applied sciences

MDPI

Article

A Novel K-Means Clustering Method for Locating Urban Hotspots Based on Hybrid Heuristic Initialization

Yiping Li [1,2], Xiangbing Zhou [2,*], Jiangang Gu [3], Ke Guo [1] and Wu Deng [2,4,*]

[1] College of Geophysics, Chengdu University of Technology, Chengdu 610059, China
[2] School of Information and Engineering, Sichuan Tourism University, Chengdu 610100, China
[3] Software Engineering College, Chengdu University of Information Technology, Chengdu 610225, China
[4] 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 rapid economic and demographic growth, traffic conditions in medium and large cities are becoming extremely congested. Numerous metropolitan management organizations hope to promote the coordination of traffic and urban development by formulating and improving traffic development strategies. The effectiveness of these solutions depends largely on an accurate assessment of the distribution of urban hotspots (centers of traffic activity). In recent years, many scholars have employed the K-Means clustering technique to identify urban hotspots, believing it to be efficient. K-means clustering is a sort of iterative clustering analysis. When the data dimensionality is large and the sample size is enormous, the K-Means clustering algorithm is sensitive to the initial clustering centers. To mitigate the problem, a hybrid heuristic "fuzzy system-particle swarm-genetic" algorithm, named FPSO-GAK, is employed to obtain better initial clustering centers for the K-Means clustering algorithm. The clustering results are evaluated and analyzed using three-cluster evaluation indexes (SC, SP and SSE) and two-cluster similarity indexes (CI and CSI). A taxi GPS dataset and a multi-source dataset were employed to test and validate the effectiveness of the proposed algorithm in comparison to the Random Swap clustering algorithm (RS), Genetic K-means algorithm (GAK), Particle Swarm Optimization (PSO) based K-Means, PSO based constraint K-Means, PSO based Weighted K-Means, PSO-GA based K-Means and K-Means++ algorithms. The comparison findings demonstrate that the proposed algorithm can achieve better clustering results, as well as successfully acquire urban hotspots.

Keywords: urban hotspots; K-means clustering; genetic algorithm; fuzzy system; particle swarm optimization; taxi GPS data

Citation: Li, Y.; Zhou, X.; Gu, J.; Guo, K.; Deng, W. A Novel K-Means Clustering Method for Locating Urban Hotspots Based on Hybrid Heuristic Initialization. *Appl. Sci.* **2022**, *12*, 8047. https://doi.org/10.3390/app12168047

Academic Editor: Giancarlo Mauri

Received: 1 June 2022
Accepted: 4 August 2022
Published: 11 August 2022

1. Introduction

Cities have traditionally been recognized as the primary drivers of economic activity and play a crucial role in human society. By evaluating the spatial distribution pattern of human economic and social activities, we can determine that humankind has entered a city-centric era [1]. The subject of "Urban Problems" has become increasingly prominent due to the rapid expansion of the size and population of major cities around the world [2]. The fundamental source of "urban problems" is the conflict between limited local resources and the needs of numerous city residents [3]. The typical phenomenon is traffic congestion [4]. Faced with the most challenging urban traffic problems, many international metropolitan management organizations hope to promote the coordination of traffic and urban development by formulating and improving traffic development strategies. The effectiveness of these strategies essentially depends on an accurate understanding of urban hotspots distribution [5]. Urban hotspots (also called human activity centers) generally have higher traffic densities than other city areas [6].

Urban hotspots are usually the clustering centers of numerous traffic trajectory points that affect population flow, movement patterns and human interaction. The urban road

network's complexity and surrounding space environment make it challenging to obtain urban hotspots. The widespread application of information technologies in the cities, such as GPS navigation, provides high-resolution Spatio-temporal perception [7]. In particular, GPS navigation services are widely used in vehicles, making it possible to accurately perceive the movement of many vehicles simultaneously in time and space [8]. The emergence of GPS trajectory data provides a new approach and method to solve the problem of urban traffic optimization [9]. As ordinary day-to-day GPS trajectory data, taxi track data not only reflects urban traffic conditions but also records people's daily travel information [10]. How to mine hidden knowledge from massive trajectory data has become an important research topic [11,12].

Many researchers have analyzed the GPS data of vehicle trajectory and drilled the connotative information behind these data, straining to find the spatial-temporal variation models reflecting vehicles distribution [13]. These models are applied to public transportation optimization analysis [14], urban population spatial and temporal dynamic distribution analysis [15], private car navigation real-time optimal path analysis [16], municipal road planning [17], urban hotspots searching [18,19], etc. Many research results have been applied to the practice of smart city construction, achieving good results in the balance of urban traffic flow and human flow load, and playing a pivotal role in emergency treatment.

These researchers use many machine learning methods based on GPS data sets, such as clustering analysis algorithm, time series algorithm, deep learning algorithm, reinforcement learning algorithm, and so on. Mariescu-Istodor and Fränti [20] proposed a grid-based method for GPS route analysis for retrieval that achieves efficient route search under noisy and variable sampling rate conditions. Zhang et al. [21] proposed a hybrid method for incremental extraction of urban road network from GPS data set, which can mine cumulative change patterns of road network. Shafabakhsh et al. [22] proposed a spatial kernel density calculation method based on GPS data to provide decision support for the allocation of medical emergency resources. Wang et al. [23] proposed a trajectory clustering method based on adaptive distance and hierarchical clustering based on the optimal cluster number to analyze similarities and anomalies in taxi GPS trajectory. Zhang et al. [24] proposed combining fuzzy C-means clustering with quantitative spatial regression to comprehensively analyze GPS data and establish the relationship model for improving the traffic service level from the perspective of time and space.

To extract the important information and mine the knowledge concealed in the data, a clustering algorithm is typically employed for preliminary data analysis [25,26]. The clustering algorithm is a fundamental and effective unsupervised learning method widely used in many fields. Clustering algorithms divide data set into different clusters according to a fixed standard (distance, similarity, etc.) [27], making the similarity of objects in the same cluster as ample as possible and the distinction of objects in the different clusters as large as possible [28]. These unsupervised clustering algorithms can be broadly divided into the following six categories: partition-based clustering , density-based clustering, model-based clustering, hierarchical clustering, grid-based clustering, and graph-based clustering [29]. Among the above algorithms, the clustering based on partition represented by the K-means algorithm has the advantages of simple structure, high efficiency, easy convergence and muscular universality. As one of the mature clustering algorithms, the K-means algorithm has achieved significant application effects in many fields. But it has disadvantages such as sensitivity to noise and outliers, difficulty in selecting cluster numbers, sensitivity to the initial clustering center, etc.

Numerous scholars have presented many ideas and approaches to locating the ideal cluster centers. The K-means++ algorithm was designed to weaken the K-means algorithm's sensitivity to the initial clustering centers; the algorithm's fundamental idea is to maximize the distance between the original clustering centers [30]. Krishna and Murty [31] combined the genetic algorithm with the K-means algorithm to obtain the Genetic K-means algorithm (GAK) converged to the global optimum. The K-means algorithm execution is regarded as

the K-means operator viewed as the search operator to replace crossover; Simultaneously, a kind of clustering-specific bias mutation operator is defined, called a distance-based mutation. Lu et al. [32] proposed a new clustering algorithm called Fast Genetic K-means Algorithm (FGAK) inspired by GAK, always converging to the global optimum eventually and running much faster. Islam et al. [33] advanced a previous genetic-searching approach called GenClust, with the intervention of fast hill-climbing of K-means and obtain an algorithm that is faster than its predecessor. Dowlatshahi and Nezamabadi-Pour [34] adapt the structure of stochastic population-based Gravitational Search Algorithm for effectively solving the multivariate data clustering problem. Some scholars have proposed that the combination of nearest-better neighborhood and fuzzy PSO is statistically superior to multi-mode optimization, indicating that the hybrid algorithm has advantages [35].

These improved K-means clustering algorithms achieved good results but still have disadvantages, such as excessive dependence on hyperparameters [36], over-sensitivity to outliers, and sensitivity to cluster center initialization. Toward alleviating these problems, this paper proposes a novel K-means clustering algorithm based on hybrid "fuzzy system-particle swarm-genetic" to automatically obtain the optimal initial centers, namely the Novel FPSO-GAK clustering algorithm. This algorithm can enhance the processing effect of automatic clustering, avoid too much uncertainty of manual configuration parameters and clustering results falling into local optimum. The initialization operation of the K-means++ was used to obtain a relatively optimal cluster center points group as one initial particle of the Particle Swarm Optimization (PSO). Then, the parameters of the PSO algorithm were optimized by a fuzzy system and the GA algorithm was implemented to search for better clustering centers. Finally, the optimal clustering center points were found after several iterations. Three unsupervised evaluation indexes, the Separation (SP), Silhouette coefficient (SC) and the Sum of Squared Error (SSE), as well as two cluster similarity indexes Centroid index (CI) and Point-level centroid similarity index (CSI), were used to evaluate the clustering results. Five taxi GPS datasets were selected and compared with GAK, PSO based K-Means (PSOK), PSO based constraint K-Means (PSOCK), PSO based Weighted K-Means (PSOWK), PSO-GA based K-Means (PSO-GAK), and K-Means++ algorithms to verify the validity of the proposedalgorithm.

The innovations and main contributions of this paper are as follows:

- A novel FPSO-GAK algorithm based on hybrid "fuzzy system-particle swarm-genetic" was designed to obtain excellent non-homogeneous search capability.
- The PSO algorithm, fuzzy system algorithm, and genetic algorithm were organically combined to obtain the optimal initial clustering centers.
- Three unsupervised evaluation indexes (SP, SC and SSE) and two cluster similarity indexes (CI and CSI) were employed to evaluate and analyze the clustering results.
- The comprehensive experiment was designed and implemented to verify the effectiveness of the novel FPSO-GAK clustering algorithm.

2. A Novel FPSO-GAK Clustering Method

2.1. The Idea of the Novel FPSO-GAK Method

K-means is an exclusive distance-based clustered partitioning algorithm that is implemented iteratively [37]. Scholars have regarded it as an effective means to discover urban functional areas in the past decades [38,39]. The K-means clustering problem on n points is NP-Hard for dimension $d \geq 2$ [40], and it is a demanding but unavoidable problem to determine the optimal clustering centers accurately. The acquisition of the optimal cluster centers depends greatly on the initial cluster centers, which is also an NP-hard problem. In practice, one of the effective methods of approaching NP-hard problems is the heuristic method. Many researchers have applied heuristic algorithms to initial cluster centers [41], exposing the issues that one single heuristic algorithm is prone to fall into local optimization and premature convergence in the process of searching. These problems are caused mainly by the homogeneity of the search strategy of one single heuristic algorithm. The proposed algorithm is formed by a suitable embedded combination of PSO, GA heuristic algorithm

and fuzzy system, which maintains the advantages of each algorithm and overcomes local optimization and premature convergence to a great extent. In order to solve the problems existing in the K-means clustering algorithm and make use of the advantages of the hybrid heuristic algorithm, this paper proposes a new K-means clustering algorithm which can obtain better clustering centers and capture urban hotspots under the premise of given cluster number.

The proposed noevel FPSO-GAK algorithm consists of three parts. Given the number of clusters K, perform the following steps: Firstly, the k-means++ initialization method is employed to obtain K initial cluster center points and treat them as one particle. Perform this process for the provided number of rounds to obtain the desired number of particles. Secondly, particles are put into the PSO algorithm. The PSO algorithm is employed to update the particle information and use the fuzzy system to update the parameter Settings of the PSO algorithm; the cycle is executed alternately until the iteration criteria are satisfied. Then, the optimized particles are put into the GA algorithm for optimization. In this stage, the GA, fuzzy, and PSO algorithms are executed alternately until the iterative convergence criteria is satisfied. Finally, the initial clustering center data corresponding to the optimal particle extracted in the above steps are input into the K-means clustering algorithm to cluster the given data. The excellent clustering centers, namely the urban hotspots, are captured.

2.2. The Flow of the Novel FPSO-GAK Clustering Method

The flow of the Novel FPSO-GAK clustering algorithm is shown in Figure 1.

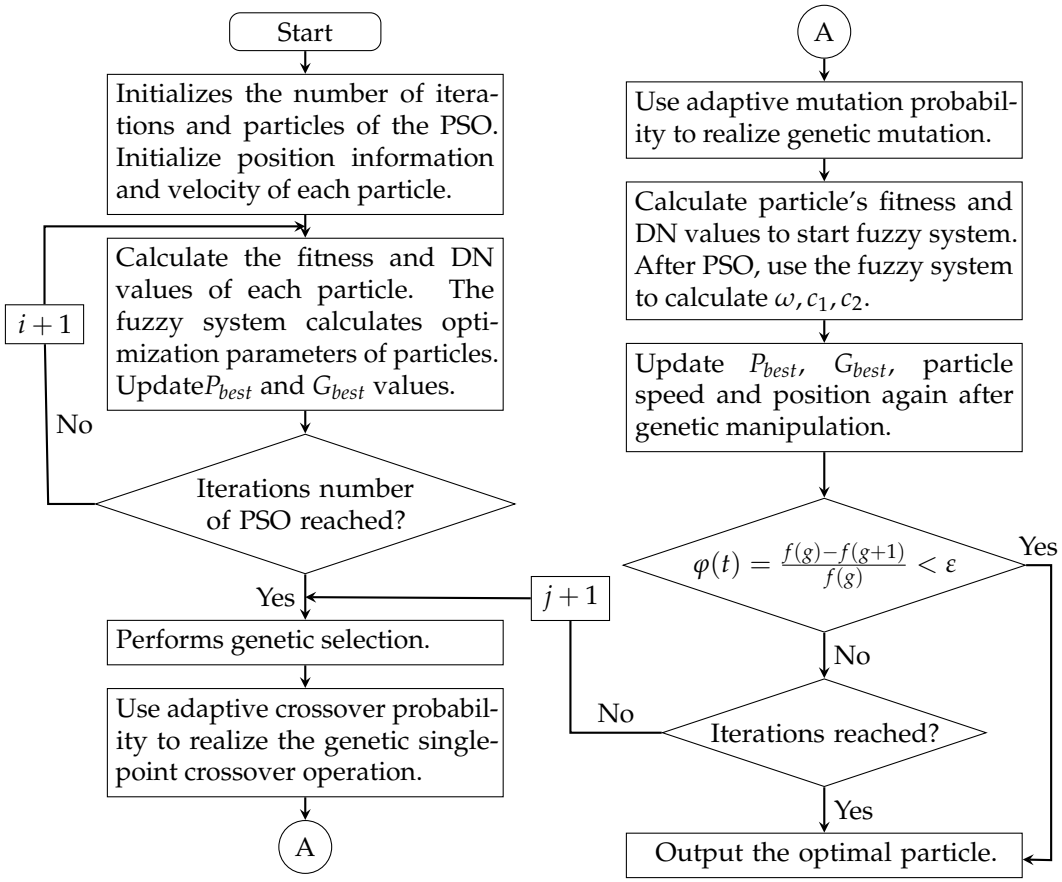

Figure 1. The flow of the novel FPSO-GAK clustering algorithm.

2.3. The Realization of the Novel FPSO-GAK Clustering Method

The detailed execution steps of the proposed clustering algorithm (pseudo-code is shown in Algorithm 1) are as follows.

Algorithm 1: Novel FPSO-GAK Clustering Algorithm.

Input:
The taxis GPS data set and the number of the data points;
The population size of PSO alogrithm, N_{swarm};
The termination condition and/or The maximum number of clustering iterations;

Output:
Clustering center points and Clustering results.

1: The population size is set as N_{swarm} (number of particles). And the initial velocity of particles is randomly generated within the given $[v_{min}, v_{max}]$; the initial velocity of different particles is different. N_{swarm} rounds of initialization operations of K-Means++ are performed in the GPS data set, and each round's initial cluster center points are taken as a particle.
2: The fitness value of each particle is calculated and sorted in descending order to find out the optimal fitness value *Pbest* of each individual and the optimal population value *Gbest*.
3: Operate the fuzzy system, put the calculation results of Formulaes (1) and (2) into the fuzzy system, and calculate the values of ω, c_1, and c_2. The above output results control the inertia weight and the particle velocity and change the particle position distribution.
4: Update the velocity and position of the particles according to Formula (4a,4b).
5: Repeat Step 1 to Step 4 until the termination conditions are met. The fitness of each particle is calculated, and the previous best even number of particles are stored in memory for subsequent genetic algorithm operations.
6: The results of Step 5 are manipulated by the genetic algorithm, including crossover, mutation, and elite operation. In particular, genetic selection has been performed in step 5, directly utilizing roulette to pick individuals for the crossover operation, but without genetic rearrangement operation.
7: Update the velocity and position of the particles according to Formula (4a,4b).
8: Determine whether the termination condition is met. If true, output the optimal particle and proceed to Step 9. Otherwise, return to Step 6 until the termination condition is met.
9: Run the K-means algorithm.

step.1 The initial parameter setting

According to the literature [42–45], one of the most remarkable features of PSO algorithms is their speed. As per experience, the initial population should range from 50 to 1,000; while the convergence will be better with a larger initial population, a population that is too large will also affect the algorithm's speed. In our experiment, the initial population N_{swarm} was between 100 and 200. N_{swarm} rounds of initialization operations of K-means++ were performed in the GPS data set, and each round's initial center points were taken as one particle. Many researchers believe that the clustering number of the K-means algorithm is usually between 2 and $\sqrt{N}$, where N is the number of data points in the data set [46,47]. The urban hotspots must be depicted at a specified density; the initial clustering number was set to around $\sqrt{\frac{N}{8}}$.

step.2 Optimizing the initial clustering center points

The fitness and DN values of each particle were calculated for the fuzzy operation. The fitness function, also known as the evaluation function, is a criterion determined by the objective function to assess the quality of individuals within a group. It is always non-negative, the greater the value, the better it is. The fitness function design should be as straightforward as feasible to reduce computation time complexity. The fitness function f in this paper is as follows:

$$f = \frac{1}{\sum_{i=1}^{K} \sum_{X_{ij} \in X_i} (X_{ij} - \overline{X_i})^2} \tag{1}$$

where, K represents the number of clusters, J means the number of the point that belong to cluster i. The following formula for DN is used to change the fuzzy system and control the maximum invariant fitness values:

$$DN = \frac{\left| \sum_{f_{PG} - f_{CG} < 0} (f_{PG} - f_{CG}) \right|}{N_{swarm}} \tag{2}$$

where PG and CG indicate the optimal population extremum/individual extremum, respectively, f denotes fitness value of a particular particle, N_{swarm} means the population size of the particle swarm. As found in related literature [48,49], the fitness value of particle swarm is normalized as the parameter of the fuzzy system according to the following formula:

$$NBF = \frac{f - f_{min}}{f_{max} - f_{min}} \tag{3}$$

Then the particle optimization parameters ω, c_1, and c_2 are calculated by the fuzzy system according to the fuzzy rules in the Table 1 and Mamdani fuzzy inference algorithm.

Table 1. Fuzzy rules for inertia weight ω, learning factor c_1 and c_2.

ω		DN				c_1		DN				c_2		DN			
		PS	**PM**	**PB**	**PR**			**PS**	**PM**	**PB**	**PR**			**PS**	**PM**	**PB**	**PR**
NBF	PS	PS	PM	PB	PB	NBF	PS	PR	PB	PB	PB	NBF	PS	PR	PB	PM	PM
	PM	PM	PM	PB	PR		PM	PB	PM	PM	PS		PM	PB	PM	PS	PS
	PB	PB	PB	PB	PR		PB	PB	PM	PS	PS		PB	PM	PM	PS	PS
	PR	PB	PB	PR	PR		PR	PM	PM	PS	PS		PR	PM	PS	PS	PS

PS: positive small; PM: positive middle; PB: positive big; PR: positive rare.

After a PSO routine runs once, the values of P_{best} and G_{best} are updated according to the sorted fitness values of each particle. Particles update velocity v_{id} and position z_{id} using the following formula:

$$v_{id}^{g+1} = \omega v_{id}^{g} + c_1 r_1 (p_{id} - z_{id}^{g}) + c_2 r_2 (p_{gd} - z_{id}^{g}) \tag{4a}$$

$$z_{id}^{g+1} = z_{id}^{g} + v_{id}^{g+1} \tag{4b}$$

where ω denotes the inertia weight, c_1 and c_2 are two positive learning factors, r_1 and r_2 are random functions to generate uniformly distributed random values in the range $[0, 1)$. The fuzzy system updates the parameter settings of the PSO algorithm, and the PSO algorithm is employed to update the particle information; the cycle is executed alternately until the iteration conditions are satisfied. The optimized particles obtained in the previous stage are sorted in descending order, and a certain number of the top particles are selected to enter the subsequent genetic algorithm. The top 30% of particles were picked for this paper.

Genetic selection operations were performed using roulette algorithms. The gene rearrangement technique based on cosine similarity was used to deal with chromosomes of unequal length and different shapes. According to the fitness value, the following adaptive crossover probability P_c formula is designed to realize the genetic single-point crossover operation:

$$P_c = \begin{cases} \frac{f_{max} - f'}{f_{max} - f_{avg}} & \text{if} \quad f' > f_{avg} \\ 1 & \text{if} \quad f' \leq f_{avg} \end{cases} \tag{5}$$

where f_{max} represents the maximum fitness value in the current population; f_{avg} represents the average fitness value in the current population; f' represents the more significant fitness value of the two chromosomes that participated in crossover operation. The crossover operation is performed when the P_c value is greater than the roulette pick probability. Formula 5 works as follows: when the fitness value of chromosomes involved in the crossover is small, crossover operation must be performed; otherwise, the probability of crossover operation is low. Then, the adaptive mutation operation is carried out. Genetic variation is the operation of gene mutation on chromosomes in a population to search for a better solution. Usually, small changes are made to the chromosomes of the population, such as mutating the chromosomes with lower fitness to expand the search space and prevent falling into local optimality. As with crossover operations, a mutation probability is required to perform chromosome mutation operations. This paper obtains the adaptive variation probability P_m by using the following formula:

$$P_m = \begin{cases} \xi_1 \times \frac{f_{max} - f}{f_{max} - f_{avg}} & \text{if} \quad f > f_{avg} \\ \xi_2 & \text{if} \quad f \leq f_{avg} \end{cases} \tag{6}$$

where ξ_1 and ξ_2 are gene mutation coefficients; f_{max} represents the maximum fitness value in the current chromosomes population; f_{avg} represents the average fitness value in the current chromosomes population; f represents the fitness value of chromosomes involved in mutations.

After performing a genetic algorithm operation, the generated particle's fitness value and DN value are calculated to start the fuzzy system. The fuzzy system calculates out the particle optimization parameters ω, c_1, and c_2 parameter. The P_{best} and G_{best} are updated according to the fitness of each particle; then, the particle swarm optimization algorithm is activated to update the particles velocity and position information. Genetic mutation operation, fuzzy system operation and particle swarm operation are performed alternately until iteration conditions are satisfied. Then, the particle with the highest fitness is selected as the initial clustering center of the subsequent K-means algorithm.

step.3 Capturing the optimal urban hotspots

On the premise that superior initial clustering centers were obtained, the K-means algorithm can converge to the final outcome more quickly and accurately. Especially when the number of clustering data points is enormous and the dimension of data is high, the initial clustering centers have a substantial influence on the K-means algorithm's efficiency. We used the particle with the highest fitness value acquired in Step 2 as the initial clustering centers of the k-means algorithm to analyze the GPS data set of urban taxis. The final clustering centers of taxi GPS data are the urban hotspots we are searching for.

2.4. Synthetic Data Test

To validate the effectiveness of the proposed algorithm, a synthetic dataset was employed for testing. The synthetic dataset (Figure 2) contains 1000 data points divided into 120 clusters with random distribution of centroids, and the data variance within the clusters is 15. There are various enhanced versions of the K-Means algorithm now available. The proposed algorithm was compared with the state-of-the-art Random Swap clustering algorithm (RS) [50], a lightweight and efficient clustering algorithm with linear time complexity depending on the size of the data. The proposed algorithm contains meta-heuristic type methods whose time complexity will be higher than the time complexity of K-means arithmetic. According to the analysis in Section 3.5, it knows that the proposed algorithm's time complexity shows a basically linear relationship with the number of data. It is estimated that the time consuming of the proposed algorithm is about 20 to 50 times that of the random-swap clustering algorithm. At the cost of time complexity, the non-homogeneous search capability of the proposed algorithm can achieve better clustering results on medium and large data sets.

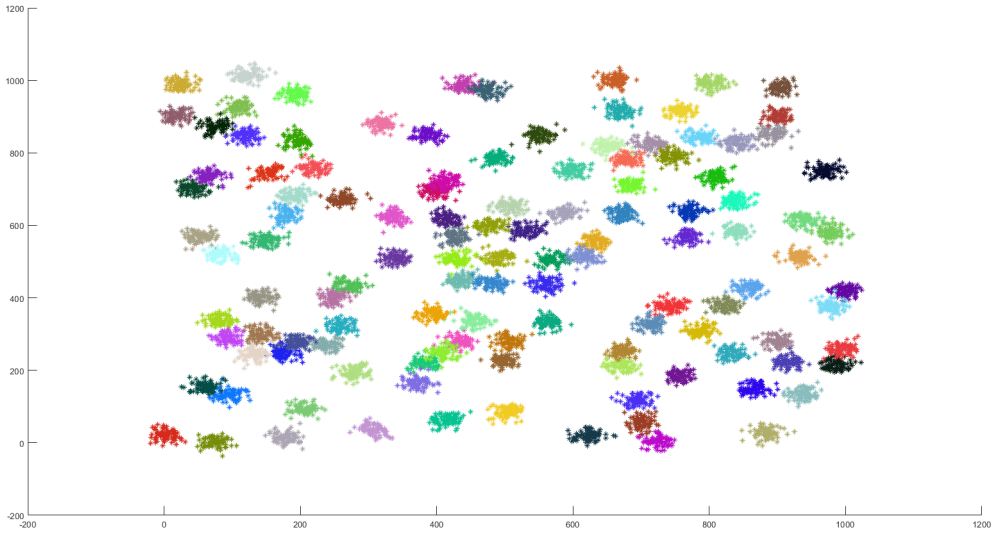

Figure 2. The synthetic dataset distribution diagram.

Thirty clustering operations were performed on the synthetic data using the proposed algorithm and random swap clustering algorithms according to the accuracy convergence condition and a fixed number of iterations, respectively (the number of clusters for each clustering operation was selected randomly between 70 and 90). The average SC, SP, SSE and ACI (reference to Section 3.6) indexes were used to evaluate the results of clustering, as shown in Table 2.

Table 2. Average indexes for clutering operations in synthetic dataset.

Algorithms	Time-Consuming	SC	SP	SSE	ACI
RS(60)	2.518 s	0.9258	265.24	277,835.5	0.2199
FPSO-GAK(85)	65.495 s	0.9365	266.79	275,294.1	
RS(5000)	206.853 s	0.9297	266.42	271,651.7	0.2213
FPSO-GAK(1000)	803.172 s	0.9383	267.13	269,894.1	

The number within parentheses represents the number of iterations performed by the algorithms.

Table 2 demonstrates that the state-of-the-art Random Swap clustering algorithm is an effective clustering algorithm that achieves excellent results with respect to all clustering indexes. The ACI index of FPSO-GAK and Random Swap Clustering algorithms is greater than 0.15, indicating a substantial difference in their clustering-level similarity. The proposed algorithm slightly outperforms the Random Swap Clustering algorithm in many indexes except the time-consuming index. The time complexity of both algorithms shows a linear relationship with the amount of data, but the proposed algorithm will take an order of magnitude more time-consuming than the Random Swap Clustering algorithm. Considering the rapid increase of computing power nowadays, it is acceptable in applications to sacrifice a certain time loss to obtain the improvement of the clustering effect. In the subsequent test and analysis of the real dataset, we will further conduct a comparative analysis of the proposed algorithm with similar types of heuristics algorithms.

3. Experiment Results and Analysis

3.1. Experiment Data Set

3.1.1. Taxis GPS Data Set

For the purpose of verifying the performance of the novel algorithm, five GPS data sets were used as shown in Table 3. GPS data sets typically contain vehicle ID numbers, longitude values, latitude values, altitude values, timestamps, GpsSpeed, etc., which are periodically recorded by the onboard GLOBAL positioning system (GPS). Currently, public

transport authorities in many cities have accumulated taxi GPS data sets for several years. In this study, we focused on GPS data sets from representative metropolitan areas in several countries, i.e., Aracaju (Brazil), Beijing (China), Chongqing (China), Rome (Italy), and San Francisco (USA). Data set Aracaju is a taxi GPS data set from Aracaju city in Brazil, which belongs to the standard data set in UCI Machine Learning Repository. Data set Beijing (China) is a taxis GPS data set from downtown Beijing, China, which is from the well-known Geolife Trajectories project [51,52]. Data set Chongqing (China) is a taxis GPS data set from urban area Chongqing, China. Data set Rome (Italy) is a taxi GPS data set from Rome city, which belongs to the data set in Crawdad and contains GPS data of about 320 taxis' trajectories [53]. Data set San Francisco (USA) is a taxi GPS data set from San Francisco, which belongs to the data set in Crawdad and contains GPS data of about 500 taxis' trajectories [53]. The data sets have been cleaned beforehand; just a subset of the original data sets has been extracted, and a few records between 1:00 and 6:00 p.m. have been eliminated.

Table 3. Summary information of Taxis GPS data sets.

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

3.1.2. Multi-Source Trajectory Data Set

In addition to the taxi vehicle GPS dataset, we validated the proposed algorithms using the multi-source trajectory dataset. The multi-source trajectory dataset is derived from the Routes 2019 subset of the MOPSI dataset (http://cs.uef.fi/mopsi/data, accessed on 5 July 2022), which contains approximately 3 million trajectory data points [54]. Trajectories encompass a variety of activities, such as walking, biking, hiking, jogging, driving, taking the bus, and others. Some of these trajectory points record life routines, such as going to work or shopping. The dataset is particularly suitable for migratory pattern mining, person activity identification, and location-based similarity. In the experiments of this paper, we select multi-source trajectory data points within the urban area for analysis, mainly including trajectory point data of vehicles, walking, jogging and biking. The spatial distribution of the multi-source data trajectory points is shown in Figure 3.

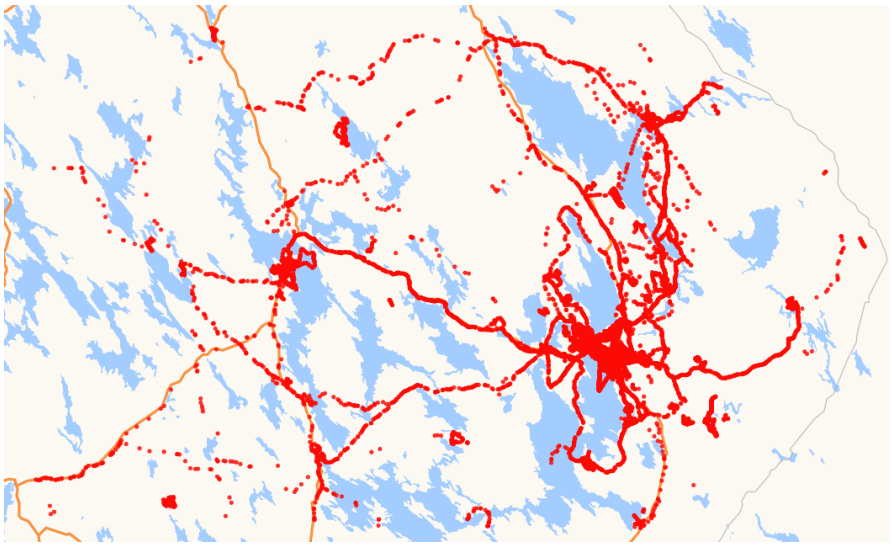

Figure 3. The spatial distribution of the multi-source data trajectory points.

3.2. Experimental Environment and Parameter Setting

The experimental environment based on a personal computer featured the following: Intel I5-8250U, dominant frequency 4 × 1.60 GHz with 8G RAM, Windows 10 operating system, and the algorithm routines were coded in MATLAB 2018b.The taxi GPS data sets contained subsets Aracaju, Beijing, Chongqing, Rome and San Francisco were used as experimental data sets. In each algorithm routine, clustering performance was evaluated using SC, SP, and SSE indicators, which paid particular attention to SC index. The above algorithm routines were trained 20 times independently in the experiment. The clustering numbers to GAK and K-means++ algorithms were initialized to 100. Then the parameters were adjusted respectively within $\sqrt{n}$ according to the amount of data sets to acquire the best experimental results. The initial parameter setting of PSO and fuzzy system are shown in Table 4.

For the PSO algorithm: the inertia weight ω keeps the particle moving with inertia, giving it the tendency to expand the search space and the ability to explore new regions. The acceleration constants c_1 and c_2 represent the weights of the statistical acceleration terms that push each particle toward the P_{best} and G_{best} positions. A low value allows the particle to hover outside the target region before being pulled back, while a high value causes the particle to make a sudden dash toward or over the target region. The higher the number of populations (NP), the higher the accuracy will be obtained, but at the same time, it will prolong the computing time. The velocity range (VRP) determines the resolution (or accuracy) in the region between the current position and the best position. If velocity maximum is too high, the particle may fly past the good solution, and if velocity minimum is too small, the particle cannot perform sufficient exploration, leading to becoming stuck in local optima.

For the GA algorithm: the adaptive cross coefficient range (ACCR) is too large, it is easy to destroy the existing favorable mode, increase the randomness and easily miss the optimal individual; it is too small to renew the population effectively. The adaptive variation coefficient range (AVCR)is too small and the diversity of the population declines too fast, which easily leads to the rapid loss of effective genes and is not easy to repair; it is is too large, the diversity of the population can be guaranteed, but the probability of higher-order mode destruction also increases.

For the fuzzy system: the fuzzy range of ω in the PSO (FRW) control the variation range in the defuzzification process. The fuzzy range of c_i in the PSO (FRC) control the variation range in the defuzzification process.

Table 4. PSO and fuzzy system parameter setting.

DS	VRP	NLDP	NC	NP	ω	c_1	c_2	FRW	FRC	ACCR	AVCR
Aracaju	0.8,-0.8	16,513	120	41	0.9	1.5	1.5	0.2,1.2	1.0,2.0	0.5,1.0	0.5,1.0
Beijing	0.8,-0.8	17,387	130	41	0.6	1.2	1.2	0.2,1.2	1.0,2.0	0.5,1.0	0.5,1.0
Chongqing	0.4,-0.4	19,149	130	41	0.6	1.1	1.1	0.2,1.2	1.0,2.0	0.5,1.0	0.5,1.0
Rome	0.1,-0.1	20,254	140	51	0.6	1.5	1.5	0.2,1.2	1.0,2.0	0.5,1.0	0.5,1.0
San Francisco	0.3,-0.3	21,826	140	51	0.8	1.6	1.6	0.2,1.2	1.0,2.0	0.5,1.0	0.5,1.0

DS: data set; NLDP: number of location data points; NC: number of clusters.

3.3. Experimental Results and Comparison Analysis

3.3.1. Taxis GPS Data Set

Figures 4–8 show the convergence diagrams for the searching optimal initial clustering centers stage of the GAK, PSOK, PSOCK, PSOWK, PSO-GAK, and the proposed algorithms utilizing Section 3.2 parameters in five distinct GPS data sets. Figures 9–13 is the convergence diagram for k-means clustering stage of 7 algorithms (the above six algorithms plus K-means++ algorithm). Tables 5–7 are the comparison tables for the average Silhouette coefficient (SC), the average degree of separation (SP), and the average SSE value of 20

training sessions respectively (plus Random Swap Clustering algorithm). In order to accurately analyze the above evaluation indexes, each algorithm was iterated 180 times in a single run to guarantee convergence. Figures 14–19 show The obtained clustering centers of Aracaju, Beijing, Chongqing, Rome, and San Francisco after the operation of six algorithms. Clustering centers outcome are shown on the Baidu Map background.

Tables 5–7 demonstrate that the proposed algorithm has superior clustering performance in terms of Silhouette coefficient, degree of separation and SSE, as well as the ability to capture better clustering results and locate more effective urban hotspots. However, in Table 7's San Francisco data set, the K-Means++ algorithm delivers just slightly worse than the proposed algorithm. It reveals that the K-Means++ algorithm can be used as a preliminary coarse-grained clustering algorithm in such applications. The tables also demonstrate that the overall clustering performance of the RS algorithm is good, and the clustering effect is better than the K-Means++ algorithm in most cases. In Table 5, the average Silhouette coefficient of the proposed algorithm is higher than other algorithms. The greater the Silhouette coefficient, the more compact within a cluster and more dispersed between the clusters, indicating that the clustering effect is superior to the competitors. In Table 6, the proposed algorithm has a higher SP value than other algorithms. The higher the SP value, the greater the separation between clustering clusters, and the higher the clustering algorithm's ability to satisfy the needs of the majority of clustering. Table 7 indicates that the average SSE value of the proposed algorithm in 20 training cycles is superior to other algorithms. Overall, the clustering results of the proposed algorithms all outperformed the comparison algorithms. The above suggests that hybrid heuristic initialization algorithms are effective, practical and feasible.

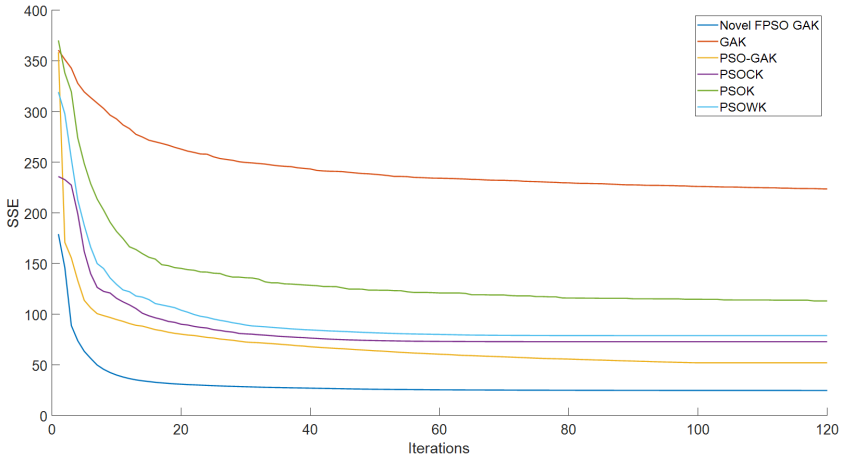

Figure 4. The SSE convergence curve for the searching optimal initial clustering centers stage of taxi GPS dataset from Aracaju(Brazil).

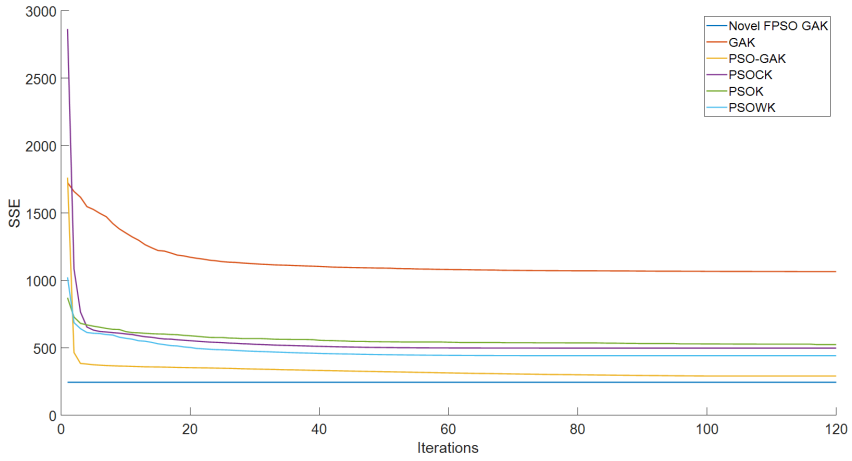

Figure 5. The SSE convergence curve for the searching optimal initial clustering centers stage of taxi GPS dataset from Beijing(China).

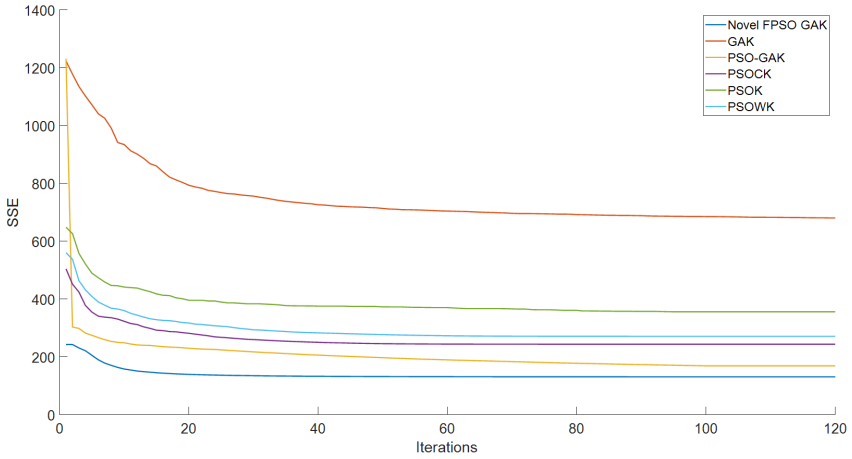

Figure 6. The SSE convergence curve for the searching optimal initial clustering centers stage of taxi GPS dataset from Chongqing(China).

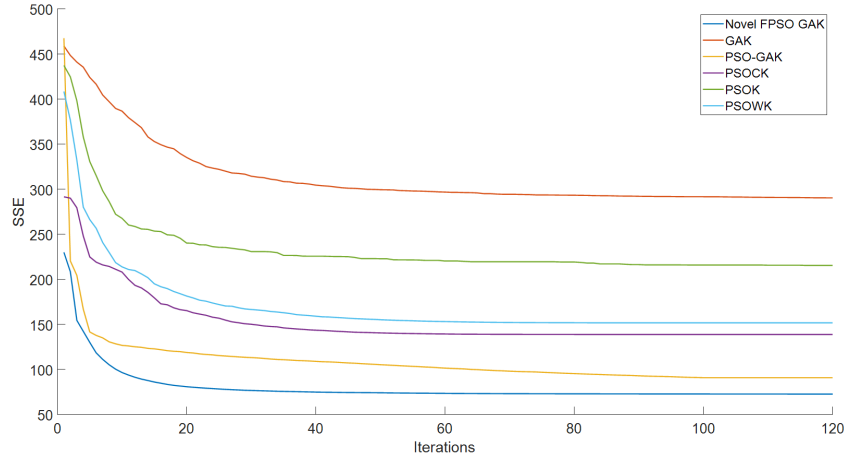

Figure 7. The SSE convergence curve for the searching optimal initial clustering centers stage of taxi GPS dataset from Roma(Italy).

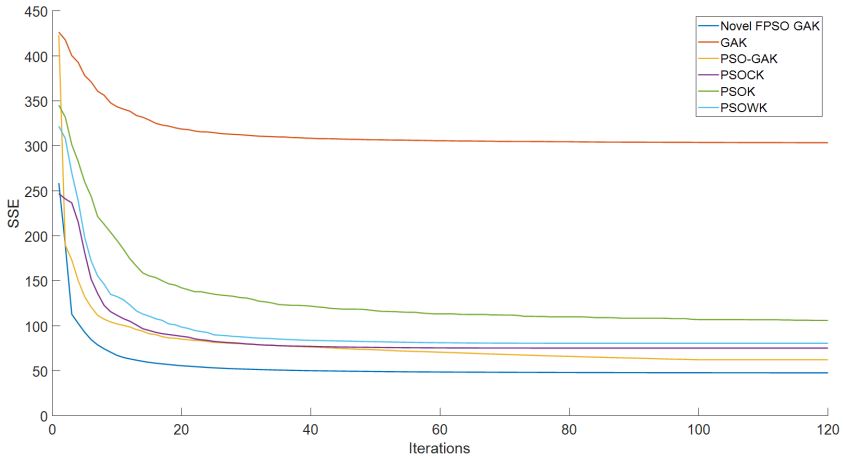

Figure 8. The SSE convergence curve for the searching optimal initial clustering centers stage of taxi GPS dataset from San Francisco (USA).

Table 5. Average Silhouette coefficient of 8 algorithms trained 20 times in 5 GPS data sets.

Algorithms	Aracaju	Beijing	Chongqing	Rome	San Francisco
K-means++	0.9339	0.8520	0.9289	0.9050	0.9183
RS	0.9301	0.8623	0.9252	0.9052	0.9190
GAK	0.9344	0.8595	0.9299	0.9065	0.9195
PSOK	0.9347	0.8616	0.9294	0.9031	0.9196
PSOCK	0.9339	0.8597	0.9286	0.9013	0.9182
PSOWK	0.9342	0.8622	0.9287	0.9042	0.9199
PSO-GAK	0.9341	0.8608	0.9296	0.9016	0.9185
FPSO-GAK	0.9360	0.8685	0.9301	0.9057	0.9209

Table 6. Average degree of separation of 8 algorithms trained 20 times in 5 GPS data sets.

Algorithms	Aracaju	Beijing	Chongqing	Rome	San Francisco
K-means++	0.0415	0.2430	0.1361	0.0756	0.0383
RS	0.0410	0.2442	0.1304	0.0760	0.0390
GAK	0.0424	0.2488	0.1355	0.0644	0.0371
PSOK	0.0439	0.2809	0.1570	0.0850	0.0434
PSOCK	0.0432	0.2748	0.1488	0.0840	0.0392
PSOWK	0.0438	0.2813	0.1636	0.0851	0.0443
PSO-GAK	0.0445	0.2843	0.1618	0.0864	0.0427
FPSO-GAK	0.0447	0.2858	0.1650	0.0865	0.0451

Table 7. Average SSE of 8 algorithms trained 20 times in 5 GPS data sets.

Algorithms	Aracaju	Beijing	Chongqing	Rome	San Francisco
K-means++	18.9475	213.4044	101.0077	51.3587	34.6147
RS	19.9865	212.5084	105.8674	51.0076	34.5113
GAK	19.0157	212.8031	100.7465	52.6603	35.1143
PSOK	19.8867	224.2299	109.4038	54.0949	36.4149
PSOCK	19.8712	223.6972	109.3995	54.1288	36.6931
PSOWK	19.5313	223.5987	110.5026	55.8484	35.7416
PSO-GAK	19.6945	224.5453	109.3748	55.1924	36.1069
FPSO-GAK	18.4654	212.3069	100.3002	50.5796	34.3267

Figures 4–8 show that the proposed algorithm undergoes minimal change during the iterative convergence process. Within the five groups of taxi GPS data sets, the pro-

posed really is in the optimal state, the optimization effect is noticeable, and its curve variations are negligible. It indicates that the proposed algorithm can obtain relatively better clustering centers by using the initialization operation of the k -means++ algorithm when initializing the populations. Assuming that the proposed algorithm is equivalent to the K-means++ algorithm, the two algorithms' convergence curves will coincide in the convergence scenario, which is not observed in the experiments. In the iterative optimization process of the proposed algorithm, the fitness value changes slightly, and the location distribution of particles and the clustering center are constantly modified to achieve a more optimal clustering centers distribution throughout the whole GPS data set. The final clustering effect can better accommodate actual requirements. The convergence curves of the proposed algorithm and PSO-GAK algorithm remain close to each other, indicating that the parameters of the proposed algorithm should be further adjusted to improve its convergence effect, which will be studied in future work.

Figures 9–13 show that the proposed algorithm can obtain better particles than the GAK, PSOK, PSOCK, PSOWK, PSO-GAK and K-means++ as the input initial clustering centers of the K-means algorithm. With optimization by the proposed algorithm, improved clustering centers can be captured, significantly reducing the K-means algorithm's sensitivity to initial clustering centers. When SSE is employed as the objective function of the k-means clustering algorithm, the clustering centers obtained by the proposed algorithm are used to perform the K-means clustering algorithm, which has faster convergence characteristics and a lower SSE value overall.

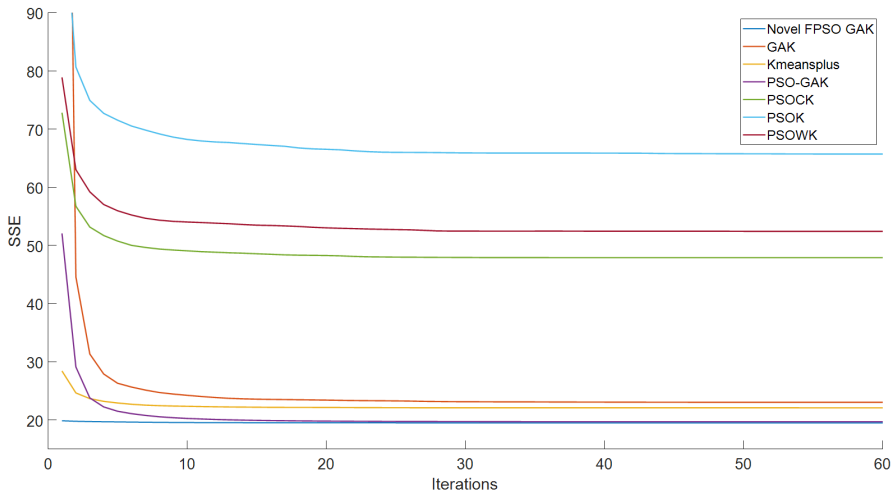

Figure 9. The SSE convergence curve of taxi GPS dataset from Aracaju(Brazil).

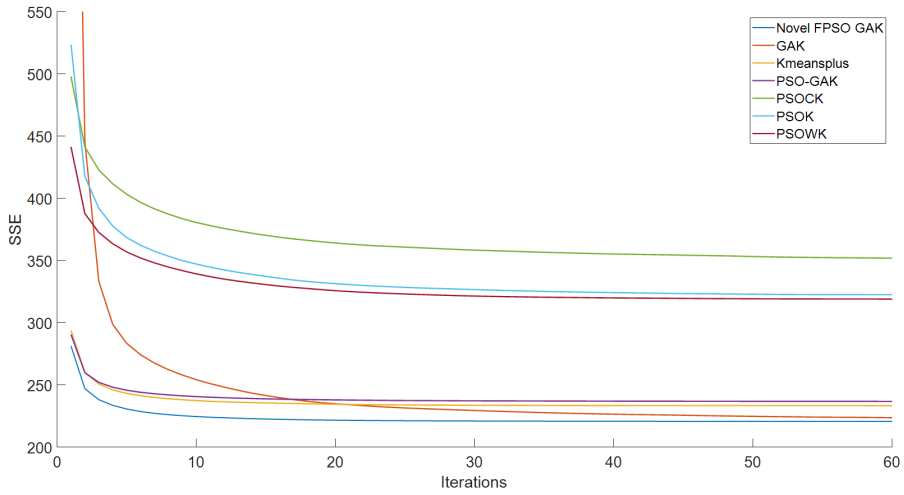

Figure 10. The SSE convergence curve of taxi GPS dataset from Beijing(China).

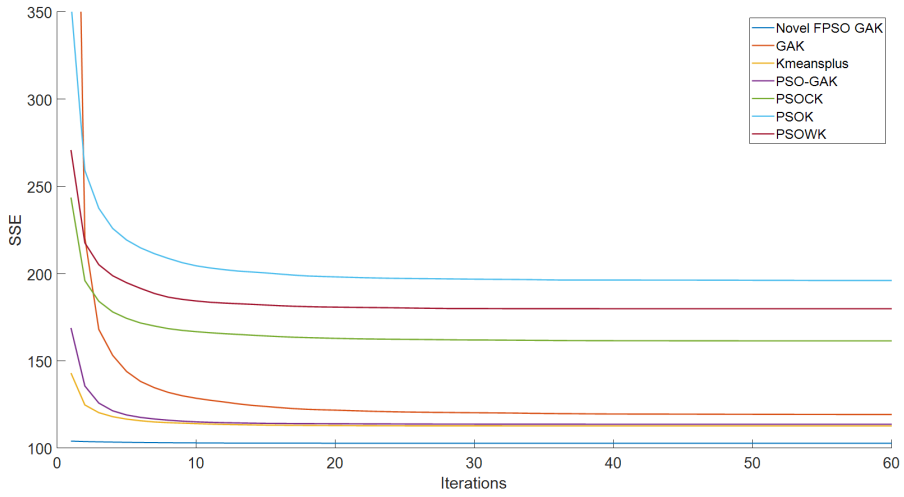

Figure 11. The SSE convergence curve of taxi GPS dataset from Chongqing(China).

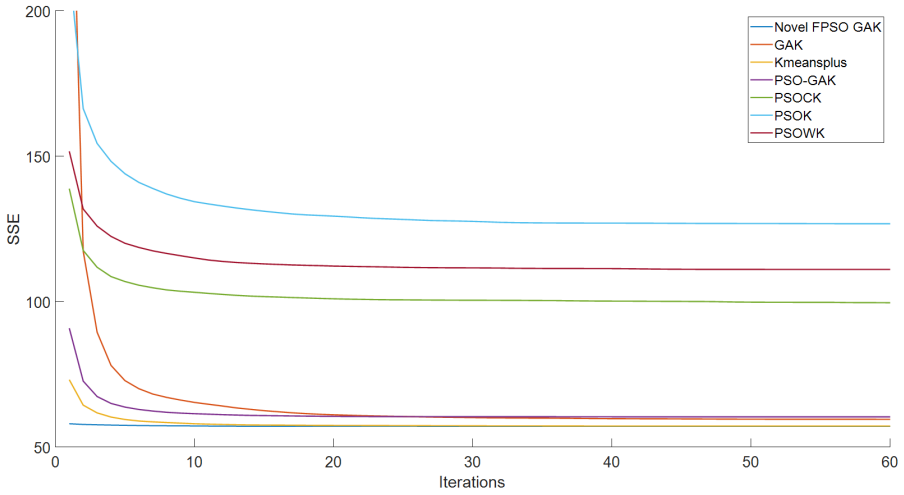

Figure 12. The SSE convergence curve of taxi GPS dataset from Roma(Italy).

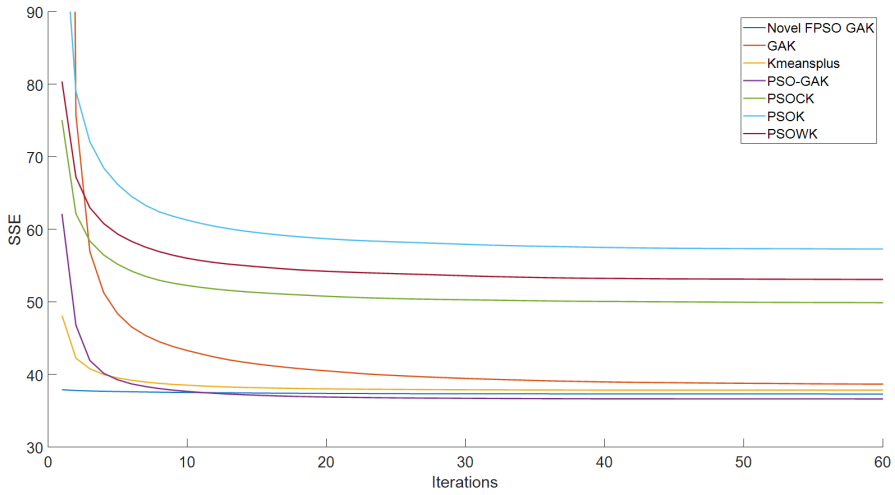

Figure 13. The SSE convergence curve of taxi GPS dataset from San Francisco (USA).

In Figure 9, the proposed algorithm begins to converge in the second generation and converges entirely in the fourth generation, and the SSE obtained is significantly smaller than other algorithms. In Figure 10, before the seventh generation, the convergence speed of the proposed algorithm is faster than other algorithms, and the SSE value is also relatively small. After the tenth generation, the SSE value of the proposed algorithm is slightly smaller than that of the K-means++ and GAK algorithms. In Figure 11 , proposed converges from about the second generation and converges entirely in the fifth generation. The SSE value is lower than that of other algorithms, showing that the clustering effect is excellent and the separation between clusters is significant, consequently demonstrating the algorithm's effectiveness. In Figure 12, the proposed converges from the second generation and entirely at about the fifth generation, with a very smooth convergence curve. In the meantime, the figure demonstrates that the SSE value is less than that of other algorithms, indicating that the clustering center captured by the proposed algorithm is effective and does not cause k-means clustering to fall into local optimal solutions. In Figure 13, the proposed converges from about the second generation and ultimately converges at the fifth generation. In the beginning stage, the SSE value of the algorithm is substantially lower than that of other algorithms, and after convergence, it is extremely stable.

Overall, experiments on various datasets indicate that the algorithm that combines the genetic algorithm, fussy system, and particle swarm algorithm virtually captures the optimum clustering centers more effectively.

3.3.2. Multi-Source Trajectory Data Set

To further validate the performance of the proposed algorithm, we employed a more complex multi-source dataset. In order to accurately analyze the above evaluation indexes, The number of clusterings is randomly generated between 130 and 150, each algorithm was iterated 120 times in a single run to guarantee convergence. Table 8 is the comparison tables for the average Silhouette coefficient (SC), the average degree of separation (SP), and the average SSE value of 20 training sessions respectively.

Table 8. Average evaluation indexes of 8 algorithms in multi-source dataset.

Algorithms	SC	SP	SSE
K-means++	0.9646	0.4415	1677.71
RS	0.9650	0.4402	1659.55
GAK	0.9633	0.4049	1668.38
PSOK	0.9640	0.4395	1687.45
PSOCK	0.9645	0.3317	1702.89
PSOWK	0.9654	0.4080	1645.93
PSO-GAK	0.9656	0.4238	1678.37
FPSO-GAK	0.9684	0.4628	1610.59

All eight algorithms converge well within less than 120 iterations in the multi-source dataset. The K-Means++ and Random Swap Clustering algorithms are much less time consuming than the other algorithms, and both achieve very good clustering results. The analysis of the experimental data shows that the Random Swap Clustering algorithm outperforms the K-Means++ algorithm. The Random Swap Clustering algorithm also outperforms single-mode heuristics in many cases. With reasonable time consumption, the evaluation indexes of the proposed algorithm are significantly better than those of the compared algorithms. This may be due to the fact that the proposed hybrid heuristic algorithm has stronger non-homogeneous spatial search capability, but also brings more computing resources and time consumption. Overall, the clustering results of the proposed algorithms all outperformed the comparison algorithms in the terms of above evaluation indexes. The above suggests that the proposed hybrid heuristic algorithms are effective.

3.4. Visual Presentation of Experimental Results

3.4.1. Visual Presentation of Taxis GPS Data Set

Figures 14–19 shows the centers of the clustering results for each algorithm in Baidu Map (put the clustering centers into the maps using the open software interface provided by Baidu Map). On the corresponding map, we intuitively perceive the distribution of clustering results based on various algorithms. According to pertinent maps, different clustering algorithms yield different clustering centers, indicating that different clustering results and identified urban hot areas are distinct. Since experiment GPS datasets are not cleaned according to road location limitations, the map may contain a very small number of outlier points.

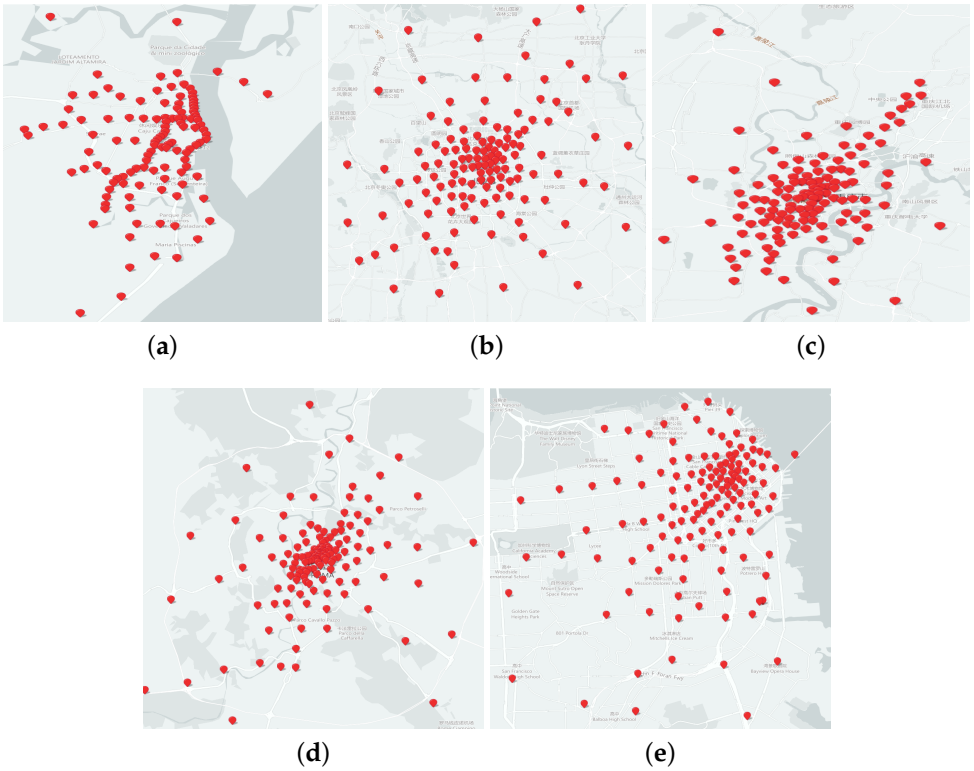

Figure 14. The urban hotspots captured by the GAK algorithm. (**a**) Aracaju; (**b**) Beijing; (**c**) Chongqing; (**d**) Roma; (**e**) San Francisco.

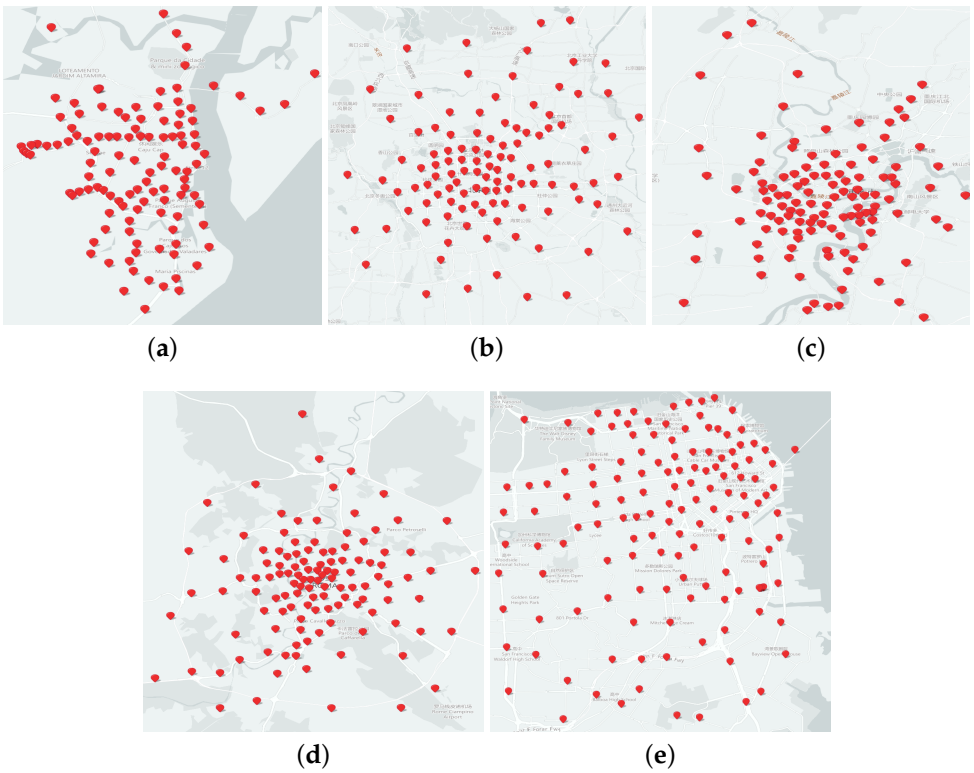

Figure 15. The urban hotspots captured by the PSO-GAK algorithm. (**a**) Aracaju; (**b**) Beijing; (**c**) Chongqing; (**d**) Roma; (**e**) San Francisco.

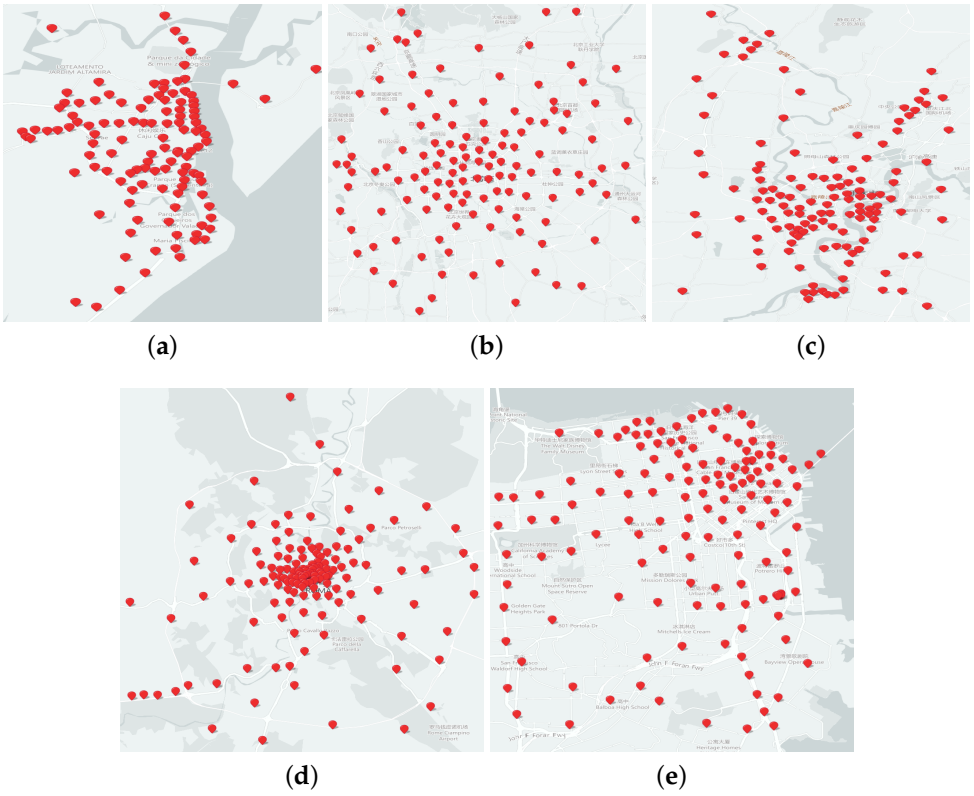

Figure 16. The urban hotspots captured by the PSOK algorithm. (**a**) Aracaju; (**b**) Beijing; (**c**) Chongqing; (**d**) Roma; (**e**) San Francisco.

Figure 17. The urban hotspots captured by the PSOCK algorithm. (**a**) Aracaju; (**b**) Beijing; (**c**) Chongqing; (**d**) Roma; (**e**) San Francisco.

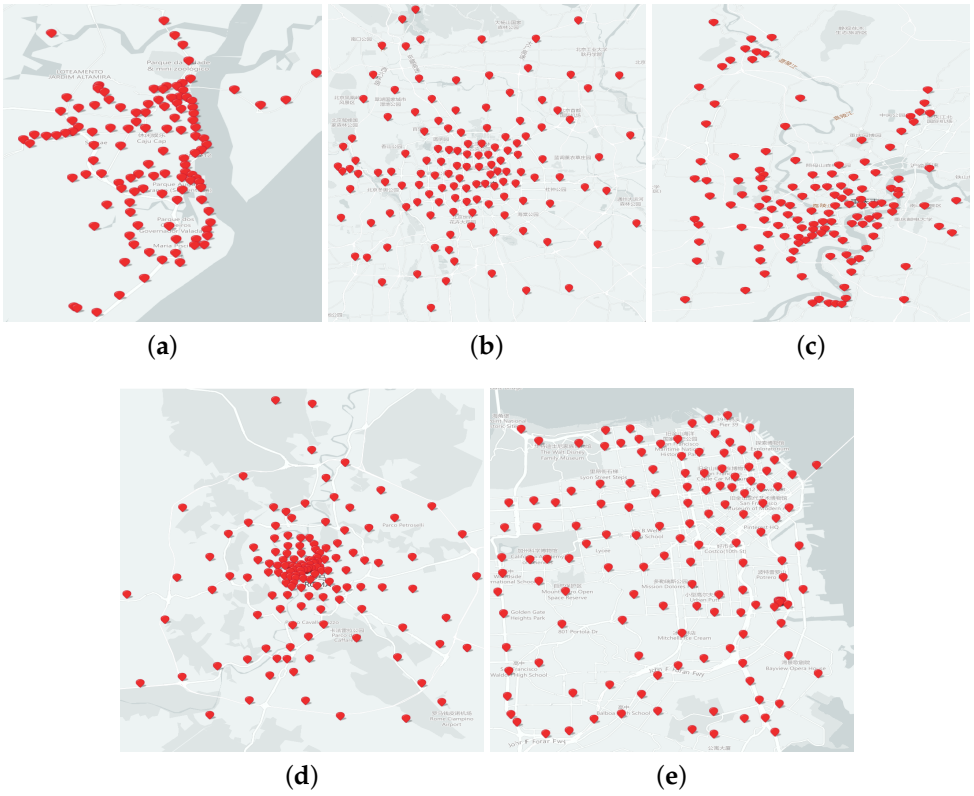

Figure 18. The urban hotspots captured by the PSOWK algorithm. (**a**) Aracaju; (**b**) Beijing; (**c**) Chongqing; (**d**) Roma; (**e**) San Francisco.

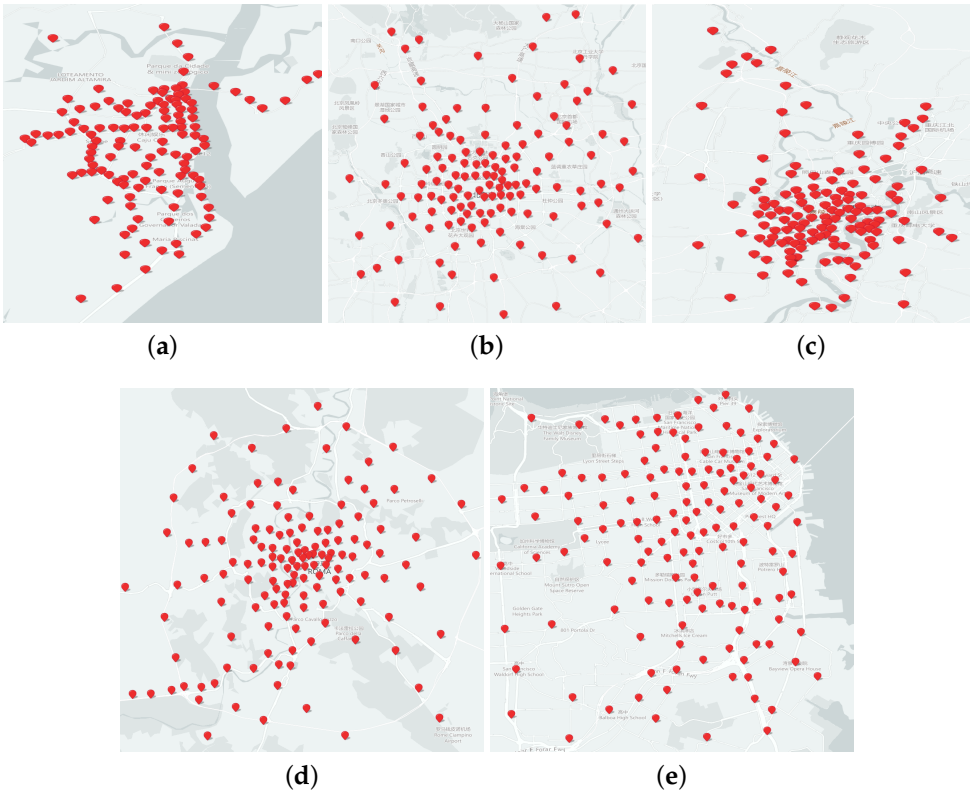

Figure 19. The urban hotspots captured by the novel FPSO-GAK algorithm. (**a**) Aracaju; (**b**) Beijing; (**c**) Chongqing; (**d**) Roma; (**e**) San Francisco.

3.4.2. Visual Presentation of Multi-Source Trajectory Data Set

Figure 20 shows the centers of the clustering results for each algorithm in multi-source trajectory data set. On the corresponding map, we intuitively perceive the distribution of clustering results based on various algorithms.

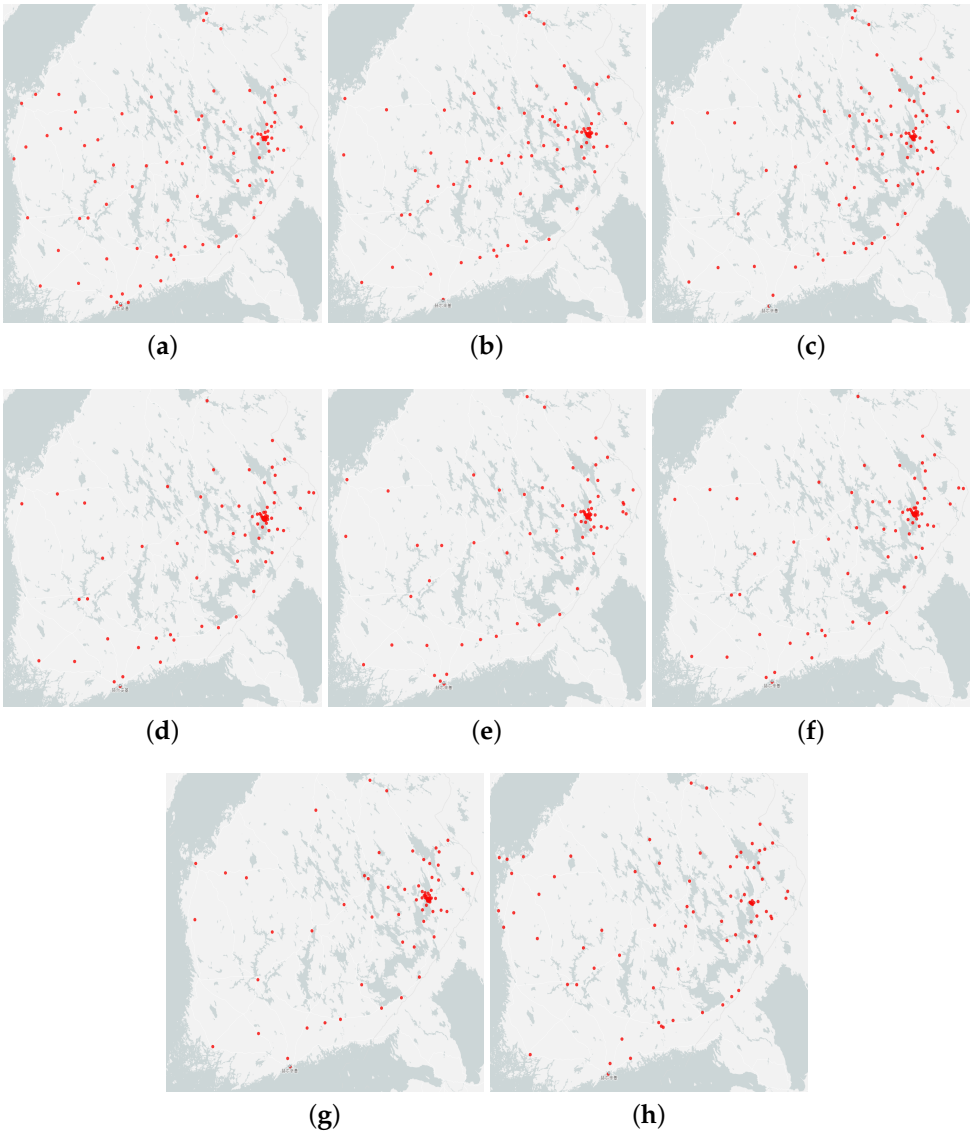

Figure 20. The urban hotspots in multi-source trajectory data set. (**a**) K-Means++; (**b**) GAK; (**c**) PSO-GAK; (**d**) PSOCK; (**e**) PSOK; (**f**) PSOWK; (**g**) RS; (**h**) FPSO-GAK.

3.5. Clustering Algorithm Complexity Analysis

The proposed noevel FPSO-GAK algorithm consists of three parts. To facilitate analysis, the following notation is used: the number of GPS data points is N, the number of clustering is K, the particles population of PSO is M. In the first part, the k-Means++ initialization method is employed to obtain particles M population. The time complexity of this part is $\mathcal{O}(P1) = MKlog(K)$. In the second part, the hybrid heuristic initialization is utilized to determine optimal initial clustering centers. The iterations number of PSO algorithm is T_1, The iterations number of GA algorithm is T_2, number of calculations to fitness function is T_3. The time complexity of the fuzzy system is related to the number of iterations and the number of fuzzy rules; it takes very little time and is negligible in terms of computational time complexity. The time complexity of the second part is $\mathcal{O}(P2) = MNKT_1 + MNKT_2 + MT_3$. In the third part, the K-Means clustering algorithm is

used to locate the urban hotspots. The iterations number of K-Means clustering algorithm is T_4.The time complexity of the third part is $\mathcal{O}(P3) = NKT_4$.

The time complexity of the proposed clustering algorithm is $\mathcal{O}(FPSO - GAK) = \mathcal{O}(P1) + \mathcal{O}(P2) + \mathcal{O}(P3) \approx NK((T_1 + T_2)M + T_4)$. In practical applications, K and M are very small relative to the number of GPS data points N. The number of iterations T of the K-means++ algorithm in which the random initialization procedure fails to acquire the better initialization tends to be extremely high for large-scale datasets. Due to the good non-homogeneous search capability of the proposed hybrid heuristic initialization algorithm, the value of $(T_1 + T_2)M$ is relatively small. Furthermore, the better-initialized centers makes T_4 would be much smaller than T. The result of their interaction makes: the complexity of the proposed algorithm is of the same order of magnitude as that of k-Means, but it will be slightly larger numerically.

In the actual experimental operation, the PSO-GAK algorithm in literature [55] consumes the most time since its complexity tends to $\mathcal{O}(N^2)$. The proposed algorithm can better adjust the search parameters automatically, and it will converge significantly faster than the PSO-GAK algorithm. The complexity of an intelligent algorithm is generally higher than the general iterative, recursive algorithm [56]. Collectively, it can be seen that the complexity of the proposed algorithm is much less than $\mathcal{O}(N^2)$ but greater than $\mathcal{O}(NKT)$. The complexity of the proposed algorithm is roughly linearly related to the number of samples, and it is very efficient and scalable for handling large data sets.

3.6. Clustering Results Similarity Analysis

External indexes have been commonly used to compare clustering algorithms by calculating how many pairs of data points are partitioned consistently between two clustering solutions. For both solutions to be consistent, a pair of points must be assigned to either the same cluster or a distinct cluster. This offers an estimation of point-level similarity but no direct information regarding cluster-level similarity. In most cases, information about the cluster-level similarity of the algorithm is of more interest than the estimation of point-level similarity. Fränti et al. [57] proposed Centroid index (CI) measures cluster-level differences of two solutions, and centroid similarity index (CSI) measures point-level differences of two solutions. The CI index has clear intuitive interpretation on how many clusters are differently located in the two solutions by an integer value. A higher CI value indicates a lower similarity between the two solutions when the number of cluster classes is determined. Considering the effect of the number of clustering, this paper employs a normalized adjusted CI index (ACI) to compare solutions with the following formula:

$$ACI = \frac{CI - abs(K_1 - k_2)}{min(K_1, k_2) - 1} \tag{7}$$

where K_1, k_2 represents the number of clusters in two solutions; $abs()$ represents the absolute value function; $min()$ represents the minimum function. Tables 9 and 10 are the comparison tables for the novel FPSO-GAK algorithm with other algorithms.

Table 9. ACI index for novel FPSO-GAK vs. other algorithms.

Algs / City	GAK		PSOK		PSOCK		PSOWK		PSO-GAK		K-Means++	
	Mean	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**
Aracaju	0.266	0.060	0.186	0.027	0.190	0.028	0.198	0.035	0.204	0.022	0.126	0.031
Beijing	0.182	0.075	0.111	0.027	0.200	0.062	0.190	0.054	0.161	0.042	0.096	0.022
Chongqing	0.307	0.042	0.151	0.022	0.208	0.029	0.198	0.025	0.207	0.026	0.131	0.037
Roma	0.364	0.039	0.162	0.030	0.249	0.023	0.243	0.027	0.272	0.025	0.181	0.024
San Francisco	0.353	0.044	0.204	0.041	0.205	0.040	0.222	0.037	0.231	0.040	0.165	0.039

mean: mean value ; std: standard deviation ; algs: algorithms. The above values are the statistical values after 20 rounds of independent running.

Table 10. CSI index for novel FPSO-GAK vs. other algorithms.

Algs / City	GAK		PSOK		PSOCK		PSOWK		PSO-GAK		K-Means++	
	Mean	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**
Aracaju	0.817	0.019	0.839	0.015	0.834	0.021	0.832	0.021	0.830	0.013	0.873	0.017
Beijing	0.737	0.018	0.750	0.022	0.746	0.018	0.747	0.018	0.753	0.020	0.753	0.019
Chongqing	0.750	0.015	0.789	0.029	0.774	0.019	0.781	0.020	0.770	0.018	0.790	0.016
Roma	0.747	0.023	0.808	0.019	0.795	0.020	0.782	0.021	0.772	0.019	0.802	0.018
San Francisco	0.700	0.021	0.722	0.024	0.721	0.027	0.727	0.024	0.717	0.026	0.745	0.021

mean: mean value ; std: standard deviation ; algs: algorithms. The above values are the statistical values after 20 rounds of independent running.

Table 9 shows that the ACI indices of FPSO-GAK and other algorithms are greater than 0.15, indicating a substantial difference in their clustering-level similarity. Obtaining the above outcomes under identical data input conditions reveals that the equivalent mathematical models of the FPSO-GAK algorithm differ significantly from those of other algorithms. However, the value of the ACI index is generally less than 0.25, indicating that the difference between equivalent mathematical models is not fundamental. The CSI indices in Table 10 are more concerned with point-level similarity, and their values indicate similar model analysis information.

4. Discussion

The K-Means algorithm can achieve very good results on convex, compressible datasets. Urban roads are usually distributed in a neighborhood manner, and the trajectory points of vehicles satisfy convexity and compressibility in a large local area. Using the K-Means algorithm to identify urban hotspots can yield effective results. K-Means algorithm is very sensitive to the initial center points. Excellent initial clustering centers significantly impact the performance of the k-Means clustering algorithm. K-Means is particularly effective in fine-tuning local cluster boundaries, but is incapable of solving global cluster locations. The organic combination of global perception and local fine-tuning is the key to obtaining good initial centers. Literature [50] proposed the method of random swap clustering, which is simple and efficient. Under the condition of a suitable KNN super parameter configuration, this method's random swap efficiency is extremely high, it is an exquisite lightweight clustering method. Inspired by the ideas of randomness and exchange, the FPSO-GAK algorithm combines a variety of heuristic algorithms with the premise of sacrificing a little efficiency to avoid the homogeneity of the random search process and can better obtain global perception ability. At the same time, local search parameters can be fine-tuned automatically to prevent over-dependence on super parameter.

In practical applications, in order to further improve the clustering efficiency, the inertia weight and learning factor can better meet the iterative requirements by adjusting fuzzy rules. At the same time, the combination mode of PSO and GA can also be changed. PSO can improve the phenomenon that GA is prone to early maturity through global optimization in the early stage. while PSO is prone to fall into local optimization in the late stage, the deficiency can be made up through the global optimization in the late stage of GA.

Compared with many existing research algorithms in this field, the proposed algorithm has a better and relatively efficient non-homogeneous search capability, which can effectively find excellent initial clustering centers and at the same time can reduce the sensitivity of K-Means algorithm to noise to a certain extent. However, the proposed algorithm is relatively complex in structure and has a relatively high time complexity compared to many lightweight algorithms. The proposed algorithm is a class of algorithms based on a heuristic mechanism, which can only guarantee to get a better solution and cannot guarantee to obtain the optimal solution for sure.

5. Conclusions

In this paper, a Novel FPSO-GAK clustering algorithm has been proposed to effectively solve the problems of difficulty in determining the sensitivity of initializing the clustering center for a K-means clustering algorithm. The Novel FPSO-GAK clustering algorithm has been employed to capture urban hotspots in metropolitan areas in several countries worldwide. After the experimental clustering operation was completed, the SC index, SP index, and SSE index were applied to evaluate the clustering performance. The proposed Novel FPSO-GAK clustering algorithm obtained better clustering results for urban hotspots in metropolitan areas over the GAK, PSOK, PSOCK, PSOWK, PSO-GAK, and K-means++ algorithm. The algorithm presented in this paper can better improve the quality of public service for people in cities. In addition, the Novel FPSO-GAK clustering algorithm proposed in this paper can also be applied to the clustering analysis of remote sensing ground object data, animal migration area analysis, urban subway data analysis, urban heating network optimization analysis, detailed analysis of power supply flow of power grid and other fields.

There are also some shortcomings of the algorithm proposed in the paper. On the one hand, we only use the GPS data of taxis. The data amount is relatively small, leading to an insufficient resolution in the refined analysis of the distribution relationship of urban hotspots. On the other hand, the Novel FPSO-GAK clustering algorithm does not add urban road constraints, leading to some hotspots falling into particular areas where traffic vehicles cannot be reached, such as large lake areas in cities. These deficiencies will be further improved in future research work.

Author Contributions: Conceptualization, W.D., X.Z., K.G. and Y.L.; methodology, W.D., X.Z. and Y.L.; software, Y.L. and J.G.; validation, Y.L. and J.G.; formal analysis, W.D., X.Z. and Y.L.; data curation, Y.L. and J.G.; writing—original draft preparation, Y.L.; writing—review and editing, W.D., X.Z. and Y.L.; visualization, Y.L. and J.G.; supervision, W.D., X.Z. and K.G. 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, 2020YFG0307, 2021YFS0407; 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.

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

References

1. Ge, D.; Long, H.; Qiao, W.; Wang, Z.; Sun, D.; Yang, R. Effects of rural–urban migration on agricultural transformation: A case of Yucheng City, China. *J. Rural. Stud.* **2020**, *76*, 85–95.
2. Cheshire, P.C.; Hay, D.G. *Urban Problems in Western Europe: An Economic Analysis*; Routledge: London, UK, 2017.
3. Leach, J.M.; Mulhall, R.A.; Rogers, C.D.; Bryson, J.R. Reading cities: Developing an urban diagnostics approach for identifying integrated urban problems with application to the city of Birmingham, UK. *Cities* **2019**, *86*, 136–144.
4. Gössling, S. Integrating e-scooters in urban transportation: Problems, policies, and the prospect of system change. *Transp. Res. Part D Transp. Environ.* **2020**, *79*, 102230.
5. Sarkar, S.; Wu, H.; Levinson, D.M. Measuring polycentricity via network flows, spatial interaction and percolation. *Urban Stud.* **2020**, *57*, 2402–2422.
6. Li, Q.; Bessell, L.; Xiao, X.; Fan, C.; Gao, X.; Mostafavi, A. Disparate patterns of movements and visits to points of interest located in urban hotspots across US metropolitan cities during COVID-19. *R. Soc. Open Sci.* **2021**, *8*, 201209.
7. Wu, D.; Wu, C. Research on the Time-Dependent Split Delivery Green Vehicle Routing Problem for Fresh Agricultural Products with Multiple Time Windows. *Agriculture* **2022**, *12*, 793
8. Li, X.; Zhao, H.; Yu, L.; Chen, H.; Deng, W.; Deng, W. Feature extraction using parameterized multi-synchrosqueezing transform. *IEEE Sens. J.* **2022**, *22*, 14263–14272.

9. Cai, L.; Jiang, F.; Zhou, W.; Li, K. Design and application of an attractiveness index for urban hotspots based on GPS trajectory data. *IEEE Access* **2018**, *6*, 55976–55985.
10. Lai, Y.; Lv, Z.; Li, K.C.; Liao, M. Urban traffic Coulomb's law: A new approach for taxi route recommendation. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 3024–3037.
11. Pan, L.; Zhang, X.; Li, X.; Li, X.; Lu, C.; Liu, J.; Wang, Q. Satellite availability and point positioning accuracy evaluation on a global scale for integration of GPS, GLONASS, BeiDou and Galileo. *Adv. Space Res.* **2019**, *63*, 2696–2710.
12. Strauss, J.; Miranda-Moreno, L.F. Speed, travel time and delay for intersections and road segments in the Montreal network using cyclist Smartphone GPS data. *Transp. Res. Part D Transp. Environ.* **2017**, *57*, 155–171.
13. 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. -Geo-Inf.* **2017**, *6*, 392.
14. Sumalee, A.; Ho, H.W. Smarter and more connected: Future intelligent transportation system. *Iatss Res.* **2018**, *42*, 67–71.
15. Zhao, P.; Hu, H. Geographical patterns of traffic congestion in growing megacities: Big data analytics from Beijing. *Cities* **2019**, *92*, 164–174.
16. Hsueh, Y.L.; Chen, H.C. Map matching for low-sampling-rate GPS trajectories by exploring real-time moving directions. *Inf. Sci.* **2018**, *433*, 55–69.
17. He, S.; Bastani, F.; Abbar, S.; Alizadeh, M.; Balakrishnan, H.; Chawla, S.; Madden, S. RoadRunner: Improving the precision of road network inference from GPS trajectories. In Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, 6–9 November 2018; pp. 3–12.
18. Bai, F.; Feng, H.; Xu, Y. Identifying the hotspots in urban areas using taxi GPS trajectories. In Proceedings of the 2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Huangshan, China, 28–30 July 2018; pp. 900–904.
19. 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.
20. Mariescu-Istodor, R.; Fränti, P. CellNet: Inferring road networks from GPS trajectories. *ACM Trans. Spat. Algorithms Syst. (TSAS)* **2018**, *4*, 1–22.
21. Zhang, Y.; Zhang, Z.; Huang, J.; She, T.; Deng, M.; Fan, H.; Xu, P.; Deng, X. A hybrid method to incrementally extract road networks using spatio-temporal trajectory data. *ISPRS Int. J. -Geo-Inf.* **2020**, *9*, 186.
22. Shafabakhsh, G.A.; Famili, A.; Bahadori, M.S. GIS-based spatial analysis of urban traffic accidents: Case study in Mashhad, Iran. *J. Traffic Transp. Eng. (Engl. Ed.)* **2017**, *4*, 290–299.
23. Wang, Y.; Qin, K.; Chen, Y.; Zhao, P. Detecting anomalous trajectories and behavior patterns using hierarchical clustering from taxi GPS data. *ISPRS Int. J. -Geo-Inf.* **2018**, *7*, 25.
24. Zhang, K.; Sun, D.; Shen, S.; Zhu, Y. Analyzing spatiotemporal congestion pattern on urban roads based on taxi GPS data. *J. Transp. Land Use* **2017**, *10*, 675–694.
25. Chen, H.; Miao, F.; Chen, Y.; Xiong, Y.; Chen, T. A hyperspectral image classification method using multifeature vectors and optimized KELM. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2781–2795.
26. Yao, R.; Guo, C.; Deng, W.; Zhao, H. A novel mathematical morphology spectrum entropy based on scale-adaptive techniques. *ISA transactions* **2022**, *126*, 691–702.
27. An, Z.; Wang, X.; Li, B.; Xiang, Z.; Zhang, B. Robust visual tracking for UAVs with dynamic feature weight selection. *Appl. Intell.* **2022**, 675–694.
28. Berkhin, P. A survey of clustering data mining techniques. In *Grouping Multidimensional Data*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 25–71.
29. Gan, G.; Ma, C.; Wu, J. *Data Clustering: Theory, Algorithms, and Applications*; SIAM: Philadelphia, PA, USA, 2020.
30. Kapoor, A.; Singhal, A. A comparative study of K-Means, K-Means++ and Fuzzy C-Means clustering algorithms. In Proceedings of the 2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT), Ghaziabad, India, 9–10 February 2017; pp. 1–6.
31. Krishna, K.; Murty, M.N. Genetic K-means algorithm. *IEEE Trans. Syst. Man, Cybern. Part B (Cybern.)* **1999**, *29*, 433–439.
32. Lu, Y.; Lu, S.; Fotouhi, F.; Deng, Y.; Brown, S.J. FGKA: A fast genetic k-means clustering algorithm. In Proceedings of the 2004 ACM Symposium on Applied Computing, Nicosia, Cyprus, 14–17 March 2004; pp. 622–623.
33. Islam, M.Z.; Estivill-Castro, V.; Rahman, M.A.; Bossomaier, T. Combining K-Means and a genetic algorithm through a novel arrangement of genetic operators for high quality clustering. *Expert Syst. Appl.* **2018**, *91*, 402–417.
34. Dowlatshahi, M.B.; Nezamabadi-Pour, H. GGSA: A grouping gravitational search algorithm for data clustering. *Eng. Appl. Artif. Intell.* **2014**, *36*, 114–121.
35. Dowlatshahi, M.; Derhami, V.; Nezamabadi-Pour, H. Fuzzy particle swarm optimization with nearest-better neighborhood for multimodal optimization. *Iran. J. Fuzzy Syst.* **2020**, *17*, 7–24.
36. Zhou, X.; Ma, H.; Gu, J.; Chen, H.; Deng, W. Parameter adaptation-based ant colony optimization with dynamic hybrid mechanism. *Eng. Appl. Artif. Intell.* **2022**, *114*, 105–139.
37. Huang, J.Z.; Ng, M.K.; Rong, H.; Li, Z. Automated variable weighting in k-means type clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 657–668.

38. Gao, S.; Janowicz, K.; Couclelis, H. Extracting urban functional regions from points of interest and human activities on location-based social networks. *Trans. GIS* **2017**, *21*, 446–467.
39. Liu, X.; Tian, Y.; Zhang, X.; Wan, Z. Identification of urban functional regions in chengdu based on taxi trajectory time series data. *ISPRS Int. J. -Geo-Inf.* **2020**, *9*, 158.
40. Grønlund, A.; Larsen, K.G.; Mathiasen, A.; Nielsen, J.S.; Schneider, S.; Song, M. Fast exact k-means, k-medians and Bregman divergence clustering in 1D. *arXiv* **2017**, arXiv:1701.07204.
41. Hatamlou, A. Black hole: A new heuristic optimization approach for data clustering. *Inf. Sci.* **2013**, *222*, 175–184.
42. Sarkar, T.; Salauddin, M.; Hazra, S.K.; Chakraborty, R. Comparative study of predictability of response surface methodology (RSM) and artificial neural network-particle swarm optimization (ANN-PSO) for total colour difference of pineapple fortified rasgulla processing. *Int. J. Intell. Netw.* **2020**, *1*, 17–31.
43. Sedighizadeh, D.; Masehian, E.; Sedighizadeh, M.; Akbaripour, H. GEPSO: A new generalized particle swarm optimization algorithm. *Math. Comput. Simul.* **2021**, *179*, 194–212.
44. Lee, J.H.; Song, J.Y.; Kim, D.W.; Kim, J.W.; Kim, Y.J.; Jung, S.Y. Particle swarm optimization algorithm with intelligent particle number control for optimal design of electric machines. *IEEE Trans. Ind. Electron.* **2017**, *65*, 1791–1798.
45. Liu, X.; Zhang, D.; Zhang, T.; Zhang, J.; Wang, J. A new path plan method based on hybrid algorithm of reinforcement learning and particle swarm optimization. *Eng. Comput.* **2021**, *39*, 993–1019.
46. Fränti, P.; Sieranoja, S. K-means properties on six clustering benchmark datasets. *Appl. Intell.* **2018**, *48*, 4743–4759.
47. Yuan, C.; Yang, H. Research on K-value selection method of K-means clustering algorithm. *J* **2019**, *2*, 226–235.
48. Niknam, T.; Amiri, B. An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis. *Appl. Soft Comput.* **2010**, *10*, 183–197.
49. Zhang, W.; Liu, Y. Multi-objective reactive power and voltage control based on fuzzy optimization strategy and fuzzy adaptive particle swarm. *Int. J. Electr. Power Energy Syst.* **2008**, *30*, 525–532.
50. Fränti, P. Efficiency of random swap clustering. *J. Big Data* **2018**, *5*, 1–29.
51. Yuan, J.; Zheng, Y.; Xie, X.; Sun, G. T-drive: Enhancing driving directions with taxi drivers' intelligence. *IEEE Trans. Knowl. Data Eng.* **2011**, *25*, 220–232.
52. Yuan, J.; Zheng, Y.; Xie, X.; Sun, G. Driving with knowledge from the physical world. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 21–24 August 2011; pp. 316–324.
53. Piorkowski, M.; Sarafijanovic-Djukic, N.; Grossglauser, M. CRAWDAD Data Set Epfl/Mobility. 24 February 2009. Available online: http://crawdad.org/epfl/mobility/20090224 (accessed on 15 January 2022).
54. Fränti, P.; Nenonen, H. Modifying Kruskal algorithm to solve open loop TSP. In Proceedings of the Multidisciplinary International Scheduling Conference (MISTA), Ningbo, China, 12–15 December 2019.
55. Garg, H. A hybrid PSO-GA algorithm for constrained optimization problems. *Appl. Math. Comput.* **2016**, *274*, 292–305.
56. Rahman, M.A.; Islam, M.Z. A hybrid clustering technique combining a novel genetic algorithm with K-Means. *Knowl.-Based Syst.* **2014**, *71*, 345–365.
57. Fränti, P.; Rezaei, M.; Zhao, Q. Centroid index: Cluster level similarity measure. *Pattern Recognit.* **2014**, *47*, 3034–3045.