sustainability

MDPI

Article

Exploring the Effects of Sampling Locations for Calibrating the Huff Model Using Mobile Phone Location Data

Shiwei Lu [1,*], Shih-Lung Shaw [1,2,3], Zhixiang Fang [1,3], Xirui Zhang [4] and Ling Yin [5]

1 State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 129 Luoyu Road, Wuhan 430079, China; sshaw@utk.edu (S.-L.S.); zxfang@whu.edu.cn (Z.F.)
2 Department of Geography, University of Tennessee, Knoxville, TN 37996, USA
3 Collaborative Innovation Center of Geospatial Technology, 129 Luoyu Road, Wuhan 430079, China
4 Information Center of Urban Planning, Land & Real Estate of Shenzhen Municipality, 8007 Hongli West Road, Shenzhen 518040, China; xrzhangchn@gmail.com
5 Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, 1068 Xueyuan Road, Shenzhen 518005, China; yinling@siat.ac.cn
* Correspondence: lusw@whu.edu.cn; Tel.: +86-27-6877-9889

Academic Editor: Marc A. Rosen
Received: 18 November 2016; Accepted: 16 January 2017; Published: 22 January 2017

Abstract: The introduction of the Huff model is of critical significance in many fields, including urban transport, optimal location planning, economics and business analysis. Moreover, parameters calibration is a crucial procedure before using the model. Previous studies have paid much attention to calibrating the spatial interaction model for human mobility research. However, are whole sampling locations always the better solution for model calibration? We use active tracking data of over 16 million cell phones in Shenzhen, a metropolitan city in China, to evaluate the calibration accuracy of Huff model. Specifically, we choose five business areas in this city as destinations and then randomly select a fixed number of cell phone towers to calibrate the parameters in this spatial interaction model. We vary the selected number of cell phone towers by multipliers of 30 until we reach the total number of towers with flows to the five destinations. We apply the least square methods for model calibration. The distribution of the final sum of squared error between the observed flows and the estimated flows indicates that whole sampling locations are not always better for the outcomes of this spatial interaction model. Instead, fewer sampling locations with higher volume of trips could improve the calibration results. Finally, we discuss implications of this finding and suggest an approach to address the high-accuracy model calibration solution.

Keywords: big data; mobile phone location data; spatial interaction model; human dynamics

1. Introduction

The introduction of the Huff model is of critical significance in urban transport, economics and business areas, which can help us understand the accessibility, business opportunities, source and distribution of customers, and give suggestions to the optimal location planning of new trading areas [1–5]. Besides, there are already many methods to analyze trading areas, such as the Ring model [6], regression model [7,8], the analog model [9], the Huff model, and so on. Among all these methods, the Huff model is a quantitative method widely used to explore the interactions in urban environment [10,11]. Before applying the Huff model, the high-accuracy calibration of the model is a crucial procedure to apply [12]. Previous studies have paid much attention to the calibration of the model [13–15]. Traditionally, the calibration of the spatial interaction model in human

mobility research is dependent on survey or questionnaire, which has a few disadvantages, such as being labor-intensive, time-consuming, and error-prone; usually having a poor response rate [16,17]; and, sometimes, lacking proper sampling mechanisms. Improper sampling methods may lead to some biases or non-representative issues [18–22], thus may influence the calibration of spatial interaction model [23,24]. Besides, the number of sampling locations is also one of the concerns in many studies. For example, Zhou et al. [25] investigates how many samples are needed for a good performance of road selection and finds that only a small number (e.g., 50–100) of training samples is needed, while Zhao et al. [26] indicates that sparse sampled call detail records data introduce some biases to human mobility research. Thus, one of the main tasks of this paper is to investigate the effects of different numbers of sampling locations for calibrating the Huff model.

Recently, researchers prefer to use larger multi-source datasets to find their better solution. Fortunately, the advent of information and communication technology (ICT) aids the acquisition of human trajectory data by lowering the cost of collecting, storing, processing, and sharing data and information [27,28]. Large volume data (such as GPS tracking data, mobile phone location data, social media check-in data, and so on), give new insights and a better understanding of human mobility and behaviors [29,30]; community detection [31–33]; urban activity space and dynamics [34,35]; and spatial interaction and modeling [36,37]. Regarding the calibration of spatial interaction model, most use all sampling locations to calibrate. For example, Yue et al. [38] and Markham et al. [39] use the whole datasets to calibrate spatial interaction models, but whether a small part of the datasets contributes to more accurate calibration results remains unresolved.

Besides, among all the big geodata, mobile phone location data are very special data because mobile phones have an extremely high penetration rate, which can be over 94% in Asian countries such as China [40], and people usually take their cell phone with them. Thus, some researchers view this type of data as a reasonable source to describe human mobility and model spatial interactions [29,31,33–35,37], and many valuable findings regarding human dynamics in the era of big data have been explored from this kind of data. For example, Gao et al. [31] propose an alternative modularity function which incorporates a calibrated gravity model to discover the clustering structures of spatial-interaction communities generated by massive mobile phone users. Liang et al. [41] analyze the collective intra-urban mobility using a modified spatial interaction model, and Simini et al. [42] propose a radiation model which predicts mobility patterns in good agreement with observed data when compared with the calibrated gravity model by using different data sources including mobile phone data. Whether the models were well calibrated to "best fit" the observed data needs to be answered before comparison and application. If so, another question is how to derive the more valuable sampling locations to get high-accuracy calibration results. Additionally, Vij et al. [43] exhibit a neutral attitude towards the volume of big data, pointing out that high quality but small volume data may be better than big data, and small volume datasets represent not only dimension reduction but also noise elimination from big data.

The calibration methodology has been widely discussed [44–46], and is not the main focus of this paper. However, we investigate the effects of sampling locations by calibrating a spatial interaction model as a case study, using mobile phone location data from the big data era, and attempt to answer the following questions:

(1) Does using all sampling locations always perform better than small volume of sampling locations to calibrate the Huff model?
(2) If not, what kinds of sampling locations are more effective for calibrating this model?

There are several contributions of this study. Firstly, the results of this paper show that small volume of sampling location dataset may perform more effective for the calibration of the Huff model than large sampling locations, which could help utilize big data better for human mobility modeling; Secondly, we propose a method to select the more effective locations from massive mobile phone towers to improve the model calibration, which could be used to guide surveys or questionnaire for

trading area analysis in real scenes. This method could save both survey time and expense in many related areas of research while achieving high-quality model calibration results. To the best of our knowledge, this is the first work to discuss whether large sampling locations are always more effective for calibrating spatial interaction model and apply business area analysis to location data from mobile phone users, opening a new area for business applications.

The rest of this paper is organized as follows. Section 2 introduces the mobile phone location dataset and the study area. Section 3 describes the method to extract the trips to business areas and the strategy to calibrate the spatial interaction model. Section 4 discusses the analysis results. Section 5 summarizes our findings and discusses future research directions.

2. Study Area and Dataset

The study area of our research is Shenzhen, one of the largest cities in China. This section provides background information of the selected largest business areas in Shenzhen and the actively tracked mobile phone location dataset.

2.1. Study Area

The study area is the city of Shenzhen, China. Shenzhen is one of China's mega-cities and has an area of approximately 2000 km^2. Shenzhen is located in the southern Guangdong Province and across the border from Hong Kong (Figure 1). As a Special Economic Zone, Shenzhen has become the fourth largest economic city in China, after Shanghai, Beijing, and Guangzhou, and has developed into an influential international city [47]. The prosperous socioeconomic status of Shenzhen makes it a good choice for business area analysis in China.

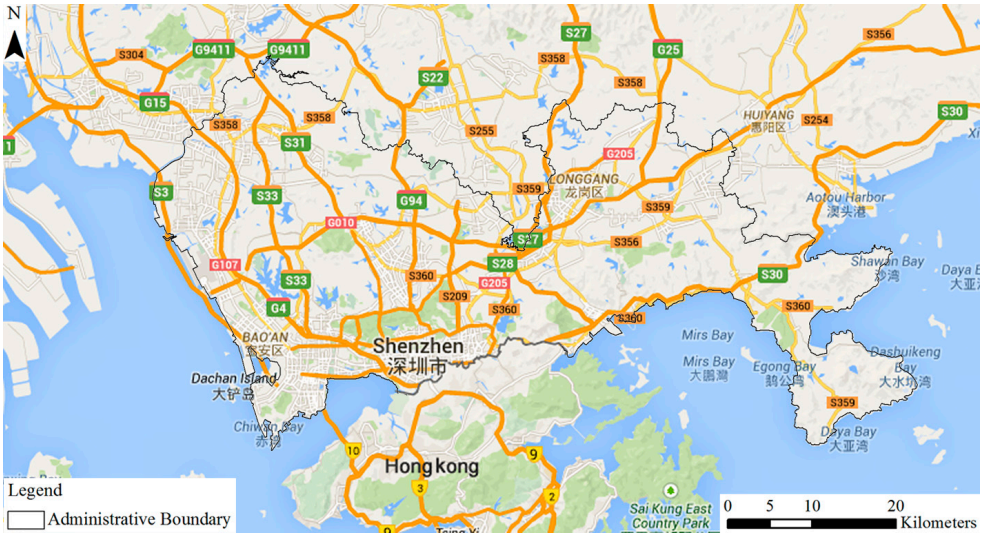

Figure 1. Location of Shenzhen (maps.google.com).

This paper uses the five biggest and most influential commercial areas in Shenzhen [48], as shown in Figure 2. The Dongmen Commercial Pedestrian Street (referred to as "D"), is one of the oldest commercial centers in Shenzhen. It is not a department store, rather it is an open-air shopping area that is interconnected with over one hundred shops. The Renmin-nan Commercial Area (referred to as "R") was the first commercial business zone when Shenzhen began its development 25 years ago. Famous shopping malls were integrated to form the first walking "air corridor" in Shenzhen. The Huaqiang-bei Commercial Area (referred to as "H") is the most prosperous shopping area in Shenzhen. It is not only a business circle for electronic products but also a center for department stores and restaurants. The Overseas Chinese Town (referred to as "O") is located in the heart of the

city. There are many specialty food streets, a sound system supermarket, a bar street, a Western fast food restaurant, a bookstore, a drugstore and other stores in the area. The Nanshan Commercial Area (referred to as "N") is the home to the Haiya department store, the Children's World Nanshan Store, Sundan electronic appliances, HOBA International Furniture Plaza, and the Wanjia Department Store. The total area of each commercial area is shown in Table 1.

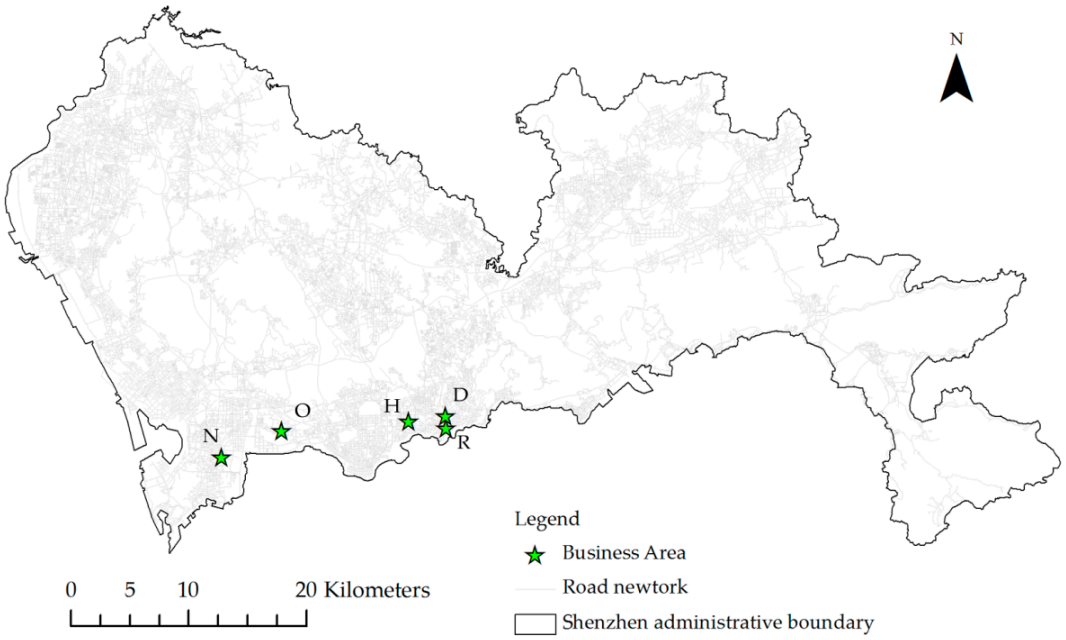

Figure 2. Locations of the five largest commercial areas in Shenzhen city.

Table 1. The total area of each commercial area.

Commercial Area	D	R	H	N	O
Area (1000 m^2)	279.10	729.20	1567.79	1155.95	1569.92

These commercial areas are patronized not only for working and purchasing goods but also for recreation. The goods include famous high-class trademark items as well as new fashions. Additionally, some areas provide cinemas and game centers. All five of these regions are well served by public transportation. These five commercial areas are the most prosperous and attractive and were used as the study areas for our research.

2.2. Data Description

The mobile phone location data used in this research are active tracking data, as shown in Table 2. Extensive work has been conducted using mobile phone location data to analyze human mobility patterns [29,49]. Data in this study are provided by a telecommunications company in Shenzhen, China, for research purposes. Each location record was generated when a mobile phone user sent or received a phone call/text message. Different from CDRs, the location of most mobile phone users in this dataset was recorded approximately every 60 min as the latitude and longitude of a nearby cell tower. In total, this actively tracked mobile phone location dataset contains location information from over 16 million anonymous phone numbers from a Friday in 2012.

For privacy concerns, this study did not obtain any personal information. Each mobile phone number was assigned a unique user ID. In addition, all mobile phone location data were collected at the mobile phone tower level such that the specific activity locations were not revealed. The density of mobile phone towers varied in different parts of the study area. Overall, cell phone towers are densely

distributed in the center of the city or areas with large populations; therefore, resulting in higher data accuracy. In suburban areas, cell phone towers are sparsely distributed and result in lower position accuracy. Nevertheless, mobile phone location data can be a reasonable data source to describe human mobility [49].

Table 2. Example of mobile phone records during the data collection period.

ID	Date	Time	Longitude	Latitude
User1	2012/**/**	07:39:27	114. *****	22. *****
User1	2012/**/**	08:21:36	114. *****	22. *****
User1	2012/**/**	08:53:36	114. *****	22. *****
...	...	...	...	...
User2	2012/**/**	03:28:41	114. *****	22. *****
...	...	...	...	...

The sign ***** ignores the minutes of a Longitude or a Latitude, and the sign **/** ignores the exact month and day due to privacy protection.

3. Methodology

In this section, we introduce how to extract O/D pairs from commercial areas and the method for calibrating the spatial interaction model. The trajectory is defined as the location sequence of an individual in space and time. Ideally, the space-time trajectory of moving objects is continuous. However, the records of mobile phone location data are not continuous due to the low temporal sampling frequency. Thus, a group of discrete location records sequenced in space and time is used to represent the trajectory of an individual.

$$Tr = \{(x_1, y_1, t_1), (x_2, y_2, t_2), ..., (x_n, y_n, t_n)\}, \quad (1)$$

where n is a set describing the spatial-temporal discrete location records. Each element (x_i, y_i, t_i) represents the coordinates (x_i, y_i) of the latitude and longitude of a nearby cell tower of this individual at time t_i.

3.1. Extracting Trips towards Commercial Areas

Before calibrating the spatial interaction model, the trips to each commercial area should be extracted. Three basic elements should be considered: the polygon scope of the commercial area, the time entering the polygon (t_A), and the time leaving the polygon (t_L).

In this paper, if an individual remained in the polygon scope of a commercial area no less than a certain time threshold, then a "stay" is formed. We set this time threshold to 1 hour due to the time resolution of our dataset. Since most of the shopping centers are open from 9:00 a.m. to 11:00 p.m., we assigned the following rules to extract the trips to commercial areas:

Rule 1: Stay duration is no less than 1 h
Rule 2: The arrival time is after 9:00 a.m.
Rule 3: The leave time is before 11:00 p.m.

If a stay meets these requirements, the location record before entering the commercial area is treated as the origin of the trip. Following the approach described above, attracted trips from cell phone towers to each commercial area were extracted. Many previous studies have used mobile phone location data to investigate the spatial interactions in complex urban environment [31,41,42]. There may be some uncertainties in the extraction of origins/destinations from mobile phone data. The location records are quite sparsely distributed in space and time from the individual perspective, due to the uneven distribution of people's phone activities [50].

However, there were zero trips from some of the origin cell phone towers to some commercial areas (called zero interactions). Thus, if the total number of trips originating from a tower was greater

than 5, then the zero interaction was added by 1. Otherwise, this cell phone tower was not considered in the study [51]. A total of 2621 cell phone towers were selected for the calibration of the spatial interaction model (hereafter, cell phone towers means these 2621 towers). Due to the requirement of dataset provider, it is not allowed to show the spatial distribution of point-based cell phone towers (hereafter, the distributions of cell phone towers are all presented by kernel density). The spatial kernel density distributions of the cell phone towers and the valid cell phone towers are shown in Figure 3.

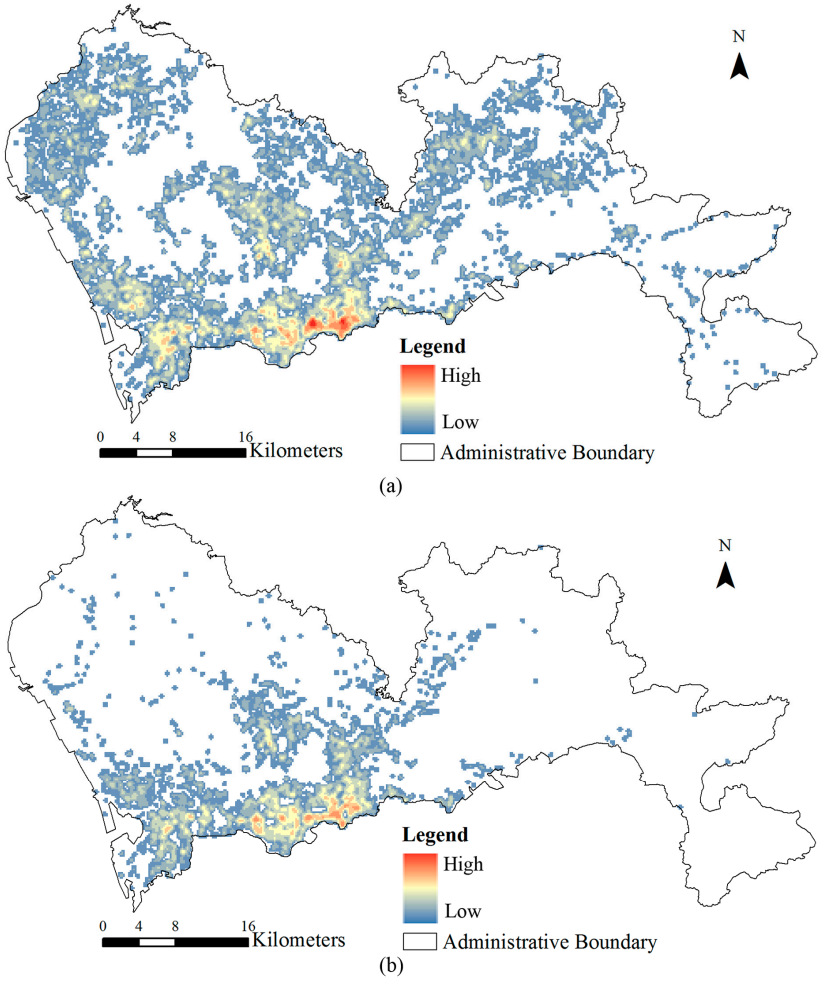

Figure 3. Spatial kernel density of total cell phone towers (**a**); and valid cell phone towers with trips towards each commercial area (**b**).

3.2. Randomly Calibrate the Huff Model

The Huff model [52,53] is a spatial interaction model that seeks to describe in a spatially explicit manner the flow of people across space to a fixed set of locations to access goods or services. The Huff model is formulated as follows:

$$T_{ij} = \frac{s_j^{\alpha} d_{ij}^{-\beta}}{\sum_k^J s_k^{\alpha} d_{ik}^{-\beta}}, \quad (2)$$

where T_{ij} (varies from 0 to 1) is the probability of residents at origin i interacting with business area j. In the Huff model, the polygon size of the commercial area (s) is used to represent the attraction according to many previous studies [3,11,38,54]; and the trip distance (d) is used as the cost; α and β are the sensitivity parameters that associate T_{ij} with attraction variable s and cost d (both of the parameters will be calibrated); and J is number of commercial areas. The Huff model is based on Newton's law of universal gravitation. Before using the Huff model to evaluate the interactions between locations and

facilities, the parameters α and β need to be calibrated to ensure the estimated flows are best fit to the observed data.

The most common methods used for Huff model calibration are the maximum likelihood method and ordinary least square regression. The descriptions of these two methods can be found in Fotheringham and O'Kelly [51]. They note that although the criteria of these two methods are different, the parameter estimates are similar, which is also verified by our study. Therefore, this paper will only choose one calibration method that uses the least square regression method derived by Fotheringham and O'Kelly [51] to calibrate the spatial interaction model.

To investigate the impacts of sampling points on the calibration of the Huff model, multiples of 30 phone towers (such as 30, 60, 90, etc.) are randomly selected due to the reason that some spatial analysis are reliable if the input samples are at least 30 [55]. For each multiple of 30 selected phone towers, we randomized the selection 500 times, as shown in Figure 4. Then, the least square regression is performed. Each random sample can derive a group of parameters, which are used to evaluate the bias between observed probability (P_{ij}) and estimated probability (T_{ij}) of all the 2621 cell phone towers. The sum of squared errors (SSE) is frequently used to measure the bias [56,57].

$$SSE = \sum_i \sum_j (T_{ij} - P_{ij})^2, \tag{3}$$

This paper uses the Huff model, one type of spatial interaction model, as an example to examine the effects of different locations and sizes of cell phone tower samples on the calibration of the model parameters. We vary the selected number of cell phone towers by multiples of 30 until we reach the total number of towers with flows to the five destinations, to calibrate the spatial interaction model and gradually answer the questions we have proposed.

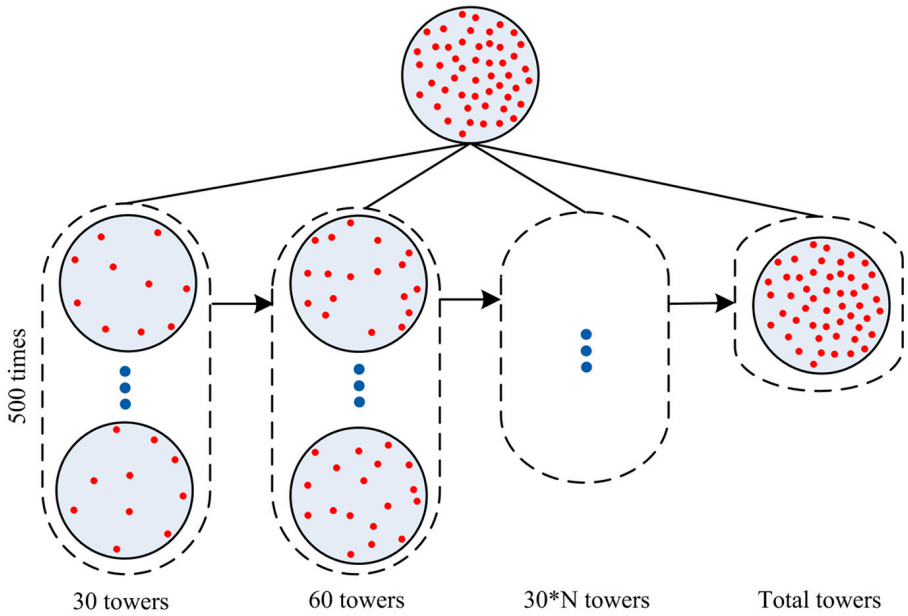

Figure 4. Random rules for selecting valid cell phone towers.

4. Results

4.1. Distribution of SSE

We calculated the distribution of SSE under each calibration parameter, as shown in Figure 5. The SSE of all the 2621 cell phone towers can be quite different with different numbers of randomly selected cell phone towers.

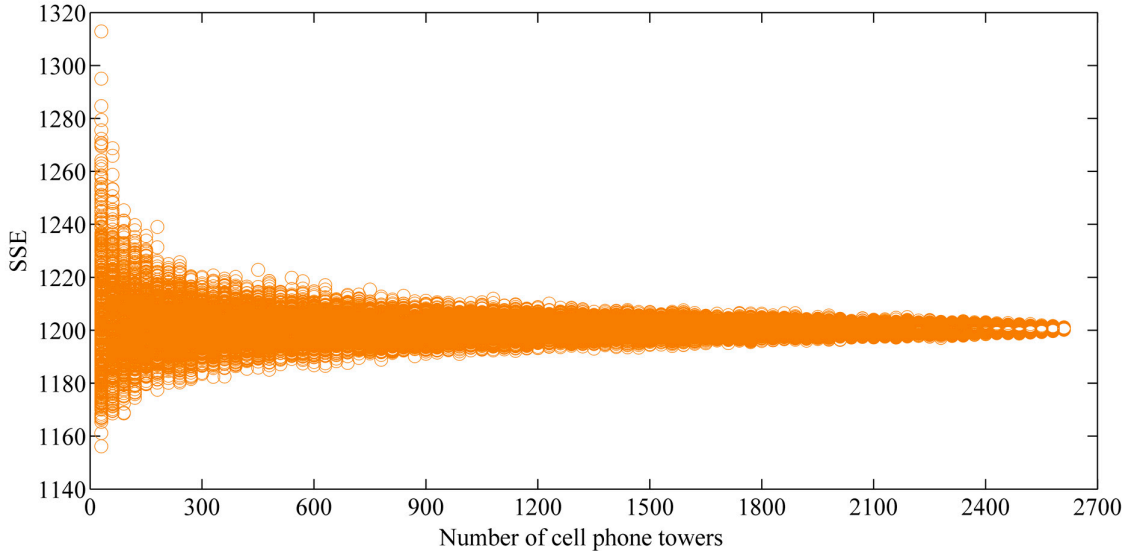

Figure 5. Distribution of *SSE*.

Firstly, as the number of random cell towers grows, the value of *SSE* is less fluctuant and closer to 1205, which is the total sum of squared errors (*TSSE*) when all cell towers were used for calibration. In particular, when more than 900 cell phone towers were used, the SSE is between 1180 and 1220. As the number of towers increases, the interval of the SSE decreases.

Secondly, the fewer random cell towers used, the more the SSE fluctuates. When the number of cell phone towers is low (such as 150 cell phone towers), we can obtain both a better and a worse calibration result than when using all cell phone towers for calibration. For example, when using 30 cell towers for calibration, the *SSE* can fluctuate from approximately 1150 to 1318. When the random dataset is 30, the selected towers can provide both a better and a worse calibration result. Some cell phone towers may appear more than once in all of the random combinations. Later, we will investigate the general characteristics of these best-performing towers.

Most importantly, few random sampling locations have the ability to improve the calibration results compared to many random sampling locations. We can conclude that it is not always that more sampling locations lead to the better solutions for calibrating spatial interaction models. In other words, when we conduct surveys or questionnaires, the locations are very important; or when we use large location data for the calibration of the spatial interaction model, not all sampling locations are valuable for calibration. The fluctuation of *SSE* is greatest when using 30 towers for calibration. In the next section, we use a random sample of 30 towers to investigate the hidden patterns of these better performing calibrations due to the most fluctuant *SSE* distribution when using this random sample.

4.2. Finding Out Which Cell Phone Towers Best Fit Each Commercial Area

Previously, we assumed that there were some common characteristics between better performing towers. Firstly, this paper measured the similarity between the estimated percentage (T_{ij}) from a location to each commercial area and the observed percentage of trips (P_{ij}) towards each commercial area. The most similar pair is considered to belong to that commercial area. The similarity index (*SI*) is measured by,

$$SI = \frac{2 \cdot min(T_{ij}, P_{ij})}{T_{ij} + P_{ij}} \quad (4)$$

where T_{ij} is the estimated percentage of flows from tower i to commercial area j, and P_{ij} is the observed percentage of flows. After each cell phone tower is tagged with their best fit commercial area, we determine whether this tower is within the tagged commercial area's Thiessen polygon (the Thiessen polygon is derived from the center of each commercial area). Due to a large number of

random selections, some towers may be selected more than once. In this case, the maximum number of the best-fit commercial area is assigned as its tagged commercial area.

Finally, each cell phone tower is classified by its best fit commercial area, as shown in Figure 6. Combined with its affiliated Thiessen polygon, we obtain the following statistical Table 3. This table illustrates that, except for "D", the other four commercial areas have a maximum percentages of best fit towers within their scope, especially for "R", "H" and "N", where the percentage of best-fit towers within their scope are 70.22%, 62.37%, and 61.35%, respectively.

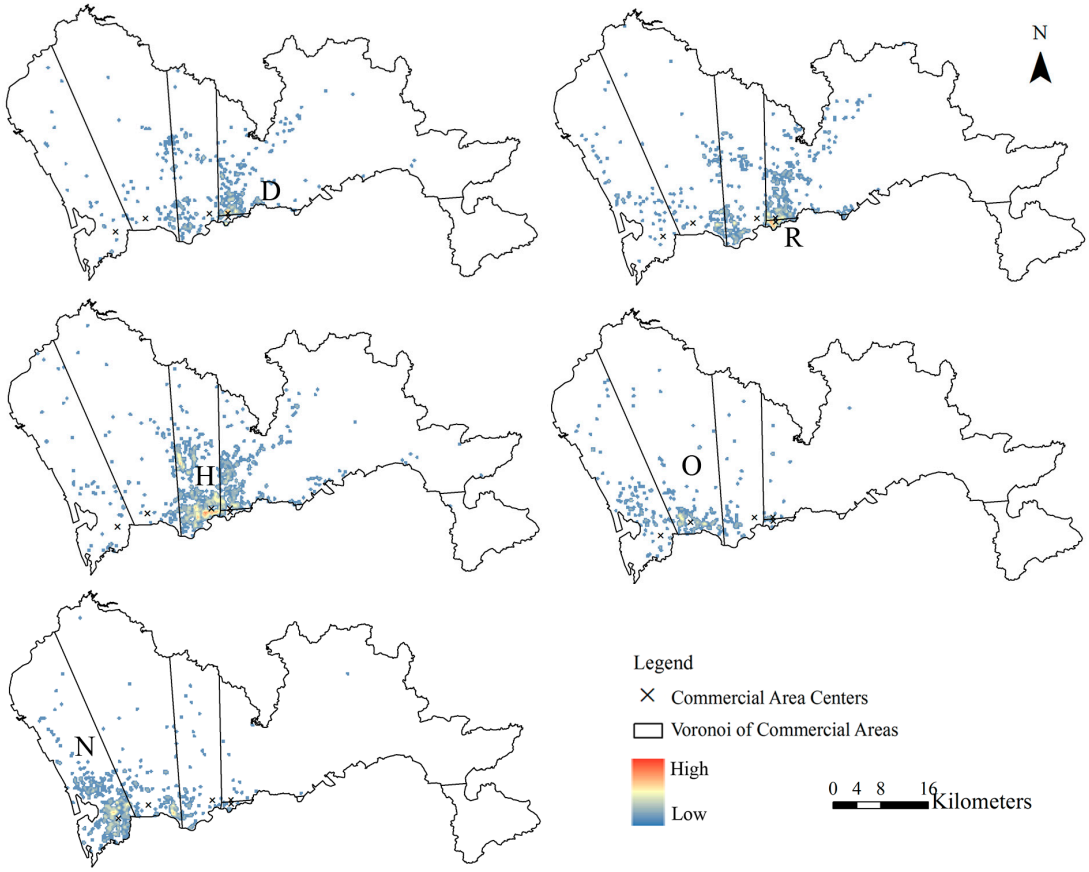

Figure 6. Spatial kernel density of cell phone towers tagged with their best fit commercial areas.

Table 3. Percentages of best-fit cell phone towers to the five commercial areas.

Cell Phone Towers	Fit D	Fit R	Fit H	Fit O	Fit N	Total
In Polygon D	23.61%	34.84%	37.70%	2.24%	1.54%	100%
In Polygon R	14.89%	70.22%	11.70%	1.06%	2.13%	100%
In Polygon H	11.99%	15.72%	62.37%	3.86%	6.05%	100%
In Polygon O	12.77%	19.70%	12.99%	28.58%	25.98%	100%
In Polygon N	3.38%	8.82%	8.26%	18.20%	61.35%	100%

However, the highest percentage of best fit towers is not always within the areas scope. For example, the percentage of best fit towers in "D" is only 23.61%, but 34.84% and 37.70% of towers in "D" are best fit for "R" and "H", respectively, which are higher than "D" itself. At the same time, although the percentage of best fit towers within polygon "O" is the highest (28.58%), 61.42% of towers are best fit for other nonadjacent commercial areas.

For most cases ("R", "H", "O" and "N"), the highest percentage of best fit towers are within their polygon scope, which reveals that this characteristic of spatial adjacency can play a role when choosing

the random sample. Next, we attempt to choose towers that best fit their adjacent commercial areas to further reveal the attributes (distance and flows) of these towers.

4.3. High-Accuracy Calibration by Using Spatial Adjacency

The distance in this paper was represented by the spatial adjacency [58]. To determine whether the tower's best fit commercial area is consistent with the tower's most adjacent commercial area, this paper divides the 2621 cell phone towers into two clusters. The two clusters are as follows:

(1) Set A: The tower's best fit commercial area is consistent with the tower's most adjacent commercial area. This subset account for 45.64% of the 2621 cell phone towers, as shown in Figure 7a.
(2) Set B: The tower's best fit commercial area is not consistent with the tower's most adjacent commercial area. This subset account for 54.36% of all the 2621 cell phone towers, as shown in Figure 7b.

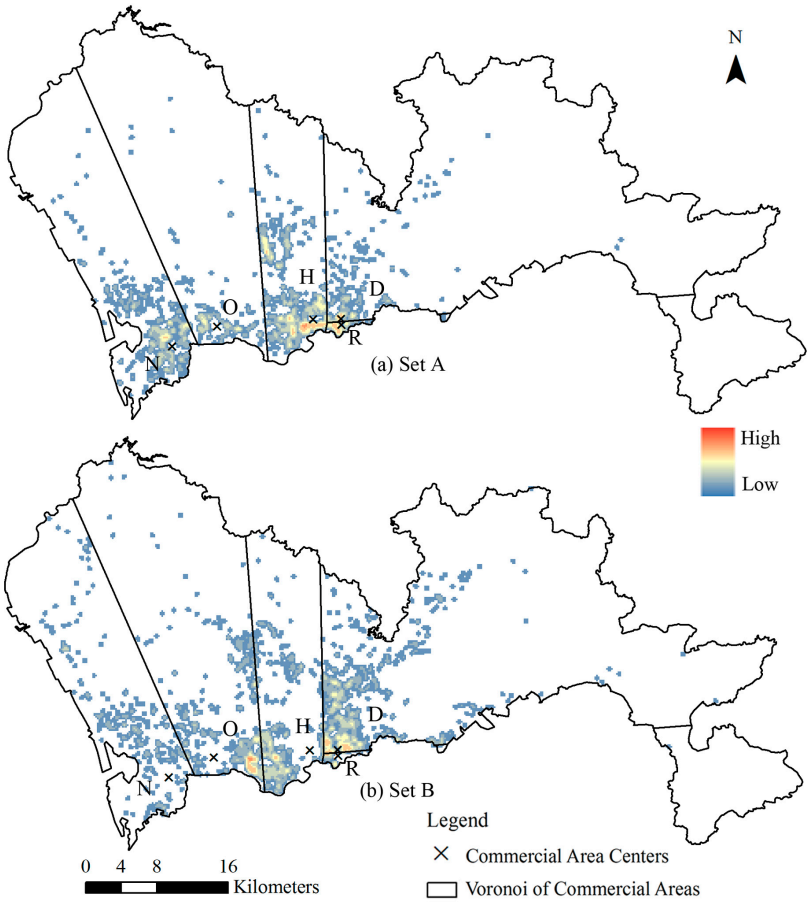

Figure 7. Spatial kernel density distribution of: Set A (**a**); and Set B (**b**).

Therefore, Set A is a complementary set of Set B. Both of the sets consist of the 2621 cell phone towers. The spatial distributions of these two sets are shown in Figure 7. It is evident that Set A and Set B are mixed in spatial distribution.

To investigates the different effects of the two sets on the calibration of the Huff model, multiples of 30 phone towers or its integer times (60, 90, etc.) are randomly selected from each set. Each multiple of 30 or its integer times of selected phone towers were randomly selected 500 times. Each time, the bias between observed P_{ij} and estimated T_{ij} of the total 2621 cell phone towers are estimated by *SSE*. The distributions of the *SSE* of the two sets are shown in Figure 8. The average value of all *SSE* and the percentage of times that *SSE* fell below 1205 and above 1205 are calculated, as shown in Table 4.

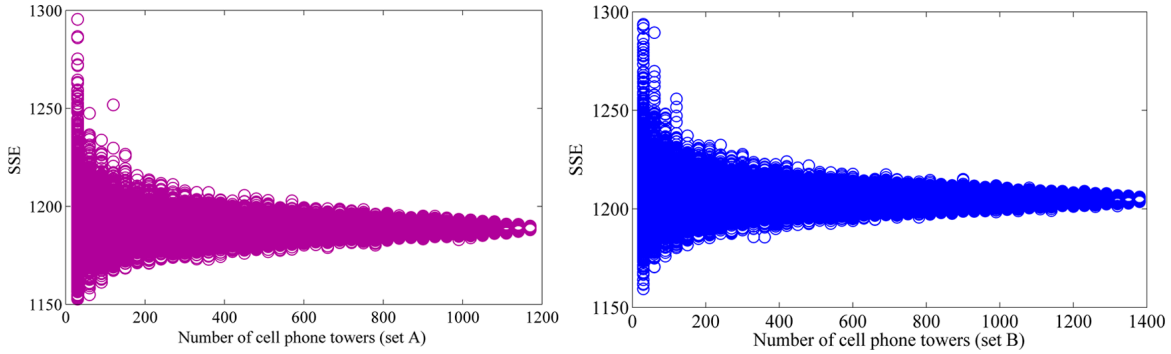

Figure 8. Distributions of SSE using cell phone towers randomly selected from Set A and Set B.

Table 4. Statistic result of the two sets.

Cell Phone Towers	Counts/Percentage	Average	Below 1205	Above 1205
Set A	1196 (45.64%)	1189.3	96.2%	3.8%
Set B	1425 (54.36%)	1205.4	9.2%	90.8%

It is obvious from Table 4 that the average of *SSE* from Set A is 1189.3, which is lower than the *TSSE* (1205). When using Set B to calibrate the Huff model, the average of *SSE* is 1205.4, which is nearly equal to *TSSE* (1205). Moreover, by using Set A, the percentage of random times that *SSE* is better than *TSSE* is 96.2%, which is significantly higher than when using Set B (with only 3.8% of random times better than *TSSE*). Therefore, using Set A (the tower's best fit commercial area is consistent with the tower's most adjacent commercial area) can result in a more effective calibration.

Until now, how to directly distinguish this kind of dataset was still unknown. From a spatial distribution point of view, Set A and Set B are well mixed. Moreover, the percentage of cell phone towers in Set A and Set B are 45.64% and 54.36%, respectively, which are both near 50%. Thus, how to easily distinguish Set A from Set B needs to be resolved. To distinguish these two sets directly, the volume of flows of each set is calculated, as shown in Table 5.

Table 5. Volume of flows of towers in each set.

	Average	≥150	[100, 150]	[50, 100]	[0, 50]
Set A	370	31.9%	5.9%	10.9%	51.1%
Set B	150	19.7%	6.5%	12.5%	61.3%

It is obvious from Table 5 that the average number of flows from each cell phone tower in Set A is 370, which is much higher than the average number of flows from each cell phone tower in Set B, 150. Further, the percentage of cell phone towers with more than 150 trips in Set A is 31.9%, which is also much higher than the percentage of cell phone towers with more than 150 trips in Set B, which is only 19.7%. Thus, the volume of trips from each tower plays a major role in distinguishing the better performing towers from all of the 2621 cell phone towers. In the next section, we will investigate how the volume of trips affects the calibration results.

4.4. High-Accuracy Calibration by Volume of Attracted Tirps

4.4.1. Calibration by Using Top 30 Cell Phone Towers with Highest Trips

According to the previous experiments, we know that volume of trips from each cell phone tower is a criterion to distinguish the better performing towers from all the 2621 cell phone towers. In this section, we select the top 30 cell phone towers with the highest number trips to the five commercial

areas. The spatial distribution of the top 30 cell phone towers is shown in Figure 9. Then, the model parameters are calibrated by these 30 towers and *SSE* is 1165.1, which is much lower than the *TSSE* (using all the towers to calibrate the model).

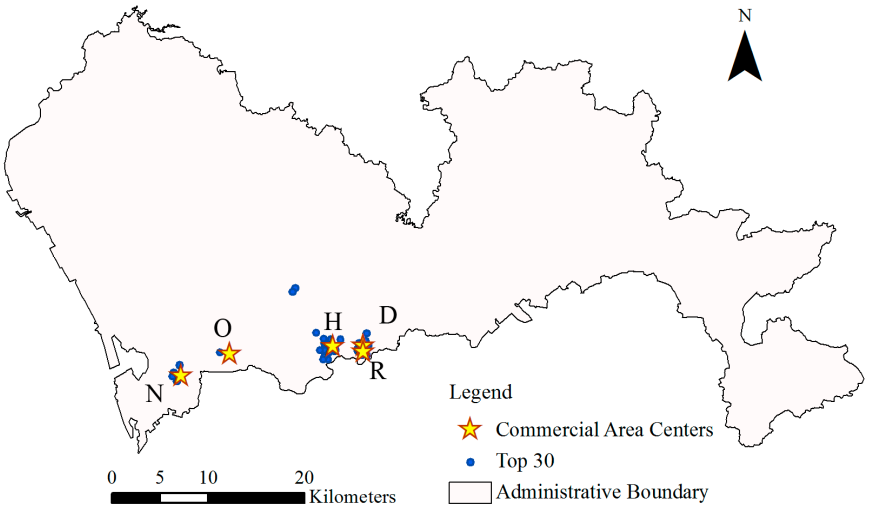

Figure 9. Top 30 flows from cell phone towers towards the business areas.

Each cell phone tower has five different distances to the five commercial areas. We divide the urban space into multiple of 3 km according to the distance of cell phone towers to five commercial areas. Each cell phone tower may be within one commercial area's 3 km buffer scope while also within another commercial area's 6 km buffer scope. If a cell phone tower is located within at least one commercial area's 3 km scope, we define it as is in the commercial area's 3 km scope. Then, we calculate whether the bias of each cell phone tower is below (better than) average or above (worse than) the average bias. The average bias is the mean value of *SSE* of the 2621 cell phone towers. As shown in Table 6, by using the top 30 cell phone towers with the highest number of trips to calibrate the model, 76.8% of the 961 cell phone towers in the commercial area's 3 km buffer scope will behave better than average and only 23.20% of cell phone towers behave worse than average. For the 742 cell phone towers in the 3 to 6 km scope, 52.07% perform better than average. When the buffer scope is over 9 km, more than 81% of cell phone towers in that scope behave worse than average, but the total number of towers within that scope is much less than within 6 km.

Table 6. Effects on towers with a different distance.

Distance (km)	Below Average	Above Average	Counts
[0, 3]	76.80%	23.20%	961
[3, 6]	52.07%	47.93%	742
[6, 9]	27.92%	72.08%	351
[9, 12]	11.42%	88.58%	254
[12, 15]	5.74%	94.26%	122
[15, 18]	9.72%	90.28%	72
[18, 21]	9.09%	90.91%	44
[21, 24]	3.23%	96.77%	31
≥24	18.18%	81.82%	44

Selecting cell phone towers with a large volume of trips for calibration can significantly benefit the model when towers are located within 6 km. In the text section, we will verify the effects on calibration of cell phone towers with a different volume of trips.

4.4.2. Calibration by Using Selected Towers with Higher Volume of Flows

From the previous experiments, we can conclude that the effects of spatial proximity are reflected by flows. The best fit towers within the areas polygon scope have higher flows and perform better. In particular, the top 30 towers with highest number of trips also behave better. Thus, we use different volumes of flow to test the effects of flows on the parameters calibration.

We select cell phone towers with more than 10 trips in multiples of 10. The distribution of the percentage of cell phone towers with specified lower bounds of trips is shown in Figure 10. As the lower bound of trips increases, cell phone towers with a small number of trips are gradually excluded. In each case, we randomly select 30 cell phone towers 500 times to calibrate the Huff model. Then, the calibrated parameters are used to calculate the *SSE* for the total 2621 cell phone towers. Each *SSE* is compared with the *TSSE*. Finally, we determine the percentage of random times where the *SSE* is lower than *TSSE*, as shown in Figure 11. The horizontal axis represents the low bound of trips, namely, the trips of selected towers that are higher than the specified value. The vertical axis represents the percentage of times where the *SSE* is lower than *TSSE* for all 500 random selections. The maximum of the low bound of trips is set to 500 because there are only 10% of towers with more than 500 trips.

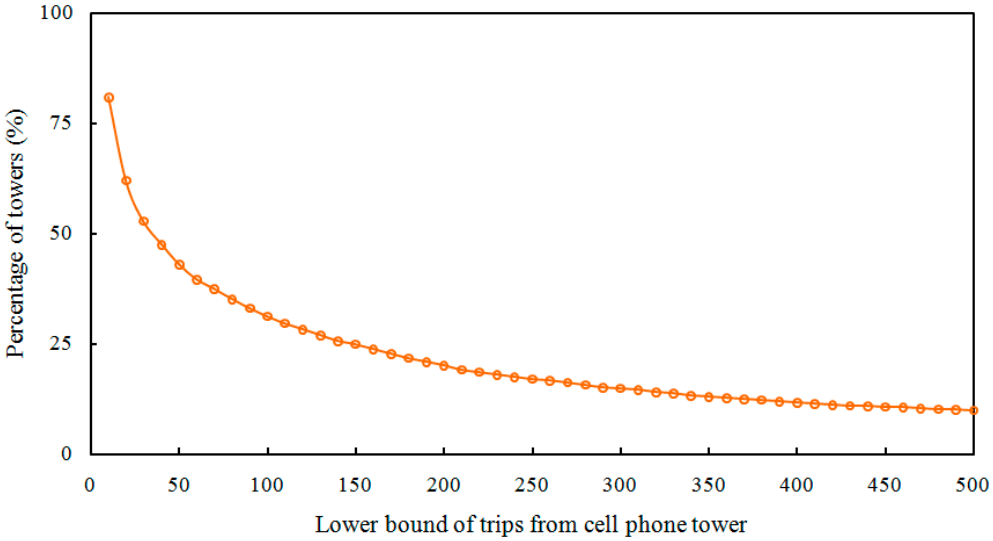

Figure 10. Percentage of cell phone towers with specified lower bounds of trips.

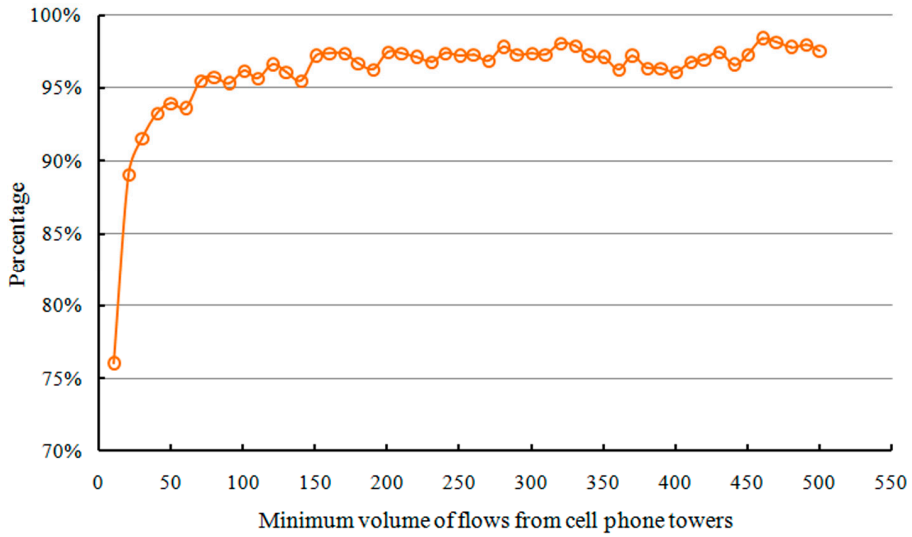

Figure 11. Percentage of times with *SSE* better than the calibration using all towers.

From Figure 11, it is clear that the percentage of times where the *SSE* is lower than *TSSE* changes significantly from 76% to 95% when the low bound of trips is increased from 10 to 70. Particularly, when the low bound of trips is higher than 70, the percentage of times where the *SSE* is lower than *TSSE* is steadily greater than 95%. This result indicates that the probability of obtaining better results is greater when using a large volume of trips from cell phone towers to calibrate the spatial interaction model. However, the question, what are the effects of the calibrated parameters on small cell phone towers when using big volume trips of cell phone towers, remains. In the next section, we will verify this effect.

4.4.3. Effects on Towers with "Small" Volume of Trips

From the experiment above, the probability of obtaining better results is higher when using a large volume of trips from cell phone towers to calibrate the spatial interaction model. However, the SSE is the overall measurement of the bias between the observed and estimated probability. When we choose the high volume of trips to calibrate the model, the effect on the small volume of trips is ignored. It may be that the overall better result is built at the expense of the small volume of trips. In this part, we select the towers with more than 10 trips as the whole candidate set to evaluate the *SSES* of towers with less than 70 trips. The number of towers with less than 70 trips is 1641. Thus, the *SSES* is the estimated and observed probability bias of these 1641 cell phone towers. We consider these 1641 towers as ones with a small volume of trips because when the low bound of trips is higher than 70, the percentage of times where the *SSE* is lower than *TSSE* is steadily greater than 95%.

Thus, we get the distributions of *SSES*, as shown in Figure 12. The horizontal axis represents the low bound of trips, namely, the trips of selected towers are higher than the specified value. The vertical axis represents the *SSES* in all 500 random selections. The maximum low bound of trips is also set as 500.

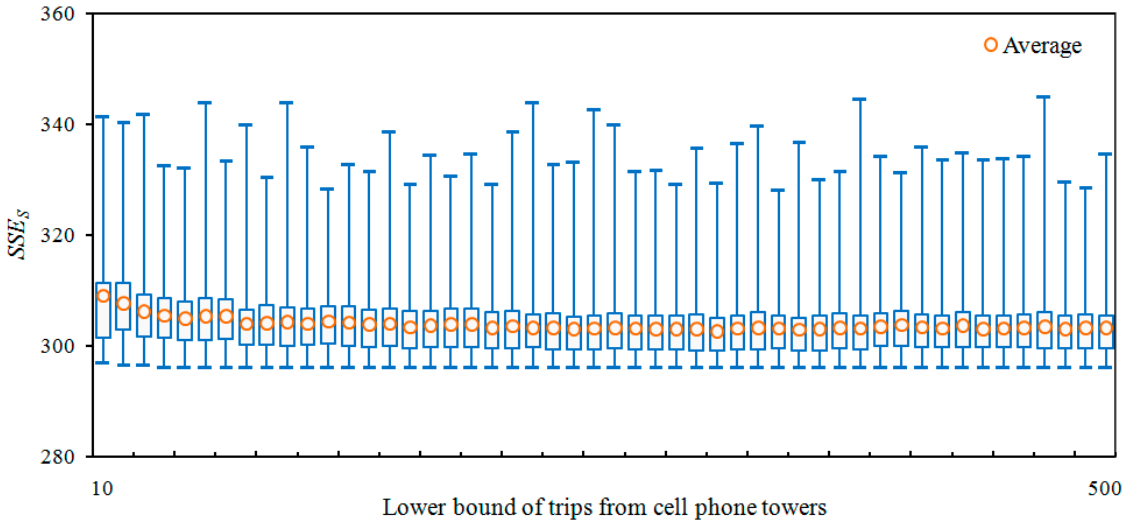

Figure 12. Distribution of *SSES* (quartile graph).

It is evident from the figure above that *SSES* is steadily distributed between 297 and 348 no matter whether the low bound of trips is greater than 10 or greater than 500. At each low bound, the *SSES* maintains a similar interval (the interval length is approximately 51). Thus, when choosing the large volume of trips to calibrate the model, the bias of the small volume of trips evaluated by *SSE* is not affected; that is, the concern that the small number of trips from cell phone towers may be sacrificed to get overall best results can be eliminated. Finally, using high volume trips from cell phone towers to calibrate the spatial interaction model is a good choice not only for obtaining better results but also for reducing computational demand.

5. Conclusions

Advancements in information and communication technology over the past two decades have produced massive and various kinds of big location data, which encourages novel insights for studies of human travel and activity patterns and other perspectives of research. However, "are large volume of sampling locations effective for calibrating spatial interaction model" is still a question for mobility research. This paper attempts to answer this question in the perspective of Huff model calibration, by using massive mobile phone location data, and some conclusions can be drawn as follows.

On the one hand, for the calibration of the Huff model, it is not "the more sampling locations are, the better calibration result is". When we take all the cell phone towers into calibration, the *SSE* is not the lowest. Moreover, the fewer random cell towers, the more fluctuant the *SSE*. However, small random sampling sizes have the ability to improve calibration results than large random samples. In the calibration of the spatial interaction model, too much sampling locations may be just as bad as too little. Some special locations hidden in the large location data are more urgent and should be used and analyzed to provide some new insights into data science.

On the other hand, when we examined the characteristics of these better performing towers, the towers that are a best fit to their adjacent commercial area are good choices, which illustrates that spatial proximity plays a role when selecting the random sample. Besides, cell phone towers with this characteristic have a larger volume of trips than the other towers. Thus, the volume of flows from cell phone towers is the measurement to distinguish the valuable locations from the poorly performing locations. When we randomly selected 30 towers with more than 70 trips, the percentage of times where the *SSE* is lower than *TSSE* is steadily higher than 95%. Moreover, when choosing the big volume of trips to calibrate the model, the bias of small volume of trips evaluated by *SSE* is not affected, that is, the concern that the small trips of cell phone towers may be sacrificed to get overall best results can be eliminated. Thus, using sampling locations with high volume trips to calibrate the spatial interaction model is a good choice not only for obtaining better results but also for reducing computational demand.

However, we do note several limitations and challenges of this research, such as:

(1) In this paper, we adopted the Huff model to define business area, and only used size to represent the attractiveness. This simplification created a mismatch between the predicted attracted areas and the observed data. Other factors such as the number of POIs, parking conditions, price level and types of companies, malls in business areas may also influence the attractiveness. In the future, additional research is needed to identify the detailed attractiveness factors and a proper spatial interaction model to better depict the relationships.

(2) Another limitation is that we have not noted the social characteristics of these better performing locations. The combinations of other factors, such as resident distribution, income, land use type and so on, may reveal the social aspects of these better performing locations, which can provide better guidance to surveying or sampling.

(3) In this paper, we investigated the effects of sampling locations on the calibration of spatial interaction model between urban environment and commercial areas. However, our findings may or may not be applicable to other land use types due to the reason that different land use patterns also play a role in the model calibration.

(4) There may be some uncertainties in the extraction of origins/destinations from mobile phone data. It is possible that the "origins" used in this paper were just some passing-by locations, due to the reason that the footprints of mobile phone subscribers were sparsely sampled in space and time [50], so it is hard to limit the "origin" as a "stay" where the subscribers have spent a certain time duration. In the future, dataset like GPS tracking data could be used to reduce the potential uncertainty in extracting the origins or destinations.

Acknowledgments: The authors would like to thank the valuable comments from anonymous reviewers. This study was jointly supported by the National Natural Science Foundation of China (Grants #41231171, #41371420, #41371377 and #41301511), the innovative research funding of Wuhan University (2042015KF0167), the Arts and Sciences Excellence Professorship and the Alvin and Sally Beaman Professorship at the University of Tennessee, and the International Science-technology Cooperation Project of Guangdong Province (2014A050503053).

Author Contributions: This research was mainly formulated and designed by Shiwei Lu, Shih-Lung Shaw and Zhixiang Fang. Ling Yin provided the dataset. Shiwei Lu and Xirui Zhang performed the experiments and analyzed the data. Shiwei Lu, Zhixiang Fang and Shih-Lung Shaw wrote the manuscript. Zhixiang Fang, Shih-Lung Shaw and Ling Yin reviewed the manuscript and provided comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Eppli, M.J.; Shilling, J.D. How critical is a good location to a regional shopping center? *J. Real Estate Res.* **1996**, *12*, 459–468.
2. Lee, M.L.; Pace, R.K. Spatial distribution of retail sales. *J. Real Estate Financ. Econ.* **2005**, *31*, 53–69. [CrossRef]
3. Suárez-Vega, R.; Gutiérrez-Acuña, J.L.; Rodríguez-Díaz, M. Locating a supermarket using a locally calibrated Huff model. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 217–233. [CrossRef]
4. Lin, T.G.; Xia, J.C.; Robinson, T.P.; Olaru, D.; Smith, B.; Taplin, J.; Cao, B. Enhanced Huff model for estimating Park and Ride (PnR) catchment areas in Perth, WA. *J. Transp. Geogr.* **2016**, *54*, 336–348. [CrossRef]
5. Luo, J. Integrating the Huff model and floating catchment area methods to analyze spatial access to healthcare services. *Trans. GIS* **2014**, *18*, 436–448. [CrossRef]
6. Applebaum, W.; Cohen, S.B. The dynamics of store trading areas and market equilibrium 1. *Ann. Assoc. Am. Geogr.* **1961**, *51*, 73–101. [CrossRef]
7. Ghosh, A.; Rushton, G. *Spatial Analysis and Location-Allocation Models*; Van Nostrand Reinhold Company: New York, NY, USA, 1987.
8. Mendes, A.B.; Themido, I.H. Multi-outlet retail site location assessment. *Int. Trans. Oper. Res.* **2004**, *11*, 1–18. [CrossRef]
9. Applebaum, W. Methods for determining store trade areas, market penetration, and potential sales. *J. Mark. Res.* **1966**, *3*, 127–141. [CrossRef]
10. Haines, G.H., Jr.; Simon, L.S.; Alexis, M. Maximum likelihood estimation of central-city food trading areas. *J. Mark. Res.* **1972**, *9*, 154–159. [CrossRef]
11. Wang, Y.; Jiang, W.; Liu, S.; Ye, X.; Wang, T. Evaluating trade areas using social media data with a calibrated huff model. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 112. [CrossRef]
12. Rodrigue, J.P.; Comtois, C.; Slack, B. *The Geography of Transport Systems*; Routledge: New York, NY, USA, 2006.
13. Batty, M. Exploratory calibration of a retail location model using search by golden section. *Environ. Plan. A* **1971**, *3*, 411–432. [CrossRef]
14. Diplock, G.; Openshaw, S. Using simple genetic algorithms to calibrate spatial interaction models. *Geogr. Anal.* **1996**, *28*, 262–279. [CrossRef]
15. Huff, D.L.; McCallum, B.M. *Calibrating the Huff Model Using ArcGIS Business Analyst*; ESRI White Paper; ESRI: Redlands, CA, USA, 2008.
16. O'Kelly, M.E. Trade-area models and choice-based samples: Methods. *Environ. Plan. A* **1999**, *31*, 613–627. [CrossRef]
17. Yue, Y.; Lan, T.; Yeh, A.G.O.; Li, Q. Zooming into individuals to understand the collective: A review of trajectory-based travel behaviour studies. *Travel Behav. Soc.* **2014**, *1*, 69–78. [CrossRef]
18. Kirby, H.R.; Leese, M.N. Trip-distribution calculations and sampling error: Some theoretical aspects. *Environ. Plann. A* **1978**, *10*, 837–851. [CrossRef]
19. Watters, J.K.; Biernacki, P. Targeted sampling: Options for the study of hidden populations. *Soc. Probl.* **1989**, *36*, 416–430. [CrossRef]
20. Goodchild, M.F. The quality of big (geo) data. *Dialogues Hum. Geogr.* **2013**, *3*, 280–284. [CrossRef]
21. Lu, S.; Fang, Z.; Shaw, S.L.; Zhang, X.; Yin, L. Quantitative analysis of the effects of spatial scales on intra-urban human mobility. *Geom. Inf. Sci. Wuhan Univ.* **2016**, *41*, 1199–1204.

22. Lu, S.; Fang, Z.; Zhang, X.; Shaw, S.L.; Yin, L.; Zhao, Z.; Yang, X. Understanding the representativeness of mobile phone location data in characterizing human mobility indicators. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 7. [CrossRef]
23. Haklay, M. How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environ. Plan. B Plan. Des.* **2010**, *93*, 3–11. [CrossRef]
24. Goodchild, M.F.; Li, L. Assuring the quality of volunteered geographic information. *Spat. Stat.* **2012**, *1*, 110–120. [CrossRef]
25. Zhou, Q.; Li, Z. How many samples are needed? An investigation of binary logistic regression for selective omission in a road network. *Cartogr. Geogr. Inf. Sci.* **2015**. [CrossRef]
26. Zhao, Z.; Shaw, S.L.; Xu, Y.; Lu, F.; Chen, J.; Yin, L. Understanding the bias of call detail records in human mobility research. *Int. J. Geogr. Inf. Sci.* **2016**, *30*, 1738–1762. [CrossRef]
27. Demirkan, H.; Delen, D. Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. *Decis. Support Syst.* **2013**, *55*, 412–421. [CrossRef]
28. Hashem, I.A.T.; Yaqoob, I.; Anuar, N.B.; Mokhtar, S.; Gani, A.; Khan, S.U. The rise of "big data" on cloud computing: Review and open research issues. *Inf. Syst.* **2015**, *47*, 98–115. [CrossRef]
29. Gonzalez, M.C.; Hidalgo, C.A.; Barabasi, A.L. Understanding individual human mobility patterns. *Nature* **2008**, *453*, 779–782. [CrossRef] [PubMed]
30. Siła-Nowicka, K.; Vandrol, J.; Oshan, T.; Long, J.A.; Demšar, U.; Fotheringham, A.S. Analysis of human mobility patterns from GPS trajectories and contextual information. *Int. J. Geogr Inf. Sci.* **2016**, *30*, 881–906. [CrossRef]
31. Gao, S.; Liu, Y.; Wang, Y.; Ma, X. Discovering spatial interaction communities from mobile phone data. *Trans. GIS* **2013**, *17*, 463–481. [CrossRef]
32. Liu, Y.; Sui, Z.; Kang, C.; Gao, Y. Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. *PLoS ONE* **2014**, *9*, e86026. [CrossRef] [PubMed]
33. Chi, G.; Thill, J.C.; Tong, D.; Shi, L.; Liu, Y. Uncovering regional characteristics from mobile phone data: A network science approach. *Pap. Reg. Sci.* **2016**, *95*, 613–631. [CrossRef]
34. Ratti, C.; Frenchman, D.; Pulselli, R.M.; Williams, S. Mobile landscapes: Using location data from cell phones for urban analysis. *Environ. Plan. B Plan. Des.* **2006**, *33*, 727. [CrossRef]
35. Xu, Y.; Shaw, S.L.; Zhao, Z.; Yin, L.; Fang, Z.; Li, Q. Understanding aggregate human mobility patterns using passive mobile phone location data: A home-based approach. *Transportation* **2015**, *42*, 625–646. [CrossRef]
36. Balcan, D.; Colizza, V.; Gonçalves, B.; Hu, H.; Ramasco, J.J.; Vespignani, A. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 21484–21489. [CrossRef] [PubMed]
37. Kang, C.; Liu, Y.; Guo, D.; Qin, K. A generalized radiation model for human mobility: Spatial scale, searching direction and trip constraint. *PLoS ONE* **2015**, *10*, e0143500. [CrossRef] [PubMed]
38. Yue, Y.; Wang, H.; Hu, B.; Li, Q.; Li, Y.; Yeh, A.G. Exploratory calibration of a spatial interaction model using taxi GPS trajectories. *Comput. Environ. Urban Syst.* **2012**, *36*, 140–153. [CrossRef]
39. Markham, F.; Doran, B.; Young, M. Estimating gambling venue catchments for impact assessment using a calibrated gravity model. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 326–342. [CrossRef]
40. Ministry of Industry and Information Technology of the People's Republic of China. Quarter Book 2016. Available online: http://www.miit.gov.cn/n1146312/n1146904/n1648372/c4802518/content.html (accessed on 3 January 2017).
41. Liang, X.; Zhao, J.; Li, D.; Xu, K. Unraveling the origin of exponential law in intra-urban human mobility. *Sci. Rep.* **2012**, *3*, 2983. [CrossRef] [PubMed]
42. Simini, F.; González, M.C.; Maritan, A.; Barabási, A.L. A universal model for mobility and migration patterns. *Nature* **2011**, *484*, 96–100. [CrossRef] [PubMed]
43. Vij, A.; Shankari, K. When is big data big enough? Implications of using GPS-based surveys for travel demand analysis. *Trans. Res. Part C Emerg. Technol.* **2015**, *56*, 446–462. [CrossRef]
44. Batty, M.; Mackie, S. The calibration of gravity, entropy and related models of spatial interaction. *Environ. Plann. A* **1972**, *4*, 205–233. [CrossRef]
45. Wu, C.B.; Sitter, R.R. A model-calibration approach to using complete auxiliary information from survey data. *J. Am. Stat. Assoc.* **2001**, *96*, 185–193. [CrossRef]
46. Roy, J.R.; Thill, J.-C. Spatial interaction modelling. *Pap. Reg. Sci.* **2004**, *83*, 339–361. [CrossRef]

47. Shenzhen Statistical Yearbook 2012. Available online: http://www.sztj.gov.cn/nj2012/indexeh.htm (accessed on 12 September 2016).
48. Yang, Y.; Du, Z.; Hua, T. Research on Trade Areas in Other Cities of Pearl River Delta. Available online: http://www.pishu.com.cn/skwx_ps/literature/628924.html (accessed on 19 January 2017).
49. Calabrese, F.; Diao, M.; Di Lorenzo, G.; Ferreira, J.; Ratti, C. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Trans. Res. Part C Emerg. Technol.* **2013**, *26*, 301–313. [CrossRef]
50. Becker, R.; Cáceres, R.; Hanson, K.; Isaacman, S.; Ji, M.L.; Martonosi, M.; Rowland, J.; Urbanek, S.; Varshavsky, A.; Volinsky, C. Human mobility characterization from cellular network data. *Commun. ACM* **2013**, *56*, 74–82. [CrossRef]
51. Fotheringham, A.; O'Kelly, M.E. *Spatial Interaction Models: Formulations and Applications*; Kluwer Academic Pub.: Boston, MA, USA, 1989; Volume 5.
52. Huff, D.L. A probabilistic analysis of shopping center trade areas. *Land Econ.* **1963**, *39*, 81–90. [CrossRef]
53. Huff, D.L. Defining and estimating a trading area. *J. Mark.* **1964**, *28*, 34–38. [CrossRef]
54. Kim, P.J.; Kim, W.; Chung, W.K.; Youn, M.K. Using new huff model for predicting potential retail market in South Korea. *Afr. J. Bus. Manag.* **2011**, *5*, 1543–1550.
55. Mitchell, A. *The ESRI Guide to GIS Analysis: Geographic Patterns & Relationships*; ESRI, Inc.: Redlands, CA, USA, 1999; Volume 1.
56. Strehl, A.; Ghosh, J.; Mooney, R. Impact of similarity measures on web-page clustering. In Proceedings of the Workshop on Artificial Intelligence for Web Search (AAAI 2000), Austin, TX, USA, 30 July–1 August 2000; pp. 58–64.
57. Haykin, S.S. (Ed.) *Kalman Filtering and Neural Networks New York*; Wiley: New York, NY, USA, 2001; pp. 221–269.
58. Gold, C.M. The meaning of "neighbour". In *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*; Springer: Heidelberg, Germany, 1992; pp. 220–235.