DOI: 10.1111/tgis.12860

RESEARCH ARTICLE

Transactions in GIS WILEY

Influence of geographical determinants on the spatial distribution of positioning uncertainties in mobile phone location data

Xiaoqing Song[1,2] | Ling Zhang[1,3,4] | Sijia Wang[1,3,4] | Yi Long[1,3,4] | Wei Jiang[2] | Qin Hao[1,3,4]

[1]Key Laboratory of Virtual Geographic Environment, Faculty of Geography Science, Nanjing Normal University, Nanjing, P.R. China

[2]School of Geography and Tourism, Anhui Normal University, Wuhu, P.R. China

[3]State Key Laboratory Cultivation Base of Geographical Environment Evolution (Jiangsu Province), Nanjing, P.R. China

[4]Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, P.R. China

Correspondence
Yi Long, Key Laboratory of Virtual Geographic Environment, Faculty of Geography Science, Nanjing Normal University, No. 1 Wenyuan Road, Qixia District, Nanjing 210023, P.R. China.
Email: longyi@njnu.edu.cn

Funding information
Natural Science Foundation of Jiangsu Province, Grant/Award Number: BK20191183; Natural Science Foundation of Anhui Province, Grant/Award Number: 2008085QD166; National Natural Science Foundation of China, Grant/Award Number: 42101419; State Key Program of National Natural Science Foundation of China Grant, Grant/Award Number: 41930104; National Key Research and Development Program of China Grant, Grant/Award Number: 2017YFB0503500

Abstract

The increasing use of mobile phone location (MPL) data in mobility research has provided many insights into people's travel behaviors. Despite these achievements, the spatial distribution of MPL data positioning uncertainties and their influence mechanism are rarely discussed. In this research, we investigate the influence of geographical determinants on the spatial distribution of the positioning uncertainties in MPL data. First, we discuss the spatial distribution trends in the positioning uncertainties of MPL data. Then we apply multiple linear regression and geographical detector (GeoDetector) models to explore the influence mechanism on the spatial distribution of the positioning uncertainties. By applying these methods to MPL data sets from a major operator in Nanjing city, we find a spatial aggregation phenomenon in the positioning uncertainties. Elevation contributes most to the spatial distribution of the positioning uncertainties. Furthermore, the influencing power of geographical factors on the spatial distribution of positioning uncertainties is nonlinearly enhanced after an interaction.

 © 2021 John Wiley & Sons Ltd

1 | INTRODUCTION

Ensuring data quality is the premise of big data research, and it is the basis of data analysis, data mining, and decision-making. Poor-quality big data will not only reduce the quality of decision-making but may even lead to catastrophic losses (Liu & Pang, 2019). Mobile phone location (MPL) data are important spatiotemporal big data. In a broad sense, MPL data can be derived from location-aware devices that are included on mobile phones (e.g., a GPS chip and Wi-Fi) or mobile positioning techniques (Zhao et al., 2018). Note that MPL data discussed in this study refer to the MPL data collected by mobile communication networks using mobile positioning techniques for billing, troubleshooting, or other technical measurement purposes. Recent years have witnessed the potential for MPL data to become a significant data source in the study of human mobility for transportation (Jiang et al., 2016; Wang, Wang, & Lian, 2020), urban planning (Widhalm, Yang, Ulm, Athavale, & González, 2015; Yu, Li, Yang, & Zhang, 2020), epidemiology (Mao, Yin, Song, & Mei, 2016; Yin et al., 2020), and sociology (Blumenstock, Cadamuro, & On, 2015; Liu et al., 2020). The quality of MPL data, especially the spatial accuracy or uncertainty, is directly related to the reliability of these human mobility research results and should be carefully examined.

The spatial accuracy of MPL data mainly depends on mobile positioning technology. In a mobile communication network, several positioning techniques are used to estimate subscribers' locations. These include the cell of origin (COO), angle of arrival, differential time of arrival, and so on (Yassin & Rachid, 2015). The COO method is simple in principle and does not consume large amounts of computing and network resources. This technology is the most common mobile positioning technology in mobile communication networks (Trevisani & Vitaletti, 2004). The positioning principle of the COO method is to take the geographical coordinates of the cell tower serving the mobile phone at a given time as an approximation of the subscriber's position (Adusei, Kyamakya, & Jobmann, 2002; Kos, Grgic, & Sisul, 2006). Therefore, the spatial accuracy of MPL data is at the cell tower level. In recent years, increasing attention has been given to evaluating the spatial accuracy of MPL data. Horn, Klampfl, Cik, and Reiter (2014) found that cell phone devices' position accuracy is approximately 500 m in urban areas with a high antenna density. In contrast, in rural and alpine regions of Austria, the position accuracy is approximately 2.5 km. Kamenjuk, Aasa, and Sellin (2017) found that the spatial accuracy is 100–1,000 m in Estonian towns (such as Tallinn and Tartu), whereas in the rest of Estonia, the accuracy falls to between 1.5 and 20 km. Pospíšilová and Novák (2016) found that the spatial accuracy reached 200 m in the urban center of Prague, while it dropped to 2,000 m in the suburbs. The Czech Republic also had similar characteristics: the positioning accuracy in the regional center was high (1.3 km), but in the rural area it was as low as 6 km. Song, Long, et al. (2020) found that the regions with high-level spatial accuracy are primarily scattered in urban areas where the building density is high. The land-use type of the moderate-level spatial accuracy areas is mostly water bodies. The very low-level spatial accuracy areas are mainly located in suburban regions where mountains or vegetation are the primary land-use types. All these results show that the spatial accuracy of MPL data is high in the urban center, where cell towers are densely distributed, but low in the suburbs, where cell towers are more sparsely distributed. This correlation suggests that the spatial accuracy of MPL data may be severely affected by the cell tower density.

The positioning inaccuracy in MPL data causes much of the uncertainty. As mentioned above, the locations documented in the MPL data are at the cell tower level. In a mobile communication network, a cell phone's signal can oscillate between neighboring or even distant cell phone towers due to load balancing or signal strength variations (Kwan, 2016). These uncertainty issues have hindered the ability to obtain reliable results in human mobility studies that are important in many geospatial applications. The research attention on MPL data uncertainty appears to have become notable only in recent years (Chen, Ma, Susilo, Liu, & Wang, 2016; Xu et al., 2020; Zhao et al., 2016, 2019). Zhao et al. (2016) compared three commonly used mobility metrics derived from call detail records (CDRs) with a data set 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, Jiang, Zhao, Song, and Li (2017) evaluated the bias of the population distribution derived from CDRs using a mobile signaling data set tracked by 24-h users as comparative data. They found that the median

relative error ranged from 25 to 30% during hours when humans were active. Xu et al. (2020) found that preprocessing methods could lead to changes in the data characteristics. They introduced much more uncertainty into the characterization and interpretation of human mobility patterns.

Many studies have focused on the sources and factors influencing the positional accuracy or uncertainty of MPL data. They are mainly divided into two aspects: the communication engineering perspective and the geographical perspective. From the communication engineering perspective, the research in this field focuses on the coverage of cell towers and the propagation path loss of wireless signals. These are closely related to the positioning accuracy of MPL data. Many studies show that the coverage of cell towers is greatly affected by the equipment factors of the positioning system, such as carrier frequency, antenna height (Hata, 1980), and the downdip antenna angle of cell towers (Zhong & Xiao, 2008). The propagation path loss of the wireless signals is also affected by the height of the mobile station, building height and other factors (Harinda, Hosseinzadeh, Larijani, & Gibson, 2019). Many models have been proposed to simulate and predict the propagation path loss of wireless signals in various channel environments (Edwards & Durkin, 1969; Harinda et al., 2019; Hata, 1980; Okumura & Ohronofi, 1968; Singh, 2012). From the geographical perspective, the research focuses on the factors influencing the positioning accuracy of MPL data. Many studies have shown that the spatial accuracy of MPL data is mainly affected by the complex geographical environment in addition to positioning technology. On the one hand, the complex geographical environment will cause the non-line-of-sight (NLOS) propagation of wireless signals, resulting in large errors (Liu, Xu, & Huang, 2018). On the other hand, the geographical environment directly affects the site selection of cell towers and then affects the positioning accuracy of the MPL data (Wang, 2017). In recent years, some studies have noted that the spatial accuracy of MPL data is affected by some geographical elements, such as terrain and vegetation (Fund, Lin, Korakis, & Panwar, 2016; Paul & Rimer, 2013). However, the exploration of the relationship between geographical factors and the spatial distribution of the uncertainty of MPL data remains in its infancy. To the best of our knowledge, no studies have explored the spatial distribution trends of the positioning uncertainty in MPL data and the influence mechanism of geographical factors on the spatial distribution of the uncertainty of MPL data. These studies are of great significance for predicting and simulating MPL data positioning accuracy and further evaluating the availability and applicability of the data (Song, Long, et al., 2020).

In this research, we investigate the influence of geographical determinants on the spatial distribution of the positioning uncertainties in MPL data. By taking high-frequency and high-positioning-accuracy GPS data as the "ground truth," we examine how the positioning uncertainty of MPL data is distributed and varies across a surface. Furthermore, we detect the geographical determinants of the positioning uncertainty and how these geographical determinants impact the spatial distribution of the positioning uncertainties. This article is organized as follows. In Section 2 we describe the data sets used in our study, which include MPL data, GPS data and potential geographical controlling factor data. In Section 3 we introduce the methods used to cluster samples and to detect the geographical determinants of positioning uncertainty. In Section 4 we discuss the results of the experiments conducted in this work in detail. Section 5 concludes.

2 | DATA SETS

2.1 | Spatiotemporal trajectory data

This study explores the influence of geographical determinants on the spatial distribution of the positioning uncertainties in MPL data. Positioning uncertainty can be understood as positioning inaccuracy, which may refer to the positioning bias of a single measurement or the degree of variation in the value of multiple measures (Shi, 2015). To obtain the positioning bias of MPL data, we used GPS data with high spatial accuracy as the reference data for comparison with MPL data. Therefore, two spatiotemporal trajectory data sets were collected in this article: MPL data and GPS data.

To ensure smooth communication and other interactive behaviors between mobile terminals and cell towers, mobile communication networks often record relevant control information. MPL data are a byproduct of this control information stored in the databases of mobile operators. The interaction events between mobile terminals and cell towers are mainly divided into call order events and location update events. The call order events include making or answering phone calls, sending or receiving short messages, and surfing the internet. When these events occur, the user's spatial and temporal location information is recorded. The location update events mainly occur in three cases: power-on or power-off updates, periodic updates, and handover updates. Power-on or power-off updates occur when a mobile phone is turned on/off by the user. When a user moves from a cell tower's service area to that of another tower, a handover update is triggered. Periodic updates are triggered when there are no interaction events between mobile terminals and cell towers (Zhao et al., 2016). The MPL data set used in our study is all from a major operator in Nanjing city. This data set is the MPL data recorded by a mobile communication network using COO positioning technology. When the above interaction events occur, the mobile communication network records the spatiotemporal location information of the cell tower connected to the mobile terminal. The location information is taken as the approximate location of the mobile terminal. Table 1 shows an example of an individual user's MPL data for one day. Each record in this MPL data set comprises the user ID (i.e., SIM card number), recording date, starting time, and coordinates (longitude/latitude in the WGS84 coordinate reference system) of the cell tower connected to the mobile terminal.

To obtain the positioning bias of MPL data, we used GPS data with high spatial accuracy as the reference data for comparison with the MPL data. As a high-precision radio navigation and positioning system based on artificial satellites, GPS can record precise spatiotemporal locations of mobile terminals. Modern smartphones are usually equipped with an embedded GPS receiver, which provides convenience and feasibility to collect individual MPL data and GPS data in the same period. GPS data acquisition in our study was performed by a GPS data collection application. Specifically, this software mainly includes a user control module for data acquisition and a background service module for data storage and calls. In the user control module, users can switch the software on and off, set the time interval of data acquisition, and upload data. The background service module supports the storage and downloading of GPS data. Table 2 shows an example of an individual user's GPS data for one day. Each record in this GPS data set comprises the phone ID, recording date, recording time, coordinates (longitude/latitude in the WGS84 coordinate reference system) of the mobile terminal, and positioning spatial accuracy ε.

It is worth noting that the aforementioned Nanjing operator planned and investigated the whole communication network before installing cell towers to achieve the best communication quality. The subsequent coverage

TABLE 1 Example of an individual user's MPL records during the data collection period

User ID	Date	Starting time	Longitude	Latitude
153********	Day 1	08:54:37	118.****	32.****
...	...	...	118.****	32.****
153********	Day 1	19:20:16	118.****	32.****

TABLE 2 Example of an individual user's GPS records during the data collection period

Phone ID	Date	Time	Longitude	Latitude	Accuracy (ε)
8671********836	Day 1	08:27:39	118.****	32.****	9
...	...	...	118.****	32.****	...
8671********836	Day 1	20:19:40	118.****	32.****	3

optimization is a process of phased and fixed-point adjustments. In the short term of our data acquisition, the equipment conditions of the whole communication network were relatively stable. To collect individual MPL data and GPS data simultaneously, we recruited 60 volunteer college students to perform data collection work for two days. Before data collection, to ensure that the samples were representative of the target population (i.e., all of the geographical environments in Nanjing), we constructed 120 itineraries so that volunteers could plan routes covering different geographical contexts, including various elevations, tall building densities, park areas, and lakes. All the smartphones and SIM cards used for data collection were provided by our research group. Each volunteer was equipped with a smartphone with the GPS data collection application mentioned above installed. Each smartphone was equipped with a SIM card from the mentioned operator. We associated the SIM card number one-to-one with the phone ID and assigned the SIM card number to the GPS record exported by the smartphone equipped with that SIM card. In addition, all the smartphones used for data collection in our study were the same brand and model to reduce the potential biases caused by equipment conditions. Volunteers used these smartphones to collect their GPS trajectory data and permitted us to access their MPL data during the data collection period. All the volunteers signed an informed consent form with us and signed a data output authorization letter with the mobile operator.

After data collection, we performed the following data preprocessing steps on the MPL and GPS data. Since the calculation of the MPL data positioning bias was based on the spatial plane coordinates, we first converted the WGS84 coordinate reference system in the original MPL and GPS data into the Beijing 1954 3-Degree Gauss–Krüger CM 117E projected coordinate system. Then we filtered the GPS data records based on the positioning spatial accuracy ε. Only the records with $\varepsilon < 3$ m in the GPS data set were retained in our study. Note that in this study we did not eliminate the oscillations in the MPL data because they are also an important manifestation of positioning uncertainties in MPL data. In addition, since the data acquisition devices and SIM cards were provided by our research group, the collected MPL data did not contain location data triggered by receiving and making calls.

Figure 1 shows the general statistical characteristics of the two data sets. There are 120 trajectories in both the MPL and GPS data sets. The numbers of records in a trajectory in the MPL data vary notably, with mean and median values of 1,018.99 and 344, respectively (Figure 1a). Comparatively speaking, the variation in the GPS trajectories is more obvious, with mean and median values of 5,822.7 and 6,933, respectively (Figure 1b). The inter-record time is measured as the duration between two consecutive records in a trajectory. The inter-record time of the MPL data ranges from a few seconds to less than 10 min (Figure 1c). The mean and median values are 41.7 and 13 s, respectively. The range of the inter-record time of the GPS data is smaller, with mean and median values of 4.21 and 3 s, respectively (Figure 1d).

2.2 | Potential controlling factors data

We collected (from various sources) a set of 12 layers of geographical variables (factors) that are available from the public domain or mobile operator and may conceivably correlate with the positioning uncertainty in the MPL data. The 12 factors can be roughly grouped into physical and human geographical factors. Their brief descriptions and sources are summarized in Table 3. The physical geographical factors mainly include the elevation, slope, aspect, distance to the nearest vegetation (DNV), distance to the nearest water body (DNWB), modified normalized difference water index (NDWI), and normalized difference vegetation index (NDVI). Topography strongly impacts the site selection of cell towers (Xie, Liu, & Yan, 2016), which indirectly affects the spatial accuracy of MPL data. Vegetation can cause NLOS propagation of wireless signals, and water bodies usually refract or reflect wireless signals, causing a loss of wireless signal propagation. These are all sources of the positioning bias of MPL data (Celidonio, Fionda, Vaser, & Restuccia, 2018; Tang, Dong, & Zhang, 2013). We used these physical geographical factors to describe the topography, vegetation, and water bodies around the volunteers. Human geographical factors include the building density, building height, road density, population density, and distance to the nearest cell tower (DNCT). Buildings can cause NLOS propagation (Omaki, Imai,

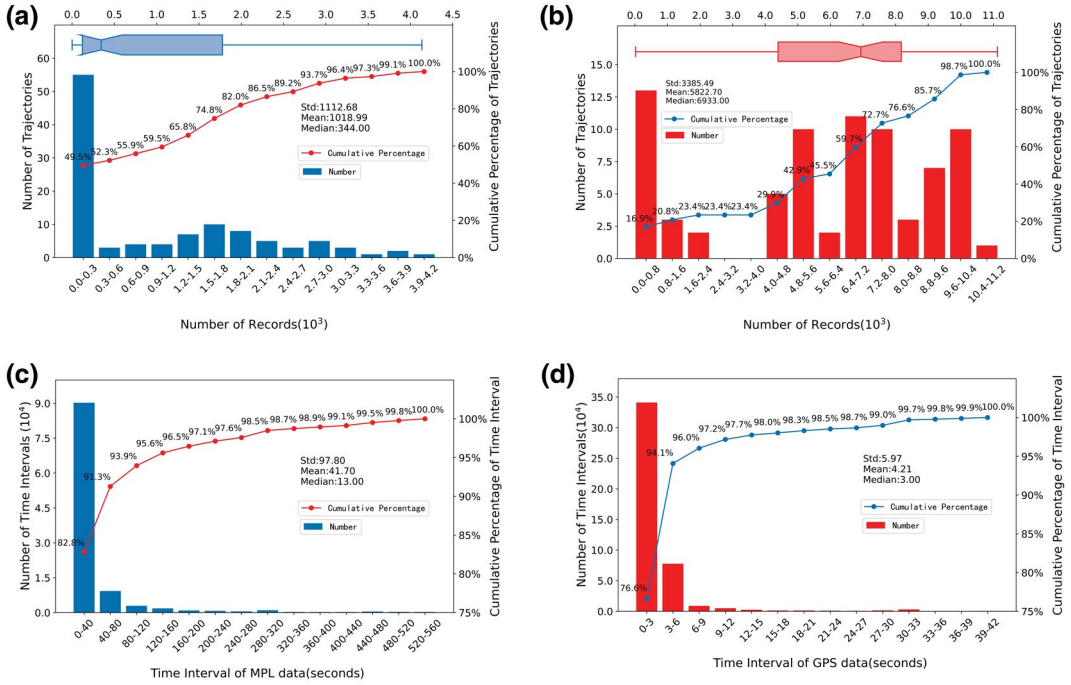

FIGURE 1 General statistics of the MPL and GPS data: (a) number of records per trajectory in the MPL data; (b) number of records per trajectory in the GPS data; (c) distribution of the time interval of the MPL data; and (d) distribution of the time interval of the GPS data

Kitao, & Okumura, 2016). As mentioned above, the spatial accuracy of the MPL data is closely related to the spatial distribution of the cell towers. In addition to these error sources, the positioning bias of the MPL data is also affected by load balance issues (Ogulenko, Benenson, Omer, & Alon, 2020). To some extent, population density and road density can reflect the degree of communication congestion. We used these human geographical factors to describe the potential error sources in the human geographical environment.

Considering that the mean spatial distance between cell towers in the study area is 95.6 m, the study area is divided into 100 m by 100 m grids. That is, the basic spatial unit is a 100 m by 100 m grid. Similarly, the 12 layers of geographical variables described in Table 3 are prepared at a resolution of 100 m by 100 m.

3 | METHODOLOGIES

3.1 | Definitions and framework

In this section we first clarify some terms and variables used in this article. Then we briefly introduce the analytical framework of our study.

3.1.1 | Definitions

To clearly describe the quantization of the positioning bias of the MPL data in our study, we defined several key concepts, as shown in Figure 2.

TABLE 3 Factors selected for analysis

Category	Factor	Factor description	Data source
Physical geographical factors	Elevation	Vertical distance above or below sea level at a location on the ground	90 m Digital Elevation Database from the NASA CGIAR-CSI Geoportal, SRTM resampled to 100 m resolution
	Slope	Degree of steepness of surface units expressed as the ratio of the vertical height to the horizontal distance of the surface	Derived from the elevation data, 100 m resolution
	Aspect	Azimuth of the projection of the normal of the sloped surface on the horizontal plane	Derived from the elevation data, 100 m resolution
	DNV	Distance to the nearest vegetation	Derived from the vegetation data (crawled from OpenStreetMap), 100 m resolution
	DNWB	Distance to the nearest water body	Derived from the water data (crawled from OpenStreetMap), 100 m resolution
	NDWI	Normalized difference water index, an index reflecting water information	Derived from the remote sensing image data (*Landsat 8* 30 m satellite digital products), 100 m resolution
	NDVI	Normalized difference vegetation index, an index reflecting vegetation coverage	Derived from the remote sensing image data (*Landsat 8* 30 m satellite digital products), 100 m resolution
Human geographical factors	Building density	Density of buildings within a spatial unit	Derived from building POI data (crawled from OpenStreetMap), 100 m resolution
	Building height	Average height of buildings within a spatial unit	Derived from building data (crawled from OpenStreetMap), 100 m resolution
	Road density	Density of roads within a spatial unit	Derived from road data (crawled from OpenStreetMap), 100 m resolution
	Population density	Population density within a spatial unit	LandScan 2020 Global Population Database, 100 m resolution
	DNCT	Distance to the nearest cell tower	Mobile operator, 100 m resolution

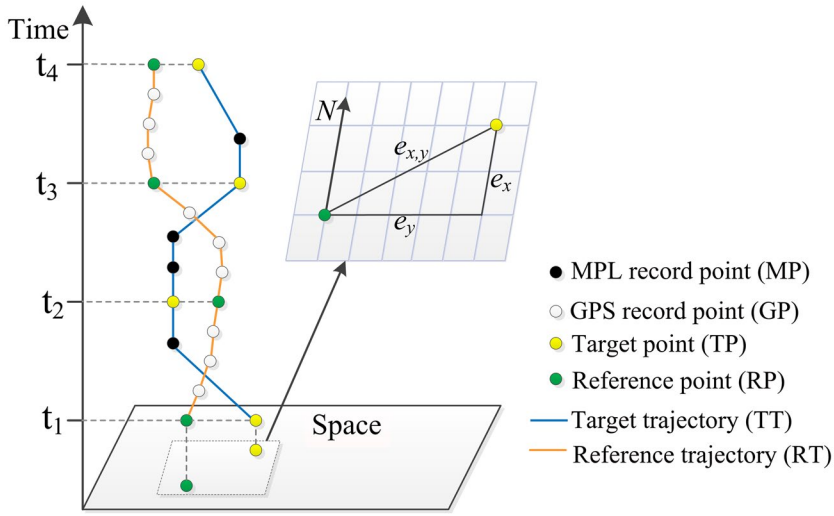

FIGURE 2 Spatiotemporal trajectories of individual MPL and GPS records

1. The *MPL record point* (*MP*) refers to the spatiotemporal location point of mobile terminals based on a mobile communication network, which is expressed as:

$$MP = (x_{MP}, y_{MP}, t_{MP}) \tag{1}$$

where x_{MP} and y_{MP} denote the vertical position and horizontal position of the mobile terminals based on the mobile communication network in the spatial plane respectively, and t_{MP} indicates the positioning time of the mobile terminals in the mobile communication network.

2. The *GPS record point* (*GP*) refers to the spatiotemporal location point of mobile terminals based on GPS, which is defined as:

$$GP = (x_{GP}, y_{GP}, t_{GP}) \tag{2}$$

where x_{GP} and y_{GP} denote the vertical position and horizontal position of the mobile terminals based on GPS positioning technology in the spatial plane respectively, and t_{GP} indicates the positioning time of the mobile terminals in the GPS.

3. In particular, when $t_{MP} = t_{GP} = t_c$, *MP* is a *target point* (*TP*), and *GP* is a *reference point* (*RP*). These are expressed as:

$$\begin{aligned} TP &= MP = (x_{TP}, y_{TP}, t_c) \\ RP &= GP = (x_{RP}, y_{RP}, t_c) \end{aligned} \tag{3}$$

where t_c is the timestamp when the mobile communication network and GPS locate the same mobile terminal at the same time; x_{TP} and y_{TP} denote the location of the *TP* in the vertical and horizontal directions of the spatial plane at this timestamp, respectively; and x_{RP} and y_{RP} denote the vertical position and horizontal position of the *RP* in the spatial plane at this timestamp, respectively.

4. The *target trajectory* (*TT*) is defined as a set of spatiotemporal location point sequences of a mobile terminal based on mobile communication network positioning in a certain period *T*:

$$TT = \{MP_1, \ldots, MP_m\} = \left\{\left(x_{MP_1}, y_{MP_1}, t_{MP_1}\right), \ldots, \left(x_{MP_m}, y_{MP_m}, t_{MP_m}\right)\right\} \tag{4}$$

where MP_i represents the spatiotemporal location point of the *i*th observation in the positioning sequence based on a mobile communication network; x_{MPi} and y_{MPi} denote the vertical position and horizontal position of the *i*th observation in the spatial plane, respectively; t_{MPi} represents the timestamp of the *i*th observation; and $i = 1, 2, \ldots, m$.

5. The *reference trajectory* (*RT*) is defined as a set of spatiotemporal location point sequences of the same mobile terminal based on GPS positioning in the same period *T*:

$$RT = \{GP_1, \ldots, GP_n\} = \left\{\left(x_{GP_1}, y_{GP_1}, t_{GP_1}\right), \ldots, \left(x_{GP_n}, y_{GP_n}, t_{GP_n}\right)\right\} \tag{5}$$

where GP_j represents the spatiotemporal location point of the *j*th observation in the positioning sequence based on GPS; x_{GPj} and y_{GPj} denote the vertical position and horizontal position of the *j*th observation in the spatial plane, respectively; t_{GPj} indicates the timestamp of the *j*th observation; and $j = 1, 2, \ldots, n$.

As shown in Figure 2, the positioning bias of MPL data at timestamp *t* can be expressed as the length of the line between *TP* and *RP*:

$$e_{x,y} = \mathrm{DIS}(TP, RP) = \sqrt{e_x^2 + e_y^2} \tag{6}$$

where $e_{x,y}$ is the positioning bias of the MPL data at timestamp *t*; DIS is the Euclidean distance function; e_x is the positioning bias in the vertical direction, $e_x = |x_{TP} - x_{RP}|$; and e_y is the positioning bias in the horizontal direction, $e_y = |y_{TP} - y_{RP}|$.

3.1.2 | Framework

This study mainly consists of trend exploration of the spatial distribution of the positioning uncertainties and influence mechanism mining. The overall analytical framework is illustrated in Figure 3.

After the preprocessing for the MPL and GPS data, the sample point pairs containing *TP* and *RP* are obtained, and the positioning biases of the *RP*s are calculated by Equation (6). The cell towers can be divided into microcell towers and macrocell towers. Microcell towers are often distributed in areas with dense buildings, and their coverage is small. Macrocell towers are often distributed in open areas, and their coverage is large. The MPL data uncertainty has different spatial distribution characteristics in microcell and macrocell tower regions, and the influence mechanism of geographical factors on its spatial distribution may also be different. Therefore, this study clustered the samples based on the type attributes of the cell towers. Through a cooperative project with the operator mentioned above, we obtained the spatial distribution of the coverage of different cell tower types. We classified the samples that fell within the coverage of the microcell towers into one cluster (i.e., samples in the microcell tower regions) and the sample points that fell within the coverage of the macrocell towers into another cluster (i.e., samples in the macrocell tower regions). The spatial distribution trend and influence mechanism of MPL data uncertainty of the two clusters were further explored.

The magnitude and patterns of influence of the geographical factors are estimated using the multiple linear regression (MLR) model and geographical detector (GeoDetector) model, as described in detail in Sections 3.2 and 3.3, respectively. We construct data set *D1* for the MLR model to explore the influence power of geographical factors on the positioning biases of the totality of sample points. Each sample in *D1* includes a positioning bias (calculated by Equation 6) and a 12-dimensional vector of the geographical factors (as shown in Table 3). To further explore the geographical determinants and influence mechanism of the spatial distribution of the MPL data positioning uncertainties, we construct data set *D2* for the GeoDetector model. As mentioned above, the study area is gridded to 100 m by 100 m resolution. We calculated the positioning bias value for each grid using the mean value of the positioning biases of sample points *RP*s within the grid. Furthermore, the standard deviation (*SD*) of these samples' positioning biases can be taken to measure the degree of positioning dispersion of this cell. Each

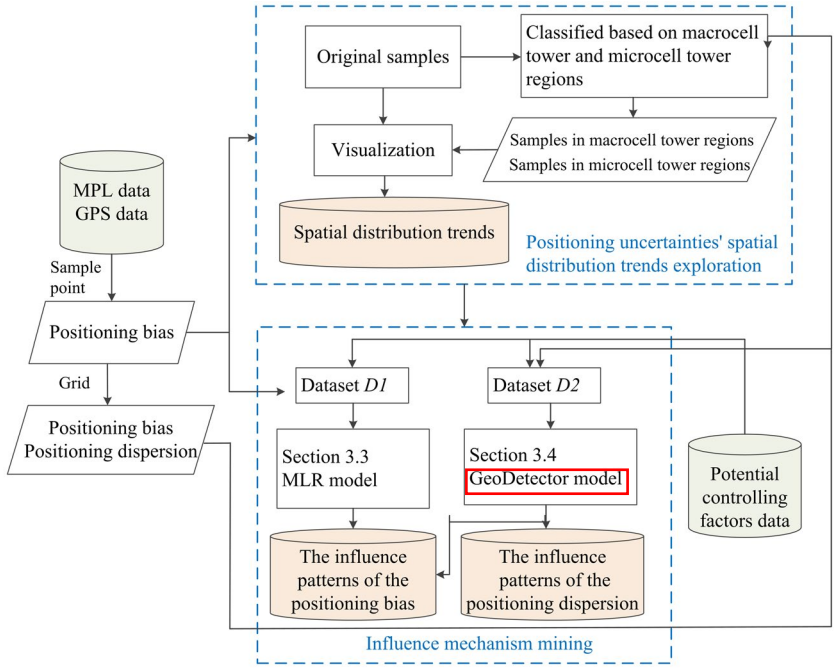

FIGURE 3 The analytical framework of this study

grid in *D2* contained a positioning bias value, a positioning dispersion value, and 12 geographical factor values. All the sample points in *D1* and all the sample grids in *D2* were classified into samples in microcell tower regions and samples in macrocell tower regions. The above analysis was repeated for different sample clusters to explore the influence mechanism of geographical factors on the spatial distribution of MPL data uncertainty in microcell and macrocell tower regions.

3.2 | MLR model

MLR models are often used to study the relationship between a dependent variable *Y* and multiple explanatory variables $\{X_1, X_2, ..., X_k\}$ and analyze the influence of explanatory variables on dependent variables. This method has been widely used in geographic analyses (Akinbile, Ogunmola, Abolude, & Akande, 2020; Ruiz-Álvarez, Alonso-Sarria, & Gomariz-Castillo, 2019). The general form of an MLR model can be expressed as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i,\ i = 1, 2, \ldots, n \tag{7}$$

where *k* is the number of explanatory variables; β_0 is the constant term; β_j ($j = 1, 2, ..., k$) is the regression coefficient of the *j*th explanatory variable; ε_i is the random error of the *i*th observation; and *n* is the number of observations. In our study, the explanatory variables are the geographical factors (as shown in Table 3), and the dependent variable *Y* is the positioning bias in data set *D1*.

The least squares method is generally used to estimate the coefficients of the explanatory variables in MLR models (Wang, Shangguan, Wu, & Guan, 2013). MLR analysis uses the fitted regression coefficient to analyze each explanatory variable's influence on the dependent variable. A non-standardized regression coefficient has dimensional influence. The influencing power of different explanatory variables on the dependent variable cannot

be compared with the non-standardized regression coefficient. Therefore, it is necessary to standardize the data to obtain standardized regression coefficients. In addition, attention should be paid to whether the regression coefficients are statistically significant.

3.3 | GeoDetector model

To further explore the patterns of influence of geographical factors on positioning uncertainty, especially the interactive influencing power among multiple factors, a GeoDetector model is employed in our study. The GeoDetector model is a statistical method for detecting the spatial variation in a geographical phenomenon and revealing the driving forces behind it (Song, Wang, Ge, & Xu, 2020; Wang et al., 2010).

The factor and interaction detectors are the core parts of GeoDetector. They reveal the influencing power of a single explanatory variable or the interactive influencing power of multiple explanatory variables with a *q*-statistic, which is computed as:

$$q = 1 - \frac{\sum_{h=1}^{l} N_h \sigma_h^2}{N\sigma^2} \tag{8}$$

where h is the spatial stratification of a single factor X (as shown in Figure 4a,b) or the crossed strata of multifactor X values (as shown in Figure 4c); N and N_h represent the number of units in the whole study area and subregion h, respectively; and σ^2 and σ_h^2 are the variances of the dependent variable Y in the entire study area and subregion h, respectively. In our GeoDetector model, the dependent variable Y is the positioning bias or the positioning dispersion in data set *D2*. The explanatory variables or factors are the geographical factors (as shown in Table 3). The value of q is between 0 and 1. The larger q is, the greater the influence of a single factor or interaction between different factors on the value of Y. For different factors $X1$ and $X2$, if $q(X1\cap X2) < \min(q(X1), q(X2))$, the interactive effect of factors $X1$ and $X2$ on Y is nonlinearly weakened; if $\min(q(X1), q(X2)) < q(X1\cap X2) < \max(q(X1), q(X2))$, the interactive effect of factors $X1$ and $X2$ on Y is nonlinearly weakened by a single factor; if $q(X1\cap X2) > \max(q(X1), q(X2))$, the interactive effect of factors $X1$ and $X2$ on Y is enhanced; if $q(X1\cap X2) > q(X1) + q(X2)$, the interactive effect of factors $X1$ and $X2$ on Y is nonlinearly enhanced; and if $q(X1\cap X2) = q(X1) + q(X2)$, the influences of factors $X1$ and $X2$ on Y are independent of each other. The ecological detector is also an important part of GeoDetector. It can be used to detect whether the two factors $X1$ and $X2$ have significant differences in terms of the influence on the spatial distribution of the dependent variable Y. GeoDetector software can be freely downloaded from http://www.geodetector.cn/?tdsourcetag=s_pctim_aiomsg.

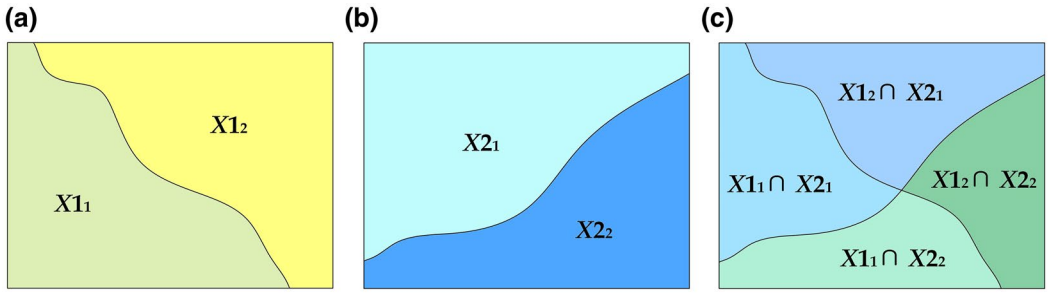

FIGURE 4 (a) Spatial stratification of a single factor $X1$; (b) spatial stratification of a single factor $X2$; and (c) crossed strata of factor $X1$ and factor $X2$ (i.e., $X1\cap X2$)

4 | ANALYSIS RESULTS

4.1 | The statistical distribution of the positioning bias

After data collection and preprocessing, we obtained 229,839 sample point pairs, each of which contained a *TP* and an *RP* (as shown in Figure 2). Equation (6) can then be used to calculate the positioning biases of the sample points *RP*. Unpaired MPL record points do not have corresponding reference points, and the positioning biases of these MPL record points cannot be measured. Therefore, these unpaired MPL record points were not considered. We calculated the positioning biases of 229,839 *RP*s. The statistical analysis results show that the minimum positioning bias in the sample set is 17.4 m, the maximum is 58,805.9 m, the mean is 450.3 m, and the *SD* is 771.19 m. Figure 5 shows the statistical distribution of the positioning biases in the sample set. Approximately half of the samples have a positioning bias less than 200 m, 77.7% have a positioning bias less than 500 m, and 95.9% have a positioning bias less than 1,500 m. Additionally, 0.5% of the samples have a significant positioning bias (i.e., greater than 5,000 m).

Theoretically, mobile phones are usually connected to the nearest cell towers. To better understand the positioning uncertainty in the MPL data, we further detect whether the *RP* of each sample point pair is connected to the nearest cell tower. We assign an attribute, *Rank*, to each sample *RP*, which indicates how close the *TP* is to the *RP* based on distance. For each sample point pair, we rank the cell towers around the reference point (*RP*) in order of distance from smallest to largest. The *Rank* is the serial number corresponding to the cell tower to which the mobile phone is currently connected. Three steps are applied to measure the *Rank* of each *RP*. We first determine the search range of the *RP*. Specifically, the search range is a circle with the *RP* as the center and the spatial distance (i.e., the positioning bias) between the *RP* and the *TP* as the radius. Then, we search for and count the number of cell towers in the circle area, denoted by *NC*. Finally, the *Rank* of the *RP* is calculated, where $Rank = NC + 1$.

As shown in Figure 6, 25.9% of the *RP*s are connected to the nearest cell towers, with a mean spatial distance of 197.6 m. Nearly half (49.85%) of the *RP*s' *Rank*s are less than or equal to 3, with a mean positioning bias of 219.9 m. The *Rank* of 89.9% of the *RP*s is no more than 30, and the mean positioning bias is 480.1 m; 6.1% of the *RP*s' *Rank*s are more than 50, and the mean positioning bias is more than 1,000 m; 3.76% of the *RP*s' *Rank*s are more than 100, and the mean positioning bias is 4,723.5 m. The above statistical results show that mobile phones

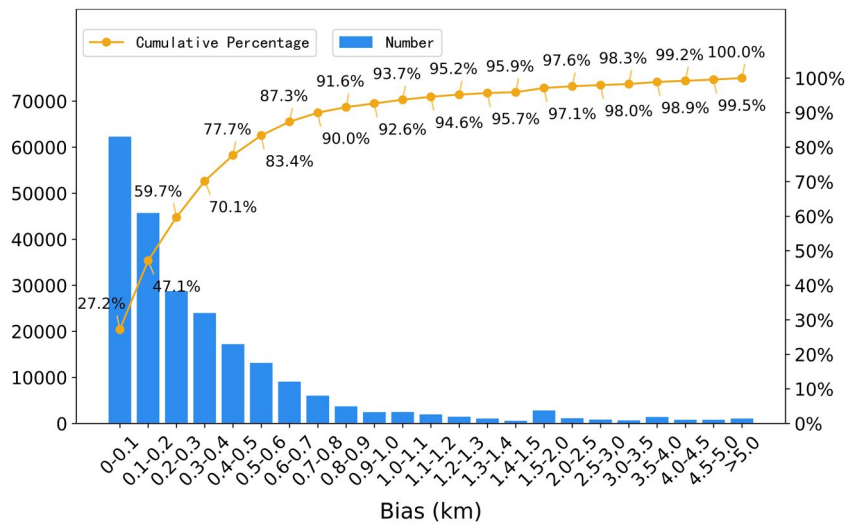

FIGURE 5 The statistical distribution of the positioning biases

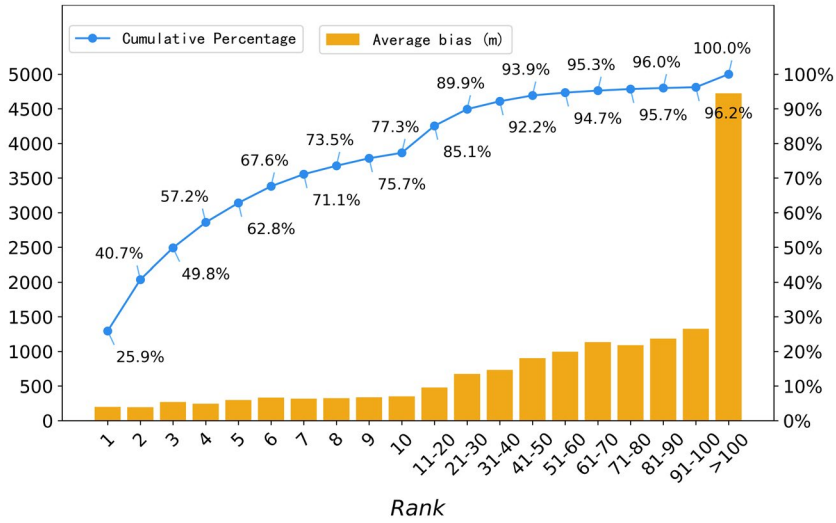

FIGURE 6 The statistical distribution of the attribute *Rank* value of the sample points

are not connected to the nearest cell towers in most cases. Even if 25.9% of the *RP*s in the overall samples are connected to the nearest cell towers, a mean positioning bias of approximately 200 m remains. These positioning biases are caused by the positioning system and cannot be eliminated.

4.2 | The spatial distribution of the positioning uncertainty

To further explore the spatial distribution trends of the MPL data positioning uncertainties, we perform a visual analysis of the spatial distribution of the 229,839 samples' positioning biases. In our study, the natural breaks classification method (Jenks, 1967) was used to grade the samples and conduct a visual analysis based on ArcGIS software.

Figure 7 shows the spatial distribution of the positioning biases of the sample set. As shown in Figure 7a, the map points are the *RP*s of the 229,839 sample point pairs, which are also GPS points. They offer the actual trajectories of volunteers on the ground. According to the values of the positioning biases, the 229,839 *RP*s are divided into 10 groups using the natural breaks classification method. The visualization results show that when the volunteers were walking around on the ground, the positioning bias of the MPL data continuously changed. When the volunteers moved inside the central area of Nanjing city (i.e., the area with dense buildings), the variation range of the positioning biases was small. Most of the positioning biases in the central area of Nanjing city were less than 500 m. However, when the volunteers travelled outside the urban area, the range of the positioning biases was large, and most of the positioning biases were greater than 1,500 m. The positioning uncertainty was noticeable in these areas. There were many *RP*s close to each other in space, but there was a large gap in the classification level of their positioning biases.

Figure 7b shows the spatial distribution trends in the positioning uncertainties in the MPL data more clearly. In this figure, the starting point of the line segment is the *RP* (i.e., the GPS point), and the end point is the corresponding *TP* (i.e., the cell tower) of the sample point pair. The length of the line segment represents the value of the positioning bias at the *RP*. From Figure 7b we can see that the positioning uncertainties are spatially aggregated. The areas with a high level of positioning bias (i.e., the yellow and red lines) often have high positioning uncertainties. On the one hand, this phenomenon is reflected in the large variation range of the positioning bias values in these areas. On the other hand, it is reflected in the bias lines diverging in all directions, which is a manifestation of the cell tower oscillation, also known as the ping-pong effect.

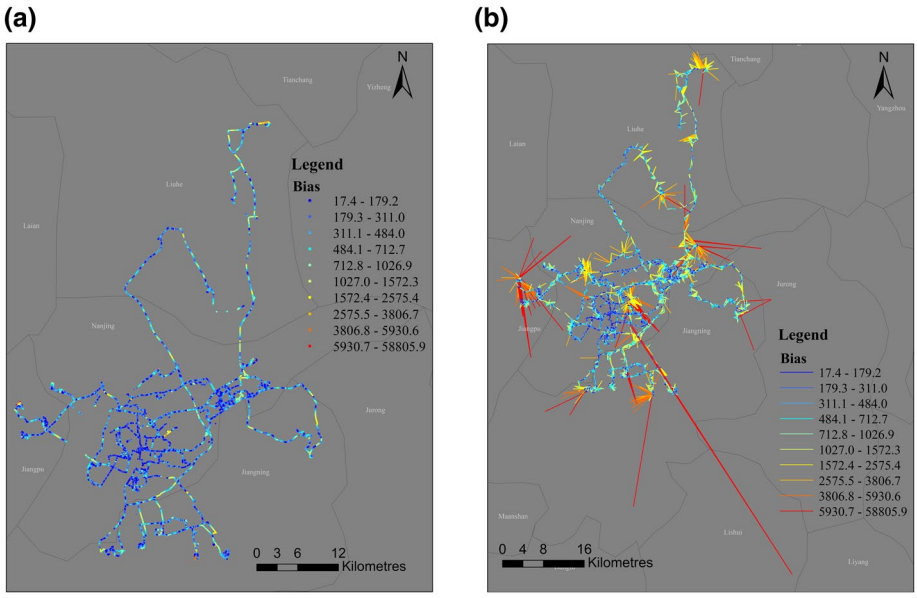

FIGURE 7 The spatial distribution of the positioning biases is shown by: (a) dots; and (b) lines

TABLE 4 Statistical results of samples in different clusters

Cluster	Minimum bias (m)	Maximum bias (m)	Average bias (m)	*SD* of biases (m)	Count	Percentage
Samples in microcell tower regions	17.4	8,281.1	321.9	433.7	129,756	56.5
Samples in macrocell tower regions	30.0	58,805.8	614.0	1,034.6	100,083	43.5

According to the type attribute of the cell towers, we classified the sample points that fall within the coverage of the microcell towers into one cluster and the sample points that fall within the coverage of the macrocell towers into another cluster. Table 4 describes the statistical results of samples in different clusters. Figure 8 shows the spatial distribution of the clustering results. Samples in microcell tower regions account for 56.5% of the total samples, and the mean positioning bias of this cluster is 321.9 m (minimum 17.4 m, maximum 8,281.1 m, and *SD* 433.7 m). Figure 8a shows the spatial distribution of samples in microcell tower regions, which are mainly distributed in areas with dense buildings. Samples in macrocell tower regions account for 43.5% of the total samples, and the mean positioning bias of this cluster is 614.0 m (minimum 30.0 m, maximum 58,805.8 m, and *SD* 1,034.6 m). Figure 8b shows the spatial distribution of samples in macrocell tower regions, which are mainly distributed in open areas (e.g., mountains, lakes, rivers and open squares). Compared with the samples in microcell tower regions, the positioning uncertainty of the samples in macrocell tower regions is more obvious.

4.3 | Patterns of influence

4.3.1 | MLR modeling results

As stated above, data set *D1* (229,839 sample points) was used for MLR analysis. The explanatory variables are the 12 factors in Table 3. Given the possible strong correlation between some variables, we used SPSS software

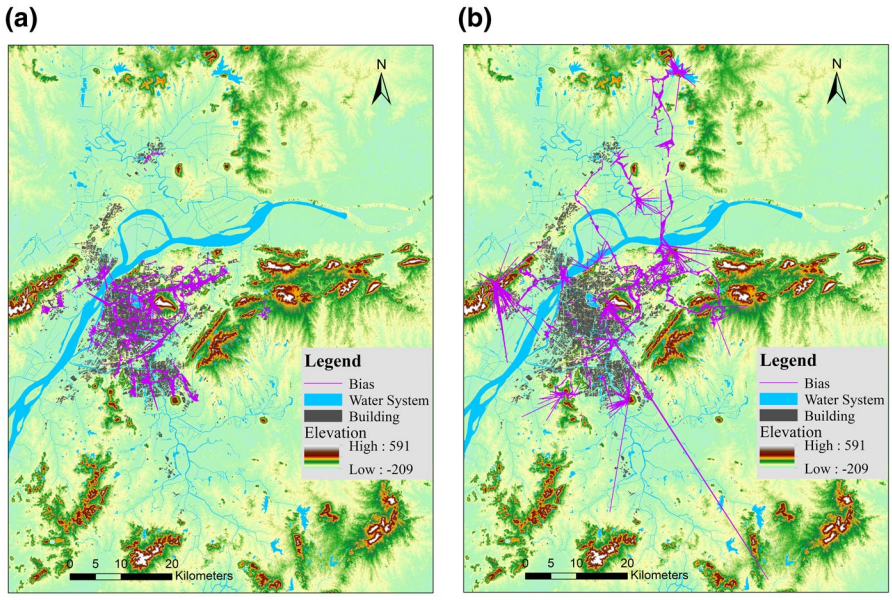

FIGURE 8 The spatial distribution of the positioning biases of samples in: (a) microcell tower regions; and (b) macrocell tower regions

for multiple stepwise regression analysis to eliminate the influence of multicollinearity. According to the fitting effect, the optimal model is selected to explore the influence of the explanatory variables on the spatial distribution of the positioning bias. We build MLR models based on all the samples and the samples of different clusters. The regression results are shown in Table 5. From the goodness of fit R^2 values of the three models, it can be seen that the R^2 values of the three models are all less than 0.3. The fitting effect of the model based on the samples in macrocell tower regions is best, while the fitting effect of the model based on the samples in microcell tower regions is worst.

From the regression coefficients and significance of the model fitted with all the samples, the spatial distribution of the positioning bias of MPL data is greatly affected by the DNCT, elevation, and NDVI, all showing a significant positive impact. The building height, DNWB, and DNV had no obvious influence on the spatial distribution of the positioning bias. However, the samples' positioning biases of different clusters have certain differences in the influencing factors. Specifically, the positioning biases of samples in microcell tower regions are most affected by the NDWI, followed by the elevation and DNCT, showing a significant positive effect. In macrocell tower regions, the positioning biases are most affected by the DNCT, followed by the elevation and NDVI, showing a significant positive effect.

4.3.2 | GeoDetector statistical results

As mentioned above, data set *D2* (7,258 sample grids) is used in the GeoDetector model to probe the influence mechanism of the positioning uncertainty. According to the equal division method, all the factors consisting of real-valued data in *D2* were discretized into ten categories (ordinal levels). The discrete values of 12 geographical factors and the mean value of the positioning biases are used to explore the influence patterns in the positioning bias spatial distribution. The discrete values of 12 geographical factors and the *SD* of the positioning biases are used to explore the influence patterns in the positioning dispersion spatial distribution.

TABLE 5 MLR results

Regression term	All samples	Samples in microcell tower regions	Samples in macrocell tower regions
Elevation	0.219***	0.181***	0.194***
Slope			0.077***
Aspect	0.039***		0.035***
DNV	0.019***	0.030***	0.008**
DNWB	−0.016***	−0.104***	0.013***
NDWI	0.211***	0.298***	0.071***
NDVI	0.212***	0.157***	0.133***
Building density	0.060***	−0.046***	0.039***
Building height	0.082***	−0.015***	
Road density	0.073***	−0.019***	0.105***
Population density	−0.050***	−0.041***	−0.080***
DNCT	0.276***	0.166***	0.288***
Adjusted R^2	.151	.077	.225
F statistic	3,704.549	1,089.570	1,731.116

Note: The values of the first 12 rows are normalized regression coefficients.

***Significance at .01 level; **significance at .05 level.

The influence patterns in the positioning bias

Figure 9 shows the results of GeoDetector analysis for MPL data positioning bias in the whole region, microcell tower regions, and macrocell tower regions. The results include three parts of geographical detectors: factor detector, ecological detector, and interaction detector. The geographical detector results show that primary explanatory variables and interactive variables vary throughout the whole region and subregions. In the whole region and the macrocell tower regions (Figure 9a,c), elevation and DNCT variables and their interaction are major contributors to the positioning bias. In the whole region (Figure 9a), the physical geographical factor elevation (15.8%) contributes most to the positioning bias, followed by the human geographical factor DNCT (12.3%). The interaction between elevation and DNCT has the highest association with positioning bias (46.7%). Similarly, elevation (21.1%) is the primary single explanatory variable in the macrocell tower regions (Figure 9c), and the interaction between elevation and DNCT is the major interactive variable (68.8%) of the positioning bias. Although the major single explanatory variable in the microcell tower regions (Figure 9b) remains elevation, the explanatory power is only 6%. The results of ecological detector demonstrate that there is no significant difference in the impacts of various factors. The maximum interactive influencing power reaches 57%, which is derived from the interaction between the physical geographical factor NDWI (2.5%) and the human geographical factor road density (4.3%). Analogously to the MLR results in the whole region and the macrocell tower regions, elevation and DNCT are important influencing factors, while the DNV, DNWB, and aspect have no obvious influence on the positioning bias. However, the major contributors measured by the factor detector (elevation and population density) in microcell tower regions are inconsistent with the MLR results (NDWI and elevation).

The influence patterns in the positioning dispersion

As mentioned in Section 3.1.2, the *SD* of the positioning biases associated with one grid can be used to measure the positioning dispersion of this spatial unit. We also use the GeoDetector model to explore the influence patterns in the geographical factors on the positioning dispersion. Figure 10 shows the results of GeoDetector analysis for MPL data positioning dispersion in the whole region, microcell tower regions and macrocell tower

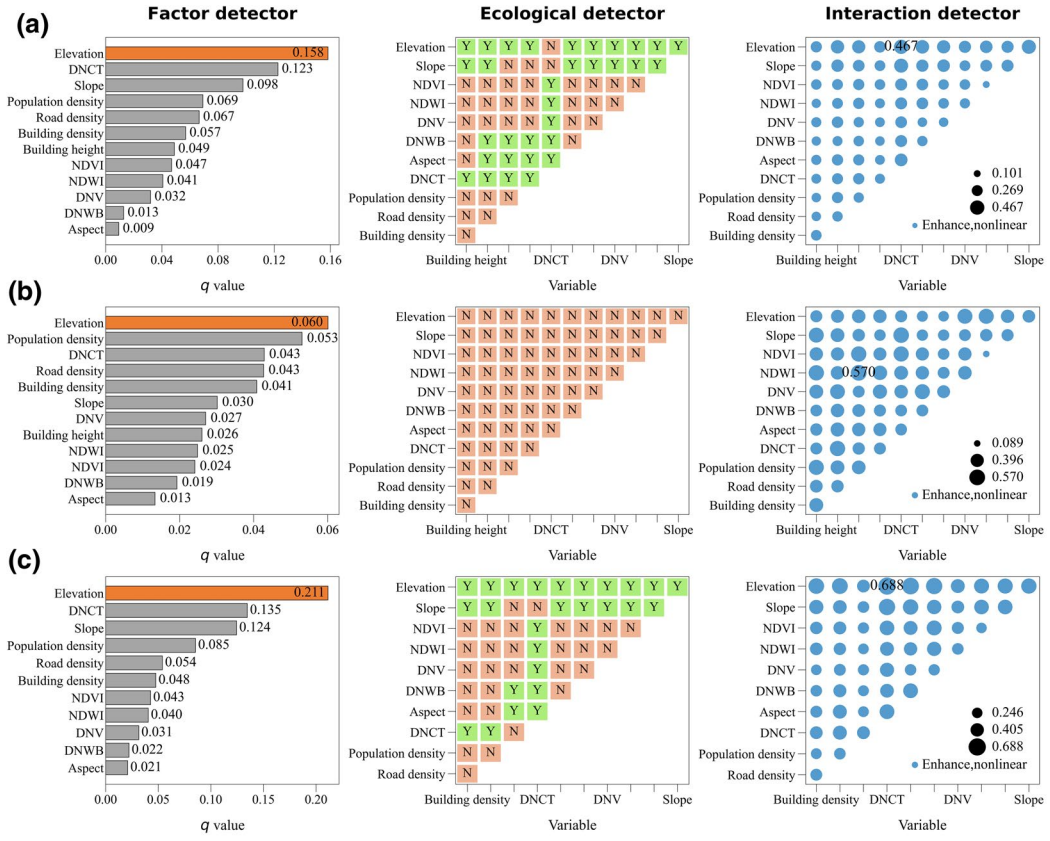

FIGURE 9 Results of GeoDetector analysis of the positioning bias of the MPL data in: (a) the whole region; (b) microcell tower regions; and (c) macrocell tower regions

regions. In the whole region (Figure 10a), the physical geographical factor elevation (6.5%) contributes most to the positioning dispersion compared with other factors. The interaction between elevation and slope has the highest contribution (54.5%), which is nonlinearly enhanced by the single variables. Population density (3.8%) is the primary single explanatory factor in the microcell tower regions (Figure 10b). Similarly to the interaction detector results of positioning bias (Figure 9b), the interaction between NDWI and road density is the primary interactive variable (74.6%) of positioning dispersion. In the macrocell tower regions (Figure 10c), elevation (22.2%) is the major contributor to positioning dispersion, and the interaction between elevation and aspect is the major interactive variable (67.8%) of positioning dispersion.

The interaction detector results for MPL data positioning bias (Figure 9) and positioning dispersion (Figure 10) of the whole region and all subregions suggest that the influencing power of any two factors is nonlinearly enhanced after an interaction. Furthermore, the factor's explanatory power with the weakest influence is also significantly nonlinearly enhanced after interacting with the primary influencing factor. For example, the interactive influencing power of elevation and aspect in the whole region reaches 0.417 (Figure 9a), which is significantly greater than the sum of the explanatory power of elevation ($q = 0.158$) and aspect ($q = 0.009$). Figures 9 and 10 show that elevation is a particularly important factor affecting the spatial distribution of MPL data positioning uncertainty. Combined with related work in the communication engineering field, elevation directly affects the site selection of cell towers (Xie et al., 2016); additionally, elevation determines radio wave propagation loss

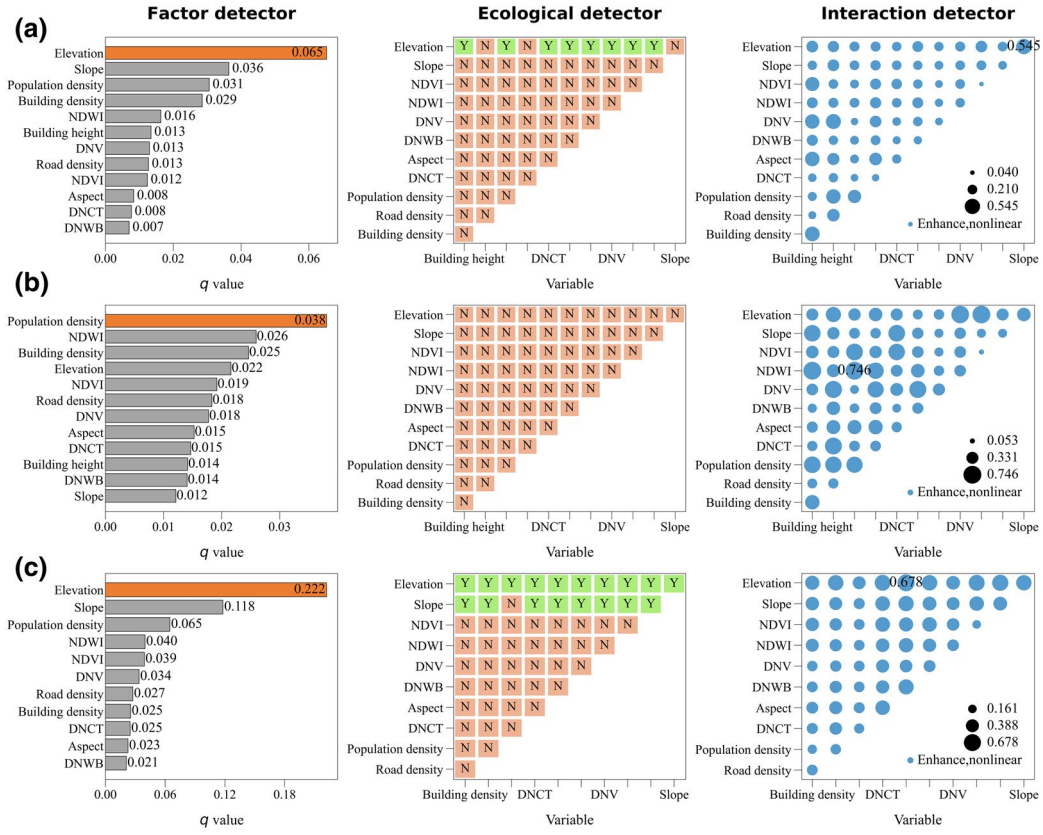

FIGURE 10 Results of GeoDetector analysis of the positioning dispersion of the MPL data in: (a) the whole region; (b) microcell tower regions; and (c) macrocell tower regions

(Ma & Wang, 2012). Either the site selection of cell towers or the radio wave propagation loss directly affects the coverage of cell towers. Therefore, elevation has a significant impact on the positioning uncertainty of MPL data.

5 | DISCUSSION AND CONCLUSIONS

The spatial accuracy of the uncertainty in MPL data is an important but often ignored issue in big spatiotemporal data analytics. This issue has been and will always be a challenge to the validity of research involving MPL data. This issue is directly related to the reliability of human mobility research results and should be carefully examined. In this research we investigate the spatial distribution of MPL data positioning uncertainties and the influence of geographical determinants on the spatial distribution of MPL data positioning uncertainties.

By using high-frequency and high-positioning-accuracy GPS data as the "ground truth," we evaluate the positioning bias of MPL data. We find that the higher the positioning biases of samples are, the lower the proportion of samples. In our study, approximately half of the total samples have a positioning bias less than 200 m, 95.9% have a positioning bias less than 1,500 m, and 0.5% of the samples have a very significant positioning bias greater than 5,000 m. In addition, we find that mobile phones are not connected to the nearest cell towers in most cases. In our study, even if 25.9% of the *RP*s in the total samples are connected to the nearest cell towers, a mean positioning bias of approximately 200 m remains.

We then examine how the positioning uncertainties of MPL data are distributed and vary across a surface. By conducting visual analysis, we find that when the volunteers moved to the central area of Nanjing city (microcell tower regions), the variation range of the positioning biases is small, and most of the positioning biases in the central area of Nanjing city are less than 500 m. However, when the volunteers traveled outside the urban area (macrocell tower regions), the range of the positioning biases is large, and most of the positioning biases are larger than 1,500 m. In addition, there is a spatial aggregation phenomenon in the positioning uncertainties, and the areas with a high level of positioning bias often have high positioning uncertainties. In these areas, the variation range of positioning bias values is large, and there are many cell tower oscillations. Compared with the samples in microcell tower regions, the positioning uncertainty of the samples in macrocell tower regions is more obvious.

To better explore the geographical determinants and influence mechanism on the spatial distribution of MPL data positioning uncertainty, MLR and GeoDetector models are applied in our study. The MLR results of the total samples show that the positioning biases are greatly affected by the DNCT, elevation, and NDVI. The same results are found in the macrocell tower regions. The positioning biases of samples in microcell tower regions are most affected by the NDWI, followed by the elevation and DNCT. From the GeoDetector statistical results, we find that elevation and DNCT variables and their interaction are major contributors to the positioning bias in the whole region and the macrocell tower regions. In the microcell tower regions, elevation is the major single explanatory variable of the positioning bias, and the interaction between NDWI and road density has the highest contribution. Similarly, elevation is the primary single explanatory factor of the positioning dispersion in the whole region and macrocell tower regions. The primary interactive variables are the interaction between elevation and slope and the interaction between elevation and aspect, respectively. Population density is the primary single explanatory factor of the positioning dispersion in the microcell tower regions, and the interaction between NDWI and road density is the primary interactive variable. Overall, elevation is a particularly important factor affecting the spatial distribution of MPL data positioning uncertainty. The influencing power of any two geographical factors on positioning bias or positioning dispersion is nonlinearly enhanced after an interaction.

The implications are manifold. First, the discovery of the spatial distribution trends of MPL data positioning uncertainty is helpful in evaluating the availability and applicability of MPL data in human mobility research. When using MPL data to study human mobility, we need to comprehensively consider the land-use type of the study area. The results (Figures 7 and 8) of the spatial distribution trends show that positioning uncertainty issues are very serious in mountainous and water body areas. MPL data are not very appropriate for studying human mobility in these areas. Identifying stop points or eliminating oscillations in these areas is a challenge and should be considered in future research. Second, the exploration of the influence mechanism of geographical factors on MPL data uncertainty is the basis for the effective prediction and simulation of MPL data positioning accuracy. There have recently been studies on simulating the spatial distribution of MPL data positioning bias based on some geographical elements (Song, Long, et al., 2020). This is helpful in analyzing the uncertainty of human mobility results based on MPL data. The influence mechanism results of this study provide support for the main influencing factors predicting MPL data positioning accuracy.

We want to mention a few limitations of this research. First, the GPS record points used as reference data are not real "ground truth," but rather a proxy. Therefore, what the study is really measuring is "relative difference" rather than actual positioning bias. To ensure that the relative difference measured in this study is as close as possible to the actual positioning bias, we filtered the GPS data based on the accuracy attribute ε of the data. High-precision filtering methods can be used to improve the positioning accuracy of GPS data and further improve the measurement accuracy of the MPL data positioning bias in the future. Moreover, the validity of the findings on the spatial trends of the MPL data positioning uncertainty is somewhat limited. The operator limited the number of subscribers who could apply for MPL data and the amount of data. In this study we performed short-term data collection according to the limited planned routes (covering different geographical contexts). Therefore, the spatial distribution of the sample point pairs might be under- or over-represented in certain places. Large-scale and long-term data acquisition efforts are expected to address this limitation. In

addition, if long-term and high-frequency data acquisition is performed, it is possible to detect the temporal distribution law of the MPL data's uncertainty. In the future, we will strive for large-scale, long-term and high-frequency MPL data collection to further analyze the temporal and spatiotemporal distribution trends of MPL data uncertainty and the influencing mechanism of geographical factors on these distributions in depth. We intend to take some equipment factors, such as the sector azimuth and antenna height of a cell tower, into account. We will also further evaluate the positioning uncertainties in MPL data sets from different mobile operators and across different study areas considering these determinants. We believe that these analyses will deepen our understanding of the veracity of MPL data used in human mobility research and provide many benefits to better use MPL data in future studies.

ACKNOWLEDGMENTS

This research is financially supported by the National Key Research and Development Program of China Grant (2017YFB0503500), State Key Program of National Natural Science Foundation of China Grant (41930104), National Natural Science Foundation of China (42101419), Natural Science Foundation of Jiangsu Province (BK20191183) and Natural Science Foundation of Anhui Province (2008085QD166).

CONFLICT OF INTEREST

The authors declare no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

DATA AVAILABILITY STATEMENT

Due to the nature of this research, participants of this study did not agree for their data to be shared publicly, so supporting data are not available.

ORCID

Ling Zhang https://orcid.org/0000-0002-3228-8374
Yi Long https://orcid.org/0000-0002-4207-5516

REFERENCES

Adusei, I. K., Kyamakya, K., & Jobmann, K. (2002). Mobile positioning technologies in cellular networks: An evaluation of their performance metrics. In *Proceedings of the Military Communications Conference*, Anaheim, CA (pp. 1239–1244). Piscataway, NJ: IEEE.

Akinbile, C. O., Ogunmola, O. O., Abolude, A. T., & Akande, S. O. (2020). Trends and spatial analysis of temperature and rainfall patterns on rice yields in Nigeria. *Atmospheric Science Letters*, *21*(3), e944. https://doi.org/10.1002/asl.944

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, *350*(6264), 1073–1076. https://doi.org/10.1126/science.aac4420

Celidonio, M., Fionda, E., Vaser, M., & Restuccia, E. (2018). NLOS mm wave propagation measurements through vegetation in urban area: A case study. In *Proceedings of the 2018 AEIT International Annual Conference*, Bari, Italy (pp. 1–6). Piscataway, NJ: IEEE.

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies*, *68*, 285–299. https://doi.org/10.1016/j.trc.2016.04.005

Edwards, R., & Durkin, J. (1969). Computer prediction of service area for VHF mobile radio networks. *Proceedings of the Institution of Electrical Engineers*, *116*, 1493–1500. https://doi.org/10.1049/piee.1969.0270

Fund, F., Lin, R., Korakis, T., & Panwar, S. S. (2016). How bad is the flat earth assumption? Effect of topography on wireless systems. In *Proceedings of the 14th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, Tempe, AZ (pp. 1–5). Piscataway, NJ: IEEE.

Harinda, E., Hosseinzadeh, S., Larijani, H., & Gibson, R. M. (2019). Comparative performance analysis of empirical propagation models for LoRaWAN 868MHz in an urban scenario. In *Proceedings of the Fifth IEEE World Forum on Internet of Things*, Limerick, Ireland (pp. 154–159). Piscataway, NJ: IEEE.

Hata, M. (1980). Empirical formula for propagation loss in land mobile radio services. *IEEE Transactions on Vehicular Technology*, *29*(3), 317–325. https://doi.org/10.1109/T-VT.1980.23859

Horn, C., Klampfl, S., Cik, M., & Reiter, T. (2014). Detecting outliers in cell phone data. *Transportation Research Record*, *2405*, 49–56. https://doi.org/10.3141/2405-07

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

Jiang, S., Yang, Y., Gupta, S., Veneziano, D., Athavale, S., & González, M. C. (2016). The TimeGeo modeling framework for urban motility without travel surveys. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(37), E5370–E5378. https://doi.org/10.1073/pnas.1524261113

Kamenjuk, P., Aasa, A., & Sellin, J. (2017). Mapping changes of residence with passive mobile positioning data: The case of Estonia. *International Journal of Geographical Information Science*, *31*(7), 1425–1447. https://doi.org/10.1080/13658816.2017.1295308

Kos, T., Grgic, M., & Sisul, G. (2006). Mobile user positioning in GSM/UMTS cellular networks. In *Proceedings of the International Symposium on Electronics in Marine*, Zadar, Croatia (pp. 185–188). Piscataway, NJ: IEEE.

Kwan, M.-P. (2016). Algorithmic geographies: Big data, algorithmic uncertainty, and the production of geographic knowledge. *Annals of the American Association of Geographers*, *106*(2), 274–282. https://doi.org/10.1080/00045608.2015.1117937

Liu, B., & Pang, L. (2019). Review of domestic and international research on big data quality. *Journal of the China Society for Scientific and Technical Information*, *38*(2), 217–226. https://doi.org/10.3772/j.issn.1000-0135.2019.02.011

Liu, D., Xu, Y., & Huang, X. (2018). Identification of location spoofing in wireless sensor networks in non-line-of-sight conditions. *IEEE Transactions on Industrial Informatics*, *14*(6), 2375–2384. https://doi.org/10.1109/TII.2017.2767631

Liu, L., Gao, X., Zhuang, J., Wu, W., Yang, B. O., Cheng, W., … Deng, O. (2020). Evaluating the lifestyle impact of China's rural housing land consolidation with locational big data: A study of Chengdu. *Land Use Policy*, *96*, *9*. https://doi.org/10.1016/j.landusepol.2020.104623

Ma, Y., & Wang, Y. (2012). Study of characteristic prediction of radio wave propagation loss on complex irregular terrain of wide-range distance. In *Proceedings of the Sixth Asia-Pacific Conference on Environmental Electromagnetics*, Shanghai, China (pp. 67–71). Piscataway, NJ: IEEE.

Mao, L., Yin, L., Song, X., & Mei, S. (2016). Mapping intra-urban transmission risk of dengue fever with big hourly cell-phone data. *Acta Tropica*, *162*, 188–195. https://doi.org/10.1016/j.actatropica.2016.06.029

Ogulenko, A., Benenson, I., Omer, I., & Alon, B. (2020). Bayesian estimate of position in mobile phone network. In *Proceedings of the Mobile Tartu International Virtual Conference* (pp. 1–20).

Okumura, T., & Ohronofi, E. (1968). Field strength and its variable in VHF and UHF land mobile service. *Review of the Electrical Communication Laboratory*, *16*, 825–873.

Omaki, N., Imai, T., Kitao, K., & Okumura, Y. (2016). Improvement of ray tracing in urban street cell environment of non-line-of-sight (NLOS) with consideration of building corner and its surface roughness. In *Proceedings of the 10th European Conference on Antennas and Propagation*, Davos, Switzerland (pp. 1–5). Piscataway, NJ: IEEE.

Paul, B. S., & Rimer, S. (2013). Wireless sensor node placement due to power loss effects from surrounding vegetation. In K. Singh & A. K. Awasthi (Eds.), *Quality, reliability, security and robustness in heterogeneous networks: QShine 2013* (Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Vol. *115*, pp. 915–927). Berlin, Germany: Springer.

Pospíšilová, L., & Novák, J. (2016). Mobile phone location data: New challenges for geodemographic research. *Demografie*, *58*(4), 320–337. https://doi.org/10.1080/17445647.2019.1709577

Ruiz-Álvarez, M., Alonso-Sarria, F., & Gomariz-Castillo, F. (2019). Interpolation of instantaneous air temperature using geographical and MODIS derived variables with Machine Learning Techniques. *ISPRS International Journal of Geo-Information*, *8*(9), 382. https://doi.org/10.3390/ijgi8090382

Shi, W. (2015). *Principles of modeling uncertainties in spatial data and spatial analyses*. Beijing, China: Science Press.

Singh, Y. (2012). Comparison of Okumura, Hata and COST-231 models on the basis of path loss and signal strength. *International Journal of Computer Applications*, *59*(11), 37–41. https://doi.org/10.5120/9594-4216

Song, X., Long, Y., Zhang, L., Rossiter, D. G., Liu, F., & Jiang, W. (2020). Spatial accuracy evaluation for mobile phone location data with consideration of geographical context. *IEEE Access*, *8*, 221176–221190. https://doi.org/10.1109/ACCESS.2020.3043317

Song, Y., Wang, J., Ge, Y., & Xu, C. (2020). An optimal parameters-based geographical detector model enhances geographic characteristics of explanatory variables for spatial heterogeneity analysis: Cases with different types of spatial data. *Giscience & Remote Sensing*, *57*(5), 593–610. https://doi.org/10.1080/15481603.2020.1760434

Tang, S., Dong, Y., & Zhang, X. (2013). On path loss of NLOS underwater wireless optical communication links. In *Proceedings of the 2013 MTS/IEEE Oceans Conference*, Bergen, Norway (pp. 1–3). Piscataway, NJ: IEEE.

Trevisani, E., & Vitaletti, A. (2004). Cell-ID location technique, limits and benefits: An experimental study. In *Proceedings of the Sixth IEEE Workshop on Mobile Computing Systems and Applications*, Windermere, UK (pp. 1–10). Piscataway, NJ: IEEE.

Wang, H., Shangguan, L., Wu, J., & Guan, R. (2013). Multiple linear regression modeling for compositional data. *Neurocomputing*, *122*, 490–500. https://doi.org/10.1016/j.neucom.2013.05.025

Wang, J. (2017). Wireless base station location technology. *China Computer & Communication*, *16*, 175–177. https://doi.org/10.3969/j.issn.1003-9767.2017.16.066

Wang, J. F., Li, X. H., Christakos, G., Liao, Y. L., Zhang, T., Gu, X., & Zheng, X. Y. (2010). Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun Region, China. *International Journal of Geographical Information Science*, *24*(1), 107–127. https://doi.org/10.1080/13658810802443457

Wang, Z., Wang, S., & Lian, H. (2020). A route-planning method for long-distance commuter express bus service based on OD estimation from mobile phone location data: The case of the Changping Corridor in Beijing. *Public Transport*, *13*(1), 101–125. https://doi.org/10.1007/s12469-020-00254-w

Widhalm, P., Yang, Y., Ulm, M., Athavale, S., & González, M. C. (2015). Discovering urban activity patterns in cell phone data. *Transportation*, *42*(4), 597–623. https://doi.org/10.1007/s11116-015-9598-x

Xie, Q., Liu, X., & Yan, X. (2016). Base station location optimization based on the Google Earth and ACIS. In *Proceedings of the International Conference on Human Centered Computing*, Colombo, Sri Lanka (pp. 487–496). New York, NY: ACM.

Xu, Y., Li, X., Shaw, S.-L., Lu, F., Yin, L., & Chen, B. Y. (2020). Effects of data preprocessing methods on addressing location uncertainty in mobile signaling data. *Annals of the American Association of Geographers*, *111*(2), 515–539. https://doi.org/10.1080/24694452.2020.1773232

Yassin, M., & Rachid, E. (2015). A survey of positioning techniques and location based services in wireless networks. In *Proceedings of the IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems*, Kozhikode, India (pp. 1–5). Piscataway, NJ: IEEE.

Yin, L., Jiang, R., Zhao, Z., Song, X., & Li, X. (2017). Exploring the bias of estimating 24-hour population distributions using call detail records. *Journal of Geo-information Science*, *19*(6), 763–771. https://doi.org/10.3969/j.issn.1560-8999.2017.06.005

Yin, L., Lin, N., Song, X., Mei, S., Shaw, S.-L., Fang, Z., … Mao, L. (2020). Space-time personalized short message service (SMS) for infectious disease control: Policies for precise public health. *Applied Geography*, *114*, 102103. https://doi.org/10.1016/j.apgeog.2019.102103

Yu, Q., Li, W., Yang, D., & Zhang, H. (2020). Mobile phone data in urban commuting: A network community detection-based framework to unveil the spatial structure of commuting demand. *Journal of Advanced Transportation*, *2020*, 8835981. https://doi.org/10.1155/2020/8835981

Zhao, Z., Shaw, S.-L., Xu, Y., Lu, F., Chen, J., & Yin, L. (2016). Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science*, *30*(9), 1738–1762. https://doi.org/10.1080/13658816.2015.1137298

Zhao, Z., Shaw, S.-L., Yin, L., Fang, Z., Yang, X., Zhang, F., & Wu, S. (2019). The effect of temporal sampling intervals on typical human mobility indicators obtained from mobile phone location data. *International Journal of Geographical Information Science*, *33*(7), 1471–1495. https://doi.org/10.1080/13658816.2019.1584805

Zhao, Z., Yin, L., Shaw, S.-L., Fang, Z., Yang, X., & Zhang, F. (2018). Identifying stops from mobile phone location data by introducing uncertain segments. *Transactions in GIS*, *22*(4), 958–974. https://doi.org/10.1111/tgis.12332

Zhong, R., & Xiao, K. (2008). Coverage optimization of GSM wireless network. *Information, Communication and Society*, *21*(1), 57–59. https://doi.org/10.3969/j.issn.1673-1131.2008.01.019

How to cite this article: Song, X., Zhang, L., Wang, S., Long, Y., Jiang, W., & Hao, Q. (2021). Influence of geographical determinants on the spatial distribution of positioning uncertainties in mobile phone location data. *Transactions in GIS*, 00, 1–22. https://doi.org/10.1111/tgis.12860