mathematics

MDPI

Article

Station Layout Optimization and Route Selection of Urban Rail Transit Planning: A Case Study of Shanghai Pudong International Airport

Pei Yin and Miaojuan Peng *

Department of Civil Engineering, School of Mechanics and Engineering Science, Shanghai University, Shanghai 200444, China
* Correspondence: mjpeng@shu.edu.cn

Abstract: In this paper, a cost-oriented optimization model of station spacing is presented to analyze the influencing factors of station spacing and layout near Shanghai Pudong International Airport. The Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm is used to cluster and analyze the high population density, and optimize the station layout in the southwest of Pudong International Airport. A spatial analysis of the land use and geological conditions in Pudong New Area is given. Combining the optimal station spacing, ideal location and spatial analysis, five routing schemes to Pudong International Airport are proposed. The DBSCAN and K-means algorithms are used to analyze the "PDIA-SL" dataset. The results show that the space complexity of the HDBSCAN is $O(825)$, and the silhouette coefficient is 0.6043, which has obvious advantages over the results of DBSCAN and K-means. This paper combines urban rail transit planning with the HDBSCAN algorithm to present some suggestions and specific route plans for local governments to scientifically plan rail transit lines. Meanwhile, the research method of station layout, which integrates station spacing, ideal location and spatial analysis optimization, is pioneering and can provide a reference for developing rail transit in metropolises.

Keywords: urban rail transit; station layout optimization; route planning; influencing factors; HDBSCAN; GIS analysis

MSC: 90B20

Citation: Yin, P.; Peng, M. Station Layout Optimization and Route Selection of Urban Rail Transit Planning: A Case Study of Shanghai Pudong International Airport. *Mathematics* **2023**, *11*, 1539. https://doi.org/10.3390/math11061539

Academic Editors: Ripon Kumar Chakrabortty, António Lopes and José R. Fernánde

Received: 19 January 2023
Revised: 15 February 2023
Accepted: 4 March 2023
Published: 22 March 2023

1. Introduction

Pudong International Airport is an essential domestic airport and one of the two major airports in Shanghai. In 2019 (COVID-19 has a great influence on urban functions [1], which leads to the decrease after 2019), the passenger throughput of Pudong International Airport reached 76.1534 million passengers, the cargo throughput was 3.6356 million tons, and the number of take-off and landing flights was 511,846 [2]. With the rapid development of the national economy and aviation technology, the passenger flow growth at the airport continues to accelerate. The pressure on landside traffic to Pudong International Airport has increased dramatically, and the contradiction between supply and demand of passenger transportation has become increasingly prominent [3]. Constructing the connection between the city and the airport is a critical task in urban transportation planning, and it is significant to optimize the station layout near the airport rationally. The optimization can promote the economic and comprehensive development of the city [4].

The development of urban rail transit in developed countries abroad started earlier, and there are many theoretical studies on the urban rail transit network design. Mohaymany and Gholami (2010) proposed a multi-model optimization method to minimize social, travel and operating cost [5]. Lai (2012) used genetic algorithm (GA) to optimize the station based on geographic information system (GIS) data [6]. Saidi et al. (2016) analyzed the route

selection, and origin-destination (OD) demand distribution in an urban rail transit network based on the total travel time cost of passengers and proposed a benefit optimization model for determining the feasibility and optimality of circular lines [7]. Compared with foreign countries, domestic research on station layout optimization started late but developed rapidly. Lv et al. (2013) established a bilevel programming model for station layout based on Alternative-Use Mode and used a simulated annealing algorithm (SAA) for analysis [8]. Chai et al. (2019) proposed a neighborhood search algorithm based on SAA for NP-hard problem whose factors increase exponentially with the network size [9]. Xu et al. (2021) proposed a bilevel multi-objective model neighborhood which integrates the situation of land use and regional traffic network and offered a maximum-minimum ant system (MMAS) combined with the Frank-Wolfe algorithm to solve the NP-Hard problem [10].

Currently, the research on clustering methods at home and abroad mainly focuses on optimizing the serial method. Serial clustering algorithms have been widely studied and applied in the field of statistical operations and data analysis, such as K-means (partition clustering algorithm) [11], CURE (a new hierarchical clustering algorithm) [12], STING (grid-based spatial clustering algorithm) [13] and DBSCAN (density-based clustering algorithm) [14]. The DBSCAN algorithm, which considers the distribution of data points, can judge any shape of clusters and effectively overcome the influence of outliers. However, this algorithm is sensitive to parameter changes, and slight changes in parameters will lead to significant differences in clustering results. And it is necessary to judge successively whether each data point is the core. Through the special processing of shared boundary points, the HDBSCAN introduces the idea of hierarchical clustering based on DBSCAN. This not only corrects the poor clustering results caused by the improper selection of the neighborhood radius (*eps*) in the DBSCAN but also shields the algorithm's sensitivity to parameters [15]. As a novel improved algorithm based on DBSCAN, the HDBSCAN is rarely used in engineering. Melvin et al. (2018) used the HDBSCAN to group biopolymer residues and grouped the residues based on the spatial proximity of the correlation matrix [16]. Ghamarian and Marquis (2019) improved solute clusters for atom probe tomography (APT) using the HDBSCAN [17]. Wang et al. (2021) used the HDBSCAN to adaptively cluster the shape features of ship trajectories, which has a good clustering effect on ship trajectories in complex waters [18]. Liu et al. (2022) fused line features on the radar signature of leg lasers and used the HDBSCAN to cluster target information and obtain the best position [19]. In summary, HDBSCAN has the most accurate and efficient performance of finding target clusters among the current density-based clustering algorithms, and there is currently no application of the HDBSCAN algorithm in rail transit network design.

At the same time, there are some problems in the design and planning of urban rail transit stations in Shanghai [20]. The specific manifestations are as follows:

(1) Preliminary analysis of passenger flow, emphasizing layout form and ignoring operational efficiency, results in narrow station spacing and the passenger flow far below expectations;
(2) Local governments are affected by the image project when encountering network planning and administrative intervention, or the adjustment of rail transit planning due to the change of local leaders is relatively standard;
(3) The increase in government debt is caused by excessive investment.

Taking Pudong International Airport as an example, the dense stations of Metro Line 2 lead to a long running time. The average time it takes for passengers to arrive at Pudong International Airport by rail transit is about 1.5 h, while the time it takes for a car is only about 1 h [21]. According to statistics, after the suspension of Metro Line 2 at 10:00 p.m., 8.8% of flights arriving at Pudong International Airport and 6.7% of departures still takes place [22]. The premature suspension of the subway caused poor service connections. Pudong International Airport has only two rail transit lines that can connect to the urban district, namely Metro Line 2 and Maglev Line. The rail transit located in the southwest of the airport is developing slowly. As a result, passengers in Songjiang, Minhang, Fengxian, Jinshan, and Pudong New Area must transfer multiple rail transit lines if they want to

take the subway to reach Pudong International Airport, which leads to inconvenience and discomfort of passengers significantly.

This paper analyzes the influencing factors of station layout, including the factors of station spacing and location. By analyzing the relationship between construction investment and income, a cost-oriented optimization model of station spacing is established, and the working principle of the HDBSCAN is introduced. Taking Pudong International Airport as an example, this paper selects large residential, office and industrial buildings with high population density in the southwest of the airport as the dataset. And this paper uses the HDBSCAN to analyze the dataset to obtain the best clusters. Finally, the ideal point for station selection is established by finding the geometric centroid in the cluster. At the same time, based on the same dataset and hardware conditions, this paper uses DBSCAN and K-means to analyze this case. The results show that the HDBSCAN has obvious advantages in terms of space complexity, silhouette coefficient and clustering effect. In addition, this paper combines the ideal locations, the geospatial conditions and the optimal station spacing. Finally, it compiles the ideal stations into five schemes of the rail transit route selection. The main contributions of this paper are as follows:

(1) This paper combines urban rail transit planning and HDBSCAN algorithm to present some suggestions and specific route schemes for the local government to scientifically plan rail transit lines.
(2) This paper comprehensively considers the construction cost, operating cost, social benefits, land development value-added along the Metro line and travel time cost of passengers, and proposes a more complete optimization model of station spacing.
(3) The method for studying the station layout, which integrates station spacing, ideal location and spatial analysis optimization, is pioneering and can provide a reference for developing rail transit in metropolises.
(4) This paper selects data points in high-population areas for analysis, and the layout planning of rail transit stations based on this dataset can relieve the traffic pressure of community residents and guarantee all-age-friendly travel in the community.

2. Analysis of Influencing Factors on Layout Optimization of Rail Transit Stations

2.1. Influencing Factors of Station Spacing

The station spacing will directly affect the passenger flow distribution of urban rail transit, and it will also change the passengers' travel decisions. When the station spacing is small, the number of passengers traveling in the unit interval is relatively large; then the passenger flow will decrease with the gradual increase of the station spacing, and the number of passengers will stop changing at a critical point; after passing the critical point, the passenger flow of the unit interval ascends with the increase of the station spacing, but the upward trend is relatively moderate [23]. The change trend of passenger flow with the increase of travel distance is approximately subject to Poisson distribution, and the formula of passenger flow with travel distance is

$$y = 0.0004x^6 - 0.017x^5 - 1.2944x^4 + 100.4x^3 - 2326.7x^2 + 18585x + 3896 \quad (1)$$

where y is the predicted full-day passenger flow of a certain travel distance, x is the travel distance of passengers by rail transit (km) [24].

The station spacing will also affect the number of stations, which will greatly affect the construction and operating cost. When the distance between stations is small, the number of stations will increase correspondingly, and the construction and operating cost will increase. Although the setting of large station spacing saves investment, it increases the walking time of passengers. Compared with other urban transportations, rail transit will reduce its attraction to passenger flow to a large extent. This will also cause part of the passenger flow to transfer to adjacent stations, increasing the transport load of adjacent stations [25]. In this paper, more consideration is given to the influence of station spacing on construction cost, which is the number of stations determined by station spacing and the

cost of a single station. The formula of average daily construction cost within the service period is

$$Z_{con} = \frac{C_{con} \cdot \left[\text{int}\left(\frac{L}{d}\right) + 1\right]}{365 \cdot n} \tag{2}$$

where Z_{con} is the daily average construction cost of all rail transit stations (CNY), C_{con} is the average construction cost of a single station (CNY), d is the distance between rail transit stations (km), L is the total length of rail transit lines (km), n is the number of years of service period (year) [26].

Each station has a certain number of staff and various supporting facilities, and the local government will inevitably pay these personnel and equipment with salary and financial allocation. The reduction of the number of stations will lead to the diminishment of relevant personnel and facilities, and the resource consumption will also decreased, thereby reducing operating costs. Therefore, the operating cost of urban rail transit can be summarized into three aspects:

(1) The salaries of operating personnel;
(2) The cost of equipment update and maintenance;
(3) The cost of daily purchase of power and other resources.

The operating cost of a rail transit station can be defined as

$$\begin{cases} Z_{ope} = C_{ope} \cdot \left[\text{int}\left(\frac{L}{d}\right) + 1\right] \\ C_{ope} = C_1 + C_2 + C_3 \end{cases} \tag{3}$$

where Z_{ope} is the average daily operating cost of all rail transit stations (CNY), C_{ope} is the average daily operating cost of a single station (CNY), C_1 is the sum of the average daily wages of all employees at a single station (CNY), C_2 is the average daily resource consumption cost of a single station (CNY), C_3 is the average daily equipment update and maintenance cost for a single station (CNY) [27].

Factors such as population density, building density and commercial areas have an important impact on passenger flow. They are directly related to the gathering intensity and distribution state of passenger distributing center, and also indirectly affect land development intensity and use patterns. The construction investment of the station is generally used to purchase resources such as construction materials, labor force, facilities and equipment, which drives the economic benefits of secondary industries. Meanwhile, the production sector will use the funds to purchase the sources and services, which will further drive the economic growth of more industries. And the social and economic growth will be a certain multiple of the initial investment as a result. The formula of social benefits generated by station investment is

$$Z_{soc} = \eta \cdot Z_{con} \tag{4}$$

where Z_{soc} is the social benefit generated by construction investment (CNY), η is the growth multiple [28].

The operation of urban rail transit will drive the growth of public and economic activities along the Metro line, thereby driving land development, which will greatly increase the added value of land. The formula of average daily social benefit is

$$Z_{land} = \sum_{j=year}^{i=1,2,3} \frac{(c_i \cdot \beta_i \cdot \theta_i)_j \cdot \pi r^2 \cdot \left[\text{int}\left(\frac{L}{d}\right) + 1\right] \cdot 10^4}{365 \cdot n_1} \tag{5}$$

where Z_{land} is the average daily social benefit (CNY) generated by land appreciation along the line; c_1, c_2 and c_3 are the average value-added of residential, commercial and industrial properties (Ten thousand CNY), respectively; β_1, β_2 and β_3 are the proportion of residential, commercial and industrial properties, respectively; θ_1, θ_2 and θ_3 are the plot

ratio of residential, commercial and industrial properties within the attractive range of stations, respectively; j is the selected data year, r is the attractive range of stations (m), and n_1 is the calculated period (year) [29].

2.2. Influencing Factors of Station Selection

Describing cities' scale mainly includes population, economy and land use [30,31]. The urban population determines the travel volume of passengers, and the land use affects the distribution of residents' travel demand, and the scale of urban rail transit and station layout [32]. Due to the high cost of urban rail transit, the smooth construction of stations is limited by the urban economy and scale [33]. To ensure the operation effect, it is necessary to attract the surrounding passenger flow to the greatest extent to ensure sufficient passenger volume, thereby promoting the healthy development. The distribution of point of interest (POI) in Shanghai considered in this paper is shown in Figure 1.

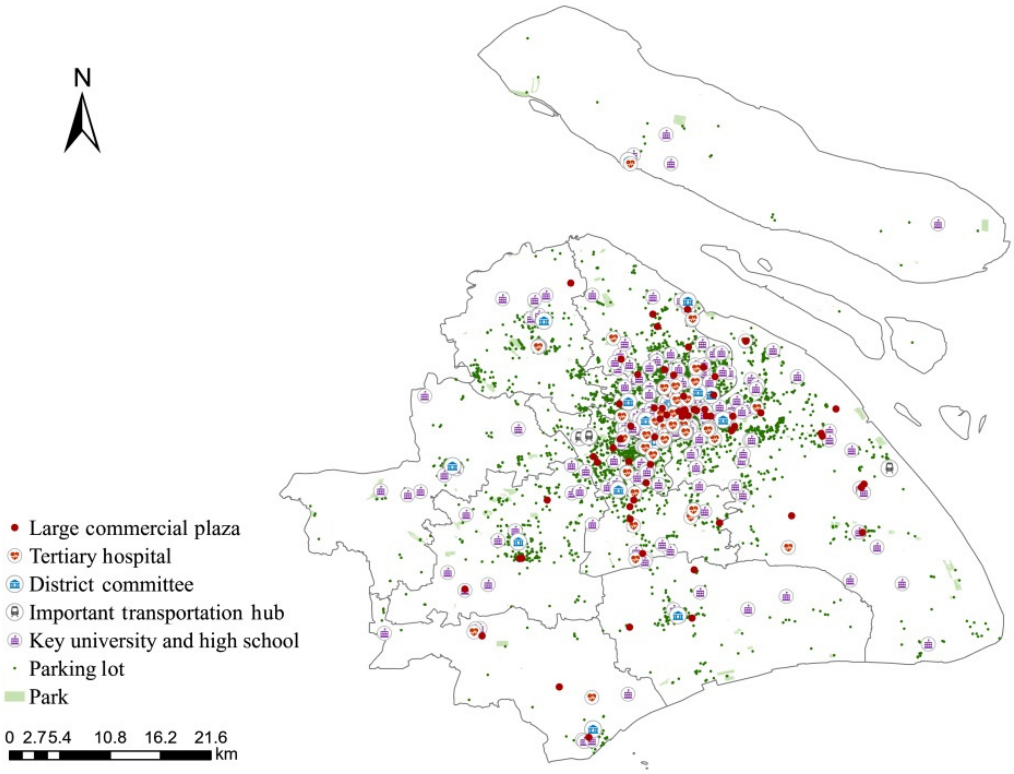


Figure 1. Distribution of POI in Shanghai considered in this paper.

The shape of cities greatly affects the layout of rail transit stations [34]. For plain cities like Shanghai, the main consideration is to connect major economic points, passenger distributing center, and comprehensive hubs to form a basic network, and then refine the plan of station layout. At the same time, various shapes and land use patterns also have an impact on passenger travel time and distance. Megacities such as Beijing and Shanghai have always adhered to the layout pattern of decentralized group [35]. Decentralization should be able to use the connections established between Metro lines to reduce the time required for travel between groups; the group development must consider the balanced layout of the central urban and suburban lines [36]. The land use patterns in Shanghai mainly include cultivated land, residential, commercial and industrial land, and a small amount of grassland, forest land and wetland [37,38]. Compared with industrial land, if there is no industrial park in the area, the district government is more inclined to sell residential or commercial land [39,40]. The distribution of land use patterns in Pudong New District is shown in Figure 2.

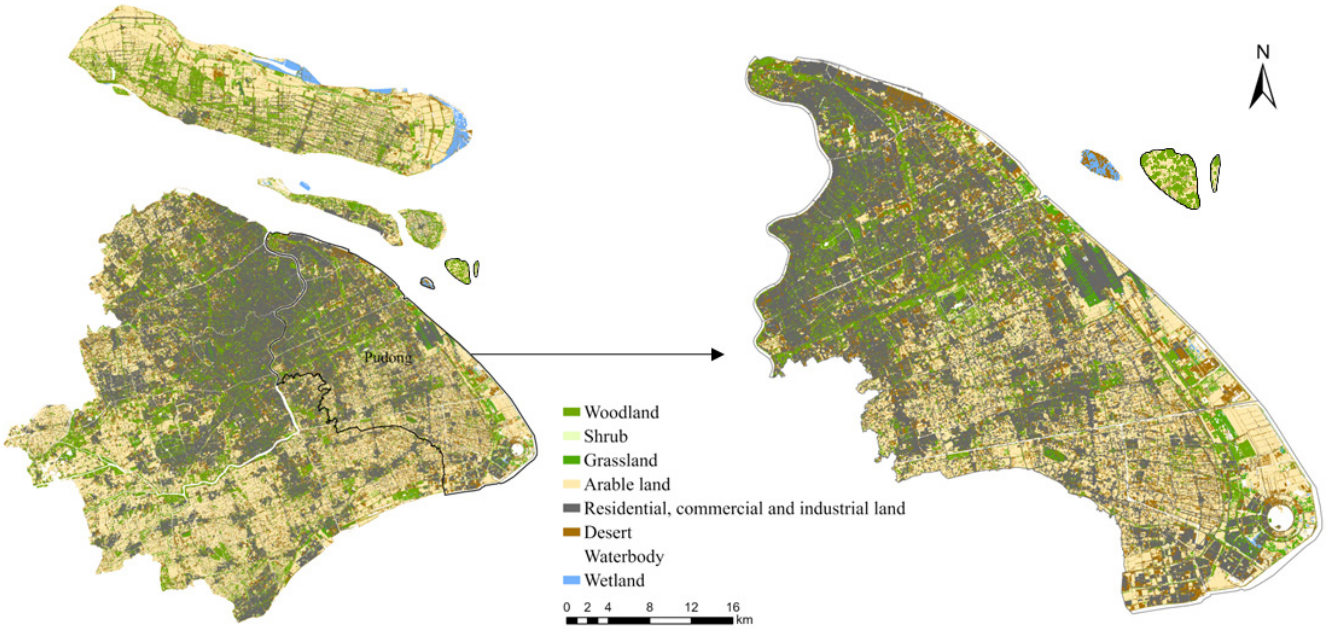


Figure 2. Distribution of land use types in Pudong New Area.

The passenger flow of each rail transit entrance and exit is different, and the passenger flow of different sections on the same line is disparate [41]. Regarding travel time, it may appear that the peak time for passenger transport is not the same, and the passenger flow entering the station is not equal to the leaving. In addition, a major task in subway line design is to determine the vertical alignment, which greatly impacts energy consumption and construction investment. The entire Metro line is divided into multiple lines by the stations, and the stations form nodes along the line. The design criteria such as the the vertical alignment's maximum and minimum slope length are different due to functional requirements. Generally, the vertical alignment of a station should be as flat as possible to ensure the safe parking of trains. Due to the complexity of the surrounding environment along the Metro lines, the vertical alignment design of the subway must deal with various constraints. Stations with high land use degree in the surrounding areas are closely related to large-scale residential and business districts, and the stations at these locations can attract passengers [42]. Regarding geological factors, mountain areas have relatively large topographical undulations, high altitudes, and diverse geomorphic features. In the planning and design of urban roads, due to frequent encounters with mountains, hills, rivers, buildings and poor geological conditions, it is necessary to avoid large-scale unfavorable geological bodies and group architectures, resulting in turning corners in the road. In the meantime, wetland protection areas should also be avoided. The Wetlands Conservation Law of the People's Republic of China requires strict control of the development and occupation of wetlands [43]. Wetlands in nature reserves are strictly prohibited from reclaimed, occupied or changed at will. During the planning process of rail transit stations, cultivated land and groundwater sources should be avoided as much as possible to avoid pollution and damage to domestic water. For the districts with poor geological conditions, the station layout and appropriate routing direction should be selected to minimize the impact on the natural environment. The 12.5 m DEM data is collected by ALOS' PALSAR sensor, and the horizontal and vertical accuracy of the data can reach 12 m. The digital elevation model (DEM) map of the 12.5 m elevation of Pudong New Area is shown in Figure 3.

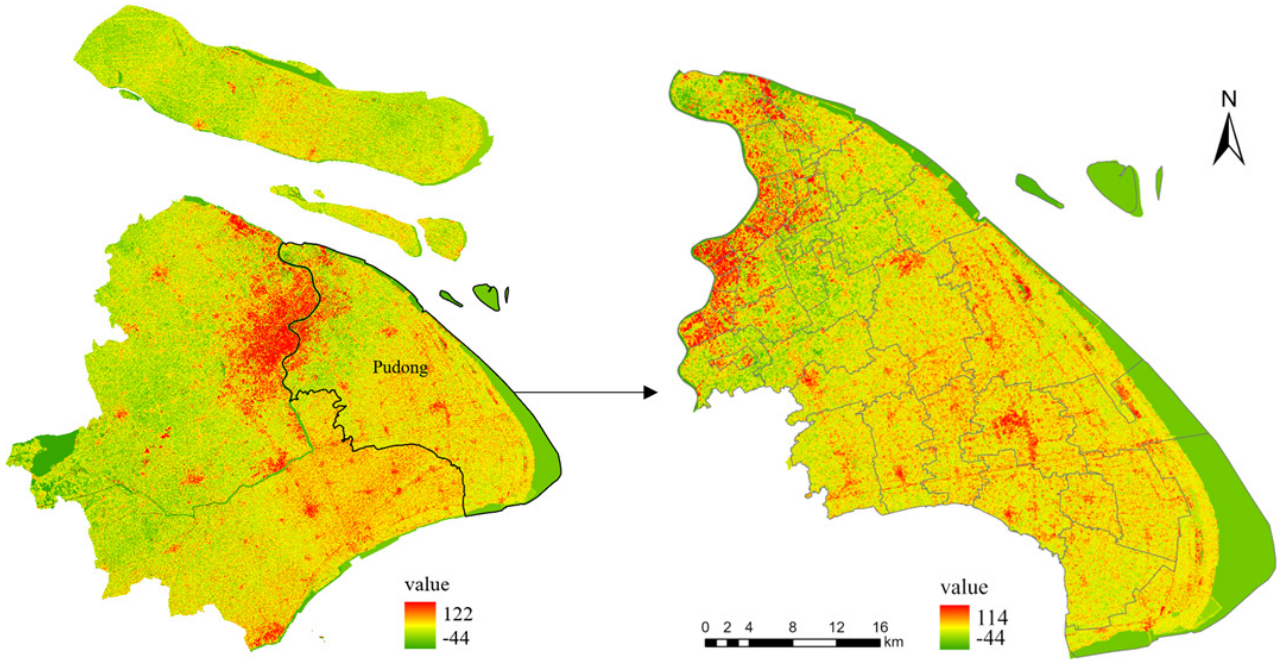


Figure 3. DEM map of the 12.5-m elevation in Pudong New Area.

3. Analysis of Influencing Factors on Layout Optimization of Rail Transit Stations

3.1. The Optimization Model of Station Spacing

The passenger flow of a station is directly related to the passenger ticket revenue. After the station is completed and put into use, the main source of its operating income is the passenger ticket revenue. The Shanghai rail transit system adopts the metered fare system: within 0–6 km, the toll is 3 CNY; between 6–16 km, the toll is 4 CNY; between 16–26 km, the toll is 5 CNY; between 26–36 km, the toll is 6 CNY; between 36–46 km, the toll is 7 CNY; between −56 km, the fee is 8 CNY. The pricing process can be simplified to the following formula, i.e.,

$$f(x) = \begin{cases} 3 & (0 < x \leq 6) \\ 4 & (6 < x \leq 16) \\ 5 & (16 < x \leq 26) \\ 6 & (26 < x \leq 36) \\ 7 & (36 < x \leq 46) \\ 8 & (46 < x \leq 56) \end{cases} \tag{6}$$

where $f(x)$ is a simplified function of the urban rail transit fare (CNY).

Based on Equations (1) and (6), the formula of rail transit fare revenue for a whole day is

$$Z_{tic} = 3\int_{\mathrm{d}}^{6} ydx + 4\int_{6}^{16} ydx + 5\int_{16}^{26} ydx + 6\int_{26}^{36} ydx + 7\int_{36}^{46} ydx + 8\int_{46}^{56} ydx \tag{7}$$

where d is the distance between rail transit stations (km).

Passenger travel time can be converted into time cost, and then calculated in parallel with other economic values. Among all passengers, travel for work and study can account for half of the entire travel population, and this half of the travel population spends one-tenth of their disposable time creating value for work and study [44]. Therefore, the formula for converting the travel time of all passengers in one day into economic benefits is

$$Z_{time} = \frac{0.5 \cdot 0.3 \cdot \mathrm{t} \cdot G}{3600} \cdot \int_{d}^{L} ydx \tag{8}$$

where G is the value created per hour (CNY) [45].

Based on the analysis of the impact of station spacing and metered fares in Shanghai, a cost-oriented optimization model of station spacing is established, and the above indicators are classified into income (Z_{inc}) and cost (Z_{cost}). The formula of maximum optimization model is

$$\begin{cases} \max Z = Z_{inc} - Z_{cost} \\ Z_{inc} = Z_{sco} + Z_{land} + Z_{tic} \\ Z_{cost} = Z_{con} + Z_{ope} + Z_{time} \end{cases} \tag{9}$$

where Z is the total benefit of rail transit (CNY).

3.2. HDBSCAN Algorithm

Density-based clustering methods have been widely used in big data analysis of geographic feature information. Based on the DBSCAN, the HDBSCAN algorithm extends DBSCAN by converting it into the hierarchical clustering method, using stability-based techniques to extract planar clusters. HDBSCAN introduces parameters minpts($mpts$) and min$clustersize$ ($mclsize$): the former represents the number of points from the random point to its nearest neighbor, and the latter represents the minimum size of the cluster, which implicitly affect the distribution of clusters detected in the cluster hierarchy.

In order to estimate the density more conveniently, HDBSCAN defines the distance measure between each point, and transforms the density space into a distance space that can be directly recognized. The HDBSCAN algorithm adopts the relevant definition of mutual reachability distance, where $core_k(x)$ is defined as the distance from point to the nearest neighbor. The formula of mutual reachability distance is

$$d_{mreach-k}(a,b) = \max\{core_k(a), core_k(b), d(a,b)\} \tag{10}$$

where $d(a,b)$ is the initial distance metric between point a and b. The size of the circle radius depends on the choice of k, and a larger value will classify more points in high-density clusters. Under the $d_{mreach-k}(a,b)$, the distance between the dense points (points with low core distance) remain unchanged. Still, the outliers will be pushed away under the influence of this parameter, which effectively reduces the outliers while maintaining the clustering effect of high-density [46].

HDBSCAN calculates the Euclidean minimum spanning tree (MST) in the dataset X, namely the proximity graph with the Euclidean distance between points as the edge weight (eps). Then, the Prim algorithm is used to help construct the MST. Every time one edge of the tree is constructed, another vertex of the adjacent edge with the smallest weight is constantly searched. Stop until the tree contains all vertices and achieve the minimum sum of all edge weights. Meanwhile, gradually remove edges whose weights greater than a falling threshold from the proximity graph, and then check the remaining connected components in the graph. The HDBSCAN algorithm takes the shortest distance between clusters as the measure, while sorting the edges of the tree in the order of increasing distance, and then iterates through the threshold until a new merged cluster is created for each edge [47]. Figure 4 shows the entire process from the calculating the MST by the Prim algorithm to construct the cluster hierarchy.

The cluster hierarchy is difficult to explain the specific clustering situation, and has many outliers that need post-processing. At the same time, any choice of cut-line location for the cluster hierarchy is also a choice for mutual reachability distance, hence a single and fixed density level. In this case, if we want to deal with clustering at variable density levels and obtain the flattened clusters, we need to extract a compressed tree of core clusters from the cluster hierarchy. To concretize this concept, the HDBSCAN introduces the new parameter $mclsize$. Once the minimum cluster size is available, the split action will ask if the number of new clusters is less than $mclsize$. The points that were split out can be declared as the breakaway points, letting the larger cluster retain its identity of parent cluster, and marking which points broke away from the initial cluster and the distance

when they left. On the other hand, if split into two clusters, each cluster size must remain equal to *mclsize*, and let the split remain in the tree. After traversing the entire hierarchy, we end up with a compressed tree with small nodes, and each node will contain information about how the size of the cluster changes as the distance shrinks.

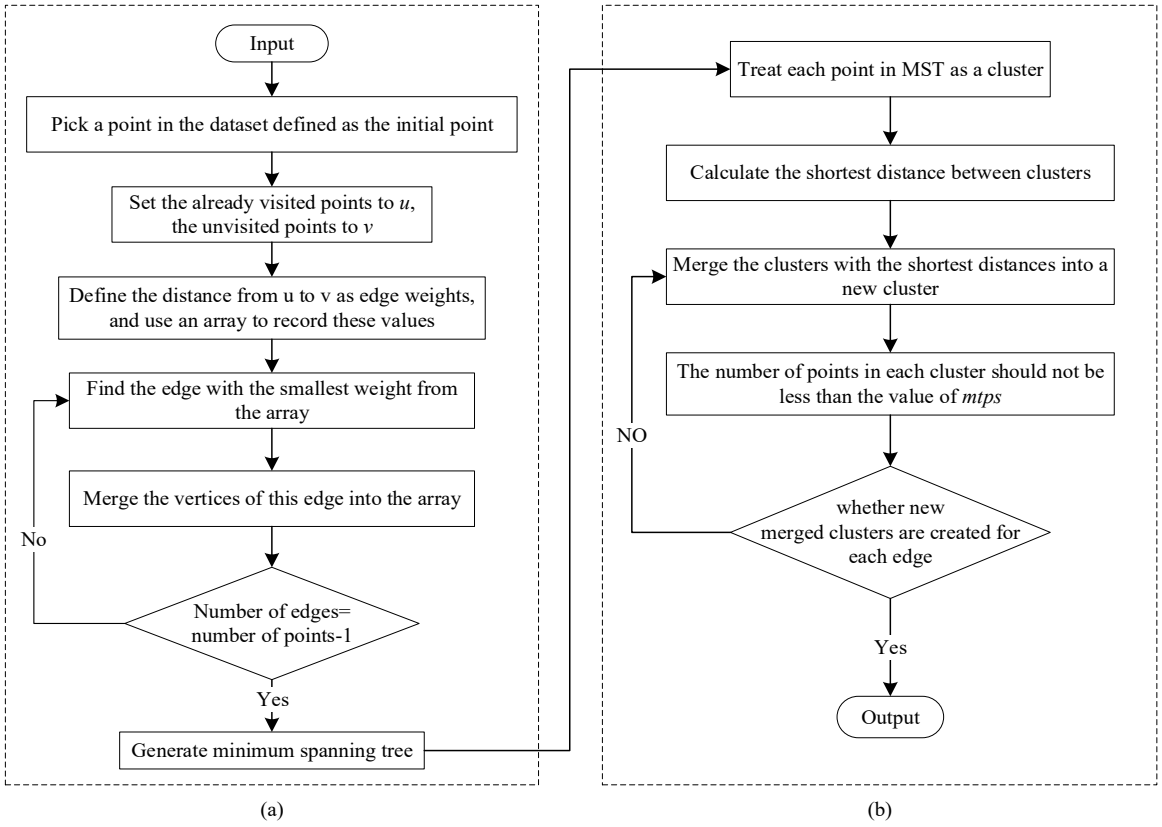


Figure 4. MST: (**a**) Prim algorithm; (**b**) Hierarchical clustering.

4. Data Collection

This paper uses Python Spider to obtain the data information of 825 large-scale residential, commercial, and office buildings on the "Anjuke" website and obtains the geodetic coordinates of the WGS-84 system in Google Earth according to the buildings' name. In order to establish a one-to-one correspondence between the geographic coordinates and the plane rectangular coordinates of the data point, this paper transforms the WGS-84 geographic coordinate into the WGS_1984_UTM_Zone_51N coordinate projection. The data locations are distributed in 11 towns under the Pudong New District, including Datuan, Hangtou, Huinan, Laogang, Nanhui New Town, Nicheng, Shuyuan, Wanxiang, Xinchang, Xuanqiao and Zhuqiao. We name the dataset "PDIA-SL", and "PDIA-SL" collects data on residential, factory and office buildings in these townships in mid, mid-high and high rise buildings [48]. Among them, buildings with 4 to 6 floors are defined as mid-rise, 7 to 9 floors are defined as mid-high rise, and 10 or more floors are defined as high-rise. Mid-rise, mid-high rise and high-rise buildings bring obvious social and economic benefits, and the most important feature is concentrating the urban population. This enables the formation of high-density areas, thereby reducing the area of construction land in megacities and megalopolis where land resources are tight.

Residential buildings refer to residential areas of various scales formed by citizens living together, surrounded by district-level roads or natural boundaries. Generally, they correspond to a residential population of 7000–15,000 and households of 2000–4000. Based on the research and analysis of urban passenger flow and land use, rail transit stations cover the social and economic centers, and these areas usually adopt high-intensity de-

velopment. The dependence between urban travel and land use is the superposition of various functional land and their spatial effects. The stations significantly impact urban passenger commuting which is mainly distributed in dense office buildings and industrial plants. Therefore, in the data selected in this paper, the building types mainly include residential, factory and office buildings. At the same time, the data points of "PDIA-SL" are all located in the southwest of Pudong International Airport, which can effectively target and study the weak rail transit development and passenger travel demand in the southwest of Pudong International Airport. The statistics of the data information are shown in Table 1, and the distribution of "PDIA-SL" dataset is shown in Figure 5.

Table 1. Statistical results of the "PDIA-SL" dataset.

Township	Types of Building			Types of Properties				
	Mid Rise	Mid-High Rise	High Rise	Residential Building	Factory Building	Office Building	House -Holds	SPR (CNY)
Huinan	98	14	35	97	27	23	78,908	26,480
Xinchang	61	4	15	58	16	6	48,915	27,832
Datuan	48	8	12	53	10	5	60,650	18,612
Hangtou	70	7	31	76	15	17	60,870	33,052
Zhuqiao	50	7	8	42	20	3	69,960	29,940
Nicheng	52	13	14	62	13	4	42,120	20,303
Xuanqiao	37	8	8	21	25	7	34,254	22,489
Shuyuan	32	2	6	35	4	1	30,318	18,052
Wanxiang	31	6	4	32	8	1	21,231	18,844
Laogang	33	5	0	31	7	0	11,030	17,576
Nanhui New Town	65	19	22	91	7	8	79,176	33,748

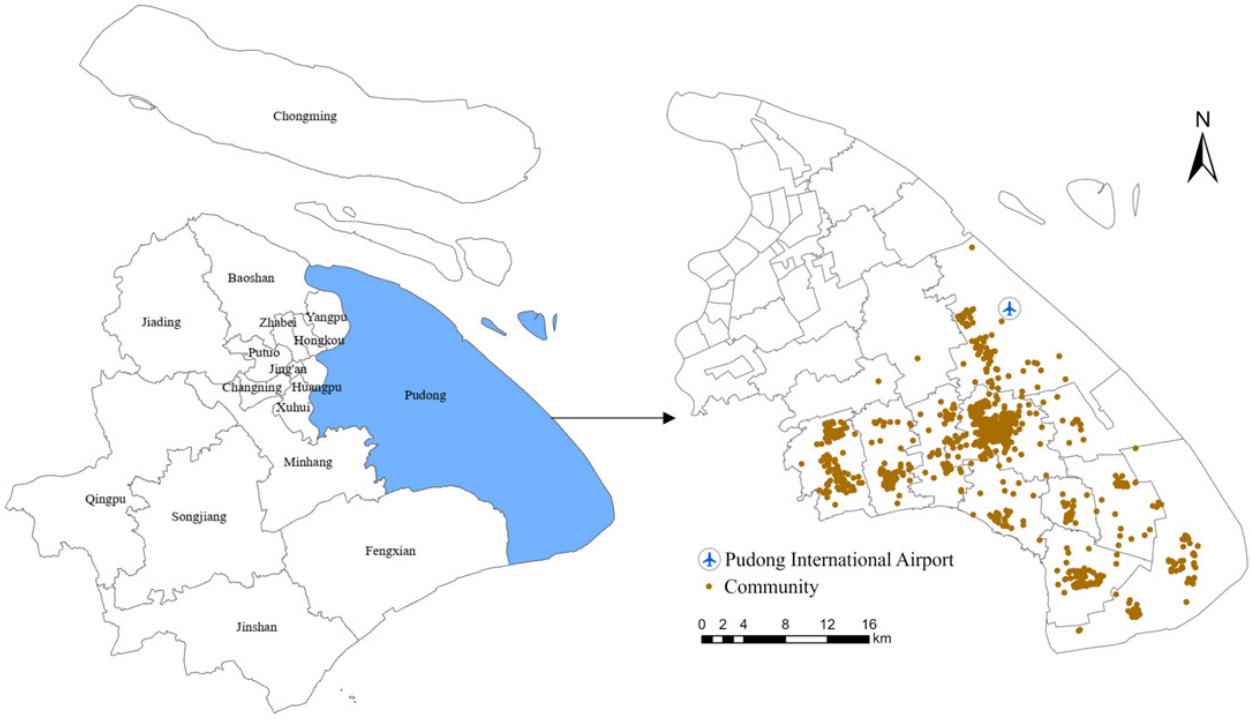


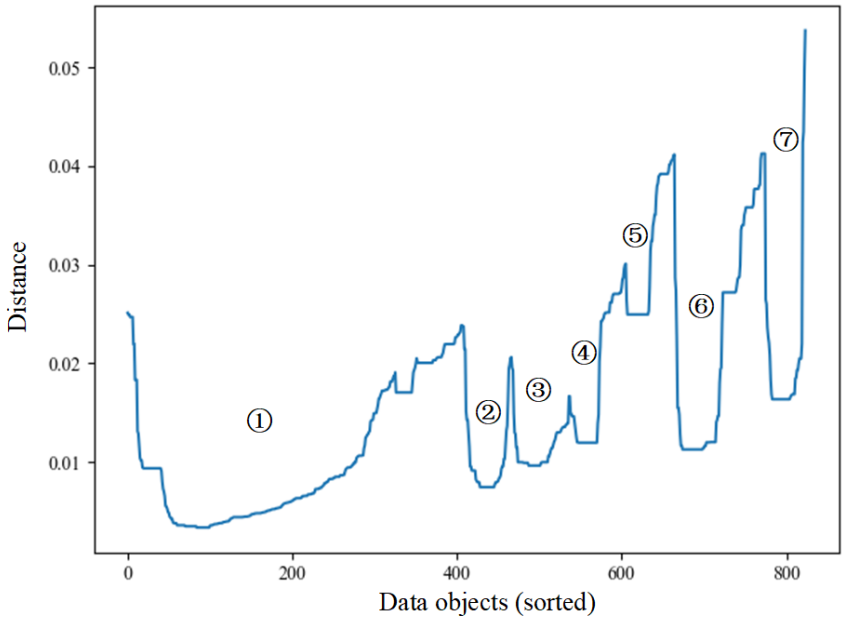
Figure 5. Distribution of the residential, factory and office buildings.

5. Case Study

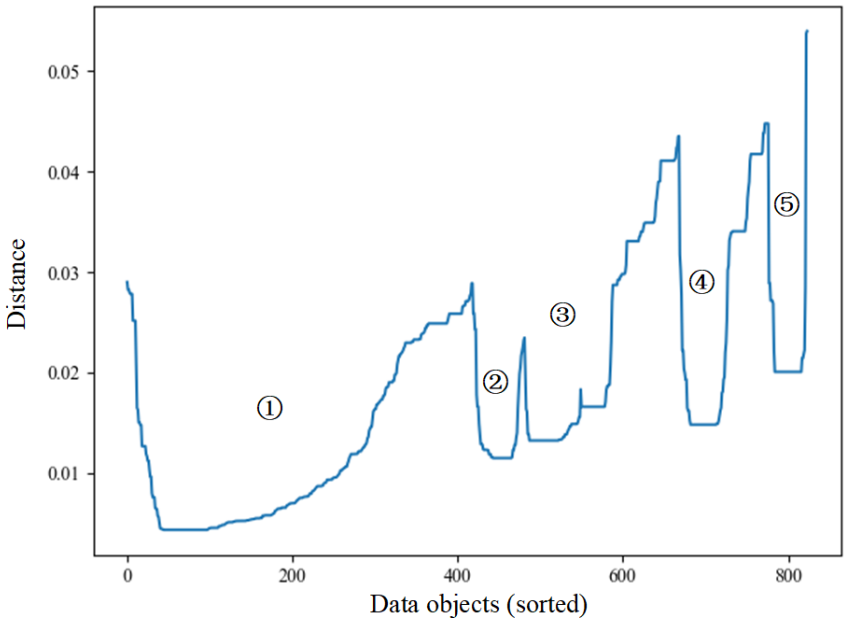
5.1. Accessibility Distance Analysis of OPTICS under Different mpts

The OPTICS algorithm is a generalization of DBSCAN. After being processed by the OPTICS algorithm, the clustering of any *mpts* corresponding to the density can be obtained theoretically. As shown in Figure 6, the clustering results are displayed as the most prominent "dent". The linear change between the dents represents the process

of data points being retrieved and formed into clusters, and we mark these dents with ordinal numbers for more apparent observation. When $mpts = 30$, the curve is very rough. Especially when the curve goes to each peak stage, the fluctuation is relatively large, which shows that the clusters are split during the process of being retrieved and clustered. It is divided into too many clusters, resulting in a small number of valid data points contained in each cluster. When $mpts = 40$, it can be clearly found that the curve becomes smooth, but the boundaries between clusters become blurred. Outliers act as bridges for connection, making two clusters mistakenly merge into a new cluster, which is not the ideal clustering result. When $mpts = 50$, the curve is relatively smooth and divided into five distinct clusters. We can regard it as the target of priority attention, and further verify it by comparing Silhouette plots. Based on the "PDIA-SL" dataset, the reachability-like graphs of $mpts = 30$, $mpts = 40$ and $mpts = 50$ are shown in Figure 6.

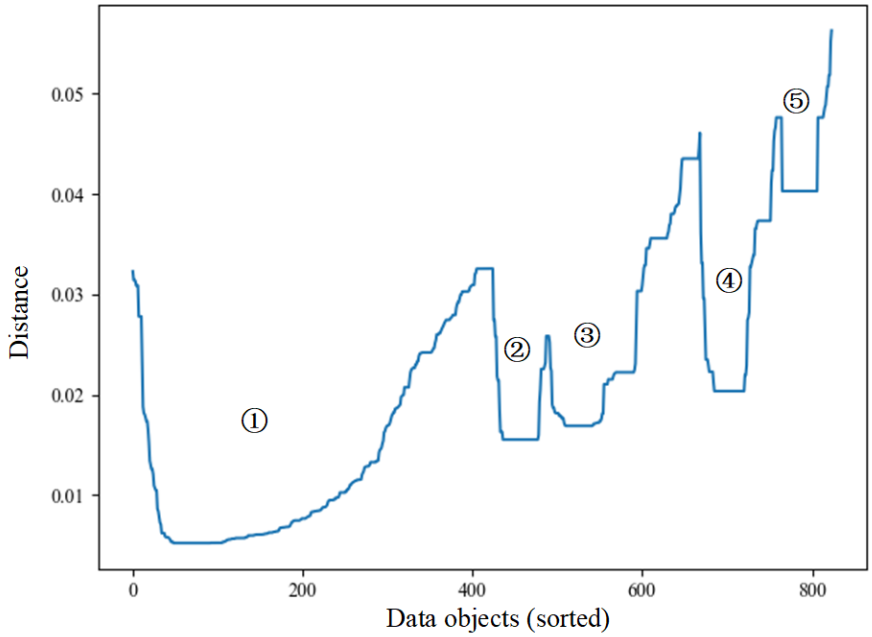


(**a**) $mpts = 30$



(**b**) $mpts = 40$

Figure 6. *Cont.*



(**c**) $mpts = 50$

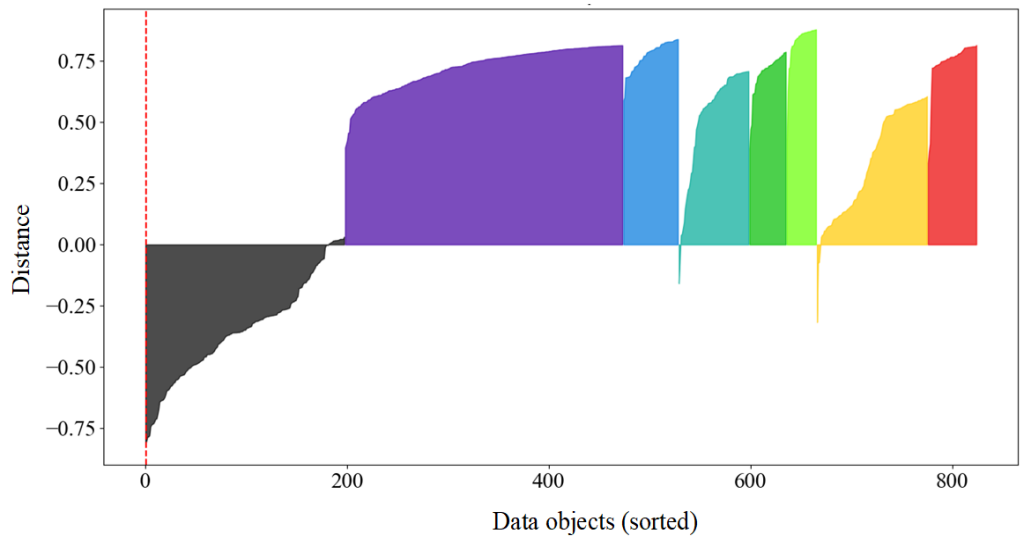
Figure 6. Reachability-like plots based on "PDIA-SL" dataset.

The Silhouette diagram can measure the similarity between the node and the cluster it belongs to compared with other clusters. The value range is between -1 and 1, and the larger the value, the more the node matches its cluster, rather than adjacent clusters. If most points have high Silhouette value, then the clustering is appropriate; if many points have low or negative values, there are too many or few clusters. We use gray to fill the clusters that maintain small and negative values in the graph, and bright colors to indicate groups with better clustering effects. As shown in Figure 7, we obtained the Silhouette diagram with $mpts = 30$, $mpts = 40$ and $mpts = 50$. When $mpts = 30$, we found that about 200 data points were processed as outliers. And all data points were divided into seven clusters after clustering, but the distribution of color blocks was too crowded. Two clusters can be retrieved in data objects between 608 and 669, which contain valid points close to the critical value of *mpts*. When $mpts = 40$, we can find that there are a large number of Silhouette values between 0 and 0.5, indicating that there are a large number of abnormal points in the data points under this condition. This is likely caused by too few clusters due to the influence of outliers. When $mpts = 50$, the effect of this clustering is relatively ideal, and the peak value of each cluster is greater than 0.6. However, there will still be more than 200 data points with negative values in this case, and we need to further analyze the clustering results.

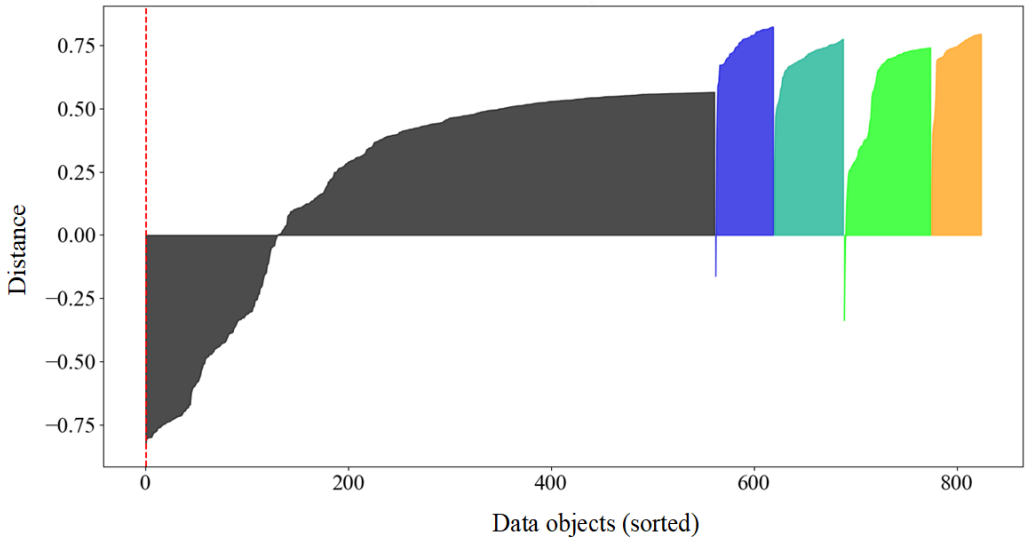
5.2. MST, Cluster Hierarchy and Extraction of Important Clusters

A rough and intuitive impression of clusters can be obtained by comparing the reachability-like diagrams under different values of *mpts* and the corresponding Silhouette plots. Still, it is inefficient to achieve ideal clustering by giving the *mpts* value. We will transform the data space, calculate the Euclidean distance between points by Prim's algorithm, calculate the edge's weight value, and generate the MST. The MST is shown in Figure 8.

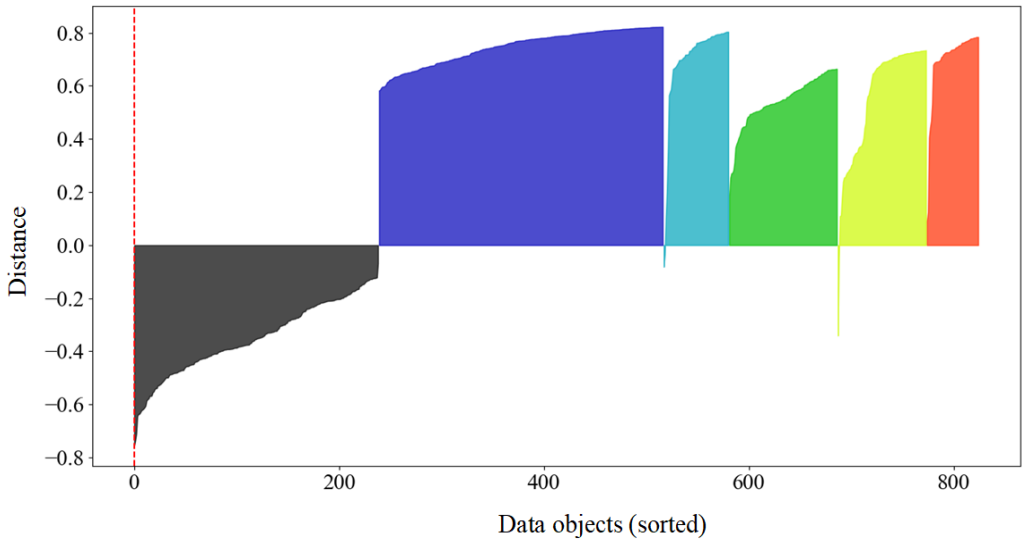
The data points in the graph are regarded as vertices, and the edge weight is equal to the mutual reachability distance between any two points. Then a threshold lowered by high stability will be considered, and any edges with weights above this threshold will be gradually discarded. Eventually a hierarchy of clusters connecting components will be established at different threshold levels. Convert the MST into a cluster hierarchy as shown in Figure 9.



(**a**) $mpts = 30$



(**b**) $mpts = 40$



(**c**) $mpts = 50$

Figure 7. Silhouette plots based on "PDIA-SL" dataset.

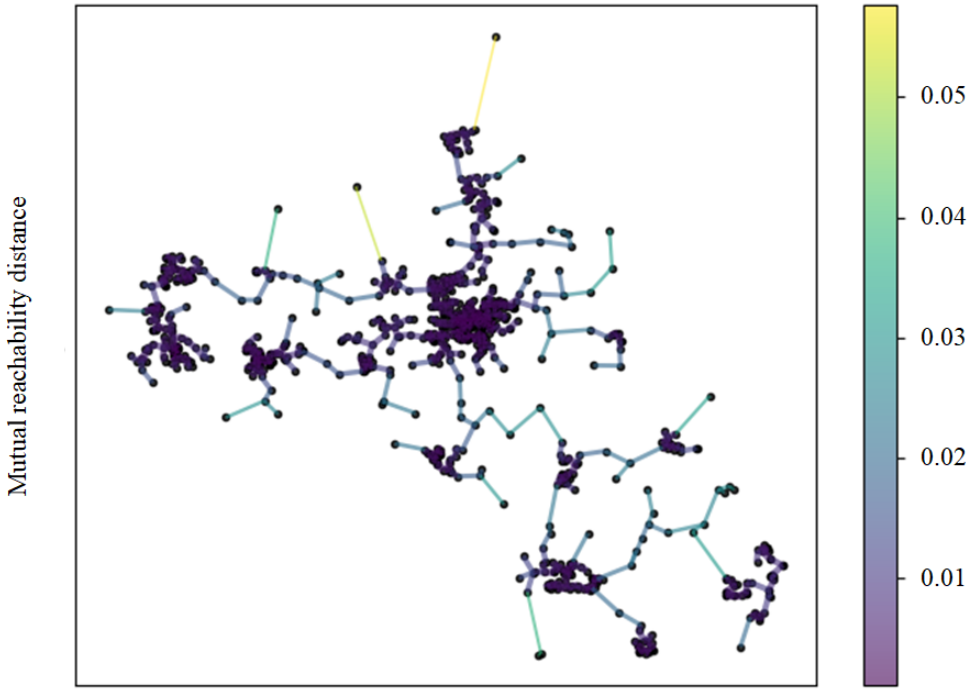


Figure 8. MST of data points using Prim's algorithm.

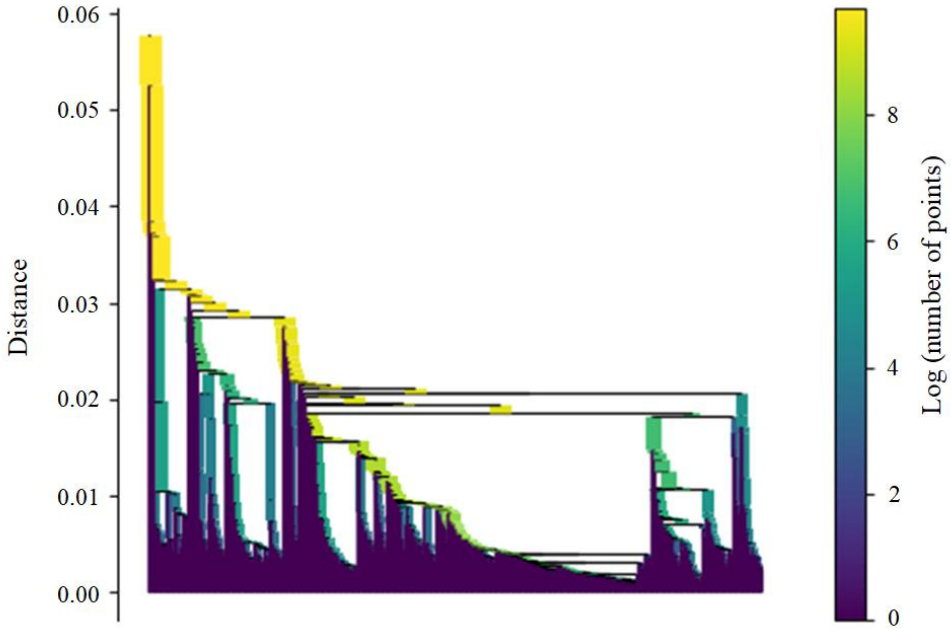


Figure 9. Single-link cluster hierarchy.

It can be observed through the cluster hierarchy that the level corresponds to each node, which corresponds to the critical value of clusters splitting into the components at different threshold levels. The splitting stops when each object forms its cluster. When the value of $d_{mreach-k}(a, b)$ is in the range of (0.03, 0.06), follow the operation method of AHP to cross-cut a dividing line at a certain level of distance, and observe the nodes and independent vertical lines below the level. In this case, it is difficult to observe the cut hierarchy due to the large number of samples. Especially when the value range of $d_{mreach-k}(a, b)$ is (0, 0.03), many vertical lines overlap together. Hence the need to further

compress the cluster tree, and we hope the selected clusters can persist and have a longer life cycle. The formula of considering the cluster persistence is

$$\lambda = \frac{1}{eps} \tag{11}$$

where $\lambda = [1, +\infty)$. When λ gradually increases (eps gradually decreases), the cluster will become smaller until it disappears completely or decomposes into sub-clusters.

In the tree, the life cycle of a cluster is defined as the length of the level where the cluster is located. Therefore, it is necessary to define a stability calculation method that can not only account for the existence of outliers in the hierarchy, but also consider the density distribution of data points in the same cluster. Assuming that there is a cluster C_i corresponding to the level of $\lambda_{\min}(C_i)$, it is defined as a maximum connected subset of $\{x \mid f(x) \geq \lambda_{\min}(C_i)\}$. The formula of relative mass excess of C_i at $\lambda_{\min}(C_i)$ is

$$\begin{cases} E_R(C_i) = \int_{x \in C_i} (\lambda_{\max}(x,\ C_i) - \lambda_{\min}(C_i))dx \\ \lambda_{\max}(x,\ C_i) = \min\{f(x),\ \lambda_{\max}(x,\ C_i)\} \end{cases} \tag{12}$$

where $\lambda_{\max}(x,\ C_i)$ is the density level at which cluster C_i splits or disappears [49].

For the cluster hierarchy, under the condition that the dataset X, cluster label and density threshold are given, the stability of C_i can be defined by the formula, i.e.,

$$S(C_i) = \sum_{X_j \in C_i} \left(\lambda_{\max}(X_j,\ C_i) - \lambda_{\min}(C_i)\right) = \sum_{X_j \in C_i} \left(\frac{1}{eps_{\min}(X_j,\ C_i)} - \frac{1}{eps_{\max}(C_i)} \right) \tag{13}$$

where $\lambda_{\min}(C_i)$ is the minimum density level at which $\lambda_{\min}(C_i)$ exists, and $\lambda_{\max}(X_j,\ C_i)$ means that the target cluster X_j will no longer belong to cluster C_i if it exceeds this density level. The $eps_{\min}(X_j,\ C_i)$ and $eps_{\min}(X_j,\ C_i)$ are the thresholds corresponding to eps. Suppose $\{C_2,\ C_3, \cdots,\ C_j\}$ is the set of all clusters in the simplified compressed cluster tree, and let $S(C_i)$ denote the stability of each cluster.

Cluster extraction is to compress a large and complex cluster hierarchy into a smaller tree with more information of each node. During traversal of the entire hierarchy, each split is asked if the number of new clusters created by the split is less than the value of *mclsize*. HDBSCAN only allows reporting of clusters with at least objects of *mclsize*, and cases where a component with less than *mclsize* is disconnected from a cluster should not be considered detached. Any stray components are flagged as outliers. We can easily perform cardinality checking and labeling of components by starting from the endpoints of the edges being removed. This simplification process can greatly reduce the number of clusters in the hierarchy, as shown in Figure 10.

Our goal is to extract the most prominent clusters as a flat solution to maximize the overall stability of the extracted clusters. The formula of calculating the maximum value of the overall stability is

$$\max_{\delta_2, \ldots, \delta_j} J = \sum_{i=2}^{j} \delta_i S(C_i) \tag{14}$$

where $\delta_i \in \{0,\ 1\}$, $i = 2,\ 3, \cdots,\ j$. δ_i indicates whether cluster Ci contains a flat solution ($\delta_i = 1$) or does not contain a flat solution ($\delta_i = 0$).

For the decision variables of $\delta_2,\ \delta_3, \cdots,\ \delta_j$, we want to prevent clusters from being selected on the same node in the subtree. Because nested clusters must be mutually exclusive, only one label can be assigned to each object. Applying dynamic programming, we assemble intermediate solutions upward into the tree to solve the subtree problem of increasing scale. Process each node except the root C_1, and decide on each node from the bottom up whether to choose C_i or the best cluster in the subtree of C_i. To make a decision at node C_i, we recursively update the total stability $\hat{S}(C_i)$ of the selected cluster. The formula of $\hat{S}(C_i)$ is

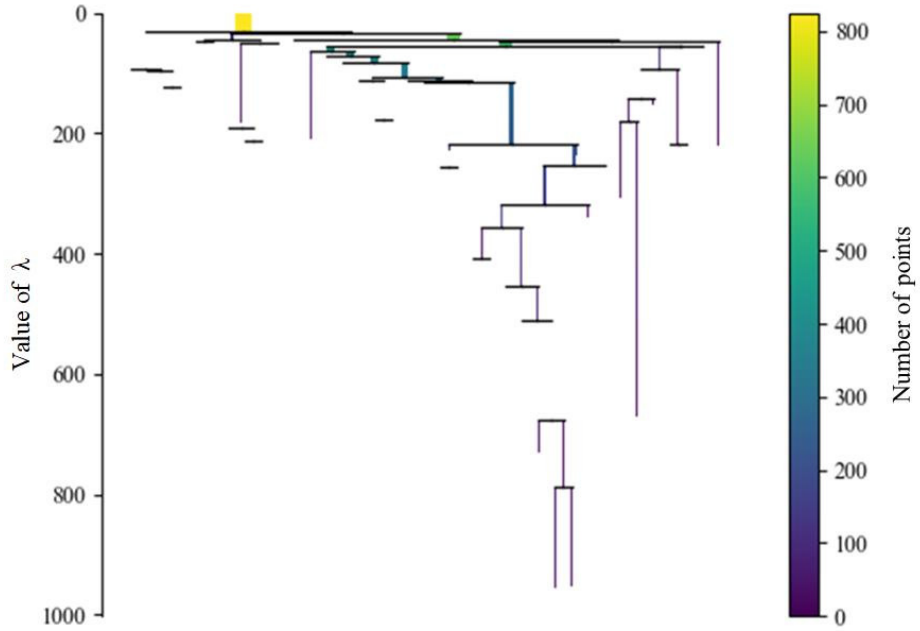


Figure 10. Single-link cluster hierarchy.

$$\hat{S}(C_i) = \begin{cases} S(C_i) & if C_i \text{ is a leaf node} \\ \max\{S(C_i),\ \hat{S}(C_{i_l}) + \hat{S}(C_{i_r})\} & if\ C_i \text{ is an intrnal node} \end{cases} \tag{15}$$

where C_{i_l} and C_{i_r} are the binary subtrees of cluster C_i [16].

We work up the tree in the order of reverse topological. If the sum stability of subclusters is greater than the cluster's stability, then the $\hat{S}(C_i)$ is the sum of the subcluster's stability. On the other hand, if the stability of a cluster is greater than the sum of subclusters, we will deselect all its descendants. After reaching the root node, we refer to the currently selected cluster as the flat cluster and return to results. The extracted cluster is also the cluster with the largest color block in the compressed tree, and we mark it with a red circle, as shown in Figure 11.

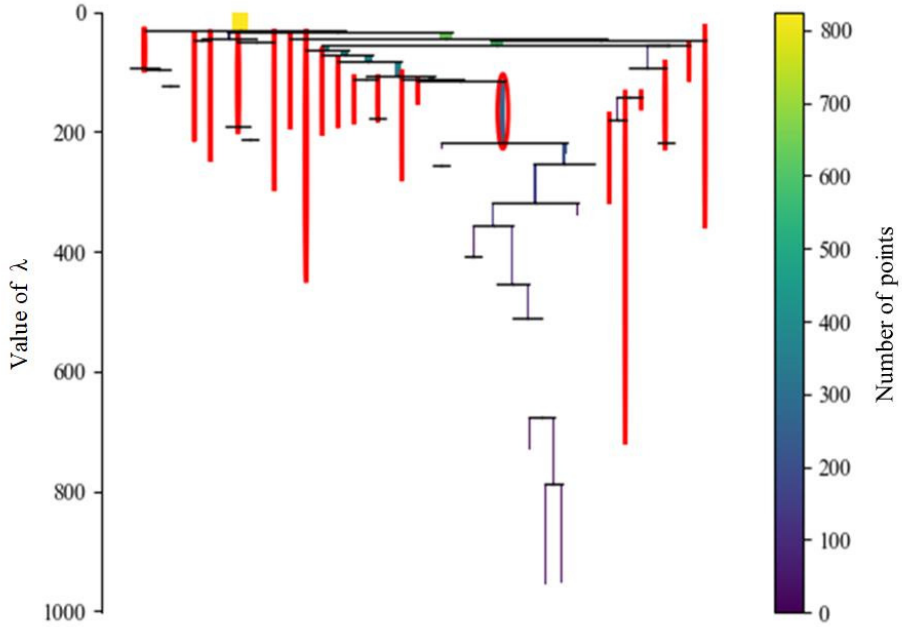


Figure 11. Important clusters.

5.3. Outlier Monitoring of GLOSH

Define the target cluster with the highest density in C_i as X_l, and X_l is the innermost data point in cluster C_i. When the threshold is increased from top to bottom through the hierarchy, X_l is the cluster that survives the longest before each cluster becomes the outliers. Therefore we define the formula for reference density and outlier, i.e.,

$$f(X_l) = \frac{1}{d_{core}(X_l)} \tag{16}$$

$$Outlier(X_i) = \frac{f_{\max}(X_i) - f(X_i)}{f_{\max}(X_i)} \tag{17}$$

where $f(X_l)$ serves as the reference density for each X_i, and

$$f_{\max}(X_i) = f(X_l) = \frac{1}{d_{core}(X_l)} \tag{18}$$

If Equation (17) is used to calculate the situation of $mclsize > 1$, X_i below the threshold will be attached to a new cluster in advance. In order to solve this problem, this paper uses the GLOSH (Global-Local Outlier Scores from Hierarchies) detection method [50], i.e.,

$$GLOSH(X_i) = \frac{\lambda_{\max}(X_i) - \lambda(X_i)}{\lambda_{\max}(X_i)} = 1 - \frac{eps_{\max}(X_i)}{eps(X_i)} \tag{19}$$

Figure 12 contains the initial dataset and the fitting curve of samples, and the GLOSH score interval is between [0, 1]. We plot the histogram of the GLOSH score with a light blue area, where the threshold is set at the quartile, and fit the histogram with the dark blue line. The higher the GLOSH score, the more likely the point is an outlier. We extract the quartiles in the figure to detect outliers, and the results show that only a few points are abnormal, which meets the requirements of ideal clustering.

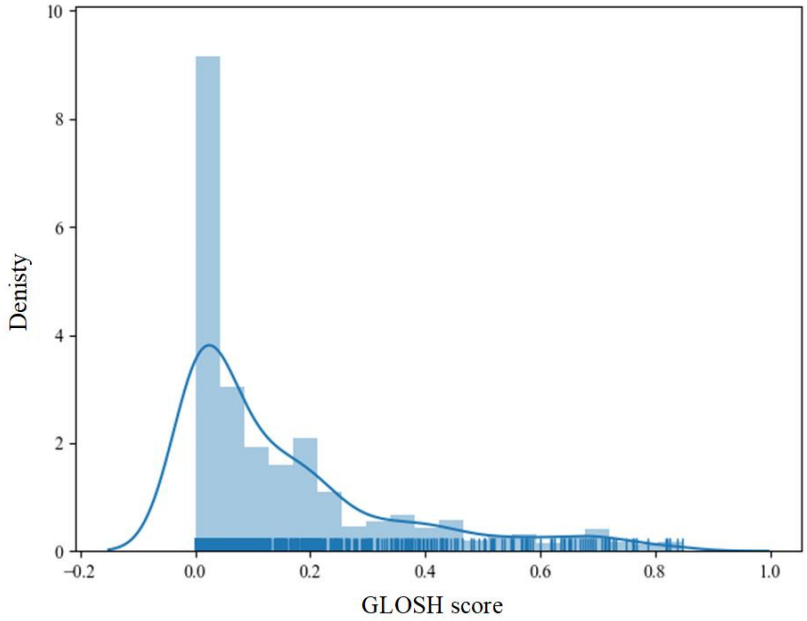
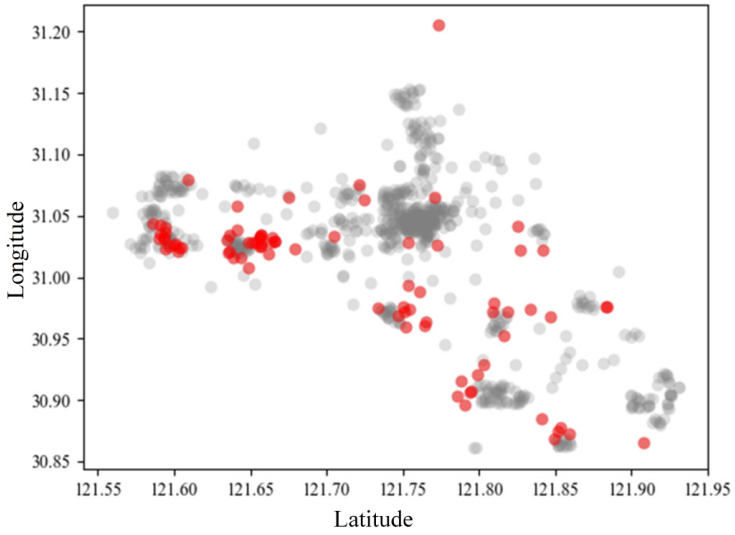
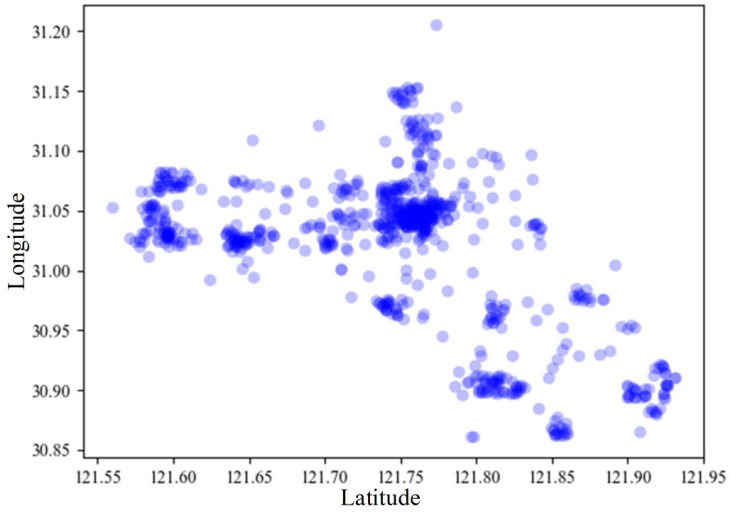


Figure 12. Bar graph of GLOSH score.

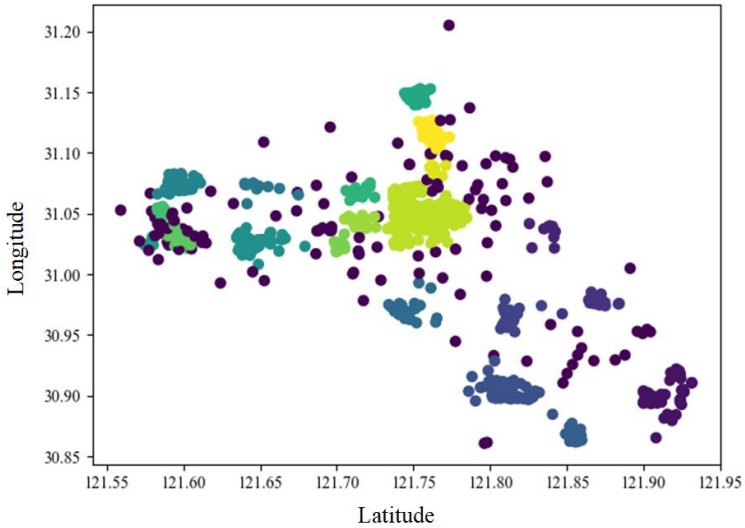
We mark the ideal clusters in the results in blue, and the outliers in red, as shown in Figure 13a,b. Through the analysis of HDBSCAN, we can get the ideal target clusters, as shown in Figure 13c. And we use different colors to distinguish and mark the clusters. Next, we still need to find the core from each cluster to represent the ideal location of the rail transit station.



(**a**) Distribution of outliers



(**b**) Distribution of ideal targets



(**c**) Ideal clustering results from the analysis of HDBSCAN

Figure 13. Result of GLOSH detection: (**a**) The data points of the red label are outliers; (**b**) The data points of the navy blue label are ideal targets; (**c**) The data points marked with non-purple color as the ideal clusters.

5.4. Determination and Correction of Ideal Points in Stations

It can be seen from Figure 13 that the ideal clustering results generated by HDBSCAN are processed by different colors, in which a total of 16 clusters are generated. Then we marked them in order with a serial number. Fit along each cluster's edges, draw various regular plane figures and label them. We use Auto CAD to calculate the geometric center of each planar figure and regard it as an ideal point for station selection, as shown in Figure 14. After calculation, the coordinates of the geometric center points of clusters 1–16 after fitting are as follows: (121.5976, 31.0753), (121.5856, 31.0325), (121.6524, 31.0250), (121.7126, 31.0714), (121.7183, 31.0411), (121.7064, 31.0283), (121.7572, 31.1520), (121.7693, 31.1127), (121.7615, 31.0450), (121.7506, 30.9729), (121.8447, 31.0290), (121.8189, 30.9656), (121.8073, 30.9027), (121.8775, 30.9663), (121.8516, 30.8654), (121.9206, 30.9027).

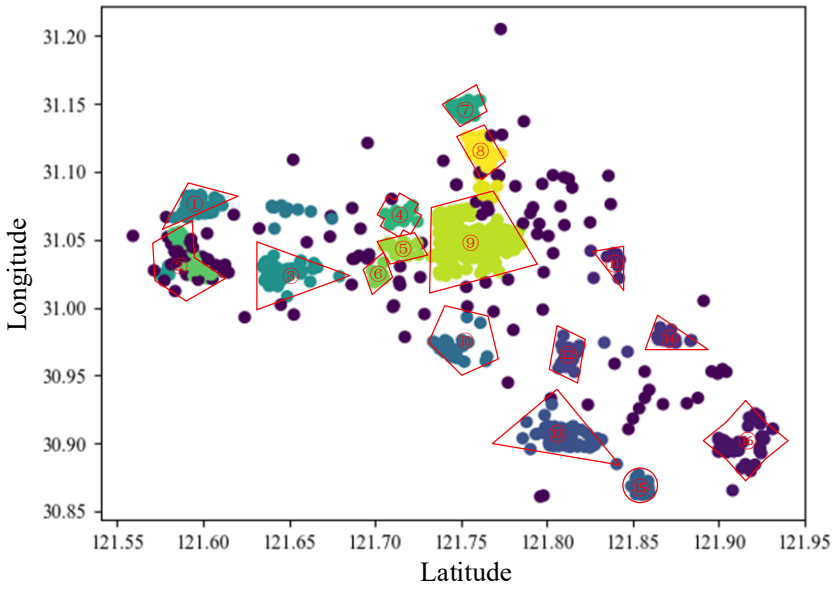


Figure 14. Fit clusters to regular images.

By sub-meter observation of the above coordinate points through Google Earth, the rail transit station should provide maximum convenience for passengers and prevent the impact of concentrated passenger flow on the station. The station's location should be about 300 m away from residential, commercial and office buildings, as well as industrial plants and large stadiums [51]. Meanwhile, groundwater sources, cultivated land and wetlands should be avoided as much as possible. The density of Metro lines and the number of stations are far less than those of the buses. Thus the layout of stations should be closely integrated with urban roads and transit networks, and should be coordinated with the urban planning of the area where the stations are located. To attract passengers to the greatest extent, facilitate the transfer between buses and subways, and further make rail transit a backbone artery with fast and large traffic volume, the rail transit stations are generally located at road intersections. Considering the above factors, we have corrected the coordinates of 16 ideal sites in Figure 14 and marked the locations in each observation scene as shown in Figure 15.

5.5. Analysis of Optimal Station Spacing

According to the cost-oriented optimization model of Equation (9), if we want to establish the relationship between the net cost and the station spacing, we need to establish the relational expressions between Z_{sco}, Z_{land}, Z_{tic}, Z_{con}, Z_{ope}, Z_{time} and the station spacing d. From the analysis of the impact of station spacing in Section 2, it can be obtained $Z_{sco} = 1.2Z_{con}$. In order to calculate the social benefits of the daily average land appreciation along the line, we investigated the data of average property (α), occupancy ratio (β) and gross plot ratio (θ) of residential, commercial and industrial land within 1200 m of attraction

range from 2009 to 2015, as shown in Table 2. To more intuitively display the changes in property values and occupancy ratios of the three types of land use, it is converted into a 3D waterfall diagram, as shown in Figure 16.

Figure 15. Correction of each ideal point in the observation map.

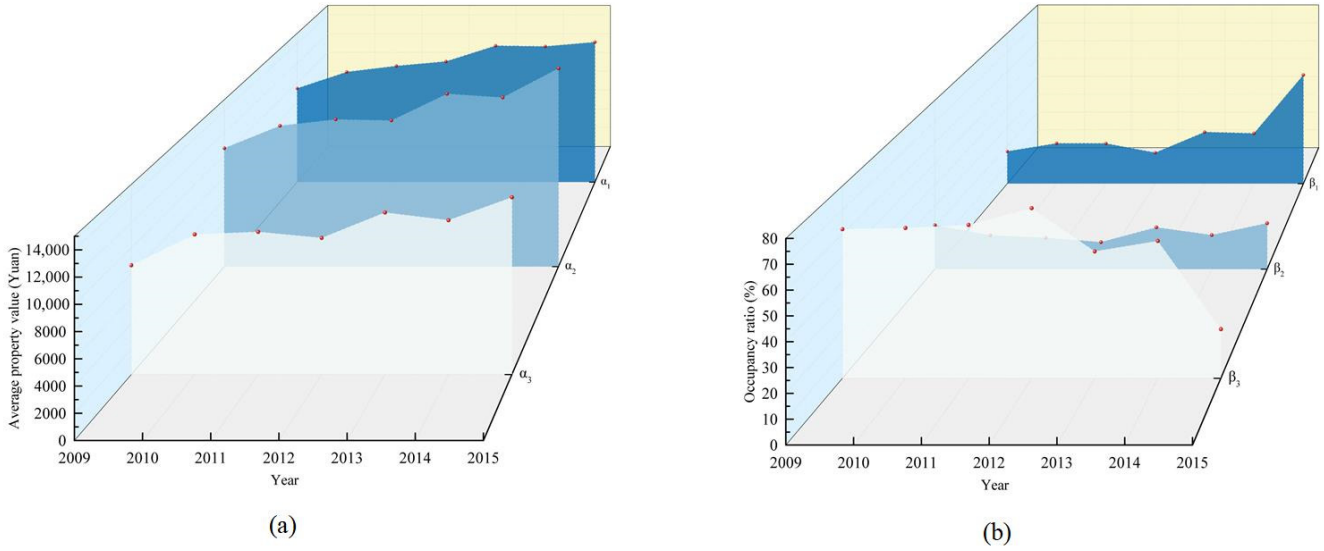


Figure 16. 3D waterfall diagram of average property value and occupancy ratio of the residential, commercial and industrial land: (**a**) Changes in average property value from 2009 to 2015 in Shanghai; (**b**) Changes in occupancy ratio from 2009 to 2015 in Shanghai.

Table 2. Average property and occupancy ratio of the residential, commercial and industrial land.

Year	c_1	β_1	θ_1	c_2	β_2	θ_2	c_3	β_3	θ_3
2009	1579.49	17.16%	1.59	2014.65	20.84%	1.97	2036.41	62.00%	1.54
2010	1658.32	21.47%	1.61	12651.34	16.01%	2.13	2436.73	62.52%	1.54
2011	608.03	21.39%	1.80	570.63	14.82%	2.38	199.38	63.79%	1.56
2012	452.28	16.46%	1.81	−93.72	12.77%	2.31	−480.51	70.77%	1.63
2013	1587.46	27.47%	1.89	2391.55	19.77%	2.51	2007.78	52.76%	1.70
2014	−69.8	26.75%	1.87	−323.21	16.13%	2.07	−612.38	57.12%	1.54
2015	433.07	57.82%	1.88	2610.45	21.75%	2.64	1810.64	20.43%	1.51

Assuming that the total length of the planned rail transit line is 56 km, substitute the above data into Equation (5). We can calculate $Z_{land} = 16281.1[\text{int}(\frac{56}{d}) + 1]$. Combining

Equations (1) and (7), the formula of the relationship between Z_{tik} and the station spacing d is

$$Z_{tik} = -1.7143d^7 + 0.0058d^6 + 0.7766d^5 - 75.3d^4 + 2326.7d^3 - 27877.5d^2 - 11688d + 6.606 \times 10^8 \quad (20)$$

According to the relevant data and experience of rail transit construction in Shanghai, the average construction cost of a single station is about 8.7752×10^7 CNY, and we set the service life of the subway to 50 years [52]. The formula for construction cost is

$$Z_{con} = 4808.33[\text{int}(\frac{56}{d}) + 1] \quad (21)$$

Up to now, the number of trains put into operation in Shanghai is about 480, the total number of stations is 508 (including the maglev line), and the body of the subway train is made of aluminum alloy [53]. Each Electric multiple units (EMU) is equipped with one driver and attendant, assuming that the drivers' salary is 1.5 times that of the attendant. And expenditures such as crew allowances and additional wages generally account for 25% of the basic salary. The basic monthly salary of attendants is about 5000 CNY, and the basic monthly salary of drivers is about 7500 CNY. There are about twenty station attendants at a subway station, with an average basic salary of 4000 CNY/month. From this calculation, it can be obtained that $C_1 = 3854.2$. According to the electricity consumption statistics of the Shanghai Metro corporation for many years, the total electricity consumption of all trains during the downtime is 2,935,032 KW·h, and the power cost of aluminum alloy trains is about 0.51 CNY/KW [54]. From this calculation, it can be obtained $C_2 = 129.94$.

According to statistics, the total cost of station equipment is 6.7162 million CNY per year [55]. In summary, the calculation formula of Z_{ope} can be obtained, i.e.,

$$Z_{ope} = 4020.34[\text{int}(\frac{56}{d}) + 1] \quad (22)$$

When passengers arrive at or leave a rail transit station, there must be a certain amount of walking time. Assuming that the average walking speed of passengers is 1 m/s, whether to choose to walk to or leave the station depends on the walking distance. The bus is the most cost-effective way for passengers to travel distance and fare. Therefore, if the walking distance exceeds the bus-stop spacing, passengers will often choose to take public transportation instead of walking. Therefore, we believe this walking distance should be less than the average spacing between bus stops, which is about 450 m. Therefore, we can conclude that the time spent by passengers walking to the rail transit station is about 450 s (t_1). Generally speaking, the time to arrive at the station from the place of origin (t_1) is equal to the time for passengers to leave the station and arrive at the destination (t_5). Assume that when passengers arrive at the station, they have fully experienced the stop time between the start and end points, and the average time that passengers need to wait at the station (t_2) is half of the time difference between the departures of the two trains [55]. In the case of not considering the delayed operation of trains, the average interval between two-way subway departures is 5 min. And we can get $t_2 = 300 \cdot \text{int}(\frac{56}{d})$. The time ($t_3$) for passengers to travel with the subway is the following formula, i.e.,

$$t_3 = \frac{v}{3.6a_1} + \frac{v}{3.6a_2} + \frac{3.6 \cdot [1000d - \frac{1}{2}a_1 \cdot \left(\frac{v}{3.6a_1}\right)^2 - \frac{1}{2}a_2 \cdot \left(\frac{v}{3.6a_2}\right)^2]}{v} \quad (23)$$

where v is the steady running speed of the subway (km/h), a_1 is the starting acceleration of the train (m/s^2), a_2 is the braking acceleration of the train (m/s^2).

Assume that the steady running speed of the train is 80 km/h, the starting acceleration is 0.9 m/s^2, and the braking acceleration is 1 m/s^2. We can calculate $t_3 = (45d - 2.75) \cdot \text{int}(\frac{56}{d})$. The stop time must meet the time requirements for passengers to get on and off the train. Secondly, reducing the stop time will improve the operating efficiency of rail transit and

reduce the travel time of passengers. Combined with the analysis of intermediate rail transit stops in various places, it is assumed that the stop time is the 30 s. We can calculate $t_4 = 30 \cdot \text{int}(\frac{56}{d})$. To sum up, the formula of total travel time is

$$t = t_1 + t_2 + t_3 + t_4 + t_5 = (327.26 + 45d) \cdot \text{int}(\frac{56}{d}) + 900 \tag{24}$$

Combine Equations (8) and (24) to calculate the total time cost of passenger travel, and the formula is

$$\begin{aligned} Z_{time} = \big((327.26 + 45d) \cdot \text{int}(\tfrac{56}{d}) + 900\big) \cdot (-5.714 \times 10^{-5} d^7 + 2.833 \times 10^{-3} d^6 + \\ 0.259d^5 - 25.1d^4 + 775.567d^3 - 9292.5d^2 - 3896d + 8.738 \times 10^6) \end{aligned} \tag{25}$$

By substituting the Z_{sco}, Z_{land}, Z_{tic}, Z_{con}, Z_{ope} and Z_{time} into Equation (9), we can obtain the functional relationship between the maximum net cost (max$Z$) and the station spacing. Meanwhile, to weaken the influence of high-power terms on the function value, this paper modifies the functional relationship. The revised formula is

$$\begin{aligned} \max Z = 13222.43[\text{int}(\tfrac{56}{d}) + 1] - (327.26 + 45d) \cdot [\text{int}(\tfrac{56}{d}) + 900] \cdot (-5.714 \times 10^{-5} \\ d^7 + 2.833 \times 10^{-3} d^6 + 0.259d^5 - 25.1d^4 + 775.567d^3 - 9292.5d^2 - \\ 3896d + 8.738 \times 10^6) - 1.7143d^7 + 0.0058d^6 + 0.7766d^5 - 75.3d^4 + \\ 2326.7d^3 - 27877.5d^2 - 11688d - 3.993 \times 10^{12} \end{aligned} \tag{26}$$

At present, we know the functional relationship between the cost and the station spacing. Then, we hope to obtain the maximum value of the objective function, that is, the distance of the optimal station spacing. In this paper, particle swarm optimization (PSO) is used to calculate the optimal global solution of the objective function. PSO is an optimization method to efficiently solve nonlinear, non-differentiable and multi-peak complex propositions [56]. The PSO algorithm uses massless and volumeless particles to simulate the birds in the flock, and the particles have only two properties, namely velocity and position. Among them, velocity represents the speed of particle movement, and position represents the direction of particle movement. The position of the particle i in N-dimensional space is expressed as $M_i = (m_1, m_2, \cdots, m_n)$, and the flight speed is expressed as $V_i = (v_1, v_2, \cdots, v_n)$. Each particle has a fitness value determined by the objective function and can mark the best position found so far (*pbest*) and the current position (X_i). The PSO algorithm is initialized as a group of random particles, and the optimal solution is found through iteration. During the iteration process, the particle updates the optimal value by tracking two extreme values (*pbest* and *gbest*). After finding these two optimal values, the particle updates its velocity and position through the following formula, i.e.,

$$\begin{cases} v_i = \omega \cdot m_i + e_1 \cdot rand() \cdot (pbest_i - m_i) + e_2 \cdot rand() \cdot (gbest_i - m_i) \\ m_i = m_i + v_i \end{cases} \tag{27}$$

where v_i is the particle velocity, m_i ($i = 1, 2, \cdots, N$) is the current position of the particle, N is the total number of particles in the group, $rand()$ is a random number between (0, 1), e_1 and e_2 are learning factors, and ω is the inertia factor. And when the value of ω is large, the global optimization ability is strong, and the local optimization ability is weak; when the value is small, the global optimization ability is weak, and the local optimization ability is strong.

The formula for calculating ω is

$$\omega = \frac{(\omega_{ini} - \omega_{end}) \cdot (G_k - g)}{G_k + \omega_{end}} \tag{28}$$

where ω_{ini} is the initial inertia weight, ω_{end} is the inertia weight when iterating to the maximum evolution algebra, and G_k is the maximum number of iterations [57].

Using the PSO algorithm to calculate Equation (26), the relationship between the number of PSO iterations and the optimal function value can be obtained. Figure 17a describes the fitness changes of the global best of the swarm at each iteration. Figure 17b shows the final state positions of particles during the iteration process, which are marked with circles of various colors. To prevent conflicts with buses, improve the overall utilization of urban transport and reduce the waste of traffic resources, the average station spacing of rail transit shoud be greater than the distance between bus stations, which is 450 m. We have obtained that the distance of the optimal station spacing is 683.6 m, which meets this requirement. In order to facilitate the determination of Metro lines, we approximately take the optimal station spacing as 680 m.

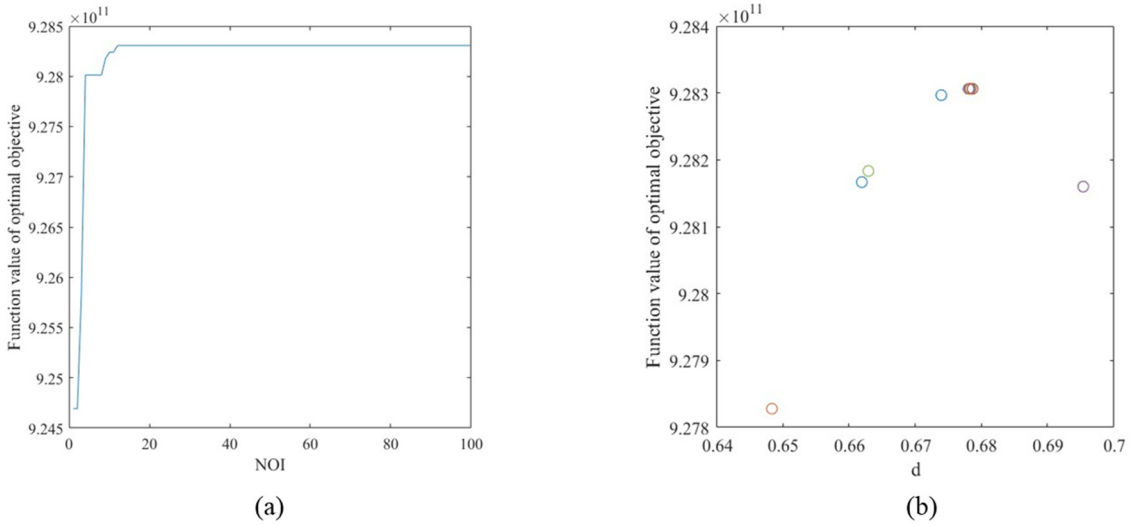


Figure 17. Global optimal value of the target function using the PSO algorithm: (**a**) The relationship between the number of PSO iteration and the optimal function value; (**b**) The optimal function value corresponds to the station spacing *d*.

5.6. Determination of Metro Lines

To simplify the calculation, we assume that the Earth is a regular sphere. The latitude difference at 31° north latitude is 111.194 km, and the longitude difference is 96.297 km [58]. Consider the calculated and corrected coordinates as ideal points of station selection, mark them in the satellite image, and connect the ideal points with smooth curves. During the connection process, attention should be paid to avoiding mountains, wetlands, cultivated land and groundwater sources, and different colors are used to distinguish the five selection lines. On the selected lines, set up the rail transit stations every 680 m. The selected location of stations should also be close to the main traffic arterial and about 300 m away from large residential, commercial, factory buildings and large gymnasiums. The ideal stations and five route selections for the expected planning are shown in Figure 18.

Route 1 passes through ideal locations 1, 4 and 8, and connects Zhoupu, Liuzao and Zhuqiao Town in Pudong New Area in the east-west direction. Route 2 passes through the ideal locations 7, 8 and 9, mainly along Zhuqiao and Huinan Town to the north of Pudong International Airport. Route 3 passes through the ideal locations 2, 3, 6, 9 and 11, which is the longest route in the east-west planning in this paper, crossing Xinchang, Xuanqiao, Huinan and Zhuqiao Town. It straddles Xinchang, Xuanqiao, Huinan and Zhuqiao Town and can be connected with the existing No. 11 Metro line and the future rail transit lines in the Minhang district. Route 4 passes through the ideal locations 10, 12, 14 and 16, and this route mainly connects Nanhui New Town. Route 5 passes through the ideal locations 4, 5, 6, 10, 13, and 15. It is the longest route selected in the north-south planning in this paper, which mainly crosses Chuansha, Xuanqiao, Datuan and Nicheng Town.

Figure 18. The ideal station location and five routes of the expected plan.

6. Comparative Analysis with DBSCAN and K-Means Clustering Results

This paper uses DBSCAN and K-means algorithms to analyze the "PDIA-SL" dataset in the above case. We will compare and analyze the clustering results of DBSCAN, K-means and HDBSCAN algorithms from the perspectives of time complexity, space complexity, silhouette coefficient and cluster view.

6.1. Time Complexity

The time complexity of an algorithm is a function that qualitatively describes the algorithm's running time and is also an indicator to measure the workload required to execute the algorithm. DBSCAN searches for clusters by checking the neighborhood of each point (eps) in the dataset. When the number of points in a certain point's neighborhood is greater than $mpts$, a cluster with this point as the core is created. Then continue to iterate and gather other clusters directly density-reachable from this cluster, and stop when no new points are added to any cluster. Most of the work of DBSCAN is to perform region queries, and the algorithm scans the entire dataset when performing region queries on each point. Therefore, its time complexity is $O(825^2)$. For large datasets, this method is very time-consuming. The K-means algorithm selects k points randomly as cluster centers, and calculates the distance between each point and center. Re-divide the center point so that the sum of the distances between all samples and this point is the smallest. The time complexity of K-means is $O(nkt)$. Among them, n represents the number of samples, k represents the number of clusters, and t represents the number of iterations [59]. For the "PDIA-SL" dataset, the algorithm's time complexity is $O(825 \times 9 \times 5)$.

The calculation process of HDBSCAN is highly complex, so we only analyze several vital steps of traversal search and iteration. When executing the K-NN query of n points, each object has to traverse the Euclidean distance calculation, and the time complexity is $O(n^2)$. By using the Prim algorithm to realize the search based on the dataset, the MST can be constructed in $O(n^2 + m)$ time complexity, where m is the edge weight of the mutual reachability graph. In this case, $m = \frac{n(n-1)}{2}$, so the time complexity is $O(n^2 + m)$. In the process of identifying and sorting $n - 1$ edges, the time complexity is $O(n \cdot \log n)$. As the MST reduced to smaller subcomponents, the clusters are relabeled, and the time complexity of this process is $O(n^2)$. In general, the total time complexity of the algorithm is $O(n^2)$, namely $O(825^2)$.

6.2. Space Complexity

Space complexity measures the storage space temporarily occupied by the algorithm during operation. Compared with its high time complexity, the space complexity of DBSCAN is only $O(825)$. Each point only needs to maintain a small amount of data information (cluster label and point identity). The space complexity of K-means is $O((n+k)\cdot m)$, and m represents the dimension [60]. It can be understood that the data information and center points need to be stored, and the space complexity is $O((825+9)\times 2)$.

In terms of main storage requirements, the HDBSCAN algorithm requires the space complexity of $O(n)$ to store datasets. Mutual reachability graph does not need to be explicitly calculated, and it only needs to store the edge weights of the MST, which requires space complexity of $O(n)$. Furthermore, only the processing hierarchy needs to be stored in time, which requires space complexity of $O(n)$. Therefore, the overall space complexity of the algorithm is $O(825)$.

6.3. Silhouette Coefficient

The silhouette coefficient is an index to evaluate the clustering effect. It combines the two factors of cohesion and separation and can be used to evaluate the impact of different algorithms or different operating modes of algorithms on the clustering results. For a certain point i in the dataset, the formula of the silhouette coefficient of the point i is

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i),\ b(i)\}} \tag{29}$$

where $a(i)$ is the average Euclidean distance from a point i to other points in the same cluster, and $b(i)$ is the average Euclidean distance from the point i to all points not in the same cluster. When there is only one point in the cluster, we define the silhouette coefficient $S(i)$ as 0. The mean $S(i)$ of all samples is called the silhouette coefficient of the clustering result, and the value of this coefficient should be between $[-1, 1]$. The larger the value is, the better the clustering effect is [61]. According to the calculation, the silhouette coefficients of HDBSCAN, DBSCAN and K-means are 0.6043, 0.3378 and 0.5488, respectively.

6.4. Cluster View

The cluster view of HDBSCAN is shown in Figure 14, and a total of 16 types of clusters are formed by polymerization. For the "PDIA-SL" dataset in the Pudong International Airport case, DBSCAN and K-means clustering generate 8 and 9 clusters respectively, as shown in Figure 19a,b. The number of clusters obtained by HDBSCAN is much larger than that of DBSCAN and K-means, and it can organize more alternative schemes in terms of station layout.

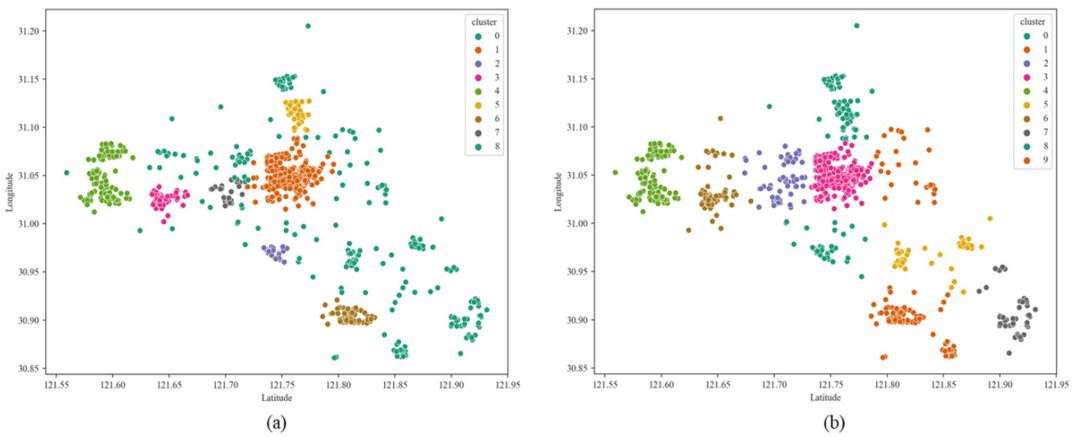


Figure 19. Clustering results: (**a**) DBSCAN analysis on the "PDIA-SL" dataset; (**b**) K-means analysis on the "PDIA-SL" dataset.

In summary, we can obtain the analysis performance and results of the three algorithms DBSCAN, K-means and HDBSCAN on the "PDIA-SL" dataset, and we list these results in Table 3 for comparison.

Table 3. Comparison of clustering results of DBSCAN, K-means and HDBSCAN.

Algorithm	Time Complexity	Space Complexity	Silhouette Coefficient	Number of Clusters
DBSCAN	$O(68025)$	$O(825)$	0.3378	8
K-means	$O(37125)$	$O(1668)$	0.5488	9
HDBSCAN	$O(68025)$	$O(825)$	0.6043	16

7. Conclusions

To more scientifically and rationally plan the station layout and route selection in the southwest of Pudong International Airport, this paper optimizes from station spacing, location and geospatial optimization.

In the process of station spacing optimization, this paper comprehensively considers the factors such as the construction cost, operating cost, social benefits, travel time cost and value-added land development along the Metro line. This paper thoroughly investigates the various data required for calculating the influencing factors and establishes a relatively complete cost-oriented optimization model of station spacing. The optimal station spacing is 683.6 m, which is greater than the station spacing of buses and avoids the waste of public transportation resources.

In the process of location optimization, this paper analyzes the factors that affect the layout of urban rail transit stations, including urban scale, economy, form and land use patterns, passenger flow distribution, and geographical environment. This paper introduces the principle and calculation process of the HDBSCAN algorithm in detail and uses HDBSCAN to analyze the "PDIA-SL" dataset. The position of the ideal points was obtained by clustering, and the ideal location of the station selection was obtained by fitting the centroid of the clusters. At the same time, this paper uses the DBSCAN algorithm and K-means to analyze the dataset "PDIA-SL", and compares the analysis results with HDBSCAN. Through the calculation, the time complexities of DBSCAN, K-means and HDBSCAN are $O(68025)$, $O(37125)$ and $O(68025)$. The space complexities are $O(825)$, $O(1668)$ and $O(825)$. The silhouette coefficients are 0.3378, 0.5488 and 0.6043. The HDBSCAN algorithm has obvious advantages regarding space complexity and silhouette coefficient. Meanwhile, the cluster view of HDBSCAN displays 16 clusters, which can organize more alternative schemes.

In the process of geospatial optimization, this paper draws the distribution of POI, land use types and elevation map of 12.5 m DEM of the Pudong New Area. This paper considers setting the station selection at about 300 m away from residential, commercial, factorial buildings and large stadiums, and avoiding groundwater, cultivated land, and wetlands as much as possible. In this paper, the station is set up at the road intersection as much as possible to attract passenger flow and facilitate the transfer between the buses and subways.

These optimization results can help urban planners and decision-makers understand the influence of various factors in station spacing optimization and location optimization. We can also realize the good clustering performance of the HDBSCAN algorithm in urban rail transit planning. The five routing schemes proposed in this paper are well integrated with these optimization results. They can provide a basis for a new round of urban rail transit development in the southwest direction of Pudong International Airport. These routing schemes aim to improve the expected benefits of passenger flow to Pudong International Airport.

Through the case study of Pudong International Airport, we want to apply the method to a more general situation. According to the practical experience of urban development at home and abroad, the planar geometric schemes of the urban arterial road network can

be summarized into five types: square grid, ring and radial, freestyle, mixed and group. The road network of square grid is the most common type, whose traffic organization is simple and road capacity is enormous; the ring and radial road network is convenient for the rapid connection between the city center and the suburbs, which is often used in megacities; the freestyle road system can better cope with complex terrain. The road system of Shanghai is composed of the above three basic schemas and cities like Shanghai also include Beijing, Changchun, Nanjing and Hefei in China. The research method of station layout, which integrates station spacing, ideal location and spatial analysis optimization, can provide a reference for developing rail transit in these metropolises. In addition, the limitation of this paper is mainly the spatial dependence and heterogeneity of geographic data information, which leads to the conclusions drawn in this paper only applicable to above-mentioned specific types of cities. In addition, the direction of future improvement may lie in the analysis of whether it can be applied to other traffic issues: Can we use this method to analyze the relationship between traffic flow and signal lights, and optimize signal light placement? Can we use this method to analyze the relationship between conveying efficiency and highway entrances and exits, and optimize the layout of entrances and exits?

Author Contributions: Conceptualization, M.P. and P.Y.; methodology, P.Y.; software, P.Y.; validation, P.Y.; formal analysis, P.Y.; investigation, P.Y.; resources, M.P.; data curation, M.P.; writing—original draft preparation, P.Y.; writing—review and editing, M.P.; visualization, P.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data, models or code generated or used during the study are available from the authors upon request.

Conflicts of Interest: The authors declare that there is no conflict of interest in the publication of this paper.

References

1. Cheng, J.; Yin, P. Analysis of the Complex Network of the Urban Function under the Lockdown of COVID-19: Evidence from Shenzhen in China. *Mathematics* **2022**, *10*, 2412. [CrossRef]
2. Li, S.; Zhou, Y.; Kundu, T.; Sheu, J.B. Spatiotemporal variation of the worldwide air transportation network induced by COVID-19 pandemic in 2020. *Transp. Policy* **2021**, *111*, 168–184. [CrossRef] [PubMed]
3. Liu, M.B.; Liao, S.M. A case study on the underground rapid transport system (URTS) for the international airport hubs: Planning, application and lessons learnt. *Tunn. Undergr. Space Technol.* **2018**, *80*, 114–122. [CrossRef]
4. Cheng, J.; Xie, Y.; Zhang, J. Industry structure optimization via the complex network of industry space: A case study of Jiangxi Province in China. *J. Clean. Prod.* **2022**, *338*, 130602. [CrossRef]
5. Mohaymany, A.S.; Gholami, A. Multimodal feeder network design problem: Ant colony optimization approach. *J. Transp. Eng.* **2010**, *136*, 323–331. [CrossRef]
6. Lai, X. *Optimization of Station Locations and Track Alignments for Rail Transit Lines*; UMI: London, UK, 2012; pp. 87–120.
7. Saidi, S.; Wirasinghe, S.C.; Kattan, L. Long-term planning for ring-radial urban rail transit networks. *Transp. Res. Part B Methodol.* **2016**, *86*, 128–146. [CrossRef]
8. Lv, Z.; He, D.; Jia, H.F.; Li, C.B. *Research on the Layout of the Integrative Urban Rail Transit Line Station*; Trans Tech Publications Ltd.: Zurich, Switzerland, 2013; pp. 1222–1229.
9. Chai, S.; Liang, Q.; Zhong, S. Design of urban rail transit network constrained by urban road network, trips and land-use characteristics. *Sustainability* **2019**, *11*, 6128. [CrossRef]
10. Xu, H.; Qiu, Z.; Yuan, H. Research on the Location of Urban Rail Transit Station based on the Integration of Land Use Layout and Regional Transportation Network. In Proceedings of the 2021 6th International Conference on Transportation Information and Safety (ICTIS), Wuhan, China, 22–24 October 2021.
11. Xia, S.; Peng, D.; Meng, D.; Zhang, C.; Wang, G.; Giem, E.; Chen, Z. A fast adaptive k-means with no bounds. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 87–99. [CrossRef]
12. Jian, S.; Pang, G.; Cao, L.; Lu, K.; Gao, H. Cure: Flexible categorical data representation by hierarchical coupling learning. *IEEE Trans. Knowl. Data Eng.* **2018**, *31*, 853–866. [CrossRef]

13. Yin, P.; Cheng, J.; Peng, M. Analyzing the Passenger Flow of Urban Rail Transit Stations by Using Entropy Weight-Grey Correlation Model: A Case Study of Shanghai in China. *Mathematics* **2022**, *10*, 3506. [CrossRef]
14. Li, M.X.; Wang, X.Q. Data analysis of blast furnace gas center based on STING grid clustering. *Mater. Sci. Eng.* **2018**, *439*, 032017. [CrossRef]
15. Khan, K.; Rehman, S.U.; Aziz, K.; Fong, S.; Sarasvady, S. DBSCAN: Past, present and future. In Proceedings of the Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014), Bangalore, India, 17–19 February 2014.
16. McInnes, L.; Healy, J.; Astels, S. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* **2017**, *2*, 205. [CrossRef]
17. Melvin, R.L.; Xiao, J.; Godwin, R.C.; Berenhaut, K.S.; Salsbury, F.R. Visualizing correlated motion with HDBSCAN clustering. *Protein Sci.* **2018**, *27*, 62–75. [CrossRef] [PubMed]
18. Ghamarian, I.; Marquis, E.A. Hierarchical density-based cluster analysis framework for atom probe tomography data. *Ultramicroscopy* **2019**, *200*, 28–38. [CrossRef]
19. Wang, L.; Chen, P.; Chen, L.; Mou, J. Ship AIS trajectory clustering: An HDBSCAN-based approach. *J. Mar. Sci. Eng.* **2021**, *9*, 566. [CrossRef]
20. Liu, P.; Yao, H.; Dai, H.; Fu, W. The Detection and Following of Human Legs Based on Feature Optimized HDBSCAN for Mobile Robot. *J. Phys.* **2022**, *2216*, 012009. [CrossRef]
21. Givoni, M.; Chen, X. Airline and railway disintegration in China: The case of Shanghai Hongqiao Integrated Transport Hub. *Transp. Lett.* **2017**, *9*, 202–214. [CrossRef]
22. Chang, Y.C. Factors affecting airport access mode choice for elderly air passengers. *Transp. Res. Part E Logist. Transp. Rev.* **2013**, *57*, 105–112. [CrossRef]
23. Aqib, M.; Mehmood, R.; Alzahrani, A.; Katib, I.; Albeshri, A.; Altowaijri, S.M. Rapid transit systems: Smarter urban planning using big data, in-memory computing, deep learning, and GPUs. *Sustainability* **2019**, *11*, 2736. [CrossRef]
24. Jiao, L.; Shen, L.; Shuai, C.; Tan, Y.; He, B. Measuring crowdedness between adjacent stations in an urban metro system: A chinese case study. *Sustainability* **2017**, *9*, 2325. [CrossRef]
25. Wu, S.S.; Cheng, J.; Lo, S.M.; Chen, C.C.; Bai, Y. Coordinating urban construction and district-level population density for balanced development: An explorative structural equation modeling analysis on Shanghai. *J. Clean. Prod.* **2021**, *312*, 127646. [CrossRef]
26. Caparros-Midwood, D.; Barr, S.; Dawson, R. Optimised spatial planning to meet long term urban sustainability objectives. Computers. *Environ. Urban Syst.* **2015**, *54*, 154–164. [CrossRef]
27. Arbex, R.O.; da Cunha, C.B. Efficient transit network design and frequencies setting multi-objective optimization by alternating objective genetic algorithm. *Transp. Res. Part B Methodol.* **2015**, *81*, 355–376. [CrossRef]
28. Liu, B.; Sun, S.; Mu, R. Performance Evaluation of Railway Logistics Based on the Layers of Entropy and Grey Correlation Degrees. In Proceedings of the International Conference of Logistics Engineering & Management, Shanghai, China, 9–11 October 2014.
29. Zheng, J. The Relationship Between Property Value and Urban Rapid Rail Transit: Based on Improved Hedonic Price Model. Ph.D. Thesis, Tsinghua University, Beijing, China, 2004.
30. Cheng, J. Analyzing the factors influencing the choice of the government on leasing different types of land uses: Evidence from Shanghai of China. *Land Use Policy* **2020**, *90*, 104303. [CrossRef]
31. Cheng, J. Data analysis of the factors influencing the industrial land leasing in Shanghai based on mathematical models. *Math. Probl. Eng.* **2020**, *2020*, 1–11. [CrossRef]
32. Lee, B.; Gordon, P.; Moore, J.E.; Richardson, H.W. The attributes of residence/workplace areas and transit commuting. *J. Transp. Land Use* **2011**, *4*, 43–63. [CrossRef]
33. Cheng, J.; Xie, Y.; Zhang, J. Analyzing the Urban Hierarchical Structure Based on Multiple Indicators of Economy and Industry: An Econometric Study in China. *CMES Comput. Model. Eng. Sci.* **2022**, *131*, 1831–1855. [CrossRef]
34. Cooper, C.H.; Harvey, I.; Orford, S.; Chiaradia, A.J. Using multiple hybrid spatial design network analysis to predict longitudinal effect of a major city centre redevelopment on pedestrian flows. *Transportation* **2021**, *48*, 643–672. [CrossRef]
35. Qiang, H.; Hu, L. Population and capital flows in metropolitan Beijing, China: Empirical evidence from the past 30 years. *Cities* **2022**, *120*, 103464. [CrossRef]
36. Zornoza-Gallego, C. Means of Transport and Population Distribution in Metropolitan Areas: An Evolutionary Analysis of the Valencia Metropolitan Area. *Land* **2022**, *11*, 657. [CrossRef]
37. Cheng, J. Analysis of commercial land leasing of the district governments of Beijing in China. *Land Use Policy* **2021**, *100*, 104881. [CrossRef]
38. Cheng, J. Residential land leasing and price under public land ownership. *J. Urban Plan. Dev.* **2021**, *147*, 05021009. [CrossRef]
39. Cheng, J.; Luo, X. Analyzing the Land Leasing Behavior of the Government of Beijing, China, via the Multinomial Logit Model. *Land* **2022**, *11*, 376. [CrossRef]
40. Cheng, J. Analysis of the factors influencing industrial land leasing in Beijing of China based on the district-level data. *Land Use Policy* **2022**, *122*, 106389. [CrossRef]
41. Caros, N.S.; Guo, X.; Stewart, A.; Attanucci, J.; Smith, N.; Nioras, D.; Zimmer, A. Ridership and Operations Visualization Engine: An Integrated Transit Performance and Passenger Journey Visualization Engine. *Transp. Res. Rec.* **2023**, *2677*, 1082–1097. [CrossRef]

42. Cheng, J. Mathematical models and data analysis of residential land leasing behavior of district governments of Beijing in China. *Mathematics* **2021**, *9*, 2314. [CrossRef]
43. Zou, Y.; Wang, L.; Xue, Z.; Jiang, M.; Lu, X.; Yang, S.; Yu, X. Impacts of agricultural and reclamation practices on wetlands in the Amur River Basin, Northeastern China. *Wetlands* **2018**, *38*, 383–389. [CrossRef]
44. Susilo, Y.O.; Dijst, M. How Far is Too Far? Travel time ratios for activity participation in the Netherlands. *Transp. Res. Rec.* **2009**, *2134*, 89–98. [CrossRef]
45. Yiu, C.Y.; Wong, S.K. The effects of expected transport improvements on housing prices. *Urban Stud.* **2005**, *42*, 113–125. [CrossRef]
46. Campello, R.J.; Moulavi, D.; Sander, J. Density-based clustering based on hierarchical density estimates. In Proceedings of the Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, Gold Coast, Australia, 14–17 April 2013.
47. Ko, K.; Cao, X.J. The impact of Hiawatha Light Rail on commercial and industrial property values in Minneapolis. *J. Public Transp.* **2013**, *16*, 47–66. [CrossRef]
48. Yin, P.; Cheng, J. A MySQL-based software system of urban land planning database of Shanghai in China. *CMES Comput. Model. Eng. Sci.* **2023**, *135*, 2387–2405. [CrossRef]
49. Dette, H.; Wu, W. Detecting relevant changes in the mean of nonstationary processes—A mass excess approach. *Ann. Stat.* **2019**, *47*, 3578–3608. [CrossRef]
50. Basora, L.; Olive, X.; Dubot, T. Recent advances in anomaly detection methods applied to aviation. *Aerospace* **2019**, *6*, 117. [CrossRef]
51. Zheng, S.; Kahn, M.E. Land and residential property markets in a booming economy: New evidence from Beijing. *J. Urban Econ.* **2008**, *63*, 743–757. [CrossRef]
52. Graham, D.J.; Carbo, J.M.; Anderson, R.J.; Bansal, P. Understanding the costs of urban rail transport operations. *Transp. Res. Part B Methodol.* **2020**, *138*, 292–316.
53. Sun, Y.; Shi, J.; Schonfeld, P.M. Identifying passenger flow characteristics and evaluating travel time reliability by visualizing AFC data: A case study of Shanghai Metro. *Public Transp.* **2016**, *8*, 341–363. [CrossRef]
54. González-Gil, A.; Palacin, R.; Batty, P.; Powell, J.P. A systems approach to reduce urban rail energy consumption. *Energy Convers. Manag.* **2014**, *80*, 509–524. [CrossRef]
55. Guo, S.; Yu, L.; Chen, X.; Zhang, Y. Modelling waiting time for passengers transferring from rail to buses. *Transp. Plan. Technol.* **2011**, *34*, 795–809. [CrossRef]
56. Cai, J.; Wei, H.; Yang, H.; Zhao, X. A novel clustering algorithm based on DPC and PSO. *IEEE Access* **2020**, *8*, 88200–88214. [CrossRef]
57. Kifana, B.D.; Abdurohman, M. Great circle distance methode for improving operational control system based on gps tracking system. *Int. J. Comput. Sci. Eng.* **2012**, *4*, 647.
58. Marini, F.; Walczak, B. Particle swarm optimization (PSO). A tutorial. *Chemom. Intell. Lab. Syst.* **2015**, *149*, 153–165. [CrossRef]
59. Pakhira, M.K. A linear time-complexity k-means algorithm using cluster shifting. In Proceedings of the 2014 international conference on computational intelligence and communication networks, Bhopal, India, 14–16 November 2014.
60. Arora, P.; Varshney, S. Analysis of k-means and k-medoids algorithm for big data. *Procedia Comput. Sci.* **2016**, *78*, 507–512. [CrossRef]
61. Dinh, D.T.; Fujinami, T.; Huynh, V.N. Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient. In Proceedings of the Knowledge and Systems Sciences: 20th International Symposium, Da Nang, Vietnam, 29 November 2019.