ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20200513124207
Preliminary communication

Route Restoration Method for Sparse Taxi GPS trajectory based on Bayesian Network

Guangyao LI, Zhengfeng HUANG*, Leyi LOU, Pengjun ZHENG

Abstract: In order to improve the availability of taxi GPS big data, we restore the chosen route for the sparse taxi GPS trajectory in this work. A trajectory restoration method based on Bayesian network is proposed. Compared with the traditional research solely based on time-spatial variables, this method additionally considers the characteristics of empty/heavy taxi status, weather conditions, drivers, vehicle running and other factors to carry out route restoration. A field case of grid network in Ningbo is taken to verify the applicability of the method, using the taxi GPS trajectory data from Ningbo Taxi Information Management Platform. The case results show that the accuracy of Bayesian network method based on multiple factors reaches 91.4%. Its performance is superior to the Multivariate logistic regression model. In addition, the proposed method is especially suitable for scenarios with a high missing rate of track data, such as a scene with timespan of about 5 min between neighbour trajectories.

Keywords: Bayesian network; missing rate; multiple factors; sparse Taxi GPS data; trajectory restoration

1 INTRODUCTION

With the progress and development of information & communication technologies as well as the popularization of navigation-related smart devices, it is convenient to obtain the GPS data nowadays. If we expand the taxi GPS data relation with other data bases, we can acquire abundant vehicle state information, such as latitude and longitude coordinates, license plate number, travel time, empty/heavy taxi status, real-time speed, the name of road section etc.[1]. Assuming the high-frequency GPS is acquired, traveller position could match with the road network. Through map-matching, we can achieve the judgment of the traveller's route trajectory. Further, we can understand the path-selection behaviour of travellers by combining with individual characteristics [2-4].

It is convenient to restore vehicles' travel route by using GPS data with fine time granularity. However, influenced by several factors such as the quality problem of the positioning equipment and the interference of the surrounding environment, the taxi GPS sampling rate is sometimes low. Yuan et al. [5] present the statistical distribution of the sampling intervals of the GPS trajectories generated by more than10000 taxies in Beijing in a week. The average time interval of the data set is 3.27 minutes. According to the result, only 34% of the data sampling interval is less than 1 minute. Low sampling would negatively affect the route restoration. For example, there is a car running at a speed of 50 km/h. The situation, no GPS data within certain 2 minutes, would lead to no vehicle track information in the 1.6 km interval. If the car is running in a multi-intersection zone, there are lots of route options and it is a hard work to restore the actual route chosen by this driver. This issue stands out more especially in urban areas, where road network is composed of short links and vehicles can travel on many different road segments in few time intervals. This kind of problem can be named "restoring the real route based on the sparse GPS trajectory". It is a special type of map-matching. For the situation of high GPS missing rate, we cannot align a sequence of observed user positions with digital road network straightly. If this problem can be solved well, the cost of data collection and storage will be reduced.

2 LITERATURE REVIEW

The sparse taxi GPS data are mainly generated in the following four scenarios: (1) the uncertainty of taxi data transmission when it travels around tall buildings, tunnels, canyons and elevated roads; (2) the positioning system of a mobile phone or car GPS device is turned off and cannot transmit location information; (3) in order to save the communication and energy cost, the taxis usually report their GPS positions to the dispatching centre with low sampling-rate data [6]; (4) GPS data lost in the process of transmission. For the last three sparse data scenarios, we can use the historical taxi GPS data to restore the travel route. It is also the scenario and strategy we focus on in this paper.

Generally, researchers adopt (1) shortest route, (2) Fréchet distance, (3) Hidden Markov Model (HMM)-related methods, (4) evaluation and optimization methods to solve this kind of path-restoration problem.

Shortest-path method is one of the earliest methods proposed to deal with route restoration based on sparse GPS data, such as the work by Bierlaire et al. [7]. Some papers use other algorithms, such as the modified *A** shortest route algorithm [8].Their method is successfully applied in sparse road network, where few optional routes are needed. But when applied to the complex urban road network, it is difficult to accurately restore the actual track by the shortest route method, because there are many candidate routes in the research range.

Brakatsoulas et al. [9] propose a map matching algorithm based on curve similarity. This algorithm uses Average Fréchet Distance (AFD) to measure the matching degree of GPS sequence and candidate section sequence, and the route with the highest matching degree is used as the final matching route. The example shows that the algorithm can be applied to the sparse GPS data with missing rate of about 30s and obtain good matching accuracy, but the application of this method is very complex.

A series of map-matching methods based on HMM are widely used in low-frequency sampling rate data sets [10]. Lou Y et al. proposed ST-Matching algorithm based on HMM [6], which can restore the vehicles' travel route by combining spatially geometric and topological structures. Compared with AFD-based algorithm, ST-Matching

improves accuracy as well as running time. Furthermore, an Interactive Voting-based Map Matching (IVMM) algorithm [5] is proposed based on ST-Matching. The IVMM performs better than ST-Matching, because its voting measure could take into account the interaction of candidate mapping points. But the shortest-path method is used to restore the route between the adjacent GPS points in the IVMM algorithm, so the matching accuracy for complex road network is not high. Experiments show that the accuracy of IVMM algorithm is only 70% when the sampling precision is 2 minutes. There are lots of other hybrid HMM methods in the path-restoration work, such as Ozdemir et al. [11], Taguchi et al. [12], etc. However, these methods may not be capable for the condition of over-high GPS missing rate.

As to the condition of trajectory span time of more than 2 minutes, other methods are generally used to solve the path-restoration problem. In the study of the track patching for incomplete vehicle location data, Wang et al. [13] constructed the utility function by considering the matching score between travel time and distance, non-linear rate of candidate route. The proposed utility is combined with Technique for Order Preference by Similarity to an Ideal Solution (TOPSIS) evaluation method to achieve the track patching. Their method requires that the field should be grid network. Zhao et al. [14] proposed a distributionally robust optimization to restore route for vehicle trajectory data. However, their data is obtained by experimentally connected vehicle, and the route-choice behavior may not be the same with real-life condition. Fu et al. [15] proposed an optimization algorithm for multi feature matching of road network. Starting from shape, distance and semantics, this algorithm can more accurately describe the feature differences between matching pairs of roads, and build regression matching model of road network by training similar feature sample sets.

In addition, some scholars improve the matching accuracy by combining GPS with multivariate data from other sensors such as WIFI [16], cellular fingerprint [17], inertial sensor [18] etc. Although these advanced devices embedded with road semantics [19] could improve the accuracy of route restoration, it also increases the cost of data acquisition.

In addition, some of the above methods do not take into account the characteristics of regional traffic operation, driver's route selection behavior characteristics, and other environmental factors, so their trajectory restoration methods are not necessarily suitable for road network scenarios with high GPS track missing rate. This paper mainly aims at the situation of high data missing rate (the time of no track point reaches 5 minutes or more), based on the collection of urban taxi GPS data, comprehensively considers the factors such as time, space, driver characteristics, environmental characteristics, operation characteristics, etc., and reconstructs the route track combining with the actual road network situation. Specifically, this paper uses the dense GPS data combined with the driver data, weather conditions and other related attributes to train the samples, so as to establish the Bayesian network model of route restoration, and then use the model to predict the sparse GPS trajectory route.

3 PROBLEM DESCRIPTION

Trajectory restoration process in our study could be described as follows.

Firstly, we conduct the variable definition. Let $Area = \{V, E\}$ denote a road network area, where E is the set of road segments and V is the set of the intersection or turning points in the road network. A road segment $e_m \in E$ contains identity code $e_m \cdot id$, its length $e_m \cdot l$, starting point $e_m start \in V$ and ending point $e_m end \in V$. A route R is a sequence of connected road segments between two specified vertices (1 and m) in the road network, i.e. R: $e_1 \rightarrow e_2 \rightarrow, ..., e_k, ..., \rightarrow e_m$, where $e_k \cdot end \equiv e_{k+1} \cdot start$ and $1 \le k < m$. Let g denote a GPS log as a sequence: $g = \{p_1, p_2, \ldots, p_i, \ldots, p_n\}$. Each track p_i is specified by its longitude $p_{i.lng}$, latitude $p_{i.lat}$, timestamp $p_{i.t}$, speed $p_{i.v}$, nearest road segment $p_{i.e}$, namely $pi = (p_{i.lng}, p_{i.lat}, p_{i.t}, p_{i.v}, p_{i.e})$. As to the situation that adjacent two points pi and p_{i+1} be across multiple road segments in the floating car GPS log, it means this point pair represent the sparse feature of GPS trajectory.

The second step is the route restoration process; it includes several sub-steps. First, extract historical GPS data which has the same starting and ending links as input samples, and conduct data correction. Then, acquire the related attribute values for the vehicles, drivers and other factors in both matched links of upstream and downstream GPS points. Finally, use the Bayesian network model to restore the route trajectory of the vehicle.

4 BAYESIAN NETWORK MODEL

Bayesian network is a directed graphical model based on probabilistic reasoning. It expresses the association of variables in the problem through a visual network model, which is suitable for the representation of uncertain knowledge. The Bayesian network consists of a set of nodes $U = \{X_1, X_2, \ldots, X_n\}$, directed arcs and the network parameter Θ. A directed arc connects the parent node and its child node, which uses a quantitative, probabilistic causality measure to represent its attribute value. The network parameter represents the conditional probability set from each parent node to its child node in the network.

$$\Theta = \left\{\theta_{X_1|pa(X_1)}, \theta_{X_2|pa(X_2)}, ..., \theta_{X_n|pa(X_n)}\right\} \quad (1)$$

In this formula: $pa(X_i)$ represents the parent node set corresponding to the child node X_i; $\theta_{X_i|pa(X_i)}$ represents the conditional probability table structured from $pa(X_i)$ to X_i.

The work of factor choice and feature extraction is the first step to construct the Bayesian network model. In other word about this step, we should measure the node variables by qualitative or quantitative methods. Structural learning is the second step. It means we should determine the structure of the Bayesian network by the methods of domain expert, prior knowledge and machine learning. Fitting the network parameters under the established network structure is the third step. Network inference is the

fourth step, where we can calculate the posterior probability of the target situation.

Structure learning is the most critical step in the construction of Bayesian network model, in which machine learning is considered to be an effective method especially in a data-rich environment. As to the machine learning method of Bayesian network structure learning, score search method is generally used. Its principle is to find the best Bayesian network structure according to certain search strategy and scoring criteria. The typical methods based on score search include exhaustive search, K_2 algorithm [20] and hill-climbing algorithm, in which K_2 algorithm is widely used. The basic idea of K_2 algorithm is as follows. First use a graph without arcs as an initial input; then examine the next node in turn according to the node significance, and each time we decide whether the previous node needs to be the parent node of the current node until all the nodes are inspected. When the added node is under inspection, the maximum of scoring function is taken as the goal to optimize the current network topology G.

There are mainly two scoring functions for the Bayesian network structure learning. One is based on the Bayesian statistics and the other is related to Mutual Information. Bayesian scoring first gives the prior probability $P(G)$ of the network structure G; then it calculates the posterior probability of G with Bayesian formula under the given data set D. The structure with the maximum posterior probability is the optimal network structure. The scoring function based on Mutual information theory mainly includes Bayesian Information Criterion (BIC) and Minimum Description Length (MDL). The scoring function is composed of the optimal parameter loglikelihood and the penalty term. Optimal parameter loglikelihood is used to evaluate the fitting degree of structure and data. Penalty term is used to avoid the over-fitting of the network and make the network relatively simple. But sometimes the network structure is too simple, and there is no parent-child relationship between nodes with obvious logical correlation [21, 22]. When the number of network nodes or the amount of sample data is too large, BIC scoring function should be the option.

Due to the moderate amount of network nodes and sample data in this paper, we applied K_2 algorithm to obtain the structure of Bayesian network and used the Bayesian scoring function in structure learning. The Bayesian scoring function is as follows:

$$K_2(G,D)=\sum_{i=1}^{n}\sum_{j=1}^{q_i}\left(\log\left(\frac{(r_i-1)!}{(N_{ij}+r_i-1)}\right)\right)+ \\ +\sum_{k=1}^{r_i} log\left(N_{ijk}!\right)+\log\left(P(G)\right) \tag{2}$$

In this formula: D represents a set of sample data; G represents the network morphology; U represents the nodes of network; $P(G)$ represents the priori probability of network morphology G (it is assumed that the probability of network morphological occurrence obeys the uniform distribution.); q_i represents the number of total states for $X_t^{'}$ s parent nodes; r_i represents the number of total states for node X_i; N_{ijk} is the total sample number corresponding to the fixed feature values of child-parent nodes for certain combinatory structure. The variable $N_{ij}=\sum_{k=1}^{r_i}N_{ijk}$.

5 INDICATORS FOR MODEL TESTING

5.1 Accuracy

We use the sample ratio of correct route restoration to the total number to reflect the accuracy of the model. Its definition is as shown in Eq. (3). In this formula, F_{ii} means that the model correctly identifies the route i. M and F represent the total number of candidate paths and samples separately.

$$A=\frac{\sum_{i=1}^{M}F_{ii}}{F} \tag{3}$$

5.2 Recall Ratio

The recall ratio represents the accuracy of specific path identification. It is used to analyse the effectiveness of the model to identify certain path. The recall ratio for path i is defined as shown in Eq. (4). The numerator represents the times of correct match about path i by the model [23]. The denominator represents the sample size of choosing path i in true world.

$$B_i=\frac{F_{ii}}{\sum_{j=1}^{M}F_{ji}} \tag{4}$$

5.3 Precision Ratio

The precision ratio is used to evaluate the level of the model about distinguishing specific categories of samples from other samples. The definition of the precision ratio corresponding to route i identification is as follows.

$$C=\frac{F_{ii}}{\sum_{j=1}^{M}F_{ji}} \tag{5}$$

5.4 Receiver Operating Characteristic (ROC) Curve

The ROC curve is frequently used to evaluate the predictive performance for the models of binary outcomes. The ROC curve is a graphical plot of the true positive rate (TPR) on the y-axis against false positive rate (FPR) on the x-axis for a threshold running from 0 to 1 [24]. The area under the ROC curve (AUC) can be used as an evaluation measure of the predictive performance. The AUC falls between 0 and 1. A larger AUC indicates the better predictive performance. Tab. 1 is the evaluation standard of ROC curve.

In this paper, the TPR represents the correct judgement ratio of a designated route, in terms of (times of correct judgement for designated route / sample size of designated route); FPR represents the sample occupancy of misjudgement, in terms of (times of error judgement for designated route from other samples/(1-sample size of designated route)). The threshold herein is used for a

designated route judgement. We would match the GPS trajectories to the designated route if the probability (obtained by Bayesian network) was larger than the threshold rather than 0.5.

Table 1 Evaluation Criteria for ROC Curve

AUC	Evaluation result of diagnostic test
0.9 ~ 1	Excellent
0.7 ~ 0.9	Good
0.5 ~ 0.7	Medium
< 0.5	Bad

5.5 Mutual Information

The mutual information [25] indicates whether the two variables X and Y have a strong relationship. The formula is as follows. If X is independent of Y, then $P(X, Y) = P(X)(Y)$, $I(X, Y) = 0$ and it means that X is not related to Y. This index would be used to analyse the sensitivity of the nodes in Bayesian network. Specifically, we analyse the sensitivity of factor nodes to route-choice node to test the rationality of factor selection in our model.

$$I(X,Y) = \int_X \int_Y P(X,Y) \log \frac{P(X,Y)}{P(X)\ (Y)} \tag{6}$$

6 DATA SOURCE AND PREPROCESSING

The data used in this paper is obtained from the Ningbo Taxi Service Information Management Platform, which has taxi GPS track data with sampling period of 15 seconds and the basic information about the drivers. This paper selects the road network around Ningbo Sports Center as the research scope. As is shown in Fig. 1, we choose an upstream link and downstream link as origin link and destination link separately to search for GPS trajectory samples. After eliminating abnormal data such as unreasonable travel time, vehicle speeds and detour routes, we obtain 15248 valid data, and the valid rate is 90 percent. The data range is from December 9, 2017 to December 31, 2018, with an average of 41 valid daily data. We apply the ArcGIS software to match these data to the road network and correlate with the track route. We obtain 18 alternative routes.

In order to reasonably select finite routes as the alternative set, we set the following principles: first order the route according to the chosen probability; then the last route within the alternative set should be no less than the selection probability of the remaining route set. The study found that if the 6 routes of the highest choice ratios were included in the analysis, the above principles could be satisfied. In addition, in order to put all the routes into comparison and meantime limits choice number of the routes, we set the 7th route as the final alternative route. This route was actually the set of remaining routes, which was not identified in Fig. 2 because it contained so many community roads.

In order to test the performance of Bayesian network model in trajectory restoration, the GPS trajectory sequence between origin and destination is deleted to create a scenario of sparse GPS trajectory sample.

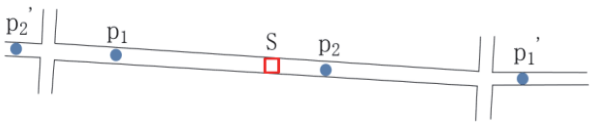
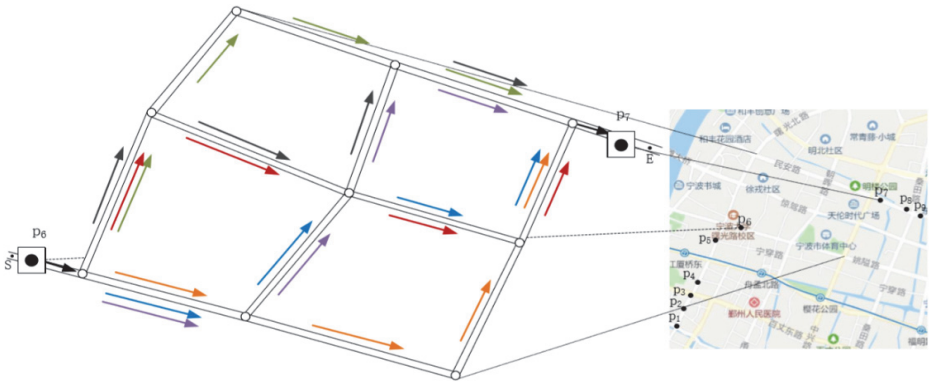


Figure 1 Relation of GPS Points and Common Origin



Factors	Route1	Route2	Route3	Route4	Route5	Route6	Route7
Taxi in-vehicle state	passengers carrying	passengers carrying	No passenger	passengers carrying	No passenger	No passenger	No passenger
Weather conditions	Good weather	Severe weather	Good weather	Good weather	Good weather	Severe weather	Good weather
Time Period	Peak hours in working day	Other time	Peak hours in working day	Other time	Peak hours in working day	Other time	Other time
Travel Time	379	931	551	414	299	400	796
Starting section speed(km/h)	33	15	11	35	18	7	22
Ending section speed(km/h)	25	8	32	5	12	13	18

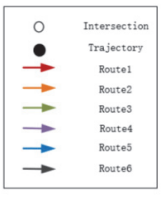


Figure 2 Survey Area and Candidate Trajectory Route

One more problem is how to unify the origin and destination locations of all samples? We map the GPS locations of upstream and downstream to the center of the road section to achieve this aim. The mapping process needs to correct the instantaneous velocity and timestamp of the starting and ending point. Fig. 2 illustrates this method. As is shown in Fig. 2, if S is the required starting point, then we should map other GPS data to this link center. First, we need to find the two adjacent GPS points. For example, p_1 and p_2 are the points that are collected when vehicle i and j passes through the starting road section; p'_1 and p'_2 are the latter valid adjacent sample points of p_1 and p_2 separately. We can calculate the speed of the vehicle i at the link center S as follows:

$$p_{is.v} = \frac{l_{i1}}{\Delta t} \tag{7}$$

In this formula: $p_{is.v}$ is the speed of the vehicle i at the link center S; l_{i1} represents the distance between p_1 and p'_1; Δt is the sampling time interval.

The timestamp correction result of the calculated vehicle i at point S is as follows:

$$p_{is.t} = p'_{is.t} + \delta \frac{l_{i2}}{p_{is.v}} \tag{8}$$

In this formula: $p_{is.t}$ is the timestamp of vehicle i at S of the route; $p'_{is.t}$ is the timestamp of vehicle i at the nearest GPS point from S; l_{i2} is the distance between the nearest GPS point and S. δ is a dummy variable. If the nearest point is at the upstream of S, then $\delta = 1$; otherwise, $\delta = -1$.

7 SELECTION OF INFLUENCING FACTORS AND THEIR FEATURES FOR ROUTE RESTORATION

For sparse GPS trajectories, various factors associated with route selection behavior are summarized as follows:
1) Operational characteristics. Taxi with various in-vehicle state will choose different routes. For instance, empty vehicles prefer to choose the route around the hot spot area [26]. In addition, the empty vehicle will drive slowly to observe potential guests.
2) Speed characteristics. There is a spatial correlation in traffic state between link and route [27]. As to each route, the taxi must pass through the starting and ending sections. The speeds on these two sections can reflect the traffic state of each route in certain degree.
3) Environmental characteristics. Weather conditions have a great influence on driver's route choice [28]. The source of weather data is from the network.
4) Driver characteristics. The route choice behaviours of a driver vary even in the same environment. These behaviours mainly depend on a driver's own characteristics (such as age, driving experience, gender, etc.); Zhang et al. use Logit model to find that young or short driving-years drivers tend to change route more frequently, and the elderly or long driving-years drivers are easier to adhere to a specific route [29].

Obviously, the drivers focus more on the taxi in-vehicle empty/heavy status, vehicle speed and weather condition when choosing a travel route. Based on the relevant literature, expert experience and further correlation analysis, we select 10 node variables including taxi in-vehicle empty/heavy state X_1, weather condition X_2, time period X_3, travel time X_4, starting section speed X_5, ending section speed X_6, driving years X_7, driving ages X_8, driver gender X_9 and the root node X_{10} (represents the alternative routes for driver to choose). These variables are used as network input nodes.

Table 2 Variable Classification and Feature-Related Sample Proportion (Training Set)

Classification	Variable	Symbolic representation	Variable value	Sample number	Sample proportion
Operating characteristics	Taxi in-vehicle state	X_1	1. no passenger	3015	26%
			2. passengers carrying	8421	74%
Environmental characteristics	Weather condition	X_2	1. severe weather	760	7%
			2. good weather	10676	93%
	Time period	X_3	1. peak hours in working day (7:00 - 8:30, 16:30 - 18:30)	1629	14%
			2. other time	9807	86%
Traffic characteristics	Travel time	X_4	1. 300 s or less	2232	19%
			2. 300 - 400s	3476	30%
			3. 400 - 500s	2235	20%
			4. 500 s and above	3493	31%
	Starting section speed	X_5	1. speed < 10 km/h	915	8%
			2. 10 km/h <= speed < 20 km/h	3087	27%
			3. 20 km/h <= speed < 30 km/h	5604	49%
			4. 30 km/h <= speed	1830	16%
	Ending section speed	X_6	1. speed < 10 km/h	801	7%
			2. 10 km/h <= speed < 20 km/h	3774	33%
			3. 20 km/h <= speed < 30 km/h	5717	50%
			4. 30 km/h <= speed	1144	10%
Driver characteristics	Driving years	X_7	1. less than 5years	1559	14%
			2. 5 - 10 years	3035	26%
			3. 10 - 15 years	2700	24%
			4. 15 years and above	4142	36%
	Driving ages	X_8	1. under 35 years old	1004	9%
			2. between 35 - 45 years old	3403	30%
			3. between 45 - 55 years old	6203	54%
			4. 55 years old and above	826	7%
	Driver gender	X_9	1. male	10673	93%
			2. female	763	7%
Target variable	Routechoice	X_{10}	1. route 1	662	6%
			2. route 2	3951	35%
			3. route 3	1298	11%
			4. route 4	1285	11%
			5. route 5	3005	26%
			6. route 6	606	5%
			7. route 7	629	6%

We extract the features of the aforementioned variables by considering sample proportion. After careful work for the test sample, we acquire discrete features for these variables in Tab. 2. Besides, we provide the variable classification and feature-related sample proportion as shown in tab.

The feature occupancies out of all the samples in the above Tab. 2 are reasonable, because their values are not far from real-world data. As to the taxi in-vehicle state, our research field is right in the downtown, which leads to a high heavy occupancy (74%) compared to the whole city. Some other feature occupancies of the indicators are relatively low, such as the occupancies of severe weather, peak hour, female drivers, which are reasonable. The feature distributions of remaining indicators are well-proportioned, such as travel time, start and end speeds, driving years, driving ages.

When we analyze the variable features of the same route-choice samples, we find these samples could correspond to lots of features for the same variable. For instance, as to the samples of route 2, the four travel-time features occupy 12.3%, 17%, 12.8% and 57.9% respectively. This phenomenon demonstrates that we should not only use sole or few factors to do the route restoration problem of GPS trajectories. We should use lots of comprehensive factors to improve the accuracy of this work.

8 RESULTS OF BAYESIAN NETWORK MODEL

In order to test the validity of Bayesian network structure model, the sample data are divided into training set and test set according to the proportion of 3:1. The training set and the test set contained 11436 and 3812 samples separately. In this paper, we use the expert knowledge combined with the K2 algorithm to set up the Bayesian network structure. First, the expert knowledge and the correlation ranking are used for the training data to determine the order of the variable nodes: weather, driver age, driver gender, time period, ending section speed, driving years, taxi in-vehicle state, starting section speed, travel time and route choice. Second, we set 4 as the maximum number of parent nodes for each node. Third, we use the Bayesian Networks Toolbox (BNT) in the software MATLAB to complete the programming of the K_2 algorithm, and acquire a Bayesian network structure as shown in Fig. 3.

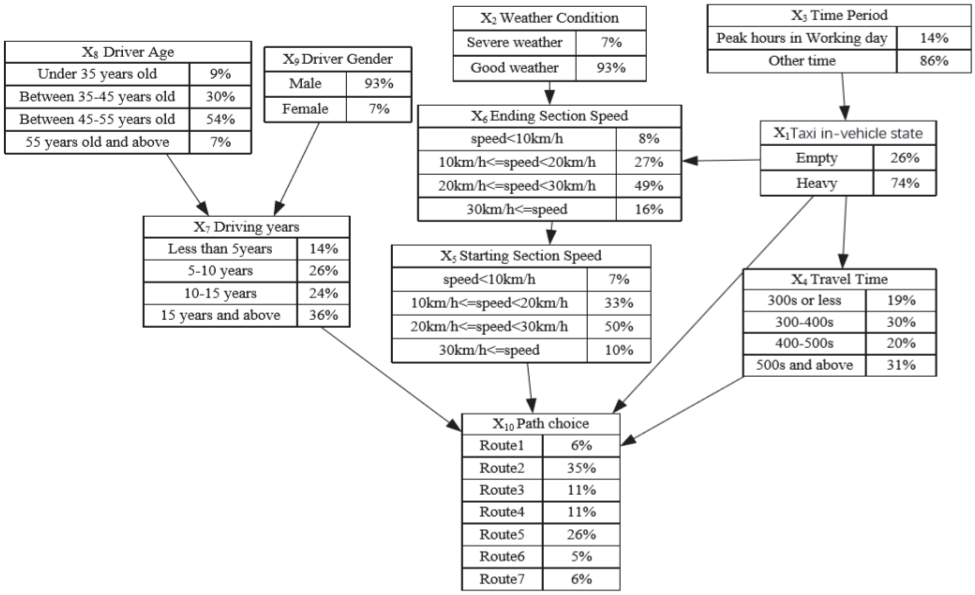


Figure 3 Result of Bayesian Network

9 TEST OF MODEL VALIDITY

9.1 Accuracy Test

According to the established Bayesian network and conditional probability table set, we used the test set to restore the route choice. According to the result, we calculated the recall and precision ratios. The confusion matrix of route identification is shown in the left of Tab. 3. Totally, the accuracy of the test set achieves 91.4%. The ratios of the recall and precision are both higher than 87%. As to route 2, these ratios even achieve 93%. It can be confirmed that the path-restoration accuracy of the model is high and reasonable.

As a comparison, we established the Multinomial Logistic Regression (MLR) model with the same factors to match the route of spare GPS trajectories. We utilized the Statistical Product and Service Solutions (SPSS) statistical software to calibrate the parameters of MLR by using the maximum likelihood estimation method. The confusion matrix of route identification is shown in the right of Tab. 3. The overall restoration accuracy of MLR model is 72%, which is lower than that of the Bayesian network. Our model takes into account the interaction of each factor, which could improve the overall restoration effectiveness. Moreover, seeing from these routes, we find the ratios of recall and precision of Bayesian network are superior to those of MLR (except for the precision ratio of route 3). It indicates that the Bayesian network is suitable for different samples and the restoration capability is more stable than the MLR model. In summary, the Bayesian network performs better at the route identification and distinction.

Table 3 Accuracy Comparison Between Bayesian Network and MLR Model

		Confusion matrix for route identification of Bayesian network							Recall ratio / %	Confusion matrix for route identification of MLR model							Recall ratio / %
		R1	R2	R3	R4	R5	R6	R7		R1	R2	R3	R4	R5	R6	R7	
Real choice route	R1	192	1	1	2	1	0	0	97	82	98	4	13	0	0	0	42
	R2	6	1211	22	30	70	6	3	90	32	1215	14	52	1	10	24	90
	R3	3	5	389	5	3	2	2	95	33	239	46	79	3	8	1	11
	R4	4	13	7	384	4	2	3	92	22	134	6	209	30	15	1	50
	R5	9	60	13	20	922	3	2	90	10	14	0	44	916	43	2	89
	R6	2	10	0	2	4	209	3	91	1	24	4	2	78	120	1	52
	R7	0	0	0	0	2	1	179	98	1	2	3	3	1	1	171	94
Precision ratio / %		89	93	90	87	92	94	93	91.4	45	70	60	52	89	61	86	71.7

9.2 ROC Curve

We calculate the TPR and FPR in terms of each threshold. We first draw the seven ROC curves respectively and then obtain the comprehensive ROC curve by averaging the six ROC curves. The AUC value of the comprehensive ROC is 0.8445. According to the evaluation standard of ROC curve, the accuracy of the Bayesian network performs well.

As shown in Fig. 4, the curve of route 7 is the closest to the upper left corner, so the prediction accuracy of this route is pretty good. The routes covered by it all have a certain circumambulation, so their travel time is significantly longer than the other 6 routes. Then the routes 1 and 6 are followed. When we take the rush hour data to do the analysis, we find that most of the drivers choosing route 6 have more than 15-year driving experience. It indicates that experienced drivers are prone to choose route 6 to avoid severe congestion in this area. As to route 1, the corresponding drivers mainly come from the driving experiences of less than 5 years and more than 10 years. As a result, we can infer that considering driving years they could make the prediction of route 1 and 6 more accurately.

Table 4 Sensitivity Analysis of Each Route Selection to Each Factor Feature

Variable	Variable value	Sensitivity /						
		Route1	Route2	Route3	Route4	Route5	Route6	Route7
Taxi in-vehicle state	1	−0.4	−18.6	−0.6	2.9	11.5	3.46	1.75
	2	0.14	6.7	0.3	−1	−4.2	−1.24	−0.63
Weather condition	1	0	0	0	0	0	0	0
	2	0	0	0	0	0	0.02	0
Time period	1	−0.02	−0.9	0	0.2	0.5	0.17	0.08
	2	0	0.1	0	0	−0.1	−0.02	−0.01
Travel time	1	−1.49	−9.3	0.9	−5.76	8.7	5.56	1.44
	2	6.19	−15.1	−1.6	0.4	8.3	2.36	−0.46
	3	−2.23	−10.1	1.6	5.8	7	−2.39	0.26
	4	−3.81	27.4	0.1	−0.4	−18.39	−4.33	−0.63
Starting section speed	1	0.93	0.1	6.6	−6.49	0.2	−2.69	1.4
	2	0.75	0.4	1.1	1.9	0	−3.45	−0.66
	3	−1.15	−2.2	−3.7	4.7	0.3	2.86	−0.78
	4	−0.77	3.9	−4.75	−4.96	−1.3	6.56	1.26
Ending section speed	1	−0.35	−1.7	−0.3	0.2	1.4	0.72	0.02
	2	0.41	−0.7	0.6	0.6	−0.7	−0.34	0.1
	3	−0.07	2.2	−0.4	−0.9	−0.5	−0.29	−0.1
	4	−0.61	7.1	−0.7	−2.33	−2.2	−0.73	−0.48
Driving years	1	6.39	-14.7	5.5	10.4	−8.9	−0.71	2.08
	2	−0.03	−2	−0.8	4.7	1	−2.99	0.08
	3	2.17	4.5	−3.05	−3.24	1.2	−2.12	0.5
	4	−3.87	4.1	0.6	−5.26	1.7	3.96	−1.2
Driving ages	1	3.29	−6.9	1.9	5.9	−3.7	−1.55	1.09
	2	−0.64	−0.1	0.1	0.4	0.3	0.08	−0.17
	3	−0.02	0.6	−0.2	−0.6	0.2	0.04	−0.02
	4	−1.25	4	-0.8	−4.08	1.4	1.28	−0.45
Driver gender	1	0.09	−0.1	0	0.2	−0.1	−0.08	0.03
	2	−1.25	1.4	0.2	−1.7	0.6	1.24	−0.39

Note: A positive value indicates an increase occurs to the previous whole-sample route-choice probability, and a negative value indicates a decrease occurs to the previous whole-sample route-choice probability.

9.3 Sensitivity Analysis

We use the mutual information value to analyse the significance of each factor in the Bayesian network. The mutual information value of each variable with respect to the target variable is calculated. Fig. 5 shows the test results of mutual information. The larger the mutual information value is, the greater the variable is to the model. It can be seen that the ranking of top 5 factor significance is as follows: travel time (0.28 bit), driving years (0.16 bit), taxi in-vehicle state (0.12 bit), starting section speed (0.10 bit).

We carry on Bayesian network reasoning to measure the influence of each factor on the match probability of specific route. The sensitivity is calculated by comparing the match probabilities between the whole samples and the samples of the specific factor feature. As shown in Tab. 4, we can know the sensitivity of each route selection with respect to each factor. Travel time, driving years and starting section speed is sensitive to the target variable. Sensitivity of other variables to the target variable is relatively weak. It is basically consistent with the analysis of aforementioned factor-significance ranking. In addition, the sensitivities of taxi in-vehicle state to route 2 and 5 are relatively high.

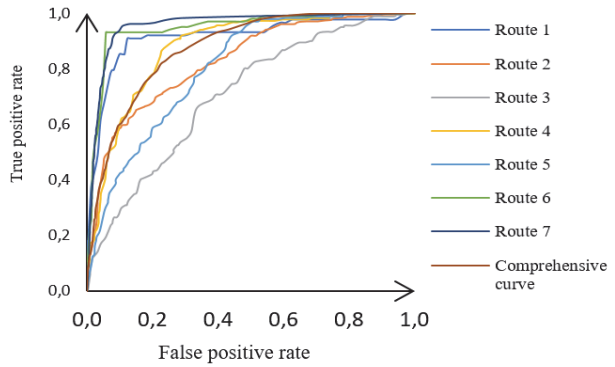


Figure 4 ROC Curve for Bayesian Network6

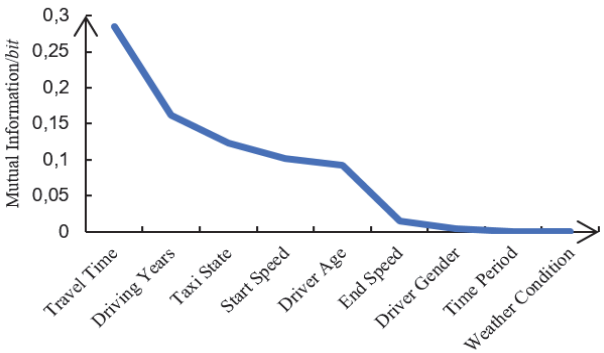


Figure 5 Mutual Information Values of the Factors

9.4 Influence of Trajectory Missing Rate on Model Prediction Accuracy

In order to study the influence of the trajectory missing rate on the restoration accuracy of our model, we take p_3, p_4, p_5 to replace previous destination p_2 separately in Fig. 6. The travel times between origin and destination are around 8 minutes, 12 minutes, 15 minutes respectively. After choosing these data missing rates, the AUC value is 0.825, 0.787 and 0.56 separately (see Fig. 7). When the missing rate approaches 12 minutes from 8 minutes, the decrease of AUC value is not much significant, and its evaluation remains "good". It indicates that our method can be used for the data missing rate of 5 minutes or above.

Figure 6 A Change of End Location to Enlarge Missing Rate

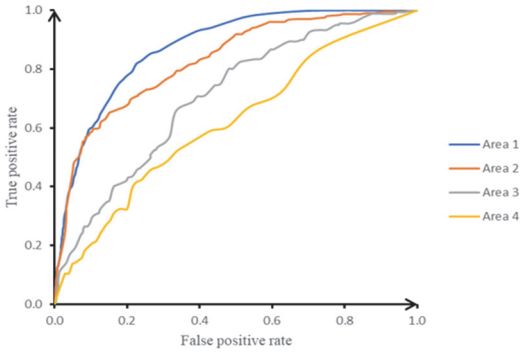


Figure7 ROC Curves of Different Data Missing Rate

10 CONCLUSION

We propose a route restoration method with respect to sparse GPS trajectory. This Bayesian network model takes lots of route-choice factors into account. This multi-factor feature is our innovation when compared to other route restoration methods. We use the taxi GPS data in Ningbo City of China to test our model. It is found that these factors can really influence the route restoration and the accuracy of our model performs better than MLR model.

1) We used the K2 algorithm to establish the Bayesian network structure based on the data of training set. The total accuracy is more than 90%. It also shows that the driver's characteristics (especially the driving years) are highly correlated with the driver's route selection.

2) Via analysis of the ROC curve of the prediction results it is indicated that our method can be used for the data missing rate of 5 minutes or above.

In general, the proposed method shows its ability and great potential for route restoration using the low-quality GPS trajectory data. However, the case employed in this study is a simple one. In fact, the roads of several denser areas are complicated in the Ningbo dataset, some branchroad segments are not well connected but there are GPS trajectories in these roads. More tests should be conducted to generate candidate routes for choice. Hence, replacing this case with a more sophisticated case is expected to lead to further improvements in future studies.

Acknowledgment

This work was supported in part by the Ningbo Natural Science Foundation of China under Grant 2019A610040, in part by the Project of Ningbo Transportation Technology under Grant HS2020000169, and in part by K. C. Wong Magna Fund in Ningbo University.

11 REFERENCE

[1] Yuan, N. J., Zheng, Y., Xie, X., Wang, Y. Z., Zheng, K., & Xiong, H. (2015). Discovering urban functional zones using latent activity trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 27(3), 712-725. https://doi.org/10.1109/TKDE.2014.2345405

[2] Chen, M., Liu, Y., & Yu, X. (2015). Predicting Next Locations with Object Clustering and Trajectory Clustering. *Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer International Publishing*. https://doi.org/10.1007/978-3-319-18032-8_27

[3] Mori, U., Mendiburu, A., Maite, Á., Lozano, J. A. (2015). A review of travel time estimation and forecasting for Advanced Traveller Information Systems. *Transportmetrica*, *11*(2), 119-157. https://doi.org/10.1080/23249935.2014.932469

[4] Lv, L., Chen, M., Liu, Y., & Yu, X. (2015). A plane moving average algorithm for short-term traffic flow prediction. *Advances in Knowledge Discovery and Data Mining. Springer International Publishing*. https://doi.org/10.1007/978-3-319-18032-8_28

[5] Yuan, J., Zheng, Y., Zhang, C., Xie, X., & Sun, G. (2010). An interactive-voting based map matching algorithm. *Eleventh International Conference on Mobile Data Management, MDM 2010, Kanas City, Missouri, USA,23-26 May 2010*. https://doi.org/10.1109/MDM.2010.14

[6] Lou, Y., Zhang, C., Zheng, Y., Wang, W., & Huang, Y. (2009). Map-Matching for low-sampling-rate GPS trajectories. *In: Proc. of the ACM-GIS*, 352-361. https://doi.org/10.1145/1653771.1653820

[7] Bierlaire, M., Chen, J., & Newman, J. (2013) A probabilistic map matching method for smartphone GPS data. *Transportation Research Part C Emerging Technologies*, *26*(1), 78-98. https://doi.org/10.1016/j.trc.2012.08.001

[8] Bozhao, L., Cai, Z., Kang, M., Su, S., Zhang, S., Jiang, L., & Ge, Y. (2020). A trajectory restoration algorithm for low-sampling-rate floating car data and complex urban road networks. *International Journal of Geographical Information Science*. 1-24. https://doi.org/10.1080/13658816.2020.1825721

[9] Brakatsoulas, S., Pfoser, D., Salas, R., & Wenk, C. (2005). On map-matching vehicle tracking data. *Proceedings of the 31st International Conference on Very Large Data Bases, Trondheim, Norway, August 30-September 2*.

[10] Gao, W. C., Li, G. L., & Ta, N. (2018). Survey of map matching algorithms. *Journal of Software*, *29*(02), 225-250.

[11] Ozdemir, E., Topcu, A. E., & Ozdemir, M. K. (2016). A hybrid HMM model for travel path inference with sparse GPS samples. *Transportation*. https://doi.org/10.1007/s11116-016-9734-2

[12] Taguchi, S., Koide, S., & Yoshimura, T. (2018). Online Map Matching With Route Prediction. *IEEE Transactions on Intelligent Transportation Systems*., 1-10. https://doi.org/10.1109/TITS.2018.2812147

[13] Wang, L. F., Chen, H., Li, Y., & Deng, Y. J. (2014). Track patching method for incomplete track in track-oriented traffic survey and analysis. *Application Research of Computers*, *31*(1), 162-165.

[14] Zhao, S. D. & Zhang, K. L. (2019). A distributionally robust optimization approach to reconstructing missing locations and paths using high-frequency trajectory data. *Transportation Research Part C: Emerging Technologies*, *102*, 316-335. https://doi.org/10.1016/j.trc.2019.03.012

[15] Fu, Z. L., Yang, Y. W., Gao, X. J., Zhao, X. Y., & Fan, L. (2016). An Optimization algorithm for multi-characteristics road network matching. *Acta Geodaetica et Cartographica Sinica*, *45*(05), 608-615.

[16] Wang, Q. H., Wang, W. Z., Li, D., & Chang, L. (2013). The Application of Electronic Map in Wireless Positioning System. *Advanced Materials Research*, 756-759, 2345-2349. https://doi.org/10.4028/www.scientific.net/AMR.756-759.2345

[17] Thiagarajan, A., Ravindranath, L., Balakrishnan, H., Madden, S., & Girod, L. (2011). Accurate, low-energy trajectory mapping for mobile devices. *In: Proc. of the USENIX*, 267-280.

[18] Guha, S., Plarre, K., Lissner, D., Mitra, S., Krishna, B., & Dutta, P. (2012). Auto Witness: Locating and tracking stolen property while tolerating GPS and radio outages. *Acm Transactions on Sensor Networks*, *8*(4), 1-28. https://doi.org/10.1145/2240116.2240120

[19] Aly, H. & Youssef, M. (2015). semMatch: road semantics-based accurate map matching for challenging positioning Data. *Journal of Marketing Research*, *51*(5), 1-10. https://doi.org/10.1145/2820783.2820824

[20] Cooper, G. F. & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, *9*(4), 309-347. https://doi.org/10.1007/BF00994110

[21] Li, S. H. & Zhang, J. (2015). Review of Bayesian networks structure learning. *Application Research of Computers*, *32*(03), 641-646.

[22] Zhang, L. W. & Guo, H. P. (2016). Introduction to Bayesian Networks. *Beijing: Science Press*.

[23] Xiao, G. N., Juan, Z. C., & Zhang, C. Q. (2017). Travel Mode Detection Based on GPS Track Data and Bayesian Network. *Statistics & Decision*, *2017*(06), 75-79.

[24] Marcot, B. G., Steventon, J. D., Sutherland, G. D., & Mccann, R. K. (2006). Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation. *Canadian Journal of Forest Research*, *36*(12), 3063-3074. https://doi.org/10.1139/x06-135

[25] Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, *29*(3), 241-288. https://doi.org/10.1016/0004-3702(86)90072-X

[26] Han, Y., Fan, S., Zhou, L., & Wang, S. K. (2019). Exploring the temporal and spatial distribution of passengers based on taxi trajectory data. *Periodical of Ocean University of China*, *49*(Sup.I), 155-162.

[27] Liu, K., Chou, P. Y., Liu, X. L., Zhang, H. C., Wang, S. H., & Lu, F. (2017). Measuring Traffic Correlations in Urban Road System Using Word Embedding Model. *Acta Geodaetica et Cartographica Sinica*, *46*(12).

[28] Zhao, X. H., Ren, G. C., Chen, C., & Rong, J. (2017). A Review on Driving Behavior under Adverse Weather Conditions. *Journal of Transport Information and Safety*, *35*(05), 70-75+98.

[29] Zhang, W. H., & Li, M. F. (2018). Drivers' Route Choice Behavior under Different Traffic Information Guidance. *Journal of Chongqing Jiao tong University* (*Natural Science*), *37*(10), 86-93.

Contact information:

Guangyao LI, postgraduate
1) Ningbo University,
Ningbo City, Zhejiang Province, China
2) Jiangsu Province Collaborative Innovation Center for Modern Urban Traffic Technologies,
Nanjing City, Jiangsu Province, China
E-mail: 1518432988@qq.com

Zhengfeng HUANG, adjunct professor
(Corresponding author)
1) Ningbo University,
Ningbo City, Zhejiang Province, China
2) Jiangsu Province Collaborative Innovation Center for Modern Urban Traffic Technologies,
Nanjing City, Jiangsu Province, China
E-mail: huangzhengfeng@nbu.edu.cn

Leyi LOU, teaching assistant
Zhejiang Sci-Tech University,
Hangzhou City, Zhejiang Province, China
E-mail: lly_spring@yeah.net

Pengjun ZHENG, Professor
1) Ningbo University,
Ningbo City, Zhejiang Province, China
2) Jiangsu Province Collaborative Innovation Center for Modern Urban Traffic Technologies,
Nanjing City, Jiangsu Province, China
E-mail: zhengpengjun@nbu.edu.cn