


## Article

# A Weight Assignment Algorithm for Incomplete Traffic Information Road Based on Fuzzy Random Forest Method

Longhao Wang <sup>1</sup>, Jing Wu <sup>1</sup>, Rui Li <sup>1</sup>, Yanjiao Song <sup>1</sup>, Jiayue Zhou <sup>1</sup>, Xiaoping Rui <sup>2,\*</sup>  and Hanwei Xu <sup>1,\*</sup>

<sup>1</sup> School of Hydrology and Water Resources, Hohai University, Nanjing 211000, China; 1809040223@hhu.edu.cn (L.W.); 1809040211@hhu.edu.cn (J.W.); 1809040210@hhu.edu.cn (R.L.); 1809040229@hhu.edu.cn (Y.S.); 1809040209@hhu.edu.cn (J.Z.)

<sup>2</sup> School of Earth Sciences and Engineering, Hohai University, Nanjing 211000, China

\* Correspondence: ruixp@hhu.edu.cn (X.R.); xuhanwei@hhu.edu.cn (H.X.)

**Abstract:** One of the keys in time-dependent routing is determining the weight for each road network link based on symmetrical and complete traffic information. To facilitate travel planning considering traffic situations based on historical global position system (GPS) trajectory data which uncover the whole road network, this paper proposes a fuzzy random forest-based road section data estimation method, which uses the third law of geography as the core idea. For different time periods, road grade, tidal lane, proximity to infrastructure (main places that affect traffic, such as schools, hospitals), and accident road sections were selected as indicators that influence the traffic. The random forest algorithm is used to build the mapping relationship between attribute data with average traffic which is obtained based on GPS data. Subsequently, the fuzzy reasoning method is used to obtain the weight of road links missing traffic information by calculating their similarities with typical road section samples. Using the road network of Suzhou City as an example, the proposed method was used to analyze estimate the average driving speeds of road sections with missing traffic information for different time periods. Experimental results show that this method can effectively avoid congested road sections and obtain high-speed travel routes.

**Keywords:** road network; weight assignment; path planning algorithm; fuzzy random forest; impact factor



**Citation:** Wang, L.; Wu, J.; Li, R.; Song, Y.; Zhou, J.; Rui, X.; Xu, H. A Weight Assignment Algorithm for Incomplete Traffic Information Road Based on Fuzzy Random Forest Method. *Symmetry* **2021**, *13*, 1588. <https://doi.org/10.3390/sym13091588>

Academic Editor: Basil Papadopoulos

Received: 1 July 2021

Accepted: 23 August 2021

Published: 29 August 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Traffic congestion has become a major problem in cities as the number of vehicles on city roads has increased. Waiting in traffic wastes time and has an adverse economic impact. To minimize the costs incurred by waiting in traffic, people can plan their routes with the shortest travel time. Towards this end, the road weights containing traffic information can be used in a route optimization algorithm. The weight value determines the search basis of the common path algorithms and affects the rationality of the optimal path. However, the traffic data in many cities cannot cover the whole road network yet. In most cities, the traffic information is collected by sensors or Taxi GPS trajectory data, the sensors are usually installed at the main roads and there are still a lot of roads lacking traffic data records, Taxi GPS trajectory data also cannot cover the whole road. This means that in the case of data missing, the weight of the whole road links cannot be generated.

The lack of traffic data on some road sections has two primary types: the partial or complete lack of road traffic data. A partial lack of road traffic data means that the road section lacks traffic data for a given date-time range and it can be inferred from the data of other dates at that time, or from historical data. A complete lack of road traffic data means that the road section has no GPS data at all. This situation requires that the missing data be filled out by data from other roads in the road network.

In the face of a partial lack of road traffic data, the weight of road network can be determined by using real-time traffic data and historical traffic data. The two commonly

used methods of weight calculation are: using up-to-date traffic information to calculate the current weight, and using historical traffic information to predict the weight. For the first method, impedance function models are usually used to determine the time cost of each road link. Examples of such models are the Bureau of Public Roads (BPR) model [1,2], the Davidson model [3–7], and the multiclass user traffic equilibrium model [6,7]. Additionally, the traffic flow fluctuation theory can be applied to determine the travel time of the road section with missing traffic data by obtaining the travel time of the known road section and establishing a dynamic allocation recursive function [8–11].

For the second method, the basic idea is to ascertain the traffic rules based on basic historical information and to use current information to predict real-time weights. Along this line of reasoning, Zhu et al. proposed a two-stage routing algorithm and weight propagation model to predict the cost of an arbitrary path on a road network and accurately detect the traffic environment [12]. Tang proposed a road network path-planning algorithm based on experience levels. By analyzing historical floating car data (FCD) to determine the experience level of road sections, significantly faster path planning results can be obtained [13–15].

## 2. Related Works

To deal with the complete lack of traffic data for road sections, some researchers generate the weight of the entire road section from incomplete road section information. Yang's study was the first to completely annotate the weights of road networks using incomplete information. Yang used regression to model the problem and took the topology of the road network into account to construct the objective function. This method attempts to find a road section with the most similar and closest distance to the road section with missing traffic information from the road sections with complete traffic data, and sets the distribution and quantity of traffic flow and infrastructure as the influence index of the predicted vehicle speed, and uses the linear method to determine road weight. Yang's approach is a good reference for the calculation of the road links' weights with incomplete information [16]. Similarly, Zambrano-Martinez applied logistic regression and proposed an equation, being able to significantly improve the curve fitting results [17]. In fact, the vehicle speed is not only impacted by traffic flow but also impacted by other factors which are not linear, it cannot reflect these complex impacts just uses a regression model to describe it.

On this basis, more complex models are proposed. Ashok presented a model for the estimation and prediction of time-dependent origin–destination (O-D) matrices, it estimated the dynamic mapping between the time-dependent O-D flow and the link volume [18]. The O-D model can be divided into two categories: direct estimation and model estimation [19]. The direct estimation method uses the expanded sampling survey, which costs a lot of time. The model estimation method uses the traffic flow to establish the travel demand model and get the predicted O-D flow. Nguyen, T.V. et al. proposed the open source DFROUTER tool. It could obtain the O-D matrix that represents the internal trips inside the target area, and it resembled the real traffic distribution in the city [20]. The O-D model is easy to be applied in small networks, but a large number of road traffic data need to be input to improve the accuracy [21]. Shang et al. constructing a space–time–state hyper network by using a triple of flow, density and speed, it estimates the level of traffic congestion and provides decisions for traffic planning [22]. The influence of various activity attributes on the spatial and temporal distribution of road traffic flow is considered. However, when the network scale is large, it is necessary to enumerate all the possible states of nodes, so the calculation cost is high. Charbonneau proposed a heuristic algorithm, the shortest path heuristic algorithm (SPT). SPT uses the minimum path heuristic to create a Steiner tree to destination, it essentially considers the closest nodes instead of all possible node. On the basis of a current solution, it optimizes step by step through neighborhood search [23]. SPT method is not easy to fall into local optimum, but the algorithm is very complex, many parameters are considered, and the efficiency is not high [24]. The above

methods are only suitable for small traffic networks, and need to be improved when there are more nodes. Zambrano-Martinez et al. proposed a route server capable of handling all the traffic in a city, and balancing traffic flows by accounting for present and future traffic congestion conditions. With the increasing popularity of autonomous vehicles, congestion problems might be more common shortly. The focus on autonomous vehicles also allowed for more predictive behavior on the road. It showed that their model improved travel times and average travel speed in Valencia, Spain by 5%. In more congested areas, this improvement was about 8% [25]. Quiroga proposed a localization system to estimate traffic conditions [26], using the precise information provided by GPS, the real-time traffic information of each section of the city is obtained through mathematical modeling, and the current traffic flow or congestion situation is judged. It analyzed the sampled GPS data by segment lengths, sampling rates, and central tendency [27]. Then the average speed, traffic flow and other information of the path are calculated according to the GPS data and mathematical model, and then the traffic situation is estimated. This method is difficult to acquire complete and symmetrical GPS data through automatic acquisition, however, if the manual acquisition method is used, it will require a lot of work.

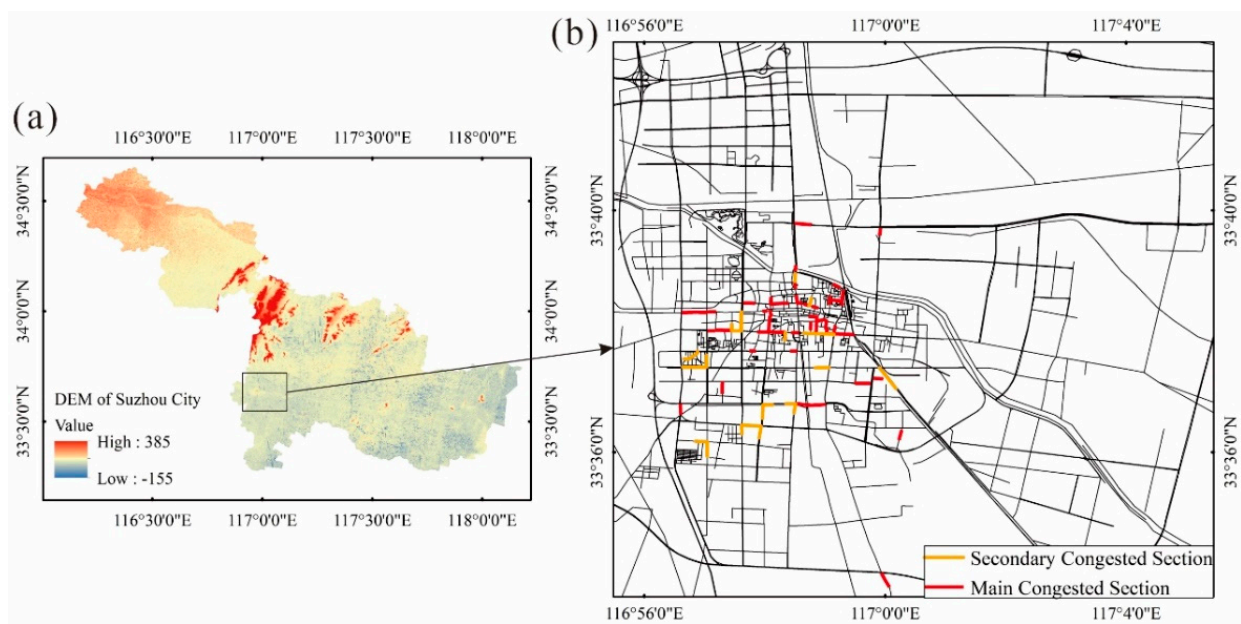
The above analysis showed that, given the complete lack of traffic data on road sections, traditional methods were suitable for small networks and were complicated. Although the new methods based on machine learning and metaheuristic techniques have improved, they could predict long-term traffic speed [28]. The method based on machine learning is fast to calculate, and it does not need a lot of input variables. However, the single machine learning method was difficult to accurately generate the impedance for incomplete information road links. These problems could be avoided. It was necessary to find the main factors affecting traffic and the similarity with different typical road sections. These were rarely considered in the existing methods. The key point was finding the data which were related to these traffic impact factors. A representative number of typical road sections could be extracted from road links as a template through expert knowledge, and we called this template a typical link pattern (TLP). These TLPs had complete information, which meant that both average speeds and the factors impacting on traffic of these road links were known. If we set various factors values to TLPs and calculated the similarity between incomplete information road link and the TLPs, the TLP which had the maximum similarity with this road link could be used to calculate weight further.

This paper differs from previous works by combining machine learning with fuzzy rules. Looking at actual situations where historical GPS data cannot cover the entire road network and there are no traffic data for some road sections [29], this paper proposes a road impedance estimation model based on fuzzy random forest, which considers the third law of geography as its guiding principle. The third law of geography says that the greater the similarities between geographical environments, the more similar the characteristics of the geographical target. Geographical similarity referred to the comprehensive similarity of two spatial locations in the geographic environment, and these two locations are not necessarily connected in space. Geographical features referred to the characteristics of the geographical variables concerned [30]. In other words, if the factors affecting traffic at specific times are similar, vehicle speeds should also be similar. In our study, the road level, driving speed, whether there are important traffic units around, whether the road is a tidal lane, whether it is an accident section, and road sampling time are all factors that affect traffic. Firstly, the random forest is used to establish the complete information road section, that is, the mapping relationship between the traffic factors and the incomplete information road section is established, and then the incomplete traffic information road section is input. Thereby, the impedance value of both complete information and incomplete road sections can be estimated.

### 3. Study Area and Data

The area this paper chose to study is Suzhou City, Anhui Province, China (Figure 1a). Suzhou City is located north of Anhui Province and is a municipal district of Yongqiao

district (Figure 1b). It is located between  $116^{\circ}09' E$ – $118^{\circ}10' E$  and  $33^{\circ}18' N$ – $34^{\circ}38' N$ . Yongqiao District is located at the junction of the provinces of Jiangsu, Henan, and Anhui. It is very convenient to get around in the city. The transportation network in Suzhou City has obvious temporal and spatial characteristics. Traffic congestion is pronounced during the morning and evening rush hours. The figure shows the road sections prone to congestion. The terrain of the study area is flat, with small undulation, various road types, and high-frequency GPS data (1–3 s) can be obtained, so it is an ideal area for the study.



**Figure 1.** Map of the road network of Suzhou City. (a) shows the terrain of Suzhou. (b) shows the congestion in Yongqiao district.

In the study, we used taxi (114 vehicles) GPS data from Suzhou Traffic Management Bureau (<http://jtj.ahsz.gov.cn/>, accessed on 22 December 2019), which has a high frequency (1–3 s). We used taxi GPS data from January to March 2014, which covered 24 h a day, including time, longitude, latitude, speed and elevation. Among them, the speed is calculated by the distance and time between two points of GPS. Table 1 shows the GPS data, as imported into ArcGIS, and with calculations of the average speed of GPS points on the road in a specific period of time, that is, the average speed of this section in that period of time. “DATA 140114” represents 14 January 2014, “TIME 11622” represents 1:16:22 of the day.

**Table 1.** High frequency GPS data and fields.

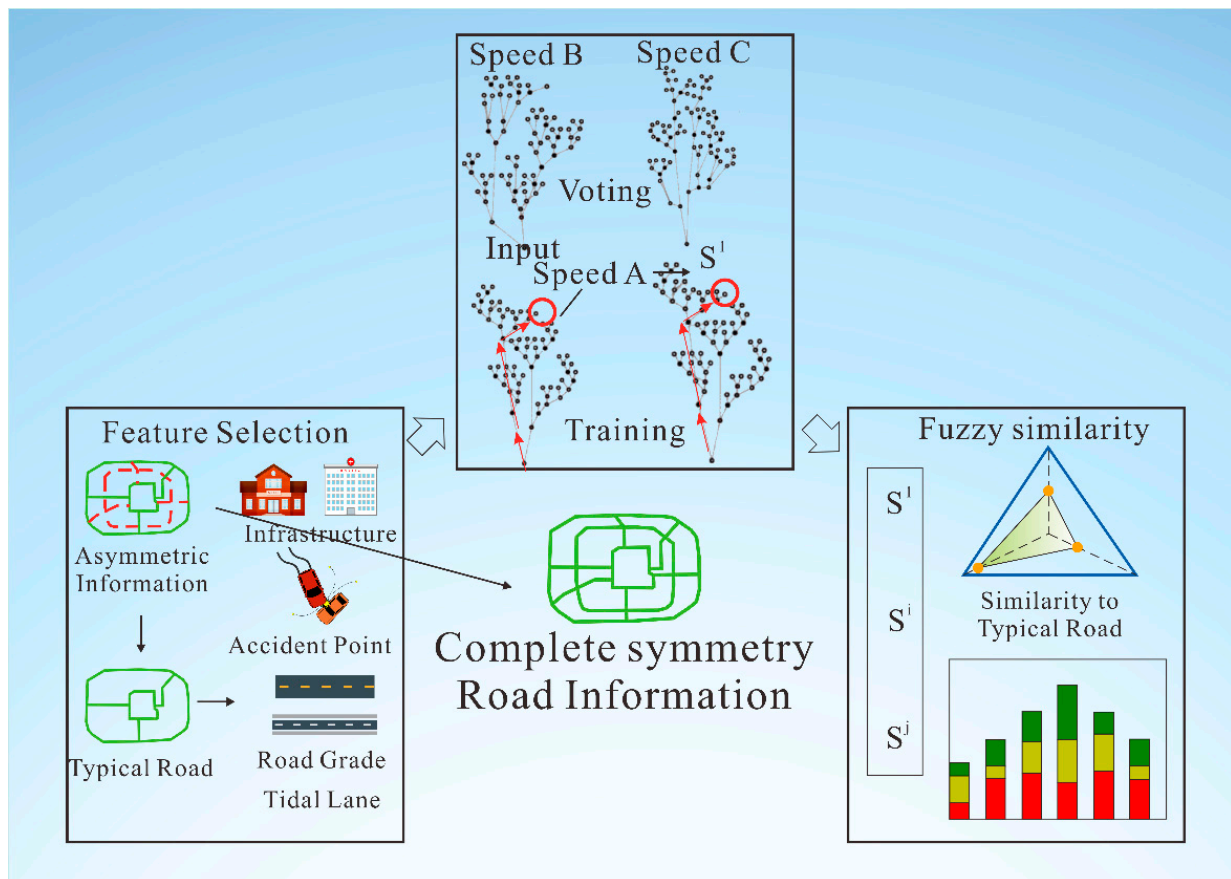
INDEX	DATA	TIME	LATITUDE	LONGITUDE	HEIGHT	SPEED
1	140115	11622	33.619021N	116.982741E	9	0
2	140115	11625	33.618906N	116.982789E	9	3
...	...	...	...	...	...	...

#### 4. Methodology

Due to the lack and asymmetry of road information, this paper brings out a new model to estimate the impedance of paths based on a random forest method. This model is based on the third law of geography, and it assumes that if the factors impact on the traffic are similar at a specified time, the vehicles speed is also similar, so the average speed of incomplete information road links can be evaluated by the full information road links through the fuzzy random forest and the average speed is the key for time-dependent routing calculating. The basic idea of this method is to first analyze the traffic



influencing factors, extract typical road sections as training data, calculate the average speed of incomplete traffic information road sections through the fuzzy random forest, and then estimate the impedance value. Finally, the road network with complete and symmetrical information is generated, as shown in the conceptual illustration of fuzzy random forest (Figure 2).



**Figure 2.** The conceptual illustration of fuzzy random forest.

The traffic condition of urban roads is produced by the combination of internal factors such as road grade and number of lanes and external factors. In the absence of traffic data, the road's speed limit is typically used as the road's weight. Since the road's speed limit is usually faster than actual traffic speeds and these road sections are used in route planning, actual travel times tend to be longer than predicted ones. Therefore, the weights of road networks used in route planning must conform to the actual traffic conditions. As an intermediate parameter, the average driving speed establishes the relationship between distance and time in path planning. In this paper, the vehicle's speed is directly related to the vehicle's travel time and many other environmental factors affecting the road network. In other words, travel time is impacted by increases or decreases in speed. Environmental factors include road locality (e.g., a vehicle's speed should be reduced when traveling on roads near schools and hospitals) and traffic conditions (e.g., average vehicles' speeds will be very low on extremely congested road sections, like those found during the morning rush hours). In path optimization, roads with poor road conditions should have increased road impedance so that path planning is less likely to use that road. Therefore, setting the weight index to the speed of the vehicle reflects the time cost of driving on that road and establishes a connection with the distance or time weight. Establishing that connection is convenient for comparing path planning results. The basic idea of the fuzzy random forest method is to analyze traffic influencing factors, extract typical road sections, estimate the

average speed of the road section, calculate the fuzzy similarity between average speeds and a typical road section that has complete information, and estimate the impedance value.

#### 4.1. Road Network Weight Analysis

Many of Suzhou City's roads have different features. There are two main factors affecting road traffic: the characteristics of the road itself (e.g., road grade, road length, road gradient, the number of lanes, etc.) and external factors (e.g., traffic volume, tidal effect, road locality, the number of traffic lights, traffic accidents, weather, etc.).

##### (1) Road grade

For the characteristics of the road itself, the road grade determines the material, flatness, design capacity, road width and other characteristics of the road. This paper's expert knowledge of the characteristics of urban roads comes from the implementation of the "urban road engineering design specifications". Table 2 shows the speed limits for different road grades, and each grade includes three roads with different number of lanes, which have different speed limits. Therefore, when road information does not cover a road section, the road grade plays a decisive role in determining the average speed of that road section. Therefore, it is an important factor in estimating the impedance of road sections with incomplete traffic information. In the case where traffic information is unavailable, this study selects the vehicle speed observed during off-peak hours as the maximum road limit speed. The impact of weather on traffic is related to the city's DEM (digital terrain model). Certain weather conditions (e.g., heavy rain and snowstorms) significantly impact average speeds. This paper does not consider the impact of road slopes because of the flatness of Suzhou City's terrain.

**Table 2.** Speed limits for China's roads.

Grade	Primary Road	Secondary Road	Third Level Road	Residential Road
Speed limit (km/h)	80	60	40	20
	70	50	30	10
	60	40	20	5

##### (2) Tidal lane

Tidal effect means that the average speed of roads of the same grade is completely different in different time periods. The tidal effect is an important external factor that affects average vehicle speeds. Additionally, average vehicle speeds on the same road grade vary over time. We analyzed the spatial-temporal characteristics of traffic using Suzhou City's historical traffic data. The traffic conditions throughout the day can be separated into four time periods. These periods are as follows: morning peak hour (7:00–9:00), evening peak hour (17:00–19:30), daytime hours (9:00–17:00) and night time (20:00–7:00 of the next day). Traffic flow is the main external factor that affects the speed of traffic, and there is a clear correlation between the two. Because it is difficult to obtain traffic flow data covering the entire road network, this paper used the average driving speeds for each road section during different times, as calculated from GPS historical data. The tidal effect has an evident regularity. This attribute can be directly assigned to the tidal road section, and the traffic condition of the road section can be determined according to the morning and evening peak times.

##### (3) Infrastructure and Accident point

Road section locality (whether near schools, hospitals, etc.) and whether the road section is an accident section are also main factors affecting traffic. Based on the third law of geography, we assumed that when the conditions of adjacent traffic units are similar at a given moment, their vehicle speeds are similar too. However, this effect is not linear. We analyzed traffic data from roads that are close to schools and hospitals, and accidental sections at various distances (less than 30 m, between 30–60 m, 60–100 m, and greater than

100 m). The results show that there is a significant difference between average road speeds of adjacent infrastructure (i.e., roads close to schools, hospitals, etc.)

The impact of weather on traffic is usually related to the DEM of a city. For example, heavy rain will have a significant impact on traffic in low-lying sections of the city, while heavy snow will have a greater impact on the traffic on slopes (e.g., bridge sections). Because of the flat terrain of this study's research area, the impact of weather on road traffic in different sections of the city is neutralized. This is reflected in the average road speeds.

#### 4.2. Road Network Weight Assignment Method Based on Fuzzy Random Forest

The amount of road network data is large, but the data coverage for some sections of the road network is lacking. It is difficult to calculate the average speed of all roads. Since travel time is affected by environmental factors, it is also difficult to directly construct a functional relationship between the impact factor and driving speeds. Therefore, we calculated the average road speeds for typical road sections in the road network using taxi GPS data, correlated with the environmental factors affecting a typical road section. This process is a typical road section sampling process. For road sections with missing data, the random forest classification method was used to obtain the average speed. In this way, the basic idea of road network weight assignment can be expressed as follows: take speed as the intermediate parameter of the road network weight, construct a mapping model between road network speeds and road environment factors, and collect data on the driving speed of typical road sections in the road network sampling using a random forest algorithm for classification mapping model training. The driving speed of other unsampled road sections can be determined according to the value of the environmental factor of the road section and the environmental factor of a typical road section. To assign a driving speed to an unsampled road section more accurately, this paper used fuzzy inference based on the random forest classification model. In this method, the similarity of environmental factors is used as the fuzzy membership degree to calculate the driving speed of the unsampled road section. Once the driving speed of the entire road network has been obtained, and the road weight matrix is constructed using that speed, and optimal path analysis is performed. The specific implementation steps are (Figure 3):

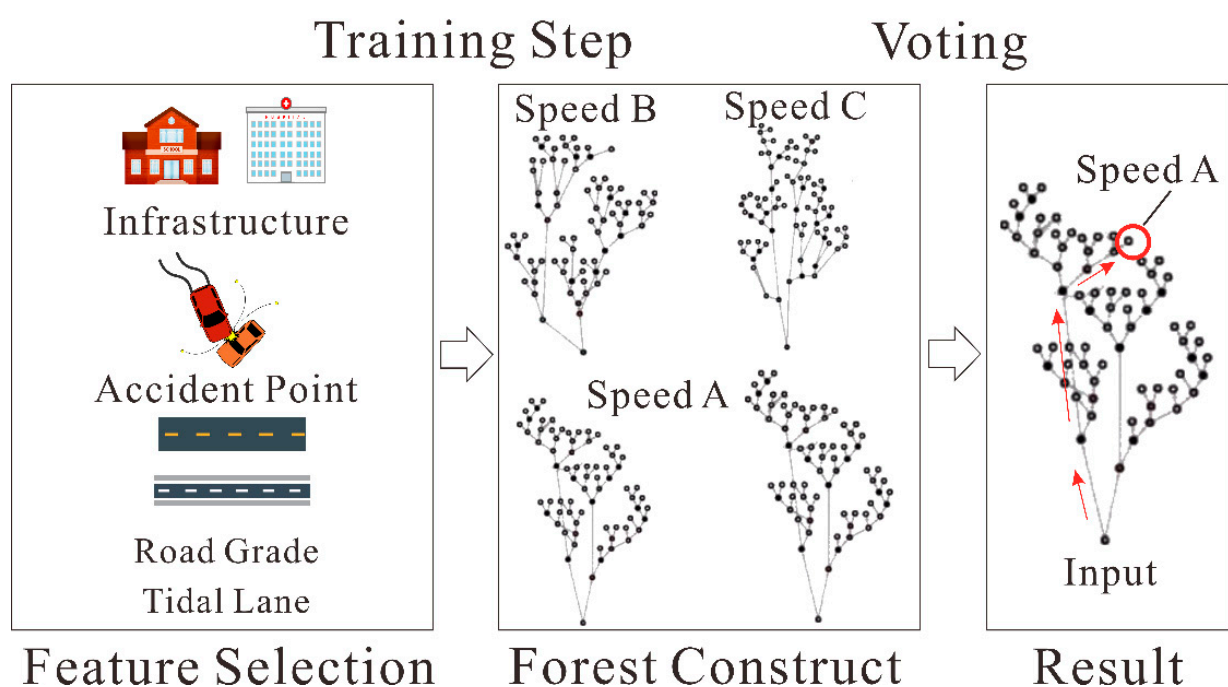


Figure 3. The concept of the random forest model.

### (1) Classified sampling

First, the characteristics of the road (e.g., road grade) and the external factors affecting traffic conditions (e.g., tidal lane, distance to traffic units, and accident sections) were selected according to the actual road conditions in the study area and the influencing factors of lane speeds. Second, with those characteristics, taxi GPS data were obtained to calculate the average vehicle speed of the corresponding road. Third, select typical road sections as samples to obtain the road speed value and its corresponding impact factors.

### (2) Random forest classification

The road grade, tidal lane, distance to transportation units, and accident road sections were the independent variables of the training samples. Speed was the dependent variable. The random forest method was used to classify the training samples.

As a supervised learning algorithm, the random forest method selects training data and constructs a classifier in a random and replacement manner, and combines multiple models to form the final overall model. In contrast to the optimal threshold classification approach used by typical decision trees, the random forest classification process uses decision trees to classify data in a random threshold manner. This classification method can produce stronger diversity, improving the model's results. The model construction and sample prediction process of the random forest method are as follows.

- a. From the original training dataset, randomly extract samples (approximately 30% of the dataset) to be the training set for a decision tree. This is repeated  $N$  times to train a total of  $M$  decision trees for random forest classification.
- b. Put the extracted samples back into the original training dataset and randomly extract the data samples again for the next level of the decision tree.
- c. Repeat Step B to train enough decision trees to form a random forest model. In theory, the greater the number of decision trees, the better the random forest model.
- d. Run the prediction dataset through the model and calculate the weight assignment result from the statistics of the multi-layer decision tree results for each group of prediction data.

Data for unsampled road sections were derived according to the prediction results of the model established by the random forest method. The concept of the random forest model is shown in Figure 3. In this model, firstly, the factors that affect the traffic are taken as the training data input, and the random forest classifier is constructed. The result is assigned to class  $A$  through voting output, where Class  $A$  represents the driving speed of the segment.

### (3) Fuzzy reasoning

The fuzzy reasoning method uses the combined weight of multiple similar typical road section samples as the weight of the unsampled road section. All samples' information is used in the fuzzy reasoning method to evaluate the traffics for information missing road sections and can avoid them being simply classified as a single class by the random forest method. The steps for fuzzy reasoning are as follows: after the unsampled road section is predicted by the random forest method (i.e., the classification result), select the typical road section sample where the difference between the dependent variable of the typical road section sample and the classification result is within a certain threshold. Then, calculate the unsampled road section and the selected road section based on the fuzzy inference method. The similarity degree for the samples of typical road sections is combined with the fuzzy membership degree of the index to obtain the road weight of the unsampled road section. The fuzzy membership degree of the index measures the similarity degree and the speed weight. Calculate the similarity between the  $h$ -th unsampled road section and the environmental factors for each typical road section with the fuzzy inference method (see Equation (1)).

$$S_{i,h} = \frac{1}{m} \sum_{j=1}^m e^{\left(\frac{|p_{jh} - p_{jt,h}|}{w_1}\right)^2} I_n(k_1) \quad (1)$$



$S_{i,h}$  represents the similarity between the  $h$ -th unsampled section and the  $i$ -th typical road section.  $p_{j,h}$  represents the  $j$ -th road attribute value of the  $h$ -th unsampled section.  $p_{j,i}$  represents the  $j$ -th road attribute of each typical road section corresponding to the  $h$ -th unsampled road section,  $k_1 \cdot w_1$  represents the correction parameters.  $m$  is the number of influencing factors.

Based on the similarity between the  $h$ -th unsampled road section and each typical road section, the weight corresponding to the  $h$ -th unsampled road section is calculated using Equation (2).

$$W_h = \frac{\sum_{i=1}^n S_{i,j} W_{i,h}}{\sum_{i=1}^n S_{i,j}} \tag{2}$$

$W_h$  represents the speed weight corresponding to the  $h$ -th unsampled road section, and  $W_{i,h}$  represents the speed weight of the  $i$ -th typical road section corresponding to the  $h$ -th unsampled road section. The specific steps are shown in Figure 4.

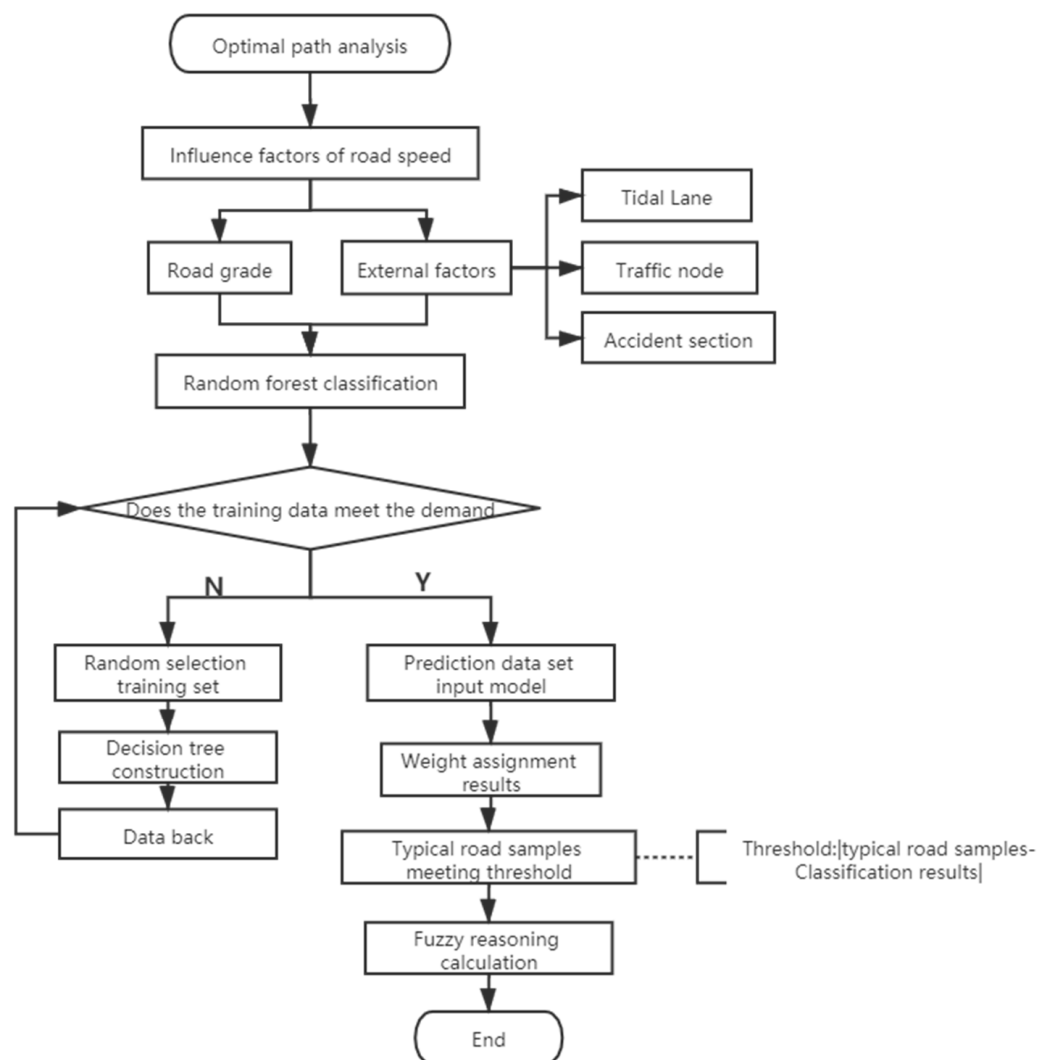


Figure 4. Road network weight assignment technology roadmap.

#### 4.3. Shortest Path Impedance Setting

The impedance setting function relationship is based on the principle that as vehicle speed increases, impedance decreases. In other words, under the effect of the influence factor (e.g., tidal lane, distance to traffic units, and accident sections), the longer a trip takes, the lower the average speed and the greater the road impedance. Setting a suitable impedance can effectively avoid congested road sections in path optimization and achieve the purpose of increasing the pass rate. The mathematical basis of the impedance setting is the ratio of the road weight to road length (Equation (3)). The speed weight, under real-world conditions, is affected by changes in the length weight. Such changes also affect the results of the shortest path planning algorithm, thereby weight changes affect the impedance, and the impedance affects the result of the shortest path planning algorithm. Path planning is calculated as follows:

$$W = \frac{speed\_allowed}{speed\_inferred/actual} \cdot road\_length \quad (3)$$

$W$  is the road impedance used for path planning.  $road\_length$  is the length of the road.  $speed\_allowed$  is the different speed limits of the roads.  $speed\_inferred/actual$  is the speed inferred or the actual speed, if the information of the road section is not complete,  $speed\_inferred$  will be used in here which is inferred from Equation (2), if the road network information is complete,  $speed\_actual$  will be used here and it is the actual speed. When the road conditions are good,  $w$  is the same as  $speed\_allowed$ . When the road conditions are poor,  $w$  is smaller than  $speed\_allowed$  and the impedance of the road section with poor conditions,  $W$  will increase.

#### 4.4. Network Analyst

After obtaining the full path impedance, a geometric network and a logical network are established based on the network analysis module of ArcGIS software. The geometric network comprises a series of topology-related feature classes and is a graphical representation of geographic features. Obtain the urban road network data from OpenStreetMap. The urban road geometric network model comprises two types of feature classes: the centerline of the road and the road junction. It can be established by the Create Geometric Network Tool in the ArcGIS desktop software, ArcCatalog. After checking the topology and establishing the network data set, enter the total road network impedance calculated in the previous step into the attribute database. Open the ArcToolbox of ArcGIS, select "Network Analyst Tools-Make OD Cost Matrix Layer," enter the network data set just created, successfully build the OD matrix, and then load the starting point and ending point position, click the solution on the network analysis toolbar, and the calculation ends.

### 5. Experiments and Results

Three experiments were conducted for various routes in the study area during the times of, morning peak, daytime, evening peak and night. The routes had different starting and ending points. A total of 392 samples were selected for typical road sections. Sample selections considered different road characteristics. Table 3 shows the sample selection process.

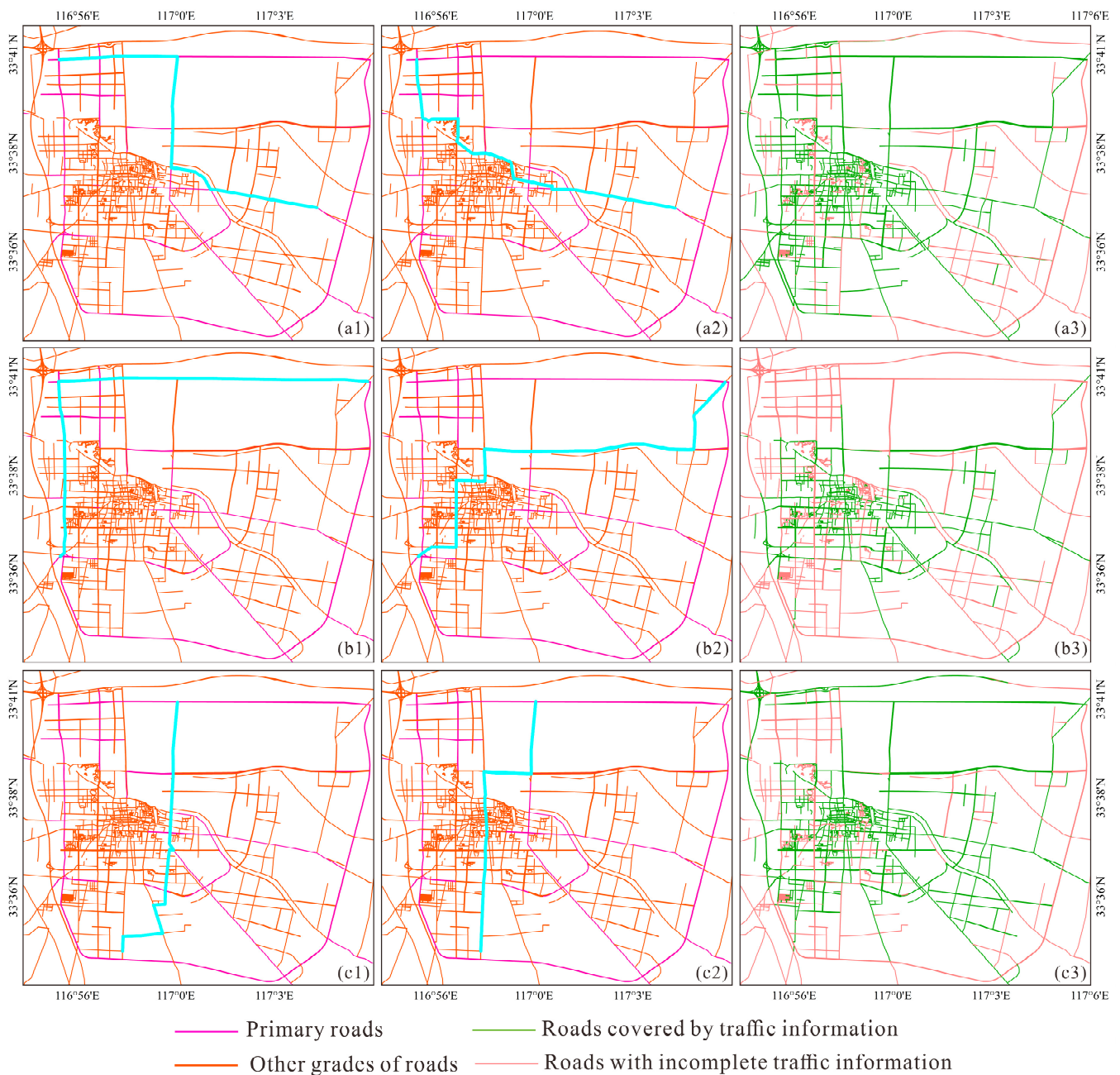
**Table 3.** Sample selection categories and data statistics.

Road Grade	Time Period	Distance to Traffic Unit	Tidal Lane	Distance from the Accident-Prone Section	Number of Samples
Trunk	Morning peak	<30 m	Yes	<30 m	5
	Daytime	30–60 m	Yes	>100 m	6
	Evening peak	>100 m	Yes	>100 m	4
	Night	>100 m	No	>100 m	3
	Morning peak	>100 m	No	>100 m	4
Primary	Daytime	>100 m	No	>100 m	4
	Evening peak	<30 m	No	>100 m	8
	Night	60–100 m	No	>100 m	6
Secondary	Morning peak	30–60 m	No	60–100 m	21
	Daytime	>100 m	No	>100 m	38
	Evening peak	>100 m	No	>100 m	12
	Night	>100 m	No	>100 m	22
Third level	Morning peak	<30 m	Yes	<30 m	3
	Daytime	>100 m	No	>100 m	11
	Evening peak	>100 m	No	>100 m	4
	Night	<30 m	No	>100 m	5
Residential	Morning peak	60–90 m	No	>100 m	2
	Daytime	<30 m	No	>100 m	6
	Evening peak	30–60 m	No	30–60 m	3
	Night	30–60 m	No	>100 m	4

A total of two tests were carried out for the whole area of the study area. For the first time, a total of three groups were tested to study the practical application of the fuzzy random forest algorithm when the road information is not covered. The selected time periods for the three groups of tests are morning peak (Figure 5a), daytime (Figure 5b) and night (Figure 5c), and different starting and ending points are selected. When there was missing traffic information for a given road section, we set the impedance to the speed limit allowed by the road class. (a2, b2, c2) and the whole road network data inferred by the fuzzy random forest as the impedance (a1, b1, c1) test, respectively; at the same time, statistics on the coverage and lack of traffic information in this period (a3, b3, c3); the path results are shown in Figure 5. Calculate the parameters of the first group and the results are compared in the following table. The shortest path in Table 4 refers to the shortest distance calculated with the road length as the impedance. Although the path is the shortest, due to the traffic congestion at that time, the average speed of the shortest path is lower and the travel time is longer than the path by our method.

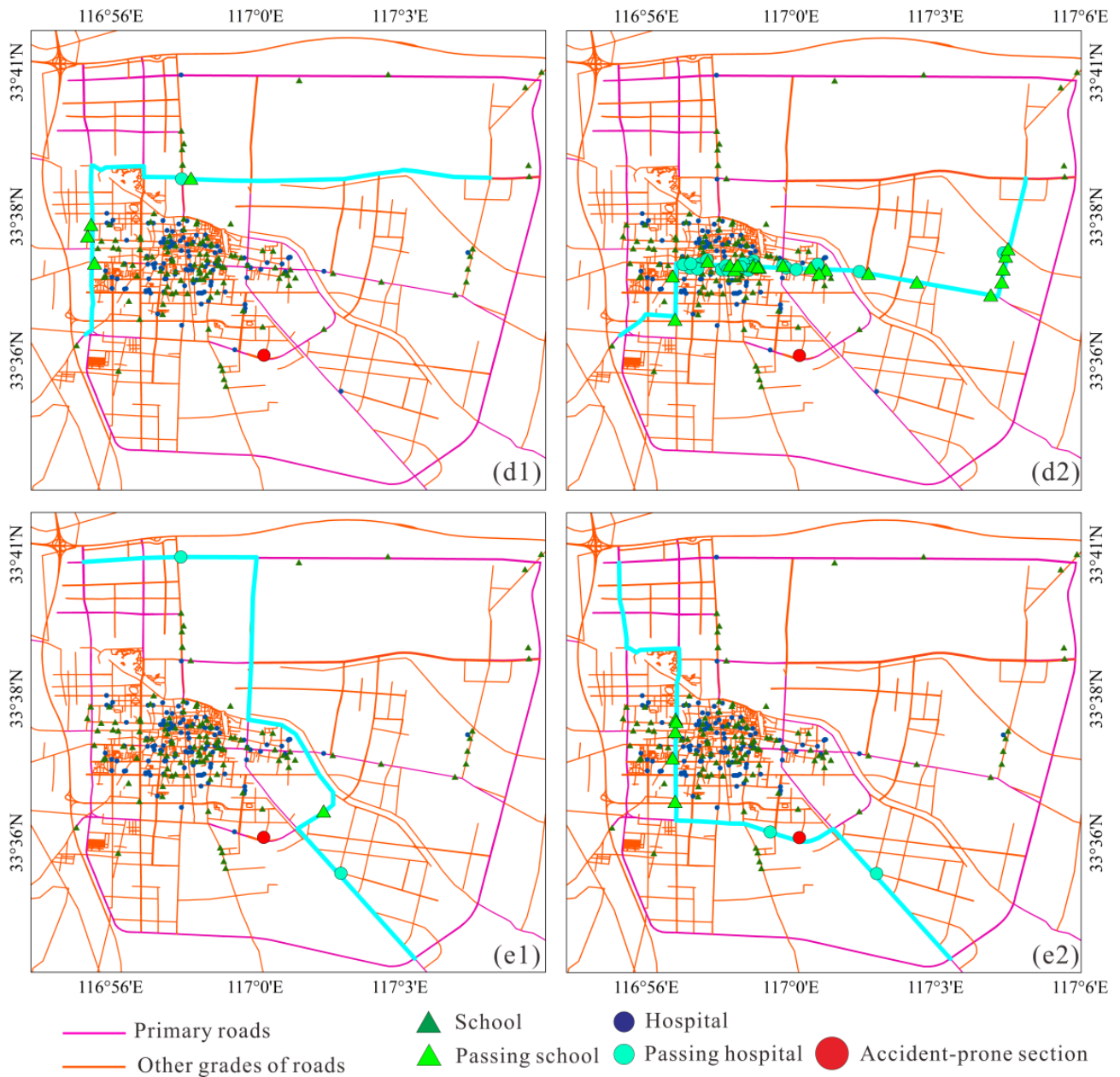
**Table 4.** Comparison of the statistics of the Test 1 path algorithm results.

Path	Path by our Method	Shortest Path
Path length (km)	19.35	18.52
Average speed (km/h)	35.50	32.14
Evaluation of travel time (min)	32.69	34.57
Time period	Morning peak	Morning peak
Primary	3	2
Secondary	4	5
Third level	0	0



**Figure 5.** Results of the first test route and lack of traffic information. (a–c) were sampled from morning peak, daytime and night respectively; (a1–c1) are static paths, which are calculated by the maximum speed allowed by the road; (a2–c2) are fuzzy paths, which are calculated by fuzzy weight; (c1–c3) are the loss of traffic information in different periods.

In the second test, two groups were tested. The selected time periods were morning peak (Figure 6d) and evening peak (Figure 6e), and different starting points and ending points were selected. Use the speed limit allowed by the road class as the impedance (d2, e2) and the whole road network data inferred by the fuzzy random forest as the impedance (d1, e1) test. At the same time, count the schools, hospitals and accident road sections passed by during the driving process, and calculate. The first two groups of parameters are compared in the following Table 5:



**Figure 6.** The results of the second test path and the passing infrastructure. (d,e) Samples were taken at the morning peak and the evening peak respectively; d1 and e1 are static paths, which are calculated according to the maximum speed allowed by the road; (d2,e2) are fuzzy paths, which are calculated by fuzzy weights; the path obtained by fuzzy impedance avoids congestion prone road sections such as schools and hospitals.

**Table 5.** Comparison of the statistics of the Test 2 path algorithm results.

Method	Path by out Method	Shortest Path
Path length (km)	24.26	23.30
Average speed (km/h)	36.52	33.28
Evaluation of travel time (min)	39.85	42.00
Time period	Evening peak	Evening peak
School	1	5
Hospital	2	2
Accident section	0	1



According to the result of the regional route's planned length, the planned route passes through the static impedance and entered the city's core. Using fuzzy impedance, the path planning result avoided roads in the city. Although the total travel distance increased slightly, the ratio of fast-driving roads to slow-driving roads was higher. The estimated travel time was also shorter. This route is also more in line with actual driving habits, indicating that the impedance model based on fuzzy inference has practical application value.

Urban road networks do not have complete traffic information covering the whole of the network. Traditional methods set the impedances of road sections with missing traffic information to their speed limits. Practically, vehicles driving on a given road often do not hit the road's speed limit. Using a fuzzy random forest model to estimate traffic data for road sections with missing traffic information can give more accurate estimates of the average driving speed on these road sections for a given time period. Consequently, the road network can be considered to have full traffic information coverage. On this basis, travel route optimization decisions that better reflect the real road conditions during a given period can be made. In the test, the fuzzy random forest method avoided the nodes in the urban core area where the theoretical value of speed limit is very high, such as schools and hospitals, but are actually congested. The result of path optimization increases the average speed during road driving process at a cost of only approximately 5% of the increased distance traveled, which ultimately reduces the driving time, greatly improves driving comfort, and has evident advantages.

## 6. Discussion

Since it is difficult just using an equation to calculate the weights covering the whole road links directly, this paper brings out a new model to estimate the impedance of the path based on a fuzzy random forest. This study mainly focuses on generating the complete and symmetrical road network coverage data based on the existing GPS data, which are used for path planning considering time. The planning result is more reasonable during the traffic peak period than using the default impedance directly when there is no GPS data coverage. Analyzing and quantifying the factors affecting traffic is a crucial issue. Whole and detailed factors can improve the accuracy of routing. The factors chosen depend on expert knowledge. Factors affecting traffic in different regions may be different.

In further research, we will apply the method to other research areas. First of all, the characteristics of different regions need to be considered to select the appropriate impact factors. For example, in areas with large terrain differences, elevation change, and snow are important. Secondly, the same factors in different regions may have different actual effects, and the weight can be adjusted according to the actual situation of different countries. In the following research, we will apply the incomplete information network construction method to other areas with high-frequency GPS data, which needs to consider other factors.

## 7. Conclusions

Weights covering all road links are the basis of routing. Traditional methods have difficulty generating accurate impedances for road links with incomplete traffic information. Using historical GPS trajectory data, this paper proposes a new road network weight distribution method based on a fuzzy random forest model. This method introduces the third law of geography into the distribution of road weights. It estimates the average driving speeds for roads with missing traffic information by constructing a mapping relationship between the average road speed and the internal and external factors that affect road traffic. Experimental results showed that this method is effective. Analysis of the traffic impact factors, and their quantifications are key to this method. Complete and detailed knowledge of the factors that affect routing can improve routing accuracy. The relationships between such factors and traffic are also important. This paper introduces several factors that affect traffic.

Choosing the appropriate environmental factors for sample selection, and the subsequent training of a fuzzy random forest model using those samples, is important in order to obtain the driving speed of the entire road network. When the fuzzy random forest algorithm is used to interpolate the data for road networks with unknown driving information, in view of, the selection of influencing factors is not set in stone due to the complex factors that affect driving speed and its many aspects. Different cities have different traffic-affecting factors. The calculation of road weights depends on the attributes of the roads. Therefore, the influencing factors of representative roads in different cities can be analyzed and tested, and the results obtained can further enrich and perfect this method.

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

**Funding:** This paper was supported by the National Key Research and Development Plan of China (No. 2019YFC1804304), the National Natural Science Foundation of China (Grant No. 41771478) and the Fundamental Research Funds for Central Universities (Grant No. 2019B02514).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The study did not report any data.

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

## References

1. Ermagun, A.; Levinson, D.M. Development and application of the network weight matrix to predict traffic flow for congested and uncongested conditions. *Environ. Plan. B Urban Anal. City Sci.* **2019**, *46*, 1684–1705. [[CrossRef](#)]
2. Zeng, Y.; Zhang, N. Review and new insights of the traffic flow lattice model for road vehicle traffic flow. In Proceedings of the 2014 IEEE International Conference on Control Science and Systems Engineering, Yantai, China, 29–30 December 2014; pp. 100–103.
3. Davidson, K.B. A flow-travel time relationship for use in transportation planning. In Proceedings of the Australian Road Research Board Conference, Brisbane, Australia, 21–25 August 1978; Volume 3, pp. 183–194.
4. Yazdani, D.; Azizi, S.M.; Eren, S. Designing a decentralized LQ controller for a traffic network using DFP optimization technique. In Proceedings of the 2006 Canadian Conference on Electrical and Computer Engineering, Ottawa, ON, Canada, 7–10 May 2006; p. 1558.
5. Potuzak, T. Division of Road Traffic Network based on Genetic Algorithm and Graph Coarsening. In Proceedings of the 2018 11th International Conference on Human System Interaction (HSI), Gdansk, Poland, 4–6 July 2018; pp. 484–490.
6. Toint, P.; Wynter, L. Asymmetric multiclass traffic assignment: A coherent formulation. In *Transportation and Traffic Theory*; Lesort, J.B., Ed.; Pergamon Press: Oxford, UK; pp. 237–260.
7. Wang, J.; Peeta, S.; He, X. Multiclass traffic assignment model for mixed traffic flow of human-driven vehicles and connected and autonomous vehicles. *Transp. Res. Part B Meth.* **2019**, *126*, 139–168. [[CrossRef](#)]
8. Fu, M.; Xue, B. A path planning algorithm based on dynamic networks and restricted searching area. In Proceedings of the 2007 IEEE International Conference on Automation and Logistics, Jinan, China, 18–21 August 2007; pp. 1193–1197.
9. Hajbabaie, A.; Benekohal, R.F. A Program for Simultaneous Network Signal Timing Optimization and Traffic Assignment. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 2573–2586. [[CrossRef](#)]
10. Li, S. Dynamic link travel time model in dynamic traffic assignment. In Proceedings of the 2003 IEEE international conference on intelligent transportation systems, Shanghai, China, 12–15 October 2003; pp. 1036–1039.
11. Singh, P.; Sharma, A.K.; Ram, S. Routing and wavelength assignment in WDM networks with dynamic link weight assignment. *Optik* **2007**, *118*, 527–532. [[CrossRef](#)]
12. Zhu, L.; Li, Z.; Chen, L.; Cheng, Y. Two-stage evolutionary algorithm for dynamic multicast routing in mesh network. *J. Zhejiang Univ.-SCI. A* **2008**, *9*, 791–798. [[CrossRef](#)]
13. Lu, J.; Cao, L. Congestion evaluation from traffic flow information based on fuzzy logic. In Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems, Shanghai, China, 12–15 October 2003; pp. 50–53.
14. Ru, H.; Xu, J.; Duan, Z.; Liu, X.; Gao, C. Modelling Travel Time After Incidents on Freeway Segments in China. *IEEE Access* **2019**, *7*, 162465–162475. [[CrossRef](#)]
15. Tang, L.; Chang, X.; Li, Q. The Knowledge Modeling and Route Planning Based on Taxi' Experience. *Acta Geod. Et Cartogr. Sin.* **2010**, *39*, 404–409.

16. Yang, B.; Kaul, M.; Jensen, C.S. Using Incomplete Information for Complete Weight Annotation of Road Networks. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 1267–1279. [[CrossRef](#)]
17. Zambrano-Martinez, J.L.; Calafate, C.T.; Soler, D.; Cano, J.-C.; Manzoni, P. Modeling and Characterization of Traffic Flows in Urban Environments. *Sensors* **2018**, *18*, 2020. [[CrossRef](#)] [[PubMed](#)]
18. Ashok, K.; Ben-Akiva, M. Estimation and prediction of time-dependent Origin-Destination flows with a stochastic mapping to path flows and link flows. *Transp. Sci.* **2002**, *36*, 184–198. [[CrossRef](#)]
19. Cascetta, E.; Postorino, M. Fixed point approaches to the estimation of O/D matrices using traffic counts on congested networks. *Transp. Sci.* **2001**, *35*, 134–147. [[CrossRef](#)]
20. Nguyen, T.V.; Krajzewicz, D.; Fullerton, M.; Nicolay, E. DFR OUTER—Estimation of Vehicle Routes from Cross-Section Measurements. In *Modeling Mobility with Open Data*; Springer: Cham, Switzerland, 2015; pp. 3–23.
21. Moreira-Matias, L.; Gama, J.; Ferreira, M.; Mendes-Moreira, J.; Damas, L. Time-evolving O-D matrix estimation using high-speed GPS data streams. *Expert Syst. Appl.* **2016**, *44*, 275–288. [[CrossRef](#)]
22. Shang, P.; Li, R.; Guo, J.; Xian, K.; Zhou, X. Integrating Lagrangian and Eulerian observations for passenger flow state estimation in an urban rail transit network: A space-time-state hyper network-based assignment approach. *Transp. Res. B-Meth.* **2019**, *121*, 135–167. [[CrossRef](#)]
23. Charbonneau, N.; Vokkarane, V.M. Routing and Wavelength Assignment of Static Multicast Demands Over All-Optical Wavelength-Routed WDM Networks. *J. Opt. Commun. Netw.* **2010**, *2*, 442–455. [[CrossRef](#)]
24. Talarico, L.; Sörensen, K.; Springael, J. Metaheuristics for the risk-constrained cash-in-transit vehicle routing problem. *Eur. J. Oper. Res.* **2015**, *244*, 457–470. [[CrossRef](#)]
25. Zambrano-Martinez, J.L.; Calafate, C.T.; Soler, D.; Lemus-Zúñiga, L.-G.; Cano, J.-C.; Manzoni, P.; Gayraud, T. A Centralized Route-Management Solution for Autonomous Vehicles in Urban Areas. *Electronics* **2019**, *8*, 722. [[CrossRef](#)]
26. Quiroga, C.; Bullock, D. Travel time studies with global positioning and geographic information systems: An integrated methodology. *Transp. Res. C Emerg. Technol.* **1998**, *6*, 101–127. [[CrossRef](#)]
27. Taylor, M.; Woolley, J.; Zito, R. Integration of the global positioning system and geographical information systems for traffic congestion studies. *Transp. Res. C Emerg. Technol.* **2000**, *8*, 257–285. [[CrossRef](#)]
28. George, S.; Santra, A. An improved long short-term memory networks with Takagi-Sugeno fuzzy for traffic speed prediction considering abnormal traffic situation. *Comput. Intell.* **2020**, *36*, 964–993. [[CrossRef](#)]
29. Liu, C.; Wang, S.; Cuomo, S.; Mei, G. Data analysis and mining of traffic features based on taxi GPS trajectories: A case study in Beijing. *Concurr. Comp. Pract. Exp.* **2021**, *33*, e5332. [[CrossRef](#)]
30. Zhu, A.; Lu, G.; Liu, J.; Qin, C.; Zhou, C. Spatial prediction based on Third Law of Geography. *Ann. GIS* **2018**, *24*, 225–240. [[CrossRef](#)]