sustainability

MDPI

Article

Multiple Binary Classification Model of Trip Chain Based on the Fusion of Internet Location Data and Transport Data

Wenjing Wang [1], Yanyan Chen [1,*], Haodong Sun [1] and Yusen Chen [2]

[1] College of Metropolitan Transportation, Beijing University of Technology, Beijing 100124, China; wenjingwang@emails.bjut.edu.cn (W.W.); sunhaodong@emails.bjut.edu.cn (H.S.)
[2] Department of Transport and Planning, Delft University of Technology, P.O. Box 5048, 2600 GA Delft, The Netherlands; Yusen.Chen@tudelft.nl
* Correspondence: cdyan@bjut.edu.cn

Abstract: Observing and analyzing travel behavior is important, requiring understanding detailed individual trip chains. Existing studies on identifying travel modes have mainly used some travel features based on GPS and survey data from a small number of users. However, few studies have focused on evaluating the effectiveness of these models on large-scale location data. This paper proposes to use travel location data from an Internet company and travel data from transport department to identify travel modes. A multiple binary classification model based on data fusion is used to find out the relationship between travel mode and different features. Firstly, we enlisted volunteers to collect travel data and record their travel trip process using a custom-developed WeChat program. Secondly, we have developed three binary classification models to explain how different attributes can be used to model travel mode. Compared with one multi-classification model, the accuracy of our model improved significantly, with prediction accuracies of 0.839, 0.899, 0.742, 0.799, and 0.799 for walk, metro, bike, bus, and car, respectively. This suggests that the model could be applied not only in engineering practice to identify the trip chain from Internet location data but also in decision support for transportation planners.

Keywords: trip chain; travel segment; data fusion; travel mode; mobile phone trip survey; multiple binary classification model

Citation: Wang, W.; Chen, Y.; Sun, H.; Chen, Y. Multiple Binary Classification Model of Trip Chain Based on the Fusion of Internet Location Data and Transport Data. *Sustainability* **2021**, *13*, 12298. https://doi.org/10.3390/su132112298

Academic Editors: Giovanni Leonardi, Javier Alonso Ruiz, Angel Llamazares and Martin Lauer

Received: 17 September 2021
Accepted: 2 November 2021
Published: 8 November 2021

1. Introduction

Travel behavior is becoming increasingly diverse and complex, and it is important to analyze travel behavior through individual travel data. This not only helps transportation planners to understand traveler behavior and thus optimize transportation services but also helps business planners to provide more accurate services through passenger profiling. The complete profile of an individual can be represented by the trip chain [1], which should include the travel mode, activity, time, and location of the whole journey. Urban transport management departments have a large amount of transport data for many different modes [2] and can analyze bus and metro trips using smart card data, but they do not have access to the complete public transportation trip chain that includes walking. With the popularity of smart phones and the mass adoption of social software, it has become possible to collect large-scale Internet location data [3,4].

There has been some research in travel mode identification based on travel feature extraction [5–7]. Some literature reviews on travel mode recognition are shown in Table 1 [8,9]. The data sources have been gradually expanded from transport surveys to cell phone terminals [10]. The data categories mainly include GPS (global positioning system, GPS) data [11,12], GIS (geographic information system) and acceleration data, and travel survey data [13,14]. Different transport features such as average speed, maximum speed, and average acceleration were extracted based on these data [15], and algorithms such as random forest [6], support vector machine [16], and Bayesian were used to estimate travel

mode [17]. These models can come to identify different modes of transportation, such as transit walking, with model accuracies ranging from 75% to 97%. However, it is not possible to use these models directly for a continuous trip chain data consisting of different travel modes. Furthermore, few studies have focused on evaluating the effectiveness of these models on Internet location data. The travel mode research can be improved from algorithms or data sources. Multisource data fusion and perception of a full trip chain is still a challenge. Whether the travel features extracted and the travel analysis models built based on mobile Internet location data match with the real travel situation needs to be tested.

Table 1. Some literature review on travel mode recognition.

Literature	Model	Output (Classes of Travel Mode)	Input (Size)	Accuracy
Reddy, S.; Burke, J.; Estrin, D. 2008. [8]	Decision tree	Walk, run, bike, motor	4 h, six users	90%
Shin, D.; Aliaga, D. 2015. [10]	Uses the sensing values from acceleration and location	Walk, train, tram, car, bus	1 day, 30 users	82%
Rasmussen, T.K. 2015. [18]	A combined fuzzy logic and GIS (geographic information system)-based algorithm	Walk, bicycle, bus, car, rail	5 days, 101 users	75.7–97.1%
Semanjski, I.; Gautama, S.; Ahas, R.; Witlox, F. 2017. [9]	A support vectors machines-based model	Walk, bike, bus, car, train	4 months, 8303 users	94%

In this paper, we aim to fill this gap by building a trip chain model based on data fusion of Internet location data and transport data. The travel management department has bus GPS data, taxi GPS data, and smart card data, etc., which are marked with travel mode. Internet location data consist of a series of latitude and longitude points, which only have location information but no travel mode markers. To fuse tagged travel data from the transport management department and untagged location data from the Internet, we enlisted volunteers to collect travel data and record their travel trip process using a custom-developed WeChat program [18]. The data collected based on this program can be used to test the evaluation results of trip chain model. To improve the recognition accuracy of the model, we fused transport data from the transport management department and spatial data from the Internet. We build a multiple binary classification model based on data fusion [19,20] and used different sets of features for different travel modes; the model accuracy is significantly improved compared to the multi-classification model with only one feature set [21].

The paper is organized as follows. First, the methodology is described, which includes the design scheme of trip data collection and the regression model for travel mode and travel properties. Then, the survey data from volunteers in our study is further explained. Following that, we present the application of our model to those individual trip data. In the final section, we draw conclusions and suggest directions for future research.

2. Methodology

A travel chain, also known as activity-based travel [22], refers to activities completed in a continuous period of time to achieve a certain purpose. A trip is defined to start from an origin station near which the previous activity has been finished [23], and end at a destination station where the next activity will take place. An example is shown in Figure 1: the commuter first travels from the home to the workplace at 8:00 by walk and bus, stays at the workplace until 18:00, and then take a taxi home. There are two trip chains in this person's commute trip. Trip Chain 1 is from the home to the workplace, including two travel segments consisting of walking and bus. Travel Chain 2 is from the workplace to the home, only a travel segment [24] of taxi.

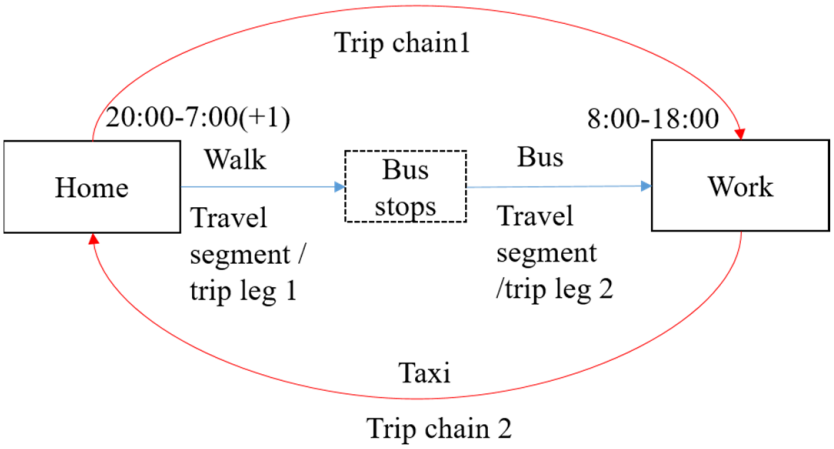


Figure 1. An example of an individual daily trip chain.

In order to analyze the travel modes in the travel chain, we will use a classification model. Binary classification is a form of classification, which is the process of predicting categorical variables, where the output is restricted to two classes [19]. We will use logistic regression, which is one of the many algorithms for performing binary classification. In transportation mode recognition, a specific binary classifier can receive trip features and predict the travel mode for that trip. For example, in the binary classification of travel mode as walking and other, we extract different travel features from the travel data as input data and then establish model to identify two types of trips: walk and other.

We assume that travel mode of a trip can be explained by travel features in this trip. In this study, we aim to test this assumption. Since not all single features are normally distributed and a non-linear relationship may exist between the independent and dependent variables [25], we take the logarithm of the variables to build the regression model if necessary. The model is presented as follows "(Equation (1))."

$$\log(y_j) = \beta_0 + \beta_1 \log(x_1) + \ldots + \beta_p \log(x_p) + \varepsilon \quad (1)$$

where y_j is the travel mode of trip j, ε represents the error term, and x_p are the different explanatory variables that represent trip properties.

As summarized in Figure 2, the methodology is divided into three parts. The first part is data collection. The aim of this paper is to build different models to identify the travel modes of Internet location data. In order to verify the accuracy of these models, real travel record data is needed. Therefore, we enlisted volunteers to collect travel data and record their travel process using a custom-developed WeChat program. Due to the unbalanced number of trip chains of different travel modes and the sample size being small, we fused the volunteer travel data with the public transportation data from the transport management department to assist modeling. Therefore, we collected three types of data, travel data from the volunteers, travel data from the transport management department, and sample location data from the Internet.

The second part is data fusion and data modeling. Because different data can extract different travel characteristics, and different modes of transportation have different travel characteristics, we established three models. Firstly, walking is identified, and then the trip chain is divided into travel segments. Secondly, the metro card data, bicycle travel data, and volunteer travel data are integrated to identify the metro and bicycle. Finally, the bus trajectory data, bus stops data, and volunteer travel data are fused to identify the bus and car. The structure of multiple binary classification models are shown in Figure 3. For each binary classification model, we use the logistic regression in Equation (1) to calculate.

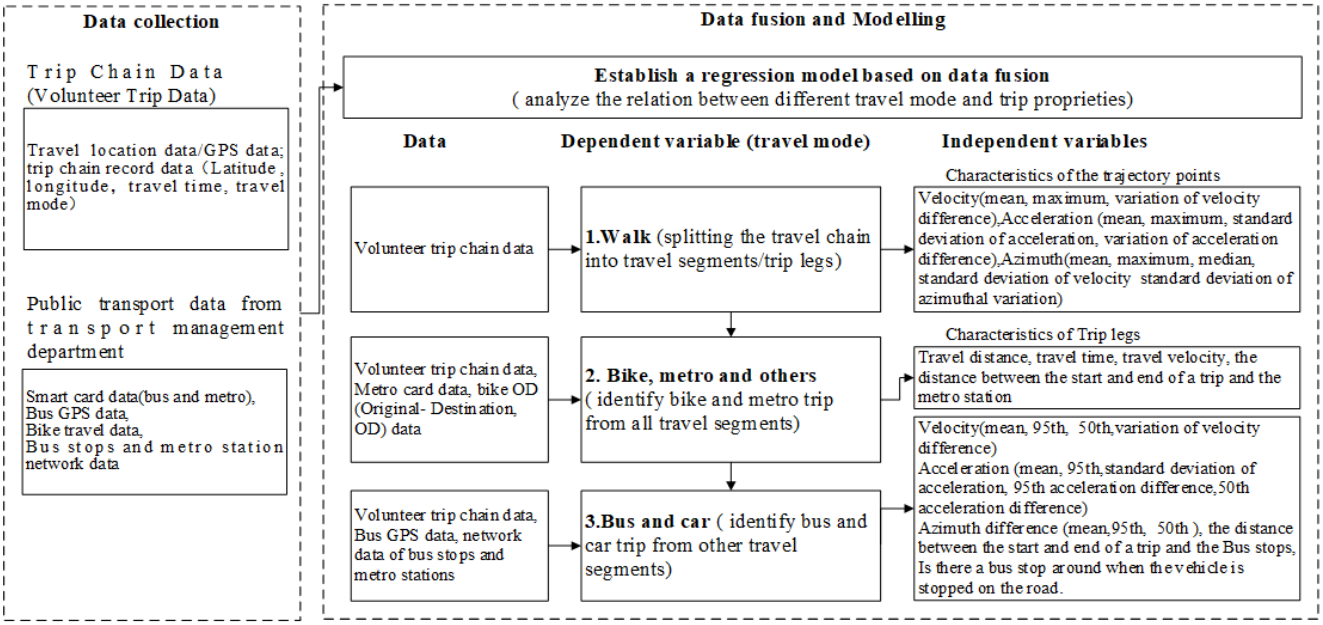


Figure 2. Main components of the developed methodology and overall research design.

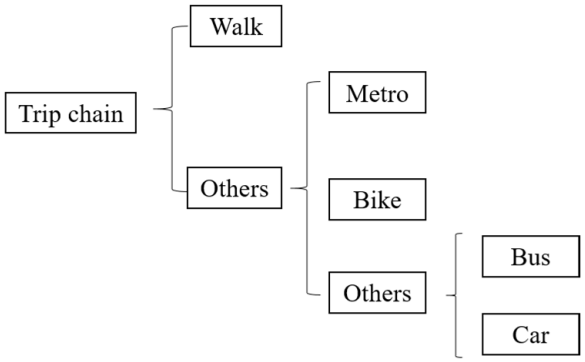


Figure 3. The structure of multiple binary classification model.

We select a group of properties that are considered to be related to different travel mode. Based on a review of the existing literature, the following properties sets are obtained (Table 2).

Table 2. Travel characteristics when connecting any two travel modes.

	Walk	Bike	Bus	Metro	Car
Walk	/	Velocity change	Waiting	Waiting	Velocity change
Bike	Velocity change	/	+Walk	+Walk	Velocity change
Bus	Velocity change	/	/	+Walk	+Walk
Metro	Velocity change	+Walk	+Walk	/	+Walk
Car	Velocity change	+Walk	+Walk	+Walk	/

2.1. Identify the Walk to Split the Trip Chain into Trip Segments

For a trip chain consisting of a series of trajectory points, one or more modes of transportation may exist to travel. It is necessary to cut the trip chain into travel segments/trip legs and then determine the mode of each travel segment. The changes in travel characteristics when interchanging or connecting any two travel modes are shown in Table 2. When the travel mode is transferred from walking to any other transportation mode, such as from walking to bicycle or bus, there is a significant change in speed or a waiting period. When the travel mode is transferred from bicycle to metro, it is connected by walking. Therefore,

we slice the trip chain into travel segments by identifying walking. That is, after identifying walking in a complete multi-mode travel segment, a number of trajectories before and after walking can be defined as a travel segment.

Taking a single trajectory point as the calculation object, the series of features such as the velocity set, acceleration set, and azimuth change set for the previous 2 min of this point are calculated to obtain the travel feature set. Applying Equation (1) and calculating the relationship between these features and travel mode of walking, we can obtain the travel mode marker for each trajectory point. The travel segment of walking can be obtained by merging a segment of trajectory that is continuously marked as walking or a segment in which more than 80% of the points are marked as walking.

2.2. Identify Metro and Bike Trips Based on Public Transport Data Fusion

After slicing the trip chain into travel segments, we focus on the travel characteristics of each travel segment. Compared to metro and bicycle trips, metro trips start and end at metro stations, and travel distances are usually greater than bicycles and at greater speeds than bicycles. Due to the limitation of data volume, in order to improve the data recognition accuracy, we can expand our training set by fusing the travel data from the transportation departments. Public transportation management has metro card data that can record the time of passengers entering and leaving the station, as well as data on the starting and ending locations of bicycles. Therefore, our preliminary selected travel characteristics are as follows.

- Travel distance;
- Travel time;
- Travel velocity;
- The distance between the start and end of a trip and the metro station.

To find whether there is a metro station within 100 m of each trajectory point, we use the Geohash method. Geohash is essentially a form of spatial indexing [26]. It converts a two-dimensional latitude and longitude into an encoding, each of which represents a certain rectangular area. In other words, all points (latitude and longitude coordinates) in this rectangle share the same Geohash code [27]. For example, Point A and Point B are in the same rectangle and they share a Geohash code WX4e5, but Point C with the number WX4ep is not in the same rectangle as them (Figure 4). This can help us to find the metro or bus station around the track point quickly.

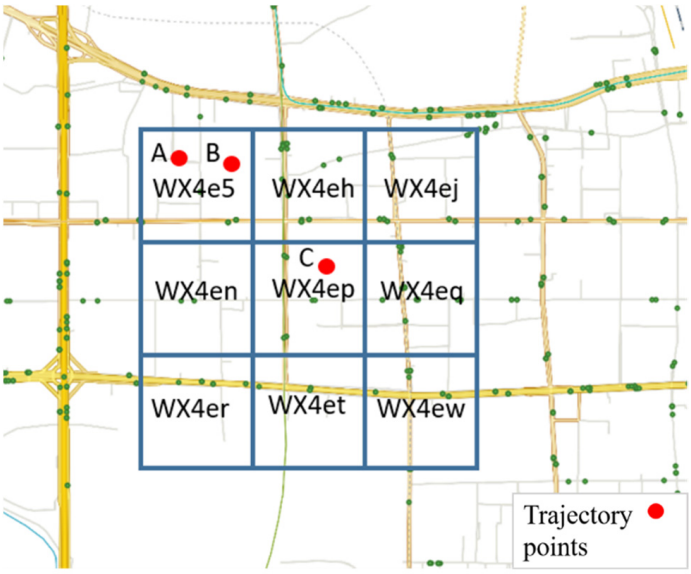


Figure 4. Sample of Geohash.

Thus, we labeled the travel segment data as metro, bicycle, and other. Next, we distinguish between car and bus from the travel segments marked as other.

2.3. Identify Bus and Car Based on Geographic Data Fusion

We use the set of travel features obtained in the first two models to distinguish between bus and car, but the recognition accuracy was not very good. Therefore, the travel features specific to bus trips were added. For example, the origin and destination of bus trips are near bus stops. Both bus and car travel processes encounter congestion or red light intersections, and there are multiple decelerations, stops, and re-accelerations. However, buses will regularly enter and exit at bus stops along the way, i.e., they slow down, stop, and reaccelerate near bus stops [28]. Therefore, we look for the trajectory points in the travel segment where all trips decelerate, stop, and re-accelerate and find whether there is a bus stop near these point. The value for calculating the ratio of stops near bus stops to all stops is given in the following Equation (2).

$$p = \frac{\sum_{i=1}^{n} s_i}{\sum_{i=1}^{n} a_i} \tag{2}$$

p represents the ratio of stops near bus stops to all stops in a trip. a_i indicates whether the ith trajectory point is a stopping point. If a_i is a stopping point, then $a_i = 1$; otherwise, $a_i = 0$. s_i indicates whether the ith trajectory point is a stopping point and is within 100 m of the bus stop. If s_i. is a stopping point and is within 100 m of one bus stop, then $s_i = 1$; otherwise, s_i. $= 0$. When the velocity of a trajectory point is less than 2 m/s, the acceleration of its previous 2 points is less than 0, and the acceleration of its last 2 points is greater than 0; then, we consider this trajectory point as a stopping point.

By calculating the number of stops at bus stops as a percentage of all stops, we obtain a new feature used to distinguish between bus and car.

In addition, the distance between the origin and destination of a trip and the bus stops also helps us to determine whether the trip is a bus trip. For example, for a bus trip, whether or not they stop near a bus stop and where the origin and destination are located at the bus stop can help us to evaluate the results. Figure 5 shows a travel segment of bus identified by applying the model we have built. The red color is a series of trajectory points, the shade of color represents the speed of the point (the greater the speed, the darker the color), and the green color is the bus stop. We can see that the segment of the trip starts and ends at a bus stop and has stops at bus stops along the way, and travel speed tends to decrease and then increase near bus stops.

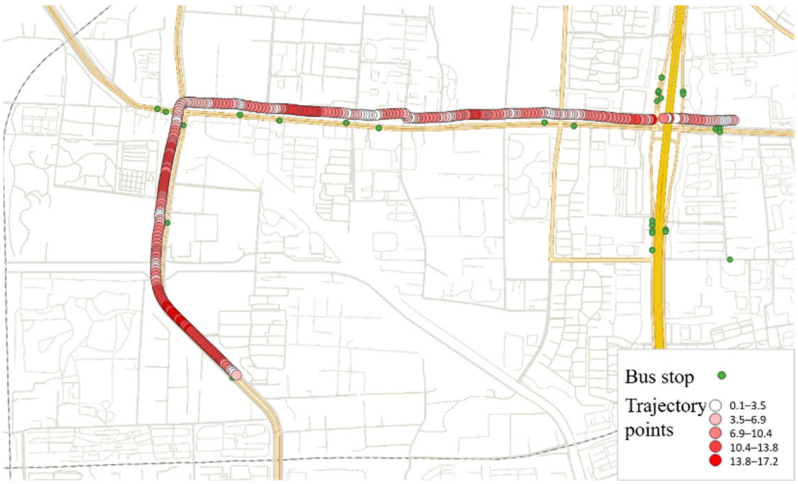


Figure 5. A travel segment of bus trip base on Internet location data.

Therefore, the travel characteristics that were chosen to initially distinguish between bus and car are as follows.

- Velocity set (mean, 95th, 50th, variation of velocity difference);
- Acceleration set (mean, 95th, standard deviation of acceleration, 95th acceleration difference, 50th acceleration difference);
- Azimuth difference set (mean, 95th, 50th);
- The distance between the origin and destination of a trip and the bus stops;
- The ratio of stops near bus stops to all stops.

3. Individual Trip Data Collection and Analysis

3.1. Real-Time Trip Chain Survey with a Smartphone

Individual traveler's real trip chain data should be acquired to test and to improve the accuracy of trip chain model. We analyze the data demand for the trip chain, which should cover different travel modes (walk, bus, bike, car, and metro). In our scheme, we developed a WeChat (a popular Chinese social media application) small program of trip chain recording and recruited volunteers to collect their trip chain record data. At the same time, these volunteers authorized us to extract their GPS data including location and time from the background of program. We collected 1125 trip chain data from April to June 2018.

Each volunteer used the WeChat small program of trip chain recording, and when they traveled, they clicked to start the trip and selected the current status and travel mode, and the background of small program automatically recorded the latitude, longitude, and time points of the travel process. When the volunteer clicked "finish", a trip chain will be generated in the background. A trip chain is a complete record of an activity (e.g., from home to workplace). It does not have to be a full day record of 24 h.

One typical public transport trip record is shown below. This contain two types of data: (1) trip chain recording data (Table 3) and (2) GPS data (latitude, longitude, and time).

Table 3. One typical public transport trip chain record (go to work by bus).

User ID:01	Start Time	End Time	Start Location	Travel Mode/Activity
Travel	7:09	7:21	Home	Walk
Activity	7:21	7:24	Bus Station 1	Waiting
Travel	7:24	7:54	Bus Station 1	Bus
Travel	7:54	8:00	Bus Station 2	Walk
Activity	8:00	8:12	Canteen	Meal
Travel	8:12	8:14	Canteen	Walk
Activity	8:14	/	Company	Work

3.2. Data

There are three sources of data for this article (Table 4), which are WeChat small program collection, public transportation management, and Internet-related companies. These data correspond to three types of uses, respectively. (1) Collecting trip chain data to build models. The trip chain data we collect through small program is used to train and build trip chain segmentation models and travel mode identification models. (2) Data fusion to improve accuracy. The smart card data [29] and bike OD data from public transportation management contain origin and destination information and travel time. Travel features from these data can be used to assist in improving the accuracy of metro and bicycle identification. Bus network data and spatial data from Internet-related companies can be used to analyze spatially relevant travel features, such as finding the nearest bus stop or metro station of a trajectory point [30]. (3) Model application and spatial analysis validation. The established model can be used to identify travel mode of the Internet

location data. Meanwhile, the spatial features of the Internet data can be used to visually validate the model results.

Table 4. Trip data used in this paper.

Data Source	Data	Data Concept	Information
WeChat small program collection	Trip data	User id Starting time Ending time Travel mode Activity GPS (Latitude, longitude, time)	1125 trip chain data from April to June 2018
Public transportation management	Smart card data	Card id Line id Boarding station Check-in time Alighting station Check-out time	More than 5 million transaction records per day
	Bike OD data	User id Starting time Ending time Total distance Starting location Ending location	1000 records
Internet-related companies	Bus network data	Line id Station name	886 bus lines
	Spatial data	Bus station location Metro station location	About 32,000 bus stops and 370 metro stations
	Internet location data	id Latitude Longitude time	About 2000 trips

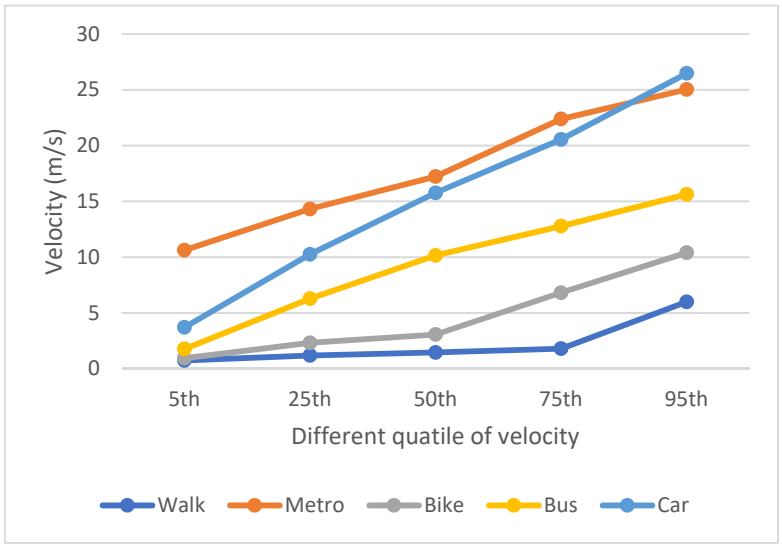
3.3. Data Analysis and Calculate Travel Properties Set

(1) Feature Calculation of Trajectory Points

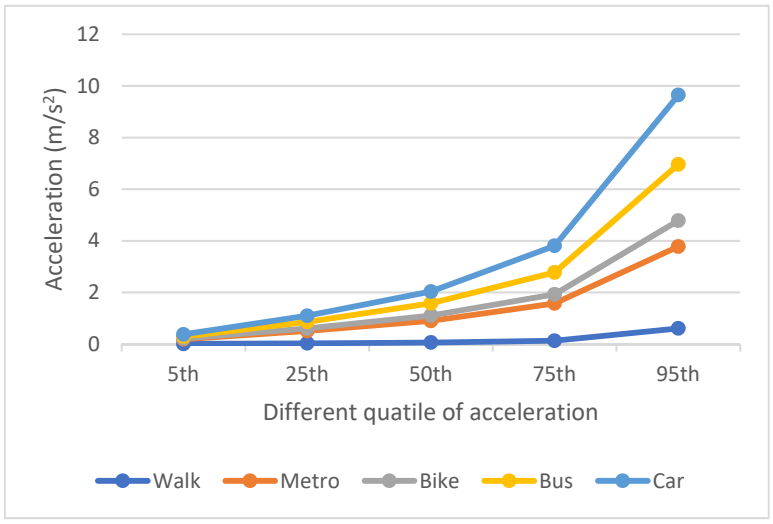
The features of each trajectory point are calculated in two-minute intervals, and we calculate the relevant features in the previous 2 min for each feature point, such as average velocity, maximum velocity, acceleration, and azimuth angle.

We present statistics on the distribution of properties set for each travel mode (Figure 6). It is tentatively inferred that the mean speed values can distinguish between walking, bus, car, and metro. The higher quartiles of speed and acceleration are easier to distinguish different travel mode than the lower quartiles. We have 56,951 marker points for 183 travel segments, including 104 segments on foot and 79 segments by other travel modes.

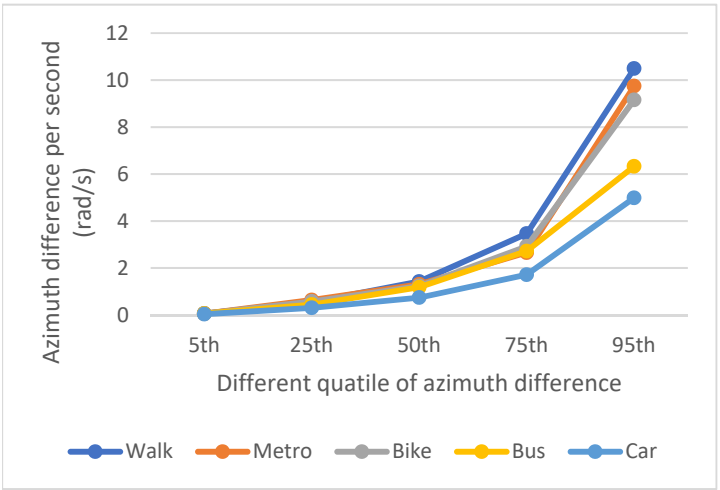
The visualization of the different quartiles of travel characteristics is used to initially distinguish which features can significantly differentiate travel modes. For example, we want to know whether maximum velocity, minimum velocity, or average velocity can distinguish travel modes. We can visualize the different quartiles of speed corresponding to different travel modes to visually determine which feature is more effective. Since there is a random error in the data, we use the 95th percentile of velocity to indicate the maximum velocity and the 5th percentile of velocity to indicate the minimum velocity.



(**a**)



(**b**)



(**c**)

Figure 6. Distribution statistics of travel properties set: (**a**) Distribution statistics of velocity, (**b**) Distribution statistics of acceleration (**c**) Distribution statistics of azimuth difference.

From Figure 6a, we can see that the 95th percentile velocity for walk, bike, bus, metro, and car are 6, 10, 16, 25, and 26 m/s, respectively. There is little difference in this value between metro and car, so we can use the 95th percentile velocity to distinguish walk, bike, and bus, but we cannot distinguish metro and car. The 25th percentile velocity of metro and car are 14 and 10 m/s respectively, so we can use the 25th percentile velocity to distinguish these two modes of travel. We can see that the 95th percentile acceleration can initially analyze different modes of transportation from Figure 6b. Since the 95th percentile acceleration for walk, bike, bus, metro, and car are 0.6, 1, 2.2, 2.7, and 3.2 m/s^2, respectively. The azimuth difference is not significant to distinguish different travel modes in Figure 6c since the values for different travel modes are relatively close.

These different characteristics and the different quartiles of the characteristics can help us to initially analyze different travel modes, while the specific feature selection needs to be further calculated in the model.

(2) Feature Calculation of Travel Segment

For each travel segment, there is a starting point and an ending point. We can determine whether the starting and ending points of the travel segment are within 100 m of a metro station. Similarly, the distance from the start and end of a travel segment to the nearest bus stop can also be calculated.

Figure 6 shows the relationship between the trajectory points (red) and the metro station (yellow) for a section of travel, and green indicates the 100 m buffer range of the metro station. We assist in determining whether the trip is a metro trip by checking whether the start or end point of the trip is located within the green buffer of the metro station. We observe whether the starting or ending point of one trip is located within the green buffer of the subway station to assist in determining whether the trip is a metro trip. From Figure 7, we can see that none of the starting and ending points are within 100 m of the metro station, so we initially obtain that this trip is not a metro trip.

Figure 7. Distance relationship between a travel segment and a metro station.

The process of calculating the distance from a trajectory point to a metro station or bus stop uses the Geohash method. The grid where the start and end point is located is calculated first, then the metro station with the same grid number is found, and then the spatial distance between the point and the metro station is calculated.

4. Results of Survey Data

4.1. Evaluation Results of Multi-Category Model

After the data analysis and travel feature extraction in the previous section, we attempt to apply a multi-classification model to distinguish multiple travel modes based on a unified set of travel features.

We use "R^2" (Equation (3)) to evaluate the accuracy of our model.

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{m} (\overline{y} - y_i)^2} \quad (3)$$

where $\hat{y}_i$ is the predicted value of y using our model, y_i is the actual value of y, and $\overline{y}$ is the mean actual value of y. R-square reflects the extent to which the fluctuation of y can be described by the fluctuation of the independent variables of our model. The value range of R-square is from 0 to 1. The closer R-square is to 1, the more accurate the model is.

As shown in Table 5 below, in the multi-classification model, we evaluate the accuracy of walk, bike, bus, car, and metro as 0.84, 0.37, 0.55, 0.35, and 0.46, respectively.

Table 5. Evaluation results of multi-category model using a unified set of features.

Travel Mode	Walk	Car	Bike	Bus	Metro	Precision
Walk	300	1	33	21	2	0.84
Car	9	84	8	141	1	0.35
Bike	2	36	40	30	0	0.37
Bus	62	67	0	164	4	0.55
Metro	15	20	0	30	55	0.46

The model performed poorly except for the recognition of walking. Therefore, we gradually improved the model by data fusion and by building specific feature sets for different travel modes.

4.2. Evaluation Results of Multiple Binary Classification Model Based on Data Fusion

We built a multiple binary classification model. Firstly, the first binary classification model classified the trip chain into walking and others. Secondly, the other data were divided into metro, bicycle, and other by two binary classification models. Finally, a binary classification model was used to distinguish between bus and car. The size of the data used in the model to identify each travel model is as follows.

The size of data used in this model (Table 6).

Table 6. Travel characteristics when connecting any two travel modes.

Travel Mode	Internet Location Data (1125 Trip Segment)	Transport Department Data	Total Size
Walk	357	0	357
Bike	108	+368	476
Metro	120	+420	540
Bus	297	+300	597
Car	243	+360	603

The evaluation results of multiple binary classification models are shown in Tables 7–9, respectively.

We use "R^2" (Equation (3)) to evaluate our multiple binary classification model based on data fusion. In our final dataset, the data are split into 70%, as a training set, and 30%, as test set. We apply the training set to build the model and use R^2 to calculate the accuracy of the model, and then apply this model to the test set to obtain the evaluated accuracy of this model on the test set.

As shown in Table 10, firstly, the first binary classification model classifies the trip chain into walking and others. In this binary classification, there are 15 travel features. The total number of data is 1125, of which 357 are for walking and 768 are for other. We use 70% of the data as the training set and 30% of the data as the test set, so there are 787 data in the training set and 338 data in the test set. The model results for walking are 0.848 and

0.832 on the training and test sets, respectively. Secondly, the other data are divided into metro, bicycle, and other by two binary classification models. The accuracy of the model for metro is 0.719 on the training set and 0.799 on the test set, and for bicycle is 0.738 and 0.769. Finally, the accuracy of the model is 0.791 and 0.721 on the training set and test set, respectively, by a binary classification model that distinguishes between bus and car.

Table 7. Estimation results of the regression model for walking.

Walk Features	Independent Variables	Coef.	$p>\|t\|$
Time difference	x_1	1.60×10^{-2}	0.001
Long time break	x_2	0.2612	0.001
Velocity	x_3	−0.0067	0.000
Average velocity	x_4	0.0216	0.000
Maximum velocity	x_5	−0.0315	0.000
Maximum velocity difference	x_6	0.0252	0.000
Azimuth	x_7	6.89×10^{-2}	0.000
Median azimuth	x_8	8.19×10^{-2}	0.027
Average azimuth	x_9	−0.0002	0.000
Standard deviation of azimuth difference	x_{10}	0.0004	0.000
Standard deviation of azimuth	x_{11}	1.20×10^{-3}	0.000
Acceleration	x_{12}	−0.0463	0.000
Maximum acceleration	x_{13}	0.0145	0.000
Median acceleration	x_{14}	−0.1362	0.000
Acceleration difference	x_{15}	0.0430	0.000
R^2	0.839		

Table 8. Estimation results of the regression model for metro and bike.

Metro and Bike Features	Independent Variables	Coef.	$p>\|t\|$
Velocity	x_1	6.33×10^{-3}	0.014
Metro station label	x_2	0.1113	0.001
Maximum velocity	x_3	0.0385	0.000
The 75th velocity	x_4	−0.1404	0.000
Median velocity	x_5	−0.1250	0.000
Travel time	x_6	0.0423	0.015
R^2 (metro)	0.899		
R^2 (bike)	0.742		

Table 9. Estimation results of the regression model for bus and car.

Bus and Car Features	Independent Variables	Coef.	$p>\|t\|$
Maximum velocity	x_1	−0.1170	0.048
Median velocity	x_2	−0.0420	0.545
Average velocity	x_3	0.1512	0.205
Standard deviation of velocity difference	x_4	0.1861	0.123
Maximum acceleration	x_5	−0.0007	0.995
Median acceleration	x_6	0.1877	0.878
Average acceleration	x_7	0.2055	0.796
Standard deviation of acceleration	x_8	−0.3476	0.059
Maximum acceleration difference	x_9	0.1052	0.053
Median acceleration difference	x_{10}	0.3685	0.199
Travel time	x_{11}	−0.0001	0.108

Table 9. *Cont.*

Bus and Car Features	Independent Variables	Coef.	$p>\|t\|$
Travel distance	x_{12}	4.83×10^{-4}	0.019
Maximum azimuth difference	x_{13}	−0.0031	0.070
Average azimuth difference	x_{14}	0.0185	0.140
Standard deviation of azimuth	x_{15}	0.0007	0.613
The number of stops	x_{16}	−0.0307	0.082
The number of stops nearby bus station	x_{17}	0.0684	0.017
The percent of bus stops	x_{18}	0.3035	0.050
OD nearby bus station	x_{19}	0.2211	0.011
R^2		0.799	

Table 10. The accuracy of model on training set and test set.

Data Sets	Total Size/Current Mode/Other Mode	Training (70%)	Test (30%)	Attributes	Classes	Accuracy of Training Set	Accuracy of Test Set
Walk	1125/357/768	787	338	15	2	0.848	0.832
Metro	1136/476/660	795	341	6	2	0.719	0.799
Bike	1080/540/540	756	324	6	2	0.738	0.769
Car	1200/597/603	840	360	19	2	0.791	0.721
Bus	1200/603/597	840	360	19	2	0.791	0.721

In general, we validate the model and the results in two ways. First, we collected both user's trip location data and the real travel mode data to verify the correctness of our model results. Second, in the model evaluation section, we used R^2 to evaluate the model accuracy.

The model we proposed performs well not only for explaining the data but also for identifying the travel mode.

5. Conclusions

In this paper, we have developed a multiple binary classification model based on data fusion to explain how different attributes can be used to model travel mode. The first binary classification model was used to identify walking and to divide trip chains into travel segments. The next two binary classification models, which fuse the metro card data and bicycle travel data, are used to identify metro and bike. The last binary classification model, which considers the distance from the trip origin and destination to the bus stop, is used to identify buses and cars. The prediction accuracy of the multiple binary classification for walk, metro, bike, bus and car is 0.839, 0.899, 0.742, 0.799, and 0.799 respectively.

We believe that our method could be used not only for explaining different modes of travel but also for applying to engineering practices to identify trip chains from Internet location data.

Existing studies on identifying travel modes have mainly used some travel features based on GPS and survey data from a small number of users. It is not possible to use these models directly for continuous trip chain data consisting of different travel modes. The first binary classification model we built slices trip chains into travel segments by distinguishing between walking and other modes. Compared with the existing models for various types of travel mode recognition, the accuracy of our model is not significantly improved. However, we provide a new methodology to improve the model accuracy by combining transport department data with Internet location data by means of data fusion of travel features. We also verify the improvement of data fusion on model accuracy by comparing the accuracy of one multivariate classification and multiple binary classification models based on data fusion. Furthermore, the current travel mode model is not validated

on Internet data; we extend the application of the travel mode identification model based on travel feature extraction on large-scale data and validate the possibility of using Internet location data to analyze trip chains.

The innovation of our study firstly lies in the new approach to modeling travel mode based on data fusion. There has been a significant amount of research on travel mode recognition, and the current research can be improved from algorithms or data sources. We fused tagged travel data from transport management department and untagged location data from Internet to improve the model training accuracy. This greatly extends the usability of travel mode recognition models. Secondly, we built different sets of features for different travel modes and used multiple binary classification models to extract travel modes step-by-step. The model accuracy was significantly improved compared to the multi-classification model with only one feature set. Thirdly, we segmented travel modes by identifying travel segments on walking and used travel data from transport management department to improve the imbalance of traffic mode data. These helped us to apply the model of travel mode to Internet location data.

This work can still be improved in a few ways. Firstly, the accuracy of the current binary classification model still needs further improvement. Several features can be added to the existing methodology in the future. Data cleansing and data governance for large scale data have the potential to improve the accuracy of travel chain data analysis [31]. Secondly, more information can be extracted from the trip chain, such as individual travel demand and transfer issues in the travel network, so that we can establish the relationship between individual travel demand and transport network in the future research.

Author Contributions: Writing—original draft, methodology, and data resources, W.W.; project administration and review, Y.C. (Yanyan Chen); writing—review and editing, H.S. and Y.C. (Yusen Chen). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (Grant No. 2018YFB1601302).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data can only be shared internally within the institute where the corresponding author works.

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

References

1. Link, C.; Liu, L.; Hou, A.; Biderman, A. Understanding Individual and Collective Mobility Patterns from Smart card Records: A Case Study in Shenzhen. 2018. Available online: https://ieeexplore.ieee.org/abstract/document/5309662 (accessed on 1 November 2020).
2. Viggiano, C.; Koutsopoulos, H.N.; Attanucci, J.; Nigel, H.; Wilson, M. Inferring Public Transport Access Distance from Smart Card Registration and Transaction Data. Available online: https://journals.sagepub.com/doi/abs/10.3141/2544-07 (accessed on 1 November 2020).
3. Wang, Y.; Correia, G.; de Romph, E.; Santos, B.F. Road Network Design in a Developing Country Using Mobile Phone Data: An Application to Senegal. *IEEE Intell. Transp. Syst. Mag.* **2018**, *31*, 2–15. [CrossRef]
4. Vishwanath, A.; Gan, H.S.; Kalyanaraman, S.; Winter, S.; Mareels, I. Personalized Public Transportation: A Mobility Model and its Application to Melbourne. *IEEE Intell. Transp. Syst. Mag.* **2015**, *7*, 37–48. [CrossRef]
5. Mäenpää, H.; Lobov, A.; Martinez Lastra, J.L. Travel mode estimation for multi-modal journey planner. *Transp. Res. Part C Emerg. Technol.* **2017**, *82*, 273–289. [CrossRef]
6. Daisy, N.S.; Millward, H.; Liu, L. Trip chaining and tour mode choice of non-workers grouped by daily activity patterns. *J. Transp. Geogr.* **2018**, *69*, 150–162. [CrossRef]
7. Bai, L.; Sze, N.N.; Liu, P.; Guo Haggart, A. Effect of environmental awareness on electric bicycle users' mode choices. *Transp. Res. Part D Transp. Environ.* **2020**, *82*, 102320. [CrossRef]
8. Reddy, S.; Burke, J.; Estrin, D.; Hansen, M.; Srivastava, M. Determining Transportation Mode On Mobile Phones. In Proceedings of the 2008 12th IEEE International Symposium on Wearable Computers, Pittsburgh, PA, USA, 28 September–1 October 2008; pp. 25–28. Available online: https://ieeexplore.ieee.org/abstract/document/4911579 (accessed on 1 November 2020).

9. Semanjski, I.; Gautama, S.; Ahas, R.; Witlox, F. Spatial context mining approach for transport mode recognition from mobile sensed big data. *Comput. Environ. Urban Syst.* **2017**, *66*, 38–52. [CrossRef]
10. Shin, D.; Aliaga, D.; Tunçer, B.; Arisona, S.M.; Kim, S.; Zünd, D.; Schmitt, G. Urban sensing: Using smartphones for transportation mode classification. *Comput. Environ. Urban Syst.* **2015**, *53*, 76–86. [CrossRef]
11. Gonzalez, P.A.; Weinstein, J.S.; Barbeau, S.J.; Labrador, M.A.; Winters, P.L.; Georggi, N.L.; Perez, R. Automating mode detection for travel behaviour analysis by using global positioning systems-enabled mobile phones and neural networks. *IET Intell. Transp. Syst.* **2010**, *4*, 37. [CrossRef]
12. Huang, Y.; Gao, L.; Ni, A.; Liu, X. Analysis of travel mode choice and trip chain pattern relationships based on multi-day GPS data: A case study in Shanghai, China. *J. Transp. Geogr.* **2021**, *93*, 103070. [CrossRef]
13. Bi, H.; Shang, W.L.; Chen, Y.; Wang, K.; Yu, Q.; Sui, Y. GIS aided sustainable urban road management with a unifying queueing and neural network model. *Appl. Energy* **2021**, *291*, 116818. [CrossRef]
14. Rasmussen, T.K.; Ingvardson, J.B.; Halldórsdóttir, K.; Nielsen, O.A. Improved methods to deduct trip legs and mode from travel surveys using wearable GPS devices: A case study from the Greater Copenhagen area. *Comput. Environ. Urban Syst.* **2015**, *54*, 301–313. [CrossRef]
15. Snellen, D. Urban Form and Activity-Travel Patterns: An Activity-Based Approach to Travel in a Spatial Context. Available online: https://research.tue.nl/en/publications/urban-form-and-activity-travel-patterns-an-activity-based-approac (accessed on 1 November 2020).
16. Zhe, L.; Jian, S.; Xunyou, N. Travel Mode Recognition Based on Smart Phone Big Data. 2016. Available online: http://en.cnki.com.cn/Article_en/CJFDTOTAL-JSYJ201612003.htm (accessed on 1 November 2020).
17. Xiao, G.; Juan, Z.; Zhang, C. Travel mode detection based on GPS track data and Bayesian networks. *Comput. Environ. Urban Syst.* **2015**, *54*, 14–22. [CrossRef]
18. Hao, L.; Wan, F.; Ma, N.; Wang, Y. Analysis of the Development of WeChat Mini Program. *J. Phys. Conf. Ser.* **2018**, 1087. Available online: https://www.researchgate.net/publication/328033998_Analysis_of_the_Development_of_WeChat_Mini_Program (accessed on 1 November 2020). [CrossRef]
19. Su, X.; Caceres, H.; Tong, H.; He, Q. Online Travel Mode Identification Using Smartphones with Battery Saving Considerations. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 2921–2934. [CrossRef]
20. Lu, Y.; Seshadri, R.; Pereira, F.; OSullivan, A.; Antoniou, C.; Ben-Akiva, M. DynaMIT2.0: Architecture Design and Preliminary Results on Real-Time Data Fusion for Traffic Prediction and Crisis Management. In Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems, Gran Canaria, Spain, 15–18 September 2015; pp. 2250–2255. Available online: https://ieeexplore.ieee.org/abstract/document/7313455 (accessed on 1 November 2020).
21. Aceto, G.; Ciuonzo, D.; Montieri, A.; Pescapé, A. Multi-classification approaches for classifying mobile app traffic. *J. Netw. Comput. Appl.* **2018**, *103*, 131–145. [CrossRef]
22. Rasouli, S.; Timmermans, H. Accounting for Heterogeneity in Travel Episode Satisfaction Using a Random Parameters Panel Effects Regression Model. *Procedia Environ. Sci.* **2014**, *22*, 35–42. [CrossRef]
23. Shen, Y.; Xu, J.; Li, J. A probabilistic model for vehicle scheduling based on stochastic trip times. *Transp. Res. Part B Methodol.* **2016**, *85*, 19–31. [CrossRef]
24. Zhou, X.; Yu, W.; Sullivan, W.C. Making pervasive sensing possible: Effective travel mode sensing based on smartphones. *Comput. Environ. Urban Syst.* **2016**, *58*, 52–59. [CrossRef]
25. Benoit, K. Linear Regression Models with Logarithmic Transformations. 2011, pp. 1–8. Available online: https://links.sharezomics.com/assets/uploads/files/1600247928973-from_slack_logmodels2.pdf (accessed on 1 November 2020).
26. Liu, J.; Li, H.; Gao, Y.; Yu, H.; Jiang, D. A geohash-based index for spatial data management in distributed memory. In Proceedings of the 2014 22nd International CFonference on Geoinformatics, Kaohsiung, Taiwan, 25–27 June 2014; pp. 5–8. Available online: https://ieeexplore.ieee.org/abstract/document/6950819 (accessed on 1 November 2020).
27. McKenzie, G.; Janowicz, K. Where is also about time: A location-distortion model to improve reverse geocoding using behavior-driven temporal semantic signatures. *Comput. Environ. Urban Syst.* **2015**, *54*, 1–13. [CrossRef]
28. Wang, W.; Chen, Y.; Liu, D.; Zhao, X. Data Mining of Individual Trip Chain Based on Mobile Phone and Data Exploration of Trip Properties. *Urban Transp. China* **2020**, *18*. Available online: http://qikan.cqvip.com/Qikan/Article/Detail?id=7103613795 (accessed on 1 November 2020).
29. Demir Alan, U.; Birant, D. Server-Based Intelligent Public Transportation System with NFC. *IEEE Intell. Transp. Syst. Mag.* **2018**, *10*, 30–46. [CrossRef]
30. Shang, W.L.; Chen, Y.; Bi, H.; Zhang, H.; Ma, C.; Ochieng, W.Y. Statistical Characteristics and Community Analysis of Urban Road Networks. *Complexity* **2020**. Available online: https://www.hindawi.com/journals/complexity/2020/6025821/ (accessed on 1 November 2020). [CrossRef]
31. Shang, W.L.; Chen, J.; Bi, H.; Sui, Y.; Chen, Y.; Yu, H. Impacts of COVID-19 pandemic on user behaviors and environmental benefits of bike sharing: A big-data analysis. *Appl. Energy* **2021**, *285*, 116429. [CrossRef] [PubMed]