sustainability

MDPI

Article

Exploring the Relationship between Land Use and Congestion Source in Xi'an: A Multisource Data Analysis Approach

Duo Wang, Hong Chen *, Chenguang Li and Enze Liu

College of Transportation Engineering, Chang'an University, Xi'an 710000, China
* Correspondence: glch@chd.edu.cn

Abstract: Traffic congestion is a critical problem in urban areas, and understanding the relationship between land use and congestion source is crucial for traffic management and urban planning. This study investigates the relationship between land-use characteristics and congestion pattern features of source parcels in the Second Ring Road of Xi'an, China. The study combines cell-phone data, POI data, and land-use data for the empirical analysis, and uses a spatial clustering approach to identify congested road sections and trace them back to source parcels. The correlations between building factors and congestion patterns are explored using the XGBoost algorithm. The results reveal that residential land and residential population density have the strongest impact on congestion clusters, followed by lands used for science and education and the density of the working population. The study also shows that a small number of specific parcels are responsible for the majority of network congestion. These findings have important implications for urban planners and transportation managers in developing targeted strategies to alleviate traffic congestion during peak periods.

Keywords: human mobility; congestion source analysis; land use; cell-phone data; machine learning

Citation: Wang, D.; Chen, H.; Li, C.; Liu, E. Exploring the Relationship between Land Use and Congestion Source in Xi'an: A Multisource Data Analysis Approach. *Sustainability* **2023**, *15*, 9328. https://doi.org/10.3390/su15129328

Academic Editor: Máté Zöldy

Received: 12 May 2023
Revised: 1 June 2023
Accepted: 6 June 2023
Published: 9 June 2023

1. Introduction

With the growth of the population and motor vehicle ownership, multiple sections of the urban road network in certain areas are often congested at the same time during peak hours. In the United States (US), traffic congestion was estimated to additionally cost drivers over 88 billion USD in the year of 2019 [1]. In China, at least 24% additional travel time is required to commute during peak periods in major cities such as Beijing, Shijiazhuang, and Chongqing [2]. Not only does traffic congestion cause direct economic loss, but it also raises a series of negative environmental issues, such as anabatic carbon emissions and deteriorative greenhouse effects [3]. The increasingly serious and extensive traffic congestion affects the efficiency and safety of road travel, and traffic congestion has gradually become a serious challenge to the sustainable development of large cities and even small and medium-sized cities.

Wang et al. [4] found that vehicles on congested roads are not scattered from individual parcels, but concentrated from some common source parcels. This finding reflects the necessity of congestion traceability, i.e., by tracing the source of congestion and analyzing various factors related to the source of congestion, adjustments can be made on the traffic demand side by means of urban renewal or control to alleviate the congestion problem.

The main method of congestion traceability studies is to trace the source of traffic flow causing road congestion by analyzing the traffic flow OD of the surrounding traffic area and constructing a dichotomous network of road use to explore the relationship between the source of congestion and the congested road [4–11]. Such congestion traceability traces the source mainly to the traveler's residence and departure traffic cell outside the road system, and there are also studies that consider the few travelers causing congestion as the source of congestion. Another type of congestion tracing involves simulating the congestion propagation process within the traffic network [12] or based on the roadway

segment importance [13] to identify the internal sources of congestion in the road system, such as the roadway segments, intersections, or ramps where the congestion originates.

In this paper, on the basis of multiple sources of data such as cell phone signaling data and POI data, congested road sections are identified and traced by road section travel time and average speed of road sections, and congested road sections originating from the same parcel are treated as a congestion group. This group is analyzed by spatial clustering, and its congestion cluster characteristics are obtained. Finally, the relationship between the land-use characteristics of the parcel and the characteristics of the congestion group is explored by using XGBoost algorithm. This paper is based on a spatial clustering analysis. In this paper, the relationship between the land-use characteristics and congestion cluster characteristics of the retroactive parcels in Xi'an Second Ring Road is analyzed in depth for 520 parcels, taking the traffic network and parcels in Xi'an Second Ring Road during the morning peak on 20 July 2021 as an example, and the causes of congestion are further explored.

The structure of this paper is as follows: Section 1 is the introduction; Section 2 is the literature review, which gives an overview of congestion tracing and land-use analysis; Section 3 is the data description, which introduces the data format used, data preprocessing method, and selected cases; Section 4 is the source tracing and analyzing method, which introduces the method framework, congestion tracing method, and XGBoost algorithm; Section 5 takes a specific case as the research object to start the analysis; Section 6 summarizes the results of the whole paper and presents the shortcomings and outlook.

2. Literature Review

2.1. Tracing the Source of Congestion

The study of congestion traceability in road traffic systems has undergone a certain developmental history. Initially, Wang et al. [4] considered the traveler's residence as a static congestion source, but this approach could only provide very limited information. Currently, the main congestion traceability research method is to first estimate the travel OD matrix, assign the travel demand to the road network, and then construct a dichotomous network model of road use to locate the top 80% of parcels contributing to congested road traffic as source parcels according to the congestion status of the road network [4,6–9]. Wang et al. [6,9] started to trace the dynamic congestion source, i.e., the departure place within a certain time period as the congestion source. Meanwhile, congestion tracing research has been extended from tracing microscopic roadway congestion sources to macroscopic regional congestion sources [8]. Depending on the transportation system under study, Wang et al. [5,7] elaborated the data and methods required for traceability, which can use mobile cellular call detail records (CDRs), cellular signaling data, radiofrequency identification (RFID) data, and metro card data to estimate travel demand, assign traffic flows on the basis of the travel OD matrix and the corresponding road network, locate the major congestion sources, construct a road use dichotomous network to explore the relationship between congestion sources and road sections, and finally propose congestion optimization methods on this basis. For macro-regional congestion traceability, Wang et al. [8] proposed to select the source parcels with the largest proportion as the city-wide congestion sources according to the travel time delay caused by each source.

Furthermore, many other scholars have promoted research in other directions of congestion traceability, arguing that travelers are the sources of congestion, and alleviating congestion by changing the path choice of travelers. For example, He et al. [10] estimated the travel demand in two networks using Beijing metro card data and San Francisco CSD, considered a small fraction of travelers experiencing severe congestion as congestion sources, and proposed a hybrid path selection model combining shortest path selection and least cost path selection for congestion sources. When congestion occurs on highways or expressways, some scholars conducted separate studies on their congestion sources considering the closed nature and special features of the roads themselves. For example, Li et al. [11] obtained OD data of regional freeways from toll station and entrance ramp

data, and then defined entrance ramps, which play a major contribution to freeway section traffic, as the main vehicle source of freeway bottlenecks.

2.2. Land Use

The land-use characteristics include socioeconomic attributes and spatiotemporal correlation and heterogeneity, in addition to the traditional 5D model (density, diversity, design, destination accessibility, and transit stop distance). In contrast, traffic-related studies are richer and more diverse, including traffic volume [14–18], travel behavior characteristics [19,20], and other traffic phenomena due to travel, such as congestion and emissions. Models to study the relationship between the two are currently dominated by various types of regression models and machine learning algorithms.

Many studies have shown that land use has a close relationship with traffic congestion. For example, Zhang et al. [21] extracted POI and real-time traffic data from an electronic map of Beijing's Fourth Ring Road area and identified major traffic congestion areas by cluster analysis; the results showed that a high proportion of commercial land use had a significant impact on traffic congestion, while a linear regression analysis found that a reasonable ratio of land-use types could effectively reduce congestion time. Qin et al. [22] extracted urban built environment features from public streetscape images and POI data, and then proposed a multigraph convolutional network structure to model the spatial dependence between traffic congestion on road networks. Bao et al. [23] explored the spatiotemporal relationship between traffic congestion and the built environment. The study found that congestion before the weekend may be more severe and have a more lasting impact on satellite cities, and then confirmed the positive impact of public transportation in alleviating traffic pressure. Shen et al. [24] investigated the interaction between land use and parking utilization in alleviating congestion in residential areas of Xi'an in 2017 and found that the imbalance between land use and parking facilities led to long-distance, cross-regional traffic, and that the difference in the availability of parking spaces between supply and demand was the main cause of traffic congestion during commuting periods. Schoeman and Schoeman [25] aimed to develop a practical assessment and development approach for evaluating the impact of land-use planning on traffic generation, emissions, and environmental factors in residential areas, with the objective of informing detailed planning and decision-making processes for selecting preferred development scenarios and guiding stakeholder actions. Yap et al. [26] investigated the impact of land-use patterns on traffic congestion in Kuala Lumpur, revealing that a high proportion of commercial land use contributes significantly to the occurrence of traffic congestion, emphasizing the importance of appropriate land-use planning to mitigate congestion. Rahman et al. [27] analyzed the causes of urban traffic congestion in US cities using a structural equation modeling framework at the mesoscale, revealing the complex nature of congestion and identifying factors such as population size, income and employment agglomeration, transportation infrastructure, mode choice behaviors, community structures, urban density, and socioeconomic factors as key influences, providing insights for policy interventions.

2.3. Research Gap and Potential Contributions

Existing studies have not conducted in-depth analysis on the source characteristics of congestion, and most of them used CSD (cellular signaling data) together with GPS (global positioning system) data for congestion tracing. In this paper, on the basis of CSD, POI (point of interest) data, and LUA (land-use allocation) data, we comprehensively analyze the relationship between the land-use characteristics and congestion patterns of the source parcels in the Second Ring Road of Xi'an, and further explore the causes of congestion.

The main purposes of this paper are as follows: (1) the relationship between land-use features of congestion source parcels and network congestion pattern features is further explored; (2) using the spatial cluster analysis technique, congested road sections traced to the same parcel are treated as a congestion pattern and their network-pattern clusters are obtained; (3) using the XGBoost algorithm, the land-use features of parcels in numerical

vector format are used as independent variables and congestion pattern features in categorical format as dependent variables, and the relationship between them is analyzed. Our research findings provide critical insights into understanding the factors contributing to traffic congestion and offer valuable guidance for developing more effective urban traffic management policies. They may also serve as a useful reference for other researchers interested in exploring the relationship between congestion and land use.

3. Data

3.1. Data Description

CSD, POI, and LUA data are employed in this study from a transportation and urban planning perspective. These data sources offer valuable insights into the travel behavior patterns of individuals, the spatial distribution of various activities, and the allocation of land for different purposes within an urban area. By integrating and analyzing these datasets, this study aims to achieve significant advances in tracing the source of congestion. Specifically, CSD can provide information on travel patterns, origin–destination flows, and congestion hotspots. POI data can help identify areas with high activity levels and potential traffic generators. LUA data can provide insights into the spatial distribution of different land uses, including residential, commercial, and industrial areas. By analyzing these datasets collectively, this study can uncover the underlying causes of congestion, identify key contributing factors, and propose targeted strategies for mitigating traffic congestion in urban areas.

(1) LUA describes the information within the study area, such as land use, POI, resident population, commuter population, and road network density. These data can all be downloaded from the Open Street Map (OSM) [28] and Autonavi development platforms [29].
(2) The CSD are provided by "Smart Footprint" company, which is affiliated with China Unicom, one of the largest telecom operators in China. It records anonymous people's travel trajectory in the way of road nodes. Data attributes include anonymous ID, monthly trip number, path ID, path node sequence, path node number, time passed by node pair, next path node number, and month (partition field) (Table 1).

Table 1. User travel trajectory table.

Notion	Description	Example
Uid	Anonymous ID	*** 92
moi_id	Month trip number	1
route_id	Route ID	1
rn_seq	Route node sequence	16
rn_id	Route node ID	35,857,550
time	Time taken to pass through the node pair	650
is_end	Is the end point?	N
mode	Mode of transportation (1—road, 2—railway, 3—airplane, 4—metro, 0—other)	1
next_rn_id	Next route node ID	35,858,257

***: anounymous hiding.

3.2. Case Selection

The research area of this study is within the Second Ring Road of Xi'an with 139,936 sections. The study area is divided into 680 parcels on the basis of their properties and the road network, and 520 effective parcels are obtained after screening, which is demonstrated in Figure 1. Meanwhile, this study uses the CSD passing through Xi'an Second Ring Road, selecting the morning peak (7:00–9:00 a.m.) on 20 July 2021 as the target period. The reason for choosing this day is that it rained heavily, and the congestion was obvious.

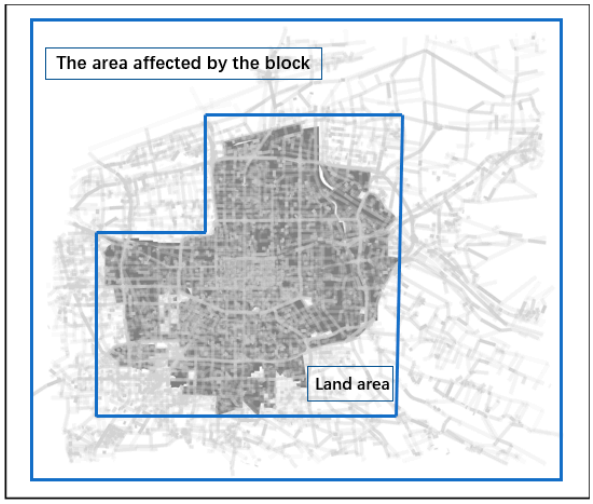


Figure 1. Research scope.

3.3. Data Preprocess

Due to the generation mechanism of CSD and the errors in track matching, there are invalid data in the dataset. These types of data will have a certain adverse impact on the subsequent congestion identification process. Therefore, we set up some rules to identify and remove them. In this study, invalid data included duplicates and error records.

Duplicate data: In this study, we encountered two types of duplicate data, those related to time and those related to location fields. Due to the erratic nature of wireless signals, it is possible for a phone to switch from one base station to another in just 1 s, resulting in multiple locations being recorded for the same time period. To address this issue, we identified these instances and kept the first recorded location while discarding the others. Additionally, when a user remains stationary or moves within a small area, multiple consecutive and identical location records are generated due to periodic position updates. To mitigate this, we retained the first and last location records while discarding the intermediate ones.

Error records: When the users' path tracks are too short or the information is too scarce, we deemed these kinds of data as errors.

4. Source Tracing and Land-Use Analysis

4.1. Methodology Framework

This paper aimed to identify the congested road sections and trace their source by analyzing multiple sources of data, considering the morning peak CSD of 20 July 2021. The methodology framework is shown in Figure 2. First, the road network is matched with the travel track table and grid table to obtain the travel track matrix. Then, due to the problem of the data source, when using vehicle GPS data, we can judge the congestion by the traffic flow and capacity. However, when mobile data are used, it is on a human scale; thus, it is difficult to judge congestion by traffic flow and infrastructure capacity. Thus, the driving time and average speed of road sections are used to identify congested road sections and trace their source through the selected vehicle tracks. The identified congested road sections originating from the same parcel are treated as a congestion group, and the congestion cluster of the group is obtained through spatial clustering analysis. These congestion cluster characteristics are used as the dependent variables in analyzing the relationship between land-use characteristics and congestion cluster characteristics using the XGBoost algorithm. Overall, this paper provides a comprehensive analysis of the causes of congestion and their relationship to land-use characteristics.

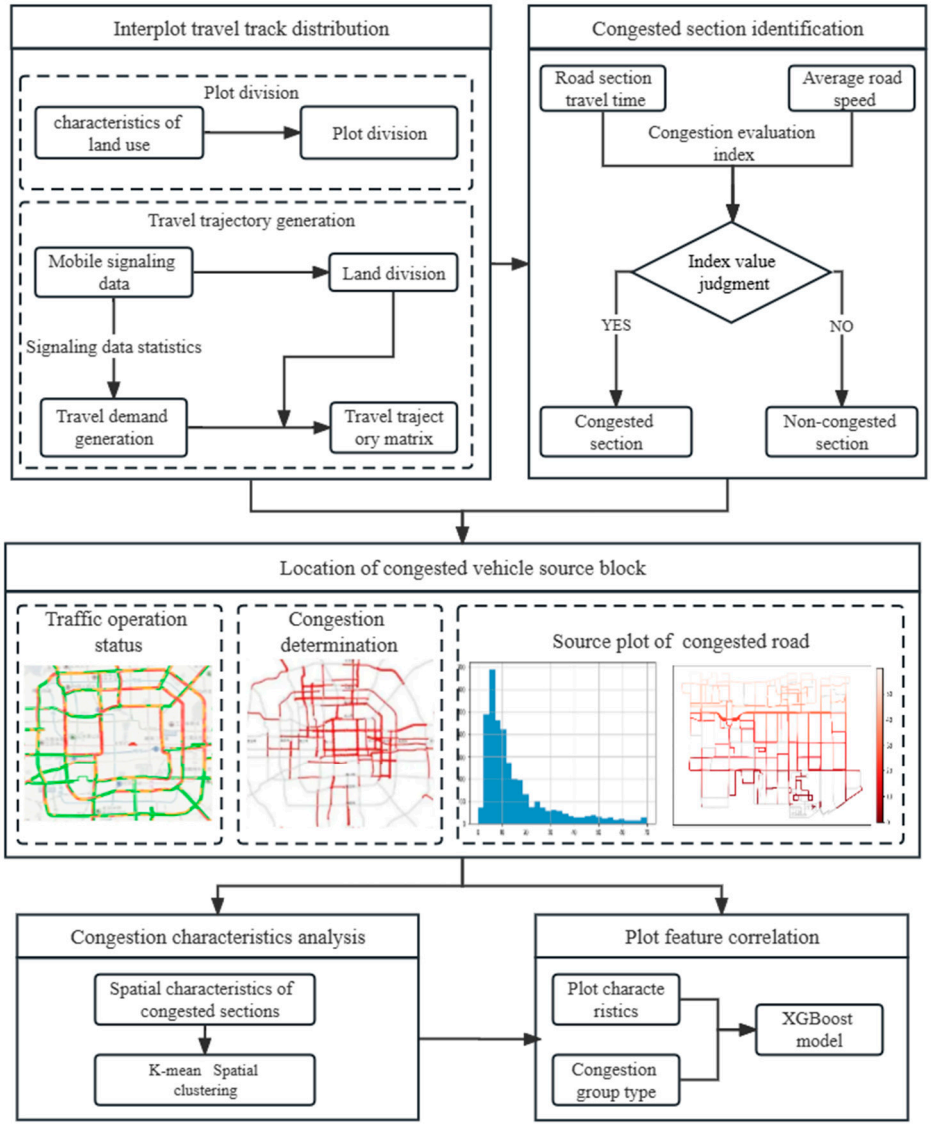


Figure 2. Specific research routes.

4.2. Congestion Traceability Method

Define P as the set of all parcels; (A, B) is the set of all travel OD from parcel A to parcel B, q_{ab} is the travel quantity from grid a ($a \in A$) to grid b ($b \in B$), r_{ab} is the path that q_{ab} goes through, $r^l{}_{ab}$ ($r^l{}_{ab} \in r_{ab}$) is the set of travel paths that pass through road 1, and Q_l a is the traffic flow of the first section. The contribution of each parcel to the traffic flow of the road in the selected time segment is as follows:

$$CO_l(A) = \sum_{S=1}^{P} \sum_{r_{ab}^l} q_{ab} \tag{1}$$

By calculating the total congestion contribution of each parcel, this study locates the global congestion source at the regional level and ranks the congestion contribution. Beginning with the parcel that has the highest congestion contribution, the contribution is accumulated successively. When the cumulative contribution reaches 15% of the total travel volume, the parcel that participates in the accumulation is defined as the urban congestion source.

4.3. Clustering Method

To investigate the spatial similarity of congestion clusters, this study treats congested road segments traced to the same parcel as a congestion group. The distribution of these groups is stored as a 0–1 matrix, and spatial clustering is performed using the k-means method. The optimal number of categories K is determined using the elbow method, which identifies the turning point of inertia (sum of squares within the group) as the best K value. Using this method, we identify the congestion clusters of the congestion groups, as described below:

Define K as the number of categories; $X = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}$ is the dataset containing n D-dimensional data points, and $C = \{c_k, i = 1, 2, \ldots K\}$ represents the K divisions organized into data objects by k-means clustering algorithm. Each division represents a class c_k, and each class has a category center μ_i. Because Euclidean distance has better applicability and efficiency in application, it is selected as the criterion of similarity and distance judgment, and the sum of squares of distances from each point in the class to cluster center μ_k is calculated.

$$J(c_k) = \sum_{x_i \in C_k} \| x_i - \mu_k \|^2 \tag{2}$$

The objective of clustering is to minimize the sum of all kinds of total distance squares $J(C) = \sum_{k=1}^{K} J(c_k)$, and the objective function is as follows:

$$\mathrm{J(C)} = \sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{J(c_k)} = \sum_{\mathrm{k=1}}^{\mathrm{K}} \sum_{\mathrm{x_i C_k}} \mathrm{Px_i - m_k P^2} = \sum_{\mathrm{k=1}}^{\mathrm{K}} \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{d_{ki} Px_i - m_k P^2}, \tag{3}$$

where $d_{ki} = \begin{cases} 1, x_i \in c_i \\ 0, x_i \notin c_i \end{cases}$.

4.4. XGBoost Methods

XGBoost, i.e., extreme gradient boosting, is an optimized distributed gradient boosting library, which aims to be efficient, flexible, and portable. XGBoost is a massively parallel boosting tree tool, which has many advantages compared with traditional methods. First, it is the fastest and best open-source toolkit available for boosting trees, being more than 10 times faster than the usual toolkit. Secondly, missing values in data are regarded as a new data type, thus not having a great impact on the model. Thirdly, compared with other machine learning methods, it combines regularization techniques with tree models and uses quantile loss functions for fitting, thus achieving better prediction accuracy. At the same time, it is more interpretable because it is able to model and visualize the relationship between features and targets thanks to the tree model. Lastly, compared with linear regression, it has the advantage that it cannot only handle nonlinear relations, while also being more robust to outliers. This is because linear regression is sensitive to outliers, whereas XGBoost uses square loss functions and quantile loss functions, which are more robust in these cases.

CART is an implementation method of decision tree, which is called the classification and regression tree. It is widely used in data mining and machine learning and can be applied to tasks such as binary classification, multiclassification, and regression analysis. As a data-driven machine learning model, XGboost uses the CART regression tree model as its tree model construction process.

Given dataset $D = \{(x_i, y_i) : i = 1, 2, L, n, x_i R^p, y_i R\}$, where n is the number of samples, each sample has P characteristics. The model can be expressed as

$$\overline{y_i} = \sum_{t=1}^{k} f_t(x_i) \; f_t \in F, \tag{4}$$

where k is the number of trees, f_t is a function in the function space F, $\overline{y_i}$ is the predicted value, x_i is the i-th data input, and F is the set of carts that are possible.

The objective function of XGBoost is shown below.

$$X_{obj} = \sum_{t=1}^{n} l(y, \overline{y}) + \sum_{k=1}^{K} q(f_k), \tag{5}$$

where $\sum_{t=1}^{n} l(y, \overline{y})$ is used to measure the difference between the predicted score and the real score; $\sum_{k=1}^{K} q(f_k)$ is the regularization term, and the regular term expression is $q(f_k) = gT + l\frac{1}{2}\sum_{j=1}^{T} w_j^2$, where T is the number of leaf nodes; g is used to control the number of leaf nodes; l makes sure that the number of leaves is not too big.

XGBoost uses a gradient lifting strategy, which preserves the existing model and adds a new regression tree to the model one at a time. The iterative process is as follows:

$$\begin{cases} \overline{y_i^0} = 0 \\ \overline{y_i^1} = f_1(x_i) = \overline{y_i^0} + \ f_1(x_i) \\ \overline{y_i^t} = \overline{y_i^{t-1}} + f_t(x_i) \end{cases} \tag{6}$$

Substituting Equation (6) into Equation (5), we can get

$$t^t = \sum_{i=1}^{n} l(y_i, \overline{y_i^{t-1}} + f_t(X_i)) + q(f_t) \tag{7}$$

To find a solution that minimizes the target function, XGBoost approximates it by using the Taylor second-order expansion of the target function at $f_t = 0$.

$$t^t = \sum_{i=1}^{n} [l(y_i, \overline{y_i^{t-1}} + f_t(X_i)) + \frac{1}{2} h_i f_t^2(X_i)] + q(f_t) \tag{8}$$

5. Case Study

5.1. Case Design

The research scope of this paper is the Second Ring Road in Xi'an, with 139,936 sections. On the basis of the main road network and the nature of the parcels, the study area was divided into 680 parcels, and 520 effective parcels were obtained after screening.

In the process of road section congestion recognition, the average speed of all road sections in the road network was ranked from low to high, and the road section with the speed in the top 10% was identified as a collection of congested road sections. The threshold was 36 km/h.

In this study, the XGBoost method under the Python environment was constructed using the Scikit-learn package. The dataset was randomly divided into a training subset and a test subset, accounting for 70% and 30% of the total data, respectively. In addition, when modeling the XGBoost method, important parameters were calibrated, namely, max_depth of the tree and min_child_weight of leaf knot number. To capture the best combination of two important parameters, we used a grid search. The optimal prediction model had max_depth = 3 and min_child_weight = 3.

5.2. Case Result

After preprocessing the data and identifying congestion, we obtained the road congestion conditions during morning rush hour throughout the study period, as shown in Figure 3. The distribution of congestion sections was discrete overall, while it was continu-

ous and dense along the First Ring Road, Second Ring Road, Third Ring Road, and Xi'an Ring Road. Additionally, congestion severity in the southern part of the network was lower compared to other directions, and there was no continuous congestion over long distances. Notably, congestion on the Xi'an Ring Road was more continuous and longer compared to other roads. This suggests that many ring roads bear more traffic volume and resulting congestion than other roads.

Figure 3. Map of congested road sections.

After identifying the congested road segments in the previous step, we analyzed the vehicle source within the research range. The analysis showed the main vehicle sources of the congested roads during the morning rush hour, as depicted in Figure 4 (average number of vehicle sources in 15 min). As shown in the figure, the vehicle source parcels in congested sections were widely dispersed, with the majority of them contributing only a small amount of vehicle flow. Conversely, the sources that significantly contributed to the main vehicle flow were limited in number and mostly located in residential areas between the First and Second Ring Roads.

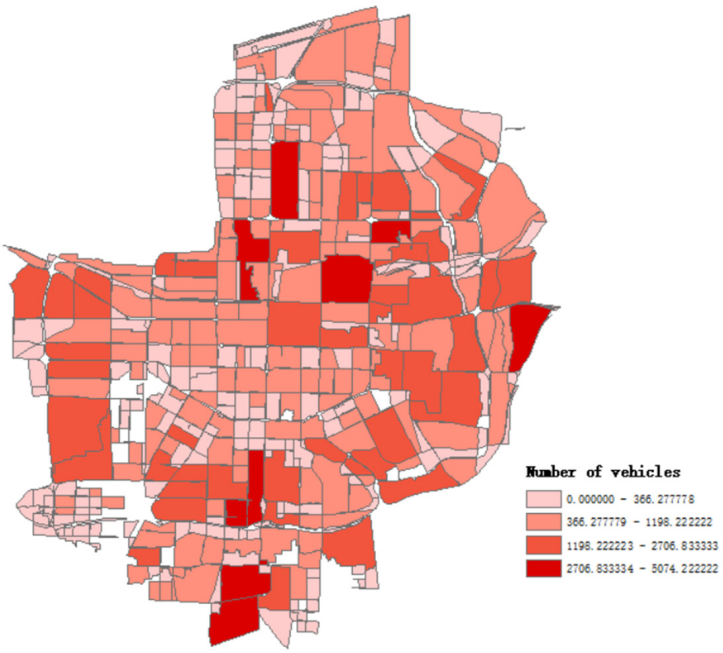


Figure 4. Source parcel distribution.

The distribution of congestion on the road network was modeled using binary vectors, with each variable representing a road section. A value of 1 indicated congestion, while a value of 0 represented no congestion. Each vector represented a congested network sourcing to one parcel. In other words, it reflected the congested roads on the network related to this parcel. Then, k-means spatial clustering analysis was carried out to find the similarity of the spatial distribution. The elbow diagram is shown in Figure 5. As can be seen from the figure, when the category was 8, the downward trend became significantly slower. According to the distribution of congestion sections in different congestion maps, the congestion maps were divided into eight clusters.

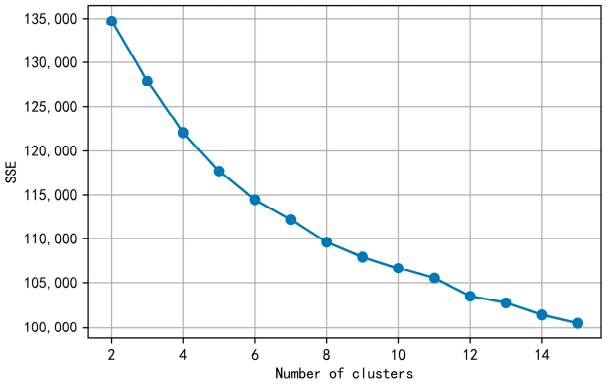


Figure 5. Elbow chart.

Statistical analysis was conducted on the categories of the congestion chart, and the quantity relationship of congested roads and corresponding sources of the eight clusters is shown in Figure 6. The horizontal axis represents the average number of congested roads within a given cluster, while the vertical axis represents the number of sources of congestion within that same cluster. The figure provides insight into the relationship between the number of parcels and the number of congestions on the network. Specifically, it shows the distribution of the number of parcels responsible for the number of congestions they generate on the network. From the diagram, it can be observed that, as the number of congested roads within a cluster increased, there was a corresponding decrease in the number of sources that contribute to the congestion. In other words, there was a negative natural logarithm correlation between the number of congested roads and the number of sources of congestion within a cluster. The results indicate that a small number of specific parcels were responsible for the majority of network congestion, while the vast majority of parcels did not contribute to congestion on the network.

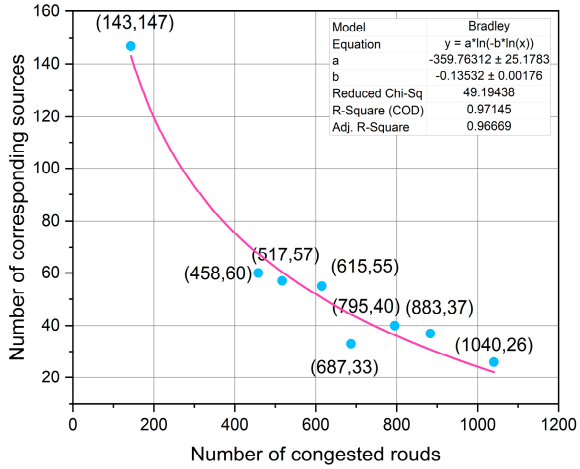
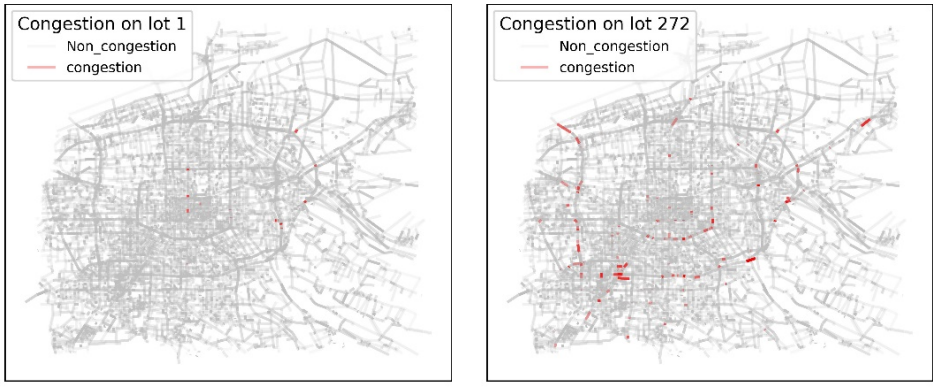
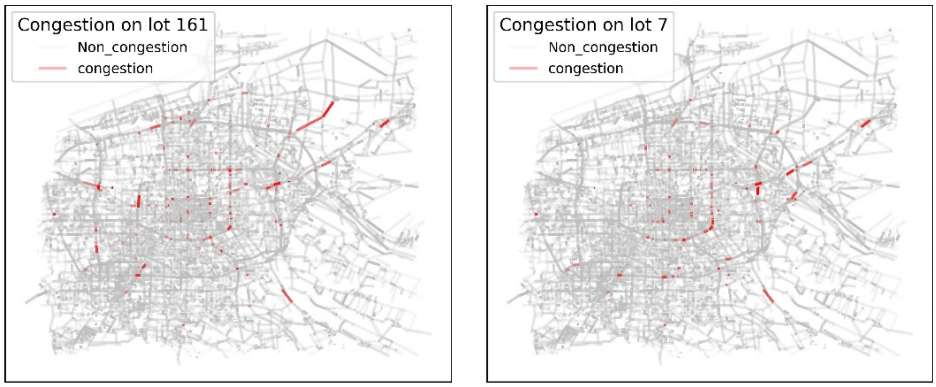


Figure 6. Quantitative correlation between congestion and source.

There were some samples of the spatial distribution of congestions of the eight clusters in Figure 7. Sequence order was based on congested road magnitude within the cluster, with fewer roads first and more roads last. As shown in Figure 7a, the congested sections were sparsely distributed, with only a small number of congested sections scattered on the Xi'an Expressway and central axis. As shown in Figure 7b, the congested sections were sparsely scattered in the south of the road network, mainly concentrated in the First Ring Road, Second Ring Road, Central Axis Road, and Xi'an Ring Highway. Most of them existed in the form of "point" congestion, with few congestion points gradually expanding in the form of "lines". There was no large-scale congestion in the form of a "network". As shown in Figure 7c, most of the congested sections were scattered and sparsely distributed on the road network, with less distribution in the southeast direction, mainly concentrated on the Second and Third Ring Roads. As shown in Figure 7d, the congested sections were scattered and sparsely distributed in the southeast of the road network, mainly concentrated on the First Ring Road, Second Ring Road, and Third Ring Road. As shown in Figure 7e, the congested sections were scattered and sparsely distributed in the southeast and northeast of the road network, mainly concentrated on the First Ring Road, Second Ring Road, and Xi'an Ring Road. As shown in Figure 7f, compared with the previous categories, the congested sections of this category were more densely distributed on the road network, mainly concentrated on the ring roads and radial road around the expressway. As shown in Figure 7g, the congested sections were densely distributed in the road network, mainly concentrated in the First Ring Road, Second Ring Road, Third Ring Road, and Xi'an Ring Road. Congestion was relatively dense and concentrated, with the congestion on the ring roads being particularly severe, forming contiguous linear congestion. As shown in Figure 7h, the congested sections were distributed in the northeast of the road network, mainly concentrated on the ring and radial roads, forming contiguous regional congestion.



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



(**c**) (**d**)

Figure 7. *Cont.*

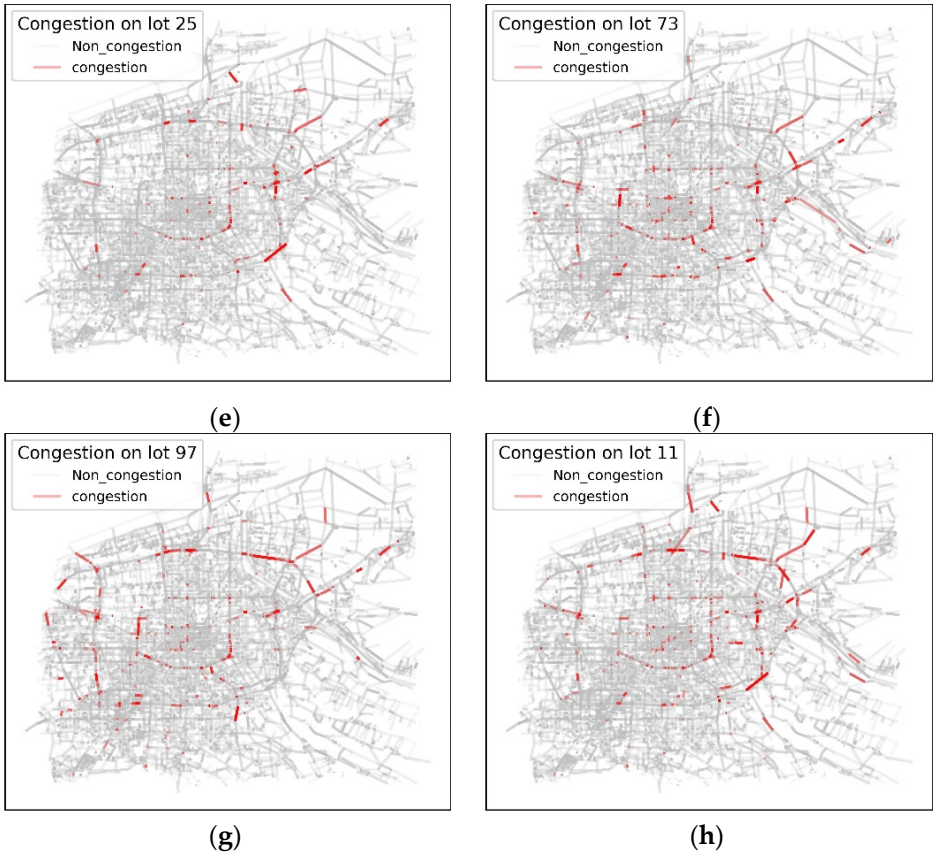


Figure 7. Sample of spatial distribution of clusters. (**a**) sample of cluster 1; (**b**) sample of cluster 2; (**c**) sample of cluster 3; (**d**) sample of cluster 4; (**e**) sample of cluster 5; (**f**) sample of cluster 6; (**g**) sample of cluster 7; (**h**) sample of cluster 8.

5.3. Correlation between Land Use and Congestion Patterns

On the basis of the main road network and properties of the parcels, 520 effective parcels were obtained after screening, as shown in Figure 8. Six variables extracted by ArcGIS were used: (1) road density, (2) working population density, (3) residential population density, (4) office area density, (5) science, education, and cultural service area density, and (6) main land-use nature. The descriptive statistics of the variables are provided in Table 2.

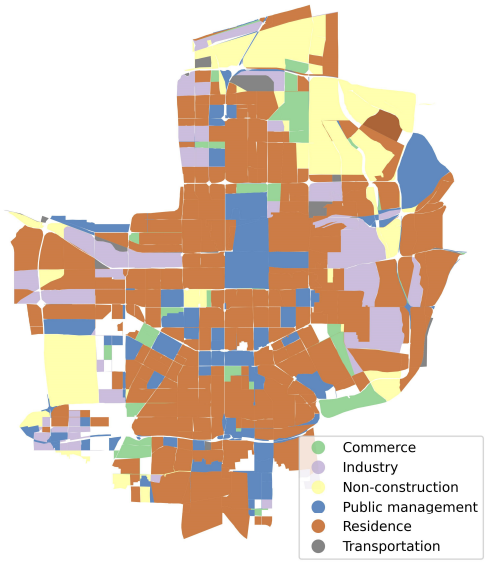


Figure 8. Land-use property map.

Table 2. Description of parcel characteristics.

Variable	Description	Ratio/Average
Road density	Continuous (m/1000 m^2)	11.68
Working population density	Continuous (people/km^2)	7.50
Residential population density	Continuous (people/1000 m^2)	13.57
Office area density	Continuous (units/km^2)	99.53
Density of science areas	Continuous (units/1000 m^2)	36.96
Main land-use nature	Discrete	(see Figure 8)

Taking the property of the land parcel as the input independent variable and the category of congestion cluster as the input dependent variable, the top 10 important indicators affecting the classification were obtained as shown in Table 3.

Table 3. Top 10 important indicators.

Characteristic	Weight
Nature of the land (residential land)	0.141731
Residential population density	0.134664
Number of science and education areas	0.114573
Working population density	0.101248
Nature of land (transportation land)	0.097864
Office space density	0.086793
Road density	0.077864
Nature of land (industrial land)	0.070127
Nature of land (public administration land)	0.067659
Nature of land (commercial land)	0.066427

The analysis of the factors that contribute to congestion clusters revealed that residential land and residential population density had the strongest impact, with correlation values of 0.141731 and 0.134664, respectively. This finding is consistent with the understanding that morning peak traffic trips are predominantly commuting trips, with individuals traveling from their place of residence to their workplace. As such, the attributes of both the residential and the work locations played a significant role in determining the location and intensity of congestion clusters.

The amount of land used for science and education, and the density of the working population were also factors significantly influencing congestion clustering, with correlation values of 0.114573 and 0.101248, respectively. This suggests that the location of educational institutions and workplaces also played an important role in determining the concentration of traffic flow during peak periods.

On the other hand, public management land and economic and commercial land were factors with the weakest influence on congestion clusters, with correlation values of 0.067659 and 0.066427, respectively. This can be attributed to the fact that traffic flow associated with these types of land use was relatively low and, therefore, had less of an impact on the overall clustering of congestion.

Overall, these findings highlight the importance of considering the relationship between land use and traffic flow when developing strategies to mitigate traffic congestion in urban areas. By taking into account the factors that contribute most significantly to congestion, urban planners and transportation managers can develop more targeted solutions to alleviate traffic flow during peak periods.

6. Conclusions

This research article investigated the characteristics of congestion sources and their relationship with land use in the Second Ring Road of Xi'an. The authors used CSD, POI, and LUA data and employed spatial cluster analysis and the XGBoost algorithm to comprehensively analyze the congestion patterns and land-use features of source parcels.

The results revealed that a small number of specific parcels were responsible for the majority of network congestion, while the vast majority of parcels did not contribute to congestion on the network.

The findings indicated that residential land and residential population density had the strongest impact on congestion clusters, followed by science and education land and working population density. When comparing different types of land use, such as public management land and economic and commercial land, it was observed that they had the least significant impact on congestion. In other words, these types of land use had a relatively weaker influence on the occurrence of traffic congestion when compared to other factors. The authors suggest that urban planners and transportation managers should consider the relationship between land use and traffic flow when developing strategies to mitigate traffic congestion in urban areas.

In future research, it would be beneficial to investigate the temporal patterns of congestion and their relationship with the land use. Additionally, the authors suggest that further research could explore the impact of other factors such as transportation infrastructure and public transit systems on congestion patterns. The findings of this study provide important insights into the factors that contribute to traffic congestion and can inform the development of more effective strategies to manage traffic flow in urban areas.

Author Contributions: Conceptualization, D.W. and E.L.; methodology, D.W., C.L. and E.L.; software, C.L.; validation, D.W. and H.C.; formal analysis, D.W. and E.L.; data preparation, D.W.; writing—original draft preparation, D.W., C.L. and E.L.; supervision, H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was funded by the "Special Support Program", Natural Science and Engineering Technology Special Funds for Young Talents in Weinan City.

Data Availability Statement: We gratefully acknowledge the provision of cell phone signaling data by Smart Step Digital Technology Co., Ltd. and China Union.

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

References

1. Traffic Congestion Costs U.S. Cities Billions of Dollars Every Year [Infographic]. Available online: https://www.forbes.com/sites/niallmccarthy/2020/03/10/traffic-congestion-costs-us-cities-billions-of-dollars-every-year-infographic/?sh=420f07404ff8 (accessed on 5 June 2023).
2. The Cities with the Biggest Traffic Jams in China. Available online: https://www.statista.com/chart/16998/the-cities-with-the-biggest-traffic-jams-in-china/ (accessed on 5 June 2023).
3. Chang, Y.S.; Lee, Y.J.; Choi, S.S.B. Is there more traffic congestion in larger cities? -Scaling analysis of the 101 largest U.S. urban centers. *Transp. Policy* **2017**, *59*, 54–63. [CrossRef]
4. Wang, P.; Hunter, T.; Bayen, A.M.; Schechtner, K.; Gonzalez, M.C. Understanding Road Usage Patterns in Urban Areas. *Sci. Rep.* **2012**, *2*, 1001. [CrossRef]
5. Wang, C.C.; Xu, Z.Z.; Du, R.H.; Li, H.F.; Wang, P. A vehicle routing model based on large-scale radio frequency identification data. *J. Intell. Transp. Syst.* **2020**, *24*, 142–155. [CrossRef]
6. Wang, P.; Wang, C.; Lai, J.; Huang, Z.; Ma, J.; Mao, Y. Traffic control approach based on multi-source data fusion. *IET Intell. Transp. Syst.* **2019**, *13*, 764–772. [CrossRef]
7. Wang, C.; Wang, P. Data, Methods, and Applications of Traffic Source Prediction. In *Transportation Analytics in the Era of Big Data*; Ukkusuri, S.V., Yang, C., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 105–120.
8. Wang, P.; Lu, H.; Tan, Q.; Xiong, Y.; Mao, Y.; Li, L. A data fusion approach for locating driver sources using mobile phone signaling data and taxi GPS data. *J. Harbin Inst. Technol.* **2018**, *50*, 96–100, 107.
9. Wang, J.; Wei, D.; He, K.; Gong, H.; Wang, P. Encapsulating Urban Traffic Rhythms into Road Networks. *Sci. Rep.* **2014**, *4*, 4141. [CrossRef]
10. He, K.; Xu, Z.Z.; Wang, P.; Deng, L.B.; Tu, L. Congestion Avoidance Routing Based on Large-Scale Social Signals. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 2613–2626. [CrossRef]
11. Li, M.L.; Yang, H.; Guo, B.; Dai, J.J.; Wang, P. Driver Source-Based Traffic Control Approach for Mitigating Congestion in Freeway Bottlenecks. *J. Adv. Transp.* **2022**, *2022*, 3536979. [CrossRef]
12. Yue, W.; Li, C.; Chen, Y.; Duan, P.; Mao, G. What Is the Root Cause of Congestion in Urban Traffic Networks: Road Infrastructure or Signal Control? *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 8662–8679. [CrossRef]

13. Wang, J.; Gu, Q.; Wu, J.; Liu, G.; Xiong, Z. Traffic Speed Prediction and Congestion Source Exploration: A Deep Learning Method. In Proceedings of the 16th IEEE International Conference on Data Mining (ICDM), Barcelona, Spain, 12–15 December 2016; pp. 499–508.
14. Ma, X.; Zhang, J.; Ding, C.; Wang, Y. A geographically and temporally weighted regression model to explore the spatiotemporal influence of built environment on transit ridership. *Comput. Environ. Urban Syst.* **2018**, *70*, 113–124. [CrossRef]
15. An, D.; Tong, X.; Liu, K.; Chan, E.H.W. Understanding the impact of built environment on metro ridership using open source in Shanghai. *Cities* **2019**, *93*, 177–187. [CrossRef]
16. Guo, J.; Zhong, S.; Yang, F.; Zhang, J.; Ran, B. Spatial and Temporal Distribution Model for Travel Origin-Destination Based on Multi-Source Data. In Proceedings of the 19th COTA International Conference of Transportation Professionals (CICTP)—Transportation in China 2025, Nanjing, China, 6–8 July 2019; pp. 5280–5292.
17. Li, S.; Lyu, D.; Huang, G.; Zhang, X.; Gao, F.; Chen, Y.; Liu, X. Spatially varying impacts of built environment factors on rail transit ridership at station level: A case study in Guangzhou, China. *J. Transp. Geogr.* **2020**, *82*, 102631. [CrossRef]
18. Shao, Q.; Zhang, W.; Cao, X.; Yang, J.; Yin, J. Threshold and moderating effects of land use on metro ridership in Shenzhen: Implications for TOD planning. *J. Transp. Geogr.* **2020**, *89*, 102878. [CrossRef]
19. Nasri, A.; Zhang, L. Impact of Metropolitan-Level Built Environment on Travel Behavior. *Transp. Res. Rec.* **2012**, *2323*, 75–79. [CrossRef]
20. Chen, P.; Shen, Q.; Childress, S. A GPS data-based analysis of built environment influences on bicyclist route preferences. *Int. J. Sustain. Transp.* **2018**, *12*, 218–231. [CrossRef]
21. Zhang, T.; Sun, L.; Yao, L.; Rong, J. Impact Analysis of Land Use on Traffic Congestion Using Real-Time Traffic and POI. *J. Adv. Transp.* **2017**, *2017*, 7164790. [CrossRef]
22. Qin, K.; Xu, Y.; Kang, C.; Kwan, M.-P. A graph convolutional network model for evaluating potential congestion spots based on local urban built environments. *Trans. Gis* **2020**, *24*, 1382–1401. [CrossRef]
23. Bao, Z.; Ng, S.T.; Yu, G.; Zhang, X.; Ou, Y. The effect of the built environment on spatial-temporal pattern of traffic congestion in a satellite city in emerging economies. *Dev. Built Environ.* **2023**, *14*, 100173. [CrossRef]
24. Shen, T.; Hong, Y.; Thompson, M.M.; Liu, J.; Huo, X.; Wu, L. How does parking availability interplay with the land use and affect traffic congestion in urban areas? The case study of Xi'an, China. *Sustain. Cities Soc.* **2020**, *57*, 102126. [CrossRef]
25. Schoeman, C.B.; Schoeman, I.M. Land use, traffic generation and emissions in formulating a simplified approach in assessing development impacts in residential areas. *Int. J. Transp. Dev. Integr.* **2019**, *3*, 166–178. [CrossRef]
26. Yap, J.Y.L.; Omar, N.; Ismail, I. A Study of Traffic Congestion Influenced by the Pattern of Land Use. *IOP Conf. Ser. Earth Environ. Sci.* **2022**, *1022*, 012035. [CrossRef]
27. Rahman, M.M.; Najaf, P.; Fields, M.G.; Thill, J.-C. Traffic congestion and its urban scale factors: Empirical evidence from American urban areas. *Int. J. Sustain. Transp.* **2021**, *16*, 406–421. [CrossRef]
28. Open Street Map (OSM). Available online: https://www.openstreetmap.org/ (accessed on 5 June 2023).
29. Autonavi Development Platforms. Available online: https://lbs.amap.com/ (accessed on 5 June 2023).