ADVANCES IN TRANSPORTATION STUDIES
An International Journal

Special Editors: Honglu Liu & Xuesong Feng

2017 Special Issue, Vol. 3

Contents

Advances in Transportation Studies an international Journal 2017 Special Issue, Vol. 3

ISBN 978-88-255-1082-9– ISSN 1824-5463-17006 – DOI 10.4399/97888255108291 – pag. 3-12

A model for matching the route of expressway vehicles with toll collection data

Z. Liu M. Sang

Transport Information Centre - China Academy of Transportation Sciences,
Hui Xin Li 240, 100029, Beijing, China
email: liuzhenhua@catsic.com

Abstract

After 30 years of rapid development, the total mileage of China's expressways has exceeded 130 thousand kilometres, and large amounts of data are stored in the toll collection system. China's expressway toll collection is implemented in the provincial transportation network. The routes crossing provinces are divided into several records in the toll collection system. License plates are the unique identifier of a vehicle used for matching routes. However, records of license plates are not good enough, so the route matching requires some other useful auxiliary information. A fuzzy matching model based on Bayesian rules is built accordingly. Bayesian matching probability is based on license plate similarity and considers the auxiliary information. The model is of high precision and effectiveness. It is valuable in expressway toll collection data analysis using big data technology.

Keywords – transportation, fuzzy matching, Bayesian rule, toll collection data

1. Introduction

The first expressway in China was put to use in 1988. After more than 20 years of development, China's expressways have exceeded 130 thousand kilometres, which can be seen in Figure 1. Large amounts of data are stored in the toll collection system. There are four toll collection modes, including average toll collection, open toll collection, closed toll collection and mixed type toll collection. China's expressways mainly use the closed toll collection mode. In this mode, the roads are all closed, and the user should pay tolls according to the travelling distance on the expressway.

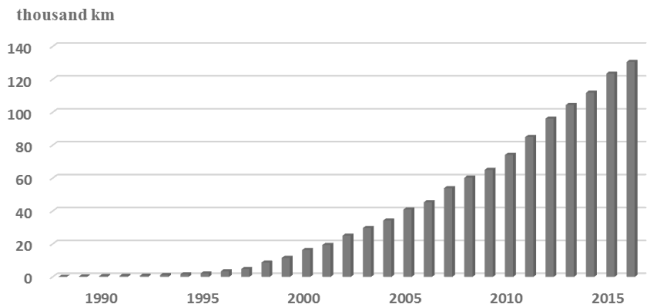

Fig. 1 - China's expressway mileage development

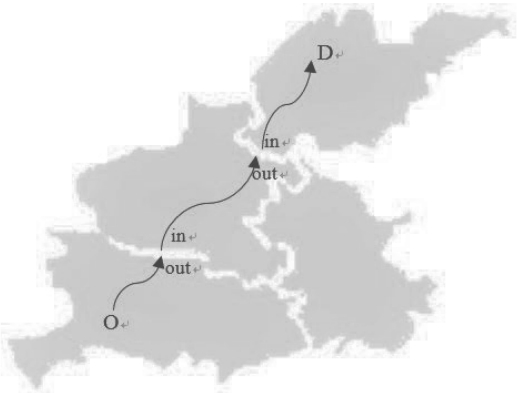

Fig. 2 - Cross-province toll collection

At the early stage of expressway development, the transportation network characteristics were not obvious, and there were many unfinished roads. The toll was collected within a road or within several roads.

With the rapid growth of expressways, the network characteristics are becoming more and more obvious. Toll collection within roads is very inconvenient for expressway users, and it also has a high labour demand. Since 2000, some provinces have begun to implement toll collection within provinces or regions. By the end of 2007, most provinces in China have implemented toll collection within provinces or regions. However, inter-provincial network toll collection is still under construction. The cross-province vehicles will create several records in the toll collection systems, so their routes will be cut into several sections as shown in Figure 2. Therefore, it is necessary to study the method of matching the routes of cross-province vehicles.

2. Features of expressway toll collection data

Most of the toll collection systems in different provinces have information about location, time, vehicle, payment and other data. Location information includes entering station, leaving station, and travelling mileage. Time information includes entering date and time and leaving date and time. Vehicle information includes license plate, vehicle type (toll type), vehicle kind (passenger vehicle or freight vehicle), axle number, and total weight. Payment information includes toll and means of payment.

The license plate is the unique identifier of a vehicle. Therefore, in theory we can judge whether the two records belong to one vehicle according to the license plate, entering station, leaving station, entering time, and leaving time. However, the quality of license plate data is not good enough. Some of the license plates are incomplete, some are missing, and some are partly wrong.

We analyse the completeness of license plate data for different provinces. Generally, there are 7 characters in a complete license plate in China, so we regard the data as complete if the license plate has 7 characters. In half of the provinces, below 50% of the license plate data are complete. For example, in Shanxi province, only 7.7% of the recorded license plates are complete, and in Liaoning Province, only 41.7% of the license plates are complete; 55.4% of the license plates have fewer than 4 characters.

The quality of the license plate data is a great obstacle in route matching. Therefore, it is necessary to find a feasible method to solve this problem.

3. Route matching model

3.1. Summary of the method

The license plate is the key piece of information for vehicle identification, so we first determine the license plate similarity of two vehicles when matching. If we cannot draw a conclusion just using the license plate, we should look for some other information. When we use other information in judging, we can introduce the method of 'Bayesian inferences'.

In Bayesian inferences, priori knowledge, the past sample information, and the current data are combined to obtain the probability of an event. This method can accurately describe the process of human judging, and it is widely used in many fields. [7, 8] Ji Hong worked out an improved weighted Bayesian model in his essay *Research on Improved Bayesian Algorithm for Anti-Spam Filtering,* and with this model the identification accuracy of junk mails was significantly heightened [4]. The essay *Thunderstorm prediction based on Bayesian classification method* combined the Bayesian classification and Fisher criteria methods to predict a storm's behaviour in the short term [3].

By analysing the sample data, we can see that the distribution characteristics of some indices of matched vehicles are quite different from those of unmatched vehicles.

The statistical characteristics of matched vehicles and those of unmatched vehicles can make a Bayesian discriminant. The license plate similarity and Bayesian discriminant will be used to make the decision.

3.2. Procedure and technical route of matching

3.2.1. General idea

In vehicle matching, due to the significant differences between the available attributes of passenger cars and freight cars, it is necessary to select indices other than license plate for the freight vehicles and passenger vehicles respectively.

The directly matched records with the license plate are regarded as samples from the population, and the statistical characteristics of the matched vehicles are obtained through an analysis of the samples. Together with the statistical characteristics of unmatched vehicles, a Bayesian classifier for vehicle features is constructed.

The Bayesian matching probability can be determined according to the index scores of the vehicles using the Bayesian formula.

Combining the license plate similarity and the Bayesian matching probability, we can get the final matching scores. We choose the largest score as the matched one.

3.2.2. Procedure

After getting the data of A province toll station and B province toll station, the records with complete and identical license plates within a reasonable time range of T are matched. The next step is to write the matched records to a new file, and write the unmatched records back to the original file.

Based on data analysis of matched records, the distribution characteristics of matched vehicle attributes are obtained. The maximum possible time difference between leaving time and entering time of matched vehicles can be marked as t.

The next step is calculating the distribution of attribute indices of unmatched vehicles.

For each entering vehicle, based on its entering time, all leaving vehicles in t time difference range are selected.

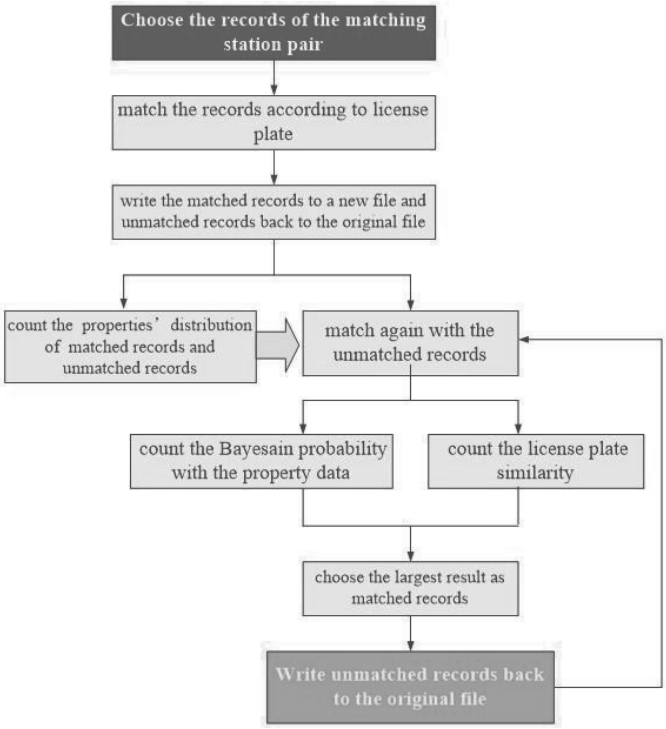

Fig. 3 - Procedure and technical route of the matching process

The Bayesian matching probability P of this entering vehicle and each of the selected vehicles is calculated.

Then, the sum of Bayesian matching probability P and license plate similarity M is calculated. The leaving vehicle with the highest score will be matched to this entering vehicle.

The procedure is shown in Figure 3.

3. 3 Key technologies

3.3.1. Bayesian discriminant

Thomas Bayes, a British mathematician, proposed Bayesian inference for the first time in 1763. It is based on subjective judgement. It can start from an inaccurate estimate and keep modifying the result according to prior knowledge, the previous sample information, and the current sample information to obtain a final accurate estimate [2][5].

a) Basic Bayesian equation

The basic Bayesian equation is shown in Equation (1).

$$P(A|B) = P(A)P(B|A)/P(B) \tag{1}$$

In Equation (1) P (A) is called prior probability. P (A|B) is called posterior probability. P (B|A) / P (B) is called probability function. It is an adjustment factor which can make the probability prediction closer to the reality.

b) Matching probability equation based on one index

We use S as a vehicles matching event and use H as a vehicles not matching event. These two events are mutually exclusive. According to total probability equation we can get equation (2).

$$P(B)=P(B|S)P(S)+P(B|H)P(H) \tag{2}$$

According to Equation (1) and Equation (2) we can get equation (3).

$$P(A|B)=P(A)P(B|A)/\left(P(B|S)P(S)+P(B|H)P(H)\right) \tag{3}$$

If we use W as the value of one index, use P (S|W) as the matching probability of W, use P (W|S) as the probability of W under the condition of matching, use P (W|H) as the probability of W under the condition of not matching, Equation (3) can be transformed to Equation (4).

$$P(S|W)=P(W|S)P(S)/\left(P(W|S)P(S)+P(W|H)P(H)\right) \tag{4}$$

If there are not any preconditions about this issue both P (S) and P (H) will be 50%. Thus, Equation (4) can be transformed to Equation (5).

$$P(S|W)=P(W|S)/\left(P(W|S)+P(W|H)\right) \tag{5}$$

c) Matching probability equation based on several indices

If there are two indices of W_1 and W_2 there are two results. One result is they are matched which we call event E_1. The other result is they are not matched which we call event E_2. The combination of two events and the probabilities are shown in Table 1 and Table 2.

If event A, event B and event C are independent，the probability of the three events occurring together is as shown in Equation (6).

$$P(ABC)=P(A)P(B)P(C) \tag{6}$$

Supposing all the events are independent, we can get the following equations.

$$P(E_1)=P(S|W_1)P(S|W_2)P(S) \tag{7}$$

$$P(E_2)=P(H|W_1)P(H|W_2)P(H) \tag{8}$$

According to Bayesian equation when W_1 and W_2 appear，the probability of matching is as follows.

$$P=P(E_1)/\left(P(E_1)+P(E_2)\right) \tag{9}$$

According to Equation (7), Equation (8), and Equation (9), we can get Equation (10).

$$P=P(S|W_1)P(S|W_2)P(S)/\left(P(S|W_1)P(S|W_2)P(S)+P(H|W_1)P(H|W_2)P(H)\right) \tag{10}$$

Tab. 1 - Event combination with two indices

Event	W_1	W_2	Result
E_1	Appear	Appear	Matched
E_2	Appear	Appear	Not Matched

Tab. 2 - Matching probabilities with two indices

Event	Probability of Event W_1	Probability of Event W_2	Matching Probability
E_1	P (S|W_1)	P (S|W_2)	P (S)
E_2	P (H|W_1)	P (H|W_2)	P (H)

If there are not any preconditions, both P (S) and P (H) will be 50%. Equation (10) can be transformed to Equation (11).

$$P = P(S|W_1)P(S|W_2)/\left(P(S|W_1)P(S|W_2)+P(H|W_1)P(H|W_2)\right) \quad (11)$$

3.3.2. How to choose the indices

We should determine which indices will be used in the process because there are many indices in the toll collection system. Different combinations of indices in Bayesian decision will have different effects. We can make some sample data and use these data to choose indices. The combination with the best matching effect will be selected.

In our research we choose three indices of 'time difference between in and out', 'total weight difference between in and out' and 'axle number difference between in and out' for freight vehicles. The matching probability equation with these indices is as Equation (12) shows.

$$P = P(S|W_1)P(S|W_2)P(S|W_3)/\left(P(S|W_1)P(S|W_2)P(S|W_3)+P(H|W_1)P(H|W_2)P(H|W_3)\right) \quad (12)$$

In Equation(12) W_1 is time difference between in and out, W_2 is axle number difference between in and out, W_3 is total weight difference between in and out, P ($S|W_1$) is the probability of W_1 in one time range of matching vehicles, P ($S|W_2$) is the probability of W_2 in one quantity range of matching vehicles, P ($S|W_3$) is the probability of W_3 in one quantity range of matching vehicles, P ($H|W_1$) is the probability of W_1 in one time range of not matching vehicles, P ($H|W_2$) is the probability of W_2 in one quantity range of not matching vehicles, P ($H|W_3$) is the probability of W_3 in one quantity range of not matching vehicles.

We choose two indices, namely, 'time difference between in and out', and 'vehicle type difference between in and out' for passenger vehicles. There is no information about total weight for passenger vehicles because the toll rate of passenger vehicles is determined by the type of vehicle, not the total weight.

3.3.3. License plate similarity algorithm

There is a license plate recognizing sub-system in the toll collection system. The license plate data are usually recognized with image processing techniques [1].

The main functions of the license plate recognizing system include collecting the license plate image, pre-processing the image, confirming the location of the image, identifying the characters of the license plate, and recognizing the characters [13].

Characters of high pixel coincidence degree are easily confused. We can see from Figure 4 and Figure 5 the probabilities of confusion with regards to '1' and '7', 'E' and 'F', and '0' and 'D' are very high in the two toll stations [6].

There are many algorithms related to string similarity, such as edit distance algorithm, longest common string algorithm, and Heckel's algorithm. Edit distance algorithm is a commonly used orderly matching algorithm [9]. The edit distance is proposed by Vladimir Levenshtein, a Russian scientist. It is the minimum number of editing operations required from one string to the other. The operation may include replacing a character with another character, inserting a character, and deleting a character. In general, the smaller the edit distance, the higher the similarity of the two strings [10][11].

The confusion probability of two characters can be obtained from the sample data. We can find all the confused characters pairs by comparison and record them, then count the number of times each confused character pair appears, and finally calculate the probability of confusion [12].

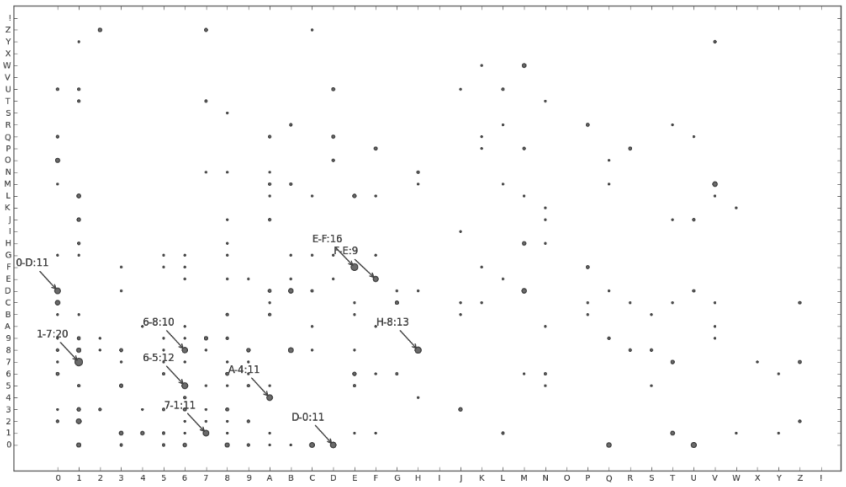

Fig. 4 - Character confusion distribution of toll station A

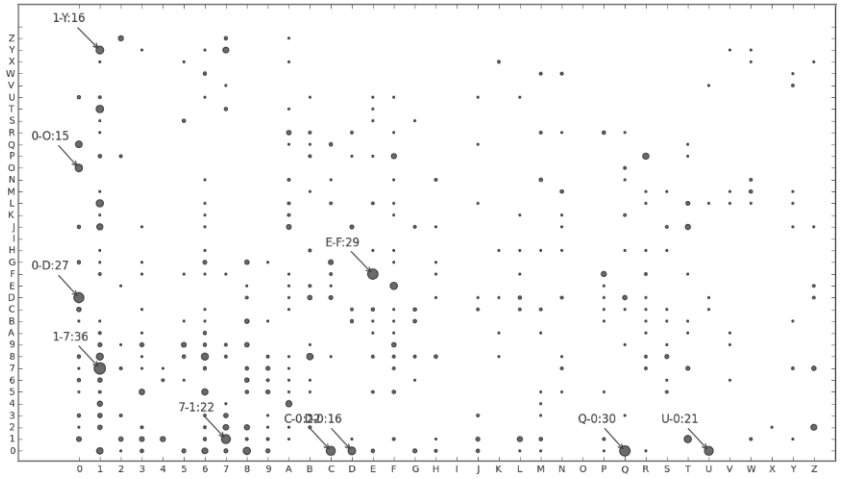

Fig. 5 - Character confusion distribution of toll station B

3.3.4. Weight of license plate similarity and Bayesian discriminant probability

We mark the license plate similarity as M and Bayesian discriminant probability as P. They are both meaningful. We need to combine them and get the final matching result. The weighted average method can be used as shown in Equation (13).

$$F = k_1 M + k_2 P \tag{13}$$

In Equation (13) F is the final matching result, M is license plate similarity, P is Bayesian discriminant probability, k_1 is the weight of M, and k_2 is the weight of P. $k_1 + k_2 = 1$.

There are some methods to obtain the weights of different indices. In this case, we use an experimental method based on sample learning. It is easy to carry out and the effect is good.

3.3.5. Reasonable time difference

It is very important to choose the appropriate time difference between in and out that can greatly reduce the computational complexity and improve the accuracy of the calculation.

The time difference window can be achieved by analysing the matched samples. Set the time difference between in and out from large to small and we can get lower quartile Q1, median, upper quartile Q3 and maximum value. IQR is calculated by subtracting Q1 from Q3. According to this experience, the abnormal value is defined as the value greater than Q3+1.5IQR or smaller than Q1-1.5IQR. (Q1-1.5IQR, Q3+1.5IQR) is the maximum possible range of time difference.

First, we need to collect enough sample data. Then, we increase the value of k from 0.1 to 0.9. We will get the results of matching with the combinations of k1 and k2. Finally, we choose the pair of k1 and k2 which will lead to the best matching effect.

4. Application of the method

4.1. Effect of matching

We used the data of Yu'e station in Henan province and Ebei station in Hubei province to test the feasibility and effect of this method. The two stations are adjacent on Jing-Gang-Ao expressway and near the border of each province. The distance between two stations is 26.2 kilometres.

The quality of the two stations' license plate data is quite different as Table 3 shows. The license plate data of Ebei station in Hubei province are of good quality, and the proportions of complete license plates in both directions are more than 96%. However, the license plate data quality of Yu'e station is poor. Especially in the out direction, the percentage of complete license plate data is less than 50%. The routes of passenger vehicles and freight vehicles are matched. Both the results of matching are satisfactory, however the result of freight vehicles' matching is better than the passenger vehicles' because there is more useful information about freight vehicles than passenger vehicles.

Tab. 3 - License plate data quality of the two stations

Station	Complete Proportion	Incomplete Proportion	Null Proportion
Yu'e station(Out)	43.89%	45.98%	10.14%
Ebei station(In)	96.67%	1.70%	1.63%
Ebei station(Out)	98.04%	1.00%	0.98%
Yu'e station(In)	73.12%	14.56%	12.32%

Tab. 4 - Matching effect of freight vehicles

Station	Matched Rate
Yu'e station(Out)	98.7%
Ebei station(In)	
Ebei station(Out)	98.2%
Yu'e station(In)	

Tab. 5 - Matching effect of passenger vehicles

Station	Matched Rate
Yu'e station(Out)	93.1%
Ebei station(In)	
Ebei station(Out)	93.6%
Yu'e station(In)	

Tab. 6 - Freight vehicle flow between four provinces on June 2015

Origin \ Destination	Hunan	Hubei	Henan	Hebei
Hunan	2497812	59966	6793	1673
Hubei	51046	2889365	69812	5853
Henan	11550	69032	3849120	78432
Hebei	2666	5899	73124	4747797

Tab. 7 - Passenger vehicle flow between four provinces on June 2015

Origin \ Destination	Hunan	Hubei	Henan	Hebei
Hunan	12731386	51182	3695	281
Hubei	54536	9639629	77215	2268
Henan	2097	57869	18090790	80670
Hebei	85	1415	84259	15503809

4.2. Application of the matching result

Using this matching method, we can get information of the real routes of vehicles on expressways. These data can be used to do many kinds of analyses with some additional information such as economy, population, environment and travelling resources.

The origin and destination data of passenger vehicles and freight vehicles between four provinces including Hunan, Hubei, Henan and Hebei are shown in Table 6 and Table 7.

We also obtained the origin and destination data between different cities and analysed the transportation connection between cities and travelling features of cities. The transportation connection between cities of the four provinces is shown in Figure 6. We can see that Wuhan city (C4201) is in the core position in cross-province transportation. Nanyang city (C4113), Xiangyang city (C4206), Changsha city (C4301) and Changde city (C4307) are in the secondary core position. There is a strong transportation link between Nanyang city (C4113) and Xiangyang city (C4206), suggesting that they belong to the same economic region.

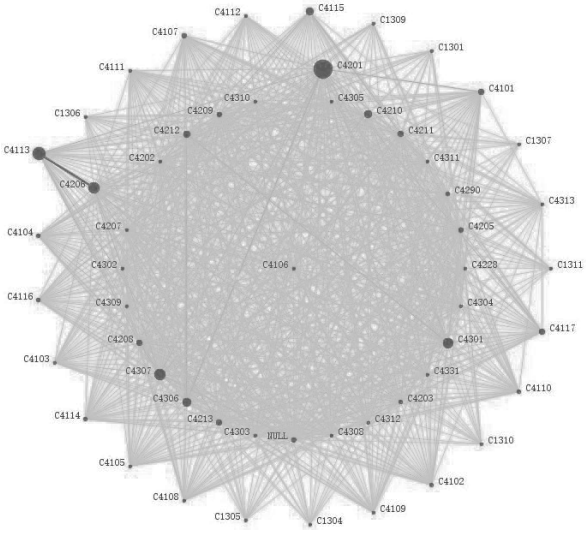

Fig. 6 - Transportation connections between cities