sensors

MDPI

Article

A Data Cleaning Method for Big Trace Data Using Movement Consistency

Xue Yang [1], **Luliang Tang** [1,*], **Xia Zhang** [2] **and Qingquan Li** [3]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China; yangxue@whu.edu.cn

[2] School of Urban Design, Wuhan University, Wuhan 430070, China; xiazhang@whu.edu.cn

[3] College of Civil Engineering, Shenzhen University, Shenzhen 518060, China; liqq@szu.edu.cn

* Correspondence: tll@whu.edu.cn; Tel.: +86-139-9568-3555

Received: 31 January 2018; Accepted: 6 March 2018; Published: 9 March 2018

Abstract: Given the popularization of GPS technologies, the massive amount of spatiotemporal GPS traces collected by vehicles are becoming a new kind of big data source for urban geographic information extraction. The growing volume of the dataset, however, creates processing and management difficulties, while the low quality generates uncertainties when investigating human activities. Based on the conception of the error distribution law and position accuracy of the GPS data, we propose in this paper a data cleaning method for this kind of spatial big data using movement consistency. First, a trajectory is partitioned into a set of sub-trajectories using the movement characteristic points. In this process, GPS points indicate that the motion status of the vehicle has transformed from one state into another, and are regarded as the movement characteristic points. Then, GPS data are cleaned based on the similarities of GPS points and the movement consistency model of the sub-trajectory. The movement consistency model is built using the random sample consensus algorithm based on the high spatial consistency of high-quality GPS data. The proposed method is evaluated based on extensive experiments, using GPS trajectories generated by a sample of vehicles over a 7-day period in Wuhan city, China. The results show the effectiveness and efficiency of the proposed method.

Keywords: data cleaning; big data; vehicle trajectory; movement consistency modeling

1. Introduction

Nowadays, big data are everywhere, from sensors that monitor traffic loads to the flood of tweets and Facebook 'likes'. Researchers use volume, velocity, variety, value, and veracity to characterize the key properties of those big data [1]. In contrast to volume, velocity, variety, and value, the fifth 'V' of big data, veracity, is increasingly recognized as a key dimension when making big data operational in various applications [2]. Big GPS trace data generated by vehicles also have the five 'V' characteristics [3] and provide us with an unprecedented window into the dynamics of urban areas [4–9]. However, the growing volume of spatial data brings significant challenges for the management of data processing. In addition, a large amount of low-quality data mixed in the raw dataset increases the uncertainty of knowledge mining. Therefore, data cleaning plays a crucial role in the research field of information science [10–12].

In this paper, we propose an efficient method for big trace data cleaning. On the basis of our previous work [12], the proposed method polishes the theory of GPS data cleaning with further development. To keep the consistency of moving objects, the entire trajectory is first partitioned into a set of sub-trajectories by movement characteristic points. Those characteristic points are extracted from trajectories based on the changes in the motion status. Then, GPS data are cleaned based on the similarities of GPS points and the movement consistency model of the sub-trajectory. The movement consistency model is built using the random sample consensus algorithm based on the high spatial consistency of high-quality GPS data. Moreover, the accuracy of cleaned data can be controlled by tuning the threshold of similarities of GPS data and the movement consistency model. The proposed method is evaluated based on extensive experiments, using GPS trajectories generated by a sample of vehicles over a 7-day period in Wuhan city, China. The results show the effectiveness and efficiency of the proposed method. In summary, the contributions of this research include: (1) High-quality GPS data can be extracted from the raw dataset using the proposed data cleaning method; (2) The method of vehicle movement consistency modeling is proposed using GPS trajectories; (3) The discussion of the relationship between the accuracy of the cleaned GPS data and the similarity threshold provides a possible way to extract GPS data based on the estimation accuracy; (4) The amount of GPS data can be compressed after data cleaning, which can greatly decrease the storage space, as well as the computing time.

2. Related Work

Based on previous studies, approaches to GPS time sequence data cleaning can be broadly classified into statistical/quantitative based and logical/constraint based [13–16]. The statistical/quantitative methods of vehicle GPS data cleaning have been widely applied to identify and clean GPS data as they are less susceptible to error stemming from sampling intervals. For example, a high density of GPS points suggests a high probability that a road is present, whereas a low density indicates that vehicles deviate far away from the road [17–19]. Therefore, low-density points are defined as outliers. To remove these outliers, several studies [17] have sorted all the data points in ascending order according to their distance from the median and chose 95% of the sorted data points as the experimental data. Kernel density was applied to compute the density of each GPS point and then all the low-density points were removed [18]. In [19], researchers proposed an adaptive density optimization method to automatically recognize outliers and then removed those outliers. In addition, an RGCPK (region growing clustering with prior knowledge) method was also proposed to delete outliers from raw traces based on the motion tendency of vehicle traces [20]. However, the existing methods proposed in articles [17–20] still cannot remove outliers mixed in the high-density points cluster.

Trajectory filtering is a typical example of the logical/constraint methods for GPS data quality management. This method has been applied to improve GPS data position accuracy, including adaptive Kalman filtering for INS/GPS [21] and particle filtering. Authors in [22] argued that they can apply various filtering techniques to a trajectory to smooth the noise and potentially decrease the error in the measurements. They also gave a detailed introduction on how to implement a filtering algorithm to smooth a trajectory using the measuring position, speed, and heading of a GPS tracking point in a trajectory. However, the kinds of methods such as Kalman filter and particle filter have the shortcomings of high complexity and computational overhead [22].

Unlike previous approaches that clean GPS trajectories based on clustering or filtering algorithms, we propose a GPS data cleaning approach through the adjustment of movement consistency of GPS data. The following section on the data cleaning model describes our method for raw GPS data cleaning. Subsequent sections discuss the experimental results and conclusions.

3. Preliminaries

3.1. Spatial Big Data: Vehicle GPS Data

Vehicle GPS data as a major part of spatial big data [23] record the position, gathering time, heading, speed, and other movement attributes of moving objects. In general, the sampling rates of those vehicle trajectories usually range from 1 s to 60 s or even longer. The GPS data cleaning method proposed in this paper focuses on finding high-quality GPS data, also known as high-accuracy GPS data, from the raw GPS database. In reality, the position accuracy of the GPS data is different because of the types of GPS receivers, collection environment, techniques (e.g., single-point positioning, precise point positioning, and difference positioning), etc. For instance, the position accuracy of raw GPS traces collected by taxis using single-point positioning technique is about 10–15 m in Wuhan, while raw GPS data generated by smartphone applications on some mobile phones have a 3–5 m accuracy. Beyond that, the accuracy of GPS data collected by the same GPS receiver also displays a difference in different environments (e.g., open area, semi-sheltered area, and sheltered area) [24,25]. At the same time, because of the influence of the error distribution law of the GPS data [26], the accuracy of each GPS point of the trajectory is likely to be different. For instance, if the accuracy of a GPS dataset is about 10 m, the accuracy of one part of such GPS data is higher than 10 m, while another portion of the GPS data shows a lower accuracy. Therefore, a raw crowdsourcing GPS database has both low-accuracy and high-accuracy GPS data; a trajectory in such a database also has both low-accuracy and high-accuracy GPS data. Although most commercial GPS receivers usually implement strong filtering techniques to obtain very smooth tracking results, a considerable amount of crowdsourced GPS data generated by low-end GPS devices are still spotty.

3.2. Discussion: Movement Consistency of Vehicle GPS Data

The GPS data records the movement of moving objects; the higher the accuracy of the GPS data, the more realistic is the moving pattern it describes. As we know, in the real world, vehicles always keep moving in a straight direction except for changing lanes or turning at intersections. Therefore, trajectories generated by those vehicles show a very smooth result when its accuracy is high, as shown in Figure 1. Figure 1a,b respectively illustrate the DGPS (Differential Global Positioning System) data with 0.1 m accuracy and its synchronous GPS data with 10 m accuracy collected by a mapping car. The model of the GPS receiver and base receiver are Trimble_R9 and NetR9, respectively. The ground truth of one of the trajectories is obtained by field measurements. As we can see from Figure 1a, the DGPS data truly reflect the movement of the mapping car; however, the GPS data cannot paint the true path of the mapping car because of the interference of some low-accuracy GPS points. Meanwhile, by comparing with the ground truth, we found that the positions of GPS points vacillate around the ground truth and some high-accuracy GPS points keep a high consistency in position and direction, as shown in Figure 1a. The DGPS data, by contrast, show a very smooth result, and most of the points are either very near to or are at the ground truth.

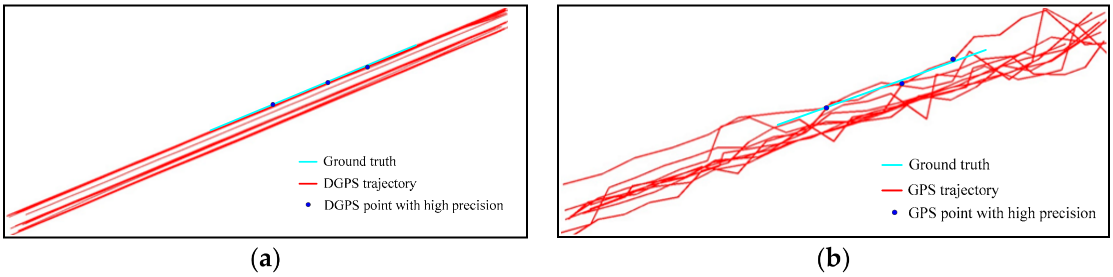


Figure 1. Movement consistency and position accuracy of the GPS data. (**a**) Differential Global Positioning System (DGPS) data overlay with the ground truth; (**b**) GPS data overlay with the ground truth.

Through the comparative results above, the high-accuracy GPS points of the trajectory present a high consistency of the movement. Based on this observation, the key techniques of GPS data cleaning from the raw crowdsourced database are to construct the consistency model of GPS points based on such consistency of high-accuracy GPS data.

4. GPS Data Cleaning Method Based on Movement Consistency

4.1. Overview

According to the analysis discussed in the previous section, the data cleaning method proposed in this paper has two steps: trajectory segmentation and movement consistency modeling, as shown in Figure 2.

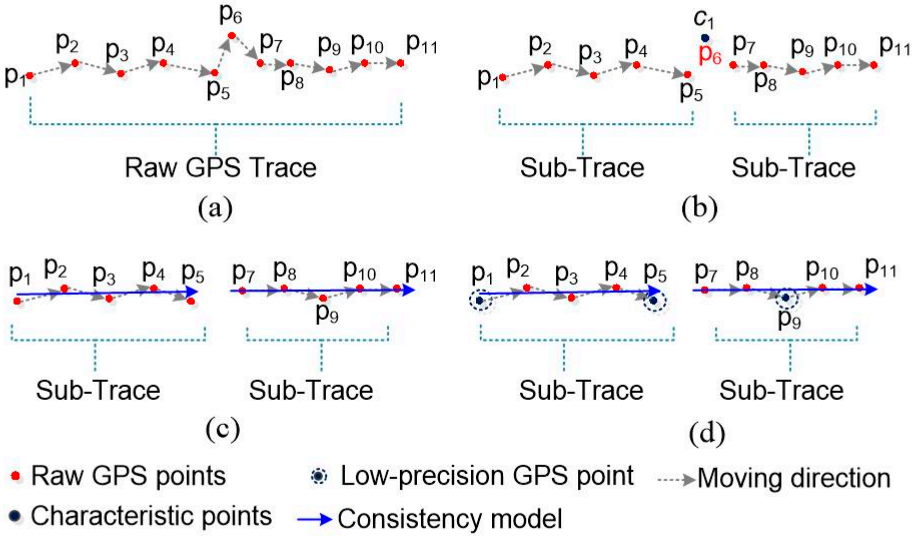


Figure 2. The methodology of GPS data cleaning: (**a**) the raw GPS trace data; (**b**) how the sub-traces are partitioned using a characteristic point p_6; (**c**) the consistency model generated for each sub-trace; and (**d**) the results of cleaning based on the similarity between the consistency model and sub-trace points.

Step 1. The whole trajectory is partitioned into a set of sub-trajectories based on the movement characteristic constraints, as shown in Figure 2a,b. These split points, also called characteristic points, are the starting and ending points of each sub-trajectory.

Step 2. The movement consistency model of each sub-trajectory is constructed using the random sample consensus algorithm based on the high spatial consistency of high-quality GPS data, as shown in Figure 2c,d. The movement consistency model is regarded as the linear position reference for cleaning points; the more similar the GPS points are to the movement consistency model, the more precise are the GPS points.

This section presents a detailed introduction of each process.

4.2. Trajectory Segmentation Based on the Changes in Motion Status of Vehicles

Trajectory segmentation is a preparatory work in spatiotemporal data mining [27]. For example, Gonzales et al. [28] identified critical points in various GPS trajectories to perform their mode classification study. In general, the whole trajectory is divided into several sub-trajectories based on the movement characteristic constraints such as position, time interval, velocity, etc. [29,30]. In this paper, we focus on GPS data cleaning based on the movement consistency. The cleaning rule of the proposed method is based on the premise that a moving object keeps moving on the same road in the same direction. Thus, trajectory segmentation aims to determine the characteristic points where the position or direction of a trajectory changes rapidly and then splits the trajectory base into the detected characteristic points.

4.2.1. The Principle of Trajectory Segmentation

The partitioning constraint factors in trajectory segmentation include position and angle, and are termed $verdis_{\mathrm{k}}$ and $angdis_{\mathrm{k}}$ in the following definitions:

Definition 1 (Position interference verdis$_{\mathrm{k}}$)**.** *Let $T_i = (p_1, p_2, \ldots, p_n)$ denote the trajectory of the object moving from p_1 to p_n. For any tracking points $p_k \in T_i$, $k = 1, 2, \ldots, n$, the vector composed by p_i and p_{i+1} presents move action, $i = 1, 2, \ldots, n$, and p_{i+2}' is the projection of p_{i+2} on the vector of p_i and p_{i+1}, then the distance between p_{i+2} and p_{i+2}' is called the position interference $verdis_k$, as shown in Figure 3.*

Definition 2 (Angle jamming angdis$_{\mathrm{k}}$)**.** *Let $T_i = (p_1, p_2, \ldots, p_n)$ denote the trajectory of an object moving from p_1 to p_n, as shown in Figure 4. For tracking points $(p_{i+1}, p_{i+2}) \in T_i$, $i = 1, 2, \ldots, n$, the vector composed by p_{i+1} and p_{i+2} is the present movement, the angle between $\boldsymbol{p_i p_{i+1}}$ and $\boldsymbol{p_{i+1} p_{i+2}}$ is the angle jamming value $angdis_k$, as shown in Figure 3.*

Definition 3 (Partitioning termination threshold a_1 and a_2)**.** *The variables a_1 and a_2 respectively represent the partitioning termination thresholds in distance and angle.*

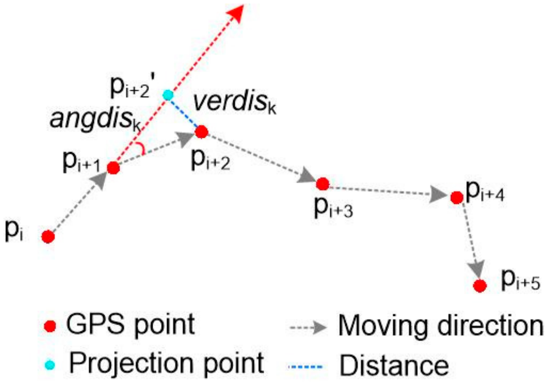


Figure 3. Trajectory partitioning based on position and angle constraints.

The main idea of trajectory segmentation for GPS data cleaning is to check the value of $verdis_{\mathrm{k}}$ and $angdis_{\mathrm{k}}$, $k = 1, 2, \ldots, n$, with respect to the present movement. This algorithm is introduced as follows:

Step 1. input the trajectory T_i $(p_1, p_2, p_3, \ldots, p_n)$;
Step 2. initialize the partitioning parameters' characteristic points C, c_1, *startIndex*, *currIndex*, *length*, a_1, and a_2, and set $c_1 = p_1$, $startIndex = 1$, $length = 1$;
Step 3. set $currIndex = startIndex + length$. If $currIndex < n$, go to Step 4; otherwise, go to Step 8;
Step 4. set $j = startIndex + 2$;
Step 5. calculate $verdis_{\mathrm{j}}$ and $angdis_{\mathrm{j}}$. If $verdis_{\mathrm{j}} > a_2 \mid\mid angdis_{\mathrm{j}} > a_1$, go to Step 6; otherwise, go to Step 3;
Step 6. push p_{j} into C and set $startIndex = j - 1$, $j = j + 1$;
Step 7. if $j < n$, go to Step 5; otherwise, go to Step 3;
Step 8. push p_n into C, and return C.

Based on this segmentation algorithm, a trajectory is divided into several sub-trajectories if any one of $verdis_{\mathrm{k}}$ and $angdis_{\mathrm{k}}$, $k = 1, 2, \ldots, n$, with respect to the present movement meet the partitioning termination thresholds a_1 and a_2. The sub-trajectories will be regarded as the basic unit for the remainder of the cleaning. It should be noted that the characteristic points are stored and managed separately from the sub-trajectories after cleaning since they could be used for trajectory compression or abnormal behavior detection.

4.2.2. Segmentation Threshold Determination

The distance and angle thresholds (a_1 and a_2) are used to determine whether the tracking point has departed from the centerline of the original route. In general, a GPS point is considered as a turning point if the vertical distance and angle between its two adjacent GPS vectors exceed the maximum width of the road or the minimal angle of the traffic turn in a city. These turning points could be considered as characteristic points that indicate the moving object has changed the moving route or direction. Beyond that, for different types of trajectories, different a_1 and a_2 values should be set for trajectory partitioning relative to their different shapes and unique characteristics. For a trajectory, the more complicated the shape, the more characteristic points are found in that partition. Thus, this study defines two deciding factors to determine the partitioning termination threshold for each trajectory. Especially, the first deciding factor is a global range of distance and angle in trajectory partitioning for all GPS data; and its value is decided by the knowledge of traffic law in a city. The second deciding factor is determined by the shape complexity of a trajectory. Both of those factors combine to determine a specific partitioning termination threshold for each trajectory as follows:

$$a_1 = \lambda_1 + g(\beta_1) \tag{1}$$

$$a_2 = \lambda_2 + g(\beta_2) \tag{2}$$

where λ_1 and λ_2 are the variables of the first deciding factor in the aspects of distance and angle, respectively. In our study, the values of λ_1 and λ_2 equate with the maximum width of the road and the minimum angle of the traffic turn in a city, respectively. The functions $g(\beta_1)$ and $g(\beta_2)$ demonstrate the relationship between the partitioning scale and the shape complexity of trajectories in the aspects of distance and angle, respectively. The variables β_1 and β_2 represent the shape complexity of a trajectory in aspects of distance and angle, respectively.

Given the significant inverse correlation between the shape complexity and the partitioning termination threshold of a trajectory, the higher the values of β_1 and β_2, and the lower the values of $g(\beta_1)$ and $g(\beta_2)$. Furthermore, the values of the partitioning termination thresholds $g(\beta_1)$ and $g(\beta_2)$ must be sensitive to the shape complexity β_1 and β_2 of trajectories within the set range. When the values of the shape complexity β_1 and β_2 of a trajectory are over a set range, then the partitioning thresholds $g(\beta_1)$ and $g(\beta_2)$ have less variation in distance and angle thresholds, respectively. Therefore, on this basis, combining the previous work for estimating the shape complexity of trajectories [31,32], the logarithmic function of the elementary function is used to model the relationship of β_1 and $g(\beta_1)$ and β_2 and $g(\beta_2)$ as follows:

$$g(\beta_1) = \log_a{}^{\beta_1} \tag{3}$$

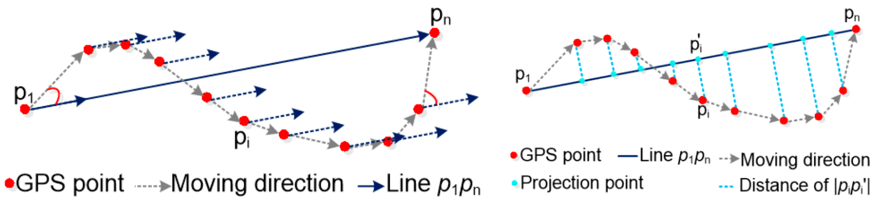
$$g(\beta_2) = \log_a^{\beta_2} \tag{4}$$

where 'a' is the base number of functions $g(\beta_1)$ and $g(\beta_2)$, $0 < a < 1$. To always keep the values of a_1 and a_2 positive, the absolute values of $g(\beta_1)$ and $g(\beta_2)$ must be smaller than the values of λ_1 and λ_2. Based on the research, the movement parameters are usually used to describe the complexity of trajectories [33]. In the movement feature set, classic descriptive statistics of movement parameters, which include the mean, standard deviation, and skewness of moving speed, the turning angle, and straightness index, are extracted from trajectories as basic movement features. In this paper, the standard deviations of projection distance and turning angle are used to represent the complexity of trajectory in position (β_1) and direction (β_2), respectively, as follows:

$$
\begin{aligned}
&\beta_2 = \sqrt{\tfrac{1}{n-2}\sum_{i=2}^{n-1}\left(\left|p_ip_i'\right| - \mu_{dis}\right)^2} \\
where& \\
&\mu_{dis} = \tfrac{1}{n-2}\sum_{i=2}^{n-1}\left(\left|p_ip_i'\right|\right) \\
&\left|p_ip_{i+1}\right| = \sqrt{(x_i - x_{i+1})^2 + (y_i - y_{i+1})^2} \\
&\left|p_1p_n\right| = \sqrt{(x_1 - x_n)^2 + (y_1 - y_n)^2}
\end{aligned}
\tag{5}
$$

$$
\begin{aligned}
&\beta_1 = \sqrt{\tfrac{1}{n-1}\sum_{i=1}^{n-1}\left(\angle\left(\overrightarrow{p_ip_{i+1}}, \overrightarrow{p_1p_n}\right) - \mu_{ang}\right)^2} \\
where& \\
&\mu_{ang} = \tfrac{1}{n-1}\sum_{i=1}^{n-1}\angle\left(\overrightarrow{p_ip_{i+1}}, \overrightarrow{p_1p_n}\right)
\end{aligned}
\tag{6}
$$

The complexity of the trajectory is positively associated with the values of β_1 and β_2, and higher values of β_1 and β_2 indicate a higher complexity of the trajectory in distance and angle. As shown in Figure 4, given a trajectory Tr_i = (p_1, p_2, . . . , p_n), p_i is the tracking point in the trajectory Tr_i and p_i = (x_i, y_i), i = 1, 2, . . . , n, $p_{1,}$ and p_n are respectively the start and end points of Tr_i. The variables β_1 and β_2 are computed by Equations (5) and (6), where $\angle(\underset{p_ip_{i+1}}{\rightarrow}, \underset{p_1p_n}{\rightarrow})$ represents the angle between the vector p_ip_{i+1} (denoted as $\underset{p_ip_{i+1}}{\rightarrow}$) and the vector p_1p_n (denoted as $\underset{p_1p_n}{\rightarrow}$), i = 1,2, . . . , n. At the same time, to avoid the extreme values of β_1 and β_2 by a looping trajectory, it is necessary to compare the positions of p_1 and p_n first. If the starting point p_1 overlaps with the ending point p_n, then p_n is replaced by p_{n-1}. This process is repeated from p_n to p_2 until a point is found that does not overlap with the starting point p_1.



(**a**) (**b**)

Figure 4. The complexity of trajectory in direction and position.

4.3. GPS Data Cleaning Based on Movement Consistency

4.3.1. The Consistency Model Construction for Each Sub-Trajectory

The sub-trajectory reflects the tendency of moving objects as they keep moving on the same road in the same direction. On this basis and combing through the discussions in Section 4.2, we find that high-accuracy vehicle tracking data are highly consistent with position and direction. For instance, tracking points of vehicles with high position accuracy always cluster together along the centerline of each lane, while also having similar headings. Thus, in this paper, we propose using this movement consistency to find high-quality GPS data from the raw GPS database. Specifically, the movement consistency model is defined as a directed line segment that belongs to the straight line l, as shown in Figure 5. Since a trajectory has been segmented into a set of sub-trajectories based on trajectory segmentation, GPS tracking points of each sub-trajectory keep with similar headings except for a few curves. Therefore, the construction of movement consistency models of each sub-trajectory equates with the generation of straight line l. At present, the least squares method is the most commonly used method for parameter estimation. However, the estimated parameters from a least squares model can be corrupted by outliers. To avoid the effects of these outliers, we use the Random Sample Consensus (RANSAC) algorithm to find GPS points with high consistency in position and heading and then get the consistency model of each sub-trajectory.

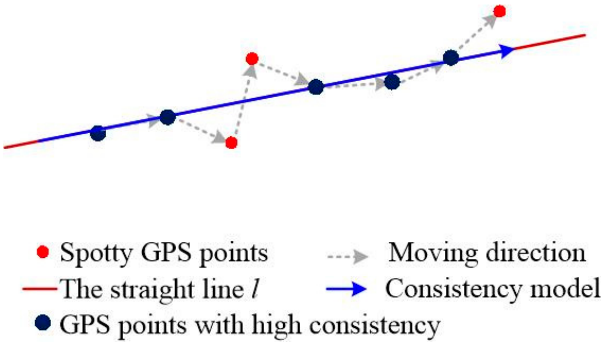


Figure 5. Construction of the consistency model by using RANSAC.

Given a sub-trajectory $STr_i = (p_i, p_{i+1}, \ldots, p_{i+t})$, $p_\mathrm{k} = (x_\mathrm{k}, y_\mathrm{k})$, $k = i, i + 1, \ldots, i + t$, $STr_i \in Tr_i$, assuming that the consistency model of STr_i belongs to the straight line l. Where x_0 and y_0 are the points that go through the consistency model, then b_0 and b_1 are the coefficients of the straight line l:

$$\begin{aligned} x &= x_0 + b_0 t \\ y &= y_0 + b_1 t \end{aligned} \tag{7}$$

The estimated model in the RANSAC algorithm is termed M^* and the same as Equation (7). The threshold τ defines a GPS tracking point pi and conforms to model M^*. The number of iterations is set as N and the parameter s is used to represent the number of data elements required to fit M^*. The concrete procedure for finding position points with high consistency using the RANSAC algorithm was obtained from [34].

4.3.2. Discussion of Similarity and Consistency Model for GPS Data Cleaning

The consistency model of each sub-trajectory is constructed based on the movement consistency of high-accuracy GPS points. Thus, for a sub-trajectory, the value of the similarity between a GPS point and its consistency model relates directly to the level of the position accuracy of it. In this study, the similarity evaluation between a GPS point and the consistency model in distance and direction is defined as Equation (8) by consulting the previous methods [35]:

$$sim_{(p_t,G)} = \omega_1 e^{-|p_t p_t'|} + \omega_2 e^{-(1-\cos(\theta_t))} \tag{8}$$

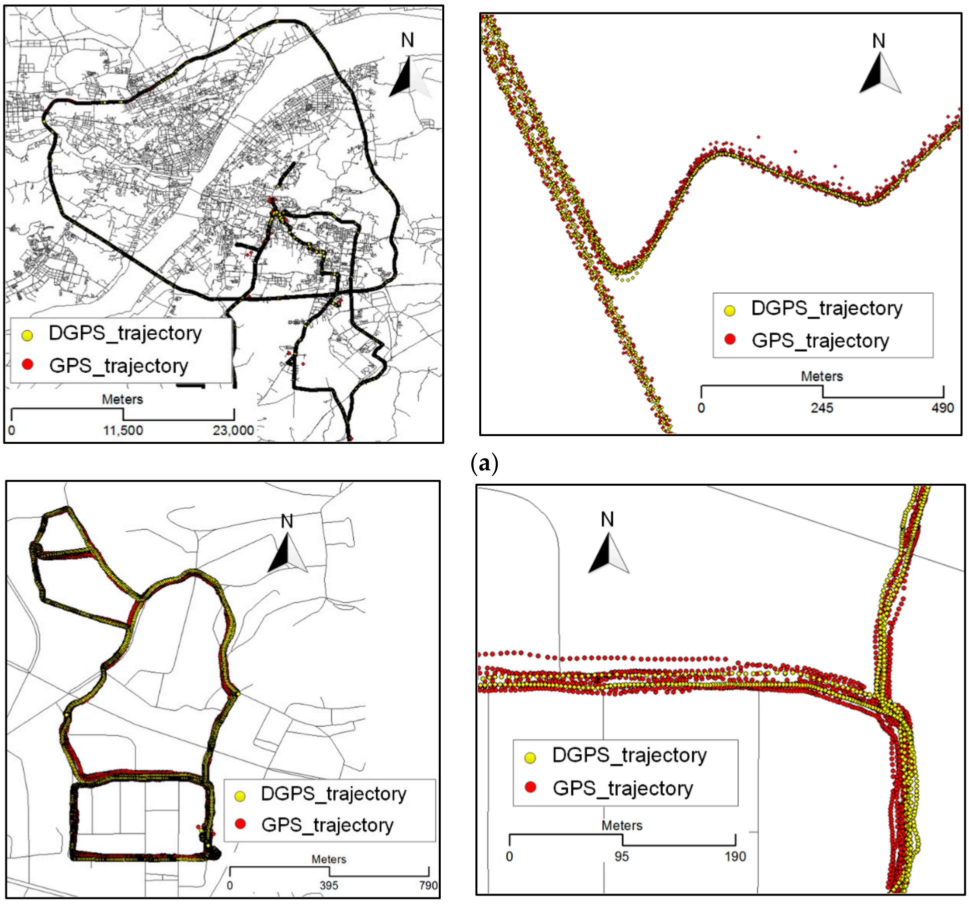
where $|p_t p_t'|$ is the distance between p_t and its projection point p_t' on the consistency model, θ_t is the angle between p_t' heading angle and the direction of the consistency model, ω_1 and ω_2 are the weight of the vertical distance and angle, $\omega_1 + \omega_2 = 1$. The similarity of GPS measurements and the consistency model range from 0 to 1.

Based on the results of similarity calculation, the high-quality GPS data are detected by setting different similarity thresholds. All cleaned GPS points will be joined back into a long trajectory and be used as raw material for information mining (e.g., road network generation, traffic flow detection, human mobility pattern mining, etc.). The similarity threshold determines the smoothness and quality of cleaned GPS points, and each similarity should correspond to an estimation accuracy of GPS data. However, since there are still many uncertainties in movement consistency construction, it is very difficult to obtain the definite relation between the similarity and the estimation accuracy of GPS data. In this paper, we use the relation of similarity (denoted as *Sim*) and position deviation (denoted as ε, also called an estimation accuracy) between GPS data and the ground truth to estimate the similarity threshold. The detailed analysis of the relation between *Sim* and *ε* by using GPS data in the real world is discussed in the next section.

5. Experimental Study

5.1. Experimental Dataset

To test the performance of our method, we experimented with real trajectory datasets. The experimental trajectory data were collected by several shuttle vehicles in Wuhan. These shuttle vehicles were equipped with the GPS logger (model: Trimble_R9), several smartphones (model: MDM6610, UBX-G6010-ST, MTK-MT6627, etc.), hand-GPS (model: SIRF systems), and an inertial measurement unit (model: POS310PCS) that recorded two kinds of traces, GPS and synchronized DGPS traces. It must be stressed that one GPS point corresponds to one DGPS point and all points represent the position of a moving object with different positional accuracies. The position accuracies of the GPS and DGPS data in an urban area were about 10–15 m and 0.05–0.1 m, respectively. The sampling rate for these data was 1 s. The time interval between two adjacent tracking points on a trajectory was not more than 360 s; otherwise, storing the trajectory was restarted from the position where it exceeded the set value. The data collection period for the shuttle vehicles was 7 days. We obtained about 140 million GPS and DGPS points, as shown in Figure 6.



(**a**)

(**b**)

Figure 6. Experimental data. (**a**) DGPS and synchronized GPS trajectories collected by IMU/DGPS systems and GPS loggers; (**b**) DGPS and synchronized GPS trajectories collected by IMU/DGPS systems, smartphones, and hand-GPS.

In our study, the highest accuracy of the cleaned data reached the meter level. The position accuracy of trajectories generated by the IMU/DGPS system reached the centimeter level. Therefore, in a follow-up experiment, the synchronized high-accuracy DGPS traces were regarded as ground truth to validate the effectiveness of the proposed method. The raw low-accuracy GPS data will be considered as the experimental data.

5.2. Parameters Discussion

The constants λ_1 and λ_2, and base 'a' for partitioning threshold determination are necessary for trajectory partitioning. Based on the above, the value of λ_1 equates with the maximum range of road width and λ_2 depends on the turning angle of vehicles in a city. The experimental traces data were collected in Wuhan. Based on the construction rule of the roads, the maximum width of the one-way road in the experimental region was about 17.5 m, so the value of λ_1 was set as 17.5 m. As the minimum angle of a traffic turn in China is about 60° and the heading error in the GPS data is about 5°–15°, we set the λ_2 to 45°. The value for base 'a' in Equations (1) and (2) ranges from 0 to 1 and affects the minimum and maximum values of $g(\beta_1)$ and $g(\beta_2)$. Based on Equations (1) and (2), the functions $g(\beta_1)$ and $g(\beta_2)$ have decreasing property with the value of trajectory complexity β_1 and β_2; and are less than zero if β_1 and β_2 are all greater than 1. To always keep the values of a_1 and a_2 as positive, the absolute minimum values of $g(\beta_1)$ and $g(\beta_2)$ must be smaller than the constants λ_1 and λ_2. Figure 7 shows the changing rules of $g(\beta_1)$ and $g(\beta_2)$ with the specific base under different values of β_1 and β_2. The base 'a' ranges between about 0 and 1.

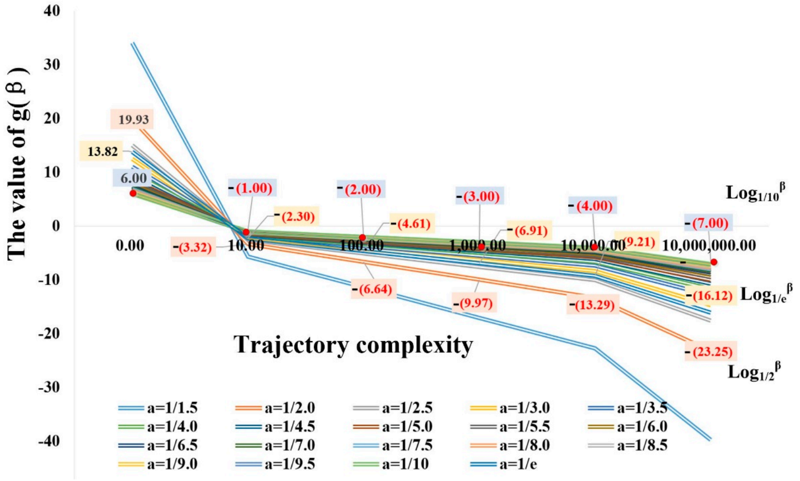
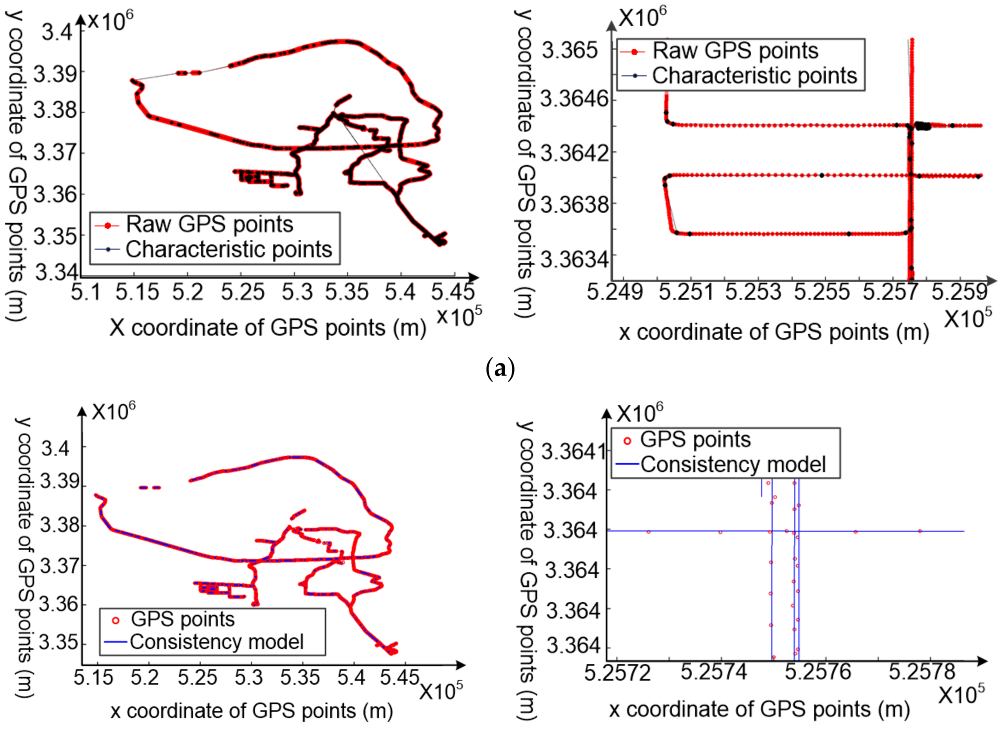


Figure 7. Discussion of base 'a' in Equation (2).

In Figure 7, the smaller the base 'a', the smaller will be the $g(\beta_1)$ and $g(\beta_2)$ change. As the constants λ_1 and λ_2 are equal to 17.5 m and 45° in this paper, base '10^{-1}' was selected as the value of 'a' in Equation (2). After trace partitioning (Figure 8a), the sub-trajectories are regarded as raw data and cleaned based on the movement consistency model. For consistency model construction, the value of τ was set as 0.1 m according to the accuracy requirement; other parameters such as N are self-adaptive (Figure 8b).



(**a**)

(**b**)

Figure 8. Experimental consistency model construction results for all vehicle movement data. (**a**) The consistency model construction results for the entire dataset; (**b**) the results of a part of the vehicle movement data.

A similarity evaluation model is used to calculate the similarity between GPS tracking points and the movement consistency model. This similarity evaluation model is used not only for estimating the similarity of GPS points and the consistency model but for cleaning threshold discussion. These two applications of similarity evaluation model are done to evaluate the similarity between GPS point and high-accuracy spatial reference in aspects of distance and angle. In this paper, the weight of distance and angle of the similarity evaluation model is estimated using the correlation between the distance and angle with measuring errors of GPS data [20]. The experimental results show that the weights in the similarity evaluation model are 0.91 and 0.09, respectively.

We use the linear regression analysis of the similarity and the position deviation of GPS measurements to derivate the relation of *Sim* and ε. With the result of multiple linear regression analysis, the relation of similarity (*Sim*) and position deviation (ε) between GPS data and the ground truth fits an exponential model, as shown in Equation (9):

$$Sim = ae^{b\varepsilon} + c \tag{9}$$

The values of parameters a, b, c in Equation (9) are determined by weights of the similarity evaluation model. The cleaning threshold with the specific estimation accuracy is obtained based on Equation (9). Based on plenty of experiment data and analyzing results, the correlation coefficient R for *Sim* and ε is about 0.942 when the values of a, b, c in Equation (9) for GPS data with 10–15 m accuracy are set as 1, −0.263, 0, respectively. Figure 9 shows the result of exponential regression of similarity and position accuracy of two different datasets collected in different environments with the same overall position accuracy. 'Dataset 1' and 'Dataset 2' were collected in an urban area on a shadowed road and a semi-shadowed road, respectively. The model of GPS receivers for collecting 'Dataset 1' and 'Dataset 2' were Trimble R9 and SIRF systems, respectively. The ground truths of these two datasets were obtained based on the CORS system by assembling the GPS receivers and CORS system together. Based on the similarity of Equation (9), we can get some cleaning thresholds by tuning the value of ε. Figure 10 shows the cleaned results of GPS points from raw GPS traces of two datasets, with its estimation accuracy set as 3 m; that is, ε equals to 3 m.

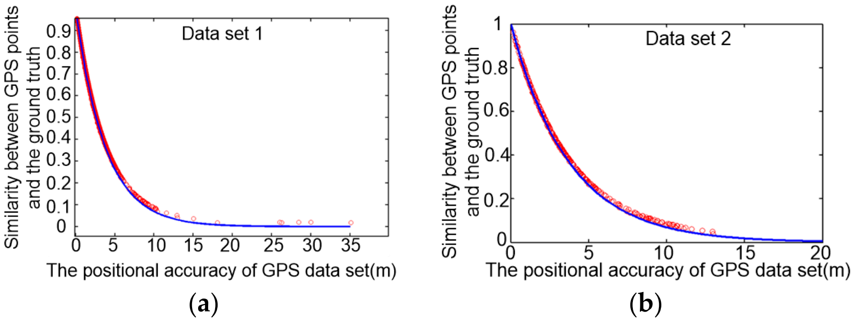


Figure 9. Linear regression results for similarity and position accuracy of the GPS data. (**a**) 'Dataset 1' (7604 GPS points) was collected in an urban area on a shadowed road; (**b**) 'Dataset 2' (6543 GPS points) was collected in an urban area on a semi-shadowed road.

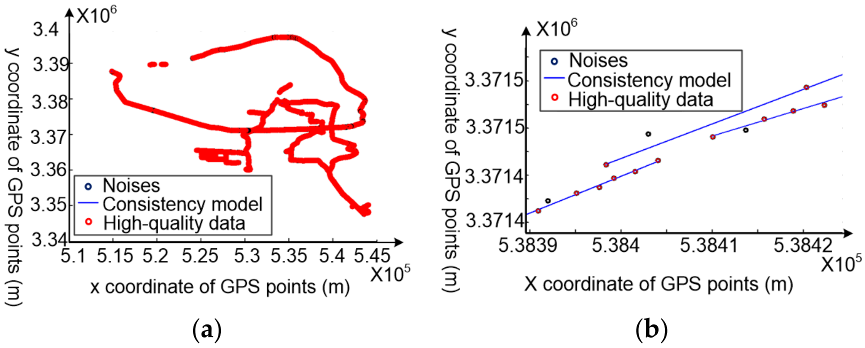


Figure 10. The cleaning results of all vehicle movement data at an estimation accuracy of 3 m. (**a**) The results of the entire dataset; (**b**) the results of part of the vehicle movement data.

5.3. Quantitative Evaluation and Discussion

To evaluate the effectiveness of the proposed method, we implemented it on the vehicle movement datasets collected in the real world. The position accuracy of those raw GPS data sets is different since the performance of the GPS devices varies. Based on field testing, the average value of the position accuracy of vehicle trajectories collected by Trimble R9, hand-held GPS, and smartphones are about 5.1 m (4.1), 5.0 m (3.6), and 9.1 m (4.7), respectively. The numerical values in parentheses are the standard deviations of each category. The raw datasets were then cleaned depending on different cleaning thresholds that were determined by the values of estimation accuracy. The experimental results for three different GPS datasets are displayed in Table 1. According to the figures given by Table 1, the accuracy and size of the cleaned GPS data are improved greatly compared with the accuracy of the raw dataset, though there is still a difference between the estimation accuracy and the real accuracy of the cleaned GPS data. In addition, based on the experimental results of three tested datasets, the accuracy of the cleaned GPS data also depends on the accuracy of the raw dataset itself. It is still a challenge for us to identify the high-accuracy GPS data from the raw datasets if there is no high-accuracy GPS data in the data in the first place. To further illustrate this point, we analyzed the distribution of accuracy for the cleaned data extracted from the vehicle trajectories collected by Trimble R9, as shown in Figure 11.

Table 1. Evaluation of the cleaned data from different datasets.

Trajectory Acquisition Device	Estimation Accuracy: ε (m)	Proportion of the Cleaned Data (%)	The Accuracy of GPS Data after Cleaning (Average Value/m)	The Accuracy of GPS Data after Cleaning (Standard Deviation/m)
Trimble R9	2	46.62	2.1	1.0
	3	58.87	2.9	1.2
	4	66.30	3.5	1.8
	5	72.48	4.1	1.8
Hand-held GPS	2	36.86	2.0	0.8
	3	41.38	2.4	1.2
	4	46.32	2.9	1.3
	5	48.76	3.7	2.3
Smartphones	2	27.43	3.8	2.4
	3	32.69	4.8	2.9
	4	40.23	5.1	3.0
	5	48.11	5.6	3.3

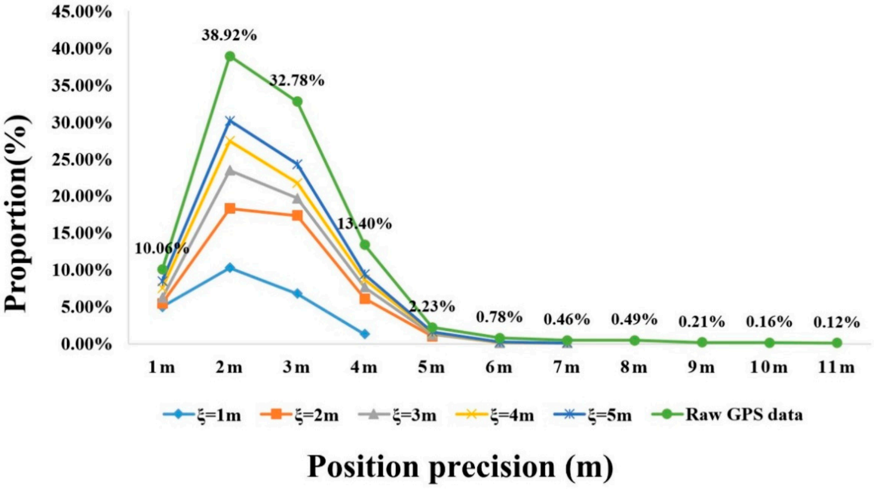


Figure 11. Comparison of the position accuracy of the cleaned data and raw GPS data in different estimation accuracy levels.

In Figure 11, the thick green solid line represents the proportion of raw GPS data in several ranges of position accuracy; the other solid lines show the proportion of cleaned data with different estimation accuracies. We observe that the proportion of GPS points that satisfy changing demands for position accuracy generally increase as the estimation accuracy falls. Although the average value and standard deviation of cleaned data based on the estimation accuracy illustrate that the proposed method is effective, a small percentage of low-position accuracy points beyond the estimation accuracy still exists in the cleaned dataset. For example, a very small subset of GPS points with 4 m accuracy is still mixed in the cleaned data when the estimation accuracy is set to about 1 m. Experimental results demonstrate that it is very difficult to find GPS data at 1 m position accuracy. The reason why the proposed method cannot strictly identify data based on the estimation accuracy is complex. The most important issue is that the GPS error follows a stable distribution; raw GPS points of a sub-trajectory include some high-accuracy points and low-accuracy points. The consistency model constructed using the RANSAC algorithm is considered as the position reference to identify the accuracy of GPS data but sometimes the position of the consistency model may be wrong, especially when there are only low-accuracy points in the sub-trajectory. In addition, the similarity threshold for cleaning is derived from the relation between GPS data and DGPS data, but there are still a lot of uncertainties caused by the collection environment, devices, techniques, etc. In the future work, we will address this problem.

To evaluate the performance of the proposed method, we conducted a qualitative comparison of position accuracy for cleaned data based on the methods discussed in the related work section (e.g., the RGCPK [20], the ADOM [19], the Kernel density method [18], and the Kalman filtering method [17]) and our method. These comparisons of the quality of cleaned data used datasets that were collected by vehicles equipped with Trimble R9. Table 2 shows the highest position accuracy results for the cleaned data from the test datasets using these methods.

Table 2. Comparisons of the previous methods for GPS data cleaning.

Methods for GPS Data Cleaning	Vehicle Trajectories Collected by Trimble R9	
	Mean Value of the Accuracy of the Cleaned Data (m)	Standard Deviation of the Accuracy of the Cleaned Data (m)
Method proposed in this paper	2.1	1.0
RGCPK	2.5	1.2
ADOM	4.5	3.2
KDE	4.6	3.3
KF	3.8	7.8

According to these results, the datasets employing the method proposed in this paper achieved the highest extracting accuracy when compared to the four other methods. Although RGCPK can also extract high-accuracy GPS data from the raw dataset, the results using RGCPK required prior knowledge to calculate the clustering threshold and filtering standard [20]. The comparison experiment also shows that methods such as ADOM and KDE (Kernel density method) can only remove low-density GPS points. However, sometimes the low-density GPS points do not equal the low-accuracy GPS points, so the cleaning effect is limited [18,19]. The KF (Kalman filtering method) is effective when the trajectory data are particularly noisy [17]. It is usually used to correct GPS data rather than to extract high-quality GPS data from raw datasets. Thus, the accuracy of the cleaned data derived using the filtering method was lowest in comparison with the other methods. Analyzing from practical applications (e.g., road network generation), the high-quality GPS data found from the raw datasets by using our method not only improves the position accuracy of road network extraction results but can also be used to detect lane-based road information.

6. Conclusions

Nowadays, the growing volume of spatial big data not only creates process management difficulties but also adds uncertainty for knowledge mining. Unlike previous approaches that clean GPS data based on clustering or filtering algorithms, in this paper, we proposed a method to clean GPS data through the adjustment of movement consistency of GPS data. The mechanism of vehicle GPS data cleaning based on movement consistency includes two steps: trajectory segmentation and consistency model construction. First, the whole trajectory is partitioned into a set of sub-trajectories by characteristic points. Those characteristic points are extracted from trajectories based on the constraints of moving distance or direction. Then, GPS data are cleaned based on the similarities of GPS points and the movement consistency model of the sub-trajectory. The movement consistency model is built using the random sample consensus algorithm based on the high spatial consistency of high-quality GPS data. Moreover, the accuracy of cleaned data can be controlled by tuning the threshold of similarities of GPS data and the local consistency model. The proposed method was evaluated based on extensive experiments, using GPS trajectories generated by a sample of vehicles over a 7-day period in Wuhan, China. Although these experimental results show the effectiveness and efficiency of the proposed method, there are still many problems and shortcomings that need further improving and refining. Due to the position accuracy of the raw GPS data being too low, the proposed method cannot find enough high-quality data from the original database according to the cleaning threshold, which is calculated by the estimation accuracy. In addition, in this paper, GPS data were collected with a high sampling rate by testing vehicles. In the real world, however, the sampling rate of most GPS data is not very high. Thus, this kind of sparse dataset also brings difficulty for data cleaning. In future work, we will address these shortcomings and continue to improve the filtering method proposed here.

Acknowledgments: The authors would like to sincerely thank the anonymous reviewers for their constructive comments and valuable suggestions to improve the quality of this article. This work was funded by the National Key R&D Program of China (2017YFB0503604, 2016YFE0200400), the National Natural Science Foundation of China (Nos. 41671442, 41571430, 41271442), and the Joint Foundation of Ministry of Education of China (6141A02022341).

Author Contributions: Xue Yang and Luliang Tang conceived and designed the algorithms of big trace data cleaning method presented in this paper. Xue Yang performed the experiments and wrote the paper. Xia Zhang and Qingquan Li contributed analysis tools and to the paper's refinement.

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

References

1. McAfee, A. Big data. The Management Revolution. *Harv. Bus. Rev.* **2012**, *90*, 61–67.
2. Saha, B.; Srivastava, D. Data quality: The other face of big data. In Proceedings of the 2014 IEEE 30th International Conference on Data Engineering (ICDE), Chicago, IL, USA, 31 March–4 April 2014; pp. 1294–1297.
3. Chatzimilioudis, G.; Konstantinidis, A.; Laoudias, C.; Zeinalipour-Yazti, D. Crowdsourcing with smartphones. *IEEE Internet Comput.* **2012**, *16*, 36–44. [CrossRef]
4. Zheng, Y.; Xie, X.; Ma, W.Y. GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory. *IEEE Data Eng. Bull.* **2010**, *33*, 32–39.
5. Van der Spek, S.; van Schaick, J.; de Bois, P.; de Haan, R. Sensing Human Activity: GPS Tracking. *Sensors* **2009**, *9*, 3033–3055. [CrossRef] [PubMed]
6. Tang, L.A.; Yu, X.; Gu, Q.; Han, J.; Jiang, G.; Leung, A.; Porta, T.L. A framework of mining trajectories from untrustworthy data in cyber-physical system. *ACM Trans. Knowl. Discov. Data* **2015**, *9*, 16. [CrossRef]
7. Rakthanmanon, T.; Campana, B.; Mueen, A.; Batista, G.; Westover, B.; Zhu, Q.; Zakaria, J.; Keogh, E. Addressing big data time series: Mining trillions of time series subsequences under dynamic time warping. *ACM Trans. Knowl. Discov. Data* **2013**, *7*, 10. [CrossRef]
8. Castro, P.S.; Zhang, D.; Chen, C.; Li, S.; Pan, G. From taxi GPS traces to social and community dynamics: A survey. *ACM Comput. Surv. (CSUR)* **2013**, *46*, 17. [CrossRef]
9. Tang, L.; Kan, Z.; Zhang, X.; Sun, F.; Yang, X.; Li, Q. A network Kernel Density Estimation for linear features in space—Time analysis of big trace data. *Int. J. Geogr. Inf. Sci.* **2016**, *30*, 1717–1737. [CrossRef]
10. Song, J.H.; Jee, G.I. Performance Enhancement of Land Vehicle Positioning Using Multiple GPS Receivers in an Urban Area. *Sensors* **2016**, *16*, 1688. [CrossRef] [PubMed]
11. Ertan, G.; Oğuz, G.; Yüksel, B. Evaluation of Different Outlier Detection Methods for GPS Networks. *Sensors* **2008**, *8*, 7344–7358.
12. Yang, X.; Tang, L. Crowdsourcing big trace data filtering: A partition-and-filter model. In Proceedings of the International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, Prague, Czech Republic, 11–19 July 2016; pp. 257–262.
13. Fan, H.; Zipf, A.; Fu, Q.; Neis, P. Quality assessment for building footprints data on OpenStreetMap. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 700–719. [CrossRef]
14. Berti-Equille, L.; Dasu, T.; Srivastava, D. Discovery of complex glitch patterns: A novel approach to quantitative data cleaning. In Proceedings of the 2011 IEEE 27th International Conference on Data Engineering, Washington, DC, USA, 11–16 April 2011; pp. 733–744.
15. Hellerstein, J.M. Quantitative Data Cleaning for Large Databases. White Paper, United Nations Economic Commission for Europe (UNECE). 2008. Available online: http://db.cs.berkeley.edu/jmh/ (accessed on 4 March 2018).
16. Bohannon, P.; Fan, W.; Geerts, F.; Jia, X.; Kementsietsidis, A. Conditional functional dependencies for data cleaning. In Proceedings of the 2007 IEEE 23th International Conference on Data Engineering, Istanbul, Turkey, 11–15 April 2007; pp. 746–755.

17. Cong, G.; Fan, W.; Geerts, F.; Jia, X.; Ma, S. Improving data quality: Consistency and accuracy. In Proceedings of the 33rd International Conference on Very Large Data Bases, Vienna, Austria, 23–27 September 2007; pp. 315–326.
18. Chen, Y.; Krumm, J. Probabilistic modeling of traffic lanes from GPS traces. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 81–88.
19. Wang, J.; Rui, X.; Song, X.; Tan, X.; Wang, C.; Raghavan, V. A novel approach for generating routable road maps from vehicle GPS traces. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 69–91. [CrossRef]
20. Tang, L.; Yang, X.; Kan, Z.; Li, Q. Lane-Level Road Information Mining from Vehicle GPS Trajectories Based on Naïve Bayesian Classification. *ISPRS Int. J. Geo-Inf.* **2015**, *4*, 2660–2680. [CrossRef]
21. Tang, L.; Yang, X.; Dong, Z.; Li, Q. CLRIC: Collecting Lane-Based Road Information Via Crowdsourcing. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 2552–2562. [CrossRef]
22. Mohamed, A.H.; Schwarz, K.P. Adaptive Kalman filtering for INS/GPS. *J. Geodesy* **1999**, *73*, 193–203. [CrossRef]
23. Jiang, Z.; Shekhar, S. *Spatial Big Data Science*; Springer International Publishing: New York, NY, USA, 2017.
24. Lee, W.; Krumm, J. Trajectory preprocessing. In *Computing with Spatial Trajectories*; Zheng, Y., Ed.; Springer: New York, NY, USA, 2011; pp. 3–33.
25. Gupta, M.; Ga1, J.; Aggarwal, C.; Han, J. Outlier detection for temporal data. *Synth. Lect. Data Min. Knowl. Discov.* **2014**, *5*, 129. [CrossRef]
26. Parkinson, B.W.; Enge, P.; Axelrad, P.; Spilker, J.J. Global Positioning System: Theory and Applications I. 1996. Available online: https://ci.nii.ac.jp/naid/10012561387/ (accessed on 4 March 2018).
27. Rasetic, S.; Sander, J.; Elding, J.; Nascimento, M.A. A trajectory splitting model for efficient spatio-temporal indexing. In Proceedings of the International Conference on Very Large Data Bases, Trondheim, Norway, 30 August–2 September 2005; pp. 934–945.
28. Gonzalez, P.A.; Weinstein, J.S.; Barbeau, S.J.; Labrador, M.A.; Winters, P.L.; Georggi, N.L.; Perez, R. Automating mode detection for travel behaviour analysis by using global positioning systems-enabled mobile phones and neural networks. *IET Intell. Transp. Syst.* **2010**, *4*, 37–49. [CrossRef]
29. Lee, J.; Han, J.; Li, X. Trajectory outlier detection: A partition-and-detect framework. In Proceedings of the 2008 IEEE 24th International Conference on Data Engineering Workshop, Cancun, Mexico, 7–12 April 2008; pp. 140–149.
30. Zhang, L.; Wang, Z. Trajectory Partition Method with Time-Reference and Velocity. *J. Converg. Inf. Technol.* **2011**, *6*, 134–142.
31. Fisher, N.I. *Statistical Analysis of Circular Data*; Cambridge University Press: Cambridge, UK, 1993.
32. Nams, V.O. Using animal movement paths to measure response to spatial scale. *Oecologia* **2005**, *143*, 179–188. [CrossRef] [PubMed]
33. Li, X. Using complexity measures of movement for automatically detecting movement types of unknown GPS trajectories. *Am. J. Geogr. Inf. Syst.* **2014**, *3*, 63–74.
34. Derpanis, K.G. Overview of the RANSAC Algorithm. *Image Rochester N. Y.* **2010**, *4*, 2–3.
35. Li, H.; Shen, I.F. Similarity measure for vector field learning. In Proceedings of the Advances in Neural Networks—ISNN 2006, Chengdu, China, 28 May–1 June 2006; pp. 436–441.

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).