Peer-reviewed version available at Entropy 2017, 19, , 165; doi:10.3390/e19040165

Article

An Entropy-based Approach for Evaluating Travel Time Predictability Based on Vehicle Trajectory Data

Tao Xu [1,*] and Xiang Li [1]

[1] Key Laboratory of Geographical Information Science, Ministry of Education, East China Normal University

* Correspondence: txucn@hotmail.com; Tel.: +86-21-54341218

Abstract: With the great development of intelligent transportation systems (ITS), travel time prediction has attracted the attentions of many researchers and a large number of prediction methods have been developed. However, as an unavoidable topic, the predictability of travel time series is the basic premise for travel time prediction has received less attention than the methodology. Based on the analysis of the complexity of travel time series, this paper defines travel time predictability to express the probability of correct travel time prediction and proposes an entropy-based method to measure the upper bound of travel time predictability. Multiscale entropy is employed to quantify the complexity of travel time series, and the relationships between entropy and the upper bound of travel time predictability are presented. Empirical studies are made with vehicle trajectory data in an express road section. The effectiveness of time scales, tolerance, and series length to entropy and travel time predictability are analysis, and some valuable suggestions about the accuracy of travel time predictability are discussed. Finally, the comparisons between travel time predictability and actual prediction results from two prediction models, *ARIMA* and *BPNN*, are conducted. Experimental results demonstrate the validity and reliability of the proposed travel time predictability.

Keywords: travel time predictability; multiple entropy; travel time series; vehicle trajectory data

1. Introduction

With the ever-increasing traffic congestion in metropolitan area, travel time for every traveler becomes complicated and irregular. How to accurately predict travel time, therefore, is of great importance to the related researchers. In the past few decades, a variety of approaches of travel time prediction have been developed with different models i.e. linear regression, ARIMA, Bayesian nets, neural networks, decision trees, support vector regression, kalman filtering, etc., Based on periodic fluctuations of historical travel time series, these approaches make use of their own derivation rules to recognize traffic patterns and predict future travel time in specific routes as precisely as possible. However, the inevitable nonstationarity of travel time series caused by high self-adapting and heterogenetic drivers or unpredictable and unusual circumstances makes it difficult to accurately predict future travel time. In addition to the performance of prediction models, the quality of input data (historical travel time series) for prediction model affects the precision of travel time prediction.

Nowadays, various data are employed for travel time prediction [14, 16, 26], e.g. vehicle trajectory data, mobile phone data, smart card data, loop detector data, video monitoring data, artificial statistical data, etc. Vehicle trajectory data are frequently collected from a large number of vehicles and consist of a huge number of GPS sample points including geographic coordinates and sample time as well as the identification of vehicle. Historical travel time series in specific trip can be from large amount of vehicle trajectory data and travel time prediction can be achieved. Based on vehicle trajectory data, many existing research works have been done for travel time prediction. Some of them evaluate the performance of prediction models [12, 16, 24, 25], some of them focus on the reliability or uncertainty of historical travel time series [1, 8, 11, 13, 21, 28], but few of them examine the quality of data from the perspective of prediction.

Travel time reliability is the consistency or dependability in travel times as measured from day-to-day or across different times of a day [15], which represents the temporal uncertainty experienced by travelers in their trip [3] or the travel time distributions under various external conditions [17]. Travel time reliability only present the certainty of historical travel rules, it has nothing to do with the accuracy of future travel time.

By this means, the aim of this paper is to evaluate the quality of historical travel time series extracted from vehicle trajectory data in terms of travel time prediction. Especially, we use the term "predictability" to denote the results of evaluation. In traffic study, some research efforts about predictability have been presented. Yue et al. [32] uses the cross-correlation coefficient between traffic flows collected at two detector stations to explain short-term traffic predictability in the form of probability. Foell et al. [7] analysis the temporal distribution of ridership demand on various date conditions and uses the F-score, an effective metric of information retrieval, to measure the predictability of bus line usage. Siddle [22] introduces travel time predictability of two specific prediction models, i.e. auto-regressive moving average, and non-linear time series analysis, in the Auckland strategic motorway network. In it, travel time predictability is used to explain the performance of specific prediction models. In addition, the predictability of road section congestion (speed) [27] and human mobility [5, 23] are measured by information entropy. Until now, there are not sufficient literatures are published about the measurement of travel time predictability by data itself. Therefore, we hope to explore travel time predictability for evaluating the characteristic of travel time series on prediction.

In this paper, the definition of travel time predictability is the possibility of correct prediction by historical travel time series with specific accuracy requirements. It indicates the influence of the complexity of historical travel time series on prediction results. E.g., travel time predictability of a travel time series is 0.9, which means that, no matter how good predictive model, we cannot predict with better than 90% accuracy the future travel time with the given travel time series. Regular commute patterns give us full of confidence for future travel time, but random traffic flow often disturb traffic rules and bring uncertainly changes for travel time prediction.

Song et al. [23] explores limits of predictability of human mobility and develops a method to measure the upper bound of predictability based on information entropy [20] and Fano's inequality [6]. In this research, mobile phone data are employed to quantify human mobility as discrete location series, and the entropy of location series is measured by Lempel-Ziv data compression [9], then Fano's inequality is used to deduce the relationships between entropy and the upper bound of predictability. Lempel-Ziv data compression algorithm is a method to measure the complexity of the nonlinear symbolic coarse-grained time series. However, travel time series usually has a continuous range of values determined by the tradeoff between accuracy and grain size. Since the complexity of time series is highly sensitive to grain size [30], the Symbolization of travel time series is never a straightforward task. Furthermore, the complexity from Lempel-Ziv algorithm can only be used for qualitative analysis and is not suitable for quantitative description [31]. Therefore, Lempel-Ziv algorithm is not suitable to measure the complexity of travel time series, and the method proposed by Song et al. [23] cannot entirely applicable to travel time predictability.

Inspired by Song et al. [23], this paper attempts to measure the complexity of travel time series and assess travel time predictability. First, travel time series are defined as a continuous variable and the Multiscale Entropy (*MSE*) [4] in different scales are measured to present the true entropy of travel time series. Then, the upper bound of predictability is calculated by the method of Song et al. [23]. Usually, MSE is used to assess the complexity of multi-value time series from the perspective of multi-time scales and has been in many fields successfully. However, *MSE* often produces some inaccurate estimations or undefined entropy which brings difficulties in evaluating the complexity of travel time series correctly. For address them, an improvement of *MSE*, the refined composite multiscale entropy (*RCMSE*) algorithm proposed by Wu et al. [29], is employed to measure the complexity of travel time series, while Wu et al. [29] demonstrates that *RCMSE* increases the accuracy of entropy estimation and reduces the probability of inducing undefined entropy. Compared with

Peer-reviewed version available at Entropy 2017, 19, , 165; doi:10.3390/e19040165

MSE, the *RCMSE* can be used to estimate entropy more accurately with the lower probability of inducing undefined entropy.

To this end, our contributions are to integrate the two methods proposed by Wu et al. [29] and Song et al. [23], and apply to the above aim to evaluate the features of entropy and predictability of travel time series.

Furthermore, this paper employs an express road section with heavy traffic flow in Shanghai, China as research area and vehicle trajectory data as data source to analyse and evaluate travel time predictability. By using large amount of vehicle trajectory data, massive trip data in the selected route are acquired, entropy and predictability are assessed. Then, the influences of time scales, tolerance, and series length on entropy and travel time predictability are discussed. At last, we employ two prediction models, *ARIMA* and *BPNN*, to predict future travel time of the selected route to verify the validity and reliability of the proposed travel time predictability.

The rest of this paper is organized as follows. The next section develops the methodology of travel time predictability. The introduction and results of empirical researches are presented in Section 3. And section 4 concludes the paper.

2. Materials and Methods

Historical travel time series are extracted from vehicle trajectory data including a large number of sample points. Each sample point of vehicle trajectory consists of vehicle ID, time stamp, longitude, latitude, speed, etc. In order to get travel time of specific trip, road matching of vehicle trajectory data is performed to specific routes with the method proposed by Li et al. [10]. By calculating the difference of time stamp of first and last sample point in origin and destination of specific trip, the set of travel time of all of trip is established. Then, based on predefined departure time interval, multiple travel time satisfying departure time settings are averaged to generate travel time series.

For any of travel time series with specific origin, destination, and route, we employ the *RCMSE* algorithm [29] to calculate multiscale entropy of travel time series and evaluate the complexity of travel time series. Then travel time predictability is defined and the relationships between the upper bound of travel time predictability and entropy of historical travel time series are presented.

2.1 Entropy of travel time series

Multiscale entropy of travel time series is presented by the refined composite multiscale entropy (*RCMSE*) algorithm which is given below.

Let $X = \{X_1, X_2, \dots, X_N\}$ be a travel time series with length of N.

Step 1. Construct m-dimensional vectors X_i^m by using Equation (1).

$$X_i^m = \{X_i, X_{i+1}, \dots, X_{i+m-1}\}, \ 1 \le i \le N - m, \tag{1}$$

Step 2. Calculate all of the Euclidean distance d_{ij}^m between any of two vectors, X_i^m and X_j^m by using Equation (2).

$$d_{ij}^m = \left\| X_i^m - X_j^m \right\|_\infty, \ 1 \le i, j \le N - m, i \ne j, \tag{2}$$

Step 3. Let r be a tolerance level. If $d_{ij}^m \le r$, X_i^m and X_j^m are called a m-dimensional matched vector pair. Let n^m be the total number of m-dimensional matched vector pairs. Similarly, n^{m+1} is the total number of $(m+1)$-dimensional matched vector pairs.

Step 4. The Sample Entropy (*SampEn*) is defined by Equation (3).

$$SampEn(X, m, r) = -ln \frac{n^{m+1}}{n^m}, \tag{3}$$

Step 5. Let $y_k^\tau = \{y_{k,1}^\tau, y_{k,2}^\tau, \dots, y_{k,p}^\tau\}$ be the k-th coarse-grained time series of X defined as Equation (4), where p is the length of the coarse-grained time series and τ is a scale factor. To obtain y_k^τ, the original time series X is segmented in N/τ coarse-grained series with each segment being of length τ. The j-th element of the k-th coarse-grained time series $y_{k,j}^\tau$ is the mean value of each segment τ of the original time series X.

$$y_{k,j}^{\tau} = \frac{1}{\tau}\sum_{i=(j-1)\tau+k}^{j\tau+k-1} X_i,\ 1 \le j \le \frac{N}{\tau}, 1 \le k \le \tau, \quad (4)$$

Step 6. Classical multiple entropy, $MSE(X,\tau,m,r)$, is defined by Equation (5).

$$MSE(X,\tau,m,r) = SampEn(y_1^{\tau},m,r), \quad (5)$$

Step 7. The improvement of classical multiple entropy, *RCMSE*, is defined by Equation (6), where $n_{k,\tau}^{m}$ is the total number of m-dimensional matched vector pairs in the k-th coarse-grained time series with the length of τ.

$$RCMSE(X,\tau,m,r) = -ln\frac{\sum_{k=1}^{\tau} n_{k,\tau}^{m+1}}{\sum_{k=1}^{\tau} n_{k,\tau}^{m}}, \quad (6)$$

Compared with *SampEn* and *MSE*, *RCMSE* algorithm can be used to estimate entropy more accurately with the lower probability of inducing undefined entropy caused by SampEn. In Equation (7), the true entropy $S(X)$ of travel time series X is denoted by $RCMSE(X,\tau,m,r)$. $S(X)$ is roughly equals to, with time scale τ, the negative logarithm of the mean of the conditional probability of new patterns (i.e. the distance between vectors is larger than r) when the dimension of the pattern changes (i.e. m to $m+1$). It describes the degree of irregularity of travel time series at different time scales and is proportional to the complexity of travel time series. Based on Equation (7), the true entropy of travel time series with different time scales can be achieved.

$$S(X) = RCMSE(X,\tau,m,r), \quad (7)$$

2.2 Travel time predictability

Based on historical travel time series, the predictability of travel time can be quantified as the probability Π that an appropriate predictive algorithm can correctly predict future travel time. We define travel time predictability as Π, and introduce the relationships between entropy and the upper bound Π^{max} of travel time predictability.

Let $X = \{X_1, X_2, \dots X_N\}$ be a historical travel time series with length of N, φ be the true value of the $(N+1)^{th}$ travel time, $\hat{\varphi}$ be the expected value, and φ_a be the predictive value with an appropriate prediction model α. Let π be the probability of $\varphi = \hat{\varphi}$ with given historical travel time series X. Equation (8) shows that π is the random value of distribution of next travel time. It can be seen that π is an upper bound of the probability distribution of predictive values. Therefore, we can demonstrate that any predicting based on historical series X cannot do better than the one that the true travel time is equals to the expected value, $\varphi = \hat{\varphi}$.

$$\begin{aligned} \pi &= P(\varphi = \hat{\varphi}|X), \\ &= \sup_x\{P(\varphi_a = x|X)\}, \\ &\ge \mathrm{P}(\varphi = \varphi_a|X), \end{aligned} \quad (8)$$

The definition of predictability Π for a travel time series with length of N is given by Equation (9), where $P(X)$ is the probability of observing a particular historical travel time series X, $\sum \pi P(X)$ presents the best success rate to predict the $(N+1)^{th}$ travel time with given travel time series X. Π may be viewed as the averaged predictability (Song et al. 2010) of a historical travel time series.

$$\Pi = \lim_{N\to\infty} \frac{1}{n}\sum_{i}^{N}\sum \pi P(X), \quad (9)$$

Next we relate entropy $S(X)$ to predictability Π to explore the upper bound of predictability Π^{max}. Based on Fano's inequality [6], the relationship between entropy and predictability is shown in Equation (10), which indicates that the complexity of X is less than or equals to the sum of the complexity of successful predicting $S(\Pi)$ and the complexity of failure predicting $(1-\Pi)\,log_2(n-1)$ where n is the number of values of X. In this paper, the unit of travel time is second, we let n be the seconds of value range of X. The equality in Equation (10) hold up if and only if Π is the maximum, i.e. $\Pi = \Pi^{max}$. In addition, entropy of Π, $S(\Pi)$ is presented by Equation (11). Therefore,

the relationship between the upper bound of travel time predictability and entropy of travel time series is presented by Equation (12). Based on the known $S(X)$ by Equation (7), we can traverse from 0 to 1 to get the optimal solution of Π^{max} with given accuracy target.

$$S(X) \leq S(\Pi) + (1-\Pi)\log_2(n-1), \tag{10}$$

$$(\Pi) = -\Pi \log_2 \Pi \ -(1-\Pi)\log_2(1-\Pi), \tag{11}$$

$$(X) = -\Pi^{max} \log_2 \Pi^{max} \ -(1-\Pi^{max})\log_2(1-\Pi^{max}) + (1-\Pi^{max})\log_2(n-1), \tag{12}$$

3. Experiments

3.1. Research area and data

In this research, an express road section with length about 6.74 kilometers in Shanghai, China is selected as research area. As shown in Figure 1, the selected route is a traffic corridor with heavy traffic flow and is a part of the express road system of Shanghai represented by gray lines. Complex traffic flow makes travel time is often changeable.

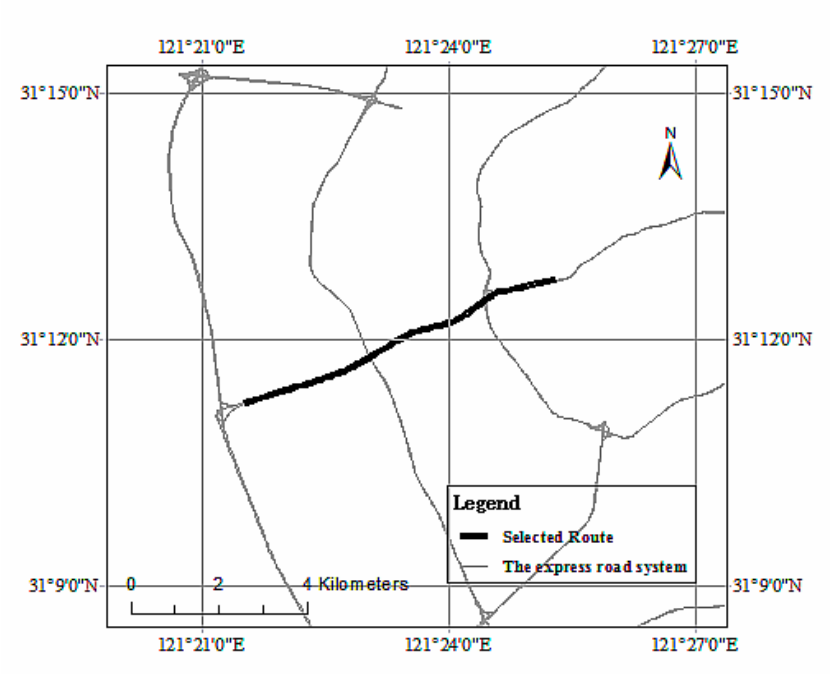


Figure 1. The selected express road section.

Taxi trajectory data covering the selected route in April 2015 are employed as data source to supply real travel time series. The total number of travel case is 20430, and the average number is about 29 per hour. It is sufficient to represent the dynamic changes of travel time. By averaging travel time of traffic cases in 5 minutes departure time interval, shown in Figure 2, 5-min travel time series from taxi trajectory data in April, 2015 can be obtained. The number of points in it is 8640. Let $X = \{X_i | 0 < i < N, N = 8640\}$ be the 5-min travel time series, where X_i is the ith sample point and N is the number of points of X. The analysis and evaluation of entropy and predictability of it are given below.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 March 2017 doi:10.20944/preprints201703.0213.v1

Peer-reviewed version available at Entropy 2017, 19, , 165; doi:10.3390/e19040165

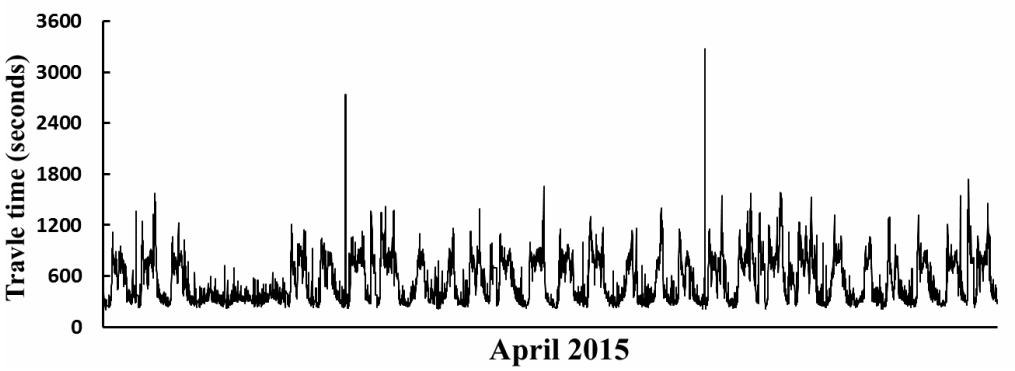


Figure 2. The 5-min travel time series from taxi trajectory data in April 2015.

3.2 Entropy and predictability

To evaluate the complexity of travel time series, the daily entropy, named S^d, and the weekly entropy, named S^w, are calculated with 24 hours subseries, i.e. 288 consecutive points, and 7*24 hours subseries, i.e. 2016 consecutive points, respectively, from 5-min travel time series. We set the difference of adjacent subseries is 1 hour (12 points) to obtain many of subseries. Then, let $S^d = \{S_j^d(X') |\ X' = \{X_{j\times 12}, X_{j\times 12+1}, \dots, X_{j\times 12+288-1}\},\ 0 < j < 696\}$, where X' is a subset of X with 288 consecutive points, and let $S^w = \{S_j^w(X'') | X'' = \{X_{j\times 12}, X_{j\times 12+1}, \dots, X_{j\times 12+2016-1}\},\ 0 < j < 552\}$ where X'' is a subset of X with 2016 consecutive points.

Let scale factor $\tau = 1$, tolerance $r = 0.1\sigma$, and dimension $m = 2$, where σ is the standard deviation of 5-min travel time series. The statistics of values of S^d and S^w are shown in Figure 3. It can be seen that the values of S^d are scattered in the range of 0.6 to 3.4 and the values of S^w are compact in the range of 1.6 to 2.3. The remarkable difference between S^d and S^w means that the complexity of daily travel time series tends to change frequently and, by contrast, the complexity of weekly travel time series is stable. In it, S^w peaks about 1.7, indicating that, on average, the probability of 2-dimension ($m = 2$) new patterns in weekly travel time series is $e^{-1.7} \approx 0.183$. S^d peaks about 1.2 and the probability of new patterns in daily travel time series is $e^{-1.2} \approx 0.301$ indicated that weekly travel time series with more complexity have smaller probability of new patterns than relatively simple daily travel time series.

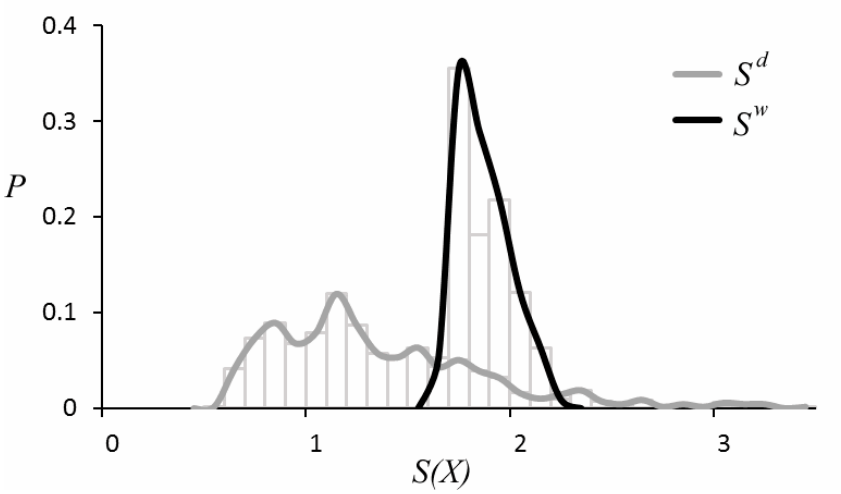


Figure 3. The entropy in weekly travel time series and daily travel time series.

Travel time predictability is the probability of accurate prediction that is determined by the complexity (entropy) and the value ranges of travel time series. In our experiments, we set the accuracy of 0.001 to calculate the upper bound of travel time predictability Π^{max} by Equation (12). So the optimal (maximum) Π^{max} can be get by traversing in the range of 0.001 to 0.999.

The statistics results of upper bound of travel time predictability in weekly travel time series Π^w and daily travel time series Π^d are shown in Figure 4. Since travel time predictability affects by entropy and the value range of series, weekly travel time series with larger entropy have smaller predictability peaking 0.95, and daily travel time series with smaller entropy have larger predictability peaking 0.99. It demonstrates that the more complex the travel time series is, the less predictability and the more difficult to correctly predict.

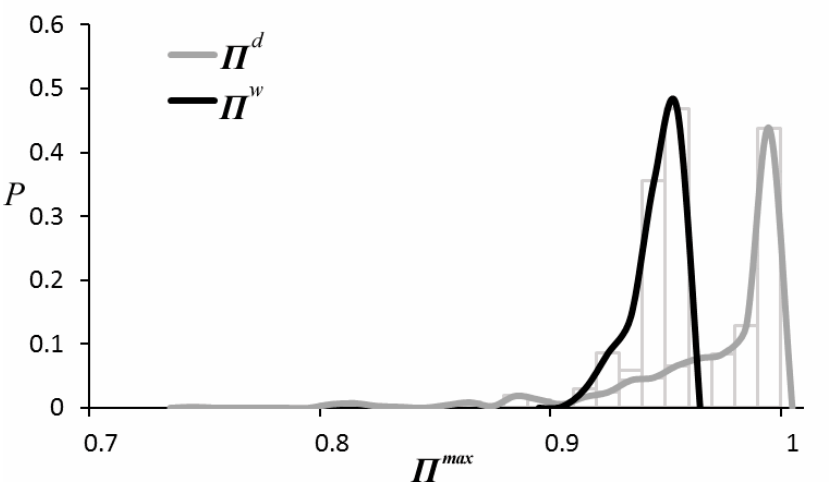


Figure 4. The upper bound of travel time predictability in weekly travel time series and daily travel time series.

3.3 Analysis and discussion

In this subsection, we analyse the effectiveness of scale factor τ, tolerance r, and series length to predictability of travel time series, and discussion the features and trends of travel time predictability. The validity of the proposed travel time predictability is verified by comparing Π^{max} and the predictive values of future travel time from two typical prediction models, *ARIMA*, and *BPNN*.

3.3.1 Time scales

The scale factor τ is the key parameter of *MSE*. It gives us a good opportunity to analyse the complexity and predictability of travel time series in multiple time scales, Figure 5 shows the entropy and predictability of 5-min travel time series with time scales of 1 to 20. The entropy is calculated by Equation (7) with $r = 0.1\sigma$, and $m = 2$. Π^{max} is calculated by Equation (12). With τ increases, entropy rises and predictability falls. There are more "new patterns" in travel time series of larger time scales. That is the complexity of travel time series of larger time scales is larger than those of smaller time scales, and travel time series of larger time scales are more difficult to correctly predict.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 March 2017 doi:10.20944/preprints201703.0213.v1

Peer-reviewed version available at Entropy 2017, 19, , 165; doi:10.3390/e19040165

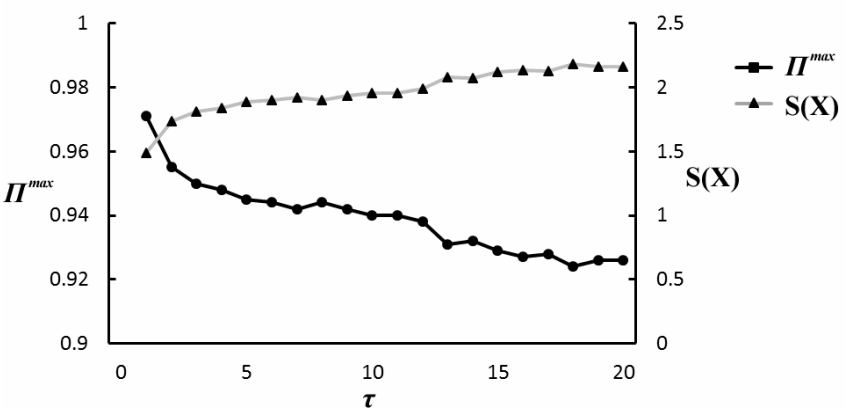


Figure 5. Entropy and predictability of 5-min travel time series with scale factor of 1 to 20.

3.3.2 Tolerance

Tolerance, r, is a key factor to evaluate the complexity of travel time series which constrains the contributions of travel time fluctuations to the complexity. We attempt to evaluate the effectiveness of r in six time scales, i.e. τ = 2, 4, 6, 8, 10, and 12. Figure 6 and 7 shows the changing trends of entropy and travel time predictability of 5-min travel time series with r of 0.01σ to 0.16σ respectively. Since when r is equals to 0.16σ, travel time predictability of six time scales reach the maximum value 0.999, the test range of r is 0.01σ to 0.16σ. In Figure 6, with the increase of r, the entropy gradually becomes smaller which is because the value gap between travel time less than r is not concerned. To six time scales, in addition to τ = 2, other values of entropy is hard to distinguished at smaller r, and ordered values of entropy can be found at larger r (about 0.06σ to 0.16σ).

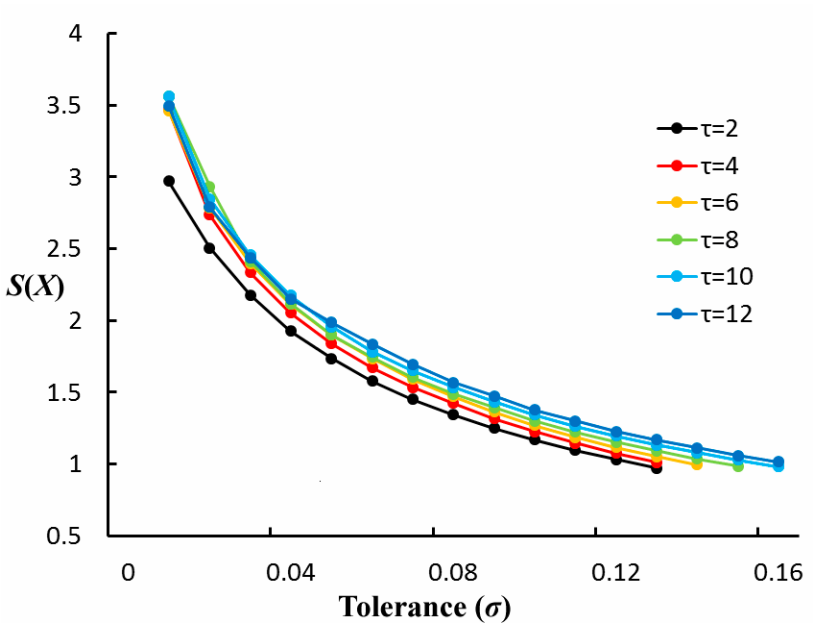


Figure 6. The effectiveness of r to entropy.

Figure 7 shows Π^{max} of 5-min travel time series with six time scales. It can be seen that there is a negative correlation between Π^{max} and S(X). The larger Π^{max} and smaller S(X) are in smaller time scales i.e. τ = 2, and the smaller Π^{max} and larger S(X) are in larger time scales. At the same tolerance

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 March 2017 doi:10.20944/preprints201703.0213.v1

Peer-reviewed version available at Entropy 2017, 19, , 165; doi:10.3390/e19040165

level, travel time series with smaller r are easier to prediction than those with larger r. With the expansion of r, Π^{max} gradually increases to 0.999. Π^{max} is limited by r. Obviously, the larger r, the higher tolerance to predictive error, the greater Π^{max} and the more accurate prediction.

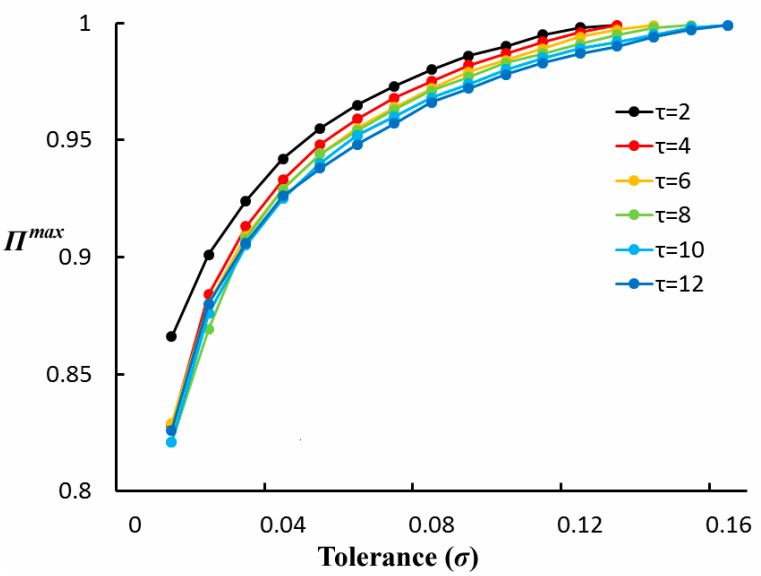


Figure 7. The effectiveness of r to travel time predictability.

For the possibility of perfect theoretical prediction, Figure 8 shows the tolerance of perfect prediction in multiple time scales. The ranges of τ are from 1 to 20. Black line represents the trends of τ with Π^{max} of 0.999. E.g. next travel time in $\tau = 6$ can be accurately predicted with $r = 0.14\sigma$ by appropriate prediction model. The growth trend of r indicates that larger τ is more difficult to predict and the perfect prediction of them need a larger tolerance ranges.

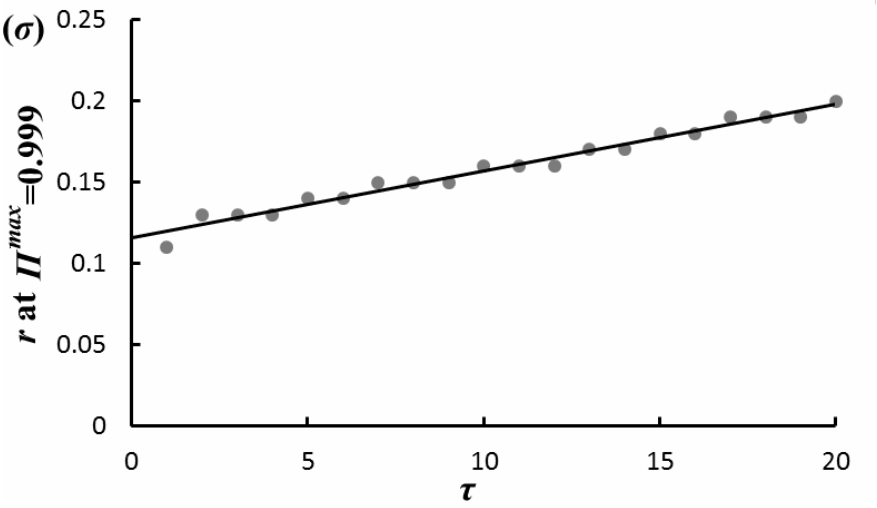


Figure 8. The effectiveness of r to perfect prediction.

3.3.3 Series length

Next we analyse the influences of series length to entropy and predictability. Figure 9 shows the entropy of travel time series in six time scales with different series length. The series length of one day 5-min travel time series is 288, others and so on. Meanwhile, $r = 0.1\sigma$, and $m = 2$. It can be seen

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 March 2017 doi:10.20944/preprints201703.0213.v1

Peer-reviewed version available at Entropy 2017, 19, , 165; doi:10.3390/e19040165

that these larger entropy are in 2-days or 3-days travel time series and the more stable trends are in more than about 14-days (i.e. two weeks) travel time series. We can think that entropy of more than 14-days travel time series is roughly independent of series length. Table 1 shows the statistics of entropy to support these conclusions, where $\overline{S(X)}$ is the average value of entropy of all of travel time series, and sd_S is the standard deviation of all of entropy. The sd_S of more than 14-days are much smaller than those of less than 14-days and the most stable series is more than 14-days travel time series with $\tau = 2$ and $\tau = 4$. Therefore, we can demonstrate that the complexity of more than 14-days travel time series is stable and is independent to series length.

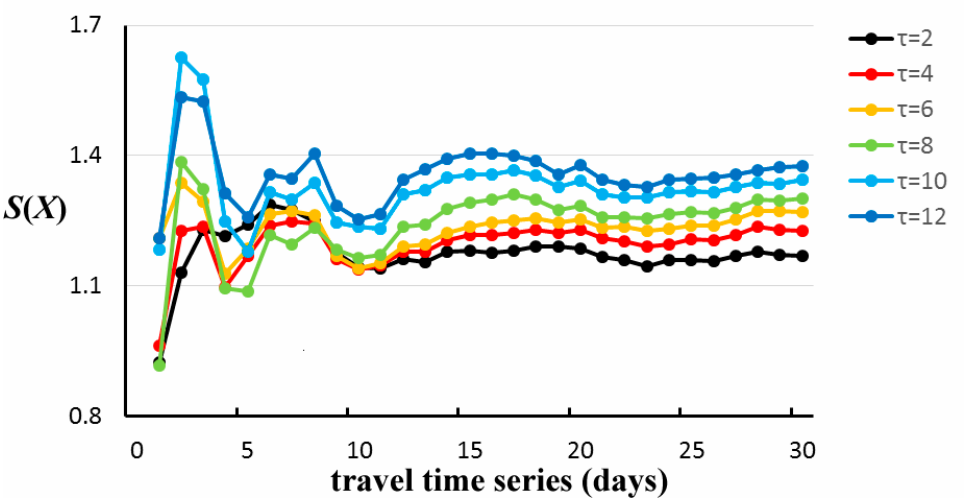
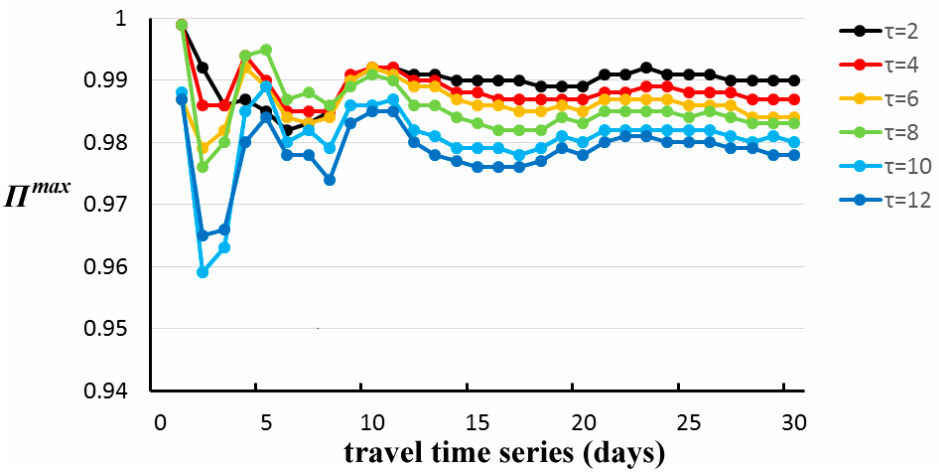


Figure 9. Entropy of travel time series with different series length.

Similar with conclusions of Figure 9, Figure 10 demonstrates the smallest value of predictability of 2-days or 3-days travel time series and the stationarity and independence of predictability of more than 14-days travel time series. In Table 1, $\overline{\Pi^{max}}$ denotes the average value of travel time predictability and $sd_{\Pi^{max}}$ denotes the standard deviation of travel time predictability. Great differences between $sd_{\Pi^{max}}$ of more than 14-days travel time series and less than 14-days travel time series present the stable predictability of more than 14-days travel time series. In addition, we can demonstrate that the most stable predictability is in more than 14-days travel time series with $\tau = 2$ and $\tau = 4$.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 March 2017 doi:10.20944/preprints201703.0213.v1

Peer-reviewed version available at Entropy 2017, 19, , 165; doi:10.3390/e19040165

Figure 10. Travel time predictability of travel time series with different series length.

Table 1. The statistics of entropy and travel time predictability

τ	$\overline{S(X)}$	sd_S		$\overline{\Pi^{max}}$	$sd_{\Pi^{max}}$	
		< 14 days	≥ 14 days		< 14 days	≥ 14 days
2	1.1752	0.0896	0.0126	0.9896	0.0045	0.0008
4	1.1966	0.0755	0.0125	0.9885	0.0040	0.0007
6	1.2334	0.0623	0.0148	0.9863	0.0040	0.0011
8	1.2415	0.1108	0.0167	0.9856	0.0058	0.0011
10	1.3261	0.1311	0.0193	0.9805	0.0089	0.0013
12	1.3571	0.0953	0.0242	0.9786	0.0066	0.0016

3.3.4 The validity of travel time predictability

To verify the validity of travel time predictability, two prediction models, i.e. AutoRegressive Integrated Moving Average (*ARIMA*) [2], and Back Propagation Neuro Networks (*BPNN*) [18], are employed to predict future travel time.

ARIMA model is a method for time series analysis and prediction. Since travel time series has obvious fluctuation differences on workday and weekend, we use seasonal *ARIMA* (*SARIMA*) model, denoted $ARIMA(p,d,q)(P,D,Q)_m$, to predict future travel time, where p is the order of the autoregressive (*AR*) part, q is the order of the moving average (*MA*) part, d is the degree of differencing for reducing the non-stationarity of time series, m is the number of periods per season, and P, D, Q refer to the *AR*, differencing, and *MA* terms for the seasonal part of the *ARIMA* model. Due to, in our experiments, the stationary and weekly change period of travel time series, we let $d = 0$, $D = 0$ and $m = 7$. By testing the autocorrelation function (*ACF*) and the partial autocorrelation function (*PACF*) of complete and seasonal part travel time series, we let $p = 3$, $q = 1$, $P = 1$, and $Q = 2$. Then, we use $ARIMA(3,0,1)(1,0,2)_7$ to predict future travel time in selected route.

As a neuro network method, *BPNN* model includes an input layer, a hidden layer, and an output layer. It can learn and store large amounts of input-output mapping by model training to represent and predict the dynamic and non-linear processes. In our experiments, BPNN model has three inputs, i.e. date, time of day, day of week and one outputs, i.e. travel time, the number of nodes in hidden layer is 7, the learning rate (η) is 0.9, and the momentum factor (α) is 0.7.

Figure 11 shows the errors of travel time prediction with *ARIMA* and *BPNN* models in 5-min travel time series. $r = 0.1\sigma$ (about 22 seconds), $m = 2$, and $\tau = 2$. We predict 50 times with *ARIMA* and *BPNN* respectively, and let e is the absolute value of the difference between predictive value and actual value. Dashed line indicates the tolerance $r = 0.1\sigma$. It can be seen that most of dots are blow it. The statistical results of travel time prediction of 5-min travel time series are shown in Table 2. $\overline{\Pi^{max}}$ denotes the average predictability of 50 prediction in 5-min travel time series. If e is less than r (blow the dashed line of Figure 11), we can think it is a successful prediction. The number of successful prediction is 40 with *ARIMA*, and is 41 with *BPNN*. Comparing with $\overline{\Pi^{max}}$ of 0.952, the success rates of prediction are lower, while their average errors, 13.46 and 12.42, are lower than tolerance r, 22.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 March 2017 doi:10.20944/preprints201703.0213.v1

Peer-reviewed version available at Entropy 2017, 19, , 165; doi:10.3390/e19040165

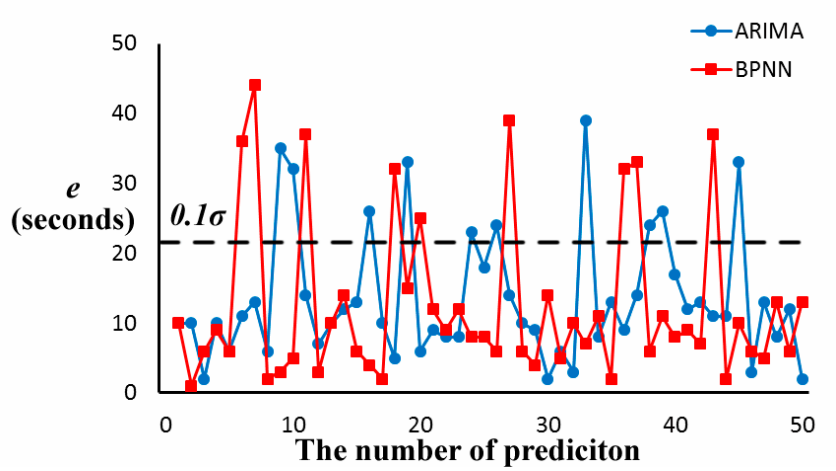


Figure 11. The errors of travel time prediction in 5-min travel time series.

Table 2. The statistics of prediction with 5-min travel time series.

Prediction model	Number of prediction	Number of success	Success rate	Average error (sec)	$\overline{\Pi^{max}}$
ARIMA	50	38	90%	13.46	0.952
BPNN	50	40	91%	12.42	

For comprehensively evaluating the relationships between travel time predictability and prediction results, two group comparisons are conducted. Figure 12 shows the comparison results between travel time predictability and prediction results in 5-min travel time series with different series length of 1 to 30-days. The average prediction results of 100 experiments of each travel time series are presented, $r = 0.1\sigma$, $m = 2$, and $\tau = 2$. Let SR_{ARIMA} be the success rate of travel time prediction by *ARIMA*, and SR_{BPNN} be the success rate by *BPNN*. Results indicate that there is a marked difference between $\overline{\Pi^{max}}$ and SR_{ARIMA}, SR_{BPNN}. Note that their change trends are consistent basically, which indicates that the accuracy of prediction is affected by the complexity of travel time series, and meanwhile, demonstrates the validity of travel time predictability.

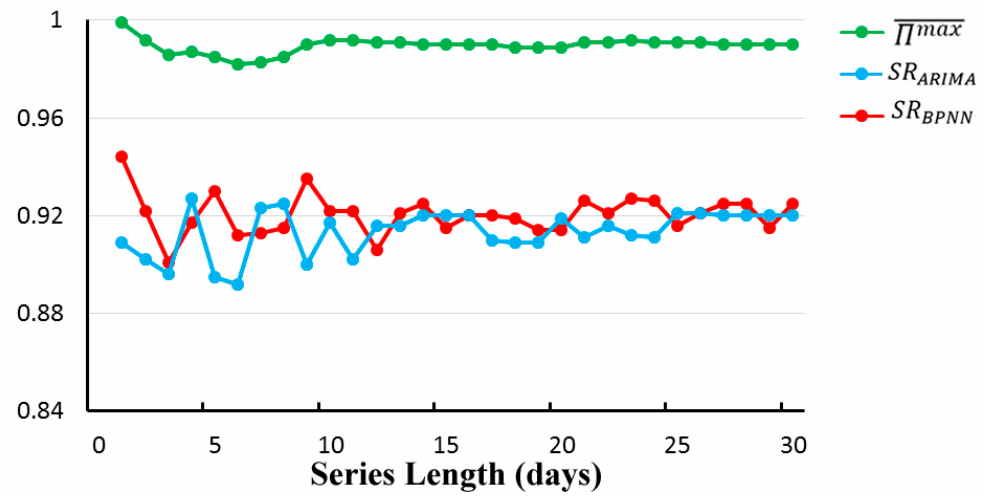


Figure 12. The comparisons between travel time predictability and prediction results in 5-min travel time series with different series length.

The same situation occurs in Figure 13. We compare $\overline{\Pi^{max}}$ and prediction results in different time scales of 1 to 20 with $r = 0.1\sigma$, and $m = 2$. With the increase of time scales, travel time predictability and two success rates of prediction decline synchronously.

The proposed travel time predictability is a valid measurement of travel time series to correct prediction, which provide an achievable target to the development of travel time prediction methods and contribute to make a differentiated scheme for travel time prediction.

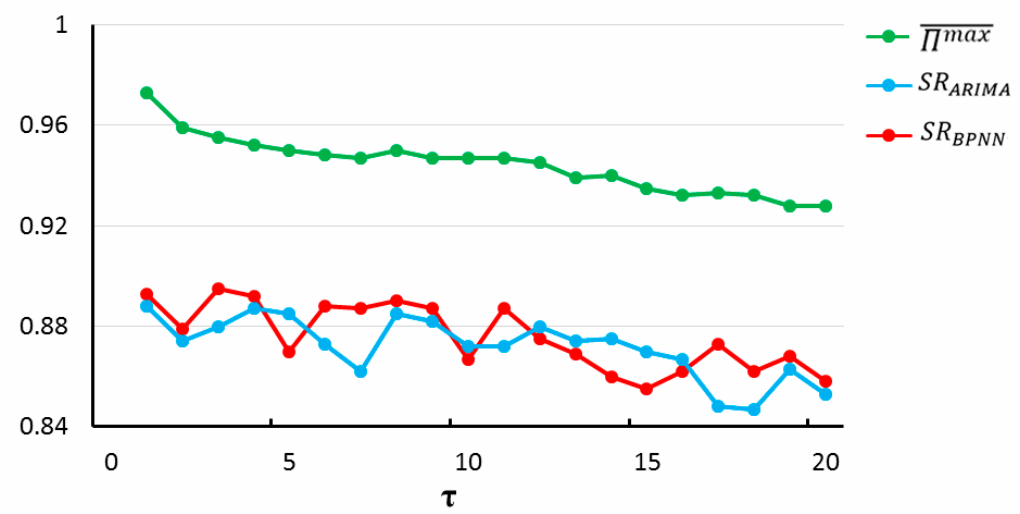


Figure 13. The comparisons between travel time predictability and prediction results in travel time series with different time scales.

4. Discussion

This paper proposes travel time predictability to describe the probability of correct prediction by the complexity of historical travel time series, and develop an entropy-based approach to measure the upper bound of travel time predictability. Multiple entropy of travel time series is calculated to evaluate the complexity, and the upper bound of travel time predictability is related to the entropy. Travel time predictability expresses the characteristics of travel time series itself and is an expected value of data-based prediction performance.

Empirical researches are conducted in an express section road about 6.74 kilometers in Shanghai, China. Data source is from taxi trajectory on April, 2015. By analyzing the effectiveness of time scales and tolerance to entropy and travel time predictability, we demonstrate that time scales and tolerance are positively related with the entropy and negative related with travel time predictability. In addition, we reveal the larger value of entropy and smaller predictability of 2-days or 3-days travel time series and the more stable values of more than 14-days travel time series. Finally, two prediction models, *ARIMA* and *BPNN*, are employed to predict travel time by historical travel time series and verify the validity and reliability of travel time predictability. Though travel time predictability is independent of prediction method, it can provide an achievable target to the development of travel time prediction methods and contribute to make a differentiated scheme for travel time prediction in diverse traffic environment.

Future efforts will be made in two directions. First, the comprehensive investigation and verification of travel time predictability should begin in variety of routes to provide the basic references for travel time prediction, which contributes to deeper traffic knowledge discovery and differentiation traffic police formulation. Second, the scope of predictability should be extended and the possibility of applying predictability to other types of time series need to be surveyed.

References

1. Abrantes, P. A. L.; Wardman, M. R. Meta-analysis of UK values of travel time: an update. *Transportation Research Part A Policy & Practice* **2011**, *45(45)*, 1-17.
2. Box, G. E. P.; Jenkins, G. M. Time series analysis: forecasting and control. *Journal of the Operational Research Society* **1971**, *22(2)*, 199-201.
3. Carrion, C.; Levinson, D. Value of travel time reliability: A review of current evidence. *Transportation Research Part A Policy & Practice* **2012**, *46(4)*, 720-741.
4. Costa, M.; Goldberger, A. L.; & Peng, C. K. Multiscale entropy analysis of complex physiologic time series. *Physical Review Letters* **2002**, *92(8)*, 705 - 708.
5. Du, Y.; Chai, Y. W.; Yang, J. W.; Liang, J. H.; Lan, J. H. Predictability of Resident activity in Beijing Based on GPS Data. *Geography and Geo-Information Science* **2015**, *31(6)*, 47-51.
6. Fano, R. M.; Hawkins, D. Transmission of Information: A Statistical Theory of Communications. *American Journal of Physics* **1961**, *29(11)*, 793-794.
7. Foell, S.; Phithakkitnukoon, S.; Kortuem, G.; Veloso, M. Predictability of Public Transport Usage: A Study of Bus Rides in Lisbon, Portugal. *IEEE Transactions on Intelligent Transportation Systems* **2015**, *16(5)*, 2955-2960.
8. Jara-Diaz, S. R. Transport economic theory. *Emerald Group Publishing Limited* **2007**, 11-49.
9. Kontoyiannis, I.; Algoet, P. H.; Suhov, Y. M.; Wyner, A. J. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *IEEE Transactions on Information Theory* **2007**, *44(3)*, 1319-1327.
10. Li, X. J.; Li, X.; Tang, D.; Xu, X. Deriving features of traffic flow around an intersection from trajectories of vehicles. The International Conference on Geoinformatics: Giscience in Change, Geoinformatics 2010, Peking University, Beijing, China, June. DBLP, 2010:1-5.
11. Li, Zh.; Hensher, D. A.; Rose, J. M. Willingness to pay for travel time reliability in passenger transport: A review and some new empirical evidence. *Transportation Research Part E Logistics & Transportation Review* **2010**, *46(3)*, 384-403.
12. Lin, H. E.; Zito, R.; Taylor, M. A review of travel-time prediction in transport and logistics. Proceedings of the Eastern Asia Society for transportation studies **2005**, *5*, 1433-1448.
13. Lint, J. W. C. V.; Zuylen, H. J. V.; Tu, H. Travel time unreliability on freeways: Why measures based on variance tell only half the story. *Transportation Research Part A Policy & Practice* **2008**, *42(1)*, 258-277.
14. Mori, U.; Mendiburu, A.; Álvarez, M.; Lozano, J. A. A review of travel time estimation and forecasting for Advanced Traveller Information Systems. *Transportmetrica A: Transport Science* **2015**, *11(2)*, 1-39.
15. Noland, R. B.; Polak, J. W. Travel time variability: a review of theoretical and empirical issues. *Transport reviews* **2002**, *22(22)*, 39-54.
16. Oh, S.; Byon, Y. J.; Jang, K.; Yeo, H. Short-term Travel-time Prediction on Highway: A Review of the Data-driven Approach. *Transport reviews* **2015**, *35(1)*, 4-32.
17. Rietveld, P.; Bruinsma, F. R.; Vuuren, D. V. Coping with unreliability in public transport chains: A case study for Netherlands. *Transportation Research Part A Policy & Practice* **2001**, *35(6)*, 539–559.
18. Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning representations by back-propagating errors. *Cognitive modeling* **1988**, *5(3)*, 1.
19. Sang, A. M.; Li, S. Q. A predictability analysis of network traffic. *Computer Networks* **2001**, *39*, 329-354.
20. Shannon, C. E. A Mathematical Theory of Communication: The Bell System Technical Journal. *Bell System Technical Journal* **1948**, *27(3)*, 3 - 55.
21. Shires, J. D.; De Jong, G. C. (2009). An international meta-analysis of values of travel time savings. *Evaluation & Program Planning* **2009**, *32(4)*, 315-325.
22. Siddle, D. Travel time predictability, *trid.trb.org (No. 554)* **2014**.
23. Song, C.; Qu, Z.; Blumm, N.; Barabási, A. L. Limits of predictability in human mobility. *Science* **2010**, *327(5968)*, 1018-1021.
24. Van Lint, J. W. C.; Van Hinsbergen, C. P. I. Short-Term Traffic and Travel Time Prediction Models. *Artificial Intelligence Applications to Critical Transportation Issues* **2012**, *22*, 22-41.
25. Vlahogianni, E. I.; Golias, J. C.; Karlaftis, M. G. Short-term traffic forecasting: Overview of objectives and methods. *Transport reviews* 2004, *24(5)*, 533-557.
26. Vlahogianni, E. I.; Karlaftis, M. G.; Golias, J. C. Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C Emerging Technologies* **2014**, *43*, 3-19.

27. Wang, J.; Mao, Y.; Li, J.; Xiong, Z.; Wang, W. X. Predictability of road traffic and congestion in urban areas. *Plos One* **2015**, *10(4)*, e0121825.
28. Wardman, M.; Batley, R. Travel time reliability: a review of late time valuations, elasticities and demand impacts in the passenger rail market in Great Britain. *Transportation* **2014**, *41(5)*, 1041-1069.
29. Wu, S. D.; Wu, C. W.; Lin, S. G.; Lee, K. Y.; Peng, C. K. Analysis of complex time series using refined composite multiscale entropy. *Physics Letters A* **2014**, *378(20)*, 1369-1374.
30. Wyner, A. D.; Ziv, J. The sliding-window lempel-ziv algorithm is asymptotically optimal. *Proceedings of the IEEE* **1994**, *82(6)*, 872-877.
31. Xie, X. X.; Li, S.; Zhang, C. L.; Li, J. K. Study on the application of Lempel-Ziv complexity in the nonlinear detecting. *Complex systems and complexity science* **2005**, *2(3)*, 61-66.
32. Yue, Y.; Yeh, A. G. O.; Zhuang, Y. Prediction time horizon and effectiveness of real-time data on short-term traffic predictability. *Proceedings of the Intelligent Transportation Systems Conference* **2007**, 962-967.