2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)
Windsor Oceanico Hotel, Rio de Janeiro, Brazil, November 1-4, 2016

Predictability Analysis on Expressway Vehicle Mobility Using Electronic Toll Collection Data

Sheng Wan, Ji Meng, Shi Fang, Xingxing Xing, Kunqing Xie and Kaigui Bian

Abstract— **This paper assesses the predictability of individual vehicle's mobility in Beijing expressway system using Electronic Toll Collection (ETC) data records. By examining the uncertainties of movements using entropy, considering both the frequencies and sequential correlations of vehicles' trajectories, we draw to the conclusion that the average limit of predictability of expressway vehicles mobility is 91%. Furthermore, we concluded that the individual property is negatively correlated to its mobility property such as visited station, average travel distance and radius of gyration. Finally, we applied Markov chain (MC) based models to predict the actual accuracy of predictability and found that MC(1) model is adequate to produce a good result and extending the order of the model won't give substantial bonus, thus the first-order Markov property of expressway vehicles' trajectory is proved. Our findings indicate that individual vehicles' mobility in expressway system is far from random and highly dependent on its historical trajectory.**

I. INTRODUCTION

With the application of electronic toll collection (ETC) system throughout China, the expressway administrators have accumulated a huge amount of raw tolling data, including the entrance time point and information related to the lane and traffic volume. Compared to traditional manual tolling system, which take the traffic flow as a whole and ignore the heterogeneity of individual driver's behavior pattern, ETC system could not only effectively promote the efficiency of tolling and alleviate traffic congestion, but also provide precise identification of each vehicle, thus make possible the research of expressway traffic patterns from the perspective of individual vehicle[1].

Without the access to ETC data with individual identification of each vehicle, previous studies of forecasts of traffic volume focused on the inlet and outlet flow at toll collectors. Zhang et al[2] used the BP algorithm and new combing methods to develop a prediction model for short-term traffic flow. Smith et al developed and tested neural network, historical average, time-series and nonparametric regression models for the freeway traffic flow forecasting problem, and found out nonparametric regression model significantly outperformed the other models[3]. In this paper however, we aim to fill this gap in knowledge of individual vehicle's mobility and evaluate the predictability of individual vehicle mobility using the ETC data from Beijing expressway system.

Previous studies on predictability analysis and forecasts of individual trajectories are important in fields such as mobile computing, epidemic modeling and disaster modeling. Various methods are applied to predict individual mobility patterns such as Markov chain(MC) models [4] [5], neural networks [6], Bayesian networks [7] and finite automation [8]. Lu et al[9] used the mobile phone data and developed a Markov chain based model to predict the travel pattern of 500000 individuals . Refering to the methods of individual mobility predictions mentioned above, we apply it to the traffic fields and make some adjustments.

We first conduct quantitative analysis of time series data of ETC and derive the highest potential accuracy of predictability, termed limit of predictability, defined by the entropy of information of a vehicle's trajectory. The limit of predictability can be calculated by solving a Fano's inequality [10][11][12].

Then we implement a series of Markov chain(MC) models to evaluate the predictive accuracy, and compare the predictive accuracy with its upper limit. We find out that individual vehicle trajectory is highly predictable with ETC data. Moreover, we find that the predictability is affected by factors such as traffic frequency, time point and also locations of toll stations in the network. The highly predictable individual traffic patterns enable development of accurate predictive models for eventually devising practical strategies to mitigate congestion and optimize traffic planning.

The rest of this paper is organized as follows. Section II presents a brief introduction of data sources and their processing. Section III defines the predictability of expressway vehicle and derives prediction algorithm based on Markov chain model. Section IV discusses the experimental results. Concluding remarks are described in Section V.

II. DATA

The ETC traffic data is provided by the project of expressway network operation monitoring based on OBU (on board unit), established by the Intelligent Information Processing Lab of Peking University and related departments of expressway systems in Beijing. The ETC system will record its entrance fee information when a car enters and exit the expressway network. Each record contains the time spot and station of both the origin and destination of a specific obumac, which identifies a unique ETC car. Thus the trajectory M_i of each ETC car i is concatenated as OD (origin-destination) pairs: $M_i = m_1, m_2, \ldots, m_n = o_1, d_1, o_2, \ldots, o_n, d_n$. In this paper however, the origin and destination are treated differently. In fact, since we concern more about the cars on the road, only the destination's predictability is assessed in this paper.

A. Spatial Topology

The expressway of Beijing is composed of a ring circuit of Beijing (the Sixth Ring Road) and eight national expressways radiated from the city center of Beijing to its neighboring

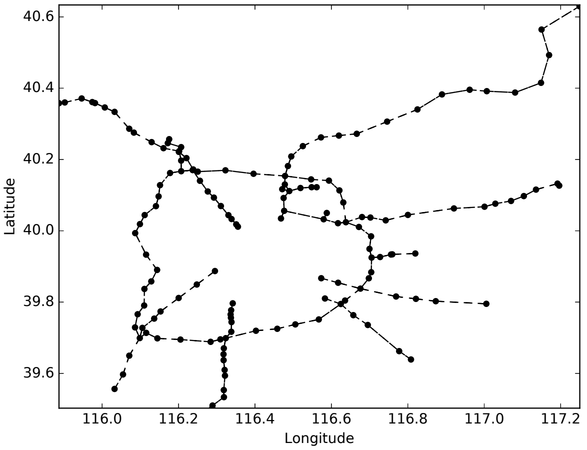


Fig. 1. Topology of Beijing highway network

provinces. The topology of expressway network contains 143 toll stations as presented in Fig. 1.

B. Length of trajectory

We collect over a million trajectories of ETC cars out of over 100 million records from ETC data during June, 2013 to December, 2013, implying approximately 50 records for each ETC car during 214 days, namely 1 entrance per 4.3 days, indicating low travel frequency on expressway. However, the distribution of the length of trajectory of expressway cars is unbalanced as plotted in Fig. 2(a), where the count of ETC cars is exponentially decreasing as the length of trajectory increases. To ensure the stability of behavior pattern of individual vehicles and avoid the illusion of high predictability stemming from vehicles which traveled too little, we exclude individual vehicles who visited less than 50 toll stations and take the rest of them as high quality vehicles. According to this rule, we select from the original sample of 1146890 vehicles and reach a sample size of 309660 vehicles. As plotted in Fig. 2(b), the selected vehicles occupy 27% of all vehicles while contribute 72% of all the ETC records.

C. Time range

To explore the predictability of expressway vehicles, one thing to be concerned is the time range within which the travel pattern of each ETC vehicles could be captured. As Fig. 3(a) and Fig. 3(b) presented, the amount of data of ETC linearly increases over time, while the number of vehicles appears slower growth, suggesting recurrent vehicles in the expressway network (Fig. 3(a)). On the other hand, the total existed OD-records and the average visited stations converges after 7 months, suggesting recurrent OD-record as well as visited stations for each individual vehicle (Fig. 3(b)). Thus it can be easily concluded that the selected time range is capable of capturing the travel pattern of each vehicle. Moreover, the next half year can avoid the famous Chinese Spring Festival travel rush, which would definitely cause exception in traffic flow and individual ETC car's travel pattern, coming at the start of the next year.

III. Methods

A. Measures of mobility.

In order to measure the mobility property of vehicles, we define three properties as: visited station N, average travel distance $\bar{D}$ and radius of gyration r_g for each vehicle. Specifically, let $M_i = m_1, m_2, ..., m_n$be the sequence of recorded station updates for ETC car i during the period of data collection, The average travel distance $\bar{D}$ is defined as:

$$\bar{D}(i) = \frac{2}{n} \sum_{j=1}^{\frac{n}{2}} |m_{2j} - m_{2j-1}| \tag{1}$$

where m_{2j} is the destination of jth record of the vehicle and m_{2j-1} the origin. The radius of gyration r_g is defined as:

$$r_g(i) = \sqrt{\frac{1}{n} \sum_{j=1}^{n} |m_j - \bar{m}|} \tag{2}$$

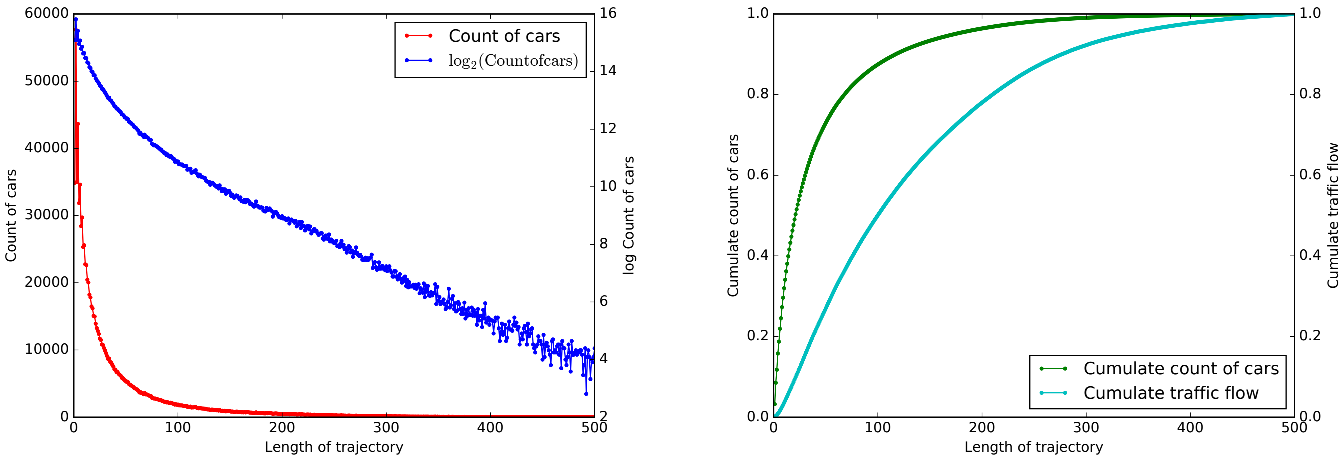
where $|m_j - \bar{m}|$ is the distance between location m_j and the mobility center $\bar{m} = \frac{1}{n} \sum_j^n m_j$.

The radius of gyration is different from the average travel distance. r_g could be smaller than $\bar{D}$ if someone travels in a limited area even if he or she covers a large distance. On the other hand, someone who travels with small steps but in a large circle would lead to a large radius of gyration but smaller average distance. It is noteworthy that the dataset only provides the locations of toll stations each individual vehicle visited, which can be used to approximate the coordinate of each vehicle. This approximation is efficient when comparing mobility between different vehicles despite some imprecision. For example, a vehicle which travels through many toll stations will have a larger $\bar{D}$ and r_g than the vehicle which travels through only one or two toll stations. The frequency distribution of N ,$\bar{D}$ and r_g are plotted in Fig. 4.

B. Measures of predictability

To measure the trajectory's predictability of expressway vehicles, we use entropy and the predictability derived from entropy as the quantity proposed by Song et al.[13] Define $X_i = x_1, x_2, ..., x_n$ as the sequence of stations visited by vehicle i. Entropy is applied here to measure the degree of uncertainty of vehicles' mobility. Larger entropy suggests greater uncertainty and less predictability of a vehicle's trajectory.

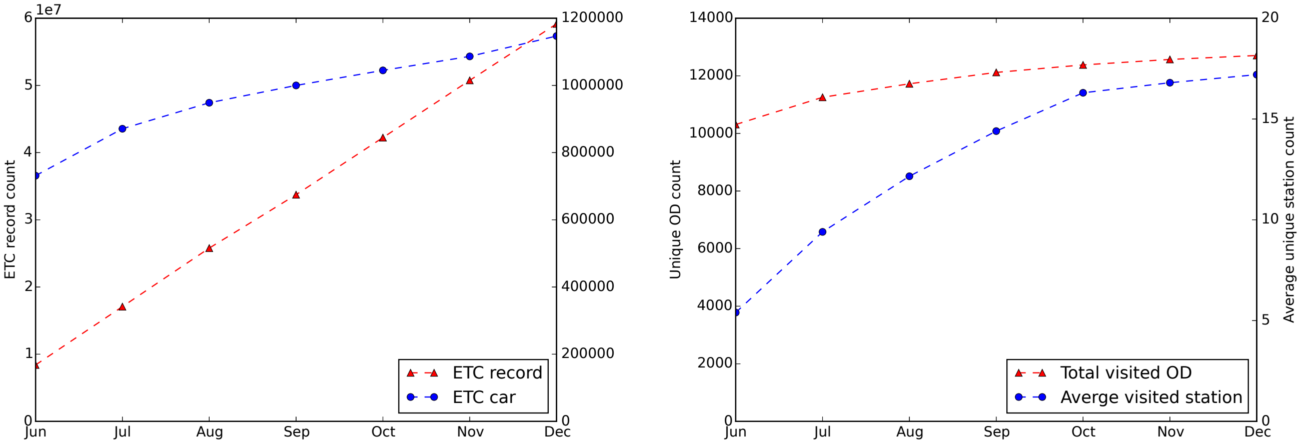
a) Entropy: Three kinds of entropy are provided as follows as uncertainty measures: (i) the random entropy, $S_i^{rand} = \log_2 L_i$ capturing the predictability of each vehicle assuming the station to be visited is uniformly distributed among L_i distinct locations in X_i; (ii) the temporal-uncorrelated entropy, $S_i^{unc} = -\sum_{k=1}^{L_i} p_k \log_2 p_k$,where p_k is the frequency at which the k^{th} distinct station is visited by vehicle i. In addition to S_i^{rand}, S_i^{unc} takes into the



(a) Distribution of length of trajectory

(b) Cumulative distribution of length of trajectory with its traffic flow

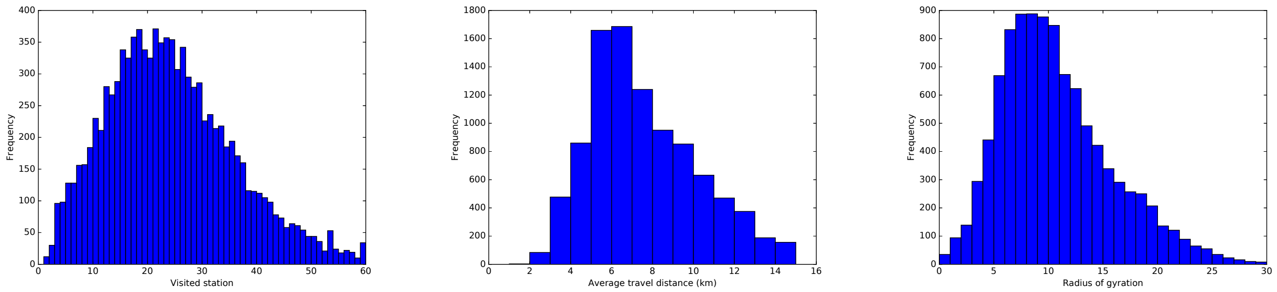
Fig. 2. Distribution and cumulative distribution relating length of trajectory



(a) ETC record linearly increases while unique ETC cars grows slower

(b) Converging counts of ODs and average visited stations

Fig. 3. Recurrent vehicles over time



(a) Distribution of visited station

(b) Distribution of average travel distance

(c) Distribution of radius of gyration

Fig. 4. Distributions of mobility properties

proportion of times each distinct station is visited, decreasing the uncertainty of the trajectory; (iii) the true entropy, $S_i^{real} = -\sum_{X_i' \subset X_i} P(X_i') \log_2 [P(X_i')]$, where $P(X_i')$ is the probability of finding a sub-sequence X_i' in X_i.

Moreover, in order to assess the impact of order n to both predictability and MC models to be presented in the next subsection, we define entropy $S_i^{real}(n)$ of different order as:

$$S_i^{real}(n) = -\sum_{X_i' \in \{X_i'' \subset X_i | \mathrm{length(X_i'')=n}\}} P(X_i') \log_2 [P(X_i')] \tag{3}$$

Where X_i' is constrained to be sub-sequence of length n in X_i and $P(X_i')$ is the probability of finding X_i' in $\{X_i'' \subset X_i | \mathrm{length(X_i'') = n}\}$.

b) Predictability: . Given the entropy E of a vehicle i, the limit of predictability $\Pi_i(E, L_i)$ of it is derived from Fano's inequality:

$$\Pi_i \leq \Pi_i^{\mathrm{Fano}}(E, L_i) \tag{4}$$

where Π_i^{Fano} is given by

$$E = H(\Pi_i^{\mathrm{Fano}}) + (1 - \Pi_i^{\mathrm{Fano}}) \log_2(L_i - 1) \tag{5}$$

and

$$H(\Pi_i^{\mathrm{Fano}}) = -\Pi_i^{\mathrm{Fano}} \log_2(\Pi_i^{\mathrm{Fano}}) - (1 - \Pi_i^{\mathrm{Fano}}) \log_2(1 - \Pi_i^{\mathrm{Fano}}) \tag{6}$$

Thus we have $\Pi_i^{\mathrm{rand}} = \Pi(S_i^{\mathrm{rand}}), \Pi_i^{\mathrm{unc}} = \Pi(S_i^{\mathrm{unc}})$ and finally, $\Pi_i^{\mathrm{max}} = \Pi(S_i^{\mathrm{real}})$.

For entropy of order n, the $\Pi(n)$ share the same equations 5 and 6.

C. Prediction algorithms

To investigate how close we can get to achieving $\Pi^m ax$ with actual prediction algorithms, we implement several variants of Markov chain (MC) based models.

In an MC-based model, the trajectory of each individual is modeled as a Markov chain of order n, which assumes that the movement of individuals between the L_i locations is a process with limited memory in the sense that the future location is visited depending only on the previous n visited location, i.e., $P(X_i^{t+1} = x^{t+1} | X_i^t = x^t, X_i^{t-1} = x^{t-1}, \ldots, X_i^1 = x^1.) = P(X_i^{t+1} = x^{t+1} | X_i^t = x^t, X_i^{t-1} = x^{t-1}, \ldots, X_i^{t-n+1} = x^{t-n+1}.)$, where X_i^t is a random variable representing the location for individual i at time t.

Given the previous n locations $X_i^t = x^t, X_i^{t-1} = x^{t-1}, \ldots, X_i^{t-n+1} = x^{t-n+1}$, the prediction is then determined by the transition matrix, P, choosing the destination location $x^{pre}(1 \leq pre \leq L_i)$ which maximizes the probability:

$$\begin{aligned} & P(X_i^{t+1} = x^{\mathrm{pre}} | X_i^t = x^t, \\ & \quad X_i^{t-1} = x^{t-1}, \ldots, X_i^{t-n+1} = x^{t-n+1}) \\ = \quad & \max_{k=1}^{L_i} P(X_i^{t+1} = x^k | X_i^t = x^t, \\ & \quad X_i^{t-1} = x^{t-1}, \ldots, X_i^{t-n+1} = x^{t-n+1}) \end{aligned} \tag{7}$$

To investigate the predictive power of different order in MC model, we vary n from 1 to 10. One thing to be mentioned is that when n grows larger, the MC(n) model may not be able to match the historical sequence in history. We define the probability of matching the historical sequence of length n as $P(n)$. The test cases that can't match won't be evaluated.

The performance of each model is evaluated by the accuracy, γ, which is the proportion of accurate predictions from all predictions made:

$$\gamma = \frac{\text{correct predictions}}{\text{predictions made}} \tag{8}$$

Using the MC models of different orders, we take the data of previous 6 months as training sample and the data of December 2013 as test sample to evaluate the predictive powers of MC models of different orders.

IV. RESULTS

A. Regularity and potential predictability

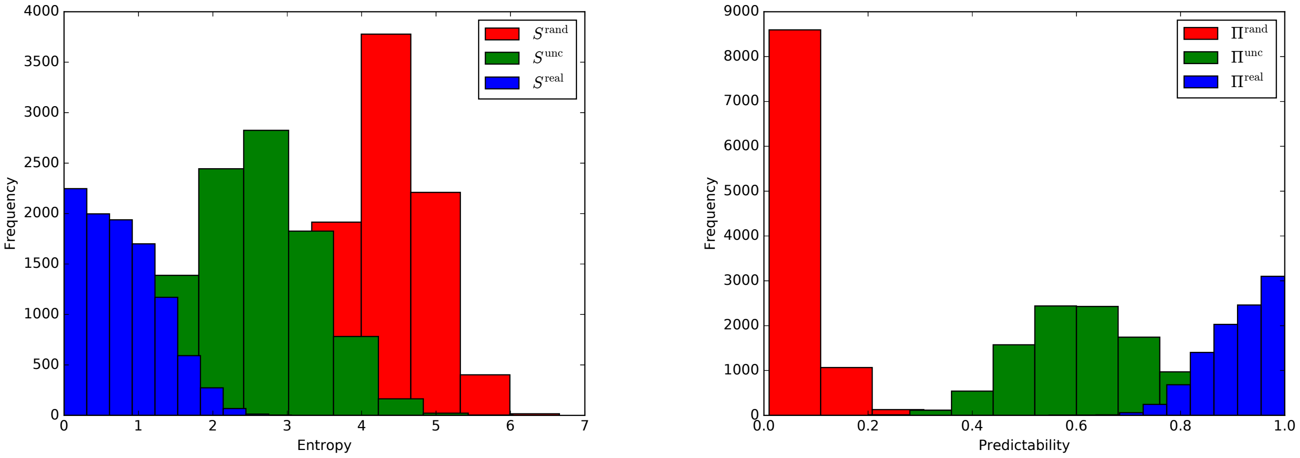
Now we focus on the potential predictability of the trajectories of individual vehicles.

Fig. 5(a) shows the distributions of users random entropy (S^{rand}), temporal uncorrelated entropy (S^{unc}) and true-entropy (S^{real}), which indicates a significant drop of the entropy of visited toll station's location considering both the frequency and sequence order correlation. The median value of S^{rand} is 4.25, suggesting that if we assume that individual vehicles randomly choose a toll station to visit the next day, a typical vehicle could be found in any of $2^{4.25} \approx 19$ locations. On the other hand, the uncertainty in a typical individuals sharply declines to $2^{S_{unc}} = 2^{2.55} \approx 5.85$ and $2^{S_{real}} = 2^{0.73} \approx 1.65$ considering the information contained in the frequency and sequence order of the trajectory.

Fig. 5(b) presents the distribution of the limit of predictability. If we don't consider the information on frequency or sequence order, the average predictability is $\Pi^{\mathrm{rand}} = 0.07$. on the other hand, if we take into account additional information on frequency and temporal correlation, the average predictability increases to $\Pi^{\mathrm{unc}} = 0.62$, and $\Pi^{\mathrm{real}} = 0.91$, respectively.

B. Vehicle mobility factors

Moreover, we examined influence of vehicle mobility to the limit of predictability. Fig. 6 shows the impacts of travel visited station, average travel distance and radius of gyration to three kinds of predictability respectively. Count of visited stations maybe the most influential mobility properties concerning predictability. As Fig. 6(a) presented, all three kinds of predictability decrease considerably as visited stations increases. Both Π^{unc} and Π^{real} both decreases linearly, while Π^{real} appears to be less steep and more robust. Fig. 6(b) shows the increases in average travel distance $\bar{D}$ cause a slight decrease in predictability, with which decreasing from 0.95 to 0.90 when the average distance $\bar{D}$ increases from 1 to 6 kilometers, then stays around 0.88 to 0.90 for $\bar{D} \in [6, 14]$, indicating the expressway vehicles with short travel distance is more predictable. In fact, the longer the distance, the more destinations is possible. In Fig. 6(c), the predictability appears similar trend as that of average distance in Fig. 6(b),



(a)

(b)

Fig. 5. The entropy and predictability distribution of ETC cars.

the two 'jumps' on Π^{rand} and Π^{unc} is smoothed in Π^{real}, implying a group of vehicles with similar radius of gyration has relatively fewer visited stations, while in the case of Π^{real}, it's less affected since it's more robust against visited stations.

C. Predictive accuracy based on Markov-chain models

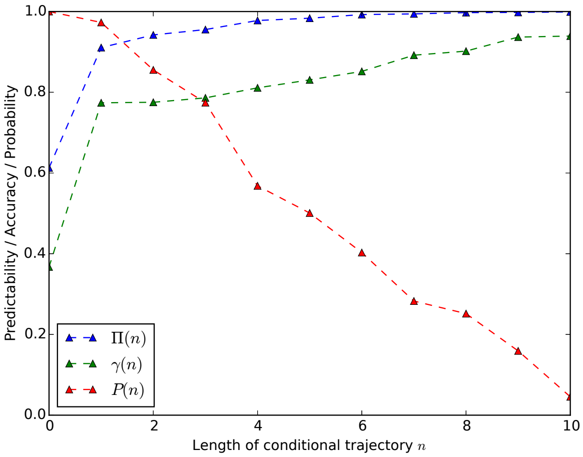


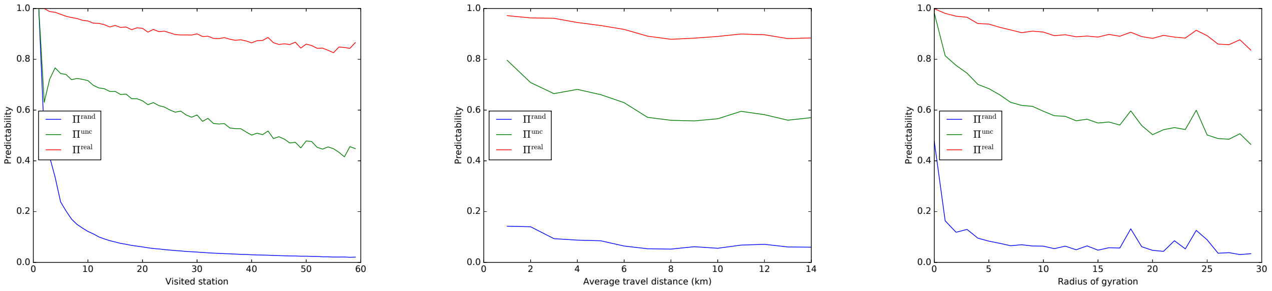
Fig. 7. Limit of predictability Π, predictive accuracy γ and probability to match historical sequence ¶ under order 1 to n

The predictability analysis in the previous section reveals that the theoretical upper bound of predictive accuracy can reach 0.91 considering the information on frequency and sequential correlation of the trajectory. In this section, we focus on the actual predictability accuracy and use MC(n) models to predict the trajectory of individual vehicles on each day. The specific methods are detailed in the Method section.

The accuracy of these models is presented in Fig. 7 and shows increasing accuracy as the order of the MC model grows. The MC-based models, MC(1) to MC(10), achieve substantially higher accuracies than the estimation method without the consideration of sequential correlation between locations, i.e., MC(0). Not surprisingly, the predictive accuracy increase as the sequential orders increase, since longer sequence contains more information on vehicle's behavior pattern. On the other hand, the longer the sequence, the harder it is to match a certain pattern in historical trajectory. Even though the predictive accuracy could approach the upper bound of predictability for MC(10), the matching rate of historical trajectory is substantially lower. From this perspective, we find that higher orders do not bring significant improvement to predictive accuracy. Balancing the predictive accuracy and matching rate, we conclude that MC(1) model is already able to produce a relatively good performance in the mobility prediction of expressway traffic flows, namely, to predict a vehicle's destination, knowing the origin of it is quite enough, no bothering for the previous station visited.

V. CONCLUSIONS

In this paper, we use the ETC data to investigate the predictability of traffic flows in Beijing expressway system. By examining the trajectories of 309660 individual vehicles on expressway during the June 2013 to December 2013, we provide a limit of predictability of 91% for the expressway vehicles destination prediction. Through the analysis of correlations between predictability and expressway vehicles' mobility properties, we found predictability's negative correlation to visited station, average travel distance and radius of gyration. By applying MC-based estimate algorithms, we found that the first order MC model (MC(1)) is able to provide a relatively good prediction accuracy of 77% and extending n to larger may not provide substantial bonus.



(a)

(b)

(c)

Fig. 6. Influence of mobility properties to predictability

To sum up, the ETC data of Beijing expressway system provides us the opportunity to investigate the predictability of expressway traffic flow from the perspective of individual vehicles, which is, as far as we know, the first attempt in this field. Our findings proves the high predictability of expressway vehicle's mobility. The results indicate that vehicles' travel pattern is highly predictable, and that vehicle's destination are substantially dependent on the last visited station, i.e. origin of it. With a good understanding of vehicles' travel patterns, we are able to identify the future driving route, thus provide theoretical and practical support for expressway traffic planning.

References

[1] Komada K, Masukura S, Nagatani T. Traffic flow on a toll expressway with electronic and traditional tollgates[J]. Physica A: Statistical Mechanics and its Applications, 2009, 388(24): 4979-4990.

[2] ZHANG H, Feng S H I. Dynamic equilibrium model of expressway toll collector based on traffic flow prediction[J]. Journal of Transportation Systems Engineering and Information Technology, 2009, 9(5): 71-76.

[3] Smith B L, Demetsky M J. Traffic flow forecasting: comparison of modeling approaches[J]. Journal of transportation engineering, 1997, 123(4): 261-266.

[4] Ross, S. M.Introduction to probability models(Academic press, 2009).

[5] Liu G, Maguire Jr G. A class of mobile motion prediction algorithms for wireless mobile computing and communication[J]. Mobile Networks and Applications, 1996, 1(2): 113-121.

[6] Liou S C, Lu H C. Applied neural network for location prediction and resources reservation scheme in wireless networks[C]//Communication Technology Proceedings, 2003. ICCT 2003. International Conference on. IEEE, 2003, 2: 958-961.

[7] Akoush S, Sameh A. Mobile user movement prediction using bayesian learning for neural networks[C]//Proceedings of the 2007 international conference on Wireless communications and mobile computing. ACM, 2007: 191-196.

[8] Petzold J, Bagci F, Trumler W, et al. Global and local state context prediction[C]//Artificial Intelligence in Mobile Systems. 2003.

[9] Lu X, Wetter E, Bharti N, et al. Approaching the limit of predictability in human mobility[J]. Scientific reports, 2013, 3.

[10] Kontoyiannis I, Algoet P H, Suhov Y M, et al. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text[J]. Information Theory, IEEE Transactions on, 1998, 44(3): 1319-1327.

[11] Fano R M, Hawkins D. Transmission of information: A statistical theory of communications[J]. American Journal of Physics, 1961, 29(11): 793-794.

[12] Brabazon, Anthony, and Michael O'Neill, eds. Natural computing in computational finance. Vol. 100. Springer Science & Business Media, 2008.

[13] Song C, Qu Z, Blumm N, et al. Limits of predictability in human mobility[J]. Science, 2010, 327(5968): 1018-1021.

[14] Lu X, Bengtsson L, Holme P. Predictability of population displacement after the 2010 Haiti earthquake[J]. Proceedings of the National Academy of Sciences, 2012, 109(29): 11576-11581.