern and some of the western regions. Notably, there are large areas with error values below 0 m in the GWR results, which are mainly distributed in the central, northeast and most of the southeast regions.

Figure 16(b) shows the spatial distribution of the predictive bias of the RF model. The very-low-level areas are scattered

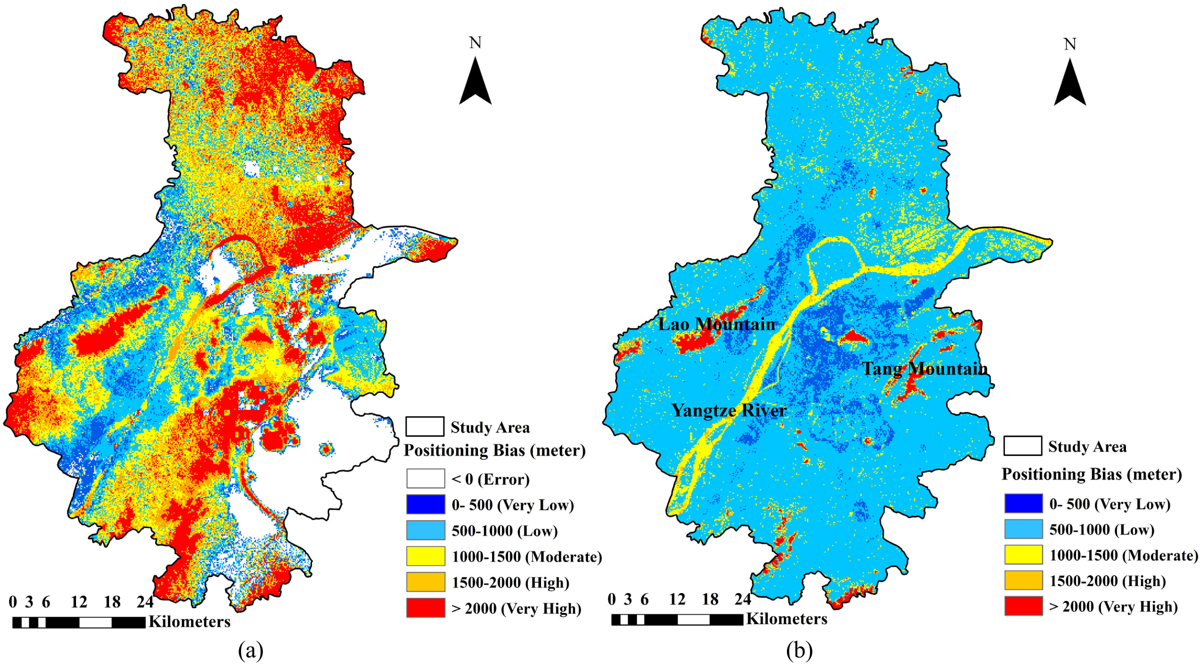


**FIGURE 16.** Predicted spatial distributions of the positioning bias of the MPL data in Nanjing city based on (a) the GWR model and (b) the RF model.

in the central parts of the study area. These zones are primarily urban areas where the building density is high. The low-level areas account for most of the study region, which is mainly cropland. The land use of the moderate level is mainly water bodies, such as the Yangtze River. The high and very high areas are mainly located in suburban regions where mountains or vegetation are the primary types of land use. For example, the very high bias areas (red areas) shown in Figure 16(b) contain two central mountains: Tang Mountain and Lao Mountain. Figure 9 and Figure 15 show that the bias is significantly associated with the population density and elevation. Notably, the elevation in the two red areas is relatively high, and the population density is relatively small. The opposite is true in the very low bias areas. As described in Section V-C, the overall prediction performance of the RF model is much better than that of the GWR model. Therefore, the results of Figure 16(b) are closer to the ground truth. Compared with that of Figure 16(b), the range in the results of Figure 16(a) is larger. It is reflected in the fact that more regions with very high values and many regions with values less than zero in Figure 16(a).

## VI. SUMMARY

The spatial accuracy of data is an essential but often neglected issue in spatiotemporal big data analytics. This issue has been and will always be challenging for the validity of research involving spatiotemporal big data usage. Evaluating the spatial accuracy of data is the basis and premise of spatiotemporal data application. In this study, we aimed to evaluate the spatial accuracy of mobile phone location data, a crucial data source in various research areas. In a case study of Nanjing, we found that seven geographical factors, including population density, elevation, building density, distance to the nearest vegetation, aspect, distance to the nearest cell tower and NDWI, significantly affected the spatial accuracy of MPL data. We built evaluation models to quantify the relationship between the seven geographical factors and the positioning bias of MPL data based on geographical weighted regression and random forest methods, respectively. The experimental results obtained from this study demonstrated the capacity of both models for assessing the positioning bias of MPL data with consideration of the positioning environment. The results of the ten-fold *CV* showed that both the GWR model and RF model exhibited good stability. However, the overall prediction performance of the RF model was much better than that of the GWR model. It was reflected in the fact that the $R^2$ of the RF model (0.85) was higher than the $R^2$ of the GWR model (0.69), and the RMSE of the RF model (158.2) was much smaller than the RMSE of the GWR model (689.69). Furthermore, there were no value less than 0 in the prediction results of the RF model, which appeared in the prediction results of the GWR model. These results demonstrated that the RF model can be useful for spatial accuracy evaluations of MPL data. We further investigated the impact of the predictive variables on the RF model. The importance ranking of the geographical variables showed that the population density, elevation and building density were the three most important factors, while the NDWI and distance to the nearest cell tower were the least important variables. Based on the above models, we simulated the spatial distribution of the positioning bias of the MPL data in the study area. Through a comparative analysis of the results, the GWR model overestimated some low biases and some moderate biases and underestimated some low biases; thus, some incorrect extreme values were generated in the results. The results of the RF model provided a basis for us to use MPL datasets in human mobility studies. This study proposed a useful spatial accuracy evaluation model for MPL data, which can provide an important theoretical basis for big data uncertainty analysis, deepen the understanding of the spatial distribution of positioning bias, and provide scientific guidance for the correct data application.

This study formulated an excellent quantitative relationship model between the positioning bias of MPL data and certain geographical factors. We believe that this quantitative relationship is not unique to the dataset used in this study but exists in other mobile phone datasets. In the future, we intend to enlarge the research scope by repeating the experiments over multiple MPL datasets from different mobile operators and across different study areas. Nevertheless, we hope this study provides some insight that can better use mobile phone location data for future human mobility studies. Moreover, further research is needed to improve this model. Some equipment factors that affect the positioning accuracy, such as the sector azimuth and antenna height of a cell tower, can be considered in the model, and it may improve the performance of the model.

## REFERENCES

[1] L. Yin, N. Lin, X. Song, S. Mei, S.-L. Shaw, Z. Fang, Q. Li, Y. Li, and L. Mao, "Space-time personalized short message service (SMS) for infectious disease control–policies for precise public health," *Appl. Geography*, vol. 114, Jan. 2020, Art. no. 102103.

[2] Z. Fang, X. Yang, Y. Xu, S.-L. Shaw, and L. Yin, "Spatiotemporal model for assessing the stability of urban human convergence and divergence patterns," *Int. J. Geographical Inf. Sci.*, vol. 31, no. 11, pp. 2119–2141, Nov. 2017.

[3] Y. Yuan and M. Raubal, "Analyzing the distribution of human activity space from mobile phone usage: An individual and urban-oriented study," *Int. J. Geographical Inf. Sci.*, vol. 30, no. 8, pp. 1594–1621, Aug. 2016.

[4] P. Wang, T. Hunter, A. M. Bayen, K. Schechtner, and M. C. González, "Understanding road usage patterns in urban areas," *Sci. Rep.*, vol. 2, no. 1, p. 1001, Dec. 2012.

[5] M. Ficek and L. Kencl, "Inter-call mobility model: A spatio-temporal refinement of call data records using a Gaussian mixture model," in *Proc. IEEE INFOCOM*, Orlando, FL, USA, Mar. 2012, pp. 469–477.

[6] D. Schulz, S. Bothe, and C. Körner, "Human mobility from GSM data—A valid alternative to GPS," presented at the Nokia Mobile Data Challenge Workshop, Jun. 2012. [Online]. Available: https://www.idiap.ch/project/mdc/publications/files/mdc-final458-schulz.pdf

[7] P. Katsikouli, M. Fiore, A. Furno, and R. Stanica, "Characterizing and removing oscillations in mobile phone location data," in *Proc. IEEE 20th Int. Symp. World Wireless, Mobile Multimedia Netw. (WoWMoM)*, Jun. 2019, pp. 1–9.

[8] Y. Xu, X. Li, S.-L. Shaw, F. Lu, L. Yin, and B. Y. Chen, "Effects of data preprocessing methods on addressing location uncertainty in mobile signaling data," *Ann. Amer. Assoc. Geographers*, pp. 1–25, Jul. 2020.

[9] L. Pospíšilová and J. Novák, "Mobile phone location data: New challenges for geodemographic research," *Demografie*, vol. 58, no. 4, pp. 320–337, 2016.

[10] R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky, "Human mobility characterization from cellular network data," *Commun. ACM*, vol. 56, no. 1, pp. 74–82, Jan. 2013.

[11] R. Zhong and K. Xiao, "Coverage optimization of GSM wireless network," *Inf. Commun.*, vol. 1, pp. 57–59, Feb. 2008.

[12] D. Liu, Y. Xu, and X. Huang, "Identification of location spoofing in wireless sensor networks in non-line-of-sight conditions," *IEEE Trans. Ind. Informat.*, vol. 14, no. 6, pp. 2375–2384, Jun. 2018.

[13] M. Hata, "Empirical formula for propagation loss in land mobile radio services," *IEEE Trans. Veh. Technol.*, vol. 29, no. 3, pp. 317–325, Aug. 1980.

[14] B. S. Paul and S. Rimer, "Wireless sensor node placement due to power loss effects from surrounding vegetation," in *Proc. Int. Conf. Heterogeneous Netw. Qual., Rel., Secur. Robustness*, 2013, pp. 915–927.

[15] F. Fund, R. Lin, T. Korakis, and S. S. Panwar, "How bad is the flat Earth assumption? Effect of topography on wireless systems," in *Proc. 14th Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, May 2016, pp. 1–5.

[16] R. Ahas, J. Laineste, A. Aasa, and Ü. Mark, "The spatial accuracy of mobile positioning: Some experiences with geographical studies in Estonia," in *Location Based Services and Telecartography*, G. Gartner, W. Cartwright, M. P. Peterson, Eds. Berlin, Germany: Springer, 2007, pp. 445–460.

[17] W. Shi, *Principles of Modeling Uncertainties in Spatial Data and Spatial Analyses*. Beijing, China: Science Press, 2015.

[18] E. Delmelle, C. Dony, I. Casas, M. Jia, and W. Tang, "Visualizing the impact of space-time uncertainties on dengue fever patterns," *Int. J. Geographical Inf. Sci.*, vol. 28, no. 5, pp. 1107–1127, May 2014.

[19] R. Golledge and R. Stimson, *Spatial Behavior: A Geographic Perspective*. New York, NY, USA: Guilford Press, 1998.

[20] L. Mao, L. Yin, X. Song, and S. Mei, "Mapping intra-urban transmission risk of dengue fever with big hourly cellphone data," *Acta Tropica*, vol. 162, pp. 188–195, Oct. 2016.

[21] Z. Zhao, S.-L. Shaw, Y. Xu, F. Lu, J. Chen, and L. Yin, "Understanding the bias of call detail records in human mobility research," *Int. J. Geographical Inf. Sci.*, vol. 30, no. 9, pp. 1738–1762, Sep. 2016.

[22] L. Yin, R. Jiang, Z. Zhao, X. Song, and X. Li, "Exploring the bias of estimating 24-hour population distributions using call detail records," *J. Geoinf. Sci.*, vol. 19, pp. 763–771, Jun. 2017.

[23] Z. Zhao, S.-L. Shaw, L. Yin, Z. Fang, X. Yang, F. Zhang, and S. Wu, "The effect of temporal sampling intervals on typical human mobility indicators obtained from mobile phone location data," *Int. J. Geographical Inf. Sci.*, vol. 33, no. 7, pp. 1471–1495, Jul. 2019.

[24] S. Hoteit, S. Secci, S. Sobolevsky, G. Pujolle, and C. Ratti, "Estimating real human trajectories through mobile phone data," in *Proc. IEEE 14th Int. Conf. Mobile Data Manage.*, Jun. 2013, pp. 148–153.

[25] S. Hoteit, G. Chen, A. Viana, and M. Fiore, "Spatio-temporal completion of call detail records for human mobility analysis," in *Proc. Rencontres Francophones sur la Conception de Protocoles, l'Évaluation de Performance et l'Expérimentation des Réseaux de Communication*. Quiberon, France: HAL Archives-Ouvertes, May 2017.

[26] R. Ahas, A. Aasa, S. Silm, R. Aunap, H. Kalle, and Ü. Mark, "Mobile positioning in space–time behaviour studies: Social positioning method experiments in estonia," *Cartography Geographic Inf. Sci.*, vol. 34, no. 4, pp. 259–273, Jan. 2007.

[27] R. Ahas, A. Aasa, A. Roose, Ü. Mark, and S. Silm, "Evaluating passive mobile positioning data for tourism surveys: An estonian case study," *Tourism Manage.*, vol. 29, no. 3, pp. 469–486, Jun. 2008.

[28] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Ranges of human mobility in Los Angeles and New York," in *Proc. Int. Conf. Pervasive Comput. Commun. Workshops*, 2011, pp. 88–93.

[29] E. Trevisani and A. Vitaletti, "Cell-ID location technique, limits and benefits: An experimental study," in *Proc. 6th IEEE Workshop Mobile Comput. Syst. Appl.*, Dec. 2004, pp. 51–60.

[30] R. Edwards and J. Durkin, "Computer prediction of service areas for VHF mobile radio networks," in *Proc. Inst. Electr. Eng.*, 1969, pp. 1493–1500.

[31] T. Okumura and E. Ohronofi, "Field strength and its variable in VHF and UHF land mobile servervice," *Rev. Electircal Commun. Lab.*, vol. 1, nos. 9–10, pp. 825–873, 1968.

[32] J. Wang, "Wireless base station location technology," *China Comput. Commun.*, vol. 16, pp. 175–177, Aug. 2017.

[33] Z. Liu and Y. Zhao, "A research on the land use structure based on the information entropy—A case study of Nanjing, Jiangsu province," *J. Shanxi Agricult. Sci.*, vol. 41, pp. 968–972 and 998, Sep. 2013.

[34] Q. Xie, X. Liu, and X. Yan, "Base station location optimization based on the Google Earth and ACIS," in *Human Centered Computing*. Cham, Switzerland: Springer, 2016, pp. 487–496.

[35] N. Omaki, T. Imai, K. Kitao, and Y. Okumura, "Improvement of ray tracing in urban street cell environment of non line-of-site (NLOS) with consideration of building corner and its surface roughness," in *Proc. 10th Eur. Conf. Antennas Propag. (EuCAP)*, Apr. 2016, pp. 1–5.

[36] M. Celidonio, E. Fionda, M. Vaser, and E. Restuccia, "NLOS mm Wave propagation measurements through vegetation in urban area: A case study," in *Proc. AEIT Int. Annu. Conf.*, Oct. 2018, pp. 1–6.

[37] S. Tang, Y. Dong, and X. Zhang, "On path loss of NLOS underwater wireless optical communication links," in *Proc. MTS/IEEE OCEANS*, Bergen, Norway, Jun. 2013, pp. 1–3.

[38] D. P. Roy, V. Kovalskyy, H. K. Zhang, E. F. Vermote, L. Yan, S. S. Kumar, and A. Egorov, "Characterization of Landsat-7 to Landsat-8 reflective wavelength and normalized difference vegetation index continuity," *Remote Sens. Environ.*, vol. 185, pp. 57–70, Nov. 2016.

[39] Y. Du, Y. Zhang, F. Ling, Q. Wang, W. Li, and X. Li, "Water bodies' mapping from Sentinel-2 imagery with modified normalized difference water index at 10-m spatial resolution produced by sharpening the SWIR band," *Remote Sens.*, vol. 354, pp. 1–19, Apr. 2016.

[40] J.-Z. Luo, W.-J. Wu, and M. Yang, "Mobile Internet: Terminal devices, networks and services," *Chin. J. Comput.*, vol. 34, no. 11, pp. 2029–2051, Dec. 2011.

[41] A. S. Fotheringham, M. E. Charlton, and C. Brunsdon, "Geographically weighted regression: A natural evolution of the expansion method for spatial data analysis," *Environ. Planning A, Economy Space*, vol. 30, no. 11, pp. 1905–1927, Nov. 1998.

[42] A. S. Fotheringham, C. Brunsdon, and M. Charlton, *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester, U.K.: Wiley, 2002.

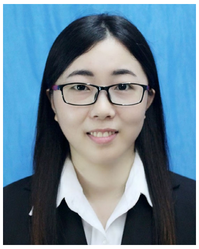
[43] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[44] C. R. Patti, S. S. Shahrbabaki, C. Dissanayaka, and D. Cvetkovic, "Application of random forest classifier for automatic sleep spindle detection," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, Oct. 2015, pp. 1–4.

[45] A. Jaiantilal. (Apr. 19, 2016). *RandomForest-MATLAB*. [Online]. Available: https://github.com/ajaiantilal/randomforest-matlab

[46] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: A classification and regression tool for compound classification and QSAR modeling," *J. Chem. Inf. Comput. Sci.*, vol. 43, p. 1947, Nov. 2003.

[47] R. Salmerón, C. Garcia, and J. Pérez, "Variance inflation factor and condition number in multiple linear regression," *J. Stat. Comput. Simul.*, vol. 88, pp. 1–20, Apr. 2018.



**XIAOQING SONG** received the bachelor's degree from the Qingdao University of Technology, Qingdao, China, in 2014, and the master's degree from Wuhan University, Wuhan, China, in 2017. She is currently pursuing the Ph.D. degree with Nanjing Normal University, Nanjing, China.

Since 2017, she has been with Anhui Normal University, as a Teaching Assistant. Her research interests include geographic information science, machine learning, and spatiotemporal data quality analysis.

**YI LONG** received the Ph.D. degree in cartography and geographical information engineering from Wuhan University, China, in 2002.

He is currently a Full Professor with the Key Laboratory of Virtual Geographic Environment, Ministry of Education, Nanjing Normal University, Nanjing, China. He is the author of two books, more than 150 articles, and more than 30 patents. His research interests include cartography, geography, and geographic information science.

**LING ZHANG** received the Ph.D. degree from the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, in 2014.

He is currently an Associate Professor with the Key Laboratory of Virtual Geographic Environment, Ministry of Education, Nanjing Normal University, Nanjing, China. His research interests include map generalization, multi-scale representation, and geoinformatics (GIS).

**DAVID G. ROSSITER** received the B.S. degree in soil science and international agronomy, the M.S. degree in computer science, and the Ph.D. degree in soil science and international agronomy from Cornell University, in 1973, 1986, and 1988, respectively.

He was retired with the Faculty of Geoinformation Science and Earth Observation, University of Twente, The Netherlands. He is currently a Guest Researcher with the ISRIC–World Soil Information, Wageningen, The Netherlands, an Adjunct Associate Professor with Cornell University, an Invited Professor with Nanjing Normal University, and a Guest Researcher with the Nanjing Soil Research Institute, Chinese Academy of Sciences. He specializes in statistical methods applied to spatial data, especially soil geography.

**FENGYUAN LIU** received the B.S. degree in geographic information science from Nanjing Normal University, Nanjing, China, in 2018, where he is currently pursuing the master's degree in cartography and geographic information systems.

His main research interests include travel geography, big data, and machine learning.

**WEI JIANG** received the B.S. degree in geographic information science from the Kunming University of Science and Technology, Kunming, China, in 2011, and the M.S. and Ph.D. degrees in geographic information science from Wuhan University, Wuhan, China, in 2013 and 2017, respectively.

From 2017 to 2020, he was a Lecturer with Anhui Normal University, Wuhu, China. His research interests include data mining using the machine learning method, fundamental studies of people's purchasing behavior, and the sentiment of tourist flow.