applied sciences

MDPI

Article

A Novel K-Means Clustering Algorithm with a Noise Algorithm for Capturing Urban Hotspots

Xiaojuan Ran [1,2], Xiangbing Zhou [2,3,*], Mu Lei [4], Worawit Tepsan [1] and Wu Deng [2,5,*]

[1] International College of Digital Innovation, Chiang Mai University, Chaing Mai 50200, Thailand; xiaojuan_ran@cmu.ac.th (X.R.); worawit.tepsan@cmu.ac.th (W.T.)
[2] School of Information and Engineering, Sichuan Tourism University, Chendu 610100, China
[3] School of Computer Science and Technology, Aba Teachers University, Wenchuan 623002, China
[4] School of Information and Engineering, Chengdu University, Chendu 610106, China; mulei@cdu.edu.cn
[5] School of Electronic Information and Automation, Civil Aviation University of China, Tianjin 300300, China
* Correspondence: zhouxb@uestc.edu.cn (X.Z.); wdeng@cauc.edu.cn (W.D.)

Abstract: With the development of cities, urban congestion is nearly an unavoidable problem for almost every large-scale city. Road planning is an effective means to alleviate urban congestion, which is a classical non-deterministic polynomial time (NP) hard problem, and has become an important research hotspot in recent years. A K-means clustering algorithm is an iterative clustering analysis algorithm that has been regarded as an effective means to solve urban road planning problems by scholars for the past several decades; however, it is very difficult to determine the number of clusters and sensitively initialize the center cluster. In order to solve these problems, a novel K-means clustering algorithm based on a noise algorithm is developed to capture urban hotspots in this paper. The noise algorithm is employed to randomly enhance the attribution of data points and output results of clustering by adding noise judgment in order to automatically obtain the number of clusters for the given data and initialize the center cluster. Four unsupervised evaluation indexes, namely, DB, PBM, SC, and SSE, are directly used to evaluate and analyze the clustering results, and a nonparametric Wilcoxon statistical analysis method is employed to verify the distribution states and differences between clustering results. Finally, five taxi GPS datasets from Aracaju (Brazil), San Francisco (USA), Rome (Italy), Chongqing (China), and Beijing (China) are selected to test and verify the effectiveness of the proposed noise K-means clustering algorithm by comparing the algorithm with fuzzy C-means, K-means, and K-means plus approaches. The compared experiment results show that the noise algorithm can reasonably obtain the number of clusters and initialize the center cluster, and the proposed noise K-means clustering algorithm demonstrates better clustering performance and accurately obtains clustering results, as well as effectively capturing urban hotspots.

Keywords: K-means clustering; noise algorithm; unsupervised evaluation; non-parametric Wilcoxon statistical analysis; urban road planning; taxi GPS data

Citation: Ran, X.; Zhou, X.; Lei, M.; Tepsan, W.; Deng, W. A Novel K-Means Clustering Algorithm with a Noise Algorithm for Capturing Urban Hotspots. *Appl. Sci.* **2021**, *11*, 11202. https://doi.org/10.3390/app112311202

Academic Editor: Ricardo Colomo-Palacios

Received: 20 October 2021
Accepted: 21 November 2021
Published: 25 November 2021

1. Introduction

Modern cities have become important engines and hubs to drive social development. A city represents the most concentrated residence of people and the gathering place of social resources. Both work and life are inseparable from urban support. In recent years, there has been a "big city disease", of which the most prominent phenomenon is urban congestion, which has become a nearly unavoidable problem for almost every large-scale city. Consequently, from the perspective of informatization and intelligence, people have successively used information technology to put forward digital cities and smart cities from the strategic level and have formulated construction schemes to meet the development needs of different cities, hoping to solve the challenges faced in the process of urban development and alleviating urban congestion. In particular, the application of the new generation of cloud computing, big data, the Internet of Things, and artificial intelligence

technology has made urban operation more intelligent and has gradually become a reality, making rail transit and urban transportation more predictable and widely applied; however, a city is a densely populated area with a high concentration of both living and vehicle operation. The growth of the world's civil vehicle sales from 2010 to 2020 is shown in Figure 1.

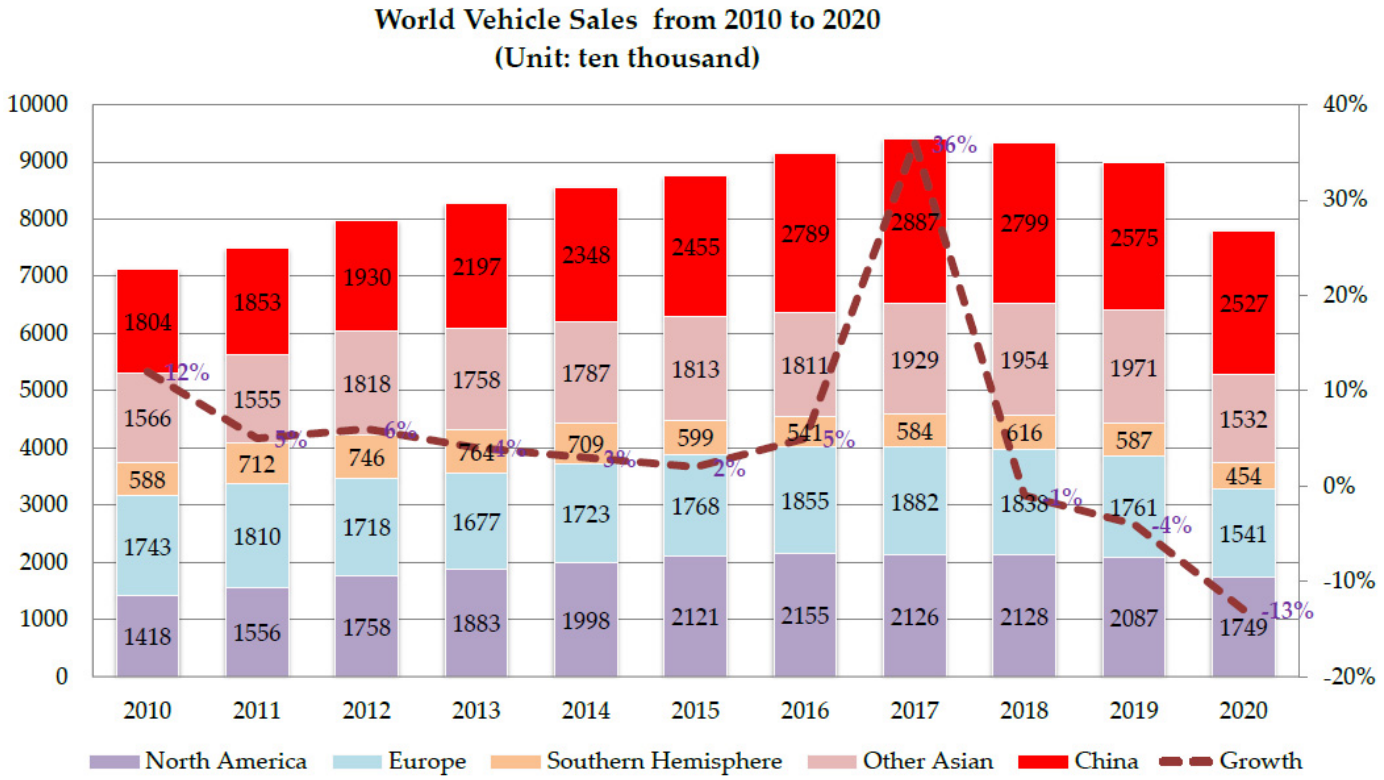

Figure 1. The growth of the world's civil vehicle Sales from 2010 to 2020.

Moreover, population flow is directly related to time, and urban congestion is still an important challenge for every city. The application of big data has served as a basic strategic digital resource in smart cities. Many researchers have analyzed the trajectory GPS data of transportation vehicles in order to mine the hidden information behind the data to reflect the urban operation status and define temporal and spatial change rules [1], in addition to use in traffic congestion status analysis [2–7], crowd movement distribution [8–10], traffic travel recommendation [11,12], and road planning [13,14], urban hotspot discovery [15–18], and so on. Such research results are directly applied to the construction of a smart city to elucidate more reasonable urban road planning and a more reasonable dispersion of vehicle flow and human flow. Such research methods usually use machine learning algorithms (such as cluster analysis and feature learning) to capture the vehicle trajectory patterns, including the origins and destinations (OD) [19–21], stops and moves (SM) [22,23], and moving objects (MO) [24,25] from the GPS data. Pongracic et al. [26] proposed a midlatitude Klobuchar correction model to correct the Klobuchar model for midlatitude users. Gu et al. [27] proposed a data-based methodology to estimate the traffic congestion of road segments between bus stops in order to improve the travel time reliability and quality of public transport services. Gao et al. [28] proposed a specific and accurate definition of traffic congestion to quantify the level of traffic congestion and constructed an image-based traffic congestion estimation framework based on a convolutional neural network. Afrin and Yodo [29] proposed a Bayesian network based on speed- and volume-related measures and a probabilistic congestion estimation approach. These models have been used to explain and discover urban operation states, crowd migration hotspots, and other urban operations.

In order to learn the valuable information hidden behind the location data, a clustering learning algorithm is a common and simple method that is used in many studies. Cluster analysis, also known as group analysis, is not only a statistical analysis method to study a classification problem (sample or index), but is also an important algorithm for data mining. Cluster analysis is composed of several patterns. Usually, a pattern is a vector of measurement or a point in multi-dimensional space. Cluster analysis is based on similarity. Patterns in a cluster have more similarity than patterns that are not in the same cluster. Clustering analysis algorithms can be divided into partition methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods. Among them, K-means clustering is the simplest, most used, and computationally efficient clustering algorithm, but it faces three major problems for any given dataset. As such, it is very difficult to find the appropriate number of clusters, optimize clustering centers, and capture global clustering results.

For the past several decades, many researchers have proposed some new ideas and methods. An improved K-means algorithm based on density (canopy K-means) has been proposed to solve the problem of determining the most suitable number of clusters and the best initial seeds [30]. An evolutionary K-means (EKM) method, based on combining K-means and a genetic algorithm, has been proposed to select parameters automatically through the evolution of partitions for solving the initiation problem of K-means [31]. The K-means++ algorithm has been proposed to quickly capture a better clustering center to find the sensitivity of the clustering process for the clustering center [32]. Fuzzy C-means (FCM) has been proposed to solve the problem of clustering edge data attribution [33–36]. An intelligent optimization algorithm with a K-means algorithm has been proposed to effectively solve the global optimization of clustering and the sensitivity of clustering center effectively [37,38]. In addition, some researchers have also proposed some improved methods for K-means algorithms and new application scenarios. For example, the depth representation extracted by depth learning technology has been proposed to improve the clustering performance of K-means clustering [39]. A competing cluster center approach has been proposed to maximize the benefits of cluster centers [40]. Ma and Zhou [41] proposed a novel sharing-based niche genetic algorithm with an initial population based on hybrid K-means clustering in order to obtain the best chromosome and perform K-means clustering. Sun et al. [42] proposed a framework to differentiate between these two types of methods with the following procedure.

These K-means clustering algorithms have adequately realized clustering and have obtained clustering results in actual engineering applications; however, some shortcomings still exist, such as a low processing efficiency, difficulty in determining the number of clusters, sensitively initializing the cluster center, and so on. In order to solve these problems, a novel K-means clustering algorithm based on a noise algorithm, namely, a noise K-means clustering algorithm, is developed here in order to improve the processing efficiency of automatic clustering and avoid both excessive manual configuration of parameter uncertainty and clustering results falling into local optimums in this paper. The noise algorithm is employed to randomly enhance the attribution of data points and output the results of clustering by adding noise judgment in order to automatically obtain the number of clusters for the given data and initialize the center cluster. Four unsupervised evaluation indexes of DB, PBM, SC, and SSE, and the nonparametric Wilcoxon statistical analysis method are employed to evaluate and analyze the clustering results and verify the distribution states and differences. Finally, five taxi GPS datasets, including Beijing, Chongqing, San Francisco, Rome and Aracaju, are selected to test and verify the effectiveness of the proposed noise K-means clustering algorithm in comparison with fuzzy C-means, K-means, and K-means plus approaches.

The innovations and main contributions of this paper are described as follows.

- A novel noise K-means clustering algorithm based on a noise algorithm is developed to capture urban hotspots.
- The noise algorithm is employed to randomly enhance the attribution of data points and output results of clustering by adding noise judgment to automatically obtain the number of clusters and initialize the center cluster.
- Four unsupervised evaluation indexes of DB, PBM, SC, and SSE are directly used to evaluate and analyze the clustering result.
- A non-parametric Wilcoxon statistical analysis method is employed to verify the distribution state and difference of clustering results.
- Comprehensive experiments are designed and executed to prove the effectiveness of the proposed noise K-means clustering algorithm with five sets of taxi GPS data.

2. Noise K-Means Clustering Algorithm

2.1. The Idea of the Noise K-Means Clustering Algorithm

A K-means clustering algorithm is an iterative clustering analysis algorithm that has been regarded as an effective means to solve urban road planning problems by scholars for the past several decades. The algorithm has been widely used in the fields of document classification, customer classification, ride data analysis, criminal network analysis, the detailed analysis of call records, and so on. It is very difficult to determine the number of clusters and sensitively initialize the cluster center. Noise can be used to simulate noise phenomena in nature. Because of its continuity, if an axis in two-dimensional noise is taken as the time axis, the result is a continuously changing one-dimensional function. As such, in order to solve the existing problems of the K-means clustering algorithm and make use of the merits of the noise algorithm, a novel noise-based K-means clustering algorithm is proposed to obtain a better clustering center and capture urban hotspots in this paper. The proposed noise-based K-means clustering algorithm consists of three parts. Firstly, the noise algorithm is employed to randomly enhance the attribution of data points and the output result of clustering by adding noise judgment in order to automatically obtain the number of clusters of the given data and initialize the cluster center. Secondly, the K-means clustering algorithm is employed to optimize the clustering center generated. It is fused with noise algorithm to form a novel noise-based K-means clustering algorithm. Finally, the proposed noise-based K-means clustering algorithm is used to obtain the clustering results for the given data and capture an excellent clustering center, i.e., and urban hotspot. Four unsupervised evaluation indexes of DB, PBM, SC, and SSE are directly used to evaluate and analyze the clustering results, and nonparametric Wilcoxon statistical analysis is employed to verify the distribution states and differences between clustering results.

2.2. The Flow of the Noise K-Means Clustering Algorithm

The flow of the proposed noise K-means clustering algorithm is shown in Figure 2.

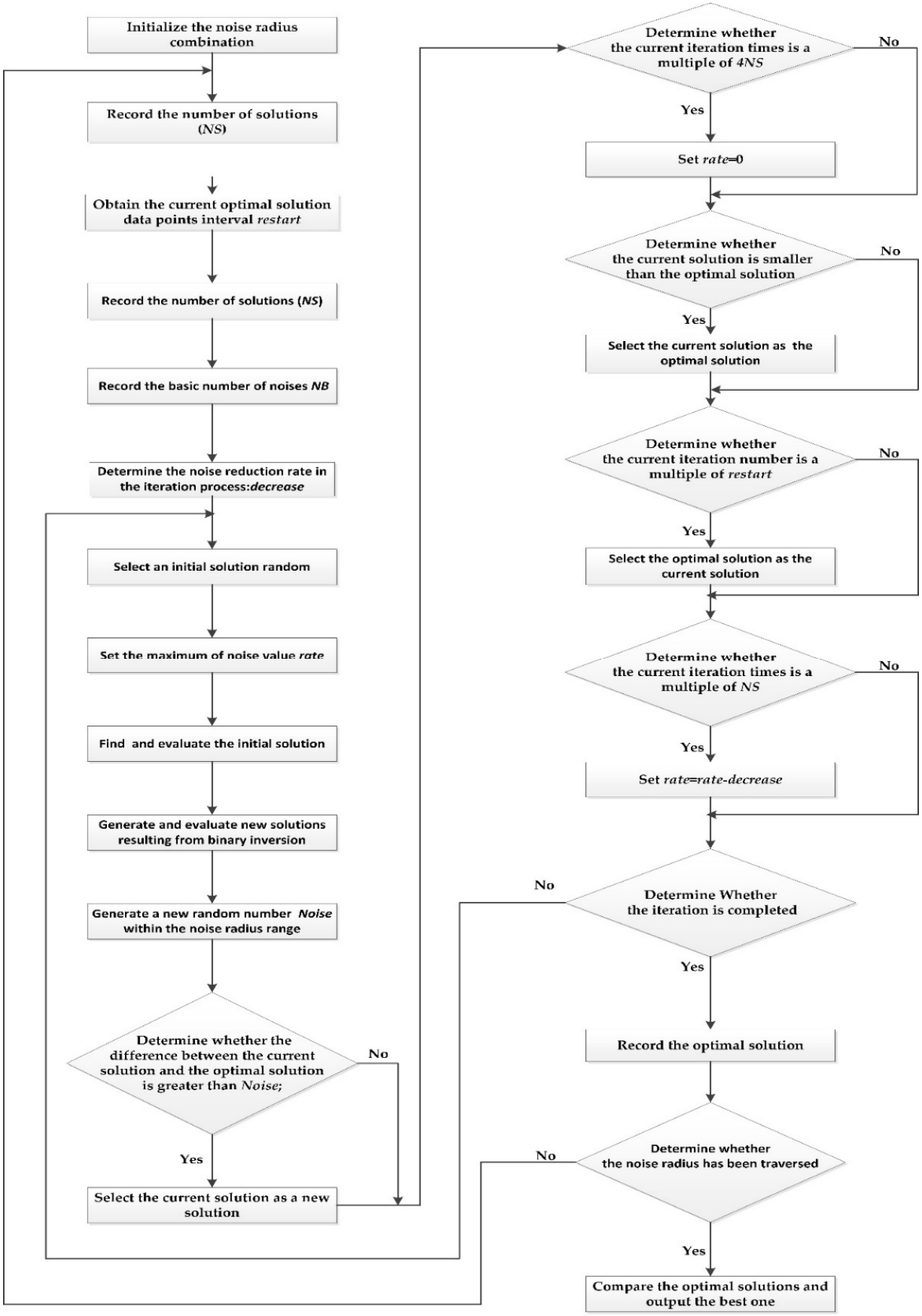

Figure 2. The flow of the noise-based K-means clustering algorithm.

2.3. The Realization of the Noise-Based K-Means Clustering Algorithm

The detailed steps of the noise-based K-means clustering algorithm are described as follows.

Step 1. Set the clustering number *K*

According to [43–46], it is generally believed that the clustering number of a K-means algorithm is between 2 and $\sqrt{N}$, where N represents the number of GPS data points. In this paper, more clusters are required to describe the distribution of urban hotspots. The GPS data points are intensive data, so the clustering number was set as $[\frac{\sqrt{N}}{2}, \sqrt{N}]$.

Step 2. Optimize the clustering number *K* using binary inversion

The clustering number K is converted from decimal to binary (the binary digits is rounded by $\sqrt{N}$). Then, one digit of the binary number is randomly flipped to generate a new binary number. Finally, the binary number is converted to a decimal and a new *K is obtained*. The binary inversion of the solution is shown in Figure 3.

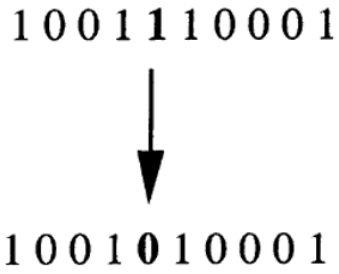

Figure 3. The binary inversion of noise.

Step 3. Optimize the clustering center

Since the location of the center point of the clustering is not fixed under the same number of clustering K values, in order to obtain a better initial distribution of the center point, the sum of squares for frror (SSE) is used to evaluate and find the optimal center of the clustering.

$$\mathrm{SSE} = \sum_{i=1}^{K} \sum_{j=1}^{N} \left(X_{ij} - \overline{X_i}\right)^2 \tag{1}$$

where K represents the number of clusters, N represents the number of taxi GPS data points in clustering, and $\left(X_{ij} - \overline{X_i}\right)^2$ represents the error sum of squares for each GPS data point.

Step 4. Output optimized clustering center and capture urban hotspots

According to the found optimal center of the clustering, it is set as the clustering center of the noise-based K-means clustering algorithm (Algorithm 1), which is used to solve the urban road planning problem in order to obtain the clustering results for the given taxi GPS data and thus capture an excellent center cluster, i.e., urban hotspots.

It can see that the complexity of the noise-based K-means clustering algorithm is related to the number of data points in each dataset. It is not directly related to the number of other noises and iteration times of the algorithm. As such, its time complexity is $\mathrm{O}(\mathrm{n}^2)$.

Algorithm 1. Noise-Based K-Means Clustering Algorithm

Input: Taxi GPS dataset and the number of taxi GPS data points, noise radius rate (which can also be generated randomly), the number of clustering iterations, and the clustering termination condition.
Output: Clustering results and new clustering center
1: Initialize the noise radius rate. Record the number of solutions (*NS*) of the current taxi GPS data.
2: Obtain the current optimal solution data points interval $restart = \sqrt{iter_{max} \times NS}$, and takes integers. $//iter_{max}$ represents the maximum number of iterations
3: Record the basic number of noises NB. Determine the noise reduction rate in the iteration process: $decrease = \frac{r_{max} - r_{min}}{\frac{iter_{max}}{NB} - 1}$.
$//r_{max}, r_{min}$ represents the maximum and minimum values of the noise radius respectively.
4: Select an integer between $[\frac{\sqrt{N}}{2}, \sqrt{N}]$ as the initial solution randomly, then evaluate it with SSE and denote it as the optimal solution.
5: Set the maximum of noise value rate. Determine whether the new solution generated by binary inversion is out of range $[\frac{\sqrt{N}}{2}, \sqrt{N}]$.
6: Generate a new random number within the noise radius range to produce *Noise*.
7: Select the current solution as the optimal solution when the difference between the current solution and the optimal solution is greater than *Noise*.
8: Set *rate* = 0 when the current iteration times is a multiple of *4NS*. Set the optimal solution as new solution when the current iteration number is a multiple of *restart*. Set *rate* = *rate-decrease* when the current iteration times is a multiple of *NS*.
9: Record the optimal solution and judge whether the iteration and the noise radius are completed.
10: Output the number of clusters and the initialization center of the given taxi GPS dataset.
11: Calculate the distance between the data point and the center point. Attribute the data points to the nearest cluster center according to the distance of the data points. Assign the data point average of each cluster as the new clustering center.
12: Calculate the SSE. Determine the termination condition of clustering.
13: Output the clustering result, which is the urban hotspots.

3. Clustering Process of the Noise K-Means Clustering Algorithm

3.1. Obtain the Clustering Number K Value and the Initial Center

The noise algorithm is used to obtain the clustering number K value and the initial center point in the given optimization objectives (such as SSE). The advantages and disadvantages of the optimal solution and the current solution are judged to join the noise, so that the data point attribution and clustering attribution output results have certain randomness. Moreover, the optimal solution is found from the current optimal solution at a certain interval of iterations, or the optimal solution is found by using "noise-free" at a certain interval of iterations. Finally, different combinations of maximum and minimum noise radius are set in terms of dataset in the algorithm: [−rate, rate] = [1, 0.9, 0.8, 0.7, 0.6,0.5, 1, 0.9, 0.8, 0.7, 0.6, 0.5, 1, 0.9, 0.8, 0.7, 0.6, 0.5, 1, 0.9, 0.8, 0.7, 0.6, 0.5, 1, 0.9, 0.8, 0.7, 0.6, 0.5; 0, 0, 0, 0, 0, 0, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4]/100.

3.2. Optimize Clustering Center

The K-means plus algorithm proposed by Arthur [32] is used to solve the sensitivity of K-means clustering center, which the computational complexity is O(log *K*). In order to obtain a better clustering number and clustering center, a K-means algorithm may be integrated with a noise algorithm to optimize the clustering center of the noise.

3.3. Obtain Clustering Result and Capture Excellent Cluster Center

K-means clustering is an unsupervised partition clustering algorithm. It takes distance as the standard of similarity measurement between data objects. That is, if the distance between data objects is smaller, their similarity is higher, and they are more likely to gather in the same class cluster. In this paper, Euclidean distance is used to calculate the distance between data objects and SSE is used to evaluate the clustering results.

4. Experiment Results and Analysis

4.1. Urban Taxi GPS Data

In order to verify the effectiveness of the proposed noise K-means clustering algorithm, five taxi GPS datasets at home and abroad are used as shown in Table 1. Taxi GPS data mainly refer to the vehicle position, direction, and speed information that is regularly recorded by the vehicle via the on-board global positioning system during travel. At present, many taxi GPS datasets exist for many cities across the world. The cities in this study, i.e., Aracaju (Brazil), San Francisco (USA), Rome (Italy), Chongqing (China), and Beijing (China), are representative cities for different continents and countries and are large-scale cities. The DB index [47], PBM index [48], SC (silhouette coefficient) [49] and SSE (sum of squares for error) are directly used to evaluate and analyze the clustering results. These evaluation methods are directly related to the number of clusters. Among them, DB index is mainly used to evaluate the performance of the noise K-means clustering algorithm. If the value of DB index is smaller, the similarity between clusters is lower and the clustering result is better. The PBM index is still used to evaluate the quality of clustering structure, where it describes clustering results and object attribution by defining quality. If the value of the PBM index is higher, the clustering effect is better. SC evaluates the clustering results by clustering the cohesion and separation. If the value of SC is greater, the clustering effect is better. SSE is used to evaluate the error probability distribution state and object attribution clustering. If the value of SSE is smaller, the clustering effect is better. The evaluation results and excellent procedures directly affect the effectiveness of urban hotspot discovery.

In Table 1, the GPS points are distributed in city areas and some are hotspots; however, it is very difficult to capture the hotspots from GPS datasets.

Table 1. Taxi GPS datasets.

Taxi GPS Dataset	Latitude and Longitude Region	Number of GPS Data Points
Aracaju (Brazil)	0.14 × 0.16	16,513
San Francisco (USA) [39]	0.10 × 0.10	21,826
Rome (Italy) [39]	0.35 × 0.50	20,254
Chongqing (China) [1]	0.60 × 0.36	19,149
Beijing (China) [40]	0.90 × 0.90	17,387

4.2. Experimental Environment and Parameter Setting

The experimental environment based on VMware featured the following: Intel Xeon E5-2658, dominant frequency 2 × 2.10 GHz with 8G RAM, Windows 2008 server, and the algorithm was coded in MATLAB 2016b. MATLAB is a commercial mathematical software produced by American MathWorks company(USA) which includes row matrix operation, drawing functions and data, implementing algorithms, creating user interfaces, connecting programs of other programming languages, and so on. It is used in data analysis, wireless communication, deep learning, image processing and computer vision, signal processing, robotics, control systems, and other fields. In our experiment, the alternative values were tested and modified for some functions to obtain the most reasonable initial values of these parameters. These selected values of the parameters take on the optimal solution and the most reasonable running time to efficiently complete the solving problem. The parameter settings of the FCM, K-means, K-means plus, and noise K-means are shown in Table 2. The number of clustering iterations was 200, and each algorithm ran 20 times independently. Generally, the evaluation results will be directly affected when the smaller clustering number of the comparison algorithm is set. FCM selected the clustering center according to the fuzzy parameters. K-means and K-means plus selected clustering center randomly. Noise-based K-means could obtain the clustering number and the initialization of the clustering center automatically. At the same time, the clustering numbers of noise-based K-means and FCM, K-means, and K-means plus were the same, and the clustering evaluation results of the corresponding clustering algorithm were also similar.

Table 2. The clustering numbers of the taxi GPS data.

Taxi GPS Dataset	The Clustering Number of FCM, K-Means, K-Means Plus	The Clustering Number of Noise K-Means
Aracaju (Brazil)	120	125
San Francisco (USA)	140	144
Rome (Italy)	135	137
Chongqing (China)	130	134
Beijing (China)	125	128

4.3. Experimental Results and Comparison Analysis

The comparison results of the maximum, average and minimum values of noise K-means, FCM, K-means and K-means plus under the evaluation of SC, BM index, DB index and SSE for the taxi GPS data are shown in Tables 3–6.

As can be seen from Tables 3–6, the proposed noise K-means clustering algorithm is used to obtain the clustering number of a given GPS dataset, which can improve the clustering effect effectively and much easier to find urban hotspots. It can also obtain the better clustering evaluation results, capture the urban hotspots, which can more effectively reflect the urban operating state through different clustering evaluation methods in the given GPS dataset. As can be seen from Table 3 for SC, the noise-based K-means clustering algorithm has a better performance, which indicates that it can better capture excellent clustering centers (urban hotspots). As can be seen from Table 4 for PBM index, the clustering results of noise K-means and K-means plus performed better in the taxis GPS data. As can be seen from Table 5 for DB index, K-means, noise K-means and K-means plus all performed well in the taxis GPS data, and there is little difference in the overall evaluation value. As can be seen from Table 6 for SSE, the noise K-means performs very well in 5 taxi GPS datasets, that because SSE is the optimization target in the clustering process of each clustering algorithm.

Table 3. Comparison results of SC evaluation values.

Taxi GPS Dataset	Algorithms	Maximum	Average	Minimum
Aracaju (Brazil)	Noise K-means	0.96083	0.95944	0.95703
	FCM	0.94635	0.94416	0.94285
	K-means	0.95827	0.95577	0.95355
	K-means plus	0.96004	0.95784	0.95586
San Francisco (USA)	Noise K-means	0.9369	0.93522	0.93314
	FCM	0.92914	0.92689	0.92494
	K-means	0.93389	0.93124	0.9287
	K-means plus	0.93502	0.93403	0.93219
Rome (Italy)	Noise K-means	0.9295	0.9275	0.9231
	FCM	0.91112	0.90662	0.90296
	K-means	0.9277	0.92535	0.92255
	K-means plus	0.92871	0.92672	0.92334
Chongqing (China)	Noise K-means	0.93896	0.93689	0.93457
	FCM	0.91979	0.91682	0.91311
	K-means	0.93673	0.93565	0.93424
	K-means plus	0.93708	0.9354	0.93195
Beijing (China)	Noise K-means	0.91192	0.90933	0.90561
	FCM	0.91979	0.91682	0.91311
	K-means	0.91125	0.91012	0.90853
	K-means plus	0.90963	0.90842	0.90668

Table 4. Comparison results of PBM index evaluation values.

Taxi GPS Dataset	Algorithms	Maximum	Average	Minimum
Aracaju (Brazil)	Noise K-means	0.03201	0.03108	0.02999
	FCM	0.01319	0.01286	0.01259
	K-means	0.03049	0.02827	0.02681
	K-means plus	0.03229	0.0308	0.02912
San Francisco (USA)	Noise K-means	0.01218	0.01172	0.01136
	FCM	0.01034	0.00958	0.00886
	K-means	0.01132	0.01041	0.00985
	K-means plus	0.01278	0.01193	0.01114
Rome (Italy)	Noise K-means	0.0236	0.022548	0.022006
	FCM	0.011832	0.010484	0.009661
	K-means	0.020837	0.019411	0.017145
	K-means plus	0.023151	0.022623	0.021651
Chongqing (China)	Noise K-means	0.0627	0.05794	0.05468
	FCM	0.03976	0.03731	0.03568
	K-means	0.06013	0.0542	0.05118
	K-means plus	0.06133	0.05753	0.0532
Beijing (China)	Noise K-means	0.06086	0.0598	0.05833
	FCM	0.03976	0.03731	0.03568
	K-means	0.06157	0.05877	0.05694
	K-means plus	0.06011	0.06036	0.05914

Table 5. Comparison results of DB index evaluation values.

Taxi GPS Dataset	Algorithms	Maximum	Average	Minimum
Aracaju (Brazil)	Noise K-means	0.08366	0.07992	0.07423
	FCM	0.17978	0.15106	0.13756
	K-means	0.08817	0.07815	0.06961
	K-means plus	0.0858	0.78368	0.06975
San Francisco (USA)	Noise K-means	0.09342	0.09039	0.08518
	FCM	0.15793	0.1459	0.12684
	K-means	0.10198	0.08553	0.08055
	K-means plus	0.09755	0.09082	0.08518
Rome (Italy)	Noise K-means	0.16462	0.10853	0.09498
	FCM	0.17961	0.15236	0.13591
	K-means	0.09967	0.09159	0.08619
	K-means plus	0.16606	0.10897	0.09622
Chongqing (China)	Noise K-means	0.11173	0.10038	0.09182
	FCM	0.17093	0.14996	0.13769
	K-means	0.10028	0.09	0.0847
	K-means plus	0.11207	0.10867	0.0922
Beijing (China)	Noise K-means	0.10447	0.09777	0.09215
	FCM	0.17093	0.14996	0.13769
	K-means	0.0925	0.08668	0.08281
	K-means plus	0.10459	0.09799	0.09305

The average running time for each algorithm when running 20 times is shown in Table 7.

As can be seen from Table 7, the average iteration time of the noise K-means clustering algorithm is longer, because the noise algorithm needs to capture the clustering number of a given dataset and initialize the clustering center. The noise K-means clustering algorithm can obtain the better clustering number of a given GPS data, and much easier to find urban hotspots. It can also obtain the better clustering evaluation results, capture the urban hotspots.

Table 6. Comparison results of SSE evaluation values.

Taxi GPS Dataset	Algorithms	Maximum	Average	Minimum
Aracaju (Brazil)	Noise K-means	19.3463	18.1769	17.4351
	FCM	28.7974	28.0926	27.3564
	K-means	21.4902	20.4493	18.193
	K-means plus	19.957	19.0552	17.4351
San Francisco (USA)	Noise K-means	34.8698	33.8601	33.1321
	FCM	41.1925	40.0136	38.6125
	K-means	39.1769	37.2797	35.5356
	K-means plus	35.6768	34.6172	34.0928
Rome (Italy)	Noise K-means	54.1587	52.2404	50.6664
	FCM	86.4097	82.9769	77.1727
	K-means	64.186	59.7377	56.1926
	K-means plus	54.7747	52.7017	51.1269
Chongqing (China)	Noise K-means	102.901	99.4066	95.5405
	FCM	151.343	144.9	135.923
	K-means	109.436	104.691	101.017
	K-means plus	108.112	101.987	99.5079
Beijing (China)	Noise K-means	151.343	144.9	135.923
	FCM	224.152	214.794	209.467
	K-means	223.885	220.258	216.065
	K-means plus	220.629	218.108	216.092

Table 7. Comparison of the average running time (s).

Taxi GPS Dataset	Clustering Algorithm	Average Running Time (s)
Aracaju (Brazil)	Noise K-means	8.03867
	FCM	21.7398
	K-means	2.3773
	K-means plus	2.7005
San Francisco (USA)	Noise K-means	10.86496
	FCM	32.4686
	K-means	3.232
	K-means plus	3.3807
Rome (Italy)	Noise K-means	9.87741
	FCM	29.5736
	K-means	3.1264
	K-means plus	3.1248
Chongqing (China)	Noise K-means	9.65121
	FCM	26.9652
	K-means	2.9121
	K-means plus	2.9302
Beijing (China)	Noise K-means	8.34432
	FCM	26.9652
	K-means	2.6924
	K-means plus	2.5276

The SSE convergence curves for the FCM, K-means, K-means plus, and noise-based K-means methods are shown in Figures 4–8.

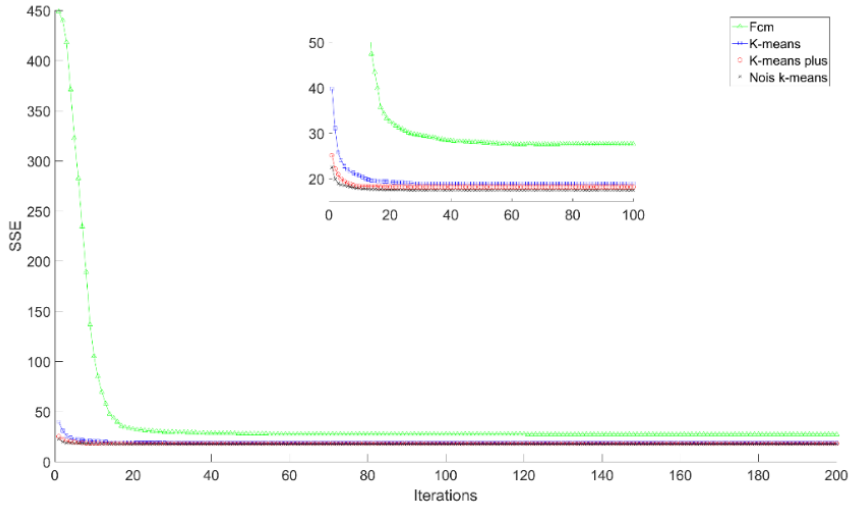

Figure 4. The SSE convergence curve for taxi GPS data (Aracaju (Brazil)).

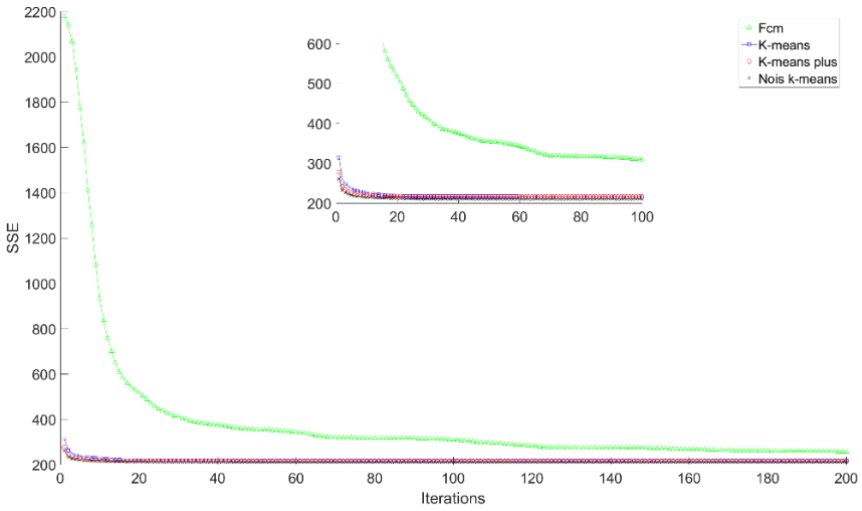

Figure 5. The SSE convergence curve for taxi GPS data (San Francisco (USA)).

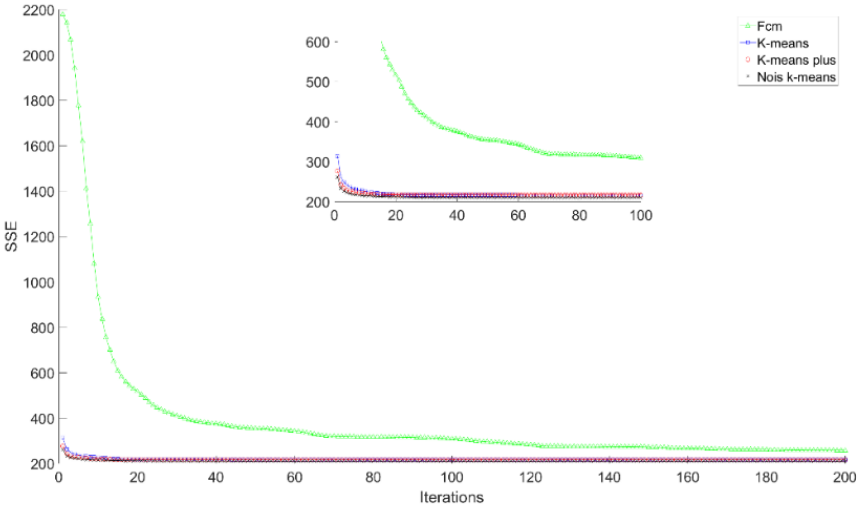

Figure 6. The SSE convergence curve for taxi GPS data (Rome (Italy)).

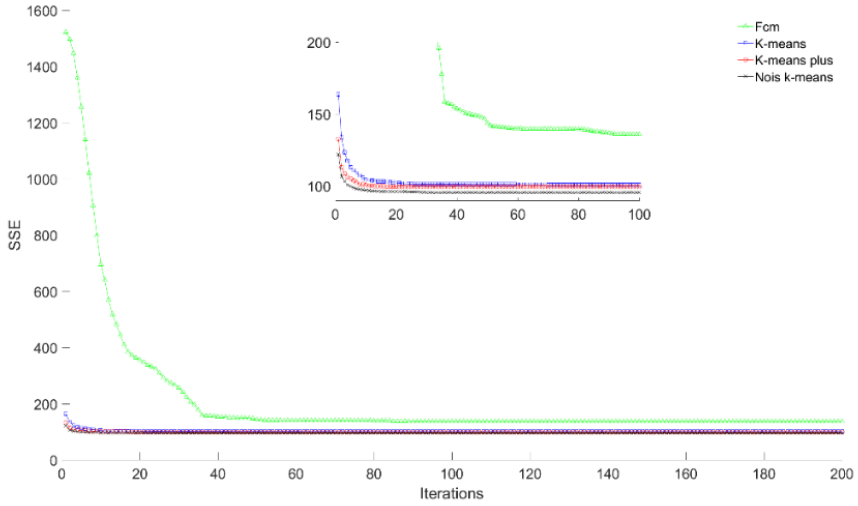

Figure 7. The SSE convergence curve for taxi GPS data (Chongqing (China)).

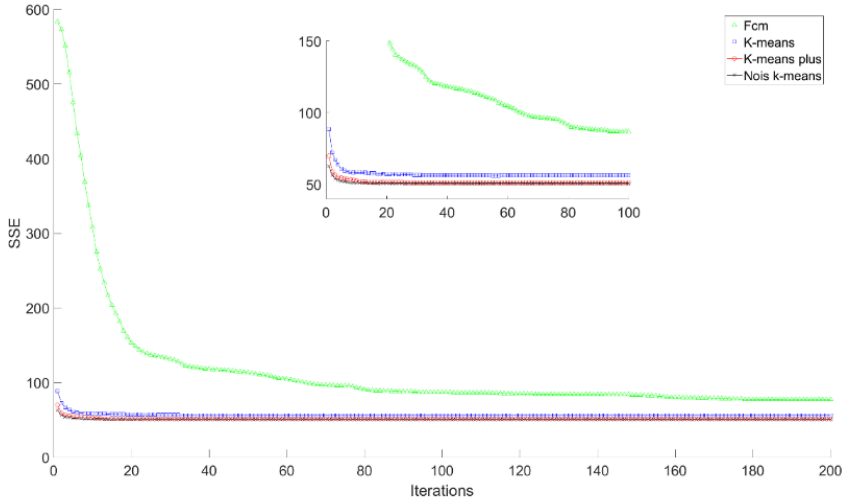

Figure 8. The SSE convergence curve for taxi GPS data (Beijing (China)).

As can be seen from Figures 4–8, all comparative clustering algorithms, except for FCM, could complete the convergence by iterating about 10 times, which shows that the proposed noise-based K-means clustering algorithm is feasible and suitable for basic partition clustering algorithms.

4.4. Visual Presentation of Urban Hotspots

In order to more intuitively display the capture of urban hotspots by clustering algorithm, a visual presentation in Amap system is used in this paper. That is, the captured cluster center location information is input into the map system through Amap API in order to realize the visual presentation of the captured urban hotspots. The obtained experiment results for the Aracaju (Brazil), San Francisco (USA), Rome (Italy), Chongqing (China), and Beijing (China) are shown in Figures 9–12.

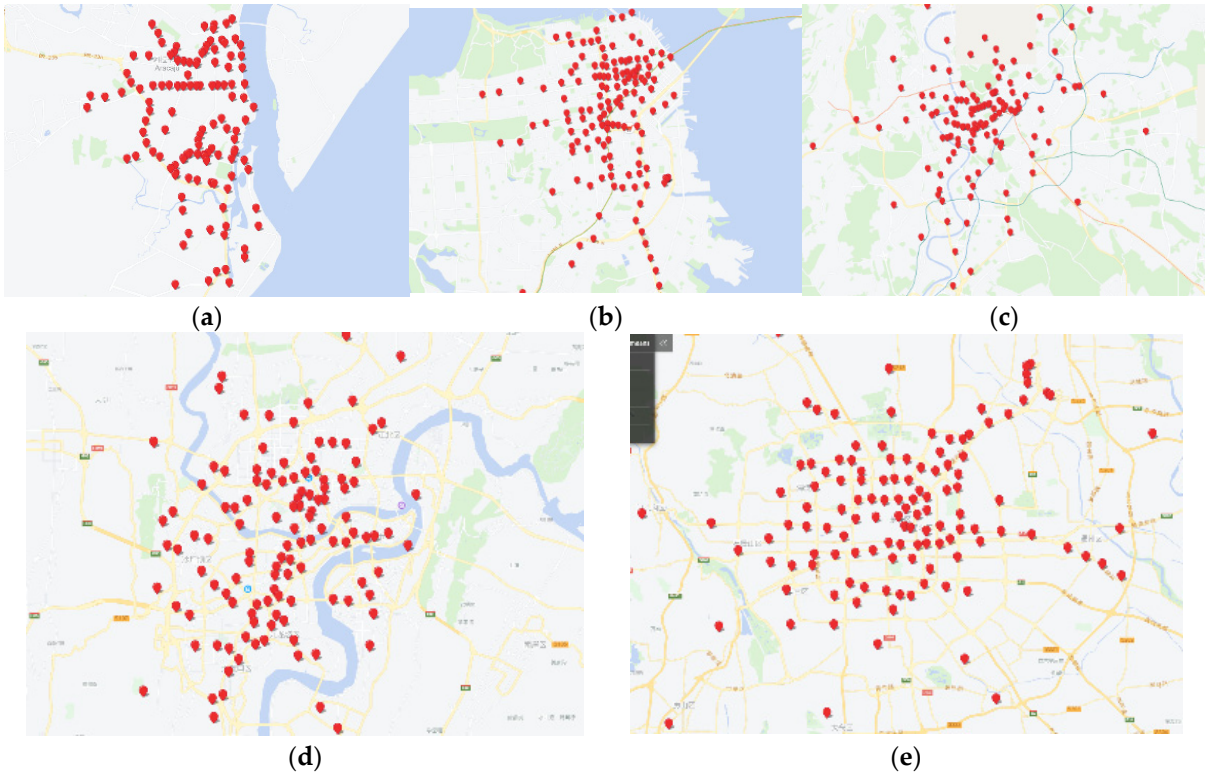

Figure 9. The city hotspot markers for the city taxi GPS data found by the FCM method: (**a**) Aracaju (Brazil), (**b**) San Francisco (USA), (**c**) Rome (Italy), (**d**) Chongqing (China), (**e**) Beijing (China).

Figure 10. The city hotspot markers for the city taxi GPS data found by the K-means method: (**a**) Aracaju (Brazil), (**b**) San Francisco (USA), (**c**) Rome (Italy), (**d**) Chongqing (China), (**e**) Beijing (China).

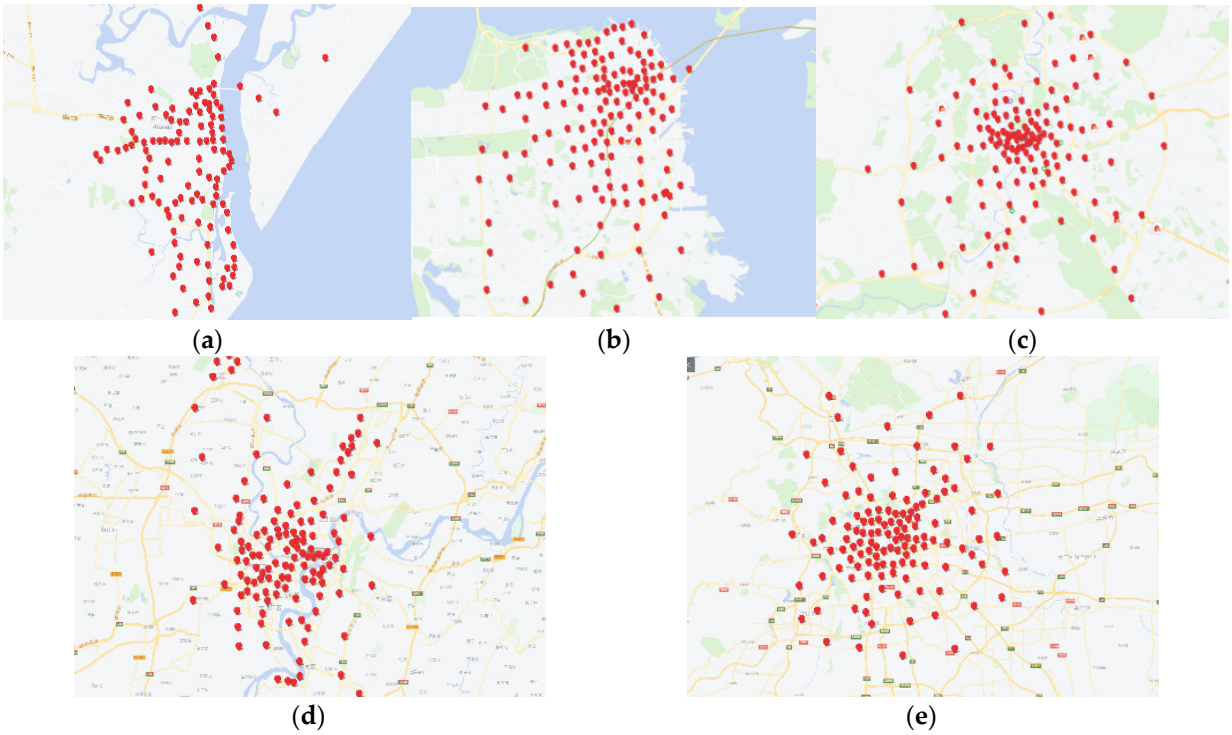

Figure 11. The city hotspot markers for the city taxi GPS data found by the K-means plus method: (**a**) Aracaju (Brazil), (**b**) San Francisco (USA), (**c**) Rome (Italy), (**d**) Chongqing (China), (**e**) Beijing (China).

Figure 12. The city hotspot markers for the city taxi GPS data found by the noise-based K-means method: (**a**) Aracaju (Brazil), (**b**) San Francisco (USA), (**c**) Rome (Italy), (**d**) Chongqing (China), (**e**) Beijing (China).

As can be seen from Figures 9–12, the spatial distributions of urban hotspots have been captured by the FCM, K-means and K-means plus clustering algorithms for the given taxi GPS data of Aracaju (Brazil), San Francisco (USA), Rome (Italy), Chongqing (China) and Beijing (China) is different. The aggregation degree of taxi operation varies significantly between different cities. From the city hotspot marking results of the taxis GPS data in the five cities obtained by FCM in Figure 9, it can be seen that the clustering effect is not ideal and that there are serious dispersion and local optimum phenomena. For example, the clustering results of San Francisco (USA) demonstrate local optimum phenomena and multiple hotspots are close together, which results in an uneven distribution of urban hotspots and deviation from urban hotspots. Thus, a potentially planned road cannot benefit many people with this approach. From the city hotspot marking results of the taxis GPS data in five cities obtained by K-means clustering in Figure 10, it can be seen that the clustering centers and numbers of clusters are not very reasonable and do not show optimal results. There are multiple local optimum phenomena, such as the serious phenomenon that multiple hotspots are close together in Rome (Italy) and Beijing (China). The phenomenon occurs such that some marker points are not clustered, so that a planned road cannot benefit a large number of people. As can be seen from the city hotspot marking results of the taxis GPS data in five cities obtained by K-means plus clustering in Figure 11, the obtained clustering effect is better than that obtained with FCM and K-means clustering. The clustering centers and number of clusters in most cities are optimal, but there is a local optimal phenomenon in Rome (Italy), which shows multiple hotspots together. A road planned with this approach could benefit many people. As can be seen from the city hotspot marking results of the taxis GPS data in five cities obtained by the noise-based K-means clustering in Figure 12, the obtained clustering effect is better than that obtained with FCM, K-means, and K-means plus clustering. The clustering centers and numbers of clusters for all five cities are the best here, and there are no groups of hotspots that are close together. The information reflected by these urban hotspots denotes shopping points, parks, stations, amusement areas, and other public places, which means that consequently planned roads benefit all people. In summary, the compared experiment results show that the noise-based K-means algorithm can reasonably obtain the number of clusters and initialize the cluster center, and the proposed algorithm shows better clustering performance and accurately obtains clustering results, in addition to effectively capturing urban hotspots.

5. Statistical Analysis of Wilcoxon

A Wilcoxon rank sum test is a non-parametric null hypothesis statistical testing method [50] that is often used to test the significant difference and distribution state of a clustering training process. When statistical validation is performed, it is usually composed of p, h and *stats*, in which p represents the results u and v of a clustering evaluation. p is the continuous distribution of data samples, which is used to test u and v under the non-null hypothesis (noise-based K-means vs. FCM, noise-based K-means vs. K-means, noise-based K-means vs. K-means plus). As $p \rightarrow 0$, the difference between u and v becomes more obvious. h is the logical value for testing 0 or 1, $h = 1$ means reject the null hypothesis, and $h = 0$ means reject the null hypothesis at α (for example, $\alpha = 0.05$, α is the significance level parameter, value range: $0 < \alpha < 1$). That is, $h = 1$ means that the difference between u and v is significant, while $h = 0$ means that the difference between u and v is not significant. *Stats* consists of two statistics, *zval* and *ranksum*. *zval* represents a normal distribution estimate of p, while *ranksum* represents a statistic [51]. The statistical analysis results of the Wilcoxon rank sum testing are shown in Tables 8–10.

Table 8. Statistical analysis results of Wilcoxon testing for noise-based K-means vs. FCM.

Taxi GPS Dataset	Clustering Result Evaluation Method	Noise-Based K-Means Versus FCM				
		k	*p*	*h*	*Stats*	
					Zval	*Ranksum*
Aracaju (Brazil)	**SC**	120	6.79561×10^{-8}	1	5.3965	610
	SSE		6.79561×10^{-8}	1	−5.3965	210
	DBI		6.79561×10^{-8}	1	−5.3965	210
	PBM		6.79561×10^{-8}	1	5.3965	610
San Francisco (USA)	**SC**	140	6.79561×10^{-8}	1	5.3965	610
	SSE		6.79561×10^{-8}	1	−5.3965	210
	DBI		6.79561×10^{-8}	1	−5.3965	210
	PBM		6.79561×10^{-8}	1	5.3965	610
Roma (Italy)	SC	135	6.79561×10^{-8}	1	5.3965	610
	SSE		6.79561×10^{-8}	1	−5.3965	210
	DBI		6.79561×10^{-8}	1	−5.3965	227
	PBM		6.79561×10^{-8}	1	5.3965	610
Chongqing (China)	SC	130	6.79561×10^{-8}	1	5.3965	610
	SSE		6.79561×10^{-8}	1	−5.3965	210
	DBI		6.79561×10^{-8}	1	−5.3965	210
	PBM		6.79561×10^{-8}	1	5.3965	610
Beijing (China)	**SC**	**125**	$\mathbf{6.79561 \times 10^{-8}}$	1	5.3965	610
	SSE		6.79561×10^{-8}	1	−5.3965	210
	DBI		6.79561×10^{-8}	1	−5.3965	210
	PBM		6.79561×10^{-8}	1	5.3965	610

Table 9. Statistical analysis results of Wilcoxon testing for noise-based K-means vs. K-means.

Taxi GPS Dataset	Clustering Result Evaluation Method	Noise-Based K-Means Versus K-Means				
		k	*p*	*h*	*Stats*	
					Zval	*Ranksum*
Aracaju (Brazil)	**SC**	120	3.41557×10^{-7}	1	5.0989	599
	SSE		9.17277×10^{-8}	1	−5.3424	212
	DBI		9.09000×10^{-2}	0	1.6906	473
	PBM		1.06456×10^{-7}	1	5.3153	607
San Francisco (USA)	SC	140	1.06456×10^{-7}	1	5.3153	607
	SSE		6.79561×10^{-8}	1	−5.3965	210
	DBI		8.59744×10^{-6}	1	4.4497	575
	PBM		6.79561×10^{-8}	1	5.3965	610
Roma (Italy)	SC	135	1.60981×10^{-4}	1	3.7735	550
	SSE		6.79561×10^{-8}	1	−5.3965	210
	DBI		1.23463×10^{-7}	1	5.2883	606
	PBM		6.79561×10^{-8}	1	5.3965	610
Chongqing (China)	SC	130	4.32000×10^{-3}	1	2.8538	516
	SSE		6.91658×10^{-7}	1	−4.9637	226
	DBI		3.98735×10^{-6}	1	4.6120	581
	PBM		9.74797×10^{-6}	1	4.4227	574
Beijing (China)	**SC**	**125**	3.36910×10^{-1}	0	−0.9603	374
	SSE		6.61044×10^{-5}	1	−3.9899	262
	DBI		7.89803×10^{-8}	1	5.3694	609
	PBM		2.13000×10^{-3}	1	3.0702	524

Table 10. Statistical analysis results of Wilcoxon testing for noise-based K-means vs. K-means plus.

Taxi GPS Dataset	Clustering Result Evaluation Method	Noise-Based K-Means Versus K-Means Plus				
		k	*p*	*h*	*Stats*	
					Zval	*Ranksum*
Aracaju (Brazil)	**SC**	120	3.38194×10^{-4}	1	3.5841	543
	SSE		2.04071×10^{-5}	1	−4.2604	252
	DBI		1.80570×10^{-1}	0	1.3390	460
	PBM		3.79330×10^{-1}	0	0.8791	443
San Francisco (USA)	SC	140	1.60981×10^{-4}	1	3.7735	550
	SSE		2.59598×10^{-5}	1	−4.2063	254
	DBI		1.80570×10^{-1}	0	1.3390	460
	PBM		2.56300×10^{-2}	1	−2.2316	327
Roma (Italy)	SC	135	9.09100×10^{-2}	0	1.6906	473
	SSE		9.61900×10^{-2}	0	−1.6636	348
	DBI		5.31000×10^{-2}	0	1.9341	482
	PBM		3.36910×10^{-1}	0	−0.9603	374
Chongqing (China)	SC	130	1.34000×10^{-3}	1	3.2054	529
	SSE		9.20913×10^{-4}	1	−3.3136	287
	DBI		3.36910×10^{-1}	0	0.9603	446
	PBM		9.03110×10^{-1}	0	0.1217	415
Beijing (China)	**SC**	**125**	7.64300×10^{-2}	0	1.7718	476
	SSE		5.62903×10^{-4}	1	−3.4489	282
	DBI		4.73480×10^{-1}	0	0.7168	437
	PBM		3.79330×10^{-1}	0	−0.8791	377

As can be seen from Tables 8–10, there are significant differences among noise-based K-means, FCM, and K-means clustering, which indicates that they are distributed differently in space; however, the differences between noise-based K-means and K-means plus clustering are not significant, especially in Roma (Italy), which rejects the significance level and indicates that their spatial distribution is similar. Furthermore, as $p \to 0$ for each group, it indicates that it is feasible and effective to automatically capture the cluster number and initialize the center cluster by use of a noise algorithm, and the urban hotspots captured by the noise-based K-means clustering algorithm more effectively represent reality.

6. Conclusions

In this paper, a novel noise-based K-means clustering algorithm has been proposed to effectively solve the problems of difficulty in determining the clustering numbers and the sensitivity of initializing the clustering center for a K-means clustering algorithm. The noise-based K-means clustering algorithm has been applied to capture urban hotspots in large cities from across the world. When the clustering operation was completed, the clustering results were evaluated by the DB index, PBM index, SC, and SSE, and the experimental results of each evaluation standard were statistically analyzed by Wilcoxon rank sum testing to obtain the significant differences for the urban hotspot distribution for each clustering algorithm. The proposed noise-based K-means clustering algorithm obtained better optimal results for urban hotspots over the FCM, K-means, and K-means plus methods. The method presented here can better serve a large number of people in large cities. In addition, the proposed noise-based K-means clustering algorithm can also be applied in the fields of the document classification, customer classification, ride data analysis, criminal network analysis, the detailed analysis of call records, and so on.

There are also some shortcomings for the method presented here. On the one hand, the amount of GPS data is too small to effectively reflect the distributions and the relationships of urban hotspots. On the other hand, it is difficult to effectively avoid specific buildings in the city, even if the optimal results of urban hotspots are used in urban road planning. As such, these problems will be further solved in future work.

Author Contributions: Conceptualization, X.Z. and X.R.; data curation, M.L.; formal analysis, W.D.; funding acquisition, X.Z. and W.D.; investigation, W.D. and X.R.; methodology, X.Z. and X.R.; project administration, M.L. and X.Z.; resources, X.R.; software, X.R. and X.Z.; validation, W.D. and W.T.; visualization, X.R.; writing—original draft, X.R. and X.Z.; writing—review and editing, W.D., W.T. and X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was jointly funded by the Sichuan Science and Technology Program, Grant/Award Numbers: 2019ZYZF0169; the A Ba Achievements Transformation Program, Grant/Award Number: 19CGZH0006, R21CGZH0001; the Chengdu Science and technology planning project, Grant/Award Number: 2021-YF05-00933-SN.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to acknowledge the UCI Machine Learning Repository, crawdad and Urban Computing GPS dataset in Microsoft Research.

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

References

1. Zhou, X.; Gu, J.; Shen, S.; Ma, H.; Miao, F.; Zhang, H.; Gong, H. An automatic k-means clustering algorithm of gps data combining a novel niche genetic algorithm with noise and density. *ISPRS Int. J. Geoinf.* **2017**, *6*, 392. [CrossRef]
2. D'Andrea, E.; Marcelloni, F. Detection of traffic congestion and incidents from GPS trace analysis. *Expert Syst. Appl.* **2017**, *73*, 43–56. [CrossRef]
3. Cui, J.; Liu, F.; Janssens, D.; An, S.; Wets, G.; Cools, M. Detecting urban road network accessibility problems using taxi GPS data. *J. Transp. Geogr.* **2016**, *51*, 147–157. [CrossRef]
4. An, S.; Yang, H.; Wang, J.; Cui, N.; Cui, J. Mining urban recurrent congestion evolution patterns from GPS-equipped vehicle mobility data. *Inf. Sci.* **2016**, *373*, 515–526. [CrossRef]
5. Li, T.; Qian, Z.; Deng, W.; Zhang, D.; Lu, H.; Wng, S. Forecasting crude oil prices based on variational mode decomposition and random sparse Bayesian learning. *Appl. Soft Comput.* **2021**, *113*, 108032. [CrossRef]
6. Shi, Y.; Da, W.; Tang, J.; Deng, M.; Liu, H.; Liu, B. Detecting spatiotemporal extents of traffic congestion: A density-based moving object clustering approach. *Int. J. Geogr. Inf. Sci.* **2021**, *35*, 1–26. [CrossRef]
7. Guo, J.; Liu, Y.; Yang, Q.; Wang, Y. GPS-based citywide traffic congestion forecasting using CNN-RNN and C3D hybrid model. *Transp. A Transport. Sci.* **2020**, *17*, 1–24. [CrossRef]
8. Yongdong, W.; Dongwei, X.; Peng, P.; Guijun, Z. Analysis of road travel behaviour based on big trajectory data. *IET Intell. Transp. Syst.* **2020**, *14*, 1691–1703. [CrossRef]
9. Dong, X.; Wang, L.; Hu, B. Analysis of spatio-temporal distribution characteristics of passenger travel behaviour based on online ride-sharing trajectory data. *J. Phys. Conf. Ser.* **2019**, *1187*, 052055. [CrossRef]
10. Siangsuebchart, S.; Ninsawat, S.; Witayangkurn, A.; Pravinvongvuth, S. Public transport gps probe and rail gate data for assessing the pattern of human mobility in the bangkok metropolitan region, Thailand. *Sustainability* **2021**, *13*, 2178. [CrossRef]
11. Cui, J.; Liu, F.; Hu, J.; Janssens, D.; Wets, G.; Cools, M. Identifying mismatch between urban travel demand and transport network services using gps data: A case study in the fast-growing Chinese city of Harbin. *Neurocomputing* **2016**, *181*, 4–18. [CrossRef]
12. Tang, J.; Gao, F.; Liu, F.; Zhang, W.; Qi, Y. Understanding Spatio-temporal characteristics of urban travel demand based on the combination of GWR and GLM. *Sustainability* **2019**, *11*, 5525. [CrossRef]
13. Luo, C.; Junlin, L.; Li, G.; Wei, W.; Li, Y.; Li, J. Efficient reverse spatial and textual k nearest neighbor queries on road networks. *Knowl. Based Syst.* **2016**, *93*, 121–134. [CrossRef]
14. Han, B.; Liu, L.; Omiecinski, E. Road-network aware trajectory clustering: Integrating locality, flow, and density. *IEEE Trans. Mob. Comput.* **2015**, *14*, 416–429.
15. Deng, W.; Shang, S.; Cai, X.; Zhao, H.; Zhou, Y.; Chen, H.; Deng, W. Quantum differential evolution with cooperative coevolution framework and hybrid mutation strategy for large scale optimization. *Knowl.-Based Syst.* **2021**, *224*, 107080. [CrossRef]
16. Hasan, S.; Ukkusuri, S.V. Urban activity pattern classification using topic models from online geo-location data. *Transp. Res. Part. C Emerg. Technol.* **2014**, *44*, 363–381. [CrossRef]
17. Iliopoulou, C.A.; Milioti, C.P.; Vlahogianni, E.I.; Kepaptsoglou, K.L. Identifying Spatio-temporal patterns of bus bunching in urban networks. *J. Intell. Transp. Syst.* **2020**, *24*, 365–382. [CrossRef]
18. Deng, W.; Xu, J.; Zhao, H.; Song, Y. A novel gate resource allocation method using improved PSO-based QEA. *IEEE Trans. Intell. Transp. Syst.* **2020**, *99*, 1–9. [CrossRef]
19. Lu, M.; Liang, J.; Wang, Z.; Yuan, X. Exploring od patterns of interested region based on taxi trajectories. *J. Vis.* **2016**, *19*, 811–821. [CrossRef]

20. Moreira-Matias, L.; Gama, J.; Ferreira, M.; Mendes-Moreira, J.; Damas, L. Time-evolving o-d matrix estimation using high-speed GPS data streams. *Expert Syst. Appl.* **2016**, *44*, 275–288. [CrossRef]
21. Huang, D.; Yu, J.; Shen, S.; Li, Z.; Zhao, L.; Gong, C. A method for bus od matrix estimation using multisource data. *J. Adv. Transp.* **2020**, *2020*, 5740521. [CrossRef]
22. Spaccapietra, S.; Parent, C.; Damiani, M.; Macêdo, J.; Porto, F.; Vangenot, C. A conceptual view on trajectories. *Data Knowl. Eng.* **2008**, *65*, 126–146. [CrossRef]
23. Luo, T.; Zheng, X.; Xu, G.; Fu, K.; Ren, W. An improved DBSCAN algorithm to detect stops in individual trajectories. *ISPRS Int. J. Geoinf.* **2017**, *6*, 63. [CrossRef]
24. Deng, W.; Xu, J.; Gao, X.; Zhao, H. An enhanced MSIQDE algorithm with novel multiple strategies for global optimization problems. *IEEE Trans. Syst. Man Cybern. Syst.* **2020**, *99*, 1–10. [CrossRef]
25. Nanni, M.; Pedreschi, D. Time-focused clustering of trajectories of moving objects. *J. Intell. Inf. Syst.* **2006**, *27*, 267–289. [CrossRef]
26. Pongracic, B.; Wu, F.L.; Fathollahi, L.; Brcic, D. Midlatitude Klobuchar correction model based on the k-means clustering of ionospheric daily variations. *GPS Solut.* **2019**, *23*, 80. [CrossRef]
27. Gu, Y.Y.; Wang, Y.D.; Dong, S.H. Public traffic congestion estimation using an artificial neural network. *ISPRS Int. J. Geoinf.* **2020**, *9*, 152. [CrossRef]
28. Gao, Y.; Li, J.L.; Xu, Z.G.; Liu, Z.Q.; Zhao, X.M.; Chen, J.H. A novel image-based convolutional neural network approach for traffic congestion estimation. *Expert Syst. Appl.* **2021**, *180*, 115037. [CrossRef]
29. Afrin, T.; Yodo, N. A probabilistic estimation of traffic congestion using Bayesian network. *Measurement* **2021**, *174*, 109051. [CrossRef]
30. Zhang, G.; Zhang, C.; Zhang, H. Improved k-means algorithm based on density canopy. *Knowl. Based Syst.* **2018**, *145*, 289–297. [CrossRef]
31. He, Z.; Yu, C. Clustering stability-based evolutionary k-means. *Soft Comput.* **2019**, *23*, 305–321. [CrossRef]
32. Arthur, D.; Vassilvitskii, S. K-means++: The advantages of careful seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, LA, USA, 77–9 January 2007; ACM, 2007; pp. 1027–1035.
33. Bezdek, J. Pattern Recognition with Fuzzy Objective Function Algorithms. *SIAM Rev.* **1981**, *25*, 442.
34. Borlea, I.-D.; Precup, R.-E.; Borlea, A.-B.; Iercan, D. A unified form of fuzzy c-means and k-means algorithms and its partitional implementation. *Knowl. Based Syst.* **2021**, *214*, 106731. [CrossRef]
35. Heil, J.; Häring, V.; Marschner, B.; Stumpe, B. Advantages of fuzzy k-means over k-means clustering in the classification of diffuse reflectance soil spectra: A case study with west African soils. *Geoderma* **2019**, *337*, 11–21. [CrossRef]
36. Bei, H.; Mao, Y.; Wang, W.; Zhang, X. Fuzzy clustering method based on improved weighted distance. *Math. Probl. Eng.* **2021**, *2021*, 6687202. [CrossRef]
37. Beg, A.H.; Islam, M.Z.; Estivill-Castro, V. Genetic algorithm with healthy population and multiple streams sharing information for clustering. *Knowl. Based Syst.* **2016**, *114*, 61–78. [CrossRef]
38. Ghezelbash, R.; Maghsoudi, A.; Carranza, E.J.M. Optimization of geochemical anomaly detection using a novel genetic k-means clustering (gkmc) algorithm. *Comput. Geosci.* **2020**, *134*, 104335. [CrossRef]
39. Huang, S.; Kang, Z.; Xu, Z.; Liu, Q. Robust deep k-means: An effective and simple method for data clustering. *Pattern Recognit.* **2021**, *117*, 107996. [CrossRef]
40. Jahangoshai Rezaee, M.; Eshkevari, M.; Saberi, M.; Hussain, O. GBK-means clustering algorithm: An improvement to the k-means algorithm based on the bargaining game. *Knowl. Based Syst.* **2021**, *213*, 106672. [CrossRef]
41. Ma, H.J.; Zhou, X.B. A GPS location data clustering approach based on a niche genetic algorithm and hybrid K-means. *Intell. Data Anal.* **2019**, *23*, S175–S198. [CrossRef]
42. Sun, H.D.; Chen, Y.Y.; Lai, J.H.; Wang, Y.; Liu, X.M. Identifying tourists and locals by K-means clustering method from mobile phone signaling data. *J. Transp. Eng. Part. A Syst.* **2021**, *147*, 04021070. [CrossRef]
43. Rahman, M.A.; Islam, M. Seed-detective: A novel clustering technique using high quality seed for k-means on categorical and numerical attributes. In Proceedings of the 9th Australasian Data Mining Conference (AusDM'11), Ballarat, Australia, 1–2 December 2011; Volume 121.
44. Liu, Y.; Wu, X.; Shen, Y. Automatic clustering using genetic algorithms. *Appl. Math. Comput.* **2011**, *218*, 1267–1279. [CrossRef]
45. Piorkowski, M.; Sarafijanovic-Djukic, N.; Grossglauser, M. Crawdad Dataset epfl/Mobility (v. 24 February 2009). Available online: http://crawdad.Org/epfl/mobility/20090224 (accessed on 7 July 2021).
46. Zheng, Y.; Liu, Y.; Yuan, J.; Xie, X. Urban computing with taxicabs. In Proceedings of the 13th International Conference on Ubiquitous Computing, Beijing, China, 17–21 September 2011; Association for Computing Machinery: Beijing, China, 2011; pp. 89–98.
47. Davies, D.L.; Bouldin, D.W. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *PAMI-1*, 224–227. [CrossRef]
48. Pakhira, M.K.; Bandyopadhyay, S.; Maulik, U. Validity index for crisp and fuzzy clusters. *Pattern Recognit.* **2004**, *37*, 487–501. [CrossRef]
49. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]

50. Nonparametric testing. In *Principles of Managerial Statistics and Data Science*; John Wiley & Sons: Hoboken, NJ, USA, 2020; pp. 533–549.
51. Zhou, X. *Research on Intelligent Clustering Learning Algorithm for GNSS Data*; Chengdu University of Technology: Chengdu, China, 2018.