Estimate the limit of predictability in short-term traffic forecasting

An entropy-based approach

Li, Guopeng; Knoop, Victor L.; van Lint, Hans

DOI
10.1016/j.trc.2022.103607

Publication date
2022

Document Version
Final published version

Published in
Transportation Research Part C: Emerging Technologies

Transportation Research Part C 138 (2022) 103607

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Estimate the limit of predictability in short-term traffic forecasting: An entropy-based approach

Guopeng Li *, Victor L. Knoop, Hans van Lint

Department of Civil Engineering and Geosciences, Delft University of Technology, Delft, 2628CN, Netherlands

ARTICLE INFO

Keywords:
Traffic forecasting
Information theory
Conditional differential entropy
Predictability analysis

ABSTRACT

Accurate short-term traffic forecasting is the cornerstone for Intelligent Transportation Systems. In the past several decades, many models have been proposed to continuously improve the predictive accuracy. A key but unsolved question is whether there is a theoretical bound to the accuracy with which traffic can be predicted and whether that limit can be directly estimated from data. To answer this question, we use core concepts in information theory to derive the limit of predictability in short-term traffic forecasting. Theoretical analysis proves that conditional differential entropy poses a rigorous lower bound of negative-log-likelihood (NLL) for probabilistic models. And the continuous form of Fano's theorem further gives a loose lower bound of mean-square-error (MSE) for deterministic models. Based on the special properties of traffic dynamics, two assumptions are made in the estimate of entropy metrics: cyclostationarity (traffic phenomena show strong periodicity) and localized spatial correlation (due to kinematic wave propagation). They allow formulating the limit of predictability as a function of longitudinal space and time-of-day which finds the most uncertain locations and periods solely from data. Experiments on univariate traffic accumulation forecasting and network-level speed forecasting show that the selected models, including some state-of-the-art deep learning models, indeed cannot outperform the estimated lower bounds but just approach them. The limit of predictability depends on time-of-day, network locations, observation range, and prediction horizon. The results reveal that the stochastic nature of traffic dynamics and improper assumptions on the prior distribution of output are two major factors that restrict the predictive performance. In summary, the proposed method estimates a trustworthy performance boundary for most traffic forecasting models. These conclusions are helpful for further studies in this domain.

1. Introduction

Short-term traffic forecasting is critically important for many key applications in traffic and transportation domain. Reliable and accurate short-term predictions of traffic quantities can help traffic managers to rapidly react and make trustworthy decisions to mitigate congestion proactively. For example, Yuan et al. (2011) and Liebig et al. (2017) show that in case urban traffic flows are dynamically guided and re-routed based on predicted traffic states, congestion can be effectively reduced during evening peak hours. Attracted by its great value in applications, researchers have proposed a wide category of methods to give more and more precise traffic predictions, e.g. Van Lint (2008), Ma et al. (2017) and Fusco et al. (2016). Although great progress has been made to improve the predictive performance in this active research field, an important question remains open: *What is the theoretical boundary*

* Corresponding author.
E-mail address: G.Li-5@tudelft.nl (G. Li).

https://doi.org/10.1016/j.trc.2022.103607
Received 24 May 2021; Received in revised form 16 December 2021; Accepted 6 February 2022
Available online 5 March 2022

of predictive accuracy for short-term traffic forecasting? The answer can tell how far we have gone in this domain and what could be the most valuable research direction in the future. Practically, it can put the results from comparing state-of-the-art predictive models into perspective.

We argue that the predictability of traffic variables is mainly governed by two factors: **observability** and **uncertainty**. In a rigorous theoretical sense, a state-space system (like traffic networks) is perfectly observable only if we *can* completely construct all the current state variables from the available measurements by using whatever assumptions on the system dynamics and how the measurements relate to those dynamics. That a system is perfectly-observable is a sufficient but not necessary condition for this system to be perfectly predictable. The necessity does not hold because the system may be not deterministic. The corresponding negative proposition is: if a system is not fully-observable, then it is not fully-predictable. Strict determinism and perfect observability together *in principle* imply perfect predictability.

However, both strict determinism and perfect observability cannot be satisfied in the traffic domain. From experience, traffic systems apparently are not fully-observable. Many latent variables, such as demand and route choice patterns, cannot be completely reconstructed from the limited information collected by sensors. It is almost impossible to get all the demand and supply information needed to predict the short-term evolution of traffic states, especially in large networks. Because of this limitation, most data-driven traffic forecasting problems (e.g. using deep learning techniques Ermagun and Levinson, 2018; Lana et al., 2018) are formulated as sequence-to-sequence regression tasks that only involve easily-observable quantities (e.g. speed and travel time), rather than as in classic state estimation and prediction tasks that explicitly estimate many underlying state variables (e.g. density Wang et al., 2006; van Hinsbergen et al., 2012). The second fact is that traffic phenomena are not deterministic but naturally *stochastic* due to all possible randomness in both supply and demand dynamics. For example, many driving and traveling behaviors, like lane-changing choices, are highly random and they could have significant impact on macroscopic traffic states (Schakel et al., 2012).

Therefore, the output of a traffic forecasting model should always be considered as an input-dependent random variable obeying a probability density distribution (PDF). In this sense, most predictive models in literature fall into one of the two categories, that is, deterministic or probabilistic. Deterministic models aim to build a point-to-point mapping. By minimizing *mean-square-error* (MSE) or *determinant of covariance matrix* (DCM), it predicts the *mean* of the output's PDF. In contract, probabilistic models describe the joint PDF of input and output random variables and learn to directly give output's PDF by minimizing *negative-log-likelihood* (NLL). We specify that this classification only depends on input–output formulation. Taking the example of a deterministic model, one may use explicit traffic modeling (Ben-Akiva et al., 1998), Kalman-filter-based methods (Wang et al., 2006; van Hinsbergen et al., 2012), or black-box deep neural networks (Ma et al., 2017). There may exist random variables inside the model (such as Bayesian networks van Hinsbergen et al., 2009)—whatever is used within such a model, if the final output is an estimate of mean value, it is a deterministic model.

NLL and MSE describe predictive uncertainty from different aspects so it is necessary to consider two corresponding metrics of predictability. Thus, we come up with the following research question central to this paper:

If traffic forecasting is formulated as a self-regressive task, given a dataset, what are the model-free, theoretical lower bounds of predictive performance for probabilistic models and deterministic models respectively?

The answer to this question is highly relevant to researchers. It gives a more objective assessment of data-driven models and puts bench-marking more and more complex models into perspective. In this paper, we use key concepts in information theory to estimate the limit of predictability in short-term traffic forecasting. Theoretically, conditional differential entropy gives the rigorous limit of the expectation of NLL. Then extended continuous form of Fano's theorem further gives a soft lower bound of the expectation of MSE/DCM. Here both metrics are indices of model-independent *average* limit of predictability. Whatever model is run on a large enough dataset, the *expectation* of NLL/MSE/DCM cannot reach the lower bound.

Another concern is that the uncertainty during some time slots and at some locations in a road network could be much higher and causes much higher predictive errors. For instance, when an on-ramp will be saturated and when a new congestion bottleneck will start is highly uncertain. Congestion propagation in a road network also largely depends on whether queues spill over some specific intersections and off-ramps (Van Lint et al., 2012; Knoop et al., 2015). Identifying the most uncertain (the least predictable) time-of-day and network locations from data is valuable for traffic managers. In this study, two special properties, cyclostationarity and localized spatial correlations, are considered in the entropy estimation scheme. So the limit of predictability can be formulated as a function of space and time-of-day. The key contributions of this paper are:

- Estimate the theoretical *spatio-temporal* lower bound of predictive error for both deterministic and probabilistic traffic forecasting models.
- Quantify how observation range and prediction horizon influence the limit of predictability.
- Identify the most unpredictable time slots and locations in a road network directly from data.
- Illustrate that the stochasticity of traffic dynamics and improper assumptions on output distribution are two major bottlenecks for further improving predictive accuracy.

The remainder of this paper is organized as follows. Section 2 presents the background knowledge and related works in literature. Section 3 describes the proposed method, including theoretical basis, implementation of spatio-temporal dependencies, and the numerical scheme to estimate the limit. Section 4 shows the results and gives analysis of numerical experiments. Section 5 finally draws conclusions and proposes several related research directions.

2. Background

2.1. Preliminaries

This subsection introduces the entropy measures in information theory and some basic concepts of discrete-time stochastic processes. In information theory, the central concept of entropy was first-time induced by Shannon to quantify the information content of a discrete random variable (Shannon, 1948). Theoretically the Shannon entropy of continuous random variables is infinity. To extend this concept, *differential entropy* of a continuous random variable V with probability distribution function $p_V(v)$ supported on $\mathcal{V}$ is proposed and defined as follows:

$$H(V) = -\int_{\mathcal{V}} p_V(v) \ln p_V(v) dv \tag{1}$$

Higher entropy means higher uncertainty. For two continuous random variables, the conditional (differential) entropy of X given Y is defined as:

$$H(X|Y) = H(X,Y) - H(Y) = -\int_{\mathcal{X},\mathcal{Y}} p_{X,Y}(x,y) \ln p_{X|Y}(x|y)\ dxdy \tag{2}$$

where $\mathcal{X}, \mathcal{Y}$ denote the support sets of X and Y. Conditional entropy measures how much additional information is carried by X when side information Y is known. It represents the average *additional* uncertainty of output. $H(X|Y) = H(X)$ if and only if X and Y are independent.

For a state-space system with n observable variables, the evolution of system state can be written as a n-dimensional time series $\{\boldsymbol{X}_t\}$, or a so-called multivariate *stochastic process*. Herein $\boldsymbol{X}_t \in \mathbb{R}^n$ represents the n-dimension system state observed at time t. When this system transits from old states to a new state, new information (uncertainty) is produced in addition to the old information carried by the historical observations. For stochastic processes, *stationarity* is one of the most important properties. A stationary process is defined as a stochastic process whose unconditional joint probability distribution of sub-sequences of any length does not change with time shifting:

$$p(\boldsymbol{X}_{t_1}, \dots, \boldsymbol{X}_{t_n}) = p(\boldsymbol{X}_{t_1+\tau}, \dots, \boldsymbol{X}_{t_n+\tau}), \quad \forall \tau, t_1, \dots, t_n \in \mathbb{R}, \quad \forall n \in \mathbb{N} \tag{3}$$

It means that statistical properties do not change with time. To facilitate the narrative, from now on we denote $\boldsymbol{X}_{t-m:t} = \{\boldsymbol{X}_{t-m}, \dots, \boldsymbol{X}_{t-1}\}$ as the past m step observations from t; $\boldsymbol{X}_{t:t+p} = \{\boldsymbol{X}_t, \dots, \boldsymbol{X}_{t+p-1}\}$ as the next p step states. m is called *observation window* and p is *prediction horizon*. When predicting $\boldsymbol{X}_{t:t+p}$ from given side information $\boldsymbol{X}_{t-m:t}$, predictive uncertainty can be measured by conditional entropy $H(\boldsymbol{X}_{t:t+p}|\boldsymbol{X}_{t-m:t})$. If $p = 1$ (1-step prediction), we have the so-called *entropy rate*:

$$S(\boldsymbol{X}_t) = \lim_{m\to\infty} H(\boldsymbol{X}_t|\boldsymbol{X}_{t-m:t}) \tag{4}$$

For stationary processes, or at least asymptotically stationary processes, both conditional entropy and entropy rate are time-independent. Information is statistically generated at a constant rate. And thus predictability is a constant.

2.2. Related works

Predictability quantification is always an important topic. For a complex system with unknown undergoing data generation process, such as traffic networks, this limit has to be estimated from collected observations (dataset). We observe three major approaches in literature.

One of the most widely-used metrics of predictability is *Lyapunov exponent* (Wolf et al., 1985) in chaos analysis. It characterizes how sensitive a deterministic process is to disturbed initial conditions or measures the stability of a stochastic process. Estimating Lyapunov exponents from time series firstly requires phase space reconstruction (PSR) through certain techniques like delayed embedding (Packard et al., 1980; Rosenstein et al., 1993). Specific to traffic time series, Nair et al. (2001) and Shang et al. (2005) use this method to analyze the chaos of scalar traffic time series and show that both univariate speed and flow series have positive maximum Lyapunov exponent, which is a signature of chaos. Some papers combine chaos analysis with other methods to predict traffic states. For example, Li et al. (2016) uses a two-level framework. Different sources of data (speed, flow, occupancy) are firstly processed in lower dimensional space, and then PSR embeds and fuses initial flow series and processed flow series into a higher dimensional space with the assistance of Bayesian estimation theory. The embedded data are then fed into a radial-basis-function (RBF) neural networks to give predictions. However, Lyapunov exponent has several shortcomings. First, in most cases extending this scheme to correlated multivariate time series is challenging. The studies mentioned above only consider univariate time series. The difficulty mainly originates from PSR. Embedding usually maps the original multivariate time series into an unnecessarily high-dimensional phase space (Lan et al., 2008), which is numerically challenging. Second, Lyapunov exponents cannot be directly related to predictive errors in state-space. Instead it gives an average separation rate. These drawbacks limit its applications.

The second strategy is maximum likelihood learning. With the development of deep neural networks (DNN) techniques, this method is becoming mainstream. It assumes that the output obeys an input-dependent prior distribution (such as Gaussian). Parameters of this distribution (like mean and variance) are learnt by a DNN through minimizing NLL. This approach enjoys many advantages. First, it allows estimating the inherent randomness of each prediction. Second, NLL minimization is easy to be implemented in an end-to-end training process so the power of DNN can be released. Specific to traffic forecasting, most papers in

literature consider traffic time series as a Gaussian process (Idé and Kato, 2009; Yuan et al., 2021). The major drawback is that we have to use distributions *"a prior"* to approximate the true but unknown distribution. The true distribution might be complex, such as a mixture, or even multi-modal. If we use a simple uni-modal distribution to approximate it, NLL may not reach a desired low value.

The third solution is entropy-based approach. One of the earliest attempts to analyze the predictability of univariate time series based on conditional entropy was proposed by Song et al. (2010), in a discrete form. The authors studied one-step predictability of human mobility based-on the mobile phone call position database. The limit of one-step predictability is defined as the maximum probability of predicting a user's correct position area in the next moment given the observations of the past trace. The Upper Bound of Predictability (UBP) is given from entropy rate by the famous Fano's theorem (Cover, 1999). The entropy rate of finite stationary time series can be estimated by Lempel–Ziv coding algorithm (Kontoyiannis et al., 1998).

This method has been widely applied in many domains to estimate the UBP of stationary univariate time series, including traffic and transportation. These studies basically use a similar strategy to process continuous variables: continuous univariate time series are discretized into several "states" to compute UBP. UBP here can be interpreted as *the maximum probability of giving a prediction whose MSE is smaller than the square of discrete size*. For example, Wang et al. (2015) investigates the UBP of traffic speed on a ring freeway. Each sensor on the network is assumed to be independent from each other and speed is discretized into a few ranges. Li et al. (2019) extends this method to continuous univariate series by measuring the similarity of two sequences. If the distance between them is smaller than a pre-defined tolerance, then they are counted as "the same". So the concept of Lempel–Ziv entropy can be extended and it can be regarded as a new metric of predictability. Li et al. (2019) uses this method to measure the UBP of travel time, etc. Some papers also try to avoid discretization by using differential entropy. For example, Darmon (2016) directly estimates differential entropy rate from stationary time series to represent the inherent unpredictability. Amigó et al. (2017) proposed an ignorance score based on differential conditional entropy to represent models' prediction quality. However, this approach has several drawbacks.

The first is the stationarity assumption. All studies above assume that traffic quantities form a stationary time series. But this does NOT hold for many traffic series. Many traffic phenomena show strong time-of-day-related periodicity. If Lempel–Ziv coding algorithm, or other entropy estimators such as non-linear embedding estimator, are directly applied to non-stationary time series, the entropy rate, and thus UBP, would be overestimated. Xiong et al. (2017) gives a systematic study on this topic. We refer the readers to this paper for more details. Second, sensors and links cannot be considered independent for network-level traffic forecasting. In many phenomena, like the spreading of congestion, the traffic state of a link is strongly correlated with its topological neighbors. We emphasize that time index and spatial correlations must be included in the estimation of limit of predictability.

To address these issues, our approach explicitly formulates conditional differential entropy as a time- and space-related quantity. Two special properties of traffic network dynamics, temporal cyclostationarity and localized spatial correlations, are used to split all data into subsets. Based on estimated conditional entropy, we derive the lower bound of NLL for probabilistic models and the lower bound of MSE for deterministic models.

3. Methodology

This section presents details of the proposed entropy-based approach. First we give theoretical analysis. Next, we show how to implement spatiotemporal correlations into a network-level predictability estimation scheme. The last subsection further introduces the used entropy estimator, the so-called kp-Nearest neighbors (kpN) estimator.

3.1. Theory

Consider two random variables $\boldsymbol{X} \in \mathbb{R}^m$ (input) and $\boldsymbol{Y} \in \mathbb{R}^n$ (output). Their joint PDF can be written as:

$$p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}) = p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}) p_{\boldsymbol{X}}(\boldsymbol{x}) \tag{5}$$

If we precisely know the conditional density function $p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y})$ for every input, then the problem is solved. We can directly use its differential entropy or covariance matrix to quantify predictive uncertainty. Unfortunately, this is infeasible in practice. When collecting data, one cannot know output distribution but just observe a series of input–output pairs. For one specific input, we have to find other input samples that are close enough in **phase-space**, and use their corresponding observed outputs to estimate the true output distribution. However, as explained in the discussion on Lyapunov exponent and PSR in Section 2.2, this is a challenging and unsolved topic. So we come up with a compromise solution. Instead of constructing a continuous PDF in probability space, the input range is relaxed according to some external evidence (prior knowledge) and a scalar *average* entropy measure is computed. This approach avoids mapping inputs into phase space and also results in sufficient samples to support entropy estimation.

Because $p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y})$ is unknown, a probabilistic model uses a prior distribution, noted as $q_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y})$, to approximate it. We have the following theorem:

Theorem 1 (*Limit of NLL*). *Consider two multivariate random variables $\boldsymbol{X} \in \mathbb{R}^m$ and $\boldsymbol{Y} \in \mathbb{R}^n$. A model estimates $p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y})$ by an approximated prior $q_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y})$, then the expectation value of NLL for any probabilistic model obeys the following inequality:*

$$\mathbb{E}_{p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})}[NLL] \geq H(\boldsymbol{Y}|\boldsymbol{X}) \tag{6}$$

Equality holds (the lower bound is reached) if and only if that $\forall \boldsymbol{x}$, $p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}) = q_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y})$ almost everywhere ("almost everywhere" means that $p - q$ has measure 0).

Proof. The expectation of NLL for one given $\boldsymbol{x}$ is:

$$-\mathbb{E}_{\boldsymbol{Y}\sim p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y})}[\ln q_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{Y})] = -\mathbb{E}_{\boldsymbol{Y}\sim p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y})}[\ln p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{Y})] + \mathbb{E}_{\boldsymbol{Y}\sim p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y})}[\ln \frac{q_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{Y})}{p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{Y})}] \tag{7}$$

The first term on the right is the entropy of $\boldsymbol{Y}$ at given $\boldsymbol{X}=\boldsymbol{x}$, the second term is *Kullback–Leibler (KL) divergence*, noted as $D_{KL}(q \mid p)$, which is non-negative because of Gibbs' inequality. $D_{KL}(q \mid p) = 0$ if and only if $p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}) = q_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y})$ almost everywhere. Now we apply expectation over input space $p_{\boldsymbol{X}}(\boldsymbol{x})$ on both side:

$$\begin{aligned}
\mathbb{E}_{p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})}[\text{NLL}] &= -\int_{\mathcal{X}} p_{\boldsymbol{X}}(\boldsymbol{x})[\int_{\mathcal{Y}} p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}) \ln p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}) d\boldsymbol{y}] d\boldsymbol{x} + \mathbb{E}_{p_{\boldsymbol{X}}(\boldsymbol{x})}[D_{KL}(q \mid p)] \\
&= -\int_{\mathcal{X},\mathcal{Y}} [p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}) p_{\boldsymbol{X}}(\boldsymbol{x})] \ln p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}) d\boldsymbol{y} d\boldsymbol{x} + \mathbb{E}_{p_{\boldsymbol{X}}(\boldsymbol{x})}[D_{KL}(q \mid p)] \\
&= H(\boldsymbol{Y}|\boldsymbol{X}) + \mathbb{E}_{p_{\boldsymbol{X}}(\boldsymbol{x})}[D_{KL}(q \mid p)]
\end{aligned} \tag{8}$$

Because of the non-negativity of KL divergence, Theorem 1 is proved. □

As the theorem says, the lower bound can be reached if and only if the output distribution of every input is perfectly modeled, no matter what the distribution is. $H(\boldsymbol{Y}|\boldsymbol{X})$ is a measure of *data uncertainty*. It describes the inherent randomness of data generation process. Higher data uncertainty means lower predictability. The distance between the NLL of a model and $H(\boldsymbol{Y}|\boldsymbol{X})$ is the *model uncertainty*, which is the additional uncertainty caused by model abstraction (Lee et al., 2017). This gap is mainly determined by how well the prior distribution can represent the true distribution. Bigger gaps imply that this probabilistic model cannot give reliable estimates of input-dependent data uncertainty. Theorem 1 gives the *optimal* lower bound for probabilistic models.

Entropy is not the only metric of uncertainty. We also want to derive a lower bound of MSE/DCM for deterministic models. We show the following theorem:

Theorem 2 (*Multivariate Fano's Theorem*)**.** *Consider two multivariate random variables $\boldsymbol{X} \in \mathbb{R}^m$ and $\boldsymbol{Y} \in \mathbb{R}^n$. If $\boldsymbol{Y}$ is predicted based on side information $\boldsymbol{X}$, then there exists a lower bound of the determinant of the expectation of covariance matrix for any point-estimate model:*

$$\det \mathbb{E}_{p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})}[(\boldsymbol{Y}-\hat{\boldsymbol{Y}})(\boldsymbol{Y}-\hat{\boldsymbol{Y}})^T] \geq \frac{1}{(2\pi e)^n} e^{2H(\boldsymbol{Y}|\boldsymbol{X})} \tag{9}$$

The lower bound is reached if and only if the error $(\boldsymbol{Y}-\hat{\boldsymbol{Y}})$ is 0-mean Gaussian and independent from $\boldsymbol{X}$.

Proof. Given an input $\boldsymbol{x}$, the point-estimate of output is $\hat{\boldsymbol{y}}$, the predictive error $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ is a random variable in $\mathbb{R}^n$. Because entropy is translation invariant ($\hat{\boldsymbol{y}}$ is a constant), we can always assume that the mean of $\boldsymbol{e}$ is 0 (un-biased estimator). If we note the covariance matrix of $\boldsymbol{e}$ as $\boldsymbol{K} = \mathbb{E}_{\boldsymbol{Y}\sim p_{\boldsymbol{Y}|\boldsymbol{Y}=\boldsymbol{x}}(\boldsymbol{y})}[\boldsymbol{e}\,\boldsymbol{e}^T]$, Cover (1999, pg.254) shows that the following inequality holds for all distributions (if $\det \boldsymbol{K}$ exists):

$$\frac{1}{2}\ln\det(2\pi e\boldsymbol{K}) \geq H(\boldsymbol{e}|\boldsymbol{X}=\boldsymbol{x}), \quad H(\boldsymbol{e}|\boldsymbol{X}=\boldsymbol{x}) = H(\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}) \tag{10}$$

Equality holds if and only if $\boldsymbol{e}$ is Gaussian. Again, we apply expectation over input space on both side:

$$\frac{1}{2}\int_{\mathcal{X}} p_{\boldsymbol{X}}(\boldsymbol{x}) \ln\det(2\pi e\boldsymbol{K}) d\boldsymbol{x} \geq H(\boldsymbol{Y}|\boldsymbol{X}) \tag{11}$$

As shown in (8), the right side is conditional entropy. Because $\ln \circ \det$ is concave, Jensen's inequality gives:

$$\frac{1}{2}\ln[(2\pi e)^n \det(\int_{\mathcal{X}} p_{\boldsymbol{X}}(\boldsymbol{x})\boldsymbol{K} d\boldsymbol{x})] \geq \frac{1}{2}\int_{\mathcal{X}} p_{\boldsymbol{X}}(\boldsymbol{x}) \ln\det(2\pi e\boldsymbol{K}) d\boldsymbol{x} \tag{12}$$

where equality holds if and only if $\boldsymbol{K}$ is independent from $\boldsymbol{X}$. The integral on the left is actually the expectation of determinant of covariance matrix (DCM). By combining (11) and (12), Theorem 2 is proved. □

Fang et al. (2019) provides an alternative proof of this theorem. When the lower bound is reached, the relationship between input and output can be written as $\boldsymbol{Y} = f(\boldsymbol{X}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \boldsymbol{K})$. $\hat{\boldsymbol{Y}} = f(\boldsymbol{X})$ theoretically can be precisely modeled by an un-biased estimator and $\mathcal{N}(0, \boldsymbol{K})$ is the inherent randomness that cannot be explained out or reduced. This lower bound is not as tight as the one in Theorem 1 because MSE and DCM measures ignore the structural information of output distribution. But it still gives a limit of any models' capability. The room of improvement for modeling can only be smaller than the gap to this limit.

Another point is that the $n \times n$ covariance matrix of $\boldsymbol{Y} = (y_1, y_2, \ldots y_n)$ is hard to learn. A probabilistic model generally assumes the prior form of each marginal distribution $p(y_i)$. So a better choice is estimating $H(y_i|\boldsymbol{X})$ for each component and obtain a series of variance limits, $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$. Their relationship is given by the following formula:

$$\det \boldsymbol{K} = \det \boldsymbol{S} \prod_{i=1}^{n} \sigma_i^2 \leq \prod_{i=1}^{n} \sigma_i^2 \tag{13}$$

where $\boldsymbol{S}$ is the correlation matrix. $\det \boldsymbol{S} \leq 1$ and $\det \boldsymbol{S} = 1$ if and only if all components of $(y_1, y_2, \ldots, y_n)$ are independent. Most deterministic models use MSE as loss function, the corresponding lower bound is:

$$\mathbb{E}[MSE] \geq \frac{1}{n}\sum_{i=1}^{n} \sigma_i^2 \tag{14}$$

To quantify the correlation in a multistep prediction, we can use the concept of *conditional mutual information*:

$$I(y_1, y_2, \dots, y_n \mid \boldsymbol{X}) = \sum_{i=1}^{n} H(y_i|\boldsymbol{X}) - H(\boldsymbol{Y}|\boldsymbol{X}) \tag{15}$$

This is a non-negative quantity. It equals to 0 if and only if all components $(y_1, y_2, \dots, y_n)$ are independent.

In summary, we have theoretically shown the lower bound of the expectation of NLL for probabilistic models and the expectation of MSE/DCM for deterministic models. If we do not consider the numerical difficulties in entropy estimation, these two limits are in-principle applicable for all traffic scenarios described by a set of observable quantities without any prerequisites. Conditional entropy is the core concept bridging them. Estimating conditional entropy requires applying expectation in a subset of input space. Next we will describe how to split the entire dataset into subsets and how to formulate conditional entropy as a function of space and time.

3.2. Spatio-temporal correlations

The spatio-temporal evolution of a traffic quantity on a road network with N links or sensors can be written as a N-dimensional time series. Assume that we collected D days of data, the dataset is noted as $\{\boldsymbol{X}_{d,t}\}$. Here d is day index and t is time-of-day. $X^i_{d,t}$ is the observed value on day d, time of day t, and link i. Considering the quasi-periodical tendency of traffic phenomena, we assume that this multivariate time series has *cyclostationarity*, which means: (1) For any m and p, the conditional entropy $H(\boldsymbol{X}_{d,t:t+p} \mid \boldsymbol{X}_{d,t-m:t})$ changes periodically every 24 h; (2) and it is Lipschitz continuous. The first point says that conditional entropy is time-of-day-dependent. And the second point allows inducing a hyperparameter called *smoothing window* δ. We estimate a conditional entropy from all samples in the interval $[t-\delta, t+\delta)$ (from all days) to represent the condition entropy at t. This smoothing window increases estimation accuracy by including more samples and it also smooths the resulted curve. For example, we prepare the following input–output set to estimate $H(\boldsymbol{X}_{t:t+p} \mid \boldsymbol{X}_{t-m:t})$:

$$\{(\boldsymbol{X}_{d,\tau-m:\tau}, \boldsymbol{X}_{d,\tau:\tau+p}) \mid d \in D \text{ and } \tau \in [t-\delta, t+\delta)\} \tag{16}$$

However, directly estimating $H(\boldsymbol{X}_{t:t+p} \mid \boldsymbol{X}_{t-m:t})$ is difficult when N is large. So the strategy of "divide and conquer" is used to further decompose the subset. We induce the assumption of *localized spatial correlation*. Notice the fact that any kinetic waves can only move bidirectionally along the road with a speed lower than a maximum positive value c_r. Not all components in $\boldsymbol{X}_{t-m:t}$ can influence the prediction of one sensor $\boldsymbol{X}^i_{t:t+p}$. We can therefore draw a *spreading cone* from the latest vertex X^i_{t+p-1} in the spatio-temporal graph. The semi-vertex angle satisfies $\tan\theta = dl/dt = c_r$. All points outside this cone are independent from $\boldsymbol{X}^i_{t:t+p}$ because their impact cannot reach location i in the next p steps. By combining cyclostationarity and localized spatial correlations, the subset for location i is:

$$\{(\boldsymbol{X}^{i,\text{input}}_{d,\tau-m:\tau}, \boldsymbol{X}^i_{d,\tau:\tau+p}) \mid d \in D \text{ and } \tau \in [t-\delta, t+\delta)\} \tag{17}$$

$$\boldsymbol{X}^{i,\text{input}}_{d,\tau-m:\tau} = \{X^j_{d,s} \mid |r(j,i)| \le c_r(\tau+p-s)\Delta t \text{ and } s \in [\tau-m, \tau)\} \tag{18}$$

where $r(j,i)$ is the directional spatial distance between two positions. $r(j,i)$ is positive if j locates at the upstream of i. Δt is time interval. $\boldsymbol{X}^{i,\text{input}}_{d,\tau-m:\tau}$ is a collection of all points in the spreading cone.

In practice not all points in a spreading cone contain effective information. For further simplification, a spreading cone can be divided into several sub-areas (Fig. 1):

(1) *self:* only the past traffic states of the target position itself.
(2) *upstream cone: self* plus the data points that locate upstream of the target location in the spreading cone.
(3) *downstream cone: self* plus the downstream data points in the cone.
(4) *up/downstream edge: self* plus the data points that are close enough to the up/downstream surface of the cone.

Theoretically the predictability of *self* should be the lowest while the others are all higher because *self* considers all links independently. By comparing the limits of *upstream cone* and *downstream cone*, we can determine that this quantity is dominated by the information from upstream, downstream, both, or neither. If results show that the prediction mainly depends on, for instance, upstream traffic states, we next compare the limits of *upstream cone* and *upstream edge* to check the possibility of further simplification. Notice that we do not try to carefully tune c_r because this is hard in practice. But choosing an estimated upper bound of propagation speed is much easier.

For univariate traffic series forecasting, which is a simpler case without spatial correlations, only cyclostationarity needs to be considered. (16) can be directly used.

In summary, given a time-of-day t, a location index i (only for multivariate series), time interval Δt, observation range m, prediction horizon p, smoothing window δ, then this input–output sample set can be prepared by the procedure above. Therefore, conditional entropy estimated from this set represents the predictability at time-of-day t and location i. Now we need a proper entropy estimator.

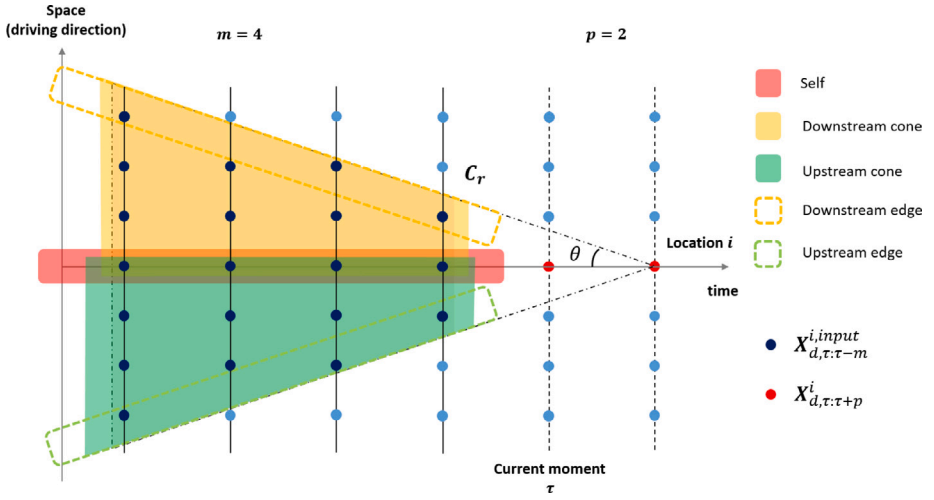

Fig. 1. Illustration of localized spatial correlation: an example of input–output pairs. The dash–dot line triangle is the spreading cone; Sub-areas are marked by different colors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.3. k-p nearest neighbors entropy estimator

Estimating differential entropy from given finite numbers of samples is a challenging topic. Entropy estimators can be roughly categorized into two groups: parametric and non-parametric approaches. For parametric estimators, the PDF's form is assumed to be known so its parameters can be learnt from the samples. However, this assumption is too strong. In most real-world cases an "a priori" known form is impractical. Consequently, non-parametric approaches have been proposed, such as embedding/non-uniform embedding estimator (Faes et al., 2011) and k-nearest neighbors estimator (Wang et al., 2009). In this study we choose the k-p nearest neighbors estimator proposed by Lombardi and Pant (2016). Compared to k-nearest-neighbors (kNN) estimator, the core innovation of kpN is that the uniform distribution for k-nearest samples is replaced by a fast decaying normal distribution whose parameters are determined by larger p-nearest neighbors. We emphasize one fact: *most traffic patterns tend to fall into several clusters and there are few rare patterns locating between them.* This property has been shown by some studies on congestion patterns recognition and classification (Lopez et al., 2017; Krishnakumari et al., 2017; Nguyen et al., 2019). So the PDF should have several peaks for these clusters and its value should be low between them. In this case, kNN estimator will overestimate the entropy. kpN can mitigate this structural error.

The algorithm is given in Algo. 1. This estimator needs to calculate one Gaussian distribution and one corresponding integral:

$$g(\boldsymbol{x}) = \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{19}$$

$$G(\boldsymbol{x}) = \int_{B(\boldsymbol{x},\epsilon)} g(\boldsymbol{x})d\boldsymbol{x} \tag{20}$$

The major drawback of kpN estimator is the relatively higher computational complexity to calculate (20). As pointed in Lombardi and Pant (2016), this process can be accelerated by using the method proposed in Cunningham et al. (2011). The kpN estimator is naturally parallel. Using GPU and multi cores can significantly reduce the running time. For an input–output set, kpN estimator gives the entropy of input and the joint entropy of input–output, their difference is the estimated conditional entropy (see (2)).

In summary, this section explains the theoretical basis of the proposed method. Both spatial and temporal factors are considered to split the entire dataset into a series of subsets. Conditional entropy is estimated from these subsets by kpN estimator. Then Theorems 1 and 2 gives two different metrics of predictability that depends on locations and time-of-day.

4. Experiment

The proposed method will be tested by using real-world datasets in this section. All data used in this paper are provided by National Data Warehouse for Traffic Information (NDW, Netherlands).

4.1. Data description

The major counter-clockwise ring freeway around Rotterdam (The Netherlands) is selected as a case study (shown in Fig. 2). Average speed V and vehicular flow Q per lane are recorded by 201 loop detectors that are not uniformly distributed. Carriageway averaging and the Adaptive Smoothing Methods (ASM) (Kawata and Minami, 1984; Treiber and Helbing, 2003) is used to estimate

Algorithm 1: kpN entropy estimator (Lombardi and Pant, 2016)

Input;
$\boldsymbol{X}_{N\times d}$, N observation samples of dimension d random variables;
k, number of nearest neighbors to calculate local probability mass;
p, number of nearest neighbors to calculate statistical quantities;
Output;
$\hat{H}(X)$, estimation of entropy;
Calculate $C = \varphi(N) - \varphi(k)$ (φ is digamma function);
for *each sample* $\boldsymbol{X}_i$ **do**
  Find p nearest neighbors $\{\boldsymbol{X}\}_i^p$ based-on chebyshev distance;
  Find the distance between $\boldsymbol{X}_i$ and its k-th nearest neighbor, noted as ϵ_i;
  Calculate mean $\boldsymbol{\mu}_i$, covariance matrix $\boldsymbol{\Sigma}_i$, and $\det(\boldsymbol{\Sigma}_i)$ from $\{\boldsymbol{X}\}_i^p$;
  Calculate the neighborhood containing $k-1$ nearest neighbors $\mathcal{B}(\boldsymbol{X}_i) = \boldsymbol{X}_i \pm \epsilon_i \boldsymbol{e}$;
  Calculate g_i in Eq. (19);
  Calculate the integral G_i in Eq. (20);
end
Calculate $\hat{H}(X) = C + \mathbb{E}[\ln G_i] - \mathbb{E}[\ln g_i]$;

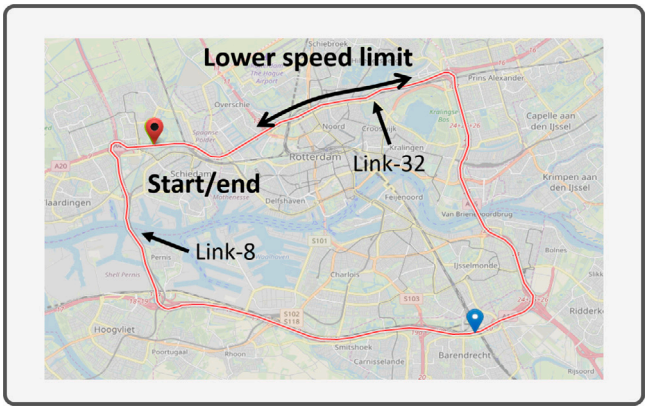

Fig. 2. The counter-clockwise ring freeway around Rotterdam. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

a continuous smoothed spatiotemporal maps of carriageway averaged speeds and flows. The calculation process explicitly considers kinematic wave theory and a wave speed estimator is employed to estimate critical parameters. The ASM is used to fill the missing data (about 3%) and to project V and Q onto uniform spatial–temporal grids. The implementation details can be found in Schreiter et al. (2010a). In this section we study the limit of predictability of the processed dataset. Data stream starts from 5:00 AM and ends at 24:00 PM every day. Considering that holidays and weekends have very different traffic patterns, we only prepared 233 workdays of data from the year of 2018. Two representative traffic forecasting tasks are formulated:

- **Univariate accumulation forecasting**: We aim to predict the total number of vehicles running on the target network (the so-called *accumulation*, noted as N_t). N_t can represent averagely how busy the highway is. It is an index of traffic demand. This is a typical univariate time series forecasting task. The ASM firstly maps non-uniform speed and flow data onto a $0.1\,\mathrm{km}\times 30\,\mathrm{s}$ uniform grid, accumulation is estimated by:

$$N_t = l\sum_{i=1}^{L} n_i \times \frac{Q_t^i}{V_t^i} \tag{21}$$

  where L is the number of uniform links; $l = 0.1\,\mathrm{km}$ is the length of link (spatial resolution); n_i is the number of lanes on each link. Then N_t is aggregated every $5\,\mathrm{min}$ by averaging to form a time series.
- **Multivariate speed forecasting**: We aim to predict speed evolution on the ring freeway in the near future. Speed describes when and where congestion emerges, evolves, and dissipates. Similarly, ASM firstly maps V onto a $0.1\,\mathrm{km}\times 30\,\mathrm{s}$ uniform grid, then the processed data is aggregated every $1.2\,\mathrm{km}$ and $4\,\mathrm{min}$ by averaging. So the entire ring freeway is divided into 35 uniform links. The processed dataset forms a 35-D time series. This is a typical network-level traffic forecasting task.

4.2. Predictive models

For accumulation forecasting, we select three baseline deterministic models: (1) **k-nearest neighbors (KNN)**: the similarity metric is Euclidean distance and the weight of averaging is inverse of distance. The optimal number of neighbors is chosen by cross validation. (2) **FCNN**: a fully-connected feed-forward neural networks with 5 hidden layers activated by *sigmoid* function. The numbers of hidden units are sequentially 64, 128, 256, 128, 64. (3) **LSTM** (Gers et al., 1999): a long–short term memory (LSTM) encoder–decoder model with 128 hidden units is used for multistep forecasting. We also construct a simple FCNN probabilistic model with 3 hidden layers. Each hidden layer contains 128 units.

For multivariate speed forecasting, three baseline deterministic models are selected: (1) **KNN**: the similarity metric is Euclidean distance and the weight of averaging is the inverse of it. Optimal number of neighbors is searched by cross validation. (2) **DCRNN**: Li et al. (2018) is one of state-of-the-art network-level traffic forecasting models that employs diffusion convolution and GRU cells to capture spatio-temporal features. (3) **STGCN**: Yu et al. (2017) is another state-of-the-art speed prediction model that has a fully convolutional structure. Here we use the variant, STGCN(cheb), proposed in the paper. Similarly, we propose a STGCN-like probabilistic model with U-Net-like skip connections (Ronneberger et al., 2015). The model details can be found in Appendix.

For probabilistic models, we pose 3 different uni-modal prior distributions:

- **Gaussian**: We assume that the marginal distribution of each component of output is Gaussian. The joint distribution is a multivariate Normal distribution. The last layer outputs mean and variance (μ, σ^2) of each component.
- **Beta distribution**: The marginal distribution of each component of output is a beta distribution $\mathcal{B}(\alpha, \beta)$ with $\alpha > 1$ and $\beta > 1$. So the joint distribution is a Dirichlet distribution. The last layer outputs the mode $\omega = (\alpha - 1)/(\alpha + \beta - 2)$, $\omega \in (0, 1)$ and the concentration $\kappa = \alpha + \beta$, $\kappa \in (2, +\infty)$. By this way the distribution is uni-modal with finite mode.
- **Inverse-Gamma distribution**: The marginal distribution of each component is an Inverse-Gamma distribution $\Gamma^{-1}(\alpha, \beta)$ with $\alpha > 2$ and $\beta > 0$ (to ensure that variance exists). The joint distribution is an Inverse-Wishart distribution. The last layer outputs mean $\mu = \beta/(\alpha - 1)$ and β of each component.

MSE is chosen as the loss function to train deterministic models and NLL is used to train probabilistic models. The dataset is split into a training set (70%), a validation set (10%), and a test set (20%). Early-stopping on validation set is used to mitigate over-fitting. For those models using recurrent encoder–decoder structure, teacher forcing (Lamb et al., 2016) method is used. Because NLL is scale-relevant, for better comparing its lower bound, all data are normalized between 0 and 1 by min–max normalization. To get the limit of MSE with true unit, one simply needs to re-scale the results.

Restricted by the limited number of samples, we cannot guarantee that the training set and the test set are drawn from the same independent identical distribution (i.i.d). The predictive performance in some moments and at some locations MAY occasionally outperform the estimated lower bound. To avoid this contradiction induced by dataset shift, we use *multi-fold* strategy. For each fold, all samples are firstly shuffled and then re-partitioned into new training/test sets. The training set is used to train the models and estimate the theoretical lower bounds; the test set is used to compute predictive errors of baseline models. This process repeats k times and their average predictive accuracy is used to validate the proposed predictability metric. k-fold method is equivalent to creating a compound model that is trained on a dataset that highly-possibly contains all samples. Meanwhile, it can guarantee that estimated predictability does not use any sample in the test set and the sub-model in each fold has not seen any sample in the test set neither. Therefore, dataset shift can be effectively reduced. For example, if the split ratio of training set is 0.7 and $k = 10$, our speed dataset contains 66 405 observations (233 days, 19 h and 4 min interval everyday) for each location, then the expectation of samples that are not included in all k-folds is $66\,405 \times (1 - 0.7)^{10} \approx 0.39$, which is negligible. The estimated lower bound is reliable.

Throughout this section, we choose a fixed smoothing window $\delta = 20\,\text{min}$.

4.3. Accumulation forecasting

Fig. 3 presents the evolution of $N(t)$ from Monday to Friday in a randomly selected week. It shows clear daily quasi-periodicity. There are two peaks that represent morning and evening peak hours respectively, but the time and the height of these two peaks are not exactly the same everyday.

Now we consider a specific accumulation forecasting task with $m = 6$ and $p = 4$ (observe what happened in the past 30 min and predict the accumulation in the next 20 min). In Fig. 4a, the estimated (average) lower bound of NLL for each prediction step is compared with those probabilistic models using different priors. Inverse-Gamma distribution is slightly better than the others. In Fig. 4b, we further compare the temporal curve of estimated limit with the best Inverse-Gamma approximation for each prediction step. This probabilistic model's NLL is indeed above the estimated limit almost everywhere. The similar temporal tendencies validate the cyclostationarity assumption. Generally speaking, accumulation time series is more uncertain during peak hours, especially during evening peak hours. The gap between the model's curve and the limit curve is the additional model uncertainty induced by Inverse-Gamma prior and model abstraction. The gap is significantly bigger for longer-term prediction. But for short-term prediction like 5 min-10 min horizon, Inverse-Gamma distribution is an acceptable prior.

Fig. 5 quantifies the influence of observation range and prediction horizon. In Fig. 5a, the prediction horizon is fixed as 20 min and input range changes from 10 min to 80 min. With the increasing of observation range, the joint conditional entropy of multistep prediction ($H(\boldsymbol{Y}|\boldsymbol{X})$) goes down, which means predictability increases because more effective information is given. Fig. 5b shows that the RMSE limit of each step increases fast with prediction horizon. The difference is more significant during peak hours. For

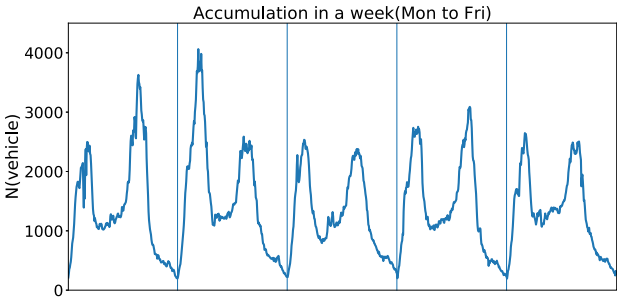

Fig. 3. The evolution of accumulation from Monday to Friday in a randomly selected week.

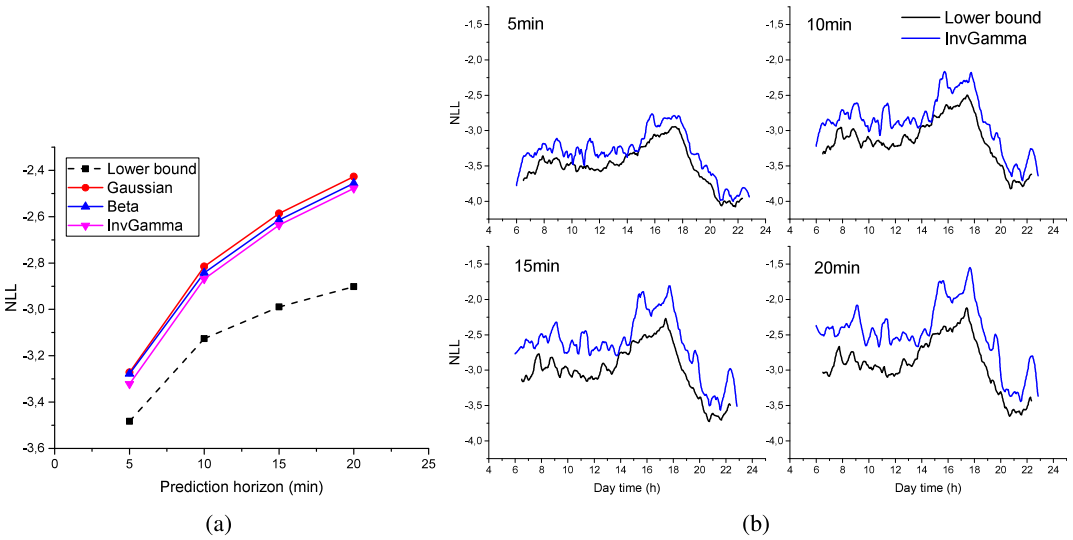

Fig. 4. (a) Comparison between the average lower bound of NLL and the performances of probabilistic models for each prediction step; (b) Comparison between the lower bound of NLL and the performances of the probabilistic model using Inverse-Gamma prior for each prediction step, along time axis. Averaging the lower bound curves in (b) gives the corresponding 4 points (black-square) in (a). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

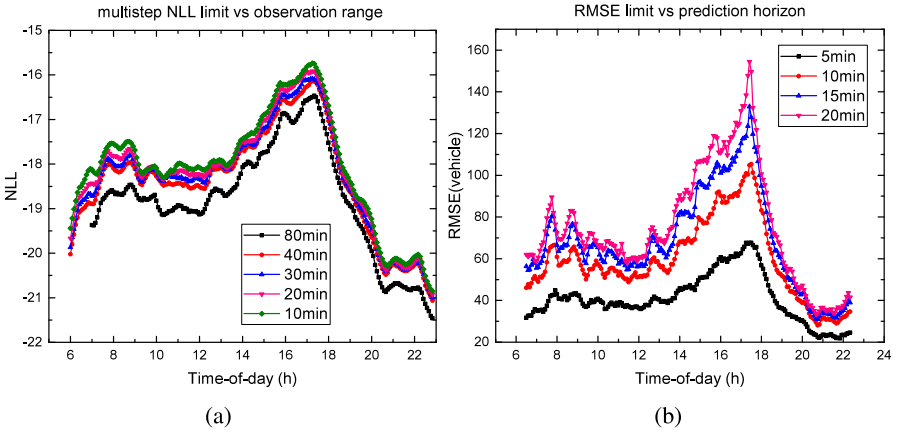

Fig. 5. (a) relationship between observation range and multistep NLL limit, $p = 4$; (b) Relationship between prediction horizon and RMSE limit of each prediction step, $m = 6$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

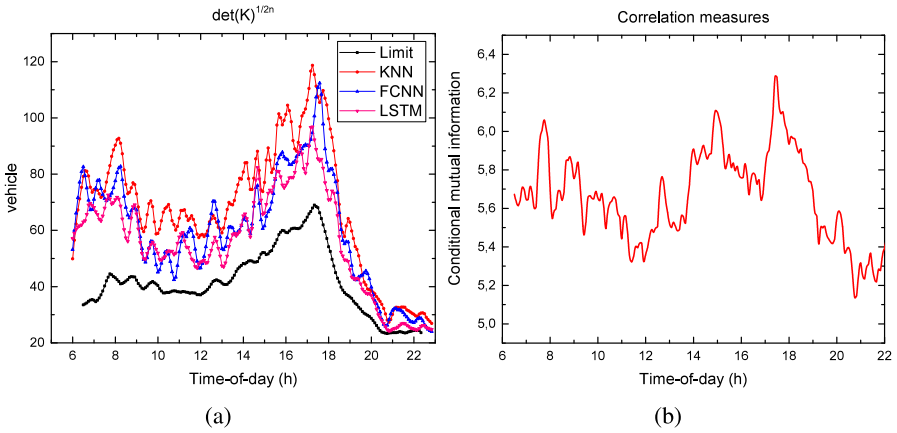

Fig. 6. (a) comparison between the lower bound of DCM and deterministic models' performances; (b) conditional mutual information for 4-step predictions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

example, at around 17:30 PM, the RMSE limit increases from 60 vehicles to maximum 160 vehicles in 20 min. This result implies that accurate long-term prediction is theoretically impossible without inducing more data.

Fig. 6 presents more analysis. 6a compares the estimated lower bound of DCM and the DCMs of deterministic models. Here we show the $2n$-th root of DCM so the unit is consistent with original data. DCMs of the three models are indeed above the estimated limit and their forms are also similar to the lower bound curve. Averagely speaking, LSTM has the best predictive accuracy. According to Theorem 2, the room of improvement for modeling cannot be larger than the gap. 6b shows the conditional mutual information of four-step predictions. The positive value proves that multistep predictions are temporally strongly-correlated. This correlation is even stronger during peak hours. Temporal correlation also makes longer-term prediction more difficult. The errors made on early steps may severely enlarge long-term predictive errors, especially in peak hours. This phenomena is consistent with Fig. 5b.

In summary, we have shown that the proposed metrics of predictability are reasonable for univariate accumulation forecasting. Next we will analyze multivariate speed prediction.

4.4. Multivariate speed forecasting

Different from univariate accumulation prediction, to implement spatial correlations in network-level speed forecasting, we need to induce a hyper-parameter, the upper bound of kinetic wave spreading speed c_r. Low-speed congestion prediction is the core of speed forecasting. Traffic flow theory tells that the maximum back-propagation speed of stop-and-go waves on highways is lower than $20\,\mathrm{km\,h^{-1}}$ (Schreiter et al., 2010b). This value is quite stable and almost the same everywhere. To explore the minimum input set in the spreading cone, we select a short segment that is frequently congested on the west of the ring freeway and test 4 sub-areas, *self*, *upstream cone*, *downstream cone*, *downstream edge*. For simplification, we fix the observation window $m = 6$ (24 min) and only calculate the lower bound of RMSE for 1-step prediction. The results are presented in Fig. 7. The curve of *self* is the highest because inputs contain the least effective information. All sensors are considered independent from each other in *self*. The *upstream cone* curve is slightly lower than *self* but the *downstream cone* curve is significantly lower than *self*. This means that the past traffic states of upstream links contain very little effective information. But downstream links have much more useful information for accurate predictions. Because the back-propagation of stop-and-go kinetic waves is important in congestion forecasting. Further, by comparing *downstream cone* and *downstream edge*, we conclude that all upstream links in spreading cone contain effective information, since the *downstream edge* curve locates between *self* and *downstream cone*. It indicates that the back-propagation speed of information may not be a constant, or it is not a constant close to c_r.

The analysis above points out that the minimum effective input for speed forecasting is *downstream semi-cone*. All the following results are calculated based on this input set.

We firstly consider a forecasting task with $m = 6$ and $p = 1$. The result is shown in Fig. 8. The lower bounds for other m and p have similar spatio-temporal distributions but different magnitudes (similar to what has been shown in Fig. 5). Temporally, there exist two less predictable peak hours: morning (7:00 AM–9:00 AM) and evening (16:00 PM–19:00 PM). Evening peak hour is even more uncertain. Spatially, there are two less predictable segments, one locates between 0 km-5 km and the other one is between 25 km-35 km. Between 35 km-40 km there is a highly predictable band (the deep blue areas). Because the speed limit is lower there (shown in Fig. 2).

Similar to accumulation forecasting, here we consider a specific speed forecasting task with $m = 6$ and $p = 4$ (observation range is 24 min and prediction horizon is 16 min). In Fig. 9, the average lower bound of NLL (over all locations and time-of-day) for each prediction step is compared with those probabilistic models using different prior distributions. Their gaps to the limit show that Beta distribution is the best approximation among the three priors while Gaussian is the worst. Speed is supported between 0 and a maximum limit. In congested areas speed is low so the Gaussian prior may cause *probability leakage*: the PDF on negative axis has no

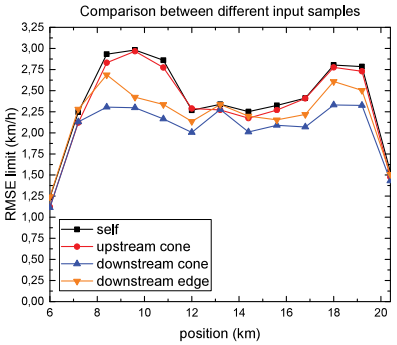

Fig. 7. The lower bound of RMSE for different input sets: $m = 6$ and $p = 1$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

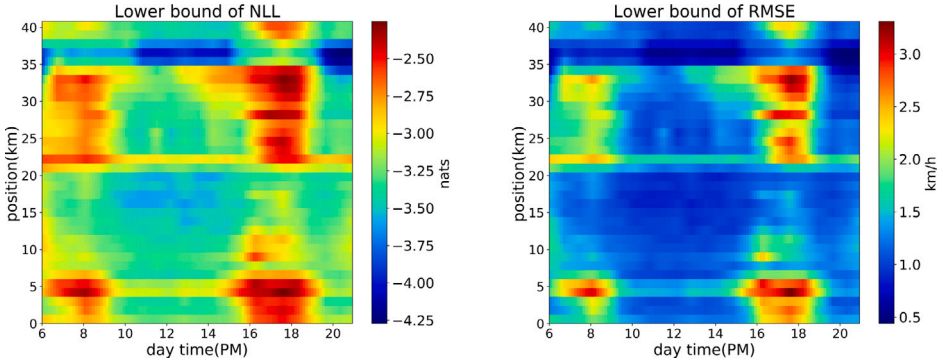

Fig. 8. The spatio-temporal lower bound of NLL (left) and RMSE (right) for speed forecasting, $m = 6$ and $p = 1$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

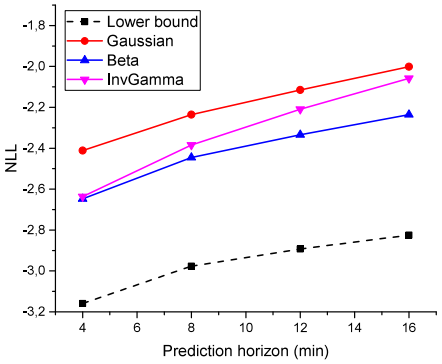

Fig. 9. Comparison between average lower bound of NLL and the performances of probabilistic models for each prediction step in speed forecasting.

meaning. So the result implies that the true distribution of speed may be highly skewed. Different from accumulation forecasting, the model uncertainty here almost does not change with prediction horizon. It implies that the true distribution of speed is complex.

Next we study this spatio-temporal limit of predictability by slicing. We will select some representative examples. For temporal predictability, we select the link with the lowest average speed (link-8) and the link with the highest standard variance of speed (link-32). Their positions are marked in Fig. 2. For spatial predictability, similarly we select the time with the lowest average speed and the highest variance of speed. They are the same time stamp, 17:30 PM, during evening peak hours. The following conclusions also hold for most other positions and time-of-day in this case study.

4.4.1. Temporal predictability

Most conclusions obtained from accumulation prediction also hold for network-level speed forecasting, such as the influence of observation range and prediction horizon. In this subsection we will not re-show all of them.

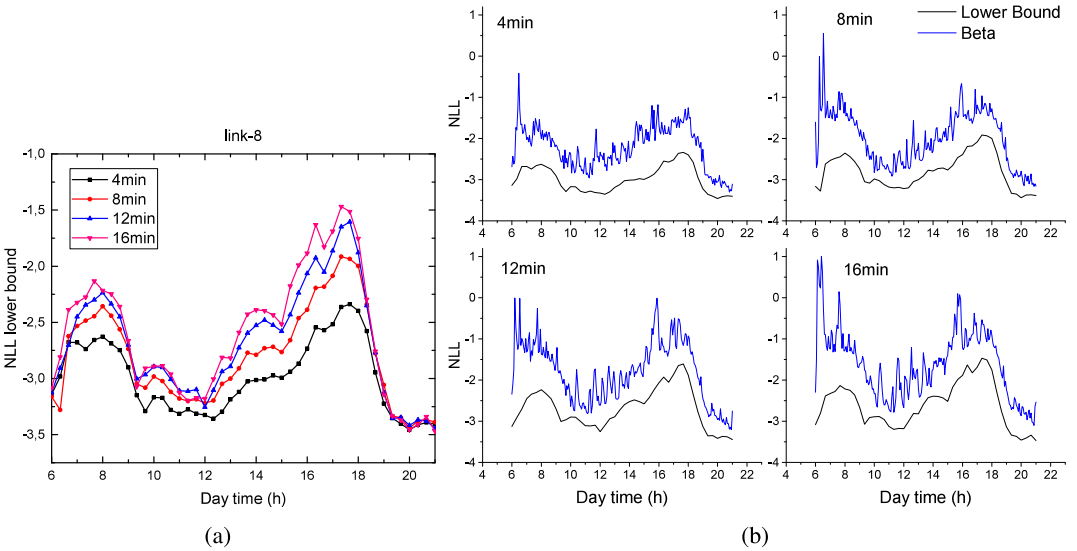

Fig. 10. (a) the lower bound of NLL for each prediction step on link-8; (b) comparison between the NLL lower bounds and the performances of Beta-prior probabilistic model on link-8. . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

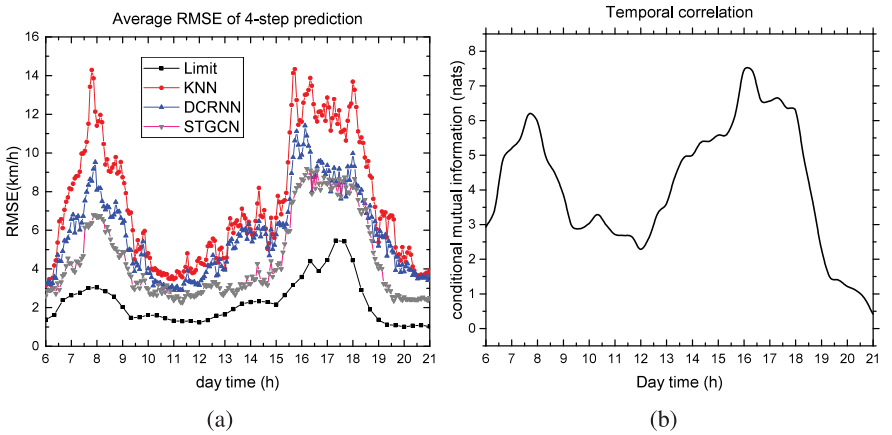

Fig. 11. Link-8: (a) Comparison between the 4-step RMSE lower bound and the RMSE of deterministic models; (b) Conditional mutual information. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 10a shows the limit of NLL for each prediction step on link-8. Speed is significantly more uncertain during morning and evening peak hours, but highly predictable at noon and in night. The lower bound increases with prediction horizon and the variation is more significant during peak hours than uncrowded time. The result means that higher uncertainty will expand quickly with the increasing of prediction horizon — long-term accurate point-estimate prediction in highly-uncertain situations is theoretically impossible, if no additional data is provided. In Fig. 10b the lower bounds are compared with the NLL of Beta-prior probabilistic model. Again we observe similar forms and uniform gaps (model uncertainty). Cyclostationarity is also a good assumption in multivariate speed forecasting. The gaps are significant, even in short-horizon forecasting. This result implies that the output distribution of this frequently congested link is complex, cannot be well approximated by a simple uni-modal prior. In Appendix another example of link-32 is presented.

Fig. 11a shows the lower bound of RMSE and the predictive errors of different deterministic models on link-8. Here the lower bound equals to the square of the arithmetic average of marginal variance (see (14)). STGCN has the best accuracy among the three baseline models. Its gap to the lower bound is relatively larger during peak hours. The gap is even considerable (about $2\,\mathrm{km\,h^{-1}}$) during free-flowing time slots (after 19:30 PM), which is very different from accumulation forecasting. Combining this result with Fig. 9 and Theorem 2, we infer that approximating speed time series by a Gaussian process is unreliable. The room of improvement for modeling is much smaller than the gap shows. Fig. 11b presents the conditional mutual information. We see that multistep predictions on link-8 are significantly correlated. The correlation is stronger during peak hours.

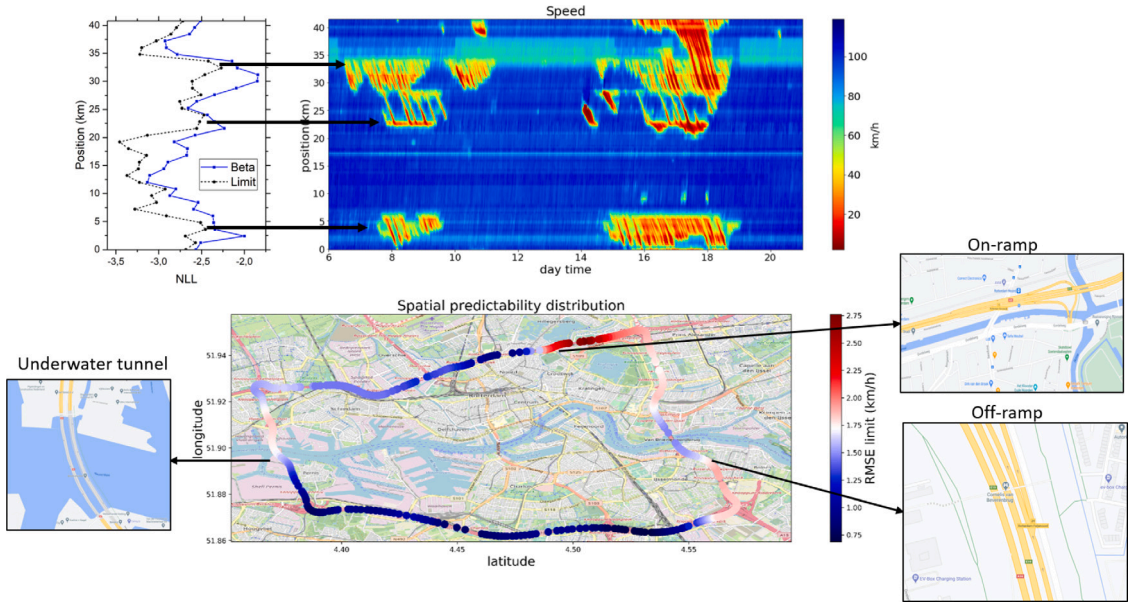

Fig. 12. Top: Comparison between the spatial predictability, NLL of Beta prior model, and the speed evolution ground-truth; Bottom: identify the most unpredictable positions on the ring freeway. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.4.2. Spatial predictability

In terms of spatial predictability, we are particularly interested in where the most uncertain locations are and why. In this subsection we do not repeatedly show the same influence of observation range or prediction horizon, but focus on analyzing the spatial distribution of predictability. The observation window is fixed as $m = 6$ (24 min) and we only consider one-step prediction. The top left figure in Fig. 12 presents the NLL limit (and thus RMSE lower bound) of different locations on the ring freeway at 17:30 PM. This limit is compared with the NLL of Beta prior model. They show similar spatial distributions. It proves that localized spatial correlation is a valid assumption in this speed forecasting task. Some positions are highly predictable (like the segment between 10 km and 18 km) meanwhile the predictability of other positions are relatively lower (like the two high peaks locate at 5 km and 30 km). The difference is significant. The spatial distribution of predictability is further compared with a representative speed evolution. The corresponding RMSE lower bound is also projected on the map to identify what are those highly unpredictable locations.

Fig. 12 shows that there are three major peaks of the predictability curve and they represent three different types of uncertain cases: (1) the highest peak on the north (at 32.4 km) corresponds to one important on-ramp connecting the ring freeway and the busy urban area around Rotterdam north station. The lack of demand data is the main reason for low predictability. How many vehicles will enter the ring freeway and when the on-ramp will be saturated is highly uncertain. (2) The peak on the west (at 4 km) is the exit of an underwater tunnel, which is also one of the major bottlenecks. The unstable driving behaviors when vehicles leave the tunnel probably cause low predictability. (3) The other one on the southeast (at 22.4 km) is an off-ramp. Stop-and-go waves tend to stop spreading here (see the top figure). Predicting how many vehicles will leave the ring freeway and how long the congestion will last is indeed highly uncertain. The analysis above identifies the most uncertain locations on this beltway. These three critical positions determine the macroscopic spreading of traffic congestion.

In many applications, studying spatial predictability is usually more important than temporal predictability, especially for data collection and highway traffic control. The distribution of predictability can help optimizing where to install sensors to maximize performance-cost ratio (Gentili and Mirchandani, 2012; Eisenman et al., 2006). There are 38 on/off-ramps that connect urban roads or other highways to this target ring freeway. But we only need to collect more data around the three critical locations mentioned above. Possible methods include installing more loop detectors, inducing more types of data (like flow), or adding speed data on the adjacent urban roads. For traffic managers, extra attention should be payed at these highly uncertain locations because they largely determine the congestion evolution of the entire beltway.

4.5. Summary of main findings

In summary, by comparing the estimated lower bound of NLL/MSE/DCM and the real performances of selected baseline models, the proposed predictability (uncertainty) metric is validated for both univariate traffic accumulation forecasting and network-level speed forecasting. Our main findings are summarized as follows:

- In accumulation prediction, Inverse-Gamma distribution is a good prior for short-term prediction. For speed forecasting, the Beta distribution offers better results. But there is still considerable distance to the NLL limit. Specifically, it turns out that approximating speed evolution as a Gaussian process is unreliable.
- In speed forecasting, the information from downstream dominates the prediction. Traffic states on upstream links have little influence on the predicted results. Since we utilize speed data only, this makes sense from a traffic flow perspective. This correlation is due to queue spill-back.
- Longer observation range and shorter prediction horizon can increase predictability. The proposed approach can quantify this relationship.
- Multistep predictions are temporally correlated. The correlation is stronger during peak hours.
- For probabilistic models, the predictive performance is mainly restricted by improper priors; for deterministic models, the maximum potential room of improvement for modeling can be quantified.

5. Conclusions and perspectives

In this paper we proposed an entropy-based method to estimate the limit of predictability for both univariate and network-level traffic forecasting. Conditional entropy gives the optimal lower bound of NLL for probabilistic model and a lower bound of MSE/DCM for deterministic models. By considering the spatio-temporal characteristics of traffic streams, both lower bounds are formulated as functions of space and day-of-time. Experiments show that cyclostationarity and localized spatial correlations are reasonable assumptions. Selected models can only approach estimated theoretical limit but cannot cross it in most cases. The influence of observation range and prediction horizon is also clarified and quantified. Longer observation windows can increase the predictability and longer prediction horizons decrease predictability. The most important contribution of this paper is that this approach gives an estimate of the boundary for a wide range of traffic forecasting models. By comparing real performances of models and the lower bound, we can infer what is the major bottleneck in modeling and estimate how much potential room remains for modeling. This approach potentially brings more than the discussion above. Here we suggest several relevant research directions.

First, the major obstacle in probabilistic forecasting is how to model the prior distribution. Currently most papers use a simple, uni-modal distribution. But these priors are not good enough in speed forecasting. To approach the estimated lower bound, exploring more complex priors, such as mixture models, is necessary and important. Second, how to formulate macroscopic traffic forecasting problem should be re-considered. Currently researchers perhaps focus too much on developing new sequence-to-sequence models (especially deep learning models) that push predictive accuracy little by little. But the remaining room of improvement by modeling may be less than expected. Our results showed that the limit of predictability of one single traffic quantity (such as speed) drops rapidly with prediction horizon during peak hours. To further improve mid-term or long-term predictive accuracy, investing more in collecting diverse, multi-scale data sources (such as trajectories, OD data, etc.) and studying how to fuse them in one model are more promising. A third highly interesting research topic is the possibility of using the spatial distribution of predictability to guide sensor installation. This sensor location problem still needs more investigation.

Finally, we emphasize that the proposed approach can still be improved. Our method uses k-fold strategy to mitigate dataset shift and avoid the failure of i.i.d assumption. However, this is not feasible in practice. There always exist new patterns and out-of-distribution samples in data streams. How to disentangle this factor and how to overcome this difficulty needs more research.

CRediT authorship contribution statement

Guopeng Li: Conceptualization, Methodology, Software, Validation, Writing – original draft. **Victor L. Knoop:** Methodology, Data visualization, Formal analysis, Supervision, Writing – review & editing. **Hans van Lint:** Conceptualization, Resources, Supervision, Writing – review & editing, Project administration, Funding acquisition.

Acknowledgments

This research is sponsored by the NWO/TTW project MiRRORS with grant agreement number 16270. We thank them for supporting this study.

Appendix A. Details about the speed forecasting probabilistic model

We built one DNN-based probabilistic speed forecasting model based-on STGCN and U-net (Ronneberger et al., 2015). U-net shows competitive performances in many computer vision tasks and it is state-of-the-art in some pixel-wise uncertainty estimation dataset, such as NYU-depth.[1] The model is composed of similar spatio-temporal convolutional module proposed in STGCN (Yu et al., 2017) and skip connections. The last layer output parameters of the assumed prior distributions. The model structure is shown Fig. A.1.

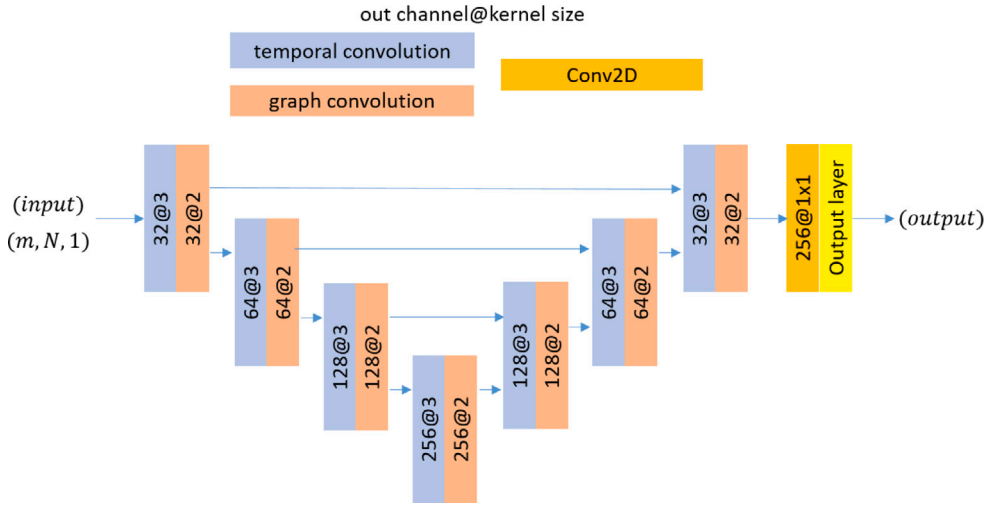

Fig. A.1. Structure of the speed forecasting probabilistic model. Here m is the observation length and N is the number of road links.

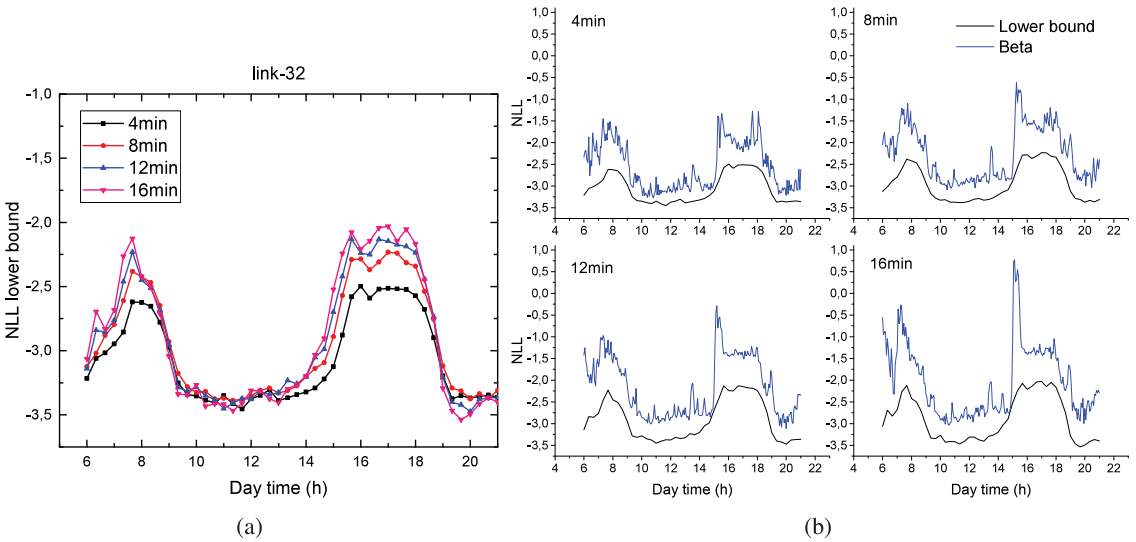

Fig. B.1. (a) the lower bound of NLL for each prediction step on link-32; (b) comparison between the NLL lower bounds and the performances of Beta-prior probabilistic model.

Appendix B. Another example: link-32

See Fig. B.1.

References

Amigó, José M., Hirata, Yoshito, Aihara, Kazuyuki, 2017. On the limits of probabilistic forecasting in nonlinear time series analysis II: differential entropy. Chaos 27 (8), 083125.

Ben-Akiva, Moshe, Bierlaire, Michel, Koutsopoulos, Haris, Mishalani, Rabi, 1998. DynaMIT: a simulation-based system for traffic prediction. In: DACCORD Short Term Forecasting Workshop. Delft The Netherlands, pp. 1–12.

Cover, Thomas M., 1999. Elements of Information Theory. John Wiley & Sons.

Cunningham, John P., Hennig, Philipp, Lacoste-Julien, Simon, 2011. Gaussian probabilities and expectation propagation. arXiv preprint arXiv:1111.6832.

Darmon, David, 2016. Specific differential entropy rate estimation for continuous-valued time series. Entropy 18 (5), 190.

[1] https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html.

Eisenman, Stacy M., Fei, Xiang, Zhou, Xuesong, Mahmassani, Hani S, 2006. Number and location of sensors for real-time network traffic estimation and prediction: Sensitivity analysis. Transp. Res. Rec. 1964 (1), 253–259.
Ermagun, Alireza, Levinson, David, 2018. Spatiotemporal traffic forecasting: review and proposed directions. Transp. Rev. 38 (6), 786–814.
Faes, Luca, Nollo, Giandomenico, Porta, Alberto, 2011. Information-based detection of nonlinear Granger causality in multivariate processes via a nonuniform embedding technique. Phys. Rev. E 83 (5), 051112.
Fang, Song, Skoglund, Mikael, Johansson, Karl Henrik, Ishii, Hideaki, Zhu, Quanyan, 2019. Generic variance bounds on estimation and prediction errors in time series analysis: An entropy perspective. In: 2019 IEEE Information Theory Workshop (ITW). IEEE, pp. 1–5.
Fusco, Gaetano, Colombaroni, Chiara, Isaenko, Natalia, 2016. Short-term speed predictions exploiting big data on large urban road networks. Transp. Res. C 73, 183–201.
Gentili, Monica, Mirchandani, Pitu B., 2012. Locating sensors on traffic networks: Models, challenges and research opportunities. Transp. Res. C 24, 227–255.
Gers, Felix A., Schmidhuber, Jürgen, Cummins, Fred, 1999. Learning to forget: Continual prediction with LSTM.
van Hinsbergen, Chris P.I.J., Schreiter, Thomas, Zuurbier, Frank S., van Lint, J.W.C., van Zuylen, Henk J., 2012. Localized extended Kalman filter for scalable real-time traffic state estimation. IEEE Trans. Intell. Transp. Syst. 13 (1), 385–394.
van Hinsbergen, C.P.I.J., Van Lint, J.W.C., Van Zuylen, H.J., 2009. Bayesian committee of neural networks to predict travel times with confidence intervals. Transp. Res. C 17 (5), 498–509.
Idé, Tsuyoshi, Kato, Sei, 2009. Travel-time prediction using Gaussian process regression: A trajectory-based approach. In: Proceedings of the 2009 SIAM International Conference on Data Mining. SIAM, pp. 1185–1196.
Kawata, Satoshi, Minami, Shigeo, 1984. Adaptive smoothing of spectroscopic data by a linear mean-square estimation. Appl. Spectrosc. 38 (1), 49–58.
Knoop, Victor L., Van Lint, Hans, Hoogendoorn, Serge P., 2015. Traffic dynamics: Its impact on the macroscopic fundamental diagram. Physica A 438, 236–250.
Kontoyiannis, Ioannis, Algoet, Paul H., Suhov, Yu M., Wyner, Abraham J., 1998. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. IEEE Trans. Inform. Theory 44 (3), 1319–1327.
Krishnakumari, Panchamy, Nguyen, Tin, Heydenrijk-Ottens, Léonie, Vu, Hai L., van Lint, Hans, 2017. Traffic congestion pattern classification using Multiclass active shape models. Transp. Res. Rec. 2645 (1), 94–103.
Lamb, Alex M., Goyal, Anirudh Goyal Alias Parth, Zhang, Ying, Zhang, Saizheng, Courville, Aaron C., Bengio, Yoshua, 2016. Professor forcing: A new algorithm for training recurrent networks. In: Advances in Neural Information Processing Systems. pp. 4601–4609.
Lan, Lawrence W., Sheu, Jiuh-Biing, Huang, Yi-San, 2008. Investigation of temporal freeway traffic patterns in reconstructed state spaces. Transp. Res. C 16 (1), 116–136.
Lana, Ibai, Del Ser, Javier, Velez, Manuel, Vlahogianni, Eleni I, 2018. Road traffic forecasting: Recent advances and new challenges. IEEE Intell. Transp. Syst. Mag. 10 (2), 93–109.
Lee, Kimin, Lee, Honglak, Lee, Kibok, Shin, Jinwoo, 2017. Training confidence-calibrated classifiers for detecting out-of-distribution samples. arXiv preprint arXiv:1711.09325.
Li, Huiping, He, Fang, Lin, Xi, Wang, Yinhai, Li, Meng, 2019. Travel time reliability measure based on predictability using the Lempel–Ziv algorithm. Transp. Res. C 101, 161–180.
Li, Yongfu, Jiang, Xiao, Zhu, Hao, He, Xiaozheng, Peeta, Srinivas, Zheng, Taixiong, Li, Yinguo, 2016. Multiple measures-based chaotic time series for traffic flow prediction based on Bayesian theory. Nonlinear Dynam. 85 (1), 179–194.
Li, Yaguang, Yu, Rose, Shahabi, Cyrus, Liu, Yan, 2018. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In: International Conference on Learning Representations.
Liebig, Thomas, Piatkowski, Nico, Bockermann, Christian, Morik, Katharina, 2017. Dynamic route planning with real-time traffic predictions. Inf. Syst. 64, 258–265.
Lombardi, Damiano, Pant, Sanjay, 2016. Nonparametric k-nearest-neighbor entropy estimator. Phys. Rev. E 93 (1), 013310.
Lopez, Clélia, Leclercq, Ludovic, Krishnakumari, Panchamy, Chiabaut, Nicolas, Van Lint, Hans, 2017. Revealing the day-to-day regularity of urban congestion patterns with 3D speed maps. Sci. Rep. 7 (1), 1–11.
Ma, Xiaolei, Dai, Zhuang, He, Zhengbing, Ma, Jihui, Wang, Yong, Wang, Yunpeng, 2017. Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction. Sensors 17 (4), 818.
Nair, Attoor Sanju, Liu, Jyh-Charn, Rilett, Laurence, Gupta, Saurabh, 2001. Non-linear analysis of traffic flow. In: ITSC 2001. 2001 IEEE Intelligent Transportation Systems. Proceedings (Cat. No. 01TH8585). IEEE, pp. 681–685.
Nguyen, Tin T., Krishnakumari, Panchamy, Calvert, Simeon C., Vu, Hai L., Van Lint, Hans, 2019. Feature extraction and clustering analysis of highway congestion. Transp. Res. C 100, 238–258.
Packard, Norman H., Crutchfield, James P., Farmer, J. Doyne, Shaw, Robert S., 1980. Geometry from a time series. Phys. Rev. Lett. 45 (9), 712.
Ronneberger, Olaf, Fischer, Philipp, Brox, Thomas, 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.
Rosenstein, Michael T., Collins, James J., De Luca, Carlo J., 1993. A practical method for calculating largest Lyapunov exponents from small data sets. Physica D 65 (1–2), 117–134.
Schakel, Wouter J., Knoop, Victor L., van Arem, Bart, 2012. Integrated lane change model with relaxation and synchronization. Transp. Res. Rec. 2316 (1), 47–57.
Schreiter, Thomas, van Lint, Hans, Treiber, Martin, Hoogendoorn, Serge, 2010a. Two fast implementations of the adaptive smoothing method used in highway traffic state estimation. In: 13th International IEEE Conference on Intelligent Transportation Systems. IEEE, pp. 1202–1208.
Schreiter, Thomas, Van Lint, Hans, Yuan, Yufei, Hoogendoorn, Serge, 2010b. Propagation Wave Speed Estimation of Freeway Traffic with Image Processing Tools. Technical report.
Shang, Pengjian, Li, Xuewei, Kamae, Santi, 2005. Chaotic analysis of traffic time series. Chaos Solitons Fractals 25 (1), 121–128.
Shannon, Claude E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27 (3), 379–423.
Song, Chaoming, Qu, Zehui, Blumm, Nicholas, Barabási, Albert-László, 2010. Limits of predictability in human mobility. Science 327 (5968), 1018–1021.
Treiber, Martin, Helbing, Dirk, 2003. An adaptive smoothing method for traffic state identification from incomplete information. In: Interface and Transport Dynamics. Springer, pp. 343–360.
Van Lint, J.W.C., 2008. Online learning solutions for freeway travel time prediction. IEEE Trans. Intell. Transp. Syst. 9 (1), 38–47.
Van Lint, Hans, Miete, Onno, Taale, Henk, Hoogendoorn, Serge, 2012. Systematic framework for assessing traffic measures and policies on reliability of traffic operations and travel time. Transp. Res. Rec. 2302 (1), 92–101.
Wang, Qing, Kulkarni, Sanjeev R., Verdú, Sergio, 2009. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. IEEE Trans. Inform. Theory 55 (5), 2392–2405.
Wang, Jingyuan, Mao, Yu, Li, Jing, Xiong, Zhang, Wang, Wen-Xu, 2015. Predictability of road traffic and congestion in urban areas. PLoS One 10 (4), e0121825.
Wang, Yibing, Papageorgiou, Markos, Messmer, Albert, 2006. RENAISSANCE - A unified macroscopic model-based approach to real-time freeway network traffic surveillance. Transp. Res. C 14 (3), 190–212.
Wolf, Alan, Swift, Jack B., Swinney, Harry L., Vastano, John A., 1985. Determining Lyapunov exponents from a time series. Physica D 16 (3), 285–317.
Xiong, Wanting, Faes, Luca, Ivanov, Plamen Ch., 2017. Entropy measures, entropy estimators, and their performance in quantifying complex dynamics: Effects of artifacts, nonstationarity, and long-range correlations. Phys. Rev. E 95 (6), 062114.

Yu, Bing, Yin, Haoteng, Zhu, Zhanxing, 2017. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. arXiv preprint arXiv:1709.04875.

Yuan, Yun, Zhang, Zhao, Yang, Xianfeng Terry, Zhe, Shandian, 2021. Macroscopic traffic flow modeling with physics regularized Gaussian process: A new insight into machine learning applications in transportation. Transp. Res. B 146, 88–110.

Yuan, Jing, Zheng, Yu, Xie, Xing, Sun, Guangzhong, 2011. Driving with knowledge from the physical world. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 316–324.