Uncertainty Quantification and Predictability Analysis for Traffic Forecasting at Multiple Scales

Li, G.

DOI
10.4233/uuid:aeb5eccb-3cd3-43cb-a796-5390d31f4f5e

Publication date
2023

Document Version
Final published version

Uncertainty Quantification and Predictability Analysis for Traffic Forecasting at Multiple Scales

Guopeng LI

This doctoral dissertation was funded by the Dutch National Data Warehouse of Traffic Information (NDW). It was also part of the MiRRORS project (with project number 16720) within the Open Technology Program, which is (partly) financed by the Applied Sciences Division of the Dutch Research Council (NWO/TTW).

Uncertainty Quantification and Predictability Analysis for Traffic Forecasting at Multiple Scales

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, Prof.dr.ir. T.H.J.J. van der Hagen,
Chair of the Board for Doctorates
to be defended publicly on
Tuesday 25 April 2023 at 10:00 o'clock
by

Guopeng LI

Diplôme d'Ingénieur,
École Nationale Supérieure de Techniques Avancées Paris, France
born in Liao Ning, China

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	Chairperson
Prof.dr.ir. J.W.C van Lint,	Delft University of Technology, promotor
Dr. V. L. Knoop,	Delft University of Technology, promotor

Independent members:

Prof.dr.ir. A. Verbraeck,	Delft University of Technology
Prof.dr. M.M. de Weerdt,	Delft University of Technology
Prof.dr.ir. M. Menendez,	New York University Abu Dhabi, United Arab Emirates
Dr. J. Kim,	The University of Queensland, Australia
Dr. A.M. Alahi,	École Polytechnique Fédérale de Lausanne, Switzerland

TRAIL Thesis Series no. T2023/5, the Netherlands Research School TRAIL

TRAIL
P.O. Box 5017
2600 GA Delft
The Netherlands
E-mail: info@rsTRAIL.nl

ISBN: 978-90-5584-324-4

Printed in the Netherlands

The meaning of the world is the separation of wish and fact.
–Kurt Gödel

Acknowledgements

Five years ago, as a fresh master's graduate, I travelled alone from Paris to Delft with a mixture of nervousness and longing, embarking on my own PhD journey. At the time, I had no idea what this journey would entail, but I knew that it would be a challenging and rewarding experience that would shape my future. Three years ago, as I gazed at the red sunset reflected on my computer monitor in a daze, I experienced doubts about whether pursuing a PhD was the right decision for me. Even as recently as one year ago, I still regretted the time lost in this pandemic and wondered whether I could have done better. I felt as though I had lost valuable time and opportunities, and I worried that I would not be able to achieve my goals. But now, as I finalize my dissertation under the warm glow of my desk lamp and recollect all the beautiful moments at Delft University of Technology, I can say with pride and certainty that this journey is magnificent. I have learned so much about myself and the world around me, and I have grown in ways that I never thought possible. This journey has become an indelible part of my life, shining as brightly as a diamond.

I am grateful for the support and guidance of my mentors, colleagues, and loved ones who have been there for me every step of the way. Without them, I would not have been able to achieve this milestone in my life.

First, I would like to express my sincere gratitude to my promotors, Hans and Victor, for your unwavering support, invaluable guidance, and kind encouragement throughout my academic journey and personal life. Your expertise, insights, and feedback have been instrumental in shaping my research and enhancing my skills. I am deeply grateful for the opportunity to learn from you. Hans, I am extremely lucky to have had you as my supervisor. Your academic taste and intuition always amaze me. Whenever I faced a challenge or a dilemma in my research, you always lead me in the right direction. I still remember how you advised me to focus on the "predictability of traffic" rather than the models themselves, and it took me two years to fully comprehend the wisdom of your suggestion. You are not only my supervisor but also a dear friend to me. Victor, you always support me without any reservations. you not only cared about my research progress but also my well-being and mental health. When I started my PhD, I knew nothing about traffic flow theory. But thanks to your patient guidance and insightful comments, I have come a long way and learned so much. I cherish our conversations, which have often been a source of comfort and inspiration for me. Whenever I feel depressed or nervous, you always encourage me. I am deeply moved

by our conversations during these times. I am looking forward to our future work together during my postdoc.

I would also like to thank the members of my dissertation committee, Prof. Alexander Verbraeck, Prof. Mathijs de Weerdt, Prof. Monica Menendez, Prof. Jiwon Kim, and Prof. Alexandre Massoud Alahi, for their valuable feedback, constructive criticism, and insightful comments. Their diverse perspectives and expertise have greatly enriched my research and contributed to the rigour and quality of my dissertation.

I am grateful to the faculty, staff, and friends in the Department of Transport and Planning. You have made my PhD journey a truly wonderful experience. The support, encouragement, and kindness are invaluable to me. I am so lucky to have the fantastic environment and the lovely people that I worked with. Panchamy and Ding, you helped me so much when I joined the DiTTlab four years ago. I cannot integrate into this team so quickly and easily without your guidance. Tin, you kindly provided so much support on coding and data retrieval, and we can always get some sparks in our minds when we discuss a topic. Prof. Alexander Verbraeck, I am so grateful that you generously provide hardware resources that greatly accelerated my research progress. Zahra and Muriel, I am so lucky to be on the same research project with you. We together did a great job! Yiru, talking with you makes me feel like basking in the spring breeze. I believe that you will have a successful PhD journey. For the many memorable excursions and enjoyable activities we shared, I would like to thank my colleagues from the DiTTlab and the TDMac-lab, Simeon, Sanmay, Saeed, Kexin, Zili, Ali, Xue, Wouter, Peter, Andreas, Samkie, Jinyang, and Lucas (Suriyana). We are an amazing team. I am grateful for the friendships we have formed.

I am also grateful to have so many friendly and excellent colleagues in our department. Paul, Nagarjun, Saman, Ziyulong, Yiyun, Joelle, Alex, Lucas, Yongqi, Bing, Yanyan, Nirvana, and many many people. The lunch talks, coffee breaks, and the exchange of diverse cultures make my life so colourful. Prof. Yasuhiro Shiomi and Zirui, I learnt a lot from our cooperation during your visits to TU Delft. Hope that we can work together again in the future! I would like also to thank the secretaries, technicians, and cleaners. You make our campus a lovely place.

I have encountered so many excellent teachers in my life. They not only impart knowledge but also shape my thinking and character. In particular, I would like to express my sincere gratitude to my mathematics teacher in high school, Mr Shaodong Wang. I hope that you rest in peace in another world.

Finally, I would like to express my deepest gratitude to my beloved family, especially my mom, Yanhong Hu. Mom, you are the greatest woman that I have ever seen in my life. Your love, support, and encouragement have sustained me throughout my life and my doctoral journey. I hope that I do make you proud.

Guopeng LI,

Delft, March 2023

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Traffic systems are critically important for modern society. They carry passengers and goods from their origins to their destinations. With the increasing level of urbanization, traffic and transportation networks are becoming larger and more complex. Higher complexity also brings higher vulnerability. The efficiency and safety of traffic systems are always facing the risk of being damaged by traffic accidents or congestion. These disturbances may cause severe consequences, such as casualties and injuries, economic and time costs, air pollution, etc. To address this challenge, developing an effective and reliable traffic management system is necessary. To this end, people collect sensor data to monitor traffic system states (perception), use observed data to describe traffic dynamics and predict the future (prediction), and then decide when and how to control traffic networks (intervention). As the intermediate module that connects perception and intervention, a reliable prediction model is one essential integral part of Intelligent Transportation Systems (ITS).

Predicting the macroscopic traffic state and the microscopic driving behaviours are fundamental for many applications. From passing an unsignalized intersection assisted by self-driving systems, reducing the waiting time around an on-ramp by real-time metering policy, to mitigating traffic congestion in large-scale road networks by proactive management, predicting how the system state will change provides critical evidence for supporting what decisions road agents and authorities should make. Driven by this application value, great effort has been devoted to developing forecasting models that can give anticipated driving behaviours or traffic states. Existing methods in the literature are numerous, and the prediction performance has been significantly improved in the past several decades.

However, besides the prediction model itself, how predictable the short-future of a traffic system is on different levels, or the so-called *predictability*, has not been deeply discussed in the literature. Traffic phenomena involve many interactive agents that are not 100% predictable by any means. The first reason is that traffic dynamics are stochastic by nature. Road agents have diverse behaviours and the traffic system is only partly-observable on all levels. Second, although most driving behaviours and traffic state patterns are highly recurring, encountering new situations is always possible in practice. Therefore, *uncertainty* governs the predictability of traffic and it should be an integral part of the traffic modelling task. Since one of the major goals of prediction is

to support traffic control, ignoring uncertainty will lead to over-confident predictions and thus decrease the robustness of decision-making. From this perspective, we argue that *uncertainty quantification* (UQ) is as important as the prediction model itself. Therefore, a systematic approach is needed to measure the boundary of any model's predictive capability and the additional risk (of getting this prediction wrong) brought by using this model in the real world. In this thesis, we explore both uncertainty quantification and predictability analysis methods.

This thesis introduction is organized as follows. First, the background on uncertainty quantification in traffic prediction at different scales is introduced, including traffic dynamics modelling (1.1.1), sources of uncertainty (1.1.2), and the three critical factors for reasonable uncertainty quantification (1.1.3). Based on the discussion, the scientific gaps are identified in Section 1.2 and the research objective and questions are come up with in Section 1.3. Next, Section 1.4 summarizes the major contributions of this thesis. Finally, Section 1.5 presents the outline of this thesis.

1.1 Research background

This section presents the background of this thesis. It will start by introducing traffic modelling methods based on their spatio-temporal scales and application purposes. Then we will scope the uncertainty in traffic systems and the corresponding requirements for quantifying it.

1.1.1 Traffic modelling and forecasting

Since the 1950s, modelling traffic flow dynamics, predicting the traffic state, and describing road agents' behaviours have been the central topics in the transport and planning domain. In practice, building a model must prioritize two key factors, the *application purpose* and the *spatial-temporal scale*. They must be considered together for determining the needed data types, the model complexity, and the modelling strategy. Therefore, the background of traffic modelling is organized along these two axes of purpose and scale.

According to Judea Pearl's arguments (Pearl, 2009), a model can be evaluated by a hierarchy of three types of problems of increasing difficulty, the so-called "ladder of causation". Each of them is closely related to one category of application purposes:

- **Level-1: association and prediction**
  The first stage is "observing and building associations". Given the desired output variable and a set of potentially relevant input variables, a model aims to directly assimilate their *joint probability distribution* from the collected observations (datasets). Such a correlation-based approach ignores the detailed data generation process. The relevant application is building a predictive model and deploying this predictor in the same environment.

- **Level-2: intervention and what-if analysis**
  The second stage is "changing by doing". Besides the naive associations, the model is encouraged to explore how manipulating one variable influences the others. This requires building the directional relationships among the variables, or the so-called *causal graph*. Such a causation-based model is necessary for what-if analysis – giving predictions in new, unseen environments and studying what will happen if the traffic system is intervened in a specific way.

- **Level-3: counterfactual and policy-making**
  The last stage is "imaging and reasoning". Besides the forward generation of output, the model can also run inversely. If we want the desired (generally optimal) output, how should we manipulate the inputs? This level is closely related to optimization and policy-making purposes, such as designing long-term land use, mitigating traffic congestion through traffic control, collision-free motion planning for autonomous vehicles, etc.

From the lowest to the highest level of this ladder, correlation-based models are replaced by causation-based models. Causation-based models are more explainable and insights into the domain knowledge can be extracted from data. In contrast, a correlation-based model cannot always give correct relationships. Their differences are compared in Table.1.1.

Table 1.1: A taxonomy of two types of models

Model	Prediction	What-if analysis	Policy-making	knowledge insights
Correlation-based	✓	✗	✗	?
Causation-based	✓	✓	✓	✓

However, building a structural causal model from data is not always possible in practice because it requires high observability of the system, especially of those confounder variables (common causes) (Schölkopf et al., 2021). For example, due to privacy issues, travellers' demand patterns cannot be completely tracked but demand is the major cause of traffic congestion (Jayakrishnan et al., 1995). To fulfil the application purpose with the existing observability gap in data, knowledge-based *assumptions* must be induced in the model.

Besides the application purpose, the level of scale of a task is another key to determining the modelling strategy. "Multi-scale" is an important characteristic of traffic systems. It means that we need to "solve problems which have important features at multiple scales of time and/or space" (Multiscale modelling, 2022). Traffic systems involve intelligent agents (human factors) who have complex decision mechanisms and behaviours at all levels. These decisions and behaviours are not completely free but are constrained by traffic rules, infrastructures, and other agents. Therefore, traffic modelling can accommodate any desired complexity at any scale. A macroscopic traffic flow model is not necessarily more sophisticated than a microscopic model, e.g. lane-changing. Choosing proper *representations* and *simplifications* at different levels, or the so-called "scale-complexity trade-off", is pivotal for traffic modelling.

Based on the discussion above and the topic of this thesis, the following overview of traffic modelling will primarily focus on what assumptions, representations, and simplifications are used, and what problems one may encounter for different application purposes.

Modelling and predicting microscopic driving behaviours (e.g. represented by trajectories) generally consider the temporal scale of several seconds (the duration of an interaction) and the spatial scale of a corridor segment or an intersection/roundabout (within the driver's eye vision). Although the scale is restricted, the confounder here–how drivers make decisions and interact with each other–is unobservable and vague. For the purposes of what-if analysis in traffic simulation and optimal motion control in autonomous driving, researchers need to make assumptions based on experiences and behavioural studies of *Human Factors* (HF). For example, the basic assumption in 1D Gipps' car-following model (Gipps, 1981) and its improved variants (we refer to the review of Ciuffo et al. (2012)) is that all drivers can keep a proper distance headway to avoid collision based on the motion of both leading and following vehicles. Besides safety, the family of Intelligent Driver Models (IDM) (Treiber et al., 2000; Kesting et al., 2010) further assume a comfortable deceleration range. IDM and its variants are still widely used in traffic simulators, such as SUMO [1]. Their safety performances are even higher than more complex reinforcement learning methods in many simulation tests (Suo et al., 2021). In more complex 2D interaction scenarios, the assumptions are mainly about the trade-off between safety, efficiency, and comfort. For example, using a game theoretical approach to represent lane-changing negotiation (the review of Ji & Levinson (2020)) or the driving behaviours inside an intersection (Zhao et al., 2022). There are many excellent reviews and studies that discuss HF in driving-behaviour modelling, such as the review of Saifuzzaman & Zheng (2014) (specifically on HF in car-following) and the generic multi-level framework proposed by Van Lint & Calvert (2018) and Calvert et al. (2020). These models are abstracted from scientifically established domain knowledge (of both traffic and cognitive or behavioural sciences) so they have excellent interpretability and generalizability. The accuracy can be improved by adding more realistic assumptions and more details.

On the other side, data-driven approaches, and especially most deep learning models, do not emphasize how to make assumptions and simplifications from domain knowledge. They focus on increasing the accommodated model complexity and learning feature representations. Compared to the HF models mentioned above, one can say that these data-driven models are accuracy-oriented. For example, some studies represent driving scenarios by rasterized multi-channel images and use computer-vision models to predict the trajectory (Nikhil & Tran Morris, 2018; Xu et al., 2018; Xie et al., 2020). Recently some researchers represent road agents and high-definite map information into nodes and describe their interactions by a dynamical graph. This lighter representation allows using Graph Neural Networks (GNN) to learn compact features. For example, VectorNet (Gao et al., 2020) uses a sequence of vectors to consistently represent map elements and trajectories, then employs a graph attention mechanism to model their interaction implicitly. Several recent studies use similar methods, e.g. Zhao et al. (2019); Huang et al. (2019a); Salzmann et al. (2020), etc. However, these models are all correlation-based (level 1). They are not adaptive to new deployment

[1] https://www.eclipse.org/sumo/

environments, and users do not understand how exactly the predictions or decisions are made (Bahari et al., 2022).

When scaling up to macroscopic traffic modelling and forecasting, since traffic flow is composed of individual road users, how to represent and simplify these road users in the traffic stream is very important for the complexity-scale trade-off. For instance, improper driving behaviours may cause traffic congestion (Hennessy & Wiesenthal, 1999). But implementing a detailed behavioural model (like those sophisticated models mentioned above) into the macroscopic traffic flow model is not always necessary in practice. The model complexity depends on the purpose.

One simplification strategy is representing the vehicle stream by a many-particle system or a continuous fluid. Instead of delicately describing each agent's behaviours, this approach depicts their collective properties (behaviours) in traffic systems, like density, average speed, and flow. The Lighthill-Whitham-Richards (LWR) model (Lighthill & Whitham, 1955b; Richards, 1956) is one of the most famous examples. The key assumption is the conservation of vehicles that can be described by a hyperbolic wave equation. The speed of the wave can be derived from a calibrated fundamental diagram. Similar examples include higher-order traffic flow models (Payne, 1971; Whitham, 2011) and gas-kinetic models (Prigogine & Herman, 1971), etc. At higher scales, like urban networks, the concept of Network Macroscopic Fundamental Diagram (NMFD) (Daganzo & Geroliminis, 2008) can further simplify the representation. We refer the readers to Helbing et al. (2009) and Johari et al. (2021) for comprehensive reviews of related methods.

Although information about the microscopic causes of traffic phenomena is partly lost due to these simplifications, the macroscopic causes still preserve in these models. For example, the over-saturated traffic demand and the back-propagating stop-and-go waves. Conversely, most machine learning approaches do not explicitly consider these macroscopic causes but (similar to the microscopic case) emphasize expanding the model complexity and assimilating the input-output association in an end-to-end way. These data-driven frameworks are more generic. They are applicable to many other dynamical systems.

Early data-driven methods include naive conditional averaging (Davis & Nihan, 1991; Smith & Demetsky, 1997), auto-regression and time series models such as linear regression (Rice & Van Zwet, 2004), Principle Component Analysis (PCA) (Xing et al., 2015), Support Vector Regression (SVR) (Castro-Neto et al., 2009), (Seasonal) Auto-Regressive Integrated Moving Average (ARIMA) (Ahmed & Cook, 1979; Williams & Hoel, 2003), Kalman filter van Hinsbergen et al. (2012), naive Bayesian methods (van Hinsbergen et al., 2009), etc. Recently, with the fast development of deep learning models, Deep Neural Networks (DNN) are getting popular in the traffic prediction domain. For example, simple Multiple-Layer Perceptron (MLP) (Sharma et al., 2018; Polson & Sokolov, 2017), pure Convolutional Neural Networks (CNN) (Zhang et al., 2019a), Recurrent Neural Networks (RNN) or combined with CNN (Ma et al., 2015; Zhao et al., 2017), State-Space Neural Networks (SSNN) (Van Lint et al., 2005), Graph Neural Networks (Kamarianakis & Prastacos, 2005; Li et al., 2018), etc. Lana et al. (2018) and Ermagun & Levinson (2018) systematically review the spatio-temporal traffic forecasting topic.

In summary, traffic modelling and forecasting must first consider the application purpose. Based on this goal, we need to trade off the model complexity for the desired scale and to determine the assumptions and representations. From this perspective, the so-called knowledge-based models and data-driven models are not two separate categories. One of the objectives of this thesis is to combine the best of two worlds– improving the interpretability and the generalizability of deep learning models by making assumptions and using causal representations that are compliant with domain knowledge.

Finally, the discussion above can also be applied to uncertainty estimation. Quantifying predictive uncertainty must consider what is the application purpose, what is the reasonable representation, and the scale-complexity trade-off. Next, the core concept of uncertainty in this thesis will be elaborated.

1.1.2 Uncertainty in traffic prediction

A perfect model does not exist. "All models are wrong but some are useful" (Box, 1976). Quantifying uncertainty is about measuring what the model does not know and what the data cannot give. Generally speaking, all uncertainty is caused by the lack of some information (or the so-called "sim2real" gap, which means the simulation-to-reality gap). The information loss can be the result of limited sensing types, measurement errors, or the assumptions and representations made in the model (whether they are implicit or explicit, data-driven or knowledge-based methods). However, from a practical perspective of traffic modelling versus data collection, we can categorize the uncertainty into two types of "unknown". One is what we in principle cannot know without *extra* domain knowledge of the underlying process or new perception technologies, the so-called *aleatoric uncertainty*. The other one is what we can know in principle but cannot know in practice due to the finite dataset size, the so-called *epistemic uncertainty*. Now we introduce these two concepts and their special role in data collection and traffic modelling.

Aleatoric uncertainty originates from the limited observability and/or the unpredictability of crucial information that affects the dynamics of the traffic process of interest, e.g. available sensor types, measurement errors, and the inherent randomness of traffic dynamics. Aleatoric uncertainty is scale, application, and case (i.e. dataset)-dependent. For example, in the case of predicting congestion dynamics along a corridor using data from loop detectors only, aleatoric uncertainty comes from unobserved microscopic interactions (leading to unexpected disturbances) and from (at best partially observable) arrival and route choice patterns. Even if we could measure these demand patterns fully, this may not imply we can predict them more than say a few minutes ahead, since we cannot observe the full information that governs the dynamics of the demand patterns themselves, such as individual choice behaviour, demographics, a large event or a pandemic, etc. Conversely, in the case of using microscopic trajectory data from GPS trackers or drones to predict the individual motion of vehicles, aleatoric uncertainty may stem from highly relevant but not observed intersection control settings and the turning signals (blinkers) of surrounding vehicles. Another source of aleatoric uncertainty may be the strong non-linearity of interactions between drivers, e.g. their unknown proficiency, driving styles, and intended directions.

Aleatoric uncertainty thus draws a "red line" on the best we can possibly achieve by developing prediction models from the given types of data. Aleatoric uncertainty is irreducible.

Epistemic uncertainty, on the other side, measures the uncertainty due to the lack of information that is potentially knowable for the given case, scale and application. It is the additional distrust brought by model abstraction and limits in data coverage. In other words, epistemic uncertainty quantifies the probability of errors due to deploying a model in situations for which it was not calibrated. For example, many traffic phenomena are recurring due to seasonality in the underlying demand patterns. e.g. day-to-day morning and evening peak hours, week-to-week workdays and weekends, and season-to-season holidays. There are, however, many factors that add noise around such recurrent patterns. Think of fluctuations due to events, weather, or long-term changes in demography or infrastructure construction. New situations where a well-trained model may fail happen occasionally. Epistemic uncertainty aims to quantify how reliable predictions with mathematical models are, and—as we will see in this thesis—it is also crucial in estimating the aleatoric uncertainty.

Epistemic uncertainty is thus due to model abstraction and the "rareness" of prevailing cases in the data. Epistemic uncertainty helps distinguish between regular and irregular patterns, which are both predictable in principle, and it supports the computation of the aleatoric uncertainty that represents those cases which are not predictable.

Aleatoric and epistemic uncertainty together quantify the concept of *predictability* in traffic prediction. When epistemic uncertainty is low, the predictability can be measured by aleatoric uncertainty alone. In the extreme case with zero aleatoric uncertainty, the underlying process is deterministic with full predictability. With very large aleatoric uncertainty, predictability is limited regardless of what model or how large datasets are used. When epistemic uncertainty is high, both the prediction and the aleatoric uncertainty estimation are unreliable. It means that the dataset is too small /sparse to provide any insights into the traffic process.

This explanation of "predictability" may sound abstract. To further clarify the concepts of aleatoric and epistemic uncertainty, we now put them in the modelling-data collection pipeline in practice.

As shown in Fig.1.1, people generally start from an initially given dataset for a specific modelling or forecasting task. Before building the model, it is necessary to evaluate whether the given dataset can satisfy the accuracy requirement of the application purpose. So the average aleatoric uncertainty (predictability) needs to be quantified by a model-free approach. If the answer is yes, then we can start building and training the model. Next, when deploying the calibrated model in real-world environments, rare or new cases may occur every now and then. Therefore, the model must not only produce a prediction but also an estimate of the input-dependent aleatoric and epistemic uncertainty. As explained before, their roles are different. Epistemic uncertainty indicates that the current traffic state has already been seen in the training dataset or that this is a rare case. Those high-value rare cases can be identified from the data stream and the dataset can be expanded continuously at a relatively smaller cost. On the other hand, aleatoric uncertainty helps to identify those inherently unpredictable samples for the given data types. If a considerable percentage of the dataset is composed of highly uncertain samples or in case the predictive accuracy needs to be radically improved, the

aleatoric uncertainty can provide clues about what new types of data should be added. Then we come back to the initial step.

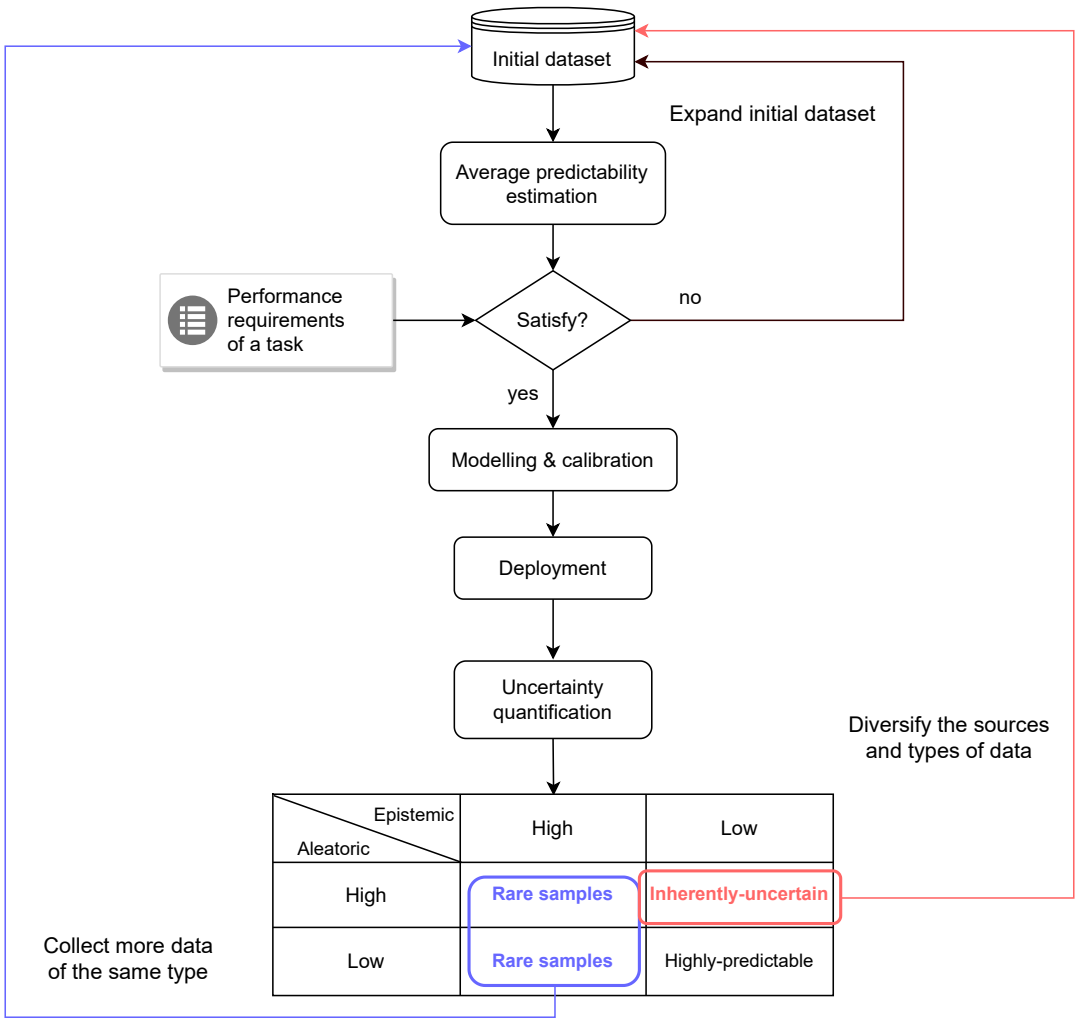


Figure 1.1: The modelling-data collection cycle and the roles of uncertainty quantification

The discussion above and the conceptual chart in Fig.1.1 point out the irreplaceable role of uncertainty quantification. Next, we will discuss what are the principles for representing and quantifying uncertainty in traffic prediction at multiple levels.

1.1.3 Three pillars of uncertainty quantification

Uncertainty Quantification (UQ) has been widely studied in many domains. Traditional fields include nuclear safety (Helton, 1993), hydrology (Beck, 1987), climate-changing (Deser et al., 2012), etc. Recent applications are mainly in computer vision, such as depth estimation (Lakshminarayanan et al., 2016) and other regression tasks. However, similar to traffic modelling, applying these existing UQ methods in traffic prediction also must consider the application purpose, the scale, and the complexity. Based on these 3 criteria, we propose that there are three key factors or the so-called

"*three pillars*" for supporting reasonable uncertainty quantification. They are *uncertainty representation*, *quantification method*, and *modelling strategy* (see Fig.1.2). Below these pillars, the available datasets, modelling techniques and related domain knowledge form the foundation.

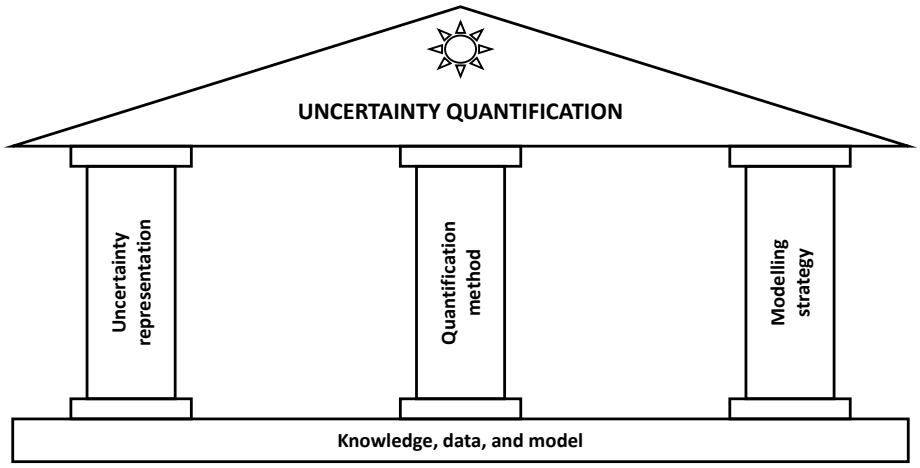


Figure 1.2: "Three pillars" of uncertainty quantification

Which UQ metrics should be used and how to represent uncertainty are the prerequisites. It depends on the specific application purpose and the model complexity of the prediction task. In practice, there are two types of metrics, *variance-based* and *entropy-based* metrics. If the desired output of the model is a scalar value (e.g. the average speed at a location) and we aim to estimate the limit of point-estimate accuracy, such as mean-square-error, then a variance-based metric is the proper choice. For probabilistic prediction tasks (e.g. estimate the probability of emerging congestion), entropy-based metrics can reflect the detailed structure of probability distributions and they are related to the boundary of log-likelihood. There are two types of representations, *parametric* and *non-parametric* representations. The parametric approach uses a small number of parameters and a prior form (e.g. mean and variance of Gaussian) to represent a distribution. It has affordable complexity when the output dimensionality is high. For example, estimating uncertainty at each location in network-level traffic forecasting generally involves hundreds of road segments. The non-parametric representation approximates the true distribution by finite elements. It is more complex than a simple prior form but more accurate. It is compatible with low-dimensionality and safety-sensitive applications, e.g. risk assessment of a single autonomous vehicle.

The quantification method is central to uncertainty estimation. Different methods consider the concept of "predictability" from different angles. Recall the flowchart in Fig.1.1. Estimating the average aleatoric uncertainty (limit of prediction accuracy) requires a model-free approach because the result indicates whether it is worth increasing the model complexity. This stage does not require a real-time and fast algorithm. In the model deployment phase, the choice of quantification method depends on the application purpose. The inference speed is important. For example, in the case of trajectory forecasting related to the safety of autonomous vehicles, lightweight and faster

methods are preferred due to the limited reaction time (Gao et al., 2020). Conversely, for long-term data accumulation and offline continuous learning purposes, inference speed is not the focus.

That modelling strategy is critically important seems a natural and almost trivial remark. However, similar to the data-driven traffic modelling field, using a correlation-based or causation-based model has not been deeply discussed in the literature yet (We refer the readers to the review of Abdar et al. (2021) and the survey of Arnez et al. (2020)). How to identify "rare samples" is naturally related to the generalizability problem, which refers to the ladder of causation discussed in section 1.1.1. Correlation-based uncertainty quantification models sometimes give confusing and unreasonable results due to their limited generalizability (Bahari et al., 2022). When the application purpose requires suppressing those spurious correlations and learning the true causal mechanism is possible, we must implement these in the quantification approach. When it is not possible, we also need to point out the drawback of the used method to avoid over-confident predictions.

In summary, the three intertwined factors, the uncertainty representation, the quantification method, and the modelling strategy must be considered comprehensively for reliable and reasonable uncertainty quantification.

1.2 Scientific gaps

This thesis addresses multiple scientific gaps in the modelling-data collection cycle in Fig.1.1 and in the methodology presented in Fig.1.2. These gaps fall into the following four research topics.

1. This thesis studies how to extract explainable dynamic spatio-temporal associations in a deep-learning-based macroscopic traffic forecasting model. This is not directly related to uncertainty quantification or predictability analysis. But this section will propose, identify, and explain several helpful principles for guiding further model designing and uncertainty quantification.

2. This thesis studies how to evaluate whether the initial given macroscopic traffic dataset can satisfy the desired accuracy requirement without building any predictor. The special properties of traffic phenomena must be implemented in the estimation scheme.

3. This thesis considers quantifying the input-dependent aleatoric and epistemic uncertainty when deploying the model in the real-time data stream. These results will give insights into the predictability of macroscopic traffic state and the role of different factors.

4. This thesis uses similar ideas and concepts to study the uncertainty in microscopic trajectory prediction tasks. In this part, we extend the previously developed methods and address the generalizability problem for intention forecasting.

Next, we will sequentially elaborate on these 4 scientific gaps.

As discussed at the end of section.1.1.1, recently many emerging methods tried to combine the interpretability of knowledge-based models with the strong function-fitting ability of deep learning models. One of the most representative ways is extracting "salient" attention weights from a deep neural network (Veličković et al., 2018) and regarding them as the quantified "influence" among road links. For example, Gated-attention networks (Zhang et al., 2018) or pair-wise spatial attention networks (Do et al., 2019) can predict traffic volume/state and give the corresponding dynamic spatial-temporal correlations. However, this method has two issues. The first one is that traffic flow theory is not considered when designing this post-hoc interpretation. We notice that many studies are exploring the possibility of using deep neural networks or other data-driven methods to solve partial differential equations (e.g. Sirignano & Spiliopoulos (2018)). Thus LWR-like model can be implemented in a deep learning model. Second, the uni-directional propagation of congestion allows modelling causation-like (but perhaps not true causation) spatial-temporal associations but most studies did not focus on these directed relationships. Therefore, designing a new asymmetric graph attention mechanism to better explain what rules the deep learning model has learnt from data is still necessary.

Estimating the average limit of predictability of univariate and multivariate time series is an important research topic. One of the most well-known methods is the so-called *Lyapunoc exponent* (Wolf et al., 1985). In Chaos analysis, delayed embedding and phase space reconstruction (PSR) gives estimated value (Packard et al., 1980; Rosenstein et al., 1993). Nair et al. (2001) and Shang et al. (2005) used this approach to analyse traffic flow time series and observed significant chaos. The drawback of this metric is the relatively high computational complexity brought by PSR (Lan et al., 2008) and it cannot directly give the lower bound of a type of prediction error. The second way is maximum likelihood learning, for example, using the Gaussian process (Idé & Kato, 2009; Yuan et al., 2021). But this approach is model-based. As discussed before, the average limit of predictability generally needs to be estimated before modelling. Another group of methods are *entropy-based* approach. Song et al. (2010) is one of the first papers that employed this approach to analyse the limit of predictability of human mobility. This strategy is also studied for the multivariate traffic data stream, such as the speed on a ring freeway (Wang et al., 2015a), travel time (Li et al., 2019), and traffic flow (Darmon, 2016). However, these methods have at least one of the following drawbacks. One, they only consider discrete stochastic processes. Two, they assume that different sensors on the road networks are independent. Three, the special spatio-temporal properties of traffic phenomena are ignored. A more comprehensive method that can address all these issues is still needed.

As for input-dependent aleatoric and epistemic uncertainty quantification, this problem has already been deeply-studied in the artificial intelligence community. We have enough tools at hand already. These widely-used methods include Bayesian neural networks (van Hinsbergen et al., 2009), Monte-Carlo dropout in inference stage (Kendall & Gal, 2017), deep ensembles (Lakshminarayanan et al., 2016), and some recently-developed one-pass models such as radical-basis function (van Amersfoort et al., 2020) or deep regression model (Amini et al., 2020). Some of these methods are also used in traffic modelling and prediction. For example, Bayesian network for traffic flow pre-

diction (Zheng et al., 2006; Fu et al., 2020), deep echo state networks (McDermott & Wikle, 2019), and deep ensembles (Deng et al., 2016; Chen et al., 2021b). The major gap here is that all these studies focus on modelling competition or performance comparison instead of understanding the role of uncertainty in the modelling-data cycle in Fig.1.1. Evaluating the predictability in model deployment and figuring out what factors restrict the predictability of macroscopic traffic has not drawn enough attention in the literature.

When we scale down to microscopic driving behaviour modelling and prediction, uncertainty quantification has radically different requirements. With the fast development of autonomous vehicles and related large-scale datasets, e.g. Waymo (Sun et al., 2020), nuScenes (Caesar et al., 2020), and Argoverse (Chang et al., 2019), trajectory-based intention and behaviour prediction has attracted a lot of attention. Uncertainty is highly related to safety evaluation. Abdar et al. (2021) gives an overview of existing studies on this topic. The gap here is that most papers directly use those parametric and correlation-based uncertainty estimation methods developed specifically for scalar variables. However, three critical factors have not been deeply discussed yet. First, vehicles move on a 2D plane restricted by arbitrary road layouts so neither classical covariance-based metric nor the parametric representation can reflect the complexity of its future 2D locations. Second, the urban driving environment is so complex that no dataset can cover all scenarios. Therefore, the generalizability of the used method to unseen new cases cannot be ignored. Third, how to relate the learnt uncertainty to our domain knowledge on driving behaviours is not the focus of the AI community. We need a novel method to fill the gaps of uncertainty representation, model generalizability, and the lack of domain knowledge in trajectory prediction.

1.3 Research objective and questions

Based on the discussion above on research background and scientific gaps, now we formulate the overall research objective of this thesis as follows:

Technically, how to systematically quantify the uncertainty of traffic prediction on multiple levels?

Scientifically, how predictable are traffic dynamics at different scales?

To the best of our knowledge, until writing this thesis, there are no systematic studies that address the issues mentioned before. This thesis aims to fill these gaps and realize the overall objective by answering the following 8 key research questions. They are clustered into 4 groups. Each group contains 1-2 research questions that correspond to one research gap.

Question 1: For macroscopic highway networks, how to build a deep-learning-based traffic forecasting model that can provide post-hoc, causation-like interpretations on spatial associations?

We start with the high-level macroscopic traffic dynamics modelling and prediction. The first research question focus on the interpretability of deep-learning-based traffic forecasting models. As discussed before, due to limited observability, causal sufficiency cannot always be satisfied. In most cases only macroscopic traffic states, such as flow and speed, are available. The true "confounder" (or the inner driving force) of congestion evolution, such as traffic demand, route choices, and microscopic driving behaviours are generally unavailable in the dataset. However, seeking post-hoc interpretations from such a model by designing novel modules is still possible. The novel module must be compatible with traffic flow theory, especially about how to model the spatial associations of congestion spreading. By considering domain knowledge, this interpretation of spatial relationships can be restricted to be causation-like. Such a method can further give principles of model construction and hyperparameters tuning. These potential findings are important for guiding further studies on predictability and uncertainty quantification of multi-scale traffic forecasting. [*Chapter 2*]

Question 2: Given a dataset that is large and representative enough, what are the model-free, theoretical lower bounds of predictive accuracy for probabilistic models and deterministic models respectively?

Question 3: How to directly estimate the spatial-temporal distribution of predictability of traffic speed before building any forecasting models?

These two research questions are related to the model-free approach of estimating the *average* limit of predictability for macroscopic highway networks. They together address one important issue: before diving into building the forecasting model, how to estimate whether the provided data at hand or the target traffic process itself are enough "predictable"? Sometimes researchers are too obsessed with improving modelling techniques but ignore where is the limit. In practice, two types of predictors are widely-used, deterministic models that give a point prediction, and probabilistic models that predict the output distribution. Correspondingly, we need to develop both theories and numerical methods to estimate the limit of mean-square-error and negative-log-likelihood. The conclusions are expected to put detailed model bench-marking into perspective. If the estimated limit can satisfy our requirement, we can go to uncertainty quantification in modelling. [*Chapter 3*]

Question 4: How to estimate the aleatoric and epistemic uncertainty of each specific prediction for highway networks?

Question 5: If the answer to question 4 suggests the predictability of highway traffic patterns is limited. What explains this limited predictability?

These two questions focus on the uncertainty quantification of a specific traffic forecasting model. Saying we need to quantify both aleatoric and epistemic uncertainty of the current input traffic states. Question 5 develops the quantification method while question 6 explains what the estimated aleatoric and epistemic uncertainty means. The answer to question 5 can point out what is missed in the current macroscopic traffic data and question 6 addresses the day-to-day recurrence of congestion patterns. The findings can guide data collection and continuous learning in a real-time data stream. [*Chapter 4*]

Question 6: In microscopic trajectory forecasting, how to give reasonable spatial uncertainty measurements that can remove the influence of arbitrary road layout?

Question 7: In the urban driving environment, how to build a generalizable motion forecasting model that can give both reliable predictions and uncertainty estimation in new scenarios?

These two research questions are about microscopic trajectory prediction. We focus on modelling the interactions among multiple road agents and predicting their behaviours in a short future around (3 s). This topic is important for developing a more realistic simulation environment for self-driving cars. The uncertainty quantification is critical for safer decision-making and motion planning. However, different from macroscopic traffic forecasting, trajectory prediction has its unique requirements (as mentioned before), such as complex urban driving scenarios, 2-dimensional distributions, and arbitrary road layouts. So the generalizability of the model and a reasonable uncertainty metric/representation is critically important. The classical correlation-based uncertainty quantification method must be extended to this problem by considering causal relationships. Further, we need to identify which types of data are indispensable but currently unavailable for modelling human driving behaviours. [*Chapter 5*]

By answering all the research questions above, we expect to obtain a bird-view and deeper understanding of the uncertainty and predictability in multi-level traffic modelling and prediction.

1.4 Contributions

By answering the sub-questions and achieving the overall research objective, the major contribution to the field of multi-scale traffic prediction is a systematic approach to quantify interpretable and reasonable uncertainty in traffic dynamics. Such an approach can comprehensively measure how predictable traffic is at different levels. In this subsection, we provide the contributions of our thesis to each scientific gap. These contributions include proposed novel models and methodology, theoretically-proved mathematical theorems, and toolkits or building blocks for different applications.

1.4.1 Traffic forecasting models

[**Chapter 2**]

- A new graph attention mechanism, the *dynamic graph convolution* (DGC) module, is proposed to learn causation-like real-time spatial associations of graph signal series. The DGC module is implemented in a recurrent neural network, the dynamic graph convolutional network (DGCN), to predict multistep macroscopic traffic speed evolution.

- The proposed DGCN model combines the advantage of a data-driven and simulation-based approach. The prediction is accurate and the learnt dynamic spatial associations are consistent with traffic flow theory.

- The DGCN also gives an empirical principle for tuning the order of the adjacency matrix (receptive field) based on the time interval and average distance between adjacent road links in graph-based prediction models. This finding can help accelerate the hyper-parameter tuning of other similar models.

1.4.2 Average predictability for highway networks

[Chapter 3]

- We prove that the conditional entropy of the output variable given the input variable gives the lower bound of negative-log-likelihood. The discrete form of Fano's theorem is extended to multivariate continuous stochastic processes, and thus gives the corresponding lower bound of mean-square-error (MSE) or determinant of the covariance matrix (DCM). Their ratio can be interpreted as the strength of temporal correlations.

- We challenge the stationarity assumption of traffic time series and propose to consider both the cyclo-stationarity and localization of traffic congestion in the estimate of conditional entropy.

- Practically, we show that the estimated average predictability is a function of time and location. This gives traffic managers and practitioners a tool to identify the most unpredictable time slots and positions from the result. The results thus provide insights into the vulnerability of highway networks.

1.4.3 Uncertainty quantification for highway networks

[Chapter 4]

- Using the beta distribution as the prior for modelling predicted highway speed, we propose a fully-convolutional probabilistic graph neural network to form deep ensembles. Both the aleatoric and epistemic uncertainty of each input sample can be quantified by variational inference.

- We show that the results of input-dependent aleatoric and epistemic uncertainty can put performance comparison into perspective. They quantify where is the boundary of modelling and how frequently rare congestion patterns happen.

- The findings offer a new pipeline for continuous data collection and modelling improvement.

1.4.4 Trajectory prediction and uncertainty

[Chapter 5]

- We proposed a novel human-driven vehicle's intention and trajectory prediction model, the so-called Uncertainty-Quantification networks (UQnet). We added a causal part to the model and a novel causal regularization to the loss function so UQnet can tell apart the spurious correlations between the driving behaviours of surrounding agents and environment biases. UQnet is state-of-the-art in the open INTERPRET challenge and shows significantly higher generalizability to new scenarios.
- We extend the 1D parametric uncertainty quantification method to 2D non-parametric models. UQnet directly uses a 2D histogram (a heatmap) to approximate the true distribution.
- We demonstrate that the proposed 2D non-parametric approach and UQnet can correctly measure the risk of intention prediction due to its higher generalizability and more reasonable uncertainty metrics. So this novel method can be used for refining/distilling a large dataset and finding out those truly rare samples for generalizability evaluation.

1.5 Thesis outline

The outline of this paper-based thesis is shown in Figure 1.3. It has a "sandwich" structure that sequentially contains the introduction, main body, conclusions and perspective. The three parts are connected by black arrows. In line with the scientific gaps and research questions, this thesis has four chapters as the main body. Each chapter answers one to several research questions and it is based on one paper that was written as the first author during the PhD. At the beginning of each chapter, we declare whether the paper is published, If yes we state where it is published. If not we state whether it is at the time under review and where it was submitted. To avoid potential confusion, the text in each chapter is completely the same as that published in the journal.

The four chapters are grouped into two parts, macroscopic and microscopic traffic modelling and prediction. They are put in two areas with different background colours. For each separate chapter, the dashed box represents that this study is not directly related to uncertainty quantification but provides an essential prerequisite. Solid boxes mean that this section develops theories, methods, and models of uncertainty quantification and predictability analysis.

The blue arrows are added to clarify the relationships between different chapters. Chapter 2 studies how to learn the causation-like real-time spatial associations in macroscopic traffic state forecasting by a deep learning model. Its findings on localized spatial associations inspire the assumptions used in Chapter 3 and its conclusions on hyper-parameter tuning from the insights of traffic flow theory guides the model

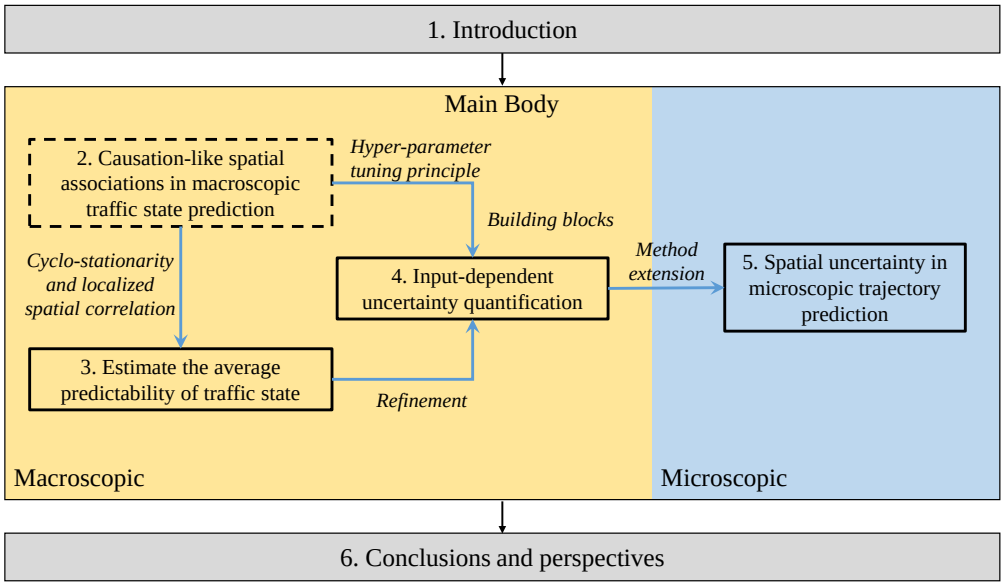


Figure 1.3: Outline of this thesis

construction in Chapter 4. Chapter 3 uses the empirical rule of congestion propagation to partition the dataset and estimate location-specific predictability. While chapter 4 directly uses the proposed attention mechanism in chapter 2 to construct the basic uncertainty quantification model. The model-based input-dependent uncertainty estimation method proposed in chapter 4 refines the model-free average predictability analysis in chapter 3 and their results are consistent. Chapter 5 addressed the same type of issue as chapter 4, but the methodology is extended to more general 2D cases in trajectory prediction and the causal mechanism is implemented.

In Chapter 6, we summarize the key findings and the overall conclusion. The implications on science and practice and some recommendations for further research topics will also be presented.

Chapter 2

Spatial associations in macroscopic traffic prediction

This chapter presents a deep learning approach that can explicitly give the learnt traffic-state-dependent spatial associations among highway links in short-term traffic forecasting. The proposed dynamic graph convolutional module combines the interpretability of knowledge-based traffic flow models and the fitting ability of data-driven deep neural networks. Implementing such a module in recurrent neural networks can give accurate and explainable predictions. The analysis proves that deeper neural networks face the risk of being confused by the spurious correlations of congestion patterns. There exists an optimal perception field that can remove the spurious correlations and allow the model to learn "causation-like" spatial associations. This conclusion is critical for designing robust prediction models and reasonable uncertainty quantification methods in the following chapters.

This chapter is published as a journal article: Li, G., V. L. Knoop, H. van Lint (2021) Multistep traffic forecasting by dynamic graph convolution: Interpretations of real-time spatial correlations, *Transportation Research Part C: Emerging Technologies*, 128, p. 103185
In section.2.4.3: Discussion on model interpretability, an extra paragraph about causal confusion is added. The other sections are the same as the published paper.

ABSTRACT

Accurate and explainable short-term traffic forecasting is pivotal for making trustworthy decisions in advanced traffic control and guidance systems. Recently, the deep learning approach, as a data-driven alternative to traffic flow model-based data assimilation and prediction methods, has become popular in this domain. Many of these deep learning models show promising predictive performance but inherently suffer from a lack of interpretability. This difficulty largely originates from the inconsistency between the static input-output mappings encoded in deep neural networks and the dynamic nature of traffic phenomena. Under different traffic conditions, such as freely-flowing versus heavily congested traffic, different mappings are needed to predict the propagation of congestion and the resulting speeds over the network more accurately. In this chapter, we design a novel variant of the graph attention mechanism. The major innovation of this so-called dynamic graph convolution (DGC) module is that local area-wide graph convolutional kernels are dynamically generated from evolving traffic states to capture real-time spatial dependencies. When traffic conditions change, the spatial association encoded by the DGC module changes as well. Using the DGC, we construct a multistep traffic forecasting model, the *Dynamic Graph Convolutional Network* (DGCN). Experiments using real freeway data show that the DGCN has a competitive predictive performance compared to other state-of-the-art models. Equally importantly, the prediction process in the DGCN and the trained parameters are indeed explainable. It turns out that the DGCN learns to mimic the upstream-downstream asymmetric information flow of typical road traffic operations. Specifically, there exists a speed-dependent optimal receptive field – which governs what information the DGC kernels assimilate – that is consistent with the back-propagation speed of stop-and-go waves in traffic streams. This implies that the trained parameters are consistent with traffic flow theory. This optimal hyper-parameter can avoid learning spurious correlations between links but capture the true cause of the future traffic state. We believe this research paves a path to more transparent deep learning models applied for short-term traffic forecasting.

2.1 Introduction

Accurate and reliable short-term traffic forecasting is one of the core functions of Intelligent Transportation Systems (ITS). Predicting the dynamic evolution of traffic has been a popular research topic for many decades, both on a single corridor (e.g. Van Lint et al. (2005)) and on large road networks (e.g. Fusco et al. (2016)). For broad and recent overviews of methods and challenges in this active research field, we refer the readers to Lana et al. (2018) and Ermagun & Levinson (2018). The discussion hereafter focuses on three key methodological challenges related to real-time traffic forecasting for travel information provision and traffic management: interpretability, observability, and uncertainty.

2.1.1 Background: network-level traffic prediction

The evolution of traffic network conditions can be viewed as a superposition of various intertwined dynamical processes, including origin-destination (O-D) traffic demand, route choice patterns, queuing and congestion backward propagation, driving behaviours that govern emerging characteristics, etc. Traffic condition forecasting requires either implicitly or explicitly considering the dynamics in both demand and supply. In the literature, many different methods are proposed to tackle short-term traffic prediction tasks. They range from simple approaches such as conditional averaging (Davis & Nihan, 1991; Smith & Demetsky, 1997), auto-regression and time series models (Ahmed & Cook, 1979; Castro-Neto et al., 2009), to sophisticated machine learning approaches (e.g. Hamner (2010); Huang et al. (2014); Polson & Sokolov (2017)), and simulation-based approaches in conjunction with sequential data assimilation methods (Wang et al., 2005; Van Hinsbergen et al., 2011). One of the biggest advantages of data-driven approaches – which encompass all methods except those using simulation-based models – is that explicitly disentangling the demand and supply dynamics is not required. Available data can be fully explored regardless of whether the model simulates the physical process between the input data (speeds, flows, travel times, trajectories, and other information) and the desired output quantity. Instead, data-driven methods aim to learn correlations between inputs and outputs. Their performances mainly depend on the type of data, model structures, and the optimization strategy, whether these mappings are explainable in traffic science. However, the interpretability of data-driven models becomes a new problem. Taking the example of deep neural networks, neither the complex model containing numerous learnable parameters nor the internal states of hidden layers are easy to explain. If a traffic prediction model is only used as a “black-box” inference engine, it may suffice for many purposes, but certainly not for traffic management and control.

Simulation-based methods offer an alternative for explainable network-level traffic forecasting since these methods explicitly delineate the prediction task into constituent sub-problems and compute the resulting traffic state by “white box” physical and behavioural models. The simulation-based approach gives a comprehensive solution for traffic state estimation, prediction, and for control optimization (Wang et al., 2008). However, this transparent methodology also comes with serious challenges.

The first challenge is the limited observability of some constituent parameters and variables. There are much more unknowns that govern the traffic dynamics than the limited bits of information derived from sensors, let alone project them in the short-term future. Particularly, demand and route choice patterns are difficult to be fully observed (Viti et al., 2014; Castillo et al., 2012; Krishnakumari et al., 2019). Some other key variables and parameters, such as vehicle densities and road capacities, are also difficult or costly to observe with or derive from sensor data. Although with vehicle counting, accumulation and thus density is available in principle, inevitable small errors may lead to unbounded biases (Xie et al., 2018; Nantes et al., 2016; Bhaskar et al., 2014).

Secondly, even if all required data were observable, because of the stochastic nature of traffic dynamics, the prediction is still influenced by uncertainty. The uncertainty comes from the predicted boundary conditions (e.g. demand and capacity constraints) and from the traffic flow process itself. Small errors in the projected inflow at a certain position, or in the expected capacity of an intersection, may cause large prediction errors in the resulting traffic states. Additionally, where and when errors are made also matters. These errors may have larger consequences at some specific locations and during some time slots. For example, the spreading of congestion in a road network largely depends on whether queues spill over some junctions and off-ramps, and these differences can be significant (van Lint et al., 2012a; Knoop et al., 2015).

To address these challenges, it is necessary to combine the best of both data-driven and simulation-based worlds: an adaptive learning solution that uses whatever available incomplete data and produces predictions that are at least partly explainable.

2.1.2 Contributions and outline

In this paper, we investigate deep learning models that are able to learn explainable dynamic spatial correlations, which can be defined as traffic-condition-dependent interactions among adjacent links. In deep learning models, spatial correlation is generally implicitly modelled by stacked convolutional or graph convolutional layers. To make this process more explainable, we propose a novel graph attention variant in the form of Dynamic Graph Convolution (DGC) modules. DGC modules assume that the spatial correlation depends on the road network connectivity, locations, and traffic states in a receptive field. In other words, it learns to mimic the location-specific and condition-dependent propagation of information. DGC modules are implemented in a recurrent encoder-decoder model to realize multistep speed forecasting. We call this prediction model a *Dynamic Graph Convolutional Network* (DGCN). By tracking the generated dynamic kernels, we explain what spatial correlations are learnt by this model in different scenarios. The key contributions of this paper are:

- **Model construction**: The dynamic convolution is extended to graphs by considering the properties of traffic networks. Based on a new DGC module, a novel multistep traffic forecasting model, the DGCN, is proposed.
- **Performance comparison**: Validated on real freeway network datasets, the DGCN shows competitive predictive accuracy compared to a selection of state-of-the-art models.

- **Model interpretation**: The generated dynamic convolutional kernels offer insights into what the model has learnt from the data. It turns out that DGC modules encode an explainable proxy to the basic traffic flow theory.

The rest of this paper is organized as follows: Section 2.2 describes the details of the DGCN. Next, section 2.3 compares the performances of DGCN against some baseline models using real freeway-network speed datasets. Section 2.4 presents model interpretations. Finally, Section 2.5 draws conclusions and gives several potential research directions.

2.2 Method: deep learning models

In this section, we first give a brief overview of deep learning models applied to short-term traffic prediction. How spatial correlations are implemented in different models is our focus. Then, specific to our purpose, some notations and the problem formulation are defined in subsection 2.2.2. Next, we introduce the basic building block, the DGC module, and combine it with an RNN encoder-decoder to construct the novel DGCN prediction model in 2.2.3.

2.2.1 Overview of related works

Recently, deep learning, especially deep neural networks (DNN), brought new possibilities to the short-term traffic forecasting domain. DNNs capture spatiotemporal features by a large amount of organized trainable parameters. Many papers propose a variety of DNN models to improve predictive performance. For example, Huang et al. (2014) proposes deep belief networks (DBN) for traffic flow prediction. Polson & Sokolov (2017) combines a linear model with $L1$ regularization and a layer activated by a *tanh* function to predict the sharp transition of traffic flow. Increasingly, standardized modules, such as convolutional layers and recurrent layers, are used to construct new DNN-based predictors. For example, Ma et al. (2015) uses long-short-term memory (LSTM) networks to predict speed evolution on a corridor. Ma et al. (2017) converts traffic dynamics to heat map images and employs deep convolutional neural networks (CNN) for speed prediction. To apply CNN on non-Euclidean structures, the convolutional operator is extended to graph convolution (Kipf & Welling, 2017; Hamilton et al., 2017). Such graph convolutional networks (GCN) are suitable for network-level traffic forecasting because a road network can be naturally represented by a graph. For example, Yu et al. (2018) combines spectral-domain graph convolution and temporal gated convolution to predict network-level traffic speed, so-called spatio-temporal graph convolutional networks (STGCN). In many of these studies, the central focus is on improving forecasting precision and reliability and much less on model interpretation. To improve the interpretability of deep learning models, an increasing number of studies attempt to explain what neural networks learn from data and particularly spatial correlations.

In the literature, there are two main methods to achieve this. The first one is implementing predefined components to represent spatial correlations. These components

stay *static* after training. The idea is similar to those variants of regression models considering spatial interactions (e.g. Kamarianakis & Prastacos (2005); Min & Wynter (2011)). For example, Li et al. (2018) considers congestion as a bidirectional diffusion process described by random walks. Diffusion convolution is embedded into gated recurrent units (GRU) resulting in a so-called diffusion convolutional RNN (DCRNN). Cui et al. (2019) proposes traffic graph convolution (TGC) to extract rich spatial features on small graphs. The design of this TGC considers the maximum speed of vehicles and results show that statistically larger kernel weights emerge at frequently congested intersections. Zhang et al. (2019b) combines a similar graph convolutional layer, a GRU encoder-decoder, and a temporal attention layer in one model, and calls this an attention graph convolutional-seq2seq (AGC-seq2seq) model. The authors define the average matrix of graph convolutional kernels as spatial dependencies and draw similar conclusions as in Cui et al. (2019). This approach can give a *static* mapping that encodes spatial dependencies that work best on average. However, spatial dependencies are fundamentally *dynamic* in traffic flows. This inconsistency restrains these models' interpretability.

The second option is letting the model learn *dynamic* spatial correlations from data. The graph attention mechanism (Veličković et al., 2018) is a promising solution in this direction. In graph attention networks (GAT), graph convolutional kernels are no longer static, but are calculated from real-time node features. For example, Zhang et al. (2018) proposes gated attention networks (GaAN) to predict traffic states. An extra soft gate is added based on GAT. Do et al. (2019) uses a simplified pairwise spatial attention layer to predict traffic volumes and directly produce dynamic spatial correlations in the form of heat maps. However, some important features of traffic flows are ignored when designing these models. First, the propagation of information is asymmetric in traffic networks. In freely flowing areas, kinematic waves move with the driving direction; in congested traffic, they move against the traffic flow. In the former case upstream information suffices for prediction; in congestion, the model needs to combine up- and downstream information. Second, how far up- or downstream the model would have to "look for information" to accurately track the evolution of congestion depends on the data granularity and the network topology. Third, some unavoidable "unseen" factors, such as an on-ramp branch that is not covered by sensors, can also influence the prediction. The ability to explain how a model gives reliable predictions by adjusting dynamic spatial correlations for incomplete data is critically important for decision-making in traffic control. Implementing the properties above may yield a more accurate and more explainable mapping.

To develop dynamical mappings for these spatio-temporal relationships, we need a spatial attention mechanism that adapts in real-time to the prevailing traffic conditions. This study aims to fill this gap by designing a new spatial attention variant considering the upstream-downstream asymmetry of traffic dynamics and the influence of incomplete data.

2.2.2 Mathematical Formulations and Preliminaries

Traffic dynamics on a road network can be written as a spatio-temporal graph:

$$\mathscr{G}(\mathscr{V}_N, \mathscr{E}, \mathbf{A}_{N\times N}; \mathbf{X}_{T\times N\times C}), \tag{2.1}$$

in which the subscripts indicate the dimensions of the tensor. $\mathscr{V}_N$ is the set of N nodes, $\mathscr{E}$ is the set of edges (connecting these nodes), and $\mathbf{A}$ is the adjacency matrix. For T discrete time steps of duration Δt, the evolution of traffic states is represented by the feature tensor $\mathbf{X}$. The feature vector of node i at time t is denoted as $\mathbf{X}_i^t \in \mathbb{R}^C$, which can be composed of C traffic variables, such as speed, flow, etc. Short-term traffic forecasting is formulated as a sequence-to-sequence regressive task in this study. On a road network $\mathscr{G}$, the input is the feature tensor observed in the past m timesteps $\mathbf{X}^{\text{obs}}_{m\times N\times C}$, the output is the predicted feature tensor of the next p timesteps $\hat{\mathbf{X}}^{\text{pred}}_{p\times N\times C}$ that maximizes the following conditional probability:

$$\hat{\mathbf{X}}^{\text{pred}}_{p\times N\times C} = \underset{\mathbf{X}^{\text{real}}_{p\times N\times C}}{\arg\max}\ Pr(\mathbf{X}^{\text{real}}_{p\times N\times C} | \mathbf{X}^{\text{obs}}_{m\times N\times C}; \mathscr{G}) \tag{2.2}$$

We denote $m\Delta t$ as the observation window and $p\Delta t$ as the prediction horizon. On a spatiotemporal graph $\mathscr{G}$, we denote the set of all nodes within k hops from a central node v_i as $\mathscr{N}_i^k$ (including the central node itself), in which k is the so-called *receptive field*. If the latent representation of v_i is a learnable function of all feature vectors within the receptive field, we have:

$$\vec{y}_i = \varphi_{\text{agg}}(\vec{x}_{\mathscr{N}_i^k}; \mathscr{G}) \tag{2.3}$$

in which $\varphi_{\text{agg}}()$ is the *k-walk graph aggregator*.

2.2.3 Dynamic graph convolutional networks

One of the earliest dynamic convolutional networks is proposed by Brabandere et al. (2016), the so-called Dynamic Filters Network (DFN). In that model, pixel-wise dynamic convolutional kernels are used to capture the movement of objects and to predict the next several frames of a video. Based on a similar idea, we design a dynamic graph convolution to capture the propagation of congestion. The structure is shown in Fig.2.1. The input of the DGC is split into two tensors, feature maps $\mathbf{X}'$ and traffic states $\mathbf{X}$. Each DGC module has three parts: (1) a filters generation network (FGN) computes dynamic graph convolutional kernels from varying feature maps; (2) the generated kernels are then applied in a local-wide graph convolution with traffic states; (3) post-processing adjusts the dimension of outputs.

The DGC module differs from the DFN (Brabandere et al., 2016) in three ways. First, it is extended from Euclidean space to graphs by involving an adjacency matrix that encodes the connectivity. This extension is necessary because the number of neighbour nodes is arbitrary on a graph. Another method is converting graphs to Euclidean images by stitching all roads in order and applying CNN (Ma et al., 2017). However, the neighbours around intersections on the graph may locate far away in the image. For extracting explainable spatial correlations from DNN, GCN is a better choice because the real geographical distances among roads are conserved. Second, the FGN uses a more computationally efficient one-step process instead of the heavy convolutional encoder-

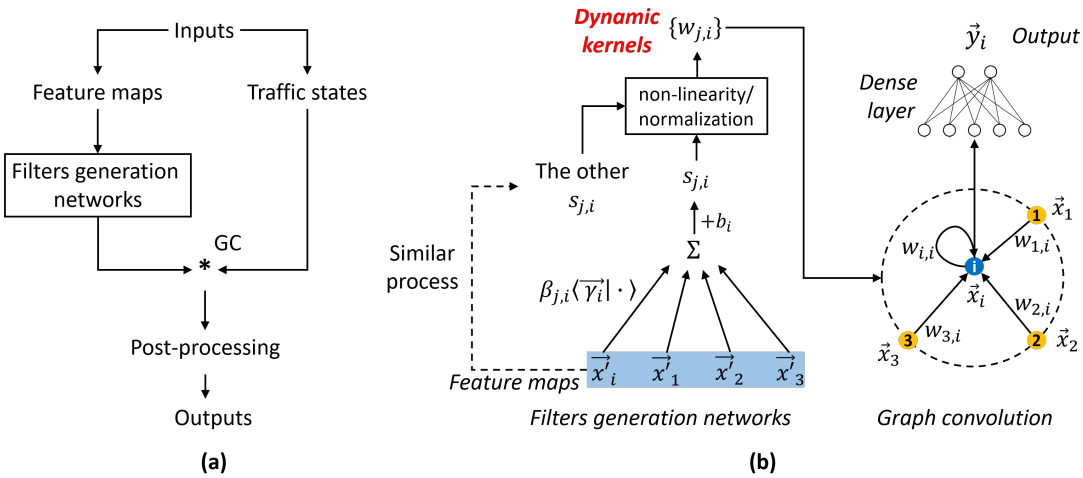


Figure 2.1: DGC module: (a) the structure of a DGC module; (b) the details of filters generation networks and the DGC graph aggregator

decoder. Third, in the FGN of the DFN, convolutional kernels are shared among all pixels. Considering the fact that each node is unique in a road network in terms of infrastructure, capacity, and link connectivity, we add position-specific parameters in the FGN of the DGC. The mathematical formula of FGN is given below:

$$\begin{cases} \mathbf{A}_{\mathbf{k}N\times N} = Ci[(\mathbf{A}_{N\times N})^k] \\ \mathbf{S}_{N\times N}(t) = (\mathbf{A}_{\mathbf{k}N\times N} \odot \mathbf{B}_{N\times N})\mathbf{X}'_{N\times C}(t)\mathbf{\Gamma}_{C\times N} + \mathbf{b}_N \\ \mathbf{W}_{N\times N}(t) = \sigma_1(\mathbf{A}_{\mathbf{k}N\times N} \odot \mathbf{S}_{N\times N}(t)) \end{cases}, \tag{2.4}$$

where $\mathbf{B}, \mathbf{\Gamma}$ and $\mathbf{b}$ are trainable parameters. $Ci[\,]$ pins all non-zero elements in a matrix to 1. $\mathbf{A_k}$ is the k-hop adjacency matrix. $\sigma_1()$ is the nonlinear activation or normalization function. Because the input feature map $\mathbf{X}'(t)$ varies at each instant, the generated graph convolutional kernel $\mathbf{W}(t)$ also evolves with time. Then the generated dynamic kernels $\mathbf{W}(t)$ is applied in a space-domain graph convolution and a shared node-wise fully-connected layer is used for post-processing:

$$\begin{cases} \mathbf{H}_{N\times D} = (\mathbf{W}_{N\times N}(t) \odot \mathbf{A}_{\mathbf{k}N\times N})\mathbf{X}_{N\times D} \\ \mathbf{Y}^{\text{out}}_{N\times C'} = \sigma_2(\mathbf{H}_{N\times D}\mathbf{V}_{D\times C'} + \mathbf{b}'_{C'}) \end{cases}, \tag{2.5}$$

in which $\mathbf{V}$ and $\mathbf{b}'$ are trainable parameters. $\sigma_2()$ is an activation function. To better clarify this process and for easier readability, equation (2.6) depicts the graph aggregator in (2.4) and (2.5) for a single node i:

$$\begin{cases} s_{j,i}(t) = (\sum_{j\in\mathcal{N}_i^k} \beta_{j,i}\langle\vec{\gamma}_i, \vec{x'}_j(t)\rangle) + b_i \\ w_{j,i}(t) = \sigma_1(\{s_{j,i}(t) \mid j \in \mathcal{N}_i^k\}) \\ \vec{h}_i = \sum_{j\in\mathcal{N}_i^k} w_{j,i}(t)\vec{x}_j \\ \vec{y}_i = FC_{\theta_{C'}}(\vec{h}_i) \end{cases} \tag{2.6}$$

The first equation in (2.6) describes how the initial dynamic graph convolutional kernels $s_{j,i}$ are generated from feature vectors. It also indicates the parameter sharing strategy: $\vec{\gamma}_i$ is the shared parameter in the very receptive field. $\beta_{j,i}$ represents the speciality of each node when generating initial weights. Each $s_{j,i}$ is a learnable k-walk graph aggregator of $\vec{x}'_{\mathcal{N}_i^k}$. For simplification, in the following parts we denote a DGC module with $N \times C'$ output dimension and k receptive field on a spatiotemporal graph $\mathscr{G}$ as follows:

$$\mathbf{Y}^{\text{out}}_{N\times C'}(t) = \text{DGC}_k(\mathbf{X}'(t), \mathbf{X}(t), C'; \mathscr{G}) \tag{2.7}$$

The DGC proposed here is essentially a variant of the graph spatial attention module (Zhu et al., 2019). In the section hereafter, we compare it with another variant, the *Graph Attention Network* (GAT) depicted by (2.8) and proposed in Veličković et al. (2018). We only discuss the single-head attention variant here, but the conclusions also hold for multi-head attention.

$$\begin{cases} \phi_w(\vec{x}_i, \vec{x}_j) = \langle FC_{\theta_{va}}(\vec{x}_i), FC_{\theta_{vb}}(\vec{x}_j) \rangle \\ w_{j,i} = \dfrac{\exp(\phi_w(\vec{x}_i, \vec{x}_j))}{\sum\limits_{j\in\mathcal{N}_i^k} \exp(\phi_w(\vec{x}_i, \vec{x}_j))} \\ \vec{y}_i = FC_{\theta_o}(w_{i,i} FC_{\theta_{va}}(\vec{x}_i) + \sum\limits_{j\in\mathcal{N}_i^k, j\neq i} w_{j,i} FC_{\theta_{vb}}(\vec{x}_j)) \end{cases} \tag{2.8}$$

Different from DGC, the GAT module computes pairwise similarities between a central node and its neighbours and then applies *softmax* normalization in each receptive field. Rearranging the feature vectors of neighbour nodes does not change the output. Although GAT can distinguish most types of node orders by overlapping multiple adjacent receptive fields, there exist traffic phenomena where GAT may fail to learn the right spatial correlations. One example is the upstream-downstream asymmetry discussed earlier. For a given road segment x_i it matters whether an upstream road segment or a downstream segment is congested. In the first case (upstream congestion) this may not affect the segment x_i at all, whereas in the other case we can expect this congestion to propagate upstream and thus affect the conditions on x_i. The GAT module does not distinguish these two cases. Other models, such as GaAN (Zhang et al., 2018), have the same limitation. At least in theory, the DGC module can explicitly learn these asymmetries.

Fig.2.2 presents a clarifying example. Consider a small corridor represented by seven nodes as shown in the left. Our focus is on the outputs of the three most central nodes here, $\vec{y}_1, \vec{y}_2, \vec{y}_3$. For simplicity's sake, we assume that all nodes can exhibit only two traffic states, a freely-flowing state represented by $\vec{x}_{\text{free}}$ versus a congested state $\vec{x}_{\text{cong}}$. We also consider a fixed size of the receptive field, $k = 2$. From a traffic flow theory perspective, the short-term evolution of these three nodes is related but quite different. Congestion propagates against the direction of flow (from right to left). Only in the case node 3 indeed becomes congested, the probability that node 2 becomes congested later in time increases dramatically. Only in that case node 2 also becomes congested, and later in time still, node 1 may become congested as well. But the three nodes are not likely to be influenced by the upstream (left) congestion because there may be a

stationary bottleneck, such as an on-ramp. Thus, for short-term prediction purposes, these three nodes need to be treated differently. However, GAT modules cannot achieve this. Because the initial weight ϕ is computed from the central node and one neighbour node, only two dynamic weights, ω_{ff} (for free-flow neighbour nodes) and ω_{fc} (for congested neighbour nodes) will be generated. No matter how these 3 congested nodes and 2 free-flowing nodes are ranged, the last equation in (2.8) gives the same output, see Fig.2.2 right top. The DGC module, in contrast to GAT, caters for these directional dynamics *by making the local weights conditional to the relative location ($\beta_{j,i}$) of all neighbour nodes*, as shown in Fig.2.2 right bottom. Each node has its unique weight. We refer to equations (2.6) versus (2.8) for comparing DGC and GAT for a single receptive field.

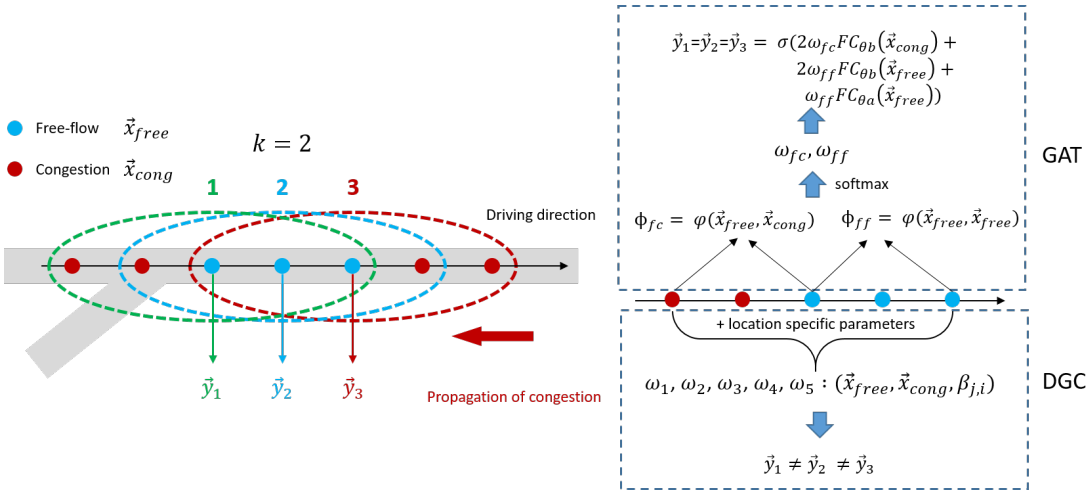
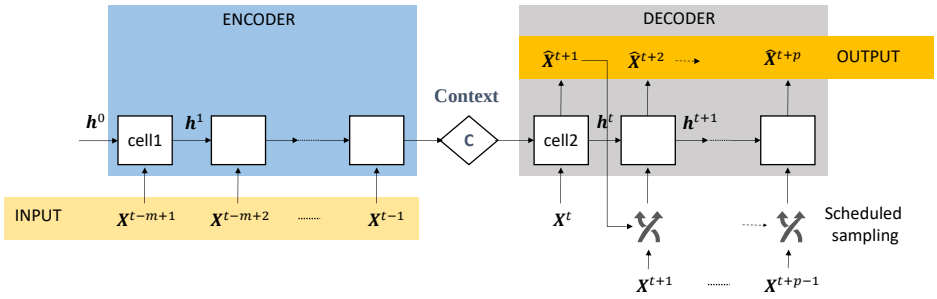
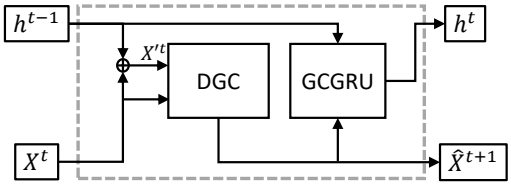


Figure 2.2: An example of directional effects: if $k = 2$, each receptive field contains three free-flow nodes and two congested nodes. GAT and GaAN give the same hidden presentation of the three central nodes, but DGC can distinguish them.

The RNN encoder-decoder (Kumar & Asger, 2018) is widely-used for seq2seq regression tasks, including applications for traffic prediction (e.g. Van Lint et al. (2005)). Each RNN cell receives the current input and the hidden state from last time step $(\mathbf{X}^t, \mathbf{h}^{t-1})$, generates the new output and hidden state $(\hat{\mathbf{X}}^{t+1}, \mathbf{h}^t)$. The encoder saves the last hidden state as the context vector $\mathbf{C}$. The context vector is copied as the first hidden state to initialize the decoder. Usually, two RNN cells of the same type but with different parameters are used in the encoder and the decoder respectively. Based on DGC modules, we propose a novel multistep traffic forecasting model, the *Dynamic Graph Convolutional Network* (DGCN). The model architecture and the inner structure of the novel RNN cell are shown in Fig.2.3. It combines a DGC module with a regular graph convolutional GRU-like (GCGRU) cell, where gates in GRU are replaced by different space-domain static graph convolutions. For each time step, the current traffic state is concatenated with the hidden state to form feature maps. The feature maps and the current traffic state are fed into DGC module to give the prediction. Then the prediction and the hidden state are put into GCGRU to produce new hidden state. Mathematical formulas are given in (2.9). Scheduled sampling (Bengio et al., 2015) is an effective curriculum learning strategy to mitigate the discrepancy between training and inference phases. In training, the decoder uses either the ground-truth or the prediction



(a)



(b)

Figure 2.3: Architecture of DGCN model: (a) RNN encoder-decoder; (b) The inner structure of DGCN cell.

given by last time step. The probability of using the ground-truth, ε, is gradually reduced to 0 with iterations during training so the model can start with learning from the correct answers and ends with giving the predictions by itself. In inference and testing, ε is reset as 0.

$$\begin{cases} \mathbf{X}^{t+1} = \mathrm{DGC}_k([\mathbf{X}^t, \mathbf{h}^{t-1}], \mathbf{X}^t, 1; \mathscr{G}) \\ \mathbf{r}^t = \sigma(\mathrm{GC}_{r,k'}([\mathbf{X}^{t+1}, \mathbf{h}^{t-1}], \mathscr{G})) \\ \mathbf{u}^t = \sigma(\mathrm{GC}_{u,k'}([\mathbf{X}^{t+1}, \mathbf{h}^{t-1}], \mathscr{G})) \\ \mathbf{c}^t = \tanh(\mathrm{GC}_{c,k'}([\mathbf{X}^{t+1}, (\mathbf{r}^t \odot \mathbf{h}^{t-1})], \mathscr{G})) \\ \mathbf{h}^t = (1 - \mathbf{u}^t) \odot \mathbf{h}^{t-1} + \mathbf{u}^t \odot \mathbf{c}^t \end{cases} \tag{2.9}$$

To make the values of dynamic kernels more stable and to give clear congestion boundaries, the adjacency matrix masking process (the 3rd equation in (2.4)) is replaced by equation (2.10) to realize *softmax* normalization in each receptive field (Vaswani et al., 2017). So the model is encouraged to explore the most important link in each receptive field and the kernel values are bounded between 0 and 1:

$$\mathbf{W}_{N\times N}(t) = softmax(\mathbf{S}_{N\times N}(t) - 10^{15}(1 - \mathbf{A}_{\mathbf{k}N\times N})) \tag{2.10}$$

In summary, we propose a DGC module to capture directional dynamics in traffic data. Subsequently, we implement these DGC modules in an RNN encoder-decoder to realize multistep traffic forecasting. In the next section, we will benchmark this new model against existing multi-step forecasting models.

2.3 Experiments

In this section, the proposed model is tested on real-world freeway networks. All data used in this study are provided by *National Data Warehouse for Traffic Information* (NDW, The Netherlands). Raw speed data are collected from loop detectors on freeways. The well-known adaptive smoothing method (ASM) (Treiber & Helbing, 2002; van Lint & Hoogendoorn, 2010) is used to fill the missing points in the sensor dataset.

2.3.1 Data description and benchmark models

The freeways around two big cities, Rotterdam and Amsterdam (Netherlands), are selected as research targets. The major clock-wise and counter clock-wise beltways around Rotterdam are respectively named as `RotCL` and `RotCC`. The two beltways are uniformly partitioned into 200 m links. `RotCL` contains 199 links and `RotCC` contains 208 links. The more complex freeway network around Amsterdam which contains multiple intersections is noted as `"AMSnet"`. `AMSnet` covers a larger area so it is uniformly partitioned into 400 m links to reduce complexity. `AMSnet` has 201 links in total. These networks are shown in Fig.2.4. Each link is represented by a node to construct graphs. Speed data of the entire year of 2018 are prepared. All holidays and weekends are removed due to the lack of congestion. 27 highly-congested weeks without severe sensor malfunction are selected. The chosen weeks are shuffled and partitioned into 3 groups: 18 weeks for training, 4 weeks for validation, and 5 weeks for testing. Only afternoon-evening peak hours between 14:00-19:00 are included in the dataset because this time period contains the most diverse and richest patterns, which makes this traffic forecasting task more challenging.

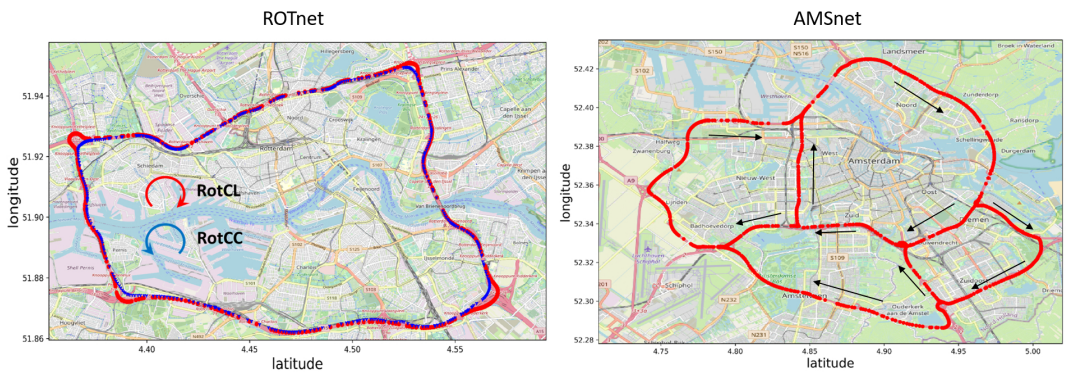


Figure 2.4: The freeway networks around Rotterdam and Amsterdam. The arrows are driving directions.

The traffic forecasting formulation is determined by 4 parameters: time interval Δt, spatial resolution l, observation steps m, and prediction steps p. For the two beltways we fix the average length of each link $l = 200$ m and define two different tasks:

- *Task-1: shorter time interval*: Data is aggregated every 2 min ($\Delta t = 2$ min), observation window is 30 min ($m = 15$) and the maximum prediction horizon is 20 min ($p = 10$).

- *Task-2: longer time interval*: Data is aggregated every 5 min ($\Delta t = 5$ min), observation window is 60 min ($m = 12$) and the maximum prediction horizon is 30 min ($p = 6$).

The four corresponding datasets are noted as `RotCC2`, `RotCC5`, `RotCL2`, and `RotCL5` (network + time interval). For `AMSnet`, we only consider *Task-1*. The DGCN has two hyperparameters: the receptive field of the DGC module k, and the GCGRU module k'. We fix $k' = 2$ and tune k for different predictive tasks and networks by grid-searching. For the two ring freeways, the optimal receptive fields of *Task-1* and *Task-2* are $k = 4$ and $k = 9$, respectively. The optimal receptive field for `AMSnet` is $k = 2$. To evaluate the performance of the DGCN, it is compared with the following benchmark models:

- **Historical average** (HA): HA is a simple model that calculates the average speed of historical data as predictions. If the training set is fixed, HA gives the same prediction every day and the performance doesn't change with the prediction horizon.

- **K-nearest neighbours** (KNN): K-nearest neighbours regression (Denoeux, 1995) calculates the predictions based on the K most similar patterns in the training set. The similarity is measured by their Euclidean distance (RMSE) and predictions are computed through the weighted average of the corresponding future evolution. Weights are normalized by the inverse of Euclidean distance. The optimal hyperparameter $K = 25$ is chosen by grid search and cross-validations.

- **DCNN**: Here we use the depth-4 model proposed in Ma et al. (2017) as a regular deep CNN baseline model. The beltways are cut off at a position and treated as a corridor. The roads in `AMSnet` are stitched together to form an image. Compared to GNN-based models, some adjacent pixels on these images may represent two geographically distant positions.

- **GAT**: The GAT module proposed by Veličković et al. (2018) is implemented in GRU cells. This model also has a recurrent encoder-decoder structure. To guarantee a fair comparison, its receptive field is set the same as the DGCN. The feature enhancement layer in GAT has 5 output units.

- **DCRNN**: This predictor proposed by Li et al. (2018) is one of the state-of-the-art multistep traffic forecasting models. Here we use the same hyperparameters setting as the dual-random walk model in their paper. This is a reasonable choice because the number of nodes and the scale of our road networks is comparable to theirs. Like the DGCN, this model also has a recurrent encoder-decoder structure.

- **STGCN**: This model proposed in Yu et al. (2018) is another state-of-the-art single-step traffic forecasting model combining spatial graph convolution and gated temporal convolution. When giving multistep predictions, the observation window moves with the updated new prediction step by step. Here we use the variant that shows the best predictive accuracy in Yu et al. (2018), called STGCN(Cheb).

- **Graph Wavenet**: This model proposed by Wu et al. (2019) combines the diffusion convolution in DCRNN and the fully-convolutional structure of STGCN and generates multistep predictions. An additional learnable self-adaptive adjacency matrix is used to capture dependencies among links. But this self-adaptive adjacency matrix is static after training.

When training the models, all speed data is firstly normalized between 0 and 1 by dividing by the speed limit $120\,\mathrm{km\,h^{-1}}$, then normalized by the z-score function. For testing, normalized data are used as input, but the output is fed into a reverse process to give true-value predictions. The errors are calculated by taking several distance measures between predictions and the corresponding ground-truth data. We choose three error measures: MAE, MAPE and RMSE. The mean average error (MAE) measures the overall accuracy; the mean average percentage error (MAPE) is particularly sensitive to errors in the contours of low-speed areas (i.e. whether the models accurately track the spatio-temporal boundaries of congestion patterns), and the root mean square error (RMSE) provides a combined measure (bias + variance) for the uncertainty in the predictions. The three metrics are used to evaluate the performance from different aspects. In training, RMSE is chosen as the loss function because of the z-score normalization. The *Adam* optimizer (Kingma & Ba, 2017) is chosen to minimize the loss function. The initial learning rate, decay rate, and scheduled sampling parameters are tuned for each model. Early stopping on the validation set is used to mitigate overfitting. Our experimental platform has one GPU (NVIDIA GeForce GTX 1070, 16GB). All deep learning models are trained and tested through parallel computation on GPU. The source code, datasets, hyperparameters setting, visualization and model interpretation tools are available online: [1].

2.3.2 Results

Table 2.1 and Table 2.2 list the overall average predictive errors of the proposed model and benchmark models on different datasets. Because the sizes of the two beltways are close, we do not distinguish their training time but give the average value. Fig.2.5 further shows the relation between each single-step errors and the prediction horizon on `RotCC` and `AMSnet`. They present how errors accumulate with prediction steps. The tendency is the same for another freeway `RotCL`. Because DNN models show significantly higher predictive accuracy than HA and KNN on `AMSnet`, we only compare deep learning models in Fig.2.5. We observe the following phenomena from the results:

- DGCN or STGCN(Cheb) achieves the best overall multistep predictive accuracy in terms of the three metrics in most forecasting tasks. They show comparable performances on the two beltways and `AMSnet`. Especially, DGCN outperforms all the other models in terms of RMSE, which is the optimization target in training. It means that DGCN is better at tackling rare congestion patterns with higher uncertainties.

[1] `https://github.com/RomainLITUD/DGCN_traffic_forecasting`

Table 2.1: Multistep predictive performances comparison on two ring freeways during peak hours 14:00 - 19:00

Model	MAE(km h^{-1})		MAPE(%)		RMSE(km h^{-1})		Time(s)
freeway	`RotCC`	`RotCL`	`RotCC`	`RotCL`	`RotCC`	`RotCL`	Average
shorter time interval: $\Delta t = 2$ min; $p = 10$							
HA	9.83	9.81	12.92	13.74	16.39	16.02	72
KNN	6.95	6.29	17.66	14.90	12.94	11.38	240
DCNN	5.84	5.18	12.07	10.94	10.42	9.04	600
GAT	6.04	5.39	13.44	11.73	10.98	9.34	1100
DCRNN	5.41	5.28	12.53	10.84	9.93	9.00	350
STGCN(Cheb)	5.16	**4.56**	11.28	**9.23**	10.69	8.90	750
Graph Wavenet	5.35	4.91	12.02	10.93	9.92	8.73	900
DGCN	**5.07**	4.68	**11.23**	9.30	**9.78**	**8.53**	650
longer time interval: $\Delta t = 5$ min; $p = 6$							
HA	9.60	9.61	**12.62**	13.39	16.26	15.85	62
KNN	7.51	6.84	18.32	15.49	13.78	12.09	27
DCNN	7.25	6.28	15.07	12.62	12.63	10.74	280
GAT	7.59	5.76	13.96	11.50	14.31	10.15	500
DCRNN	6.26	5.70	14.44	12.89	11.40	10.10	250
STGCN(Cheb)	**5.66**	5.34	12.79	10.97	11.40	10.03	400
Graph Wavenet	6.46	5.62	14.61	10.83	11.79	9.63	500
DGCN	5.90	**5.15**	12.70	**10.31**	**11.12**	**9.57**	350

Table 2.2: Multistep predictive performances on `AMSnet`: $\Delta t = 2$ min, $m = 15$, $p = 10$

Model	MAE(km h^{-1})	MAPE(%)	RMSE(km h^{-1})	Time(s)
HA	5.98	17.74	11.61	65
KNN	4.65	14.42	11.57	330
DCNN	4.14	11.02	8.52	610
GAT	3.86	10.15	8.17	1500
DCRNN	3.82	9.36	8.05	550
STGCN(Cheb)	3.76	9.48	8.03	820
Graph Wavenet	**3.67**	9.31	8.11	1040
DGCN	**3.67**	**8.98**	**7.83**	680

- In Fig.2.5, for the beltway `RotCC`, the slope of the KNN curve is the smallest in terms of the three accuracy metrics. The predictive errors of STGCN(Cheb)

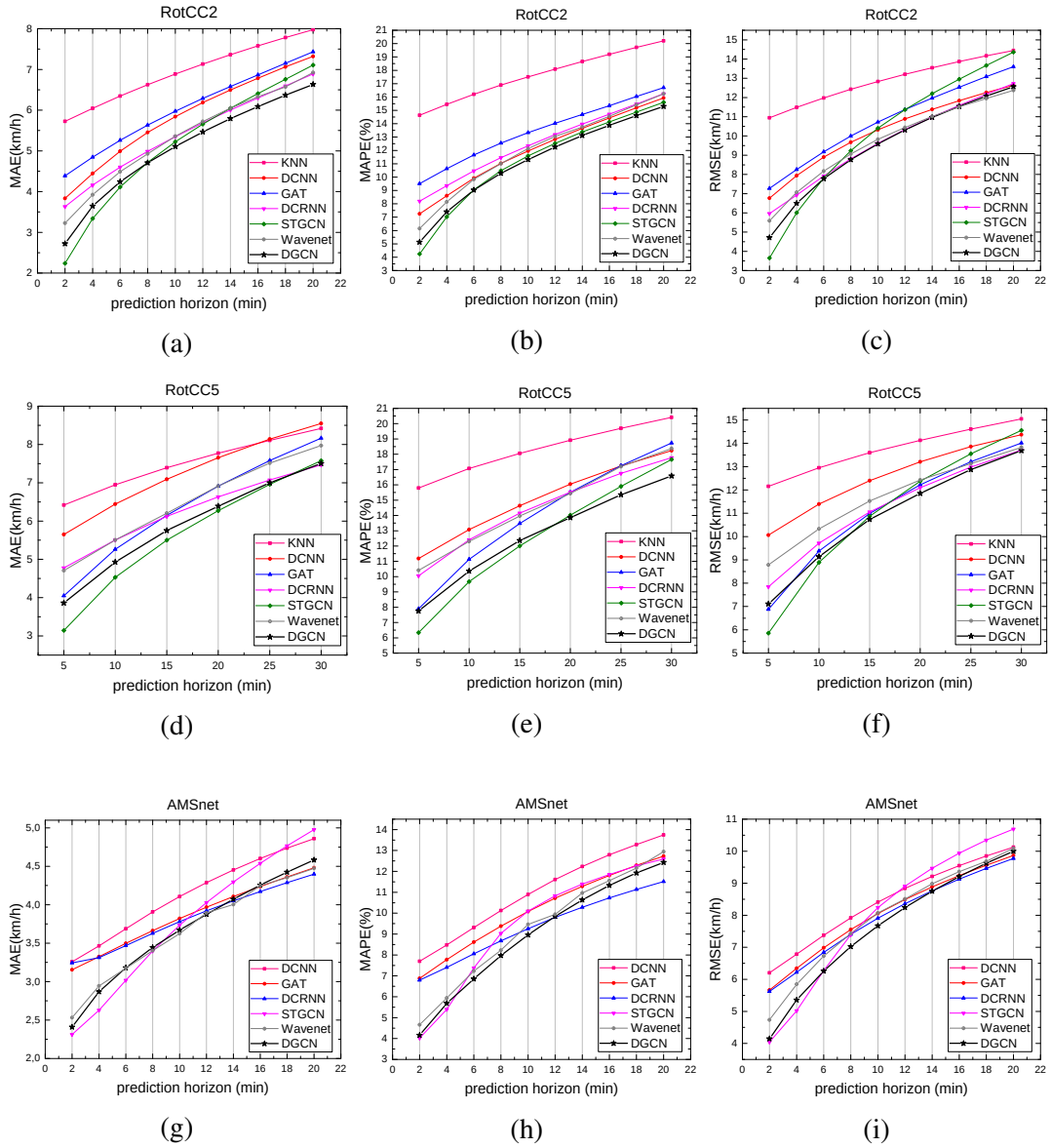


Figure 2.5: The relations between MAE/MAPE/RMSE and the prediction horizon for each single time step.

increase the fastest with prediction steps, especially for RMSE. The other models are between them. The reason is that STGCN(Cheb) is a one-step prediction model employing a fully-convolutional structure. But all the other GNN-based deep learning models need to balance the multistep predictive errors and the prediction horizon. DCNN also has a regular fully-convolutional structure, but all predictive steps are generated together as an image so the multistep errors are also balanced. Although the overall average errors of DGCN and STGCN(Cheb) are close, STGCN(Cheb) performs better for shorter-horizon prediction while DGCN is more accurate for longer-horizon prediction.

- Having similar GRU-like structures containing graph attention modules, DGCN consumes less running time than GAT but consistently gives better predictions.

The computational efficiency originates from the parameters sharing in the DGC module: the calculation of pairwise similarities is replaced by collective matrix operations.

- Generally speaking, deep learning models give better predictions than HA and KNN. This advantage is more significant on `AMSnet` than on the two ring freeways because the congestion patterns on a net-like graph are more complex due to spatial correlations among roads. DNN models consume longer training time than HA and KNN. However, the running time can be further reduced by using more powerful GPU clusters.

To summarize, we showed that DGCN is able to give satisfactory predictions on real-world datasets. The results support the theoretical discussion in Section 2.2: DGC performs better than GAT when tackling complex directional patterns. Next, we will focus on the most important advantage of the proposed model: DGCN has high interpretability.

2.4 Model interpretation

Now we explain what DGCN learns from data by looking inside of DGC module. Section 2.4.1 relates the optimal receptive field to the back-propagation speed of stop-and-go waves and shows again that the upstream-downstream asymmetry of models has a significant influence on traffic forecasting performance. Section 2.4.2 directly gives the relation between dynamic graph convolutional kernels and speed data. It shows how DGC “understands" traffic dynamics in different situations.

2.4.1 Optimal receptive field

To explore how predictive errors change with the receptive field, we formulate a series of forecasting tasks: the observation step $m = 10$ and prediction step $p = 3$ are fixed. The time interval changes from 1 min to 5 min by 1 min. For each of the five tasks above, the receptive field k gradually increases from 1 to 12. Thus, we can obtain five MAE - k curves. In Fig.2.6, the five curves are overlapped together. Notice that the x-axis is the real physical receptive field, $k \times l$. The left figure shows that the five curves have a similar tendency: with the increasing receptive field, MAE drops rapidly until it reaches an optimal point, then increases slowly. The longer the time interval is, the larger the optimal receptive field is. The right figure shows that the relation between Δt and k_{opt} can be approximated by a linear equation:

$$k_{opt} \times l \approx 19.8\,\text{km}\,\text{h}^{-1}\Delta t \tag{2.11}$$

To explain this relation and visualize the effect of the receptive field directly, we set different k values, re-train the DGCN models on `RotCC2`, and compare their predictions with the ground-truth. The results are shown in Fig.2.7 ($\Delta t = 2\,\text{min}$). For a small receptive field $k = 2$, the model tends to predict the backpropagation of stop-and-go waves

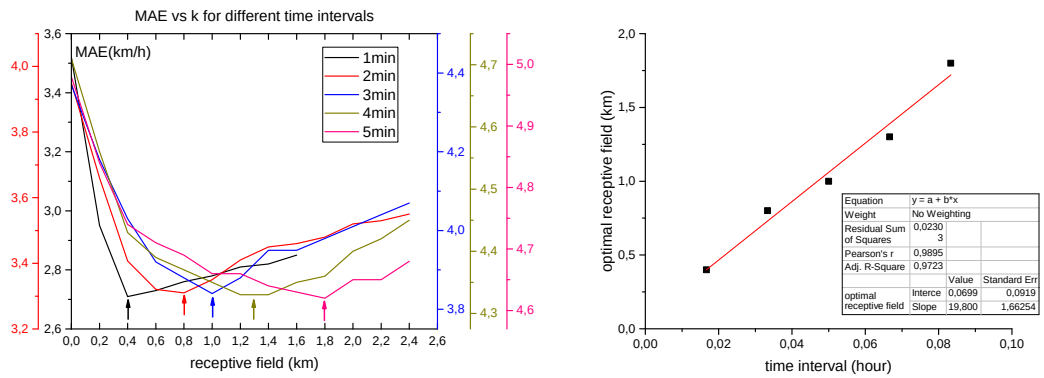


Figure 2.6: Left: MAE-receptive field relations for different time intervals on `RotCC`; Right: the linear relation between the optimal receptive field and the time interval.

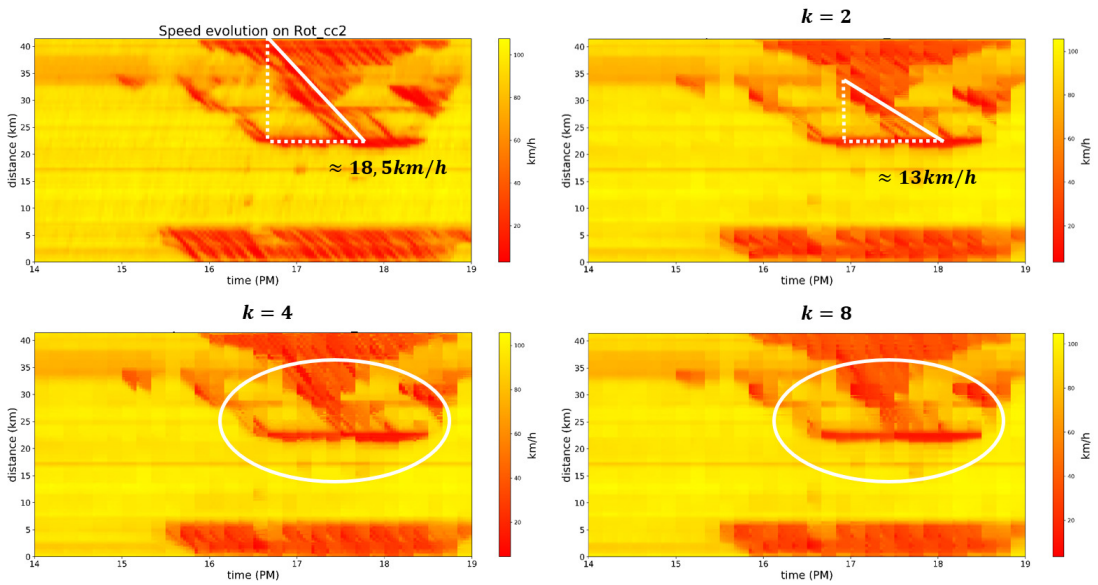


Figure 2.7: Ground-truth (top left) and 10 min predictions given by DGCN with different receptive field k during peak hours on `RotCC2`: an example on 15-01-2018.

but the constrained receptive field results in mismatching patterns. The speed of the stop-and-go wave measured on this prediction is about $13\,\mathrm{km\,h^{-1}}$, which is lower than the true value. As for $k = 4$, the receptive field is large enough so the model can give precise predictions with clear stop-and-go wave boundaries. Even the merging of congestion (see the area in the circle) is successfully predicted because DGC considers the speciality of each location. DGC learns from historical data that congestion frequently stops spreading at that position because there is an off-ramp, even if this off-ramp is not included in the dataset. When k is too large, for example, $k = 8$, each receptive field contains multiple stop-and-go waves so the model is confused about which wave will cause the incoming congestion. So the congestion boundary becomes blurred. Fig.2.7 clearly illustrates how DGC mimics the directional flow of information. To predict the traffic state at a position after Δt, only the most adjacent stop-and-go wave should be included in the receptive field. Thus, we can estimate the maximal speed of

stop-and-go waves by an empirical inequality:

$$k_{\text{opt}} \geq |v_s|_{\max} \frac{\Delta t}{l} \tag{2.12}$$

Comparing (2.12) with (2.11), we can estimate the upper bound of the speed of stop-and-go waves on `RotCC`: about $19.8\,\text{km}\,\text{h}^{-1}$. In Fig.2.7, this estimated average speed is about $18\,\text{km}\,\text{h}^{-1}$. Considering the receptive field is discretized by links' length $l = 0.2\,\text{km}$ in our model, $19.8\,\text{km}\,\text{h}^{-1}$ is a reasonable upper bound. This analysis explains why we choose $k = 4, k = 9$ for the two ring freeways and $k = 2$ for `AMSnet` in our experiments. It also gives a hyperparameter tuning principle in a traffic forecasting model. For a road network, if the time interval is small enough to show the fine structure of stop-and-go waves, the shorter links are, or the higher v_s is, the bigger receptive field should be chosen.

2.4.2 Dynamic spatial correlations

In many deep learning based traffic forecasting models, temporal dependencies and spatial correlations are modelled separately by different modules (e.g. Cui et al. (2019); Zhang et al. (2019b); Yu et al. (2020)). But in DGCN, the spatial correlations depend on historical and current traffic states. For simplification, we take the first decoder cell as an example. The encoder encrypts observed traffic conditions in the past $m-1$ time steps into the context vector $\mathbf{C}$, and $\mathbf{C}$ is concatenated with the current input $\mathbf{X}^t$. Thus, each node is associated with two features, (c_i, x_i). They represent the historical information and the current traffic condition respectively. Because of the *softmax* normalization in DGC, all weights are non-negative and the sum of weights in each receptive field equals to 1. So we can interpret generated dynamic graph kernels $\mathbf{W}(t)$ as relative spatial correlations. For example, $W_{i,j}(t)$ represents the "influence" of node-j on node-i for the prediction of next time step $t+1$ (see Fig.2.1). We call it an "influence coefficient". To help visualise multi-dimensional dynamic filters in a single figure, we define a directional distance vector on the ring freeways:

$$\vec{d}_k = (-k, -k+1, -k+2, \ldots, -1, 0, 1, \ldots, k-2, k-1, k) \tag{2.13}$$

For node i, the spatial correlation vector is defined by:

$$\vec{J}_i^t = (W_{i,i-k}^t, W_{i,i-k+1}^t, W_{i,i-k+2}^t, \ldots, W_{i,i+k-2}^t, W_{i,i+k-1}^t, W_{i,i+k}^t) \tag{2.14}$$

The filter value is defined by their inner product divided by k:

$$f_i^t = \frac{1}{k} \langle \vec{d}_k, \vec{J}_i^t \rangle, \quad f_i^t \in [-1, 1] \tag{2.15}$$

For more complex graphs, the definition is exactly the same, but the numbers of downstream/upstream adjacent nodes (the length of $\vec{d}_k$ and $\vec{J}_i^t$) depend on locations. Following the concept of spatial attention, f_i^t is called an *attention coefficient*. It is a variable of space and time. More negative f_i^t means that upstream links are more important, while more positive f_i^t indicates that traffic states of downstream links dominate the

prediction. In the following parts, spatial correlation vectors $\vec{J}_i^t$ and attention coefficients f_i^t extracted from decoder cells are compared with real-time predictions v_i^{t+1} to seek model interpretations.

Fig.2.8 shows the relations between the average attention coefficient $\bar{f}$ and the predicted speed v for different datasets. $\bar{f}$ is calculated over all links with the same speed range, which is uniformly aggregated every $5\,\mathrm{km\,h^{-1}}$, from $0\,\mathrm{km\,h^{-1}}$ to $120\,\mathrm{km\,h^{-1}}$. In Fig.2.9, each column of the heat map represents the average spatial correlation vector $\vec{J}$ in the corresponding speed range. Here we only show `RotCC` as two clarifying examples in Fig.2.9 because the number of adjacent links is not the same for different nodes on `AMSnet`. Fig.2.10 further directly compares the ground-truth of speed and the corresponding evolution of dynamic attention coefficient on `RotCC2` during the peak hour of a randomly selected workday. In Fig.2.11, we select two representative scenarios on `AMSnet`: an on-ramp and an off-ramp. The speed is compared with the real-time spatial attention distribution to precisely show how DGC gives predictions under these more complex situations. The following conclusions can be drawn from the results:

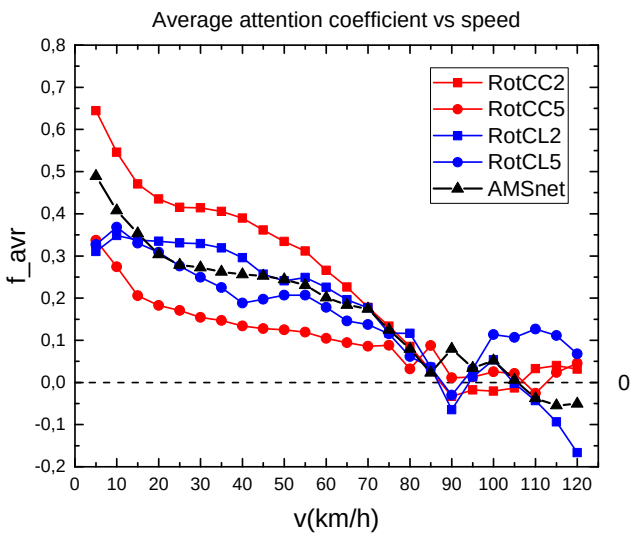


Figure 2.8: $\bar{f}$-v relations on different datasets

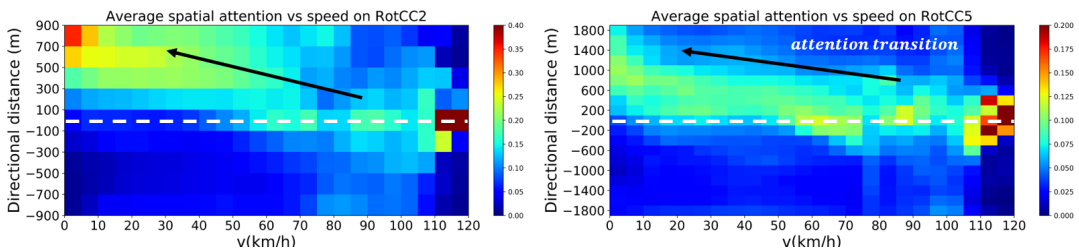


Figure 2.9: $\vec{J}$-v relations on `RotCC`: each column is the average $\vec{J}$ in that speed range. y-axis is the directional distance.

1. Given that congestion is defined by a speed value lower than $70\,\mathrm{km\,h^{-1}}$, then Fig.2.8 shows that the average attention coefficient $\bar{f}$ gradually shifts to a higher

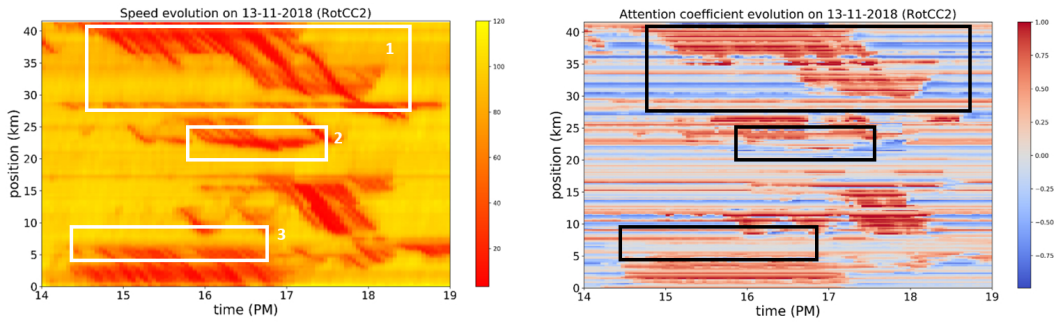


Figure 2.10: Comparison between the speed ground-truth and the dynamic attention coefficient.

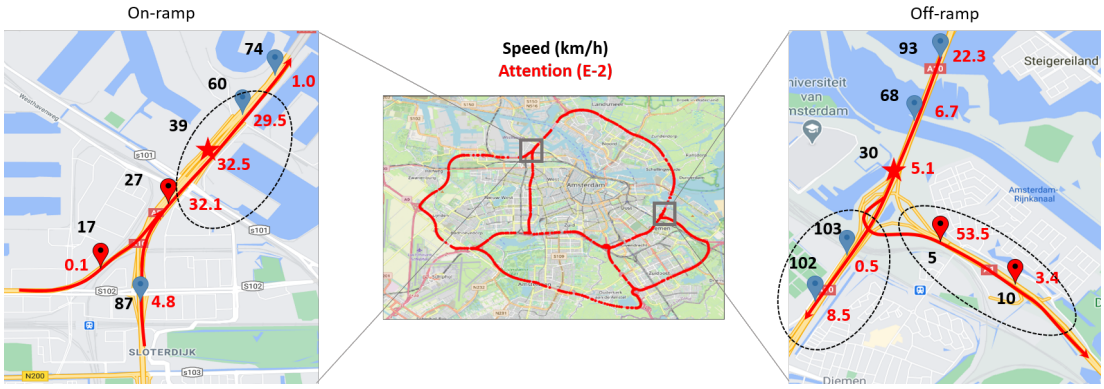


Figure 2.11: Comparison between the speed ground-truth and the dynamic attention on `AMSnet` ($k = 2$): The left figure is an on-ramp and the right one is an off-ramp. The arrows represent driving directions (downstream). Stars represent the central nodes in the receptive field. Black numbers are speed, red numbers are dynamic attention values. Red nodes and blue nodes respectively represent congested and free-flowing traffic states.

positive value with the decreasing of v. It means that DGC tends to credit higher spatial importance to downstream links in low-speed regions. We call this phenomenon *attention transition*.

2. In Fig.2.9, the heat map shows that in free-flowing areas the average spatial attention distributes more or less evenly around central links, while for congested areas spatial attention gradually converges to farther downstream links with the decreasing of speed (the top-left red spot). We call this phenomenon *attention convergence*.

3. Attention transition and convergence statistically explain how DGC "understands" traffic dynamics. In case all adjacent links within the receptive field of a specific link i are freely flowing, the model infers that link i will remain in a free-flow state in the next time step. But when congestion occurs, the model switches attention to downstream links to assimilate information that describes backwards propagating jam waves and other congestion dynamics. The more speed decreases, the farther away downstream the attention is shifted. This mechanism is

consistent with basic traffic flow theory and in line with the analysis in section 2.4.1.

4. Fig.2.10 presents how DGCN learns the unique property of each location. In general, as explained above, the dynamic graph convolutional kernels pay more attention to downstream links in low-speed congested areas, like area-1. But at some special positions, like in area-2 and 3, even if the speed is very low, the attention still keeps neutral (close to 0) to stop the back propagation of stop-and-go waves and to resist the change of traffic states. These positions are important on-ramps causing bottlenecks or off-ramps stopping the shock waves from spill-back. DGCN learns these infrastructure differences solely from incomplete data. DFN and GAT cannot do this.

5. Fig.2.11 shows similar but more complicated phenomena. For the on-ramp on the left, the left branch from the west is congested but another branch from the south and the downstream links are free-flowing. The DGC deducts that there will be a standing bottleneck due to the merging of two traffic streams. So the spatial attention focuses on the central node (neutral). For the off-ramp shown in the right figure, the downstream branch to the southeast is congested but another branch and upstream links are not congested. The DGC automatically increases its attention to the congested downstream branch but credits very low attention to the other one. It predicts that the jam wave will spill back and continue spreading. The analysis shows that DGC is able to treat more complex graphs and give explainable real-time spatial correlations. It allows for delicately studying the role of intersections in a network.

2.4.3 Discussion on model interpretability

So far we have explained what dynamic spatial correlations the DGC has learned from traffic data and the resulting statistical relationships between trainable parameters and speed. We emphasize that the attention convergence and transition phenomena highlighted in the results section are *emerging* results of the learning process. In this final subsection we return to how this relates to traffic dynamics, which can be compactly described by the LWR kinematic wave model (Lighthill & Whitham, 1955a):

$$\begin{aligned} \frac{\partial \rho(x,t)}{\partial t} + \nabla_x q(x,t) &= s(x,t) \\ \Rightarrow \frac{\partial \rho(x,t)}{\partial t} + c(\rho)\nabla_x \rho(x,t) &= s(x,t) \end{aligned} \tag{2.16}$$

in which ρ is density, q is flow, $c(\rho) = \partial q/\partial \rho$ is the kinematic wave speed, and s is the source term governing in- and outflows at on-ramps and off-ramps. Once traffic density increases beyond some critical density ρ_c, the wave speed $c(\rho) < 0$, which means kinematic waves move against the direction of flow. In flows with densities below ρ_c, disturbances move *with* the flow ($c(\rho) \geq 0$). Recall that this *directional* dynamics is used to justify the location-specific design of the DGC module. It turns out this choice really pays off. Using the spatial attention mechanism, the DGC module indeed seems to have encoded traffic state-dependent wave speeds which determine

which up- and downstream nodes offer the most relevant information for the prediction. In freely-flowing conditions, it uses a mix of up- and downstream information; in congestion, it assimilates more information from downstream. This dynamic state-dependent behaviour cannot be achieved by the benchmark models.

Another tentative possibility is to consider the DGC model as an alternative to real-time macroscopic simulation models (e.g. Wang et al. (2005); Van Hinsbergen et al. (2011)), which combine continuous traffic flow models as described by (2.16) with sequential Bayesian estimators (Kalman filters, particle filters, etc). Li et al. (2018) explains that diffusion graph convolution is equivalent to the discrete form of a Laplacian operator (a second-order differential) on graphs. Similarly, extended spatial-domain graph convolution can also be treated as the discrete form of an unknown nonlinear spatial operator with a source term parameterized inside:

$$\frac{\partial \rho}{\partial t} + \mathscr{S}_x(\rho, s) = 0 \tag{2.17}$$

To approximate this equation and the implicit mapping in $\mathscr{S}_x$ by neural networks, some basic rules need to be fulfilled. For one, the kinematic wave model in (2.16) is a conservation equation. Nothing is conserved in the relation (2.17) learnt by DGCN, unless one would explicitly add such conservation rules. A related issue is that the maximum speed of jam waves should constrain the adaptive receptive field. In many practical cases, however, flow data are not available, whereas speeds are increasingly available, both through infrastructure-based sensing and through probe vehicle data. The dynamics of average speed can be described as the superposition of a convective process and other higher-order effects due to location-specific reasons. Our results suggest that these dynamics can be more or less learnt by DGCN just from data.

Finally, we point out that the results shown in Fig.2.6 and Fig.2.7 are relevant to the "causal confusion" in correlation-based deep learning models. Without restrictions on the direction of information flowing, most DNNs cannot distinguish correlation and causation. Learning false causation or spurious correlation, such as which stop-and-go wave will cause the congestion here, can damage the robustness of the model. The optimal receptive field also suggests that deeper is not necessarily better in traffic forecasting. Domain knowledge is always important for traffic dynamics modelling.

2.5 Conclusion and outlook

In this paper, we demonstrated a multistep short-term traffic forecasting model using dynamic graph convolution. The key feature of the proposed model is that it combines the road graph embedded in their structures and the RNN component which allows learning explainable traffic dynamics.

Experiments on freeway networks show that the proposed model, DGCN, can give satisfactory short-term predictions. The core innovation, the dynamic graph convolution (DGC) module, is a “grey-box” that allows us to unravel what heuristic dynamic spatial correlations the model has learnt. Two points are demonstrated. First, designing a suitable directional features extraction module and choosing a so-called adaptive receptive

field is critically important for predicting the fine structures of congestion patterns. Second, we observed two phenomena, graph attention transition and convergence, that show how the DGC learns dynamic spatial correlations that make sense from a traffic flow perspective. The model has learnt speed-and-location-dependent, kinetic wave speeds so as to assimilate the right data under the right conditions. In practice, this model can give both network-level traffic state prediction and dynamic spatial dependencies among links. The predicted states can be used for prediction-based services, such as estimated time of arrival and congestion forewarning. In traffic control, spatial correlations can help practitioners to diagnose where the predicted congestion spreads from, which is important for transparent decision-making.

This study leads to several potential research directions. For the deep learning technique itself, arguably the dynamic graph convolution module is not deep enough. The mechanism can be easily extended to multi-head attention, or multi-scale DGC modules with different receptive fields to capture hierarchical information flow. This may improve both predictive performance and interpretability. For example, this study focuses on continuous traffic streams on freeway networks. Urban road networks may have more bottlenecks because of traffic lights so the traffic flows are less continuous. Spatial correlations in these scenarios need further studies.

Second, the DGC in this paper is myopic by design. It considers traffic dynamics from a localized and short-term view. Some other factors resulting in emerging congestion patterns or long-term variations, such as sudden demand peaks, traffic accidents, and weather, are not considered. Whereas our model can predict the dynamics of congestion once it has started, it is not suitable for predicting the onset of congestion, which is highly uncertain. This problem restricts the prediction horizon to 30 min or less during busy peak hours. For multi-step forecasting, it is very sensitive to previous predictions and unexpected new bottleneck activation quickly increases prediction errors.

A third interesting path of investigation is to add flows to the inputs. We hypothesize that in that case a DGC-based model may be designed such that it is explainable as a real-time macroscopic traffic flow model with adaptable state variables and parameters. The big difference with classic simulation-based traffic data assimilation methods is that DGC-based approaches may be better able to learn the dynamics from incomplete data than say an LWR-like model in combination with a sequential Monte Carlo method. On the other hand, we also may be overly optimistic about the level of potential interpretability.

Finally, and this combines all previous points, maybe it is possible to redesign these dynamic graph convolution models such that they are able to learn the dynamics of multiple demand and supply processes, such as the dynamics of the traffic state itself, the dynamics of key parameters like capacity, and the dynamics of demand. We believe it is better to develop methods that use the best of both worlds of data-driven and simulation-based models. These results can help design similar frameworks and develop more explainable deep learning based traffic forecasting models for decision-making support.

Acknowledgements

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

Chapter 3

The average predictability of macroscopic traffic speed

In the previous chapter, we obtained an empirical relationship that describes which links may causally influence the future traffic state at a location. This chapter aims to directly estimate the average predictability of highway traffic state from data. The empirical rules are used to partition the dataset and reduce the computational complexity of the proposed entropy-based approach. Theoretically, we proved that conditional entropy gives the lower bound of negative-log-likelihood for probabilistic forecasting models. Then the k-p-nearest algorithm is used to estimate the predictability of speed as a function of time and locations. The results illustrate that the average predictability is significantly lower at some special locations and during evening peak hours.

This chapter is published as a journal article: Li, G., V. L. Knoop, H. van Lint (2022a) Estimate the limit of predictability in short-term traffic forecasting: An entropy-based approach, *Transportation Research Part C: Emerging Technologies*, 138, p. 103607. The content is the same as the journal paper.

ABSTRACT

In the past several decades, many models have been proposed to continuously improve 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 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 the time of day, network locations, observation range, and prediction horizon. Results reveal that the stochastic nature of traffic dynamics and improper assumptions on the prior distribution of output are two major factors restricting predictive performance. In summary, the proposed method estimates a trustworthy performance boundary for most traffic forecasting models.

3.1 Introduction

Short-term traffic forecasting is critically important for many key applications in the 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); 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 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 & 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 travelling behaviours, like lane-changing choices, are highly random and they could have a 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 two categories, that is, deterministic or probabilistic. Deterministic models aim to build 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 contrast, 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 the input-output formulation. Taking the example of a deterministic model, one may use explicit traffic modelling (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 the 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 data set, 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 cause 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., 2012a; 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 3.2 presents the background knowledge and related works in literature. Section 3.3 describes the proposed method, including its theoretical basis, implementation of spatio-temporal dependencies, and the numerical scheme to estimate the limit. Section 3.4 shows the results and gives an analysis of numerical experiments. Section 3.5 finally concludes and proposes several related research directions.

3.2 Background

3.2.1 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 $\mathscr{V}$ is proposed and defined as follows:

$$H(V) = -\int_{\mathscr{V}} p_V(v) \ln p_V(v) dv \tag{3.1}$$

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

$$H(X|Y) = H(X,Y) - H(Y) = -\int_{\mathscr{X},\mathscr{Y}} p_{X,Y}(x,y) \ln p_{X|Y}(x|y)\, dxdy \tag{3.2}$$

where $\mathscr{X},\mathscr{Y}$ 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:

$$p(\boldsymbol{X}_{t_1},...,\boldsymbol{X}_{t_n}) = p(\boldsymbol{X}_{t_1+\tau},...,\boldsymbol{X}_{t_n+\tau}),\ \ \forall \tau, t_1,...,t_n \in \mathbb{R},\ \ \forall n \in \mathbb{N} \tag{3.3}$$

It means that statistical properties do not change with time. To facilitate the narrative, from now on we denote $\boldsymbol{X}_{t-m:t} = \{\boldsymbol{X}_{t-m},...,\boldsymbol{X}_{t-1}\}$ as the past m step observations from t; $\boldsymbol{X}_{t:t+p} = \{\boldsymbol{X}_t,...,\boldsymbol{X}_{t+p-1}\}$ 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*:

$$S(\boldsymbol{X}_t) = \lim_{m\to\infty} H(\boldsymbol{X}_t|\boldsymbol{X}_{t-m:t}) \tag{3.4}$$

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.

3.2.2 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 the 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 disturbing 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 analyse 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 the 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 radial-basis-function (RBF) neural networks to give predictions. However, the 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 for 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 the literature consider traffic time series as a Gaussian process (Idé & 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 the desired low value.

The third solution is the 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 the 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 the 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. (2015a) investigates the UBP of traffic speed on a ring freeway. Each sensor on the network is assumed to be independent of 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 the Lempel-Ziv coding algorithm, or other entropy estimators such as the non-linear embedding estima-

tors, 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 neighbours. We emphasize that time index and spatial correlations must be included in the estimation of the 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.

3.3 Methodology

This section presents details of the proposed entropy-based approach. First, we give a 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 neighbours (kpN) estimator.

3.3.1 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:

$$p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}) = p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y})p_{\boldsymbol{X}}(\boldsymbol{x}) \tag{3.5}$$

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 the 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 the Lyapunov exponent and PSR in subsection 3.2, 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 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:

$$\mathbb{E}_{p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})}[NLL] \geq H(\boldsymbol{Y}|\boldsymbol{X}) \tag{3.6}$$

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:

$$-\mathbb{E}_{\boldsymbol{Y}\sim p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y})}[\ln q_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{Y})] = -\mathbb{E}_{\boldsymbol{Y}\sim p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y})}[\ln p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{Y})] + \mathbb{E}_{\boldsymbol{Y}\sim p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y})}[\ln \frac{q_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{Y})}{p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{Y})}] \tag{3.7}$$

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:

$$\begin{aligned}
\mathbb{E}_{p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})}[\mathrm{NLL}] &= -\int_{\mathscr{X}} p_{\boldsymbol{X}}(\boldsymbol{x})[\int_{\mathscr{Y}} p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}) \ln p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}) d\boldsymbol{y}] d\boldsymbol{x} + \mathbb{E}_{p_{\boldsymbol{X}}(\boldsymbol{x})}[D_{KL}(q \mid p)] \\
&= -\int_{\mathscr{X},\mathscr{Y}} [p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}) p_{\boldsymbol{X}}(\boldsymbol{x})] \ln p_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}) d\boldsymbol{y} d\boldsymbol{x} + \mathbb{E}_{p_{\boldsymbol{X}}(\boldsymbol{x})}[D_{KL}(q \mid p)] \\
&= H(\boldsymbol{Y}|\boldsymbol{X}) + \mathbb{E}_{p_{\boldsymbol{X}}(\boldsymbol{x})}[D_{KL}(q \mid p)]
\end{aligned} \tag{3.8}$$

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 modelled, 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:*

$$\det \mathbb{E}_{p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})}[(\boldsymbol{Y}-\hat{\boldsymbol{Y}})(\boldsymbol{Y}-\hat{\boldsymbol{Y}})^T] \geq \frac{1}{(2\pi e)^n} e^{2H(\boldsymbol{Y}|\boldsymbol{X})} \tag{3.9}$$

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):

$$\frac{1}{2}\ln\det(2\pi e\boldsymbol{K}) \geq H(\boldsymbol{e}|\boldsymbol{X}=\boldsymbol{x}), \quad H(\boldsymbol{e}|\boldsymbol{X}=\boldsymbol{x}) = H(\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}) \tag{3.10}$$

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

$$\frac{1}{2}\int_{\mathscr{X}} p_{\boldsymbol{X}}(\boldsymbol{x})\ln\det(2\pi e\boldsymbol{K})d\boldsymbol{x} \geq H(\boldsymbol{Y}|\boldsymbol{X}) \tag{3.11}$$

As shown in (3.8), the right side is conditional entropy. Because $\ln \circ \det$ is concave, Jensen's inequality gives:

$$\frac{1}{2}\ln[(2\pi e)^n \det(\int_{\mathscr{X}} p_{\boldsymbol{X}}(\boldsymbol{x})\boldsymbol{K}d\boldsymbol{x})] \geq \frac{1}{2}\int_{\mathscr{X}} p_{\boldsymbol{X}}(\boldsymbol{x})\ln\det(2\pi e\boldsymbol{K})d\boldsymbol{x} \tag{3.12}$$

where equality holds if and only if $\boldsymbol{K}$ is independent from $\boldsymbol{X}$. The integral on the left is actually the expectation of the determinant of the covariance matrix (DCM). By combining (3.11) and (3.12), 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 $\boldsymbol{Y} = f(\boldsymbol{X}) + \varepsilon, \;\; \varepsilon \sim \mathscr{N}(0,\boldsymbol{K})$. $\hat{\boldsymbol{Y}} = f(\boldsymbol{X})$ theoretically can be precisely modelled by an unbiased estimator and $\mathscr{N}(0,\boldsymbol{K})$ 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 on any model's capability. The room of improvement for modelling 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 obtaining a series of variance limits, $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$. Their relationship is given by the following formula:

$$\det\boldsymbol{K} = \det\boldsymbol{S}\prod_{i=1}^{n}\sigma_i^2 \leq \prod_{i=1}^{n}\sigma_i^2 \tag{3.13}$$

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 the loss function, the corresponding lower bound is:

$$\mathbb{E}[MSE] \geq \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2 \tag{3.14}$$

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

$$I(y_1, y_2, ..., y_n \mid \boldsymbol{X}) = \sum_{i=1}^{n} H(y_i|\boldsymbol{X}) - H(\boldsymbol{Y}|\boldsymbol{X}) \tag{3.15}$$

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

3.3.2 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 the 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 conditional entropy at t. This smoothing window increases estimation accuracy by including more samples and it also smooths the resulting curve. For example, we prepare the following input-output set to estimate $H(\boldsymbol{X}_{t:t+p} \mid \boldsymbol{X}_{t-m:t})$:

$$\{(\boldsymbol{X}_{d,\tau-m:\tau}, \boldsymbol{X}_{d,\tau:\tau+p}) \mid d \in D \text{ and } \tau \in [t-\delta, t+\delta)\} \tag{3.16}$$

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 of $\boldsymbol{X}^i_{t:t+p}$ 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:

$$\{(\boldsymbol{X}^{i,\text{input}}_{d,\tau-m:\tau}, \boldsymbol{X}^{i}_{d,\tau:\tau+p}) \mid d \in D \text{ and } \tau \in [t-\delta, t+\delta)\} \tag{3.17}$$

$$\boldsymbol{X}^{i,\text{input}}_{d,\tau-m:\tau} = \{X^{j}_{d,s} \mid |r(j,i)| \leq c_r(\tau+p-s)\Delta t \text{ and } s \in [\tau-m, \tau)\} \tag{3.18}$$

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 the time interval. $\boldsymbol{X}^{i,\text{input}}_{d,\tau-m:\tau}$ is a collection of all points in the spreading cone.

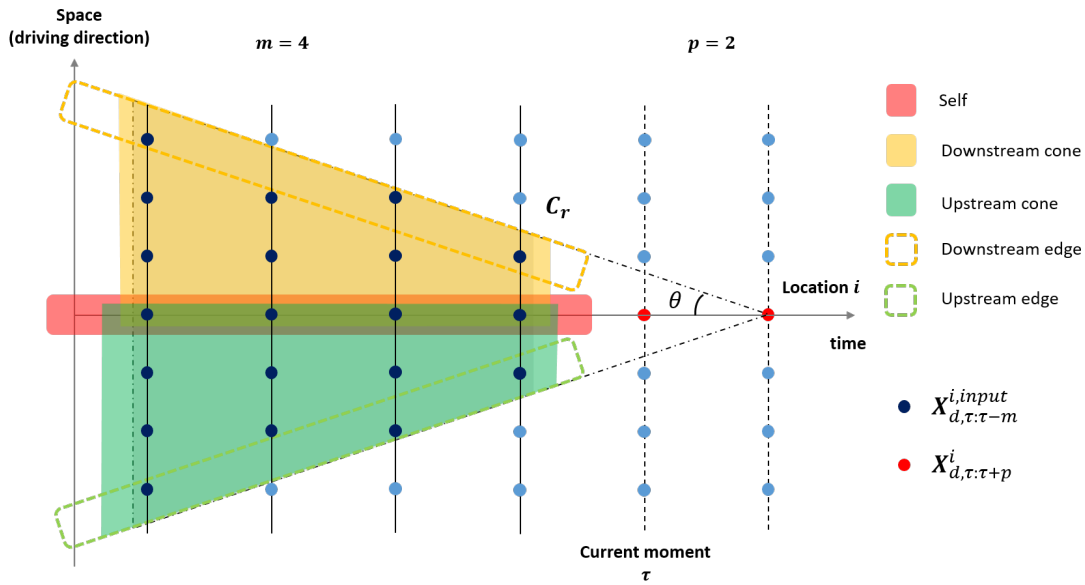


Figure 3.1: Illustration of localized spatial correlation: an example of input-output pairs. The dash-dot line triangle is the spreading cone; Sub-areas are marked by different colors.

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.3.1):

(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. (3.16) 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.

3.3.3 k-p nearest neighbours 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 neighbours estimator (Wang et al., 2009). In this study, we choose the k-p nearest neighbours estimator proposed by Lombardi & Pant (2016). Compared to the k-nearest-neighbours (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 neighbours. We emphasize one fact: *most traffic patterns tend to fall into several clusters and there are few rare patterns located between them.* This property has been shown by some studies on congestion pattern 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.3.1. This estimator needs to calculate one Gaussian distribution and one corresponding integral:

$$g(\boldsymbol{x}) = \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{3.19}$$

$$G(\boldsymbol{x}) = \int_{\mathscr{B}(\boldsymbol{x},\varepsilon)} g(\boldsymbol{x})d\boldsymbol{x} \tag{3.20}$$

The major drawback of kpN estimator is the relatively higher computational complexity to calculate (3.20). As pointed out in Lombardi & 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 Eq. (3.2)).

Algorithm 3.1 kpN entropy estimator, from Lombardi & Pant (2016)

Input:
$\boldsymbol{X}_{N\times d}$, N observation samples of dimension d random variables
k, number of nearest neighbours to calculate local probability mass
p, number of nearest neighbours 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 $\mathscr{B}(\boldsymbol{X}_i) = \boldsymbol{X}_i \pm \varepsilon_i \boldsymbol{e}$
Calculate g_i in Eq.3.19
Calculate the integral G_i in Eq3.20
end for
Calculate $\hat{H}(X) = C + \mathbb{E}[\ln G_i] - \mathbb{E}[\ln g_i]$

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 theorem 1 and 2 give two different metrics of predictability that depends on locations and time of day.

3.4 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).

3.4.1 Data description

The major counter-clockwise ring freeway around Rotterdam (The Netherlands) is selected as a case study (shown in Fig.3.2). 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 & Minami, 1984; Treiber & Helbing, 2003) is used to estimate 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. The data stream starts at 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 how busy the highway is (on average). 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 $0.1\,\text{km} \times 30\,\text{s}$ uniform grid, accumulation is estimated by:

$$N_t = l \sum_{i=1}^{L} n_i \times \frac{Q_t^i}{V_t^i} \tag{3.21}$$

  where L is the number of uniform links; $l = 0.1\,\text{km}$ is the length of link (spatial resolution); n_i is the number of lanes on each link. Then N_t is aggregated every 5 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 $0.1\,\text{km} \times 30\,\text{s}$ uniform grid, then the processed data is aggregated every 1.2 km and 4 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.

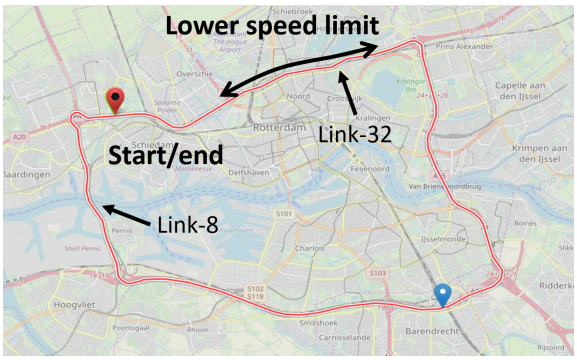


Figure 3.2: The counter-clockwise ring freeway around Rotterdam.

3.4.2 Predictive models

For accumulation forecasting, we select three baseline deterministic models: (1) **k-neareast neighbours (KNN)**: the similarity metric is Euclidean distance and the weight of averaging is inverse of the distance. The optimal number of neighbours is chosen by cross-validation. (2) **FCNN**: a fully-connected feed-forward neural network 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. The optimal number of neighbours is searched by cross-validation. (2) **DCRNN**: (Li et al., 2018) is one of the state-of-the-art network-level traffic forecasting models that employ 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 an STGCN-like probabilistic model with U-Net-like skip connections (Ronneberger et al., 2015). The model details can be found in the 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 the mean and variance (μ, σ^2) of each component.
- **Beta distribution**: The marginal distribution of each component of output is a beta distribution $\mathscr{B}(\alpha, \beta)$ 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 $\kappa = \alpha + \beta$, $\kappa \in (2, +\infty)$. In this way, the distribution is uni-modal with a 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 the validation set is used to mitigate over-fitting. For those models using recurrent encoder-decoder structure, the 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 the 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 either. Therefore, dataset shift can be effectively reduced. For example, if the split ratio of the training set is 0.7 and $k = 10$, our speed dataset contains 66405 observations (233 days, 19 hours and 4 minutes interval everyday) for each location, then the expectation of samples that are not included in all k-folds is $66405 \times (1 - 0.7)^{10} \approx 0.39$, which is negligible. The estimated lower bound is reliable.

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

3.4.3 Accumulation forecasting

Fig.3.3 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 every day.

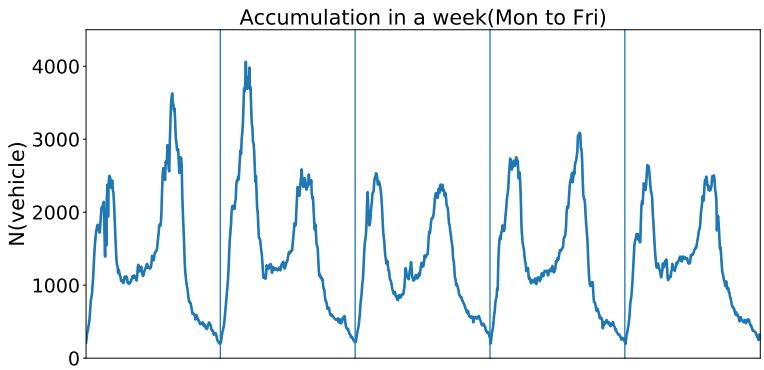
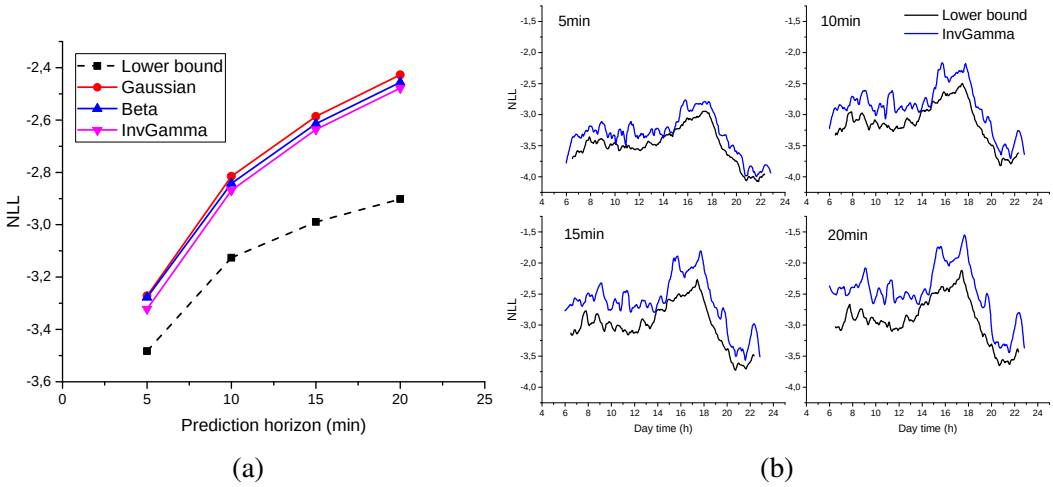


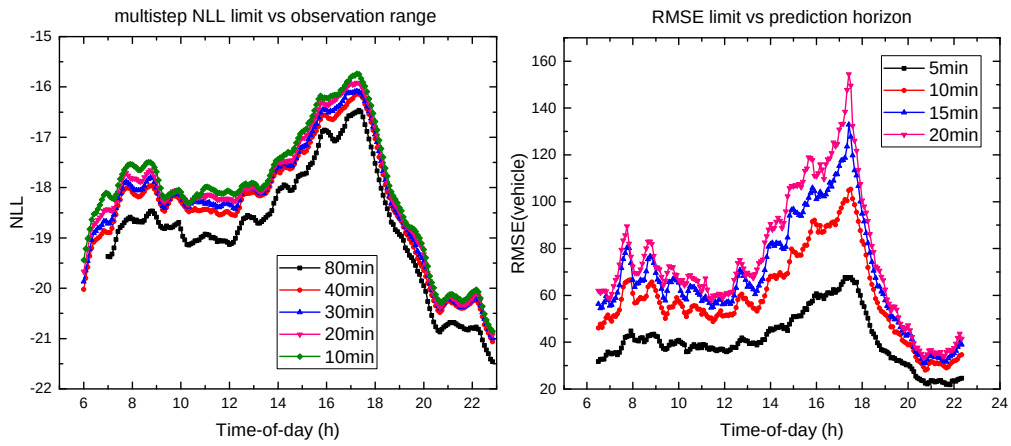
Figure 3.3: The evolution of accumulation from Monday to Friday in a randomly selected week

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.3.4a, 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.3.4b, we further compare the temporal curve of the 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.



(a) (b)

Figure 3.4: (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)



(a) (b)

Figure 3.5: (a) relationship between observation range and multistep NLL limit, $p = 4$; (b) Relationship between prediction horizon and RMSE limit of each prediction step, $m = 6$.

Fig.3.5 quantifies the influence of observation range and prediction horizon. In Fig.3.5a, 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.3.5b shows that the RMSE limit of each step increases fast with the prediction horizon. The difference is more significant during peak hours. For example, at around 17:30 PM, the RMSE limit increases from 60 vehicles to a maximum of 160 vehicles in 20 min. This result implies that accurate long-term prediction is theoretically impossible without inducing more data.

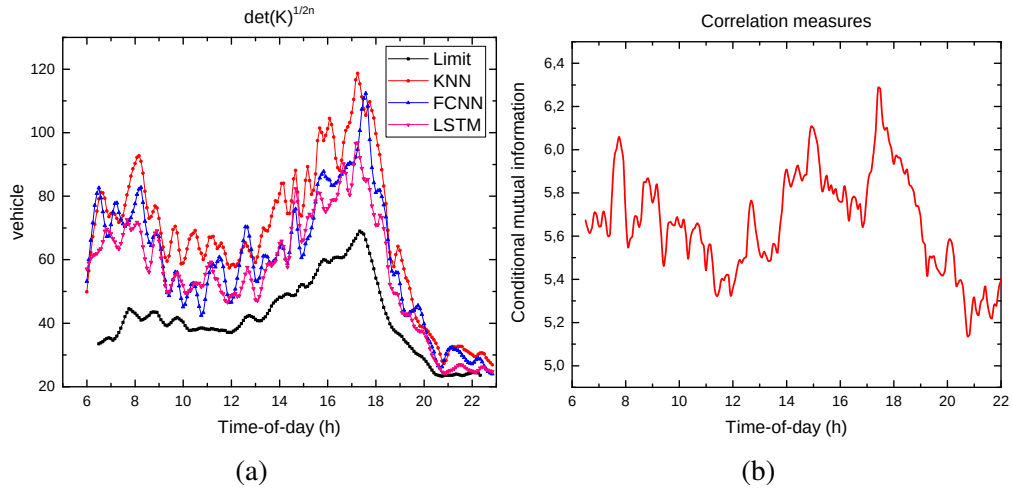


Figure 3.6: (a) comparison between the lower bound of DCM and deterministic models' performances; (b) conditional mutual information for 4-step predictions.

Fig.3.6 presents more analysis. Fig.3.6a 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 the 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 modelling cannot be larger than the gap. Fig.3.6b 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 phenomenon is consistent with Fig.3.5b.

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

3.4.4 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.3.7. 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 the 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.

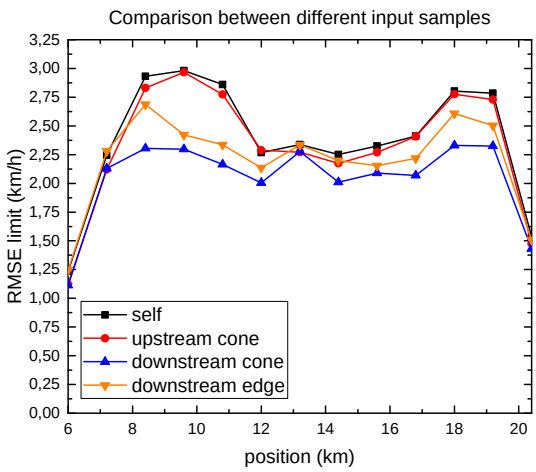


Figure 3.7: The lower bound of RMSE for different input sets: $m = 6$ and $p = 1$

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 first consider a forecasting task with $m = 6$ and $p = 1$. The result is shown in Fig.3.8. The lower bounds for other m and p have similar spatio-temporal distributions but different magnitudes (similar to what has been shown in Fig.3.5). 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.3.2).

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.3.9, 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 the negative axis has no 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 the prediction horizon. It implies that the true distribution of speed is complex.

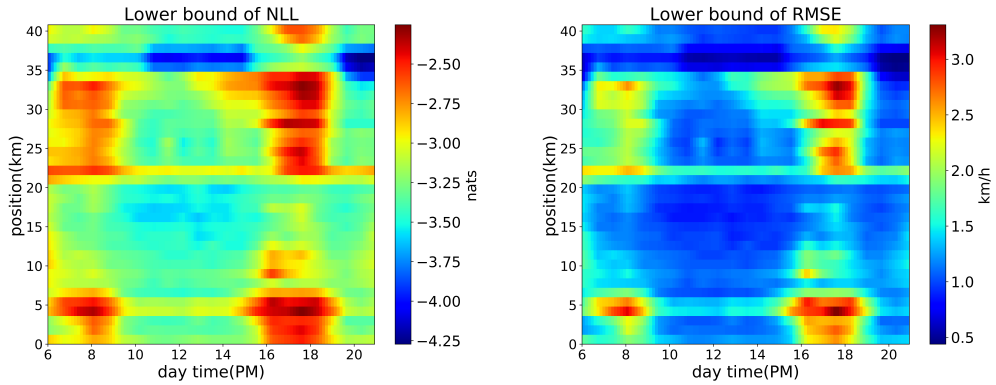


Figure 3.8: The spatio-temporal lower bound of NLL (left) and RMSE (right) for speed forecasting, $m = 6$ and $p = 1$

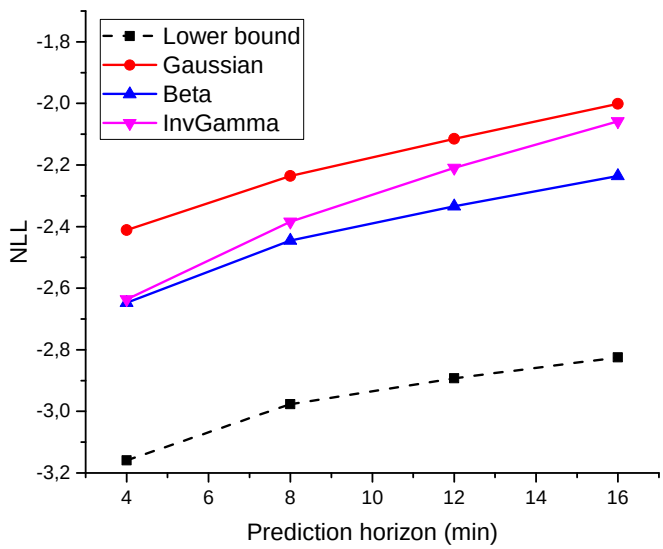


Figure 3.9: Comparison between average lower bound of NLL and the performances of probabilistic models for each prediction step in speed forecasting

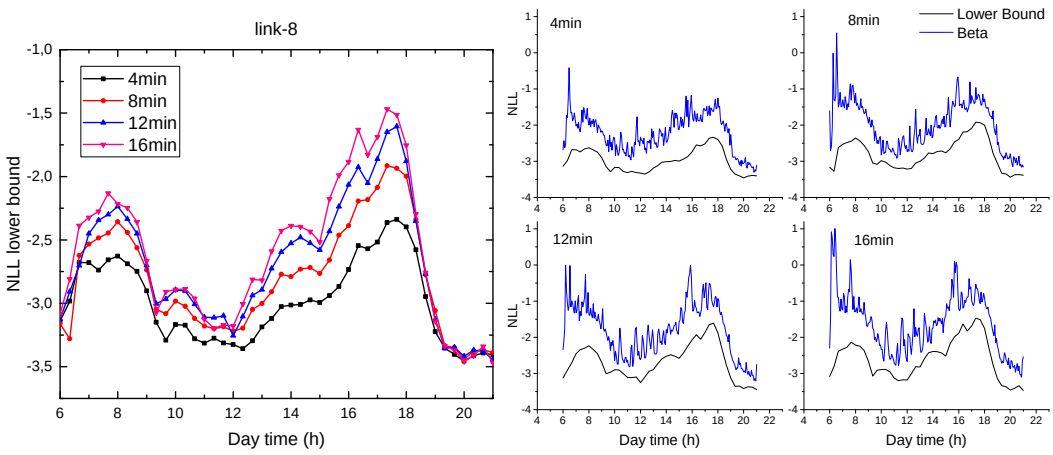
Next, we study this spatio-temporal limit of predictability by slicing. We will select some representative examples. For temporal predictability, we select the link with the lowest average speed (link-8) and the link with the highest standard variance of speed (link-32). Their positions are marked in Fig.3.2. 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.

Temporal predictability

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

Fig.3.10a shows the limit of NLL for each prediction step on link-8. Speed is signifi-

cantly more uncertain during morning and evening peak hours, but highly predictable at noon and at night. The lower bound increases with the 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 prediction horizon – long-term accurate point-estimate prediction in highly-uncertain situations is theoretically impossible if no additional data is provided. In Fig.3.10b the lower bounds are compared with the NLL of the 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, and cannot be well approximated by a simple uni-modal prior. In the appendix, another example of link-32 is presented.



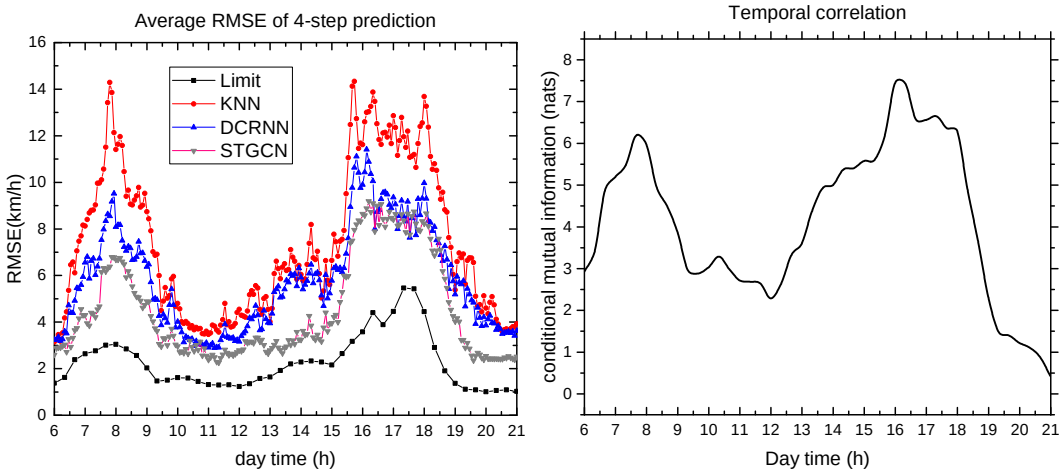
(a) (b)

Figure 3.10: (a) the lower bound of NLL for each prediction step on link-8; (b) comparison between the NLL lower bounds and the performances of the Beta-prior probabilistic model on link-8.

Fig.3.11a shows the lower bound of RMSE and the predictive errors of different deterministic models on link-8. Here the lower bound is equal to the square of the arithmetic average of marginal variance (see (3.14)). 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.3.9 and theorem 2, we infer that approximating speed time series by a Gaussian process is unreliable. The room of improvement for modelling is much smaller than the gap shows. Fig.3.11b presents the conditional mutual information. We see that multistep predictions on link-8 are significantly correlated. The correlation is stronger during peak hours.

Spatial predictability

In terms of spatial predictability, we are particularly interested in where are the most uncertain locations and why. In this subsection, we do not repeatedly show the same



(a) (b)

Figure 3.11: Link-8: (a) Comparison between the 4-step RMSE lower bound and the RMSE of deterministic models; (b) Conditional mutual information.

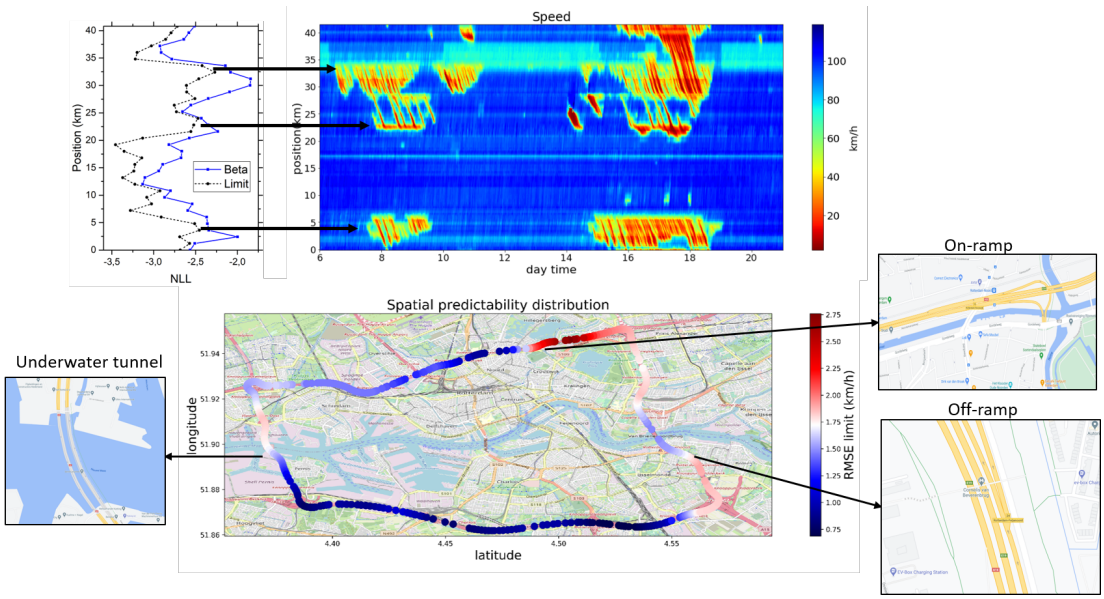


Figure 3.12: Top: Comparison between the spatial predictability, NLL of Beta prior model, and the speed evolution ground-truth; Bottom: identify the most unpredictable positions on the ring freeway.

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.3.12 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 the 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 is 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.3.12 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 behaviours 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 optimise where to install sensors to maximize performance-cost ratio (Gentili & 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 adjacent urban roads. For traffic managers, extra attention should be paid to these highly uncertain locations because they largely determine the congestion evolution of the entire beltway.

3.4.5 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 a 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 ranges and shorter prediction horizons 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 modelling can be quantified.

3.5 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 the probabilistic model and a lower bound of MSE and 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 the estimated theoretical limit but not cross it in most cases. The influence of observation range and prediction horizon is also clarified and quantified. Longer observation windows can increase 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 modelling and estimate how much potential room remains for modelling. 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 problems 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 modelling may be less than expected. Our results showed that the limit of predictability of one single traffic quantity (such as speed) drops rapidly with the 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 a k-fold strategy to mitigate dataset shift and avoid the failure of the 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 as well.

Acknowledgements

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

Chapter 4

Uncertainty quantification in network traffic forecasting

The previous chapter addressed the average predictability estimation in a model-free approach. Now we present a detailed analysis of predictive uncertainty for each sample when deploying the model in the data stream. The deep-ensemble-based approach shows that the irreducible randomness mainly restricts the predictability of traffic state in traffic dynamics instead of the emerging rare cases. The proposed method also explains that the lack of microscopic driving behavioural data leads to a bifurcated future, which is the cause of the limited predictability.

This chapter is submitted to a journal (under review).

ABSTRACT

Uncertainty quantification in living data stream must consider both the inherent randomness in traffic dynamics, the so-called aleatoric uncertainty, and the additional distrust caused by the lack of knowledge in deployment, the epistemic uncertainty. They together depict how predictable macroscopic traffic is. In this chapter, we use deep ensembles of probabilistic graph neural networks to estimate both types of uncertainty in network-level speed forecasting. Experimental results reveal that, although rare congestion patterns always emerge in real-life, the short-term predictability of traffic state is mainly restricted by the irreducible stochasticity and bi-modality in traffic dynamics instead of data shortage or imperfect modelling. The future traffic state substantially bifurcates. We argue that the improvement room for better modelling is limited and investing in diversifying data types is more important.

4.1 Introduction

The quality of many travel services and traffic management functions, such as arrival time estimation (ETA) (Van Lint et al., 2005), real-time route planning (Gehrke & Wojtusiak, 2008; Liebig et al., 2017), and congestion control (Chen et al., 2000), relies on accurate short-term traffic state forecasting. The application value greatly drives the development of new models to continuously improve predictive performance. These methods range from classical traffic-simulation-based approaches (Ben-Akiva et al., 1998; Qiao et al., 2001; Wang et al., 2006, 2016), regressive and time-series models (Castro-Neto et al., 2009; Davis & Nihan, 1991), to recently popular deep learning models (Van Lint et al., 2002; Ma et al., 2017; Li et al., 2017). Currently, machine learning and deep neural networks (DNNs) are being actively studied and widely applied for predicting the traffic state of large road networks.

However, like other non-trivial prediction tasks, traffic forecasting is also coupled with *uncertainty*. The evolution of traffic state in any road network is not deterministic but inherently stochastic for any observers. The unavailability of some critical information (to agencies performing the forecasting task), such as operational and tactical driving behaviours and route choices, induces randomness in both supply and demand dynamics. This limited observability implies that the input and the output of any traffic forecasting model must be formulated as a set of *correlated* random variables. The output should be a probability distribution function (PDF) instead of a scalar value. To develop forecasting models equipped with uncertainty quantification, it is necessary to clarify what the uncertainty represents and how to model it.

From the perspective of modelling and application, predictive uncertainty is typically categorized into two parts, the so-called *aleatoric* uncertainty and *epistemic* uncertainty (Der Kiureghian & Ditlevsen, 2009; Kendall & Gal, 2017). Aleatoric uncertainty (from alea, Latin for 'dice') represents the inherent randomness of a stochastic process and the measurement error in data collection, which cannot be explained out by expanding the dataset. This uncertainty draws a lower bound of predictive accuracy for any models using the same (types of) inputs (Li et al., 2022a). For example, for a Gaussian process, the variance gives the limit of mean-square-error for any predictors that give a point estimate. On the other hand, epistemic uncertainty stems from the abstraction of the used model and the lack of data to learn the output distribution. Take the same example, the used predictor may be not good enough to model the Gaussian process or there is not enough data provided to calibrate the parameters (such as mean and variance). This will induce additional uncertainty. The idea is that—in principle—epistemic uncertainty can be reduced by expanding the training set, and/or by adding (supposedly known) sophistication to the applied model.

Uncertainty Quantification (UQ) has been widely studied for many risk-sensitive systems with partially-observable information, such as nuclear safety (Helton, 1993), hydrology (Beck, 1987), meteorology (Deser et al., 2012), etc. If a system is explicitly simulated through physical models, then all uncertainty pertaining to the parameters, the model structure, and the data used to calibrate and identify these respectively, can influence the confidence of the output. Examples in the traffic domain include the wide heterogeneity of car-following parameters in microscopic traffic simulations (Sharma et al., 2019) and the distribution of fundamental diagram parameters (e.g. capacity,

critical density) used in LWR-type models (Li et al., 2012). The output distribution for predictions using simulation models is typically obtained by simulating the *forward* propagation of such 'errors' (sources of uncertainty) (van Lint et al., 2012b).

At the other end of the spectrum, many *inverse* data-driven UQ methods ignore how uncertainty propagates within the model but directly build the joint probability of input-output from collected datasets. With the development of deep learning techniques, this end-to-end approach is getting popular and shows better performance in UQ, but the cost is low interpretability and poor generalizability. For distinguishing aleatoric and epistemic uncertainty, the two approaches above are in-principle the same. For each input, we need to obtain an ensemble of output distributions (Section.4.2 will give a brief overview) and decompose the total uncertainty (such as entropy, variance, or even skewness and kurtosis) into explainable (epistemic) and unexplainable (aleatoric) parts via conditioning (Brillinger, 1969).

Specific to network-level short-term traffic forecasting, we observe that most proposed models in the literature use those easily-observable macroscopic variables (speed, flow etc.) as inputs (Ma et al. (2017); Li et al. (2017), etc.) and they specifically focus on modelling improvement and accuracy comparison. Disentangling input-dependent aleatoric and epistemic uncertainty can address a critically important but in many cases ignored question:

How predictable is macroscopic traffic state, and why?

The two types of uncertainty must be considered together to evaluate the 'predictability'. If aleatoric uncertainty dominates, then using better models or collecting more data of the same type cannot significantly improve the prediction accuracy. Conversely, epistemic uncertainty can be regarded as a measure of how much predictive accuracy can still be achieved by better modelling or collecting more data. if epistemic uncertainty is higher, then investing in modelling techniques and expanding datasets are worthwhile. In summary, quantifying and comparing these two sources of uncertainty is informative for whether analysts and scientists should focus on expanding the source of heterogeneous data, or on more sophisticated modelling techniques.

In this study, we employ probabilistic models and Deep Ensembles (DE) (Lakshminarayanan et al., 2016) to quantify both aleatoric and epistemic uncertainty associated with network-level, multi-step highway speed forecasting. The major contributions of this paper are summarized as follows:

- Use a deep ensemble of graph neural networks to quantify aleatoric and epistemic uncertainty in highway speed forecasting.
- Reveal that the predictability of traffic speed is mainly restricted by the inherent randomness of traffic dynamics, especially the bifurcation of long-term traffic state.
- Conclude that collecting more macroscopic traffic data or developing more complex correlation-based models cannot significantly improve prediction accuracy.

This paper is organized as follows. We first briefly overview the related works in the literature in section 4.2. Section 4.3 presents the uncertainty quantification method and details of the proposed model. In section 4.4 and 4.5, the model is tested on a real-world highway network and the analysis of the experimental results are presented. Finally, section 4.6 draws conclusions and proposes several relevant research directions.

4.2 Overview

Quantifying uncertainty in prediction models is an important topic in many domains. Recently deep neural networks (DNN) are getting popular and show better performance. In this subsection, we first summarize those UQ methods using DNNs in general, and then briefly review UQ methods and their applications in the traffic domain.

In principle, the key to quantifying aleatoric and epistemic uncertainty by DNNs is getting an ensemble of diverse output distributions that have close losses. Bayesian neural networks (BNN) are amongst the most popular methods to this end. However, directly training a BNN is difficult. For a given training set $\mathscr{D}$, the exact form of the posterior distribution of model parameters $p(\theta|\mathscr{D})$ is intractable. So one has to approximate it by a variational distribution. Several methods have been developed. Monte-Carlo dropout (MC-dropout) (Kendall & Gal, 2017) is one of the most widely-used methods. Dropout layers (Srivastava et al., 2014) is implemented in DNNs and these random dropouts are also enabled during inference. So we can obtain an ensemble of outputs by running the same model several times. Extra modification of predictive models is not required. However, Foong et al. (2019) show that MC-dropout is a relatively worse approximation for deeper BNNs (compared to shallow models), and not robust to data set to shift. Other approximation techniques include Markov chain Monte-Carlo (MCMC) (Kupinski et al., 2003), variational inference (VI) (Swiatkowski et al., 2020), Taylor-expansion based Laplacian approximation (Ritter et al., 2018), to name a few.

Deep ensembles (Lakshminarayanan et al., 2016) constitute another powerful approach for both improving predictive accuracy and estimating uncertainty. Specifically, random-initialization shows surprisingly good performance and robustness (Ovadia et al., 2019), and achieves on par with the state-of-the-art in many tasks. The high performance of DE can be interpreted from the perspective of 'loss landscapes'. Fort et al. (2019) show that random-initialization DE is able to explore multiple local optima of loss functions whereas a single optimization trajectory often falls into just one. During inference, we do not sample weights from one model but get multiple outputs from an ensemble of deterministic models (with different but fixed weights). The major drawbacks of DE are high computational resource requirements and slow inference speed. Some techniques like DE distillation (Malinin et al., 2019) are proposed to mitigate these problems. But to achieve comparable performance, pre-training a DE is still necessary in these methods.

Recently, some studies try to realize UQ through one deterministic model and one single forward pass, such as inducing radical-basis function (van Amersfoort et al., 2020) or deep evidential regression (Amini et al., 2020). But the effectiveness and

performance of these methods need to be tested on more tasks (Malinin et al., 2020). For a recent and comprehensive review of uncertainty quantification in deep learning models, we refer the readers to Abdar et al. (2021).

Specific to traffic state estimation and prediction, uncertainty quantification is fundamentally important. One of the most common approaches is using stochastic differential equations to predict traffic states (Chu et al., 2011; Tahmasbi & Hashemi, 2013). However, some critical parameters, like the number of incoming vehicles or the diffusion coefficient in Tahmasbi & Hashemi (2013), are not observable or hard to calibrate in practice. So many recent studies employ data-driven approaches. A classical data-driven method is Kalman filter (Liu et al., 2006; Van Hinsbergen et al., 2011), in which total uncertainty is decomposed into process error and measurement error. It was applied to quantify predictive uncertainty of univariate flow rates (Guo et al., 2014), speeds (Guo & Williams, 2010), and travel times (Van Hinsbergen et al., 2011). Heteroscedastic Gaussian Process (HGP) (Rodrigues & Pereira, 2018) is another data-driven method to quantify predictive uncertainty. It gives time-varying uncertainty in large-scale crowd-sourced traffic data, such as speed.

DNN-based UQ methods are also widely used in traffic forecasting. Bayesian neural networks are used to give confidence intervals of univariate time series, such as travel time (van Hinsbergen et al., 2009) or traffic flow prediction (Zheng et al., 2006). Graph neural networks (GNN) extend grid-like convolutional neural networks to graph structures (Kipf & Welling, 2016) and have been applied to network-level traffic forecasting because road networks can be naturally represented by a graph. In the literature, many GNN-based traffic forecasting models are proposed and some of them aim at quantifying predictive uncertainty, such as Bayesian GNN (Fu et al., 2020), deep echo state networks (McDermott & Wikle, 2019), and ensemble-based approach (Del Ser et al., 2020; Chen et al., 2021b). However, these papers above focus more on accuracy benchmarks and only present total uncertainty or confidence level. Aleatoric/epistemic uncertainties and their influences on traffic forecasting are not deeply discussed.

Different from the papers above, this study focuses on understanding the quantified predictive uncertainty from a traffic dynamics perspective rather than for just performance comparison.

4.3 Method

4.3.1 DE-based uncertainty quantification

In Section 4.2, we mentioned that deep ensemble is one of the most robust UQ methods and explained why from the perspective of loss landscape (Fort et al., 2019). In this section, we first briefly explain how DE works. Then we introduce what uncertainty metrics should be used and how uncertainty is quantified in this study.

Assume that a training dataset $\mathscr{D} = \{(\boldsymbol{x}_s, \boldsymbol{y}_s)\}_{s=1}^{S}$ is given. Here $\boldsymbol{X}$ and $\boldsymbol{Y}$ are input and output variables separately. S is the number of samples. An ensemble of K randomly-initialized probabilistic models are independently trained by using this dataset, noted as $\{\mathscr{M}_i\}_{i=1}^{K}$. Now for one specific test input $\boldsymbol{x}^*$, each model predicts the probability

distribution of output so the K models give an ensemble of distributions $\{p_i(\boldsymbol{y}^*|\boldsymbol{X} = \boldsymbol{x}^*)\}_{i=1}^K$. We briefly note it as $\{p_i(\boldsymbol{y}^*)\}_{i=1}^K$ from now on.

If the training set includes enough 'similar' input samples, they compose a considerable part of the total loss function during training. Therefore, the models can well assimilate the output distribution, regardless of the initialization of parameters. The ensemble of output distributions is consistent. In this case, we say that the epistemic uncertainty is low and the estimation of $p(\boldsymbol{y}^*)$ is reliable. $p(\boldsymbol{y}^*)$ itself represents the irreducible aleatoric uncertainty. On the other hand, if the test point $\boldsymbol{x}^*$ is a rare sample that was seldom or almost never observed in training, then these samples contribute a small part or zero to the training loss, which means models' behaviours differ a lot around this rare point due to weaker constraint in training. So the DE will give diverse distributions and thus the estimate of $p(\boldsymbol{y}^*)$ is unreliable. This is a so-called high epistemic uncertainty case. Fig.4.1 illustrates these concepts by a simple task, saying learning the Gaussian relationship $p(y|x) = \mathcal{N}(\mu(x), \sigma^2(x))$. The training set contains more samples in the centre but fewer samples around the boundary. The magnitude of noise σ^2 decays with $|x|$. $x^* \approx 0$ is a typical low epistemic and high aleatoric uncertainty test point. While $x^* \approx -3.5$ has significantly higher epistemic uncertainty due to the lack of data so we cannot give reliable predictions. In this example, collecting more data or improving modelling techniques between -2 and 2 cannot increase the prediction accuracy because the endogenous noise is irreducible. But collecting more data around the borders can significantly reduce epistemic uncertainty and thus improve the total performance.

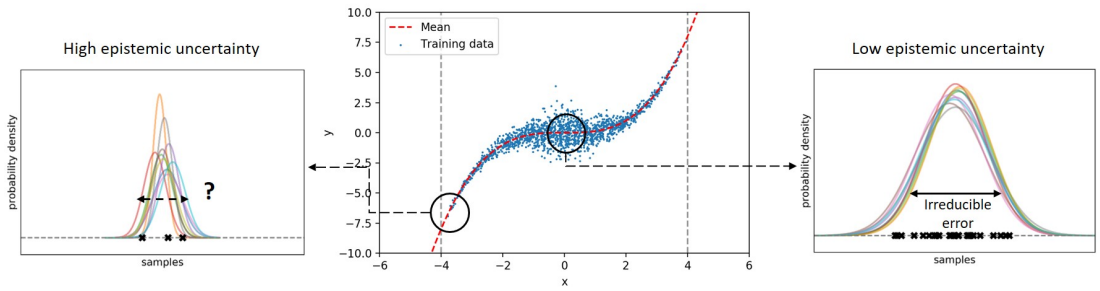


Figure 4.1: A 1D example of aleatoric uncertainty and epistemic uncertainty. Here 10 different models are used to learn $p(y|x)$ from the given training data. The left and the right figure shows two ensembles of predicted distributions at two different test points, $x = -3.5$ (left) and $x = 0$(right)

The discussion above gives a qualitative analysis of the DE-based UQ method. For quantitative estimation, three questions must be answered:

(1) What metrics should be used to measure uncertainty?

(2) How to quantitatively estimate them?

(3) What prior distributions should be used for a traffic forecasting task?

For the first question, there are two widely-used metrics to represent the uncertainty of a continuous scalar random variable, *variance* and *differential entropy*. For variance

metrics, if we note (μ_i, σ_i^2) as the mean and variance of each distribution in the ensemble, then the well-known *law of total variance* decomposes total uncertainty into epistemic and aleatoric parts as follows (Lakshminarayanan et al., 2016):

$$\mathrm{Var}(y^*) = \underbrace{\mathbb{E}(\sigma_i^2)}_{\text{aleatoric}} + \underbrace{\mathrm{Var}(\mu_i)}_{\text{epistemic}} \tag{4.1}$$

On the other side, differential entropy is defined as follows:

$$H(y^*) = -\int_Y p(y^*) \ln p(y^*) dy^* \tag{4.2}$$

Here $p(y^*)$ is the posterior distribution of output, which can be approximated by the average distribution of the ensemble (if K is large enough):

$$p(y^*) \approx \frac{1}{K} \sum_{i=1}^{K} p_i(y^*) \tag{4.3}$$

Similarly, Malinin et al. (2020) shows that the total entropy can be decomposed into the following terms:

$$H(y^*) = \underbrace{\mathbb{E}[H_i(y^*)]}_{\text{aleatoric}} + \underbrace{\mathbb{E}[D_{KL}(p_i(y^*) \mid\mid p(y^*))]}_{\text{epistemic}} \tag{4.4}$$

Here $H_i(y^*)$ is the differential entropy of each distribution in the ensemble and $D_{KL}(p_i(y^*) \mid\mid p(y^*))$ is the *Kullback–Leibler (KL) divergence* from each distribution to the (average) posterior distribution. KL divergence measures the directed 'distance' from the first distribution to the second one. It is non-negative. The definition is:

$$D_{KL}(p_i(y^*) \mid\mid p(y^*)) = \int_{p(y^*)>0} p_i(y^*) \ln \frac{p_i(y^*)}{p(y^*)} dy^* \tag{4.5}$$

For the second question, Eq.(4.1) and Eq.(4.4) give how to quantify both types of uncertainty. They have very similar forms. The aleatoric uncertainty is measured by the average variance or differential entropy of the ensemble distribution, and the epistemic term measures the *diversity* of the distribution ensemble by the variance of their mean values (point estimate) or by their average distance to the posterior distribution. They are consistent with the example shown in Fig.4.1. However, the epistemic terms in Eq.(4.1) and Eq.(4.4) have different advantages. Variance is scale-dependent but KL-divergence is not. Therefore, Eq.(4.1) is suitable for estimating the remaining room of accuracy improvement and Eq.(4.4) is better for objectively measuring the 'rareness' of input.

For the third question, the answer depends on what quantity we want to predict. Since traffic flow [veh/h] is not a state variable (i.e. low flows may coincide with both free-flowing traffic and heavy congestion), and density [veh/km] is difficult to measure directly, speed [km/h] is an appropriate index of congestion. In the literature, there are two ways to learn the output distributions. The first one is the *parametric* approach, which means we assume that the prior distribution can be represented by a small set of

parameters, such as Gaussian (Yuan et al., 2021; Rodrigues & Pereira, 2018). The second way is non-parametric histogram regression (Gu et al., 2021). The output space is discretized and the model learns the probability that the prediction falls into each interval. Parametric probabilistic models are usually easier to train, but only some simple priors can be used. The histogram regression method, on the other hand, theoretically can approximate any distributions. But the training is harder and it requires higher memory due to a large number of intervals. Generally, the non-parametric approach is only suitable for those tasks with only one or two scalar outputs, such as microscopic trajectory forecasting (Gilles et al., 2022) or dense object detection (Lin et al., 2017).

In this study, we explore both approaches. First, the parametric approach is used for quantifying two types of uncertainty. Here we choose *Beta distribution* prior due to the boundness of speed:

$$b(v;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} v^{\alpha-1}(1-v)^{\beta-1} \tag{4.6}$$

where $B(\alpha,\beta)$ is a Beta function that normalizes the density function. $\alpha > 1$ and $\beta > 1$ must be satisfied to make sure that the distribution is bounded everywhere. The mean and variance are:

$$\mu = \frac{\alpha}{\alpha+\beta} \tag{4.7}$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{4.8}$$

and its differential entropy is given by:

$$H = \ln B(\alpha,\beta) - (\alpha-1)\psi(\alpha) - (\beta-1)\psi(\beta) + (\alpha+\beta-2)\psi(\alpha+\beta) \tag{4.9}$$

$\psi()$ is the digamma function. For the standard Beta distribution, v is defined between 0 and 1. In practice, we can re-scale it between 0 and the speed limit. To summarize. the uncertainty quantification process is as follows:

(1) Train an ensemble of probabilistic forecasting models that are randomly initialized.

(2) Calculate the mean, variance, and differential entropy of each output distribution.

(3) The average of the mean values gives the predicted speed.

(4) For variance-based uncertainty metrics, directly use Eq.(4.1).

(5) For entropy-based metrics, first approximate the posterior distribution by Eq.(4.3), then use numerical methods to calculate the integrals in Eq.(4.5).

We also trained a histogram-regression model to explain what factors restrict predictability. The speed is discretized into 1 $\mathrm{km\,h^{-1}}$ intervals from 0 to the speed limit so the task is converted to a classification problem. The learnt probability of all classes (mesh unit) together gives the approximated output distribution.

The discussion above shows the uncertainty quantification method used in this paper. Next, we will formulate the traffic forecasting problem and introduce the used graph neural networks.

4.3.2 Problem formulation

It is difficult to forecast the evolution of congestion patterns solely from speed. On a spatio-temporal map, a speed drop (congestion) appears suddenly, which is the result of a mismatch between traffic demand and supply (capacity) on a particular road section. In practice, we observe that the flow usually goes up about 10-20 min before a speed drop around an on-ramp (it indicates that probably more vehicles are coming). So we expect that including flow in input can increase the prediction accuracy.

Given a highway network with N links represented by a graph $\mathscr{G}$. Its adjacency matrix is noted as $\mathbf{A}_{N\times N}$. The time interval between two adjacent observations (δt) is fixed. We aim to predict the *marginal* distribution of speed (V) at each link i and time stamp t in the next T steps from the observed speed and flow (Q) in the past P steps. The problem formulation is modelling the following conditional probability distribution:

$$p_{i,t}(V_{i,t} \mid \mathbf{V}_{P\times N}, \mathbf{Q}_{P\times N}, \mathbf{A}_{N\times N}) \quad \forall\, 1 \le i \le N;\ 1 \le t \le T \tag{4.10}$$

4.3.3 Model structure

Considering that the output is the probability distribution of multi-step *speed*, we prefer using a fully convolutional model instead of an RNN-based model so rolling out flow predictions in the decoder can be avoided. The proposed model is adopted based on the attention-based spatial-temporal graph convolutional networks (ASTGCN) proposed by Guo et al. (2019). This model is mainly composed of stacked spatio-temporal blocks (ST-block) and residual connections.

One ST-block's structure is shown in Fig.4.2. It sequentially contains a dynamic graph convolutional module proposed in Li et al. (2021), a temporal attention layer, a normal convolutional layer along the time axis, and an extra residual connection. Batch normalization (Santurkar et al., 2018) is applied at the end of each ST-block. The output shape of every ST-block is (P,N,F). Here F is the feature dimension. 10 ST-blocks are stacked together to extract spatio-temporal features from inputs. More details of each layer in the ST-block can be found in the appendix B.

After the stacked ST-blocks, the hidden representation is converted to the desired output shape by an output module. The two different output module structures are shown in Fig 4.3. For modelling Beta distributions, we learn its mode ω and reduced concentration κ:

$$\begin{aligned} \kappa &= \alpha + \beta - 2 \\ \omega &= \frac{\alpha - 1}{\kappa} \end{aligned} \tag{4.11}$$

After applying two normal 2D convolutional layers, the hidden features are split into h_1 and h_2, and then they are activated by the following functions to give ω and κ:

$$\begin{aligned} \omega &= \text{sigmoid}(h_1) \\ \kappa &= \exp(h_2 + 3) \end{aligned} \tag{4.12}$$

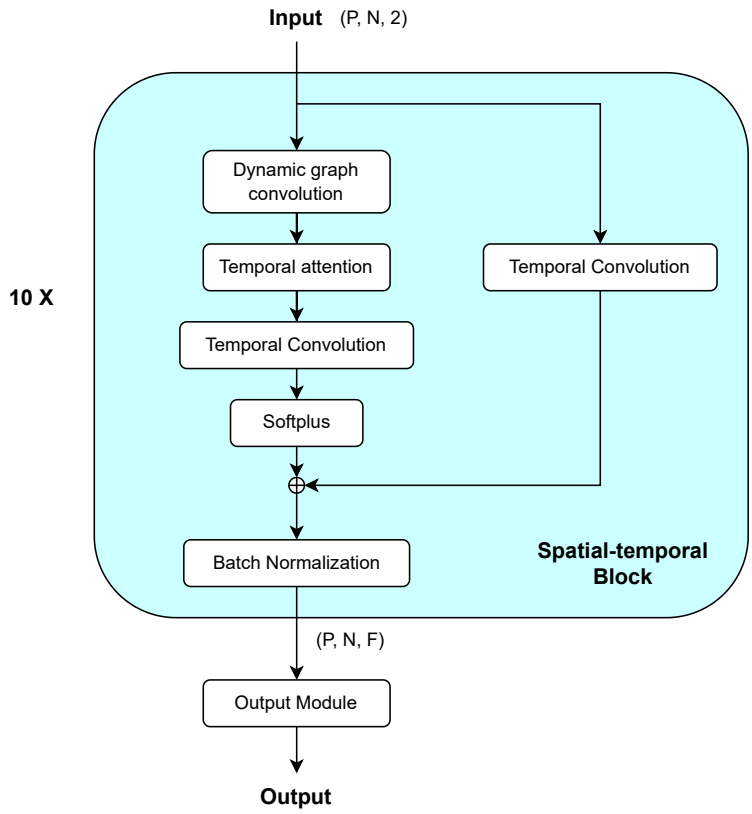


Figure 4.2: The structure of the proposed model

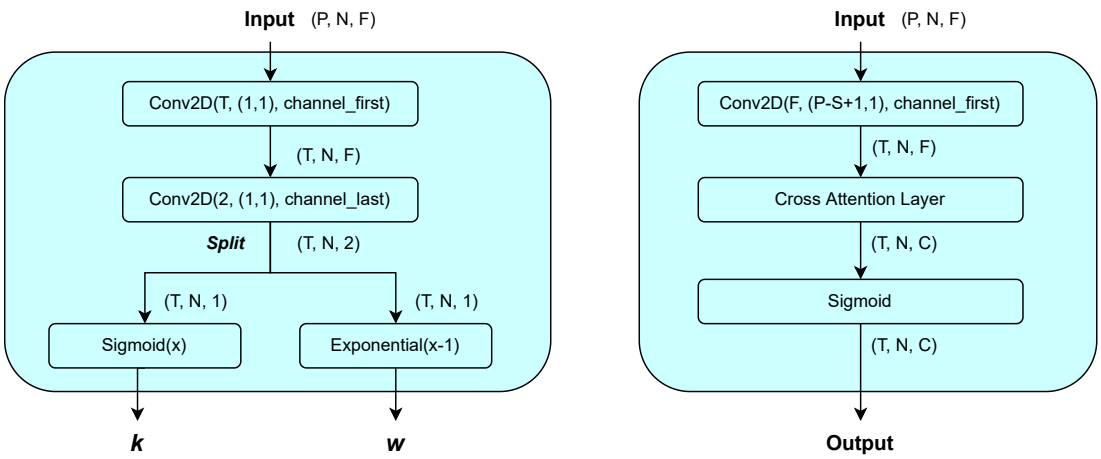


Figure 4.3: The structure of the output module. The left one is for Beta distributions and the right one is for histogram regression.

Compared to directly learning a and b, this scheme is numerically more stable and converges faster (similar to the activation strategy used for Gaussian prior. See Kendall & Gal (2017) for more details). The loss function is negative-log-likelihood (NLL), defined as follows:

$$NLL = -\sum_{i,t} \ln b(V_{i,t}^{\text{label}}; \omega_{i,t}, \kappa_{i,t}) \tag{4.13}$$

For histogram regression, assume that the speed is uniformly discretized into C intervals. The output module firstly adjusts the dimension to the desired shape by a 2D convolutional layer, then cross attention layers (Huang et al., 2019b) map the features to each interval. Considering that every location in the highway network has its unique properties, such as the number of lanes, speed limit etc., the cross-attention layer shall not share parameters among different road links. Otherwise, the model will give similar predictions for all locations (the model fails). This non-shared strategy greatly increases the complexity of the model and decreases the speed of training. That is why we only train one histogram-regression model for interpreting the estimated predictability instead of training an ensemble of such models for uncertainty quantification.

More details of the cross-attention layer can be found in the appendix B. The final output is activated by the sigmoid function. Focal loss (Lin et al., 2017) is used to learn the unbalanced distribution of labels. Here Y_p is the ground-truth of class (0 or 1) and $\hat{Y}_p$ is the predicted probability of each class.

$$L = -\frac{1}{P}\sum_{p}(Y_p - \hat{Y}_p)^2 f(Y_p, \hat{Y}_p) \tag{4.14}$$

$$f(Y_p, \hat{Y}_p) = \begin{cases} \ln \hat{Y}_p & \text{if } Y_p = 1 \\ (1 - Y_p)^4 \ln(1 - \hat{Y}_p) & \text{else} \end{cases} \tag{4.15}$$

4.4 Experiments

In this section, experiments are carried out on a real-world highway network. All data used in this paper are collected and processed by the Dutch National Data Warehouse (NDW—www.ndw.nu). The speed and flow data are measured by dual inductive loop detectors on the freeways. The raw data are aggregated to average speed and average flow per lane for each loop detector. These loop detectors are not uniformly distributed on the target network. We then apply the Adaptive Smoothing Method (ASM) (Schreiter et al., 2010a) to project the aggregated data onto a uniform grid and to fill in the missing values. In this study, we use the final processed data.

The highway network around Amsterdam (the Netherlands) is selected as a case study. The network is shown in Fig.4.4. It consists of 9 highways that connect the Amsterdam city centre, the surrounding suburban areas, several smaller towns, and Schiphol international airport. This busy network contains rich and diverse congestion patterns. All highways are uniformly partitioned into 400m length links, and we consider specific driving directions only (marked in Fig.4.4), resulting in a network of 193 links ($N = 193$). Both speed and flow are aggregated every 4 min by averaging. The observed speed and flow in the past 60 min ($P = 15$) is used to predict the evolution of speed in the next 40 min ($T = 10$).

The data for the entire year of 2018 is chosen as the training set. To focus on predicting traffic congestion, only congested periods are considered. The data preparation method is as follows. For one moment t, if any position's speed is lower than $40\,\mathrm{km\,h^{-1}}$ within $(t - 20\,\mathrm{min}, t + 20\,\mathrm{min})$, then this sample is added in the data set. So the prepared data

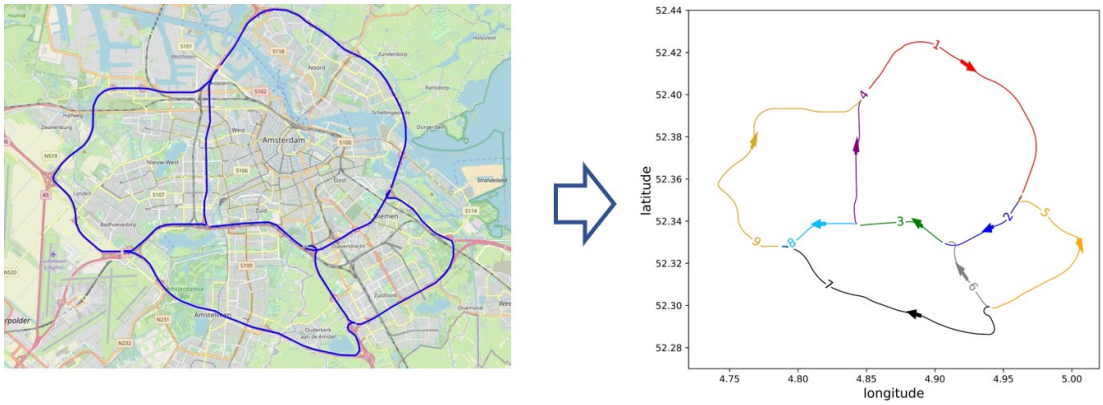


Figure 4.4: The selected highway network around Amsterdam. Arrows represent driving directions. The number of each road is also marked.

set includes highly diverse patterns ranging from pre- to post-congestion scenarios. Completely free-flowing traffic states are excluded, which makes this prediction task more challenging but also more valuable for uncertainty estimation. We randomly select 15% samples from the training set as the validation set. To mimic the real-world model deployment and continuous data collection environment, two different test sets are prepared, Mar. 1st - May 31st of 2019 (noted as `2019-test`) and Mar. 1st - May 31st of 2022 (noted as `2022-test`). `2019-test` is before the Covid-19 pandemic and `2022-test` is after ending the lockdown measurements in the Netherlands. We expect that their congestion patterns are somehow different. The number of samples for all used data sets is listed in Table 4.1. `2022-test` has fewer samples than `2019-test` because it has less congestion.

Table 4.1: Number of samples of used datasets

	Training	Validation	2019-test	2022-test
Nb. of samples	12964	1830	4053	3163

Following the recommendation given in Lakshminarayanan et al. (2016), we train maximum of 15 randomly-initialized Beta-based models for uncertainty quantification and 1 histogram-regression model for assimilating true distributions. The input speed and flow data are normalized by the z-score function for all datasets. Details of experimental settings and the graph neural networks can be found in the open source code: add before submission.

4.5 Results and discussion

In this section, we will sequentially present the experimental results of predictive accuracy, aleatoric uncertainty analysis, and epistemic uncertainty analysis. From these results, we will answer the research question central to this study: how predictable is macroscopic traffic state, and why.

4.5.1 Accuracy

Here we consider three widely-used accuracy metrics, mean-absolute-error (MAE, which is L1 loss), mean-absolute-percentage-error (MAPE), and root-mean-square-error (RMSE, which is L2 loss):

$$\mathrm{MAE} = \frac{1}{TN}\sum_{i,t}|V_{i,t} - \hat{V}_{i,t}| \tag{4.16}$$

$$\mathrm{MAPE} = \frac{1}{TN}\sum_{i,t}\frac{|V_{i,t} - \hat{V}_{i,t}|}{V_{i,t}} \tag{4.17}$$

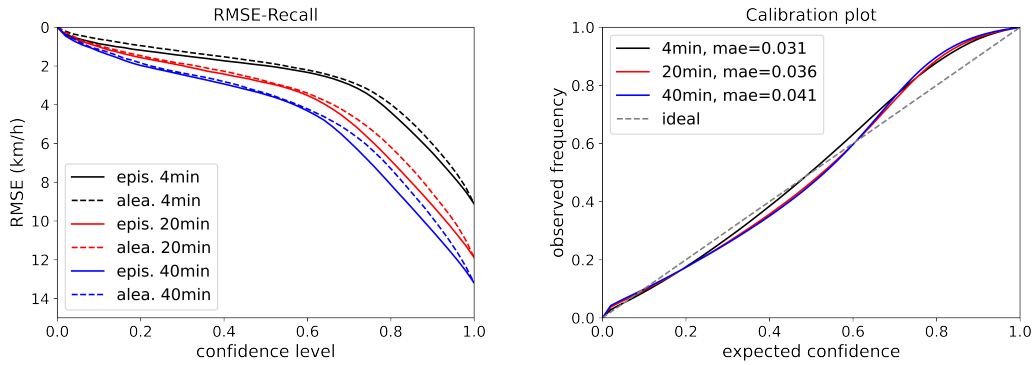
$$\mathrm{RMSE}^2 = \frac{1}{TN}\sum_{i,t}(V_{i,t} - \hat{V}_{i,t})^2 \tag{4.18}$$

Table.4.2 compares the prediction accuracy of the Beta-prior models used in this study with the original ASTGCN model that minimizes MSE (one model, not an ensemble). Overall, They have very close performances. We observe that the accuracy does not improve much for $K > 10$, which is consistent with the conclusion in Lakshminarayanan et al. (2016). We emphasize that the focus of this study is quantifying uncertainty instead of the accuracy benchmark. We do not discuss deeply on the accuracy results.

Table 4.2: Performances of different ensemble sizes on both test sets

Ensemble size K	MAE($\mathrm{km\,h^{-1}}$)	MAPE(%)	RMSE($\mathrm{km\,h^{-1}}$)	NLL
2019-test				
1	5.95	15.96	11.77	-2.45
3	5.89	16.15	11.57	-2.46
5	5.83	16.27	11.55	-2.48
10	5.77	16.17	11.46	-2.49
15	5.76	16.12	11.46	-2.49
ASTGCN	5.89	16.25	11.31	–
2022-test				
1	5.22	15.73	10.80	-2.51
3	5.07	12.15	10.37	-2.59
5	5.02	12.23	10.34	-2.60
10	4.96	12.12	10.25	-2.63
15	4.96	12.15	10.24	-2.64
ASTGCN	4.99	13.05	10.92	–

Before analyzing uncertainty, we need to evaluate how well the distributions are modelled. Here we evaluate the quality of uncertainty quantification by using the standard proposed in Kendall & Gal (2017). The RMSE-recall curve and calibration plots on the test set `2019-test` are presented in Fig.4.5. Fig.4.5a shows how RMSE improves by removing those speed predictions with aleatoric or epistemic uncertainty larger than a percentile threshold. We have two observations here. First, all curves are mono-



(a) (b)

Figure 4.5: (a) RMSE-Recall curves and (b) calibration plots for different single predictive steps on `2019-test`

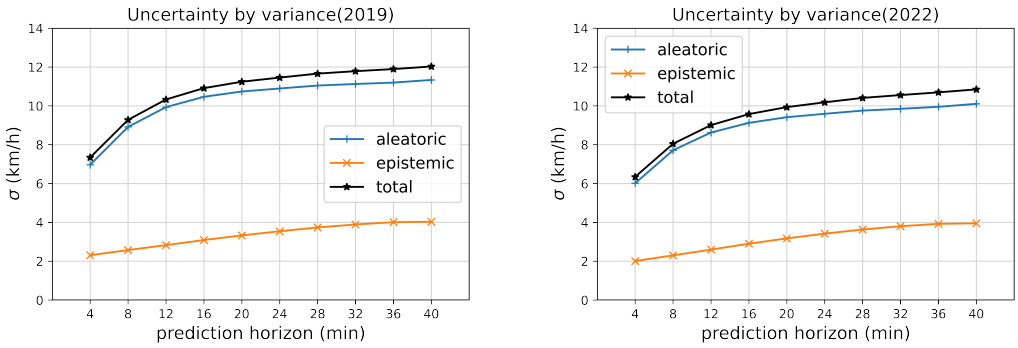
tonically decreasing in Fig.4.5a, which means that estimated uncertainty is strongly correlated with RMSE loss. Second, the curves of epistemic and aleatoric uncertainty models are very similar for the three selected predictive steps. In other words, if only one type of uncertainty is explicitly modelled, the model will always attempt to model the total uncertainty by compensating for the other one. This result is also reported in Lakshminarayanan et al. (2016). Fig.4.5b further clarifies how well the true distribution is modelled by Beta priors on average. The x-axis depicts the percentile of predicted distribution (expected confidence) and the y-axis is the percentage of observations that are indeed below this percentile (frequency). In the ideal case, the true distribution is perfectly modelled so the relationship between frequency and expected confidence should be $y = x$. In Fig.4.5b, we see that MAE distance to the ideal case increases with the prediction horizon. But the three curves are all close to the ideal line. It means that the uncertainty estimation is reliable on average.

The results above prove that the used Beta-prior model can well depict the output distributions in most cases. Next, we will focus on analyzing aleatoric and epistemic uncertainty respectively.

4.5.2 Predictability of traffic congestion

Fig.4.6 shows the relationship between two types of uncertainty and prediction horizon. On average, both epistemic and aleatoric uncertainty, and thus total uncertainty, increases with the prediction horizon on both test sets. For all prediction steps, aleatoric uncertainty is significantly higher than epistemic uncertainty, which means that the total predictive error (RMSE) is overwhelmingly determined by the inherent randomness of traffic dynamics. The remaining improvement room for better modelling and expanding the dataset is limited.

Next, we compare different uncertainty metrics. In Fig.4.7, the distributions of two types of uncertainty measured by entropy metrics and variance metrics on both test sets are presented. We see that using either variance or differential entropy, the distributions of aleatoric uncertainty are highly consistent. The `2022-test` set has lower



(a) (b)

Figure 4.6: Relationship between the average aleatoric uncertainty, epistemic uncertainty, total uncertainty of each predictive step, and the prediction horizon. (a) `2019-test`; (b) `2022-test`. Notice that uncertainty is measured by σ here.

average aleatoric uncertainty than `2019-test`. However, the distributions of epistemic uncertainty measured by variance or entropy are significantly different. Variance is scale-dependent so the epistemic uncertainty is positively correlated with the aleatoric uncertainty. The top right distributions are indeed very similar to those in the top left figure. There is no significant difference between the distributions of epistemic uncertainty measured by variance for the two test sets. The average epistemic uncertainty of `2022-test` is even slightly lower than `2019-test`. But the entropy-based epistemic uncertainty metric is scale-independent. The bottom right figure clearly shows that `2022-test` has higher epistemic uncertainty than `2019test` (see the two arrows, they mark the peaks of the two distributions). The difference originates from the fact that entropy-based metrics consider the diversity of distributions instead of only the diversity point estimates. For both test sets, most samples locate at the low epistemic uncertainty end. Only 11 samples are identified as out-of-distribution cases due to large-scale loop detector failures.

In summary, the analysis above answers the first part of the research question. The predictability of the macroscopic traffic state (RMSE lower bound) is mainly determined by the irreducible aleatoric uncertainty. If using variance-based metrics, 2022 and 2019 have almost the same epistemic uncertainty. While using entropy-based metrics, we observe a significant shift between the 2019 and 2022 test sets (based on the 2018 training set). This result demonstrates that the variance-based metric can give the improvement room of RMSE accuracy. In contrast, the entropy-based metric is more suitable for measuring how 'rare' the current traffic state is.

4.5.3 Bi-modality of speed forecasting

This subsection answers the second part of the research question: why is aleatoric uncertainty so high, and what causes low predictability? The key is using those concepts in traffic flow theory.

First of all, we visualize the statistical relationship between the estimated aleatoric uncertainty and the predicted speed. Considering that different locations on the highway

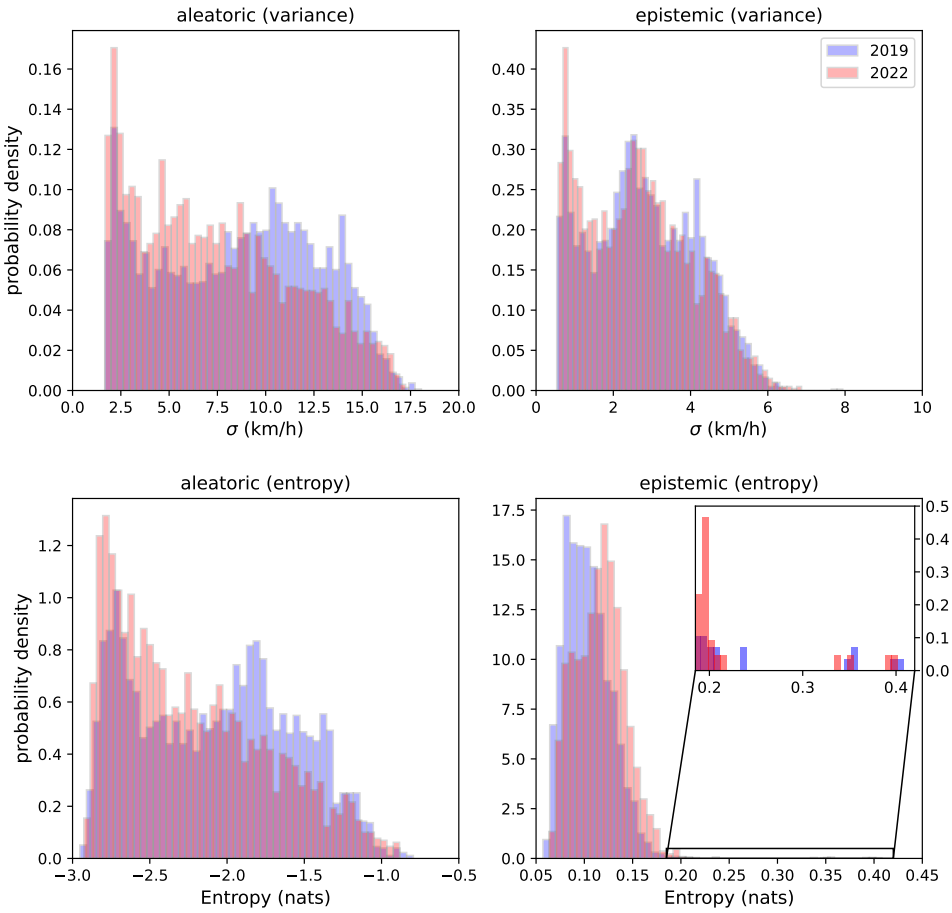


Figure 4.7: Distributions of prediction uncertainty on two test sets. Left column: aleatoric uncertainty; right column: epistemic uncertainty; top row: variance metrics; bottom row: entropy metrics.

network have different fundamental diagrams, it is reasonable to study this relationship for each specific link of the highway network. We manually checked all 193 links. Three representative examples are given in Fig.4.8. These three links are around three frequently congested on-ramps (marked on the right figure). The aleatoric uncertainty-speed relationship has a consistently similar 'inverse U' shape. The model gives relatively lower aleatoric uncertainty for both congested (low speed) and free-flowing (high speed) predictions. Free-flowing prediction is even more certain. However, the inherent randomness is significantly higher around the transition (capacity) state (medium speed, $50\,\mathrm{km\,h^{-1}}$-$60\,\mathrm{km\,h^{-1}}$), which represents the boundary between congested and free-flowing areas.

To explain this inverse-U relationship, we use the trained histogram-regression model to explicitly show the evolution of the approximated distribution of speed. Fig.4.9 presents an example at link-55. When the prediction horizon is short, the distribution is

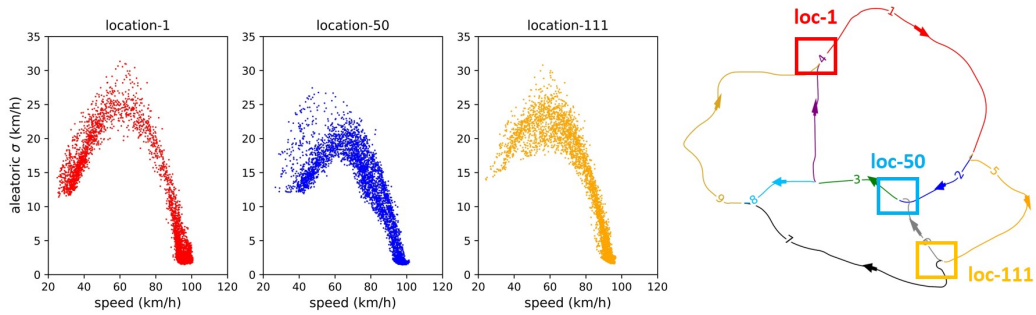


Figure 4.8: Relationships between the predicted speed and aleatoric uncertainty at three positions around congestion bottlenecks. Here we only visualize the result of 20 min predictions, but the conclusions hold for all prediction horizons.

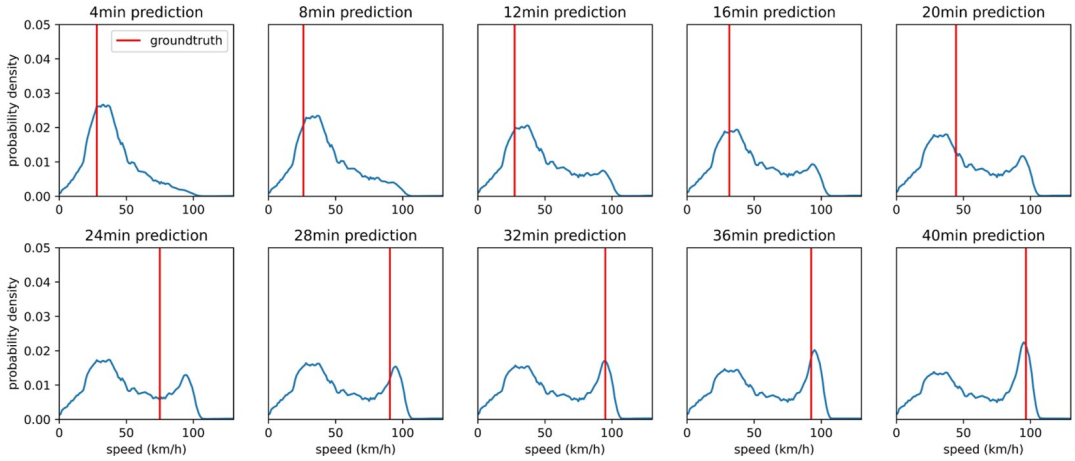


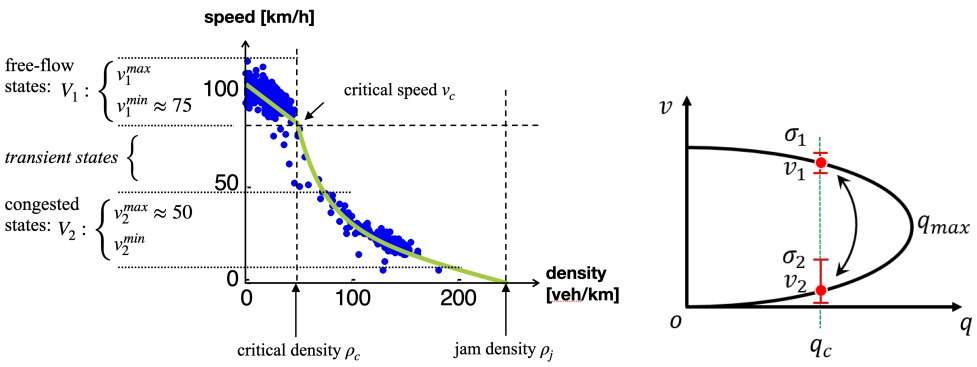
Figure 4.9: An example of predicted pdfs given by histogram-regression model at link-55. The red lines are the evolution of the groundtruth (labels).

uni-modal (only one local maximum) and highly concentrated around the label, which means that it can be well approximated by a Beta distribution. With the increasing prediction horizon, the variance increases and the distribution gradually show stronger bi-modality. One mode is the congested state and the other one is the free-flowing state. When such a bi-modal distribution is approximated by a uni-modal Beta distribution by minimizing NLL loss, the mean value (predicted point value) will locate at the middle ($50\,\mathrm{km\,h^{-1}}$-$60\,\mathrm{km\,h^{-1}}$, between two local maxima). The result directly interprets the inverse-U relationship observed in Fig.4.8.

Traffic flow theory explains the observed bi-modality in terms of the statistics of observed traffic states and the underlying explanatory mechanism. Fig.4.10a shows a typical speed-density relationship with 4-minute aggregate data from a typical location upstream of a major congested bottleneck, and a stylized (Smulders) fundamental diagram (FD) drawn over the data. The data illustrate that one can roughly distinguish three types of observed traffic states: those related to freely-flowing traffic (speed interval V_1); (highly) congested conditions (speed interval V_2) and transient states between these. Clearly, transient states occur less frequently than either of the other two states. These statistics make sense. Transient states include passing shock waves between larger spatio-temporal regions of either free-flowing or congested traffic, in

which due to acceleration and deceleration average speed is somewhere between V_1 and V_2. Most prolifically, this effect manifests as back-propagating stop-and-go wave patterns travelling upstream (often many kilometres). The precise frequency and onset of such waves are notoriously difficult to predict since they emerge from microscopic perturbations (e.g. sudden brake actions, cut-ins), which are unpredictable solely from average flow and speed data.

Another explanation is based on the perspective of capacity drop, as shown in Fig.4.10b. When congestion occurs, capacity drops from q_{max} to q_c so traffic states may collapse to either intersected point on the fundamental diagram. The transient state between them is unstable and seldom observed. But which interval the traffic state will fall depends on microscopic driving behaviours that cannot be known from macroscopic data. We refer the readers to Helbing et al. (2009) for comprehensive explanations.



(a) (b)

Figure 4.10: (a) A typical fundamental diagram measured at a location upstream of a major bottleneck. An approximate (Smulders) speed-density relationship is drawn over the measurements. Most observations fall into two intervals: free-flow states V_1 or (heavily) congested states V_2. (b) Fundamental diagram and capacity drop.

We now answer the second part of the question. The long-term predictability (aleatoric uncertainty) of highway congestion is low because of the impossibility to predict the precise time/frequency of congested wave patterns from speed and flow data, no matter what model is used and/or how big the flow/speed data set is. This uncertainty grows rapidly with the prediction horizon and will cause the bifurcation of future traffic states.

4.6 Conclusion and perspective

In this paper, we use a Deep Ensemble of graph convolution-based neural networks to quantify aleatoric and epistemic uncertainty in network-level speed forecasting. In this last section, we will first summarize the main findings and then give several related research topics.

This study illustrates that, for the case of using a full year of speed and flow data, the irreducible aleatoric uncertainty is significantly higher than the epistemic uncertainty in highway speed forecasting, although the test set of 2022 has more rare samples.

The evolution of speed itself is highly stochastic on highway networks so long-term congestion has substantially low predictability, given we just consider flow and speed as inputs. Further analysis shows that the true distribution of predicted speed has significant bi-modality, which can be explained by traffic flow theory. This bi-modality cannot be predicted solely from averaged speed and flow data because they do not have enough information about the *causation* of the bifurcation. We also showed that compared to uncertainty-based metrics, entropy-based metrics are better indicators of the 'rareness' of samples.

These findings justify the conclusion that neither investment in collecting more 'rare' (corner-case) speed and flow data, nor the development of more sophisticated models will lead to substantial improvement in traffic forecasting, *since this may reduce epistemic uncertainty only.* However, we also emphasize this conclusion needs to be substantiated with empirical evidence on different scales and different types of road networks.

We close with several ideas for further research. First, enriching the diversity of data types may be beneficial in reducing aleatoric uncertainty. For example, adding microscopic traffic quantities, such as vehicle trajectory data, may increase the predictability of congestion because they are the potential causation of the bi-modality of speed. How to effectively fuse these data from different levels into one predictive framework is a challenging topic. Second, in this paper, the rareness metric is directly computed from all predicted points. But in practice, we may be more interested in low-speed areas. So this definition could be combined with congestion extraction and classification techniques, such as Nguyen et al. (2019), to quantitatively study the recurrence of congestion patterns and have a bird-view of the long-term evolution of traffic. A third idea pertains to histogram-regression models. This method can also be combined with deep ensembles to quantify uncertainty. One of the biggest advantages of this model is that we do not need to assume the prior form of output distribution so the uncertainty estimation may be more precise. However, this model has higher computational complexity and it is more difficult to train. How to address these challenges also needs more investigation.

Acknowledgment

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

Disclosure statement

There is no potential conflict of interest reported by the authors.

Chapter 5

Uncertainty quantification in motion prediction

This chapter extends the previously-established uncertainty quantification method to the motion forecasting task. Considering the arbitrary road layout and the diverse interaction scenarios, entropy-based metrics are used to represent uncertainty and the correlation-based deep learning model is replaced by a designed causal model. Experiments show that inducing causal inference significantly improves the generalizability prediction accuracy. Further analysis points out that the major bottlenecks in trajectory forecasting are the unknown driving styles, unrecognizable intended directions, and the lack of domain knowledge of speed-dependent driving behaviours' heterogeneity.

This chapter is an article submitted to a journal and it has been pre-printed on SSRN:
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4241523
The content is the same as the pre-printed version.

ABSTRACT

Predicting the trajectories of road agents is fundamental for self-driving cars. Trajectory prediction contains many sources of uncertainty in data and modelling. A thorough understanding of this uncertainty is crucial in a safety-critical task like autopiloting a vehicle. We need to distinguish between the uncertainty caused by partial observability of all factors that may affect a driver's near-future decisions, the so-called *aleatoric uncertainty*, and the uncertainty of deploying a model in new scenarios that are possibly not present in the training set, the so-called *epistemic uncertainty*. This paper proposes a new framework to systematically quantify both sources of uncertainty. Specifically, to approximate the spatial distribution of an agent's future position, we propose a 2D histogram-based deep learning model combined with deep ensemble techniques for measuring both aleatoric and epistemic uncertainty by entropy-based quantities. The proposed Uncertainty Quantification Network (UQnet) employs a causal part to enhance its generalizability. Experiments on the INTERACTION dataset show that UQnet significantly improves the generalizability of the missing rate compared to the previous state-of-the-art. Further analysis shows that high aleatoric uncertainty cases are mainly caused by heterogeneous driving behaviours and unknown intended directions. Based on this aleatoric uncertainty component, we estimate the lower bounds of mean-square-error and final-displacement-error as indicators for the predictability of trajectories. Furthermore, we use epistemic uncertainty to identify rare cases in the test set. Our results illustrate that domain knowledge of speed-dependent driving behaviour is essential for adapting a model from low-speed to high-speed situations. Our paper contributes to motion forecasting with a new framework, that recasts this problem in terms of model generalization, and puts forward methods to quantify the resulting uncertainty.

5.1 Introduction

5.1.1 Background

Predicting human-driven vehicles' intentions, behaviours, and trajectories is an essentially important topic for engineers and researchers in many domains (Lefèvre et al., 2014). For example, in the context of autonomous driving, motion forecasting is indispensable for developing safe and smooth self-driving systems (Yoon et al., 2019). Accurate and reliable trajectory prediction is critical for motion planning (Wang et al., 2021). Trajectory prediction is also very important for multi-scale traffic modelling. Many phenomena in traffic flow originate from microscopic driving behaviours. For instance, heterogeneous car-following behaviours may cause road capacity drop (Yuan et al., 2018) and improper lane-changing around an on-ramp can lead to traffic congestion (Daamen et al., 2010; Leclercq et al., 2016). Building a reactive microscopic traffic simulator from real-world data requires high-quality trajectory prediction in the forward simulation (Bergamini et al., 2021). Many applications, such as dynamic traffic signal control (Chen et al., 2020) and collaborative platoon cruising (Hallé & Chaib-draa, 2005), can benefit from such a reactive simulator, and thus improve the macroscopic efficiency of traffic networks.

Recently, emerging Artificial Intelligence (AI) techniques and public trajectory datasets, such as Waymo (Sun et al., 2020), nuScenes (Caesar et al., 2020), Argoverse (Chang et al., 2019), etc. together stimulate the fast development of trajectory forecasting (Rudenko et al., 2020). In the literature, numerous Deep Neural Networks (DNN) have been proposed to continuously improve prediction performance. Significant progress has been made by applying AI to trajectory-based vehicle motion prediction. We refer the readers to Huang et al. (2022) for a comprehensive survey.

Although trajectory data are fundamental for motion forecasting, they clearly do not disclose the full complexity of driving behaviours. First, observed trajectories are just the final results of the underlying interactions between vehicles and the environment (including other vehicles). Many important factors, such as driving styles, vehicle characteristics and all the factors that influence these (e.g. turn signals), are not observable using trajectories alone. Second, there always is a non-zero probability of encountering rare behaviours (e.g. a risky cut-in) or circumstances (e.g. combination of high demand and signal malfunctioning) which may lead to different resulting interactions and thus different trajectories than present in the training data set. The model faces a much higher risk of getting it wrong in such a situation. Either way, *uncertainty* is unavoidable in trajectory-based motion prediction. Ignoring uncertainty will make the self-driving system unaware of potential danger and thus lead to accidents. Before quantifying uncertainty, it is necessary to clarify different types of uncertainty and their special roles in motion prediction.

5.1.2 Uncertainty in motion forecasting

From a modelling perspective, the total predictive uncertainty can be categorized into two types, **aleatoric** uncertainty and **epistemic** uncertainty (Der Kiureghian & Ditlevsen, 2009).

Aleatoric uncertainty represents the inherent randomness in the collected data and the underlying process itself. For motion prediction, aleatoric uncertainty has two major sources. The first one, which we call endogenous *interaction uncertainty*, is due to the fact that human drivers do not share all the information needed to predict the intended future behaviours of all other (relevant) drivers. As a result, they can at best make an informed guess about the driving strategies of their surrounding drivers before the interaction starts, and use this guess to for example cooperate or not (Lutteken et al., 2016). The consequence of this interaction uncertainty is that future trajectories may be completely different even with similar starting conditions. In a lane-changing case, for example, the following vehicle on the adjacent lane may yield to the cutting-in vehicle, or drive more aggressively and force the other driver to abandon the lane-change (Wang et al., 2015b). The second source of aleatoric uncertainty is due to the discrepancy between the information perceived by the so-called demonstrators (in this case the human drivers) and the imitators (sensors that observe these demonstrators), a problem which is also referred to as *sensor shift* (Etesami & Geiger, 2020). Restricted by sensoring methods, imitators do not perceive as much or even the same information as the demonstrators. For example, if trajectories are collected by drones, important stimuli such as traffic lights or turn signals are not observable due to the bird-eye-view position. Without these stimuli, the uncertainty around the predicted future state is much larger, particularly at decision points on the road (e.g. intersections). Also, data sets collected by sensors installed on vehicles, such as cameras, Lidar, Radar, etc, maybe "blind" to relevant signals that can potentially influence the demonstrators' decisions, for example, sound, glare, in-vehicle information, etc.

In summary, aleatoric uncertainty is the result of limitations in observability, either from the perspective of demonstrators (the agents whose path we aim to predict) or the imitators (the sensors partially measuring this path). We cannot distinguish between these components from data only, and the combined aleatoric uncertainty is thus *irreducible*, regardless of the size of the data set. The total aleatoric uncertainty can be understood as a lower bound of predictive accuracy for all models using the same type(s) of data. Aleatoric uncertainty is therefore also called *data uncertainty*.

Epistemic uncertainty, or alternatively, *knowledge uncertainty*, represents the uncertainty that in principle could be reduced to zero with the data available. Nonzero epistemic uncertainty is due to the "rareness" of the prevailing situation. In the case of trajectory prediction, it measures whether enough samples have already been seen to support reliable trajectory prediction. Under many circumstances, road users behave in similar ways (Makansi et al., 2021), which implies a prediction of this behaviour can be reliable most of the time. However, rare cases typically pertain to unsafe and high-risk situations. Finding out these "corner cases" by the prediction model itself is the key to building an "honest and trustworthy" self-driving system that can clearly tell what it does not know. Quantifying the "rareness" of these samples is also important for anomaly detection (Laxhammar & Falkman, 2013), continuous learning (Ebrahimi et al., 2019), and evaluating the generalizability of a model.

Although aleatoric and epistemic uncertainty quantification (UQ) has already been widely-studied, especially in combination with deep learning models (Abdar et al. (2021) provides a comprehensive review), applying it in motion forecasting has radically different requirements than in many other domains (e.g. traffic prediction).

First, a driver's intention is rigorously restricted by *arbitrary* layouts of road networks. Therefore, the spatial distribution (of probable trajectories) cannot always be approximated by simple priors (e.g. a 2D or mixture Gaussian) (Gilles et al., 2021). Second, the longitudinal and lateral components of 2D coordinates can be strongly correlated, yielding inflated or deflated uncertainty. If we use the covariance matrix as in classical UQ approaches, those non-diagonal elements are not intuitively explainable. Further, if the prediction has strong multi-modality (multiple local maxima), those diagonal elements are not meaningful either, regardless of the determinant of the covariance matrix. Fig.5.1 illustrates the problem. The left and right scenarios in Fig.5.1 represent different layouts (Fork versus T-junction) but exactly the same decision problem (left or right?) and thus the same degree of uncertainty in motion prediction. However, the co-variance in x and y ($\mathrm{Var}(x)$, $\mathrm{Var}(y)$), and the determinant of the covariance matrix in the left case (the fork) are all significantly smaller than the right case (the T-junction). We thus need to find a more reasonable *scalar* uncertainty metric that can exclude this artefact due to road layouts, in which the angle of two diverging or merging roads may inflate or deflate the estimated uncertainty.

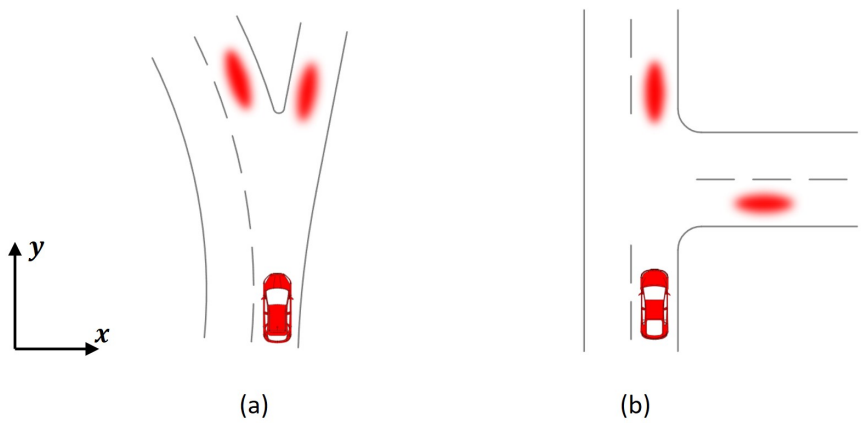


Figure 5.1: The spatial probability distribution of the target vehicle's future position in two different scenarios. They have different covariance matrices but the same differential entropy.

Third, we argue that epistemic uncertainty estimation is closely related to *domain adaptation* in motion prediction. The aim of quantifying epistemic uncertainty is detecting those *rare driving behaviours* (represented by trajectories) that have not been seen in the training set. However, trajectories are strongly *correlated* with the topology of lane networks. If the model cannot adapt itself to those unseen test lane networks by learning the correct *causal effects* of surrounding road agents, many normal driving behaviours in new scenarios will be incorrectly recognized as rare samples. This will lead to significant degradation in both generalization prediction performance (Hu et al., 2021) and epistemic uncertainty estimation.

In summary, the three points above call for a new approach to UQ for motion planning that we present in this paper. We propose a suitable uncertainty metric; a new non-parametric uncertainty quantification method; and an adaptable model to find out true rare interactions during inference.

5.1.3 Contributions and outline

Inspired by some recent works on motion prediction (Gu et al., 2021; Gilles et al., 2021) and UQ method (Malinin & Gales, 2018), we propose a non-parametric approach to estimate both aleatoric and epistemic uncertainty in human drivers' trajectory forecasting. Instead of using a set of parameters to assimilate the closed-form distribution of future positions, our model directly learns a mesh-grid 2D histogram (a heatmap) to approximate any distribution. This heatmap-based model is combined with deep ensemble techniques to quantify predictive uncertainty. Scalar entropy-based quantities are used as metrics. The conditional differential entropy and the mutual information represent aleatoric and epistemic uncertainty respectively. To make the estimate of epistemic uncertainty more reliable, a regularization net is added to the predictor to suppress (control for) spurious correlations so that the model is more robust in new scenarios. Experiments on the INTERACTION dataset (Zhan et al., 2019) show that the proposed UQnet method indeed has better generalizability. Based on the estimated uncertainty, we further analyze the predictability of vehicle trajectories and illustrate how speed is a key factor in model generalization.

The major contributions of this paper are summarized below:

- Propose a deep-ensemble-based non-parametric approach for quantifying both aleatoric and epistemic uncertainty measured by entropy quantities in single-agent trajectory prediction.
- Induce a causal regularization to enhance the generalizability and make the estimated epistemic uncertainty more reliable.
- Use quantified uncertainty for analyzing the predictability of trajectory and detecting out-of-distribution cases. The results further give insights into the role of speed in model generalization and vehicle interaction.

The paper is organized as follows. We first briefly overview the relevant works in Section 5.2. Next we introduce the proposed method and the used model UQnet in Section 5.3 and 5.4. Section 5.5 shows the experimental results and the corresponding analysis. Finally, we draw our conclusions and give several research directions in the last section.

5.2 Overview

In this section, we will present a short overview of related studies in the literature and bring out the distinctiveness of the proposed method.

There are different approaches to predicting vehicle trajectories. One popular approach is to apply simple physics-model-based methods in combination with (e.g. Kalman filtering, Prevost et al. (2007)), in which typical assumptions such as constant yaw rate and/or acceleration are made to simplify the problem with prior co-variance structures that model unobservable deviations (Houenou et al., 2013; Ammoun & Nashashibi,

2009). These models describe the projected movements in explanatory terms but typically suffer from limitations in prediction horizons and accuracy (Lefèvre et al., 2014).

On the other hand, recently-popular data-driven models, especially deep-learning models, directly assimilate the collected data and show better performances in many scenarios. For example, Alahi et al. (2016) proposed to model interactions between pedestrians with social pooling in a "black-box" way. Here "social" represents the common scene rules and the influence of neighbour agents. Further studies improve the social mechanism by inducing generative adversarial networks (Gupta et al., 2018) or considering multi-agent dynamic features (Zhao et al., 2019). Some studies abstract agents as nodes, treat pairwise influences as edges, and use graph neural networks to model the interaction, e.g. (Vemula et al., 2018; Ma et al., 2019). Huang et al. (2019a) further seek "post-hoc" interpretation from learnt graph attention weights, saying higher attention means larger influence. *Trajectron*++ (Salzmann et al., 2020) developed a modular and graph-structured recurrent model as the encoding channel to generate multi-modal predictions from incorporated agent dynamics. VectorNet (Gao et al., 2020) proposed an alternative vector-based lightweight representation to reduce model complexity. It has been widely used in many following works (e.g. Liang et al. (2020)).

However, the models above highly rely on independent identical distribution (i.i.d.) assumption. Compared to physical models, correlation-based DNNs are fragile when being deployed in new environments because of causal confusion (De Haan et al., 2019), which means the model learns spurious correlations (overfits the problem). Recently some studies tried to improve DNN's generalizability by inducing more robust causal relationships. Chen et al. (2021a) proposed to use counterfactual analysis to alleviate the spurious (specious) correlation of environmental bias. The plug-in module consistently improves the performance of baseline models. Hu et al. (2021) constructed a structural causal model (SCM) to learn invariant features across different scenarios, the so-called causal-based time series domain generalization (CTSDG) model. In Liu et al. (2022), the input is decomposed into invariant variables, style confounders, and spurious features in hidden space. By training the model to suppress spurious features, the robustness was significantly improved. Kumor et al. (2021) came up with a theoretical criterion that determines the feasibility of learning a demonstrator's trajectory under sensor shift from the perspective of causal models. In brief, combining DNNs with a causal model is expected to increase both the robustness and the transparency of data-driven trajectory prediction models.

Additional to limits in generalizability, uncertainty is another important topic in motion forecasting, but it has not drawn as much attention as developing prediction models. Arnez et al. (2020) reviewed different UQ approaches that potentially can be used for autonomous vehicle applications. For example, Makansi et al. (2019) proposed a sampling-fitting two-stage strategy to learn the mixture Gaussian distribution of a vehicle's future position, which can naturally represent aleatoric uncertainty. In Pang et al. (2021), Bayesian neural network is used to quantify the total uncertainty brought by weather for trajectory forecasting. Djuric et al. (2020) directly give estimates of both aleatoric and epistemic uncertainties measured by marginal variance for each prediction step in vehicle trajectory forecasting. The model was also deployed in a real-world autonomous vehicle test. Tang et al. (2022) further considers the prediction uncertainty for safer decision-making and motion planning in high-risk scenarios. The

authors assume that the predicted position obeys a Gaussian distribution and use Deep Ensembles to explicitly quantify both aleatoric and epistemic uncertainty measured by the covariance matrix. The studies above pioneered this domain. They are all parametric approaches, which means the prior form of spatial distribution must be closed-form (such as Gaussian, Laplace, or Mixture Density). However, this assumption does not always hold. For example, vehicles tend to drive along the lane centerlines so the form of its spatial distribution highly depends on the road layouts and lane connectivity. Recently, heatmap-based models pave the path to non-parametric UQ. Based on the vectorized representation, DenseTNT (Gu et al., 2021) converts the 2D regression problem to a classification task and directly learns the probability that the target vehicle will appear in each small mesh unit. This method is anchor-free and it can approximate any 2D distribution. GoHome (Gilles et al., 2021), similarly, generates lane-centralized heatmap from high-definite maps and trajectories. The authors explicitly point out that this approach can be used to quantify aleatoric and epistemic uncertainty. In summary, we believe that there are already enough tools towards more reasonable UQ in motion prediction.

From the overview above, we find there are two key gaps in the literature. First, most papers use a variance/covariance matrix to quantify uncertainty, which is not reasonable for a 2D distribution with multiple local maxima (multi-modal). Second, domain adaptation and uncertainty quantification are regarded as two separate topics. We argue that they are highly entangled, especially for quantifying epistemic uncertainty. Our method tries to overcome the first drawback by using entropy-based metrics for heatmap outputs and to bridge the second gap by inducing causation in deep-learning models.

5.3 Method

This section describes the proposed method in detail. We will sequentially introduce the 2D histogram and Deep-Ensemble-based uncertainty quantification method and how to combine this technique with Granger causality.

5.3.1 Uncertainty quantification

Suppose that the input and the output of a model are two random variables, respectively noted as $\boldsymbol{X}$ and $\boldsymbol{Y}$. The training dataset contains N collected input-output samples, noted as $\mathscr{D}_N = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$. $\mathscr{D}_N$ is used to train a prediction model with inner random variables $\boldsymbol{\Theta} \sim p_\Theta(\boldsymbol{\theta})$, noted as $p_Y(\boldsymbol{y}) = \mathscr{M}(\boldsymbol{x}, \boldsymbol{\theta})$. Now given a test input $\boldsymbol{x}^*$, we aim to quantify the uncertainty of the output distribution $p_Y(\boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x}^*)$ (briefly denoted as $p(\boldsymbol{y}^*)$ and the specific random variable is noted as $\boldsymbol{Y}^*$ from now on). In this study we use *differential entropy* $H(\boldsymbol{Y}^*)$, which is a scalar metric, to represent the total prediction uncertainty. It is defined by:

$$H(\boldsymbol{Y}^*) = -\int_{\mathscr{Y}} p(\boldsymbol{y}^*) \ln p(\boldsymbol{y}^*) d\boldsymbol{y}^* \tag{5.1}$$

For the multi-modal distributions shown in Fig.5.1, the two local maxima are separated by road boundaries so apparently both cases have the same entropy. This metric can exclude the influence of road layouts. The unit of differential entropy is "nats". Amini et al. Amini et al. (2020) derive that the total entropy can be decomposed into the following terms:

$$H(\boldsymbol{Y}^*) = \underbrace{\mathbb{E}_{p_\Theta(\boldsymbol{\theta})}[H(\boldsymbol{Y}^*|\Theta=\boldsymbol{\theta})]}_{\text{aleatoric}} + \underbrace{\mathbb{E}_{p_\Theta(\boldsymbol{\theta})}[D_{KL}(p(\boldsymbol{y}^*|\Theta=\boldsymbol{\theta}) \mid\mid p(\boldsymbol{y}^*))]}_{\text{epistemic}} \tag{5.2}$$

Here $p(\boldsymbol{y}^*)$ is the posterior distribution marginalized by $\boldsymbol{\theta}$, given as follows:

$$p(\boldsymbol{y}^*) = \mathbb{E}_{p_\Theta(\boldsymbol{\theta})}[p(\boldsymbol{y}^*|\Theta=\boldsymbol{\theta})] \tag{5.3}$$

And $D_{KL}(p||q)$ is the *Kullback-Leibler* divergence that measures the directed "distance" from distribution p to q, which is non-negative. The definition is:

$$D_{KL}(p(\boldsymbol{y}^*|\Theta=\boldsymbol{\theta}) \mid\mid p(\boldsymbol{y}^*)) = \int_{\mathscr{Y}} p(\boldsymbol{y}^*|\Theta=\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{y}^*|\Theta=\boldsymbol{\theta})}{p(\boldsymbol{y}^*)} d\boldsymbol{y}^* \tag{5.4}$$

So Eq.(5.2) can be interpreted as follows. The first aleatoric term is the *conditional entropy* that measures the average entropy (uncertainty) across an ensemble of distributions. While the second epistemic term is the *mutual information* that measures the average distance from each distribution to the average (posterior) distribution, which reflects how diverse the ensemble of distributions are. During inference, if the model has already seen similar inputs enough times in training, then the output distributions will be consistent for different model parameters $\boldsymbol{\theta}$. If the input is a rare or new case, different $\boldsymbol{\theta}$ will give diverse distributions because it contributes little or even zero to the training loss. This can be illustrated by the simple 1D example in Fig.5.2. Different models will give consistent predictions at $x = 0$ but diverse outputs at $x = -3.5$. Here we also point out one special property: Different from covariance measures, the epistemic term in (5.2) is **scale-independent**. It is independent of the mean or variance of each distribution, which makes it naturally convenient to represent "rareness" objectively.

The discussion above shows the principle of UQ, that is learning $p_\Theta(\boldsymbol{\theta})$ that gives consistent predictions when enough samples are provided but diverse outputs for rare samples. The existing methods in the literature include Bayesian neural networks, Monte-Carlo dropout (Kendall & Gal, 2017), and Deep Ensemble (DE) (Lakshminarayanan et al., 2017), etc. Among these UQ methods, the deep ensemble is still the most robust approach and it is state-of-the-art in many uncertainty quantification tasks (e.g. monocular depth estimation, Poggi et al. (2020)). An ensemble of randomly-initialized models are trained independently. During inference, the input is passed into the trained models in parallel to get their corresponding output distributions. Fort et al. (2019) interpret the advantage of DE from the perspective of loss landscape. The authors show that deep ensembles can explore different local minima of the loss while other methods usually fall into only one. In this study, we also choose the deep ensemble strategy.

Most studies in the literature assume that the prior form of the output distribution can

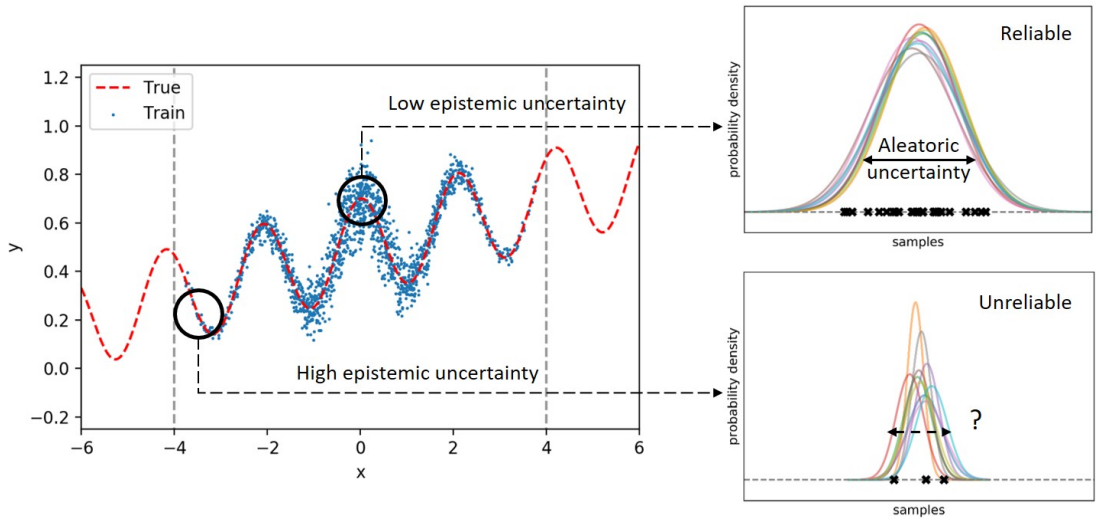


Figure 5.2: A 1D example of aleatoric and epistemic uncertainty. Here we learn $y = f(x)$ from noisy data. Both the magnitude of noise and the number of samples are higher in the middle but decay with $|x|$.

be described by a set of parameters $\boldsymbol{\gamma} = \mathscr{M}(\boldsymbol{x}^*, \boldsymbol{\theta})$, such as mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ for Gaussian (Makansi et al., 2019), or concentration $\boldsymbol{\alpha}$ for Dirichlet distribution, etc. However, the output distribution can be much more complex due to arbitrary road layouts. To alleviate this restriction, we directly approximate a 2D spatial distribution by a mesh-grid heatmap $\hat{\boldsymbol{Y}}_{h\times w} = \mathscr{M}(\boldsymbol{x}^*, \boldsymbol{\theta})$. The value at a pixel $\hat{Y}_{i,j}$ represents the probability that the vehicle will be present within that specific square. So the regression problem is converted to a dense multi-class classification problem. An ensemble of N models give N 2D distributions $\{\hat{\boldsymbol{Y}}_n\}_{n=1}^N$ for one input $\boldsymbol{x}^*$. Then the posterior distribution $p(\boldsymbol{y}^*)$ can be approximated by the element-wise average distribution:

$$p(\boldsymbol{y}^*) \approx \boldsymbol{Y}_m = \frac{1}{N}\sum_{n=1}^{N} \hat{\boldsymbol{Y}}_n \tag{5.5}$$

And the conditional differential entropy of each heatmap can be approximated by (here x and y are coordinates):

$$H_n = -\int_{\mathscr{X},\mathscr{Y}} \hat{Y}_n(x,y) \ln \hat{Y}_n(x,y) dxdy \tag{5.6}$$

The KL divergence between each distribution to the posterior distribution can be approximated by:

$$D_{KL}(\hat{\boldsymbol{Y}}_n || \boldsymbol{Y}_m) = \int_{\mathscr{S}} \hat{Y}_n(s) \ln \frac{\hat{Y}_n(s)}{Y_m(s)} ds \tag{5.7}$$

Where $\mathscr{S}$ is the non-zero support set of $\boldsymbol{Y}_m$ (to avoid 0 division). Therefore, according to (5.2), the aleatoric term can be approximated by the average of conditional differential entropy:

$$\mathbb{E}_{p_\Theta(\boldsymbol{\theta})}[H(\boldsymbol{Y}^* | \Theta = \boldsymbol{\theta})] \approx \frac{1}{N}\sum_{n=1}^{N} H_n \tag{5.8}$$

And the epistemic term can be approximated by the average of KL divergence:

$$\mathbb{E}_{p_\Theta(\boldsymbol{\theta})}[D_{KL}(p(\boldsymbol{y}^*|\Theta=\boldsymbol{\theta}) \mid\mid p(\boldsymbol{y}^*))] \approx \frac{1}{N}\sum_{n=1}^{N} D_{KL}(\boldsymbol{Y}_n||\boldsymbol{Y}_m) \tag{5.9}$$

In summary, (5.8) and (5.9) give the estimates of aleatoric and epistemic uncertainty from deep ensembles respectively. For the 2D integrals in (5.6) and (5.7), we use *Simpson's rule* to calculate them numerically. We refer the readers to Cruz-Uribe & Neugebauer (2002) for the approximation error bound of this method. Besides the numerical error of integral, the accuracy of epistemic term estimation (5.9) largely depends on the size of deep ensembles N. If N is not large enough, the absolute error can be high due to the bad approximation of (5.5). However, if we are only interested in the relative rareness of samples, this approach works well with a smaller ensemble size.

5.3.2 Causal regularization

The UQ method introduced above completely ignores what assumptions are used in the model. Considering $\boldsymbol{X}$ and $\boldsymbol{Y}$ as two *correlated* random variables works well when the identical independent distribution assumption holds but becomes fragile for out-of-distribution (OOD) samples. This is especially problematic for model generalization and epistemic uncertainty quantification. Specific to microscopic traffic modelling, the input $\boldsymbol{X}$ can be decomposed into three variables:

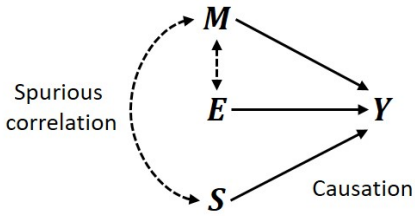
(1) $\boldsymbol{M}$: Maps information, including the geometries of lanes and their connectivity, traffic rules, etc.

(2) $\boldsymbol{E}$: The observed trajectory of the target vehicle.

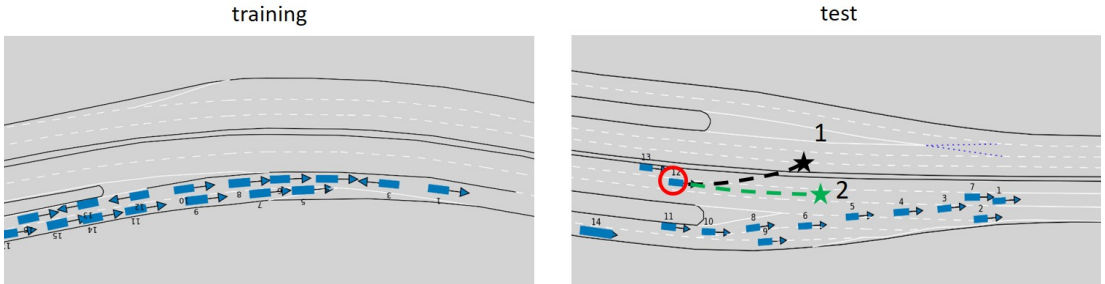
(3) $\boldsymbol{S}$: The observed trajectories of the surrounding agents (including vehicles and pedestrians).

One of the most significant spurious correlations is between $\boldsymbol{M}$ and $\boldsymbol{S}$ (shown in Fig. 5.3a). For example, in the training scenario shown in Fig.5.3b, a driver simply follows the surrounding vehicles OR follows the lane centerline work perfectly fine most of the time. This may cause severe over-fitting to the scenario biases. When deploying the trained model in a new merging where two roads are not parallel, this correlation does not exist anymore. For the specific case on the right figure, we observed that different deep learning models will give two separate groups of predictions: follow the other vehicles (off-road black star) or follow the lane (in-road green star). It means that the DE will unreasonably identify this case as "rare". In Bahari et al. (2022), the authors also show that most state-of-the-art motion forecasting models can be confused by new road layouts and give off-road outputs.

Inducing *causation* is one plausible way to improve generalizability. Because $\boldsymbol{M}$, $\boldsymbol{E}$, and $\boldsymbol{S}$ are given for each specific prediction, we formulate the relationship as *Granger causality* (Granger, 1980). We say the variable "Granger-causes" $\boldsymbol{Y}$ if involving this



(a)



(b)

Figure 5.3: (a) The causal model of trajectory prediction. Solid lines are causation and dotted lines are spurious correlations. (b) In the training set scenario, the driving direction and the average traffic flow are highly correlated. But the correlation does not hold in the test scenario so a correlation-based model may fail.

variable can improve the prediction accuracy of $\boldsymbol{Y}$ (or equivalently reducing the uncertainty of $\boldsymbol{Y}$). Maps information and the past trajectory of the target vehicle must be given otherwise we can predict nothing. So here we only consider the causal effect of $\boldsymbol{S}$. We induce an additional regularization net which only takes $\boldsymbol{M}$ and $\boldsymbol{E}$ as inputs, noted as $\widetilde{\boldsymbol{Y}} = \mathscr{M}_{\text{rg}}(\boldsymbol{M},\boldsymbol{E}|\boldsymbol{M})$. The output of the normal predictor net (including $\boldsymbol{S}$) is noted as $\hat{\boldsymbol{Y}}$. The causal effect of inducing $\boldsymbol{S}$ should make the output distribution (1) equally or more concentrated and (2) stay within the drivable area given by excluding surrounding agents. Mathematically we can quantify and punish the causal effect by:

$$\arg\min L_1(\hat{\boldsymbol{Y}}_{01}, \min(\hat{\boldsymbol{Y}}_{01}, \widetilde{\boldsymbol{Y}}_{01})) \tag{5.10}$$

L_1 is L1 norm, and the subscript "01" means re-normalizing the initial distribution linearly between 0 and 1. min here is the element-wise minimum. The objective function in (5.10) measures how much probability of $\hat{\boldsymbol{Y}}$ is "leaked" from $\widetilde{\boldsymbol{Y}}$.

Another problem is that the training dataset is not balanced. The "leakage" is nearly 0 for most training samples. Only some sparse driving behaviours, such as lane-changing, U-turning, or waiting before an intersection area may break the correlation between $\boldsymbol{S}$ and $\boldsymbol{M}$. We need to increase the contributions of these samples to the total loss. Assume that the loss function between the ground-truth and the prediction is $L(\hat{\boldsymbol{Y}},\boldsymbol{Y})$. We apply a prediction-dependent weight for each specific sample. The weight increases with the effect given in (5.10). So the total loss function is constructed as

follows (λ is a positive constant):

$$L_{\text{total}}(\widetilde{\boldsymbol{Y}},\hat{\boldsymbol{Y}},\boldsymbol{Y}) = (1+\underbrace{\lambda L_1(\hat{\boldsymbol{Y}}_{01},\min(\hat{\boldsymbol{Y}}_{01},\widetilde{\boldsymbol{Y}}_{01}))}_{\text{regularization}})\ \underbrace{L(\hat{\boldsymbol{Y}},\boldsymbol{Y})}_{\text{predictor loss}} + \underbrace{L(\widetilde{\boldsymbol{Y}},\boldsymbol{Y})}_{\text{regularization loss}} \tag{5.11}$$

Now we give an intuitive explanation of the regularization term. $L(\widetilde{\boldsymbol{Y}},\boldsymbol{Y})$ trains the regularization net independently. Without information about surrounding agents, the regularization net is encouraged to explore the "reachable area" $\widetilde{\boldsymbol{Y}}$. If the output of the predictor is within this reachable area, then the regularization term (5.10) is 0, which means this is a normal case. When $\hat{\boldsymbol{Y}}$ is partly or completely out of the reachable area, the regularization term will assign a higher weight to $L(\hat{\boldsymbol{Y}},\boldsymbol{Y})$. In this way, we expect that the model can focus more on those sparse cases and better learn the influences from maps and surrounding vehicles.

In summary, the proposed method employs heatmap-based deep ensembles to quantify predictive uncertainty. A regularization net and a causation-based regularization term are added to enhance the model's adaptability. Next, we will present the structure of the proposed model.

5.4 UQnet model

Training an ensemble of deep-learning-based motion forecasting models is time-consuming. Running the inference also requires higher memory and a longer time. Therefore, fast and lightweight models are preferred. In this study, we use the representation proposed in VectorNet (Gao et al., 2020). VectorNet abstracts map elements (such as lane centerlines, crosswalks, etc.) and agents' trajectories into splines. Each spline is represented by a series of end-to-head connected vectors. Compared with other representation methods, such as rasterized images, this vector-based representation significantly reduces the input size but preserves the most important factors in motion forecasting.

The backbone of our predictor is constructed based on DenseTNT (Gu et al., 2021). Its structure is shown in Fig.5.4. The predictor and the counterfactual net share the same sub-graph neural networks (encoder) to extract features from each spline separately. Then we added a laneGCN-like (Liang et al., 2020) graph attention module (*Lane-Attention*) to explicitly learn the connectivity of roads. The basic idea is that upstream/downstream, left/right adjacent lanes have different influences on the central lane. We use an attention-based module instead of the parameter-fixed laneGCN module proposed in Liang et al. (2020) so the influence of one lane on another depends on their own features as well. Meanwhile, a multi-layer perceptron (MLP) is used to convert each coordinate (x_t, y_t) to hidden representation $\boldsymbol{H}_C$. The extracted hidden representations of maps elements ($\boldsymbol{H}_M$), coordinates ($\boldsymbol{H}_C$), and trajectories ($\boldsymbol{H}_E$ for target vehicle, $\boldsymbol{H}_S$ for surrounding agents) are passed into the predictor and the regularization net.

The regularization net decoder was constructed based on the proposed causal relationship. A cross-attention layer (*Map2Target*) passes map information to the target vehicle. Then another cross attention layer (*Feature2Pos*) generates the probability

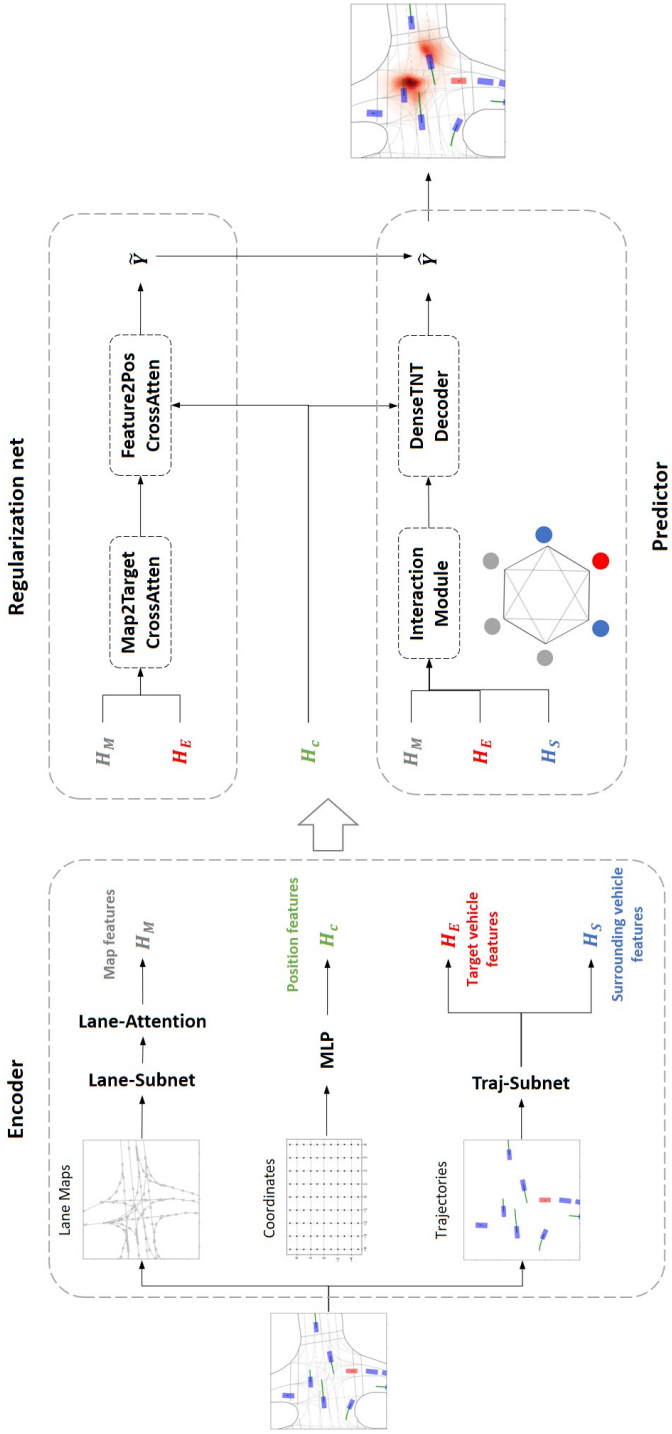


Figure 5.4: Model structure of UQnet

density of each location from the concatenated output of *Map2Target* layer. In the predictor decoder, stacked multi-head self-attention graph layers are employed to learn the interactions among these splines. At least two layers are needed to learn their state-dependent dynamic relationships. Then the same decoder used in DenseTNT (Gu et al., 2021) generates the prediction heatmap $\hat{\boldsymbol{Y}}$. For more details, we refer to the appendix C and the open source code [1]. For convenience, we name our model as *Uncertainty*

[1] https://github.com/RomainLITUD/UQnet-arxiv/

Quantification networks (**UQnet**).

UQnet uses *Focal loss* (Lin et al., 2017) as $L()$ to measure the error between the ground-truth and the predicted distribution. It is defined as follows:

$$L_{\text{fl}}(\hat{\boldsymbol{Y}},\boldsymbol{Y}) = -\frac{1}{P}\sum_{p}(Y_p - \hat{Y}_p)^2 f(Y_p,\hat{Y}_p) \tag{5.12}$$

$$f(Y_p,\hat{Y}_p) = \begin{cases} \ln \hat{Y}_p & \text{if } Y_p = 1 \\ (1-Y_p)^4 \ln(1-\hat{Y}_p) & \text{else} \end{cases} \tag{5.13}$$

Focal loss can well address imbalanced samples. The ground-truth label is constructed by the objection detection technique proposed in Zhou et al. (2019) and also used in GoHome (Gilles et al., 2021). For example, if the ground-truth location is (x_g, y_g). We add an extra Gaussian noise $\boldsymbol{\varepsilon}$. Its mean is the ground-truth location and the covariance matrix has the form $\sigma_\varepsilon^2 \boldsymbol{I}$. For every pixel $Y_{i,j}$ of the heatmap with centre coordinate (x_c, y_c), its probability is determined as follows:

$$Y_{i,j} = \frac{1}{2\pi\sigma_\varepsilon^2}\exp[-\frac{(x_c - x_g)^2 + (y_c - y_g)^2}{2\sigma_\varepsilon^2}] \tag{5.14}$$

This method increases the number of non-zero pixels in the heatmap and thus can significantly accelerate the training process. However, it also induces extra errors. In this study, we choose $\sigma_\varepsilon = 0.7\,\text{m}$ so the Gaussian roughly covers the average length of a car. It also means that we manually increase the mean-square-error (MSE) by $0.49\,\text{m}^2$ and the lower bound of differential entropy is the entropy of this white noise, around 2.12. We need to correct this in uncertainty estimation. Replacing L in (5.11) by (5.12), we have the total loss function to train UQnet. The value of λ depends on the size of the output heatmap and its spatial resolution. In inference, the regularization net can be removed and we only keep the predictor. The lanescore module in DenseTNT decoder is also used to increase the convergence speed, but we do not consider lanescore in inference or uncertainty quantification.

5.5 Evaluation

In this section, the proposed method will be evaluated on the open INTERACTION (Zhan et al., 2019) dataset. INTERACTION collects trajectories of road agents by drones in diverse urban traffic scenarios and high-definite maps are also provided. In the INTERPRET single-agent track prediction challenge, all agents' trajectories in the past 1 s are provided to predict the track of the target vehicle in the next 3 s. The data providers split all these cases into three groups. The training set contains 47584 cases in 12 scenarios. The validation set has 11794 cases from the same scenarios. The test set has 22644 cases and around 30% of them are collected from new scenarios to measure the generalizability of the model. For fairness, labels of the test set are not provided.

For each case, all trajectories of agents and lane splines are centred to the target vehicle's current position and re-oriented according to its driving direction (y-axis points at

the yaw angle). Each lane's centerline is evenly split into 5 head-to-end vectors along the driving direction. Each centerline vector has 8 features, $\boldsymbol{f}_i = [\boldsymbol{x_s}, \boldsymbol{x_e}, j, c, l, w]$. The first two are positions of start and end points. The integer j is the order of the vector. We also incorporate lane-level features. $c \in \{0,1\}$ represents whether this lane intersects with another non-connected lane. w is 1/2 width of the lane, ranging from 1.5 m to 4.5 m and l is the total length of the lane. For trajectory vectors we use the similar representation $\boldsymbol{f}_a = [\boldsymbol{x_s}, \boldsymbol{x_e}, \boldsymbol{v}, \text{agent\_type}, t]$. $\boldsymbol{v}$ is the average speed between 2 sequential timestamps. The integer agent type can be a vehicle (1) or pedestrian/cyclist (-1) and t is the timestamp. From the description above we can construct the input for UQnet. For example, if the current case has n_m lanes and n_a agents, Then the maps input is $\boldsymbol{M} \in \mathrm{R}^{n_m \times 5 \times 8}$ and the trajectories input is $\boldsymbol{T} \in \mathrm{R}^{n_a \times 9 \times 8}$. To generate the heatmap, we consider a rectangle area around the target vehicle that covers $y \in [-12\,\mathrm{m}, 75\,\mathrm{m}]$ and $x \in [-23\,\mathrm{m}, 23\,\mathrm{m}]$. In training, we set the spatial resolution to 1 m so the coordinate input has the shape $(46 \times 87, 2)$. During training and validation, the target vehicle is randomly selected among all vehicles that have complete 4 s records.

UQnet learns the spatial distribution of the last position after 3 s. We can use different sampling strategies to get predicted positions from the heatmap. The models on the leader board are ranked by Missing Rate (MR) so here we use a naive local-maximum sampling strategy (Gilles et al., 2021) to greedily generate the most possible k final positions ($k = 6$ for the INTERPRET challenge). MR is calculated as follows. If the predicted final position of the target agent is out of a given lateral or longitudinal area of the ground truth, it is "missed". MR measures how many percentage of predictions are missed. The lateral threshold is 1 m and the longitudinal threshold is a piece-wise function depending on the velocity of the target agent at the current moment:

$$\mathrm{th}(v) = \begin{cases} 1 & v < 1.4\,\mathrm{m\,s^{-1}} \\ 1 + \frac{v-1.4}{11-1.4} & 1.4\,\mathrm{m\,s^{-1}} \leq v \leq 11\,\mathrm{m\,s^{-1}} \\ 2 & v > 11\,\mathrm{m\,s^{-1}} \end{cases}$$

An MLP with 1 hidden layer completes the trajectory from the predicted last position. For uncertainty quantification, according to the empirical suggestion given in Kendall & Gal (2017), 7 randomly-initialized UQnets are trained in parallel (5-10 models are proper choices). Then the method proposed in section 5.1 is used to quantify epistemic and aleatoric uncertainty.

Table.5.1 compares the MR of UQnet with other models on the leader board [2]. Here UQnet is the performance of the first model (not the best one). UQnet (predictor only) does not use the causal regularization term. We see that the generalizability MR is significantly improved from the previous-best 11.07% to 6.86% (an improvement of 38%) and the overall MR is reduced from 4.91% to 3.64% (an improvement of 26%). UQnet's regular MR reaches the average level. If the causal regularization is not used, the regular MR is slightly better but the generalizability MR drops a lot. These results support that inducing the regularization term indeed largely improves the generalizability of the predictor. Next, we will focus on analysing the estimated uncertainty.

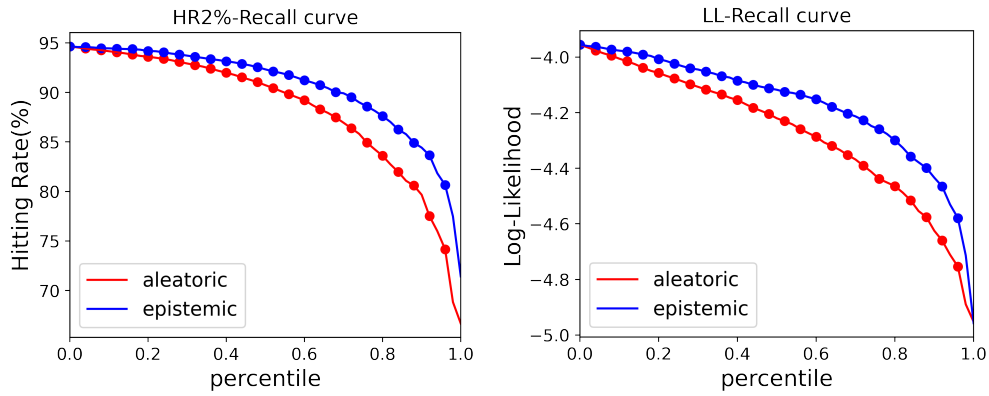
[2] http://challenge.interaction-dataset.com/leader-board

Table 5.1: Comparison with other models (Missing Rate, %)

Model	Regular MR	Generalizability MR	Overall MR
UQnet	1.96	**6.86**	**3.64**
UQnet (predictor only)	1.84	10.31	4.75
GoHOMEGilles et al. (2021)	**1.37**	11.68	4.91
Multimodal Transformer	2.00	11.07	5.11
HDGTJia et al. (2022)	1.42	13.47	5.56
DenseTNTGu et al. (2021)	2.80	12.00	5.96

5.5.1 Precision-recall analysis

We first show the precision-recall curve for both aleatoric and epistemic uncertainty on the validation set. Here the precision is represented by hitting rate HR (HR=1-MR) and log-likelihood (LL). We only sample 2 positions from the heatmap to calculate $HR_2\%$. Fig.5.5 shows how precision improves by preserving those cases with aleatoric or epistemic uncertainty lower than a specific threshold. For example, in Fig.5.5a, a point on the aleatoric uncertainty curve (the red line) at the 0.2 percentile depicts the hitting rate, considering the samples with the lowest 20% aleatoric uncertainty *only*. We see that the estimated uncertainty is negatively correlated with precision, which means the estimated uncertainty can indeed reflect the prediction confidence (Kendall & Gal, 2017).



(a) (b)

Figure 5.5: Precision-Recall curve on the test set: (a) Hitting rate and (b) log-likelihood

Next, we will analyse aleatoric and epistemic uncertainty in detail. To facilitate the discussion below, now we split the validation set and the test set into the following three groups:

1 **Validation set**: All samples in the validation set are from exactly the same scenarios of the training set, including mergings, intersections, and roundabouts. It serves as a baseline.

2 **In-Distribution (ID) test cases**: Those test cases that are collected from the same scenarios as the training set and validation set.

3 **Out-of-Distribution (OOD) test cases**: Those test cases that are collected from new scenarios, including completely new types of scenarios, e.g. diverging.

5.5.2 Aleatoric uncertainty and predictability

The distributions of estimated aleatoric uncertainty with added Gaussian noise for the three subgroups are shown in Fig.5.6a. They have similar shapes that are highly concentrated around 2.5 nats and the value is always higher than $H(\boldsymbol{\varepsilon}) = 2.12$ due to the added noise. We must rectify this estimation error. Unfortunately, although the added noise $\boldsymbol{\varepsilon}$ is independent of the prediction, differential entropy is not linearly additive. Note the rectified random variable of position as $\hat{\boldsymbol{Y}}_r = \hat{\boldsymbol{Y}} - \boldsymbol{\varepsilon}$, then $H(\hat{\boldsymbol{Y}}_r) \neq H(\hat{\boldsymbol{Y}}) - H(\boldsymbol{\varepsilon})$. We must find a roundabout to correct this. Aleatoric uncertainty is closely related to the concept of the "limit of predictability", which can be represented by the lower bound of accuracy for any model. In our previous study we show that the aleatoric uncertainty measured by conditional entropy gives the lower bound of negative-log-likelihood (NLL) and from that, we can further derive the limit of MSE (Li et al., 2022a):

$$\mathrm{MSE}^2 = \det\boldsymbol{\Sigma} \geq \frac{1}{(2\pi e)^2} e^{2H} \tag{5.15}$$

MSE is linear and additive for independent random variables. So we can use (5.15) to derive the lower bound of MSE for the rectified prediction:

$$\mathrm{MSE}_{lb}(\hat{\boldsymbol{Y}}_r) = \mathrm{MSE}_{lb}(\hat{\boldsymbol{Y}}) - \sigma_\varepsilon^2 = \frac{e^{H(\hat{\boldsymbol{Y}})}}{2\pi e} - 0.49 \tag{5.16}$$

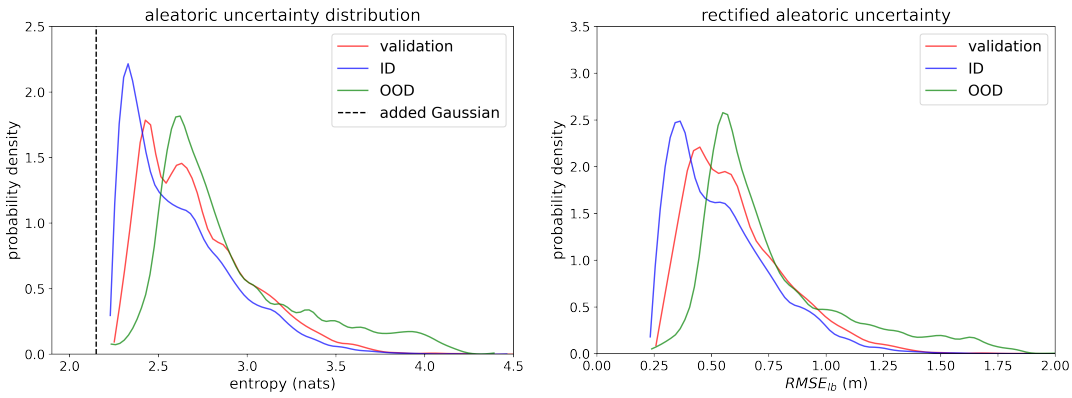
Now we can use the formula above to convert entropy measures to MSE_{lb} (noted as σ_{lb}^2), or its square root RMSE_{lb}, to represent rectified aleatoric uncertainty equivalently. The results are shown in Fig.5.6b. The peak now locates at around 0.5 m. It is necessary to clarify that the σ_{lb}^2 is different from prediction MSE. For the two cases shown in Fig.5.1, they have the same σ_{lb}^2 derived from the conditional entropy but (b)'s prediction MSE is apparently higher.

In most trajectory forecasting tasks, people prefer to use final-displacement-error (FDE), which is the euclidean distance. For given σ_{lb}^2, 2D symmetric Laplace distribution minimizes the L2 norm (Eltoft et al., 2006) (for 1 sampled prediction), noted as $p_L(x,y|\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \sigma_{lb}^2 \boldsymbol{I})$. However, multivariate Laplace distribution contains the modified Bessel function of the second kind. Its mean-absolute-deviation does not have a closed form. So we can only estimate the lower bound of FDE by numerically calculating the following integral:

$$\mathrm{FDE}_{lb} = \int \sqrt{x^2 + y^2} \cdot p_L(x,y|\mathbf{0}, \sigma_{lb}^2 \boldsymbol{I}) dxdy \tag{5.17}$$

The estimated lower bound of average FDE on the validation set is compared with the performances of some models in Fig.5.7. It measures the limit of predictability

from a different angle. However, we emphasize that the analysis of predictability is for a single prediction. In practice, we generally sample multiple predictions for safer motion planning. This strategy can significantly increase the coverage of outputs.



(a) (b)

Figure 5.6: Distributions of aleatoric uncertainty for three different groups. (a) Differential entropy metric with Gaussian noise; (b) Denoised rectified RMSE lower bound metric

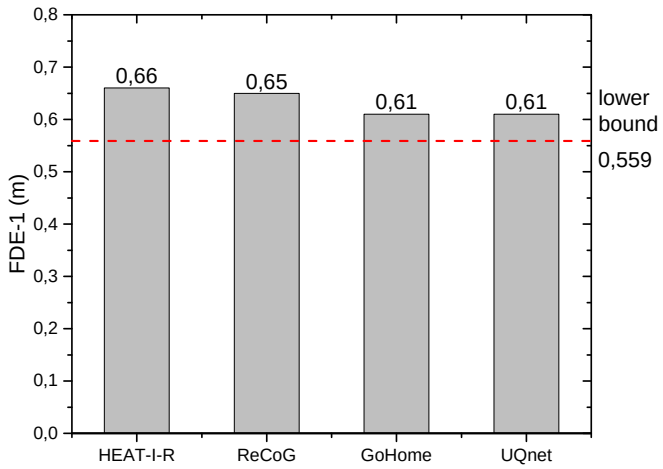
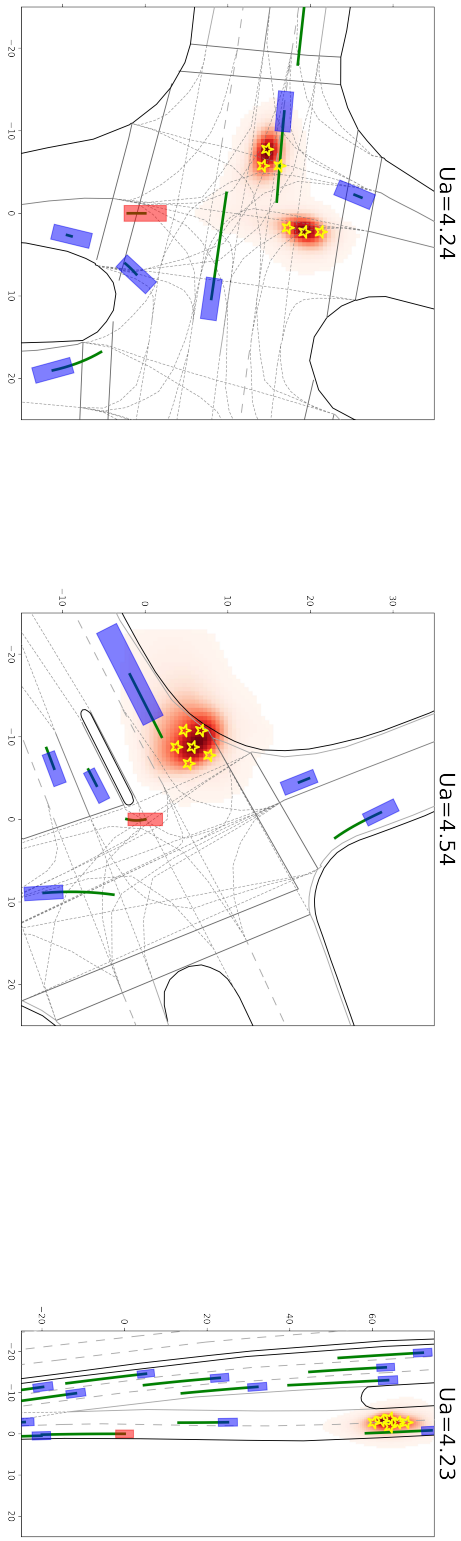


Figure 5.7: The estimated lower bound of FDE for one prediction and FDEs of HEAT-I-R Mo et al. (2021), ReCoG Mo et al. (2020) and GoHome Gilles et al. (2021). All metrics are evaluated on the validation set.

Fig.5.8 shows three cases that have low epistemic uncertainty but high aleatoric uncertainty. Recall that low epistemic uncertainty means that the predicted heatmap is reliable and these cases are common. They represent three different types of missing information. Case-(a) has strong bi-modality due to the lack of turn signal. In the real world, the driver is expected to flash the left turn signal or do nothing. One modality would disappear. The high aleatoric uncertainty is caused by the special bird-view of the drone. In case-(b), the target vehicle is predicted to do a U-turn following a large vehicle. The heterogeneity of this steering manoeuvre itself is high. It highly depends on the driver's unknown proficiency. In case-(c), the target vehicle is driving at a high speed ($79.2\,\mathrm{km\,h^{-1}}$) and the leading car on its adjacent lane is significantly slower



(a) (b) (c)

Figure 5.8: Three examples of low epistemic uncertainty but high aleatoric uncertainty. The red vehicle is the target vehicle and blue vehicles are surrounding vehicles. Green lines are their trajectories in the past 1 s. The heatmaps represent the spatial distribution of the target vehicle's position after 3 s. Yellow stars are the 6 most possible sampled positions. U_a gives aleatoric uncertainty value (nats).

(46.8 km h^{-1}). Their gross longitudinal distance headway is about 21.5 m. We expect that the target vehicle will approach or even overtake the left neighbour in the next 3 s. The interaction uncertainty is high. The same acceleration will always yield the same speed difference. However, the higher the initial speed is, the bigger the difference of position becomes. The aleatoric uncertainty of case-(a) can be reduced by adding turn signals, but adding sensors cannot help case-(b) and (c).

5.5.3 Epistemic uncertainty and rareness

Fig.5.9 presents the distributions of epistemic uncertainty. Different from Fig.5.6, OOD's distribution is significantly different from the other two groups. The proposed method identifies more samples from new diverging/off-ramp scenarios as "new cases". But even in OOD, most samples locate at the low-epistemic end.

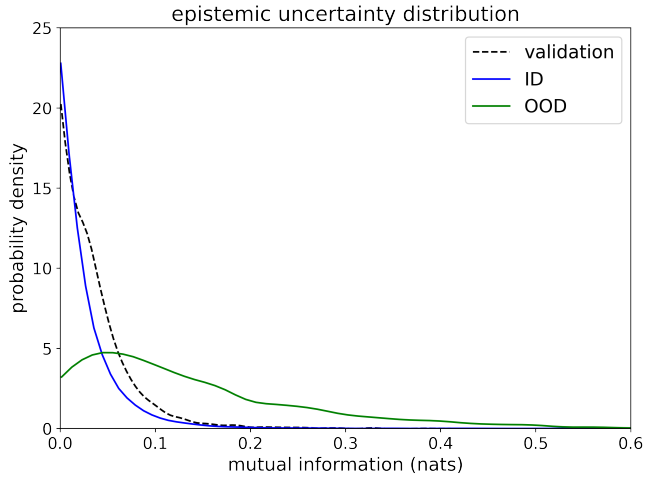
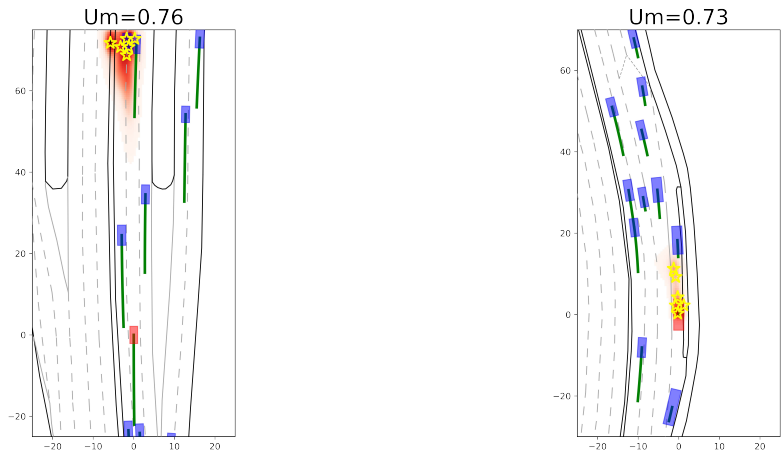


Figure 5.9: The distributions of epistemic uncertainty

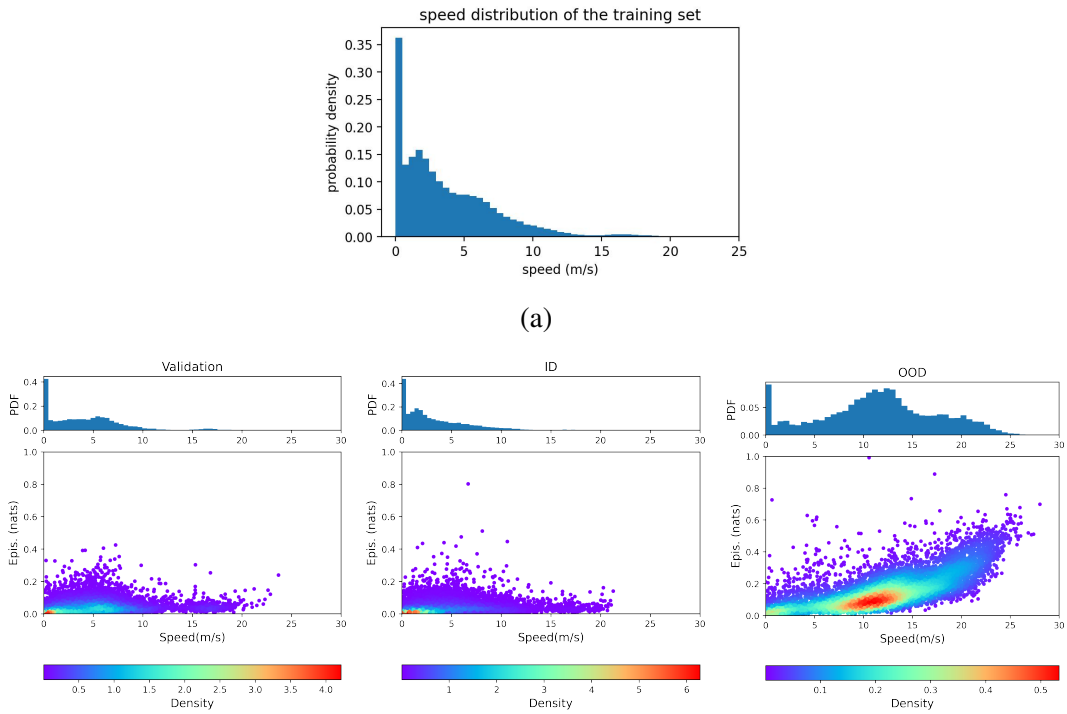
We found that most of these "rare cases" are from one specific scenario: the target vehicle is approaching the diverging decision point at a very high speed. Fig.5.10a shows one example. We see that the spatial distribution disperses in a larger area because different models in the ensemble give inconsistent predictions. Some predictions are even off-road. The result is reasonable because the training set does not contain any diverging/off-ramp scenario. Although our model can effectively adapt to same-type scenarios, it cannot be generalized to a completely new situation. Fig.5.10b shows another rare case. The target vehicle stops on the right-most lane but its leading and following vehicles are still moving. It is reasonable to infer that this is a rare abnormal case.

One may argue that, since intersections and roundabouts are composed of similar diverging and merging sub-structures, the model might also be applied to (on/off) ramps. However, the difference in operational speeds prohibits this generalisation. When the speed is low (congested), epistemic uncertainty is small because the surrounding vehicles' influences are more important than geometry (the map elements). The target vehicle's choices are largely restricted. However, when the speed is high, driving behaviours are completely different. Human drivers will tend to keep long-distance head-



(a) (b)

Figure 5.10: Two examples of high epistemic cases. U_m is the estimated epistemic uncertainty (nats).



(a)

(b) (c) (d)

Figure 5.11: (a) shows the distribution of the target vehicle's speed in the training set. (b) (c) and (d) show both the speed distribution and the scatter density plots of speed-epistemic uncertainty relationships for different groups

ways and make faster and more determined decisions to avoid collision (Toledo et al., 2009) because the braking distance is proportional to the square of speed. Fig.5.11 directly shows the relationship between the speed and epistemic uncertainty. The val-

idation set and ID groups have very similar speed distributions (like the training set). Most of the samples are low-speed cases. There are not many cases where the target vehicle's speed is higher than $10\,\mathrm{m\,s^{-1}}$. While the average speed of the OOD group is significantly higher. Speed and epistemic uncertainty are positively correlated for OOD. Especially, those cases with speed higher than $20\,\mathrm{m\,s^{-1}}$ have significantly higher epistemic uncertainty. INTERPRET challenge splits the training set, the in-distribution test set, and the generalizability test set by scenario types, but the difference in speed is ignored and not balanced. However, pure data-driven models are not likely to be generalizable from low-speed to high-speed cases because their driving behaviours are substantially different. For example, the experiments in Huang et al. (2018) show that the intra-driver heterogeneity of car-following behaviours under high-speed situations is significantly higher than in low-speed situations. When the speed is high, even the same driver shows very different behaviours in keeping distance headways in repeated runs. Adding such extra domain knowledge established by traffic researchers is the key to further improving the generalizability.

Below we summarize the major findings in this section as follows:

- UQnet has good prediction performance. Compared with other correlation-based models, it has better generalizability due to the added causal regularization.
- The proposed heatmap-based UQ method gives estimates of both aleatoric and epistemic uncertainties.
- In this context,
  - Aleatoric uncertainty can be understood as a (reasonable) limit of predictability for any motion forecasting model.
  - Epistemic uncertainty is representative of new or rare interaction cases in the test set.
- Our results demonstrate that one of the major obstacles of domain generalization for motion forecasting is properly encoding the speed-dependency of the (causal) relationships between variables.

5.6 Conclusion and perspective

In this paper, we proposed a novel non-parametric spatial uncertainty quantification method. UQnet ensembles can give accurate predictions and reasonable measures of aleatoric and epistemic uncertainty. From aleatoric uncertainty, we estimate the lower bound of final-displacement error, which can measure the limit of predictability for trajectories. On the other hand, epistemic uncertainty can quantitatively identify which cases have not been seen in training. The main difficulty to improve models' generalization capabilities lies in improved modelling speed-dependent driving behaviours.

We offer two more tentative conclusions. First, we observed that aleatoric uncertainty is inherently high in many situations and our results seem to suggest that current AI

models already harness the maximum value in commonly available trajectory data sets. In other words, using trajectory data and maps information only, more model sophistication is not likely to significantly improve the prediction performance for in-distribution situations. Combining and fusing more (different) data sources offers a more promising path to better predictions. For example, adding turn signals (blinkers) may effectively reduce much of the (aleatoric) interaction uncertainty. Second, our results support the idea that data-driven models should be combined with extra domain knowledge to gain better generalizability. A key example is the notion that operational speed matters for which of the available data (e.g. geometry versus surrounding vehicle kinematics) are most informative for motion prediction, and thus for the uncertainty associated with that prediction.

Finally, an important avenue for further research is the idea that epistemic uncertainty could be used as a tool to recognise rare events in training sets. Instead of collecting as much data as possible to improve generalisation, one could rather focus on collecting sufficiently representative and heterogeneous data sets, in combination with applying adaptive learning techniques. This enables models to exploit continuous and guided online learning.

Acknowledgements

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

Chapter 6

Conclusions and perspectives

This section presents the conclusions and perspectives. The answers to the research questions and the main findings will be presented in Section 6.1. Then the overall conclusion will be given in section 6.2. Sections 6.3 and 6.4 respectively discuss the implications to science and to practice. Outlooks and future research directions related to this thesis will also be discussed.

6.1 Key findings

First of all, we present the answer to each individual research question posed in Section 1.3. Under each question, we first give a short take-out answer and then elaborate on the detailed response.

Question 1: For macroscopic highway networks, how to build a deep-learning-based traffic forecasting model that can provide post-hoc, causation-like interpretations on spatial associations?

The key is combining the domain knowledge from traffic flow theory with the data-driven learning paradigm. A deep-learning model must adhere to physical constraints.

In chapter 2, we built such a "grey-box" deep-learning model that can give both accurate network-level traffic predictions and explainable spatial associations. Two key points are highlighted in this study. First, the structure of the proposed dynamic graph convolutional (DGC) module allows us to explicitly model the directional propagation of spilling-back stop-and-go waves in a highway network. Second, the hyperparameters of DGC must be carefully tuned. More specifically, we found that "deeper" (more layers) graph neural networks do not necessarily mean "better" in traffic forecasting. "Deeper" means that each link can retrieve the information from farther locations. Due to the repeating feature of stop-and-go waves in traffic congestion, such a wider range may mislead the model to learn spurious correlations between the current link and another location, even if that location physically cannot have any influence on the current link. Therefore, both the accuracy of new congestion patterns and the interpretability of the model will degrade. This is the so-called "causal confusion".

To avoid causal confusion, one has to manually set the receptive field. We theoretically discussed and experimentally demonstrated that there exists an optimal range of the receptive field, which is between the time interval times the speed of stop-and-go waves ($\Delta t \times v$) and the two times of it ($2\Delta t \times v$). A wider information reception will make the model confused about which stop-and-go wave causes the next step of congestion, while a smaller reception field definitely fails to see the incoming congestion. In this way, the designed DGC can learn causation-like influences instead of just correlations among a set of highway links. Experiments on real-world highway networks around Rotterdam and Amsterdam demonstrate that the model indeed successfully gives explainable spatial associations. The visualization shows that DGC behaves very similarly to solving an LWR-like partial differential equation. The role of a restricted receptive field here is the same as the well-known Courant–Friedrichs–Lewy (CFL) condition in the explicit finite-element method. When congestion happens, the model will identify whether the current position is a standing bottleneck (like at an on-ramp) or a simple corridor. For congestion bottlenecks, the model will focus more on the central link to predict whether the congestion bottleneck will continue. If the location is on a corridor, then the model will look ahead to the downstream direction for estimating whether the nearest stop-and-go wave can reach the current location in the short future.

In summary, both designing the structure of the model and tuning its hyper-parameters must consider the scientifically established traffic flow theory. Otherwise, unrestricted information flow hinders users from seeking interpretations. These findings are not

directly related to uncertainty quantification itself but the methods and models used in the following chapters must be established according to the proposed principles.

Question 2: Given a dataset that is large and representative enough, what are the model-free, theoretical lower bounds of predictive accuracy for probabilistic models and deterministic models respectively?

The conditional differential entropy gives the lower bound of negative-log-likelihood (NLL) for probabilistic forecasting models, and the variance of a Gaussian distribution that has the same entropy as the NLL gives the lower bound of mean-square-error (MSE) for deterministic forecasting models.

In chapter 3, we theoretically prove that the conditional differential entropy poses the limit of NLL for any models that predict the probability distribution of the output. Then the well-known discrete form of Fano's theorem was extended to a continuous form. The corresponding inequality gives the lower bound of MSE for deterministic forecasting models. These are "hard boundaries" for any models using the same dataset. We can use them to represent the average predictability of a traffic quantity before diving into model construction.

Question 3: How to directly estimate the spatial-temporal distribution of predictability of traffic speed before building any forecasting models?

The key is using the spatio-temporal properties of traffic dynamics to partition the dataset into a series of time-of-the-day and location-related subsets. After reducing the dimensionality, we can use a numerical entropy estimator and the theorems to estimate the limit of predictability.

We need to combine the spatio-temporal associations given in chapter 2 with the theorems proved in chapter 3. Traffic flow and traffic congestion on the highway have two special properties. Spatially they are localized due to the limited spreading speed of congestion and temporally they are cyclo-stationary because of the underlying fluctuated day-to-day demand patterns. Therefore, the entire dataset is partitioned based on the propagation of congestion and time of the day so the dimensionality can be dramatically reduced. Practically, we employed the k-p nearest neighbour (kpN) algorithm to estimate these lower bounds directly from data points. Experiments showed that the obtained limit of predictability seems reliable for both macroscopic production prediction and congestion prediction. The results reveal that many traffic forecasting models are approaching the limit. Investing in the collection of diverse data types and multi-modal data fusion is more valuable than developing more sophisticated modelling techniques. Further, our approach is able to find out the most uncertain time slots (peak hours) and locations on a highway network. These locations are generally on ramps or tunnels around Rotterdam.

Question 4: How to estimate the aleatoric and epistemic uncertainty of each specific prediction for highway networks?

Training an ensemble of probabilistic traffic forecasting models and both types of uncertainty for each prediction can be estimated from the corresponding ensemble of output distributions.

In chapter 4, we address the uncertainty encountered when deploying a traffic speed forecasting model and expanding the dataset continuously. The Deep Ensemble (DE) is employed to estimate both types of uncertainty. The basic idea is "voting" by randomly-initialized models. Diverse results mean high epistemic uncertainty. For each input, the DE gives a set of probability distributions. The average distribution can potentially represent the aleatoric uncertainty while the "inconsistency" of distributions (can be measured by both mutual information or the variance of mean values) gives an estimate of epistemic uncertainty. Experiments showed that Beta-prior is a good approximation on average and this approach can indeed give reliable uncertainty quantification. We also compared two metrics of uncertainty, variance and entropy. We conclude that the entropy metric is more suitable for quantifying the rareness of the input sample.

Question 5: If the answer to question 4 suggests the predictability of highway traffic patterns is limited. What explains this limited predictability?

Due to the lack of demand data and microscopic driving behaviours, the bi-modality of future traffic state is substantially unpredictable.

In chapter 4, we found that, although there is a significant dataset shift from the year 2019 to 2022 (before and after Covid-19), aleatoric uncertainty is still overwhelmingly higher than epistemic uncertainty and this irreducible inherent randomness increases rapidly with the prediction horizon. This result means that the predictability of macroscopic traffic state in real-world deployment is mainly restricted by the limited observable data types instead of imperfect modelling or the shortage of data (of the same type). Further, the proposed histogram-regression model directly demonstrates why the long-term predictability is low. Due to the capacity drop, traffic flow generally rapidly jumps between congested and free-flowing status and forms stop-and-go waves. The predicted distribution of speed is bi-modal. The true causation of this phenomenon lies in the very microscopic driving behaviours and traffic demand, which are impossible to be accurately predicted solely from speed and flow measurements.

Question 6: In microscopic trajectory forecasting, how to give reasonable spatial uncertainty measurements that can remove the influence of arbitrary road layout?

Using 2D-histogram to approximate the true distribution and using the entropy-based uncertainty metrics.

In chapter 5, we show that using the 2D histogram representation and the entropy-based metric together can give reasonable estimates of spatial uncertainty in motion prediction. They do not need any assumptions on the form of lane centerlines. Entropy-based uncertainty metrics are more meaningful than covariance-based metrics because they are scale-independent and can rule out the influence of road layout (like the angle of merging). It is consistent with the decision-making process in a naturalistic driving environment.

Question 7: In the urban driving environment, how to build a generalizable motion forecasting model that can give both reliable predictions and uncertainty estimation in new scenarios?

Inducing causal mechanism can significantly improve both the generalizability prediction accuracy and the reliability of epistemic uncertainty estimation. Adding extra knowledge on speed-dependent driving behaviours is indispensable for further improving the generalizability.

In chapter 5, we argue and demonstrate that exploring the causal effect instead of just correlation in a data-driven approach is the key to improving both the generalizability and the reliability of uncertainty quantification. The proposed UQnet model employs a causal regularization module to enhance its adaptability to new interaction scenarios. Experiments on the INTERPRET open challenge show that the causation-based module significantly improves the predictive accuracy in new test scenarios and thus the corresponding epistemic uncertainty estimation is also more robust. Analysis of aleatoric uncertainty demonstrates that low predictability cases can be caused by the lack of intended direction (which can be indicated by turning signals) or the unknown driving style of the human driver. The previous one can be reduced by adding new signals but the latter one cannot. Epistemic uncertainty comparison of the in-distribution test set and out-of-distribution test set shows that low-speed driving behaviours are radically different from high-speed situations. We conclude that splitting the dataset only according to the road topology of interaction scenarios is not a proper choice for testing the generalizability of trajectory forecasting models. The training set should cover a certain amount of data in both high-speed and low-speed ends.

6.2 Overall conclusion

To summarize, this thesis achieves three things. First, we developed a systematic uncertainty quantification approach for traffic modelling and prediction that considers the "three pillars". Second, we complete the modelling-data collection cycle mentioned in the introductory chapter. Third, we quantified the influence of data collection and modelling on the limit of predictability for traffic forecasting on different levels, and illustrate that the current bottleneck of trajectory prediction and traffic state prediction is the limited diversity of data sources. Next, we will elaborate on them respectively.

- **Insights on uncertainty quantification:** The three pillars of uncertainty quantification are closely related to each other. Both variance/covariance or entropy can be used for quantifying aleatoric and epistemic uncertainty via conditional decomposition. Variance/covariance metrics are good choices for uni-modal distributions and point-estimate models. In contrast, entropy-based metrics are proper for multi-modal or multi-dimensional distributions (such as trajectories) and probabilistic predictors. For the quantification method, the proposed model-free approach can give a trustworthy estimate of the average limit of predictability. Considering the maximum speed of information propagation on the spatial-temporal map can significantly reduce computational complexity. As for input-specific uncertainty, the deep ensemble has already been proven to be an efficient and robust method. From a different angle of view, the parametric representation suits variance/covariance metrics and shows advantages in inference speed. Non-parametric representation, on the other hand, naturally fits entropy metrics.

It has better accuracy but also suffers from large computational complexity and memory requirements. However, the modelling strategy of the UQ approach, must be considered when the available datasets *allow* building a causation-based model. The significant improvement of generalizability error of UQnet shows that both accuracy and the quality of uncertainty estimation can benefit from this.

- **Insights on continuous data collection:** This thesis builds a protocol for "initial dataset–modelling and deployment–continuous data collection" based on uncertainty quantification. The major conclusion here is that we should focus on finding "valuable data" instead of "big data" in the traffic domain. Most samples in the living data stream are repeating low-value data. Putting all data we can access in the database is a waste of money and time. We provide an effective approach to finding those valuable rare samples from a large dataset.

- **Insights on the predictability of traffic:** The quantitative results in this thesis give the predictability of multi-level traffic systems. For macroscopic highway state prediction, although there is a diverse range of congestion patterns and new patterns always emerge in model deployment, the inherent stochasticity of the traffic stream caused by the limited observability of demand and driving behaviours makes the predictability drops fast with time. This is the major bottleneck of traffic state prediction. For microscopic trajectory and intention forecasting, we found that the exposed intended direction is the most important factor for reducing the interaction uncertainty. For scenario generalization problems, the speed is as important as scenario types, otherwise collecting more data cannot enhance the transferability and the predictability of the motion forecasting model.

We believe that these conclusions are valuable for rethinking the model-data relationship in the traffic modelling domain.

6.3 Implications for practice

The methods developed in this thesis are dedicated to making traffic prediction models more reliable and explainable. The most important implication for practice is that:

Researchers in public and commercial organisations should (and now can) evaluate explicitly whether they should invest in models, data or both.

In practice, collecting data is costly. For example, road authorities need to install many loop detectors to monitor the traffic state of a large highway network; autonomous driving companies and laboratories employ people to drive a vehicle equipped with expensive sensors or use drones to collect microscopic trajectory data. On the other hand, constructing models that use the available data at hand is relatively cheaper. In the short term, we often see a great performance improvement. But then defeating the state-of-the-art model will be more and more difficult. We will face the diminishing of marginal utility. We emphasize that obsessing over neither expanding datasets

nor developing more complex modelling techniques without quantitative guidance can always bring the expected return.

For macroscopic traffic networks, the one-sentence implication is "stop investing in developing more complex forecasting models that only use average speed, density, or flow data under *i.i.d. assumption for only improving prediction accuracy*". Although continuously collecting diverse congestion patterns is useful for many other studies and applications, such as getting insights about the regularity of city- or nation-level traffic supply, accuracy improvement is not one of them. In chapters 3 and 4 we quantify and analyze the limit of predictability. The results showed that the widely-studies deep learning models are already approaching this estimated limit. Further reducing the accuracy by 0.1% hardly matters. Meanwhile, we do believe that traffic forecasting is still an attractive and valuable research topic. However, more attention should be paid to the following directions. (1) Improving the interpretability of models. As discussed in chapter 2, there is a possibility to combine data-driven traffic forecasting models with traffic flow theory. We can find a balance point between accuracy and transparency. (2) Enhancing the transferability of prediction models. Although macroscopic traffic congestion patterns depend on road network topology, demand, and travel choices, there exist some shared common features. For example, in chapter 2 we show that the speed of the back-propagating stop-and-go waves is a constant always everywhere. Extracting these invariant features from several datasets may increase the adaptability of a model and thus reduce the training and deployment cost. (3) Traffic forecasting must be combined with traffic control because they are sequentially inherent. Therefore, accuracy is not the only requirement of traffic forecasting. A good model must allow some key operations, such as running forward simulation, what-if analysis, counterfactual analysis, etc. The directions above need more investment.

For microscopic trajectory and driving behaviour modelling and forecasting, due to the fast development of autonomous driving vehicles, this domain is drawing more attention. Here we give two implications for practice. First, different from traffic state forecasting, accuracy is always critically important because it is closely related to safety issues. Besides those downstream tasks (prediction, motion planning, interaction strategy, etc.), perception is still the main source of uncertainty. Trajectories of road agents are relatively easier to be obtained so we see that commercial companies and research institutes launched many open motion-prediction challenges based on trajectory data and map information. But with the lack of some information in the perception module, no-collision prediction is almost impossible. For example, in chapter 5 we mentioned that showing a driver's intention to others is the key to reducing interaction risk and one of the most common approaches is flashing the turn signal. However, we still do not have a reliable dataset of this type of signal. There are two reasons. The first one is that using the pure computer-vision method to recognize turning signals is still an open issue. Complex urban light environments (especially during the night) and diverse car turn signals make recognition very difficult. Second, although these signals can be easily obtained from vehicle manufacturers, almost all companies refuse to provide them due to privacy issues and competition concerns. This barrier causes further problems for connected vehicles. Here we recommend learning from the aviation industry. Governments and companies should develop standards for data-sharing among different manufacturers and the privacy protection of drivers. The shared and flowing data can significantly accelerate the development of this domain.

Second, autonomous driving is currently an AI-dominating field (and we do think that it should be), but contributions from traditional vehicle and traffic engineering cannot be ignored. In chapter 5, we discovered that the INTERACTION dataset does not consider the role of speed in-vehicle interaction for splitting in-distribution and out-of-distribution test sets. This will weaken the generalizability of the model. Close cooperation between AI researchers and traffic engineers can avoid this kind of "roundabouts".

6.4 Implications for science and recommendations

Modelling (whatever phenomena) is one of the most important branches in traffic and transportation studies. The most important implication for the science is that our results suggest that *quantifying the uncertainty of model outcomes is as important as modelling itself*. Like many other domains, traffic modelling is also currently heavily influenced by emerging AI techniques. We must keep in mind that these methods are hammers and pincers in our toolkit. When one of them can be used to solve a problem or address an issue, we use it, not vice versa.

Here we list several recommendations for related future research topics based on this thesis.

- *Causation-based modelling*:
  For a long time, simulation-based methods and data-driven methods are regarded as two different worlds by many people. The simulation-based method is criticized for its huge human effort requirement, unrealistic abstraction, and relatively poor performances in real-world cases. On the other side, data-driven methods, especially deep-learning models, has unsatisfactory generalizability, low interpretability, and high fragility to new scenarios. However, the recently-emerging tendency to combine causality discovery and causal inference with machine learning brings new possibilities. Both old and new domains can potentially bring together the advantages of both approaches.

- *Multi-modal data fusion*:
  Diversifying the data types and sources is critically important for increasing the possible predictability of traffic and implementing causation in data-driven models. However, how to fuse these multi-modal data forms together in a single framework is challenging. For example, in macroscopic traffic modelling, how to efficiently implement trajectory data and traffic demand in the predictor deserves more investigation. The difficulty is that these datasets are on different scales. For autonomous vehicles, multi-modal data may include trajectories, map information, videos recorded by cameras, point clouds sensored by Lidar/Radar, or even sound signals. Utilizing these types of data to better model human-driven vehicles and other road users' behaviours is crucial. Learning from those multi-modal models in the AI domain, such as CLIP (Li et al., 2022c) may give some clues.

- *Sensor network optimization*:
  Besides inducing multi-modal data, another cheaper and more practical solution is optimizing the way that we are currently using for data collection. For example, the proposed method in chapter 3 can identify the least predictable locations on a highway network. Then installing more loop detectors at these locations may be helpful. Or conversely, given a fixed number of detectors (same investment), how should these detectors be distributed in a road network to maximize the potential predictability? Developing a numerical scheme to optimize this can reduce the cost of data collection.

- *What-if analysis and traffic control*:
  The ultimate goal of traffic prediction is for supporting traffic control. In practice, we need to predict several possible futures and prepare their corresponding management plans. In chapter 4, the histogram-regression model is able to predict the bifurcation of future traffic states. Based on this result, it is possible to do the "what-if" analysis. We can consider the most and the least severe congestion that may happen in the short future and then study how to mitigate them.

- *Continuous learning and elastic models*:
  Another great challenge that one may encounter in model deployment is the reusability of forecasting models. Currently, with the increasing size of data, most models need to be completely re-calibrated on the newest dataset. This leads to huge time and money costs. We need to develop a more adjustable model that can continue learning from the on-live data stream without re-training. There are two possible directions. The first one is dataset distillation. Instead of blindly expanding the dataset, we use uncertainty quantification to extract the most valuable samples and maintain a much smaller "core dataset". The size of this core dataset will increase much slower than the original dataset because the occurrence of rare samples is low in practice. So the re-training cost can be reduced. The second way is using the online-learning and continuous learning techniques in the AI community. How to apply them in the traffic domain needs more exploration.

- *Safety assessment of autonomous vehicles*:
  For trajectory forecasting and autonomous vehicles, the non-parametric UQnet proposed in chapter 5 is useful for assessing safety and conflict probability. Safety assessment is closely related to prediction problems. All safety or collision indicators, such as time-to-collision (TTC) and deceleration-rate-to-avoid-collision (DRAC), are based on some simplified assumptions on the target vehicle's future positions. The output of UQnet is the probability distribution of a vehicle's future position. If we apply it to all road users in a case, then we can calculate the probability that the distance between two vehicles is smaller than a specific threshold. This can be a more accurate safety indicator. However, how to validate it needs to be studied.

Appendices

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 below.

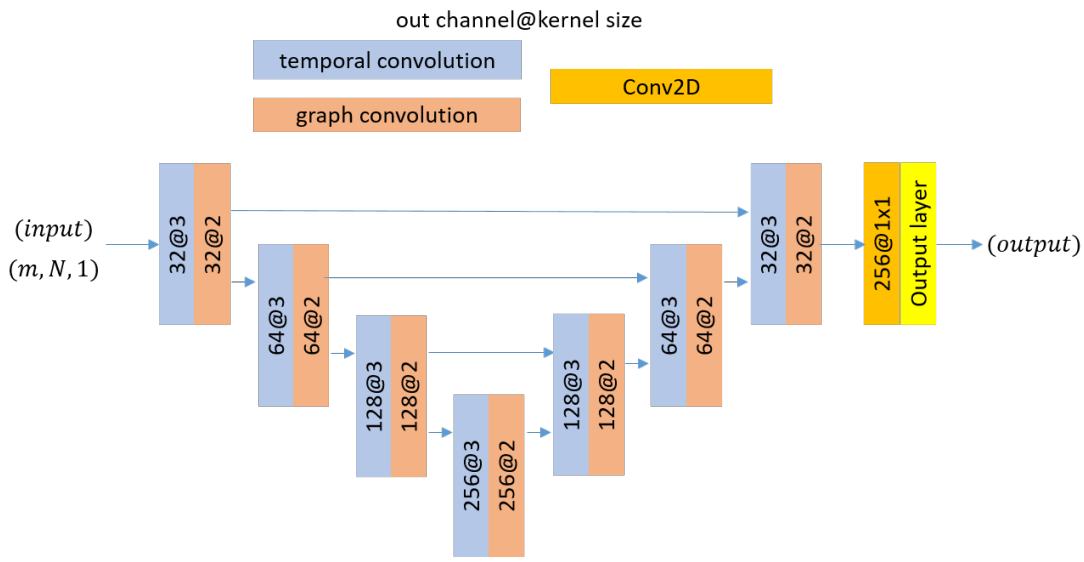
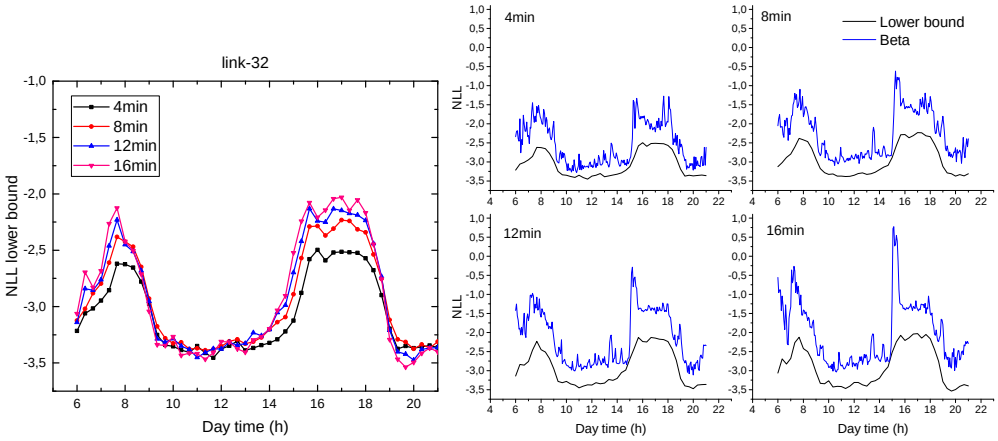


Figure A.1: Structure of the speed forecasting probabilistic model. Here m is the observation length and N is the number of road links.

Here we show another example of the NLL lower bound on link-32.

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



(a) (b)

Figure A.2: (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

Detailed structure of the Beta-regression graph neural networks

The input tensor is concatenated speed and flow at N links in the past P time steps, so the input shape is $(P,N,2)$. We firstly present the details of one ST-block. The first module in a ST-block is the dynamic graph convolutional (DGC) module proposed in Li et al. (2021). It learns input-dependent kernels instead of static kernels for applying graph convolution. We refer the readers to the paper for more details. This DGC module has two hyperparameters, output dimension F_{out} and the order of adjacent matrix k. The DGC module is applied to every time steps. We briefly note it as:

$$\boldsymbol{H}_{(P,N,F_{out})} = \mathrm{DGC}(\boldsymbol{X}_{(P,N,F_{in})}; k, F_{out}) \tag{B.1}$$

The temporal attention layer is a global attention layer along the time axis, it does not have any hyper-parameters and the output has the same shape as input. We note the input $\boldsymbol{X}_{(P,N,F_{in})}$ and its transpose $\boldsymbol{X}^T_{(P,F_{in},N)}$, then the layer writes:

$$\boldsymbol{Q} = \boldsymbol{X}^T\boldsymbol{W}^q, \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}^k, \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}^v \tag{B.2}$$

$$\boldsymbol{H} = \mathrm{softmax}(\boldsymbol{W}^c\boldsymbol{Q}\boldsymbol{K}^T)\boldsymbol{V} \tag{B.3}$$

The trainable parameters are $\boldsymbol{W}^q \in \mathbb{R}^{F_{in}\times N\times N}$, $\boldsymbol{W}^k \in \mathbb{R}^{F_{in}}$, $\boldsymbol{W}^v \in \mathbb{R}^{F_{in}\times F_{in}}$, $\boldsymbol{W}^c \in \mathbb{R}^{P\times P}$. Then the output $\boldsymbol{H}$ has the same shape as $\boldsymbol{X}$. We briefly noted the process above as:

$$\boldsymbol{H}_{(P,N,F_{out}=F_{in})} = \mathrm{TA}(\boldsymbol{X}_{(P,N,F_{in})}) \tag{B.4}$$

The two temporal convolutional layers share the same hyperparameters, the length of the kernel L. Their number of channels are the same as input and the zero-padding is used then these two layers do not change the tensor shapes. The activation, batch normalization, and residual connection are shown in Fig.4.2.

In summary, one ST-block has only three hyperparameters, k, F_{out}, L. In this study we choose $k = 5, F_{out} = 64, L = 5$. After applying 10 ST-blocks the output has the shape $(P,N,F_{out} = 64)$.

The output module for learning Beta distribution is easy to understand. Next we present the cross-attention layer used in the histogram-regression module. The input tensor's

shape is $\boldsymbol{X}_{T,N,F_{in}}$. The speed range is uniformly discretized into C intervals, noted as $\boldsymbol{V}_{C,1}$. Considering that different locations have different numbers of lanes and speed limits, the cross attention to each interval should be location-dependent. So the query speed tensor $\boldsymbol{V}$ must be duplicated N times, noted as $\boldsymbol{Z}_{N,C,1}$. Then the cross-attention layer writes:

$$\boldsymbol{Q} = \boldsymbol{Z}\boldsymbol{W}^q, \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}^k \tag{B.5}$$

$$\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{K}^T \tag{B.6}$$

The trainable parameters are $\boldsymbol{W}^q \in \mathbb{R}^{1\times F_{in}}$ and $\boldsymbol{W}^q \in \mathbb{R}^{F_{in}\times F_{in}}$. Then the output tensor has the shape (T,N,C).

Appendix C

Detailed structure of the proposed UQnet

Many layers in UQnet are graph self-attention layers or cross-attention layers (Velickovic et al., 2017). We start from the more general cross-attention layer. Assume that we have a query input $\boldsymbol{X}_1 \in \mathrm{R}^{n_a \times d_a}$ and a feature input $\boldsymbol{X}_2 \in \mathrm{R}^{n_b \times d_b}$, and an adjacency matrix representing the connectivity $\boldsymbol{A} \in \mathrm{R}^{n_a \times n_b}$ then the output of a cross-attention layer is:

$$\boldsymbol{Q} = \boldsymbol{X}_1 \boldsymbol{W}^q, \boldsymbol{K} = \boldsymbol{X}_2 \boldsymbol{W}^k, \boldsymbol{V} = \boldsymbol{X}_2 \boldsymbol{W}^v \tag{C.1}$$

$$\boldsymbol{H} = \mathrm{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T - 1e^7 \times (1 - \boldsymbol{A})}{\sqrt{d_h}})\boldsymbol{V} \tag{C.2}$$

where $\boldsymbol{W}^q \in \mathrm{R}^{d_a \times d_h}$, $\boldsymbol{W}^k, \boldsymbol{W}^v \in \mathrm{R}^{d_b \times d_h}$ are trainable parameters, the output is $\boldsymbol{H} \in \mathrm{R}^{d_a \times d_h}$. It can also be extended to multi-head attention layers by, for example, concatenating. We briefly note such a cross-attention layer with m attention heads as $\boldsymbol{H} = \mathrm{Attention}(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{A}, d_h, m)$. If $\boldsymbol{X}_1 = \boldsymbol{X}_2$, the cross-attention layer becomes a graph self-attention layer, briefly noted as $\boldsymbol{H} = \mathrm{SelfAttention}(\boldsymbol{A}, d_h, m)$. The input of UQnet includes (the shape of each tensor is also given):

1. $\boldsymbol{T}(26, 9, 8)$: trajectories of all agents in the selected rectangle. The first one is the target vehicle.
2. $\boldsymbol{M}(55, 5, 8)$: map elements within the same range.
3. $\boldsymbol{F}(N, 2)$: coordinates of a mesh-grid. N depends on the resolution of the 2D-histogram. It can be different during training and inference.
4. $\boldsymbol{A}(81, 81)$: adjacency matrix that controls connectivity and information flow directions. It may vary in different layers according to the requirements. In this paper we need the following adjacency matrices:
   (a) $\boldsymbol{J}$, matrix of ones.
   (b) $\boldsymbol{A}_a$, all agents and lanes are connected bi-directionally.
   (c) $\boldsymbol{A}_{l/r/p/f}$, each lane only receive information from their left, right, previous, or following lanes and itself.

(d) $\boldsymbol{A}_t$, all lanes are connected bi-directionally among them and only passes information to the target vehicle.

The encoder and decoder structures is shown in Fig.C.1 and Fig.C.2. The *DenseTNT decoder* is given in DenseTNT (Gu et al., 2021). We refer the readers to their paper and the open-source code for more details.

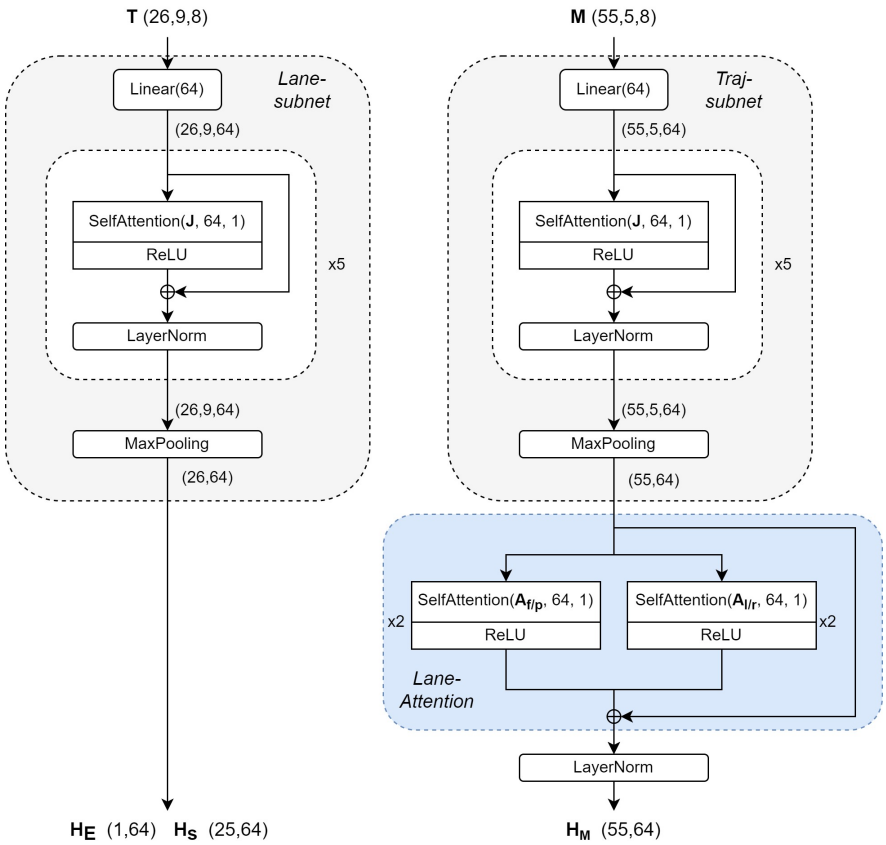


Figure C.1: The encoder structure of UQnet

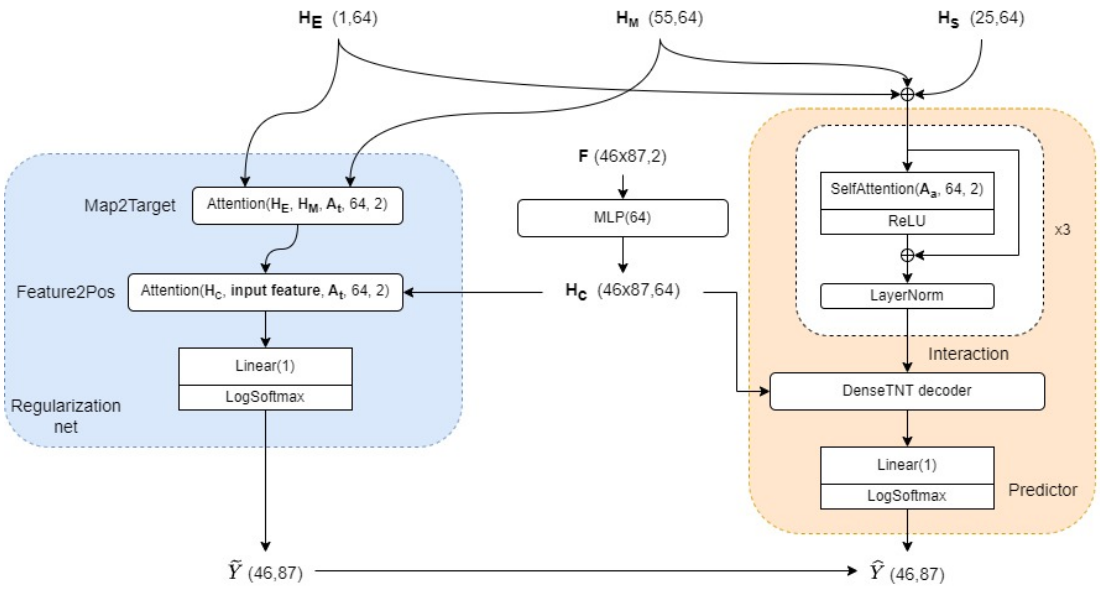


Figure C.2: The decoder structure of UQnet

Bibliography

Abdar, M., F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, et al. (2021) A review of uncertainty quantification in deep learning: Techniques, applications and challenges, *Information Fusion*.

Ahmed, M. S., A. R. Cook (1979) *Analysis of freeway traffic time-series data by using Box-Jenkins techniques*, 722.

Alahi, A., K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, S. Savarese (2016) Social lstm: Human trajectory prediction in crowded spaces, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 961–971.

Amigó, J. M., Y. Hirata, K. Aihara (2017) On the limits of probabilistic forecasting in nonlinear time series analysis ii: differential entropy, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(8), p. 083125.

Amini, A., W. Schwarting, A. Soleimany, D. Rus (2020) Deep evidential regression, *Advances in Neural Information Processing Systems*, 33.

Ammoun, S., F. Nashashibi (2009) Real time trajectory prediction for collision risk estimation between vehicles, in: *2009 IEEE 5th International Conference on Intelligent Computer Communication and Processing*, IEEE, pp. 417–422.

Arnez, F., H. Espinoza, A. Radermacher, F. Terrier (2020) A comparison of uncertainty estimation approaches in deep learning components for autonomous vehicle applications, *arXiv preprint arXiv:2006.15172*.

Bahari, M., S. Saadatnejad, A. Rahimi, M. Shaverdikondori, A. H. Shahidzadeh, S.-M. Moosavi-Dezfooli, A. Alahi (2022) Vehicle trajectory prediction works, but not everywhere, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17123–17133.

Beck, M. B. (1987) Water quality modeling: a review of the analysis of uncertainty, *Water resources research*, 23(8), pp. 1393–1442.

Ben-Akiva, M., M. Bierlaire, H. Koutsopoulos, R. Mishalani (1998) Dynamit: a simulation-based system for traffic prediction, in: *DACCORD Short Term Forecasting Workshop*, Delft The Netherlands, pp. 1–12.

Bengio, S., O. Vinyals, N. Jaitly, N. Shazeer (2015) Scheduled sampling for sequence prediction with recurrent neural networks, in: *Advances in Neural Information Processing Systems*, pp. 1171–1179.

Bergamini, L., Y. Ye, O. Scheel, L. Chen, C. Hu, L. Del Pero, B. Osiński, H. Grimmett, P. Ondruska (2021) Simnet: Learning reactive self-driving simulations from real-world observations, in: *2021 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, pp. 5119–5125.

Bhaskar, A., T. Tsubota, E. Chung, et al. (2014) Urban traffic state estimation: Fusing point and zone based data, *Transportation Research Part C: Emerging Technologies*, 48, pp. 120–142.

Box, G. E. (1976) Science and statistics, *Journal of the American Statistical Association*, 71(356), pp. 791–799.

Brabandere, B. D., X. Jia, T. Tuytelaars, L. V. Gool (2016) Dynamic filter networks, in: *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 667–675.

Brillinger, D. R. (1969) The calculation of cumulants via conditioning, *Annals of the Institute of Statistical Mathematics*, 21(1), pp. 215–218.

Caesar, H., V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, O. Beijbom (2020) nuscenes: A multimodal dataset for autonomous driving, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631.

Calvert, S. C., W. J. Schakel, J. van Lint (2020) A generic multi-scale framework for microscopic traffic simulation part ii–anticipation reliance as compensation mechanism for potential task overload, *Transportation Research Part B: Methodological*, 140, pp. 42–63.

Castillo, E., A. Rivas, P. Jiménez, J. M. Menéndez (2012) Observability in traffic networks. plate scanning added by counting information, *Transportation*, 39(6), pp. 1301–1333.

Castro-Neto, M., Y.-S. Jeong, M.-K. Jeong, L. D. Han (2009) Online-svr for short-term traffic flow prediction under typical and atypical traffic conditions, *Expert systems with applications*, 36(3), pp. 6164–6173.

Chang, M.-F., J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, et al. (2019) Argoverse: 3d tracking and forecasting with rich maps, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8748–8757.

Chen, B.-S., S.-C. Peng, K.-C. Wang (2000) Traffic modeling, prediction, and congestion control for high-speed networks: A fuzzy ar approach, *IEEE Transactions on Fuzzy Systems*, 8(5), pp. 491–508.

Chen, C., H. Wei, N. Xu, G. Zheng, M. Yang, Y. Xiong, K. Xu, Z. Li (2020) Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 3414–3421.

Chen, G., J. Li, J. Lu, J. Zhou (2021a) Human trajectory prediction via counterfactual analysis, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9824–9833.

Chen, X., H. Chen, Y. Yang, H. Wu, W. Zhang, J. Zhao, Y. Xiong (2021b) Traffic flow prediction by an ensemble framework with data denoising and deep learning model, *Physica A: Statistical Mechanics and its Applications*, 565, p. 125574.

Chu, K.-C., L. Yang, R. Saigal, K. Saitou (2011) Validation of stochastic traffic flow model with microscopic traffic simulation, in: *2011 IEEE International Conference on Automation Science and Engineering*, IEEE, pp. 672–677.

Ciuffo, B., V. Punzo, M. Montanino (2012) Thirty years of gipps' car-following model: Applications, developments, and new features, *Transportation research record*, 2315(1), pp. 89–99.

Cover, T. M. (1999) *Elements of information theory*, John Wiley & Sons.

Cruz-Uribe, D., C. Neugebauer (2002) Sharp error bounds for the trapezoidal rule and simpson's rule, *J. Inequal. Pure Appl. Math*, 3(4), pp. 1–22.

Cui, Z., K. Henrickson, R. Ke, Y. Wang (2019) Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting, *IEEE Transactions on Intelligent Transportation Systems*.

Cunningham, J. P., P. Hennig, S. Lacoste-Julien (2011) Gaussian probabilities and expectation propagation, *arXiv preprint arXiv:1111.6832*.

Daamen, W., M. Loot, S. P. Hoogendoorn (2010) Empirical analysis of merging behavior at freeway on-ramp, *Transportation Research Record*, 2188(1), pp. 108–118.

Daganzo, C. F., N. Geroliminis (2008) An analytical approximation for the macroscopic fundamental diagram of urban traffic, *Transportation Research Part B: Methodological*, 42(9), pp. 771–781.

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

Davis, G. A., N. L. Nihan (1991) Nonparametric regression and short-term freeway traffic forecasting, *Journal of Transportation Engineering*, 117(2), pp. 178–188.

De Haan, P., D. Jayaraman, S. Levine (2019) Causal confusion in imitation learning, *Advances in Neural Information Processing Systems*, 32.

Del Ser, J., I. Lana, E. L. Manibardo, I. Oregi, E. Osaba, J. L. Lobo, M. N. Bilbao, E. I. Vlahogianni (2020) Deep echo state networks for short-term traffic forecasting: Performance comparison and statistical assessment, in: *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, pp. 1–6.

Deng, D., C. Shahabi, U. Demiryurek, L. Zhu, R. Yu, Y. Liu (2016) Latent space model for road networks to predict time-varying traffic, in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1525–1534.

Denoeux, T. (1995) A k-nearest neighbor classification rule based on dempster-shafer theory, *IEEE transactions on systems, man, and cybernetics*, 25(5), pp. 804–813.

Der Kiureghian, A., O. Ditlevsen (2009) Aleatory or epistemic? does it matter?, *Structural safety*, 31(2), pp. 105–112.

Deser, C., A. Phillips, V. Bourdette, H. Teng (2012) Uncertainty in climate change projections: the role of internal variability, *Climate dynamics*, 38(3), pp. 527–546.

Djuric, N., V. Radosavljevic, H. Cui, T. Nguyen, F.-C. Chou, T.-H. Lin, N. Singh, J. Schneider (2020) Uncertainty-aware short-term motion prediction of traffic actors for autonomous driving, in: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2095–2104.

Do, L. N., H. L. Vu, B. Q. Vo, Z. Liu, D. Phung (2019) An effective spatial-temporal attention based neural network for traffic flow prediction, *Transportation research part C: emerging technologies*, 108, pp. 12–28.

Ebrahimi, S., M. Elhoseiny, T. Darrell, M. Rohrbach (2019) Uncertainty-guided continual learning in bayesian neural networks, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 75–78.

Eisenman, S. M., X. Fei, X. Zhou, H. S. Mahmassani (2006) Number and location of sensors for real-time network traffic estimation and prediction: Sensitivity analysis, *Transportation Research Record*, 1964(1), pp. 253–259.

Eltoft, T., T. Kim, T.-W. Lee (2006) On the multivariate laplace distribution, *IEEE Signal Processing Letters*, 13(5), pp. 300–303.

Ermagun, A., D. Levinson (2018) Spatiotemporal traffic forecasting: review and proposed directions, *Transport Reviews*, 38(6), pp. 786–814.

Etesami, J., P. Geiger (2020) Causal transfer for imitation learning and decision making under sensor-shift, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 10118–10125.

Faes, L., G. Nollo, A. Porta (2011) Information-based detection of nonlinear granger causality in multivariate processes via a nonuniform embedding technique, *Physical Review E*, 83(5), p. 051112.

Fang, S., M. Skoglund, K. H. Johansson, H. Ishii, Q. Zhu (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.

Foong, A. Y., D. R. Burt, Y. Li, R. E. Turner (2019) On the expressiveness of approximate inference in bayesian neural networks, *arXiv preprint arXiv:1909.00719*.

Fort, S., H. Hu, B. Lakshminarayanan (2019) Deep ensembles: A loss landscape perspective, *arXiv preprint arXiv:1912.02757*.

Fu, J., W. Zhou, Z. Chen (2020) Bayesian spatio-temporal graph convolutional network for traffic forecasting, *arXiv preprint arXiv:2010.07498*.

Fusco, G., C. Colombaroni, N. Isaenko (2016) Short-term speed predictions exploiting big data on large urban road networks, *Transportation Research Part C: Emerging Technologies*, 73, pp. 183–201.

Gao, J., C. Sun, H. Zhao, Y. Shen, D. Anguelov, C. Li, C. Schmid (2020) Vectornet: Encoding hd maps and agent dynamics from vectorized representation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11525–11533.

Gehrke, J. D., J. Wojtusiak (2008) Traffic prediction for agent route planning, in: *International conference on computational science*, Springer, pp. 692–701.

Gentili, M., P. B. Mirchandani (2012) Locating sensors on traffic networks: Models, challenges and research opportunities, *Transportation research part C: emerging technologies*, 24, pp. 227–255.

Gers, F. A., J. Schmidhuber, F. Cummins (1999) Learning to forget: Continual prediction with lstm.

Gilles, T., S. Sabatini, D. Tsishkou, B. Stanciulescu, F. Moutarde (2021) Gohome: Graph-oriented heatmap output for future motion estimation, *arXiv preprint arXiv:2109.01827*.

Gilles, T., S. Sabatini, D. Tsishkou, B. Stanciulescu, F. Moutarde (2022) Gohome: Graph-oriented heatmap output for future motion estimation, in: *2022 International Conference on Robotics and Automation (ICRA)*, IEEE, pp. 9107–9114.

Gipps, P. G. (1981) A behavioural car-following model for computer simulation, *Transportation Research Part B: Methodological*, 15(2), pp. 105–111.

Granger, C. W. (1980) Testing for causality: A personal viewpoint, *Journal of Economic Dynamics and control*, 2, pp. 329–352.

Gu, J., C. Sun, H. Zhao (2021) Densetnt: End-to-end trajectory prediction from dense goal sets, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15303–15312.

Guo, J., W. Huang, B. M. Williams (2014) Adaptive kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification, *Transportation Research Part C: Emerging Technologies*, 43, pp. 50–64.

Guo, J., B. M. Williams (2010) Real-time short-term traffic speed level forecasting and uncertainty quantification using layered kalman filters, *Transportation Research Record*, 2175(1), pp. 28–37.

Guo, S., Y. Lin, N. Feng, C. Song, H. Wan (2019) Attention based spatial-temporal graph convolutional networks for traffic flow forecasting, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 922–929.

Guopeng, L., V. L. Knoop, H. van Lint (2020) Dynamic graph filters networks: A gray-box model for multistep traffic forecasting, in: *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, pp. 1–6.

Gupta, A., J. Johnson, L. Fei-Fei, S. Savarese, A. Alahi (2018) Social gan: Socially acceptable trajectories with generative adversarial networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2255–2264.

Hallé, S., B. Chaib-draa (2005) A collaborative driving system based on multiagent modelling and simulations, *Transportation Research Part C: Emerging Technologies*, 13(4), pp. 320–345.

Hamilton, W. L., Z. Ying, J. Leskovec (2017) Inductive representation learning on large graphs, in: *Advances in Neural Information Processing Systems*, pp. 1024–1034.

Hamner, B. (2010) Predicting travel times with context-dependent random forests by modeling local and aggregate traffic flow, in: *2010 IEEE International Conference on Data Mining Workshops*, IEEE, pp. 1357–1359.

Helbing, D., M. Treiber, A. Kesting, M. Schönhof (2009) Theoretical vs. empirical classification and prediction of congested traffic states, *The European Physical Journal B*, 69(4), pp. 583–598.

Helton, J. C. (1993) Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal, *Reliability Engineering & System Safety*, 42(2-3), pp. 327–367.

Hennessy, D. A., D. L. Wiesenthal (1999) Traffic congestion, driver stress, and driver aggression, *Aggressive Behavior: Official Journal of the International Society for Research on Aggression*, 25(6), pp. 409–423.

Houenou, A., P. Bonnifait, V. Cherfaoui, W. Yao (2013) Vehicle trajectory prediction based on motion model and maneuver recognition, in: *2013 IEEE/RSJ international conference on intelligent robots and systems*, IEEE, pp. 4363–4369.

Hu, Y., X. Jia, M. Tomizuka, W. Zhan (2021) Causal-based time series domain generalization for vehicle intention prediction, in: *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*.

Huang, W., G. Song, H. Hong, K. Xie (2014) Deep architecture for traffic flow prediction: deep belief networks with multitask learning, *IEEE Transactions on Intelligent Transportation Systems*, 15(5), pp. 2191–2201.

Huang, Y., H. Bi, Z. Li, T. Mao, Z. Wang (2019a) Stgat: Modeling spatial-temporal interactions for human trajectory prediction, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6272–6281.

Huang, Y., J. Du, Z. Yang, Z. Zhou, L. Zhang, H. Chen (2022) A survey on trajectory-prediction methods for autonomous driving, *IEEE Transactions on Intelligent Vehicles*.

Huang, Y.-X., R. Jiang, H. Zhang, M.-B. Hu, J.-F. Tian, B. Jia, Z.-Y. Gao (2018) Experimental study and modeling of car-following behavior under high speed situation, *Transportation research part C: emerging technologies*, 97, pp. 194–215.

Huang, Z., X. Wang, L. Huang, C. Huang, Y. Wei, W. Liu (2019b) Ccnet: Criss-cross attention for semantic segmentation, in: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 603–612.

Idé, T., S. Kato (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.

Jayakrishnan, R., W. K. Tsai, A. Chen (1995) A dynamic traffic assignment model with traffic-flow relationships, *Transportation Research Part C: Emerging Technologies*, 3(1), pp. 51–72.

Ji, A., D. Levinson (2020) A review of game theory models of lane changing, *Transportmetrica A: transport science*, 16(3), pp. 1628–1647.

Jia, X., P. Wu, L. Chen, H. Li, Y. Liu, J. Yan (2022) Hdgt: Heterogeneous driving graph transformer for multi-agent trajectory prediction via scene encoding, *arXiv preprint arXiv:2205.09753*.

Johari, M., M. Keyvan-Ekbatani, L. Leclercq, D. Ngoduy, H. S. Mahmassani (2021) Macroscopic network-level traffic models: Bridging fifty years of development toward the next era, *Transportation Research Part C: Emerging Technologies*, 131, p. 103334.

Kamarianakis, Y., P. Prastacos (2005) Space-time modeling of traffic flow, *Computers & Geosciences*, 31(2), pp. 119–133.

Kawata, S., S. Minami (1984) Adaptive smoothing of spectroscopic data by a linear mean-square estimation, *Applied spectroscopy*, 38(1), pp. 49–58.

Kendall, A., Y. Gal (2017) What uncertainties do we need in bayesian deep learning for computer vision?, *arXiv preprint arXiv:1703.04977*.

Kesting, A., M. Treiber, D. Helbing (2010) Enhanced intelligent driver model to access the impact of driving strategies on traffic capacity, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1928), pp. 4585–4605.

Kingma, D. P., J. Ba (2017) Adam: A method for stochastic optimization.

Kipf, T. N., M. Welling (2016) Semi-supervised classification with graph convolutional networks, *arXiv preprint arXiv:1609.02907*.

Kipf, T. N., M. Welling (2017) Semi-supervised classification with graph convolutional networks, in: *ICLR 2017 : International Conference on Learning Representations 2017*.

Knoop, V. L., H. Van Lint, S. P. Hoogendoorn (2015) Traffic dynamics: Its impact on the macroscopic fundamental diagram, *Physica A: Statistical Mechanics and its Applications*, 438, pp. 236–250.

Kontoyiannis, I., P. H. Algoet, Y. M. Suhov, A. J. Wyner (1998) Nonparametric entropy estimation for stationary processes and random fields, with applications to english text, *IEEE Transactions on Information Theory*, 44(3), pp. 1319–1327.

Krishnakumari, P., T. Nguyen, L. Heydenrijk-Ottens, H. L. Vu, H. van Lint (2017) Traffic congestion pattern classification using multiclass active shape models, *Transportation Research Record*, 2645(1), pp. 94–103.

Krishnakumari, P., H. van Lint, T. Djukic, O. Cats (2019) A data driven method for od matrix estimation, *Transportation Research Part C: Emerging Technologies*.

Kumar, S., M. Asger (2018) A study of state-of-the-art neural machine translation approaches, *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 4(1), pp. 135–139.

Kumor, D., J. Zhang, E. Bareinboim (2021) Sequential causal imitation learning with unobserved confounders, *Advances in Neural Information Processing Systems*, 34, pp. 14669–14680.

Kupinski, M. A., J. W. Hoppin, E. Clarkson, H. H. Barrett (2003) Ideal-observer computation in medical imaging with use of markov-chain monte carlo techniques, *JOSA A*, 20(3), pp. 430–438.

Lakshminarayanan, B., A. Pritzel, C. Blundell (2016) Simple and scalable predictive uncertainty estimation using deep ensembles, *arXiv preprint arXiv:1612.01474*.

Lakshminarayanan, B., A. Pritzel, C. Blundell (2017) Simple and scalable predictive uncertainty estimation using deep ensembles, *Advances in neural information processing systems*, 30.

Lamb, A. M., A. G. A. P. Goyal, Y. Zhang, S. Zhang, A. C. Courville, Y. Bengio (2016) Professor forcing: A new algorithm for training recurrent networks, in: *Advances In Neural Information Processing Systems*, pp. 4601–4609.

Lan, L. W., J.-B. Sheu, Y.-S. Huang (2008) Investigation of temporal freeway traffic patterns in reconstructed state spaces, *Transportation Research Part C: Emerging Technologies*, 16(1), pp. 116–136.

Lana, I., J. Del Ser, M. Velez, E. I. Vlahogianni (2018) Road traffic forecasting: Recent advances and new challenges, *IEEE Intelligent Transportation Systems Magazine*, 10(2), pp. 93–109.

Laxhammar, R., G. Falkman (2013) Online learning and sequential anomaly detection in trajectories, *IEEE transactions on pattern analysis and machine intelligence*, 36(6), pp. 1158–1173.

Leclercq, L., V. L. Knoop, F. Marczak, S. P. Hoogendoorn (2016) Capacity drops at merges: New analytical investigations, *Transportation Research Part C: Emerging Technologies*, 62, pp. 171–181.

Lee, K., H. Lee, K. Lee, J. Shin (2017) Training confidence-calibrated classifiers for detecting out-of-distribution samples, *arXiv preprint arXiv:1711.09325*.

Lefèvre, S., D. Vasquez, C. Laugier (2014) A survey on motion prediction and risk assessment for intelligent vehicles, *ROBOMECH journal*, 1(1), pp. 1–14.

Li, G., V. L. Knoop, H. van Lint (2021) Multistep traffic forecasting by dynamic graph convolution: Interpretations of real-time spatial correlations, *Transportation Research Part C: Emerging Technologies*, 128, p. 103185.

Li, G., V. L. Knoop, H. van Lint (2022a) Estimate the limit of predictability in short-term traffic forecasting: An entropy-based approach, *Transportation Research Part C: Emerging Technologies*, 138, p. 103607.

Li, G., Z. LI, V. Knoop, H. van Lint (2022b) Uqnet: Quantifying uncertainty in trajectory prediction by a non-parametric and generalizable approach, *Available at SSRN 4241523*.

Li, H., F. He, X. Lin, Y. Wang, M. Li (2019) Travel time reliability measure based on predictability using the lempel–ziv algorithm, *Transportation Research Part C: Emerging Technologies*, 101, pp. 161–180.

Li, J., Q.-Y. Chen, H. Wang, D. Ni (2012) Analysis of lwr model with fundamental diagram subject to uncertainties, *Transportmetrica*, 8(6), pp. 387–405.

Li, M., R. Xu, S. Wang, L. Zhou, X. Lin, C. Zhu, M. Zeng, H. Ji, S.-F. Chang (2022c) Clip-event: Connecting text and images with event structures, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16420–16429.

Li, Y., X. Jiang, H. Zhu, X. He, S. Peeta, T. Zheng, Y. Li (2016) Multiple measures-based chaotic time series for traffic flow prediction based on bayesian theory, *Nonlinear Dynamics*, 85(1), pp. 179–194.

Li, Y., R. Yu, C. Shahabi, Y. Liu (2017) Diffusion convolutional recurrent neural network: Data-driven traffic forecasting, *arXiv preprint arXiv:1707.01926*.

Li, Y., R. Yu, C. Shahabi, Y. Liu (2018) Diffusion convolutional recurrent neural network: Data-driven traffic forecasting, in: *ICLR 2018 : International Conference on Learning Representations 2018*.

Liang, M., B. Yang, R. Hu, Y. Chen, R. Liao, S. Feng, R. Urtasun (2020) Learning lane graph representations for motion forecasting, in: *European Conference on Computer Vision*, Springer, pp. 541–556.

Liebig, T., N. Piatkowski, C. Bockermann, K. Morik (2017) Dynamic route planning with real-time traffic predictions, *Information Systems*, 64, pp. 258–265.

Lighthill, M., G. Whitham (1955a) On kinematic waves ii: A theory of traffic flow on long crowded roads, *Proc. R. Soc*, A 229(1178), pp. 317–345.

Lighthill, M. J., G. B. Whitham (1955b) On kinematic waves ii. a theory of traffic flow on long crowded roads, *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 229(1178), pp. 317–345.

Lin, T.-Y., P. Goyal, R. Girshick, K. He, P. Dollár (2017) Focal loss for dense object detection, in: *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988.

Liu, H., H. Van Zuylen, H. Van Lint, M. Salomons (2006) Predicting urban arterial travel time with state-space neural networks and kalman filters, *Transportation Research Record*, 1968(1), pp. 99–108.

Liu, Y., R. Cadei, J. Schweizer, S. Bahmani, A. Alahi (2022) Towards robust and adaptive motion forecasting: A causal representation perspective, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17081–17092.

Lombardi, D., S. Pant (2016) Nonparametric k-nearest-neighbor entropy estimator, *Physical Review E*, 93(1), p. 013310.

Lopez, C., L. Leclercq, P. Krishnakumari, N. Chiabaut, H. Van Lint (2017) Revealing the day-to-day regularity of urban congestion patterns with 3d speed maps, *Scientific Reports*, 7(1), pp. 1–11.

Lutteken, N., M. Zimmermann, K. J. Bengler (2016) Using gamification to motivate human cooperation in a lane-change scenario, in: *Proceedings of the IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, Series Using gamification to motivate human cooperation in a lane-change scenario, Rio de Janiero, Brasil.

Ma, X., Z. Dai, Z. He, J. Ma, Y. Wang, Y. Wang (2017) Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction, *Sensors*, 17(4), p. 818.

Ma, X., Z. Tao, Y. Wang, H. Yu, Y. Wang (2015) Long short-term memory neural network for traffic speed prediction using remote microwave sensor data, *Transportation Research Part C: Emerging Technologies*, 54, pp. 187–197.

Ma, Y., X. Zhu, S. Zhang, R. Yang, W. Wang, D. Manocha (2019) Trafficpredict: Trajectory prediction for heterogeneous traffic-agents, in: *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 6120–6127.

Makansi, O., E. Ilg, O. Cicek, T. Brox (2019) Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7144–7153.

Makansi, O., J. von Kügelgen, F. Locatello, P. Gehler, D. Janzing, T. Brox, B. Schölkopf (2021) You mostly walk alone: Analyzing feature attribution in trajectory prediction, *arXiv preprint arXiv:2110.05304*.

Malinin, A., S. Chervontsev, I. Provilkov, M. Gales (2020) Regression prior networks, *arXiv preprint arXiv:2006.11590*.

Malinin, A., M. Gales (2018) Predictive uncertainty estimation via prior networks, *arXiv preprint arXiv:1802.10501*.

Malinin, A., B. Mlodozeniec, M. Gales (2019) Ensemble distribution distillation, *arXiv preprint arXiv:1905.00076*.

McDermott, P. L., C. K. Wikle (2019) Deep echo state networks with uncertainty quantification for spatio-temporal forecasting, *Environmetrics*, 30(3), p. e2553.

Min, W., L. Wynter (2011) Real-time road traffic prediction with spatio-temporal correlations, *Transportation Research Part C: Emerging Technologies*, 19(4), pp. 606–616.

Mo, X., Y. Xing, C. Lv (2020) Recog: A deep learning framework with heterogeneous graph for interaction-aware trajectory prediction, *arXiv preprint arXiv:2012.05032*.

Mo, X., Y. Xing, C. Lv (2021) Heterogeneous edge-enhanced graph attention network for multi-agent trajectory prediction, *arXiv preprint arXiv:2106.07161*.

Multiscale modelling, Multiscale modelling — Wikipedia, the free encyclopedia, [Online; accessed 24-September-2022], URL `https://en.wikipedia.org/wiki/Multiscale_modeling`.

Nair, A. S., J.-C. Liu, L. Rilett, S. Gupta (2001) Non-linear analysis of traffic flow, in: *ITSC 2001. 2001 IEEE Intelligent Transportation Systems. Proceedings (Cat. No. 01TH8585)*, IEEE, pp. 681–685.

Nantes, A., D. Ngoduy, A. Bhaskar, M. Miska, E. Chung (2016) Real-time traffic state estimation in urban corridors from heterogeneous data, *Transportation Research Part C: Emerging Technologies*, 66, pp. 99–118.

Nguyen, T. T., P. Krishnakumari, S. C. Calvert, H. L. Vu, H. Van Lint (2019) Feature extraction and clustering analysis of highway congestion, *Transportation Research Part C: Emerging Technologies*, 100, pp. 238–258.

Nikhil, N., B. Tran Morris (2018) Convolutional neural network for trajectory prediction, in: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0.

Ovadia, Y., E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, J. Snoek (2019) Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift, *arXiv preprint arXiv:1906.02530*.

Packard, N. H., J. P. Crutchfield, J. D. Farmer, R. S. Shaw (1980) Geometry from a time series, *Physical review letters*, 45(9), p. 712.

Pang, Y., X. Zhao, H. Yan, Y. Liu (2021) Data-driven trajectory prediction with weather uncertainties: A bayesian deep learning approach, *Transportation Research Part C: Emerging Technologies*, 130, p. 103326.

Payne, H. J. (1971) Model of freeway traffic and control, *Mathematical Model of Public System*, pp. 51–61.

Pearl, J. (2009) Causal inference in statistics: An overview, *Statistics surveys*, 3, pp. 96–146.

Poggi, M., F. Aleotti, F. Tosi, S. Mattoccia (2020) On the uncertainty of self-supervised monocular depth estimation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3227–3237.

Polson, N. G., V. O. Sokolov (2017) Deep learning for short-term traffic flow prediction, *Transportation Research Part C: Emerging Technologies*, 79, pp. 1–17.

Prevost, C. G., A. Desbiens, E. Gagnon (2007) Extended kalman filter for state estimation and trajectory prediction of a moving object detected by an unmanned aerial vehicle, in: *2007 American control conference*, IEEE, pp. 1805–1810.

Prigogine, I., R. Herman (1971) Kinetic theory of vehicular traffic, Tech. rep.

Qiao, F., H. Yang, W. H. Lam (2001) Intelligent simulation and prediction of traffic flow dispersion, *Transportation Research Part B: Methodological*, 35(9), pp. 843–863.

Rice, J., E. Van Zwet (2004) A simple and effective method for predicting travel times on freeways, *IEEE Transactions on Intelligent Transportation Systems*, 5(3), pp. 200–207.

Richards, P. I. (1956) Shock waves on the highway, *Operations research*, 4(1), pp. 42–51.

Ritter, H., A. Botev, D. Barber (2018) A scalable laplace approximation for neural networks, in: *6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings*, vol. 6, International Conference on Representation Learning.

Rodrigues, F., F. C. Pereira (2018) Heteroscedastic gaussian processes for uncertainty modeling in large-scale crowdsourced traffic data, *Transportation research part C: emerging technologies*, 95, pp. 636–651.

Ronneberger, O., P. Fischer, T. Brox (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, M. T., J. J. Collins, C. J. De Luca (1993) A practical method for calculating largest lyapunov exponents from small data sets, *Physica D: Nonlinear Phenomena*, 65(1-2), pp. 117–134.

Rudenko, A., L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, K. O. Arras (2020) Human motion trajectory prediction: A survey, *The International Journal of Robotics Research*, 39(8), pp. 895–935.

Saifuzzaman, M., Z. Zheng (2014) Incorporating human-factors in car-following models: a review of recent developments and research needs, *Transportation research part C: emerging technologies*, 48, pp. 379–403.

Salzmann, T., B. Ivanovic, P. Chakravarty, M. Pavone (2020) Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data, in: *European Conference on Computer Vision*, Springer, pp. 683–700.

Santurkar, S., D. Tsipras, A. Ilyas, A. Madry (2018) How does batch normalization help optimization?, *Advances in neural information processing systems*, 31.

Schakel, W. J., V. L. Knoop, B. van Arem (2012) Integrated lane change model with relaxation and synchronization, *Transportation Research Record*, 2316(1), pp. 47–57.

Schölkopf, B., F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, Y. Bengio (2021) Toward causal representation learning, *Proceedings of the IEEE*, 109(5), pp. 612–634.

Schreiter, T., H. van Lint, M. Treiber, S. Hoogendoorn (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, T., H. Van Lint, Y. Yuan, S. Hoogendoorn (2010b) Propagation wave speed estimation of freeway traffic with image processing tools, Tech. rep.

Shang, P., X. Li, S. Kamae (2005) Chaotic analysis of traffic time series, *Chaos, Solitons & Fractals*, 25(1), pp. 121–128.

Shannon, C. E. (1948) A mathematical theory of communication, *The Bell system technical journal*, 27(3), pp. 379–423.

Sharma, A., Z. Zheng, A. Bhaskar (2019) Is more always better? the impact of vehicular trajectory completeness on car-following model calibration and validation, *Transportation research part B: methodological*, 120, pp. 49–75.

Sharma, B., S. Kumar, P. Tiwari, P. Yadav, M. I. Nezhurina (2018) Ann based short-term traffic flow forecasting in undivided two lane highway, *Journal of Big Data*, 5(1), pp. 1–16.

Sirignano, J., K. Spiliopoulos (2018) Dgm: A deep learning algorithm for solving partial differential equations, *Journal of computational physics*, 375, pp. 1339–1364.

Smith, B. L., M. J. Demetsky (1997) Traffic flow forecasting: comparison of modeling approaches, *Journal of transportation engineering*, 123(4), pp. 261–266.

Song, C., Z. Qu, N. Blumm, A.-L. Barabási (2010) Limits of predictability in human mobility, *Science*, 327(5968), pp. 1018–1021.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov (2014) Dropout: a simple way to prevent neural networks from overfitting, *The journal of machine learning research*, 15(1), pp. 1929–1958.

Sun, P., H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al. (2020) Scalability in perception for autonomous driving: Waymo open dataset, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2446–2454.

Suo, S., S. Regalado, S. Casas, R. Urtasun (2021) Trafficsim: Learning to simulate realistic multi-agent behaviors, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10400–10409.

Swiatkowski, J., K. Roth, B. Veeling, L. Tran, J. Dillon, J. Snoek, S. Mandt, T. Salimans, R. Jenatton, S. Nowozin (2020) The k-tied normal distribution: A compact parameterization of gaussian mean field posteriors in bayesian neural networks, in: *International Conference on Machine Learning*, PMLR, pp. 9289–9299.

Tahmasbi, R., S. M. Hashemi (2013) Modeling and forecasting the urban volume using stochastic differential equations, *IEEE Transactions on Intelligent Transportation Systems*, 15(1), pp. 250–259.

Tang, X., K. Yang, H. Wang, J. Wu, Y. Qin, W. Yu, D. Cao (2022) Prediction-uncertainty-aware decision-making for autonomous vehicles, *IEEE Transactions on Intelligent Vehicles*.

Toledo, T., H. N. Koutsopoulos, M. Ben-Akiva (2009) Estimation of an integrated driving behavior model, *Transportation Research Part C: Emerging Technologies*, 17(4), pp. 365–380.

Treiber, M., D. Helbing (2002) Reconstructing the spatio-temporal traffic dynamics from stationary detector data, *Cooper@ tive Tr@ nsport@ tion Dyn@ mics*, 1(3), pp. 3–1.

Treiber, M., D. Helbing (2003) An adaptive smoothing method for traffic state identification from incomplete information, in: *Interface and Transport Dynamics*, Springer, pp. 343–360.

Treiber, M., A. Hennecke, D. Helbing (2000) Congested traffic states in empirical observations and microscopic simulations, *Physical review E*, 62(2), p. 1805.

van Amersfoort, J., L. Smith, Y. W. Teh, Y. Gal (2020) Simple and scalable epistemic uncertainty estimation using a single deep deterministic neural network.

van Hinsbergen, C. I., J. Van Lint, H. Van Zuylen (2009) Bayesian committee of neural networks to predict travel times with confidence intervals, *Transportation Research Part C: Emerging Technologies*, 17(5), pp. 498–509.

Van Hinsbergen, C. P., T. Schreiter, F. S. Zuurbier, J. Van Lint, H. J. Van Zuylen (2011) Localized extended kalman filter for scalable real-time traffic state estimation, *IEEE transactions on intelligent transportation systems*, 13(1), pp. 385–394.

van Hinsbergen, C. P. I. J., T. Schreiter, F. S. Zuurbier, J. W. C. van Lint, H. J. van Zuylen (2012) Localized extended kalman filter for scalable real-time traffic state estimation, *Ieee Transactions on Intelligent Transportation Systems*, 13(1), pp. 385–394.

van Lint, H., O. Miete, H. Taale, S. Hoogendoorn (2012a) Systematic framework for assessing traffic measures and policies on reliability of traffic operations and travel time, *Transportation research record*, 2302(1), pp. 92–101.

van Lint, H., O. Miete, H. Taale, S. Hoogendoorn (2012b) Systematic framework for assessing traffic measures and policies on reliability of traffic operations and travel time, *Transportation Research Record*, 2302, pp. 92–101.

Van Lint, J. (2008) Online learning solutions for freeway travel time prediction, *IEEE Transactions on Intelligent Transportation Systems*, 9(1), pp. 38–47.

Van Lint, J., S. C. Calvert (2018) A generic multi-level framework for microscopic traffic simulation—theory and an example case in modelling driver distraction, *Transportation Research Part B: Methodological*, 117, pp. 63–86.

Van Lint, J., S. Hoogendoorn, H. J. van Zuylen (2002) Freeway travel time prediction with state-space neural networks: modeling state-space dynamics with recurrent neural networks, *Transportation Research Record*, 1811(1), pp. 30–39.

Van Lint, J., S. Hoogendoorn, H. J. van Zuylen (2005) Accurate freeway travel time prediction with state-space neural networks under missing data, *Transportation Research Part C: Emerging Technologies*, 13(5-6), pp. 347–369.

van Lint, J. W. C., S. P. Hoogendoorn (2010) A robust and efficient method for fusing heterogeneous data from traffic sensors on freeways, *Computer-Aided Civil and Infrastructure Engineering*, 25(8), pp. 596–612.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin (2017) Attention is all you need, in: *Advances in neural information processing systems*, pp. 5998–6008.

Velickovic, P., G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio (2017) Graph attention networks, *stat*, 1050, p. 20.

Veličković, P., G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio (2018) Graph attention networks, in: *ICLR 2018 : International Conference on Learning Representations 2018*.

Vemula, A., K. Muelling, J. Oh (2018) Social attention: Modeling attention in human crowds, in: *2018 IEEE international Conference on Robotics and Automation (ICRA)*, IEEE, pp. 4601–4607.

Viti, F., M. Rinaldi, F. Corman, C. M. Tampère (2014) Assessing partial observability in network sensor location problems, *Transportation research part B: methodological*, 70, pp. 65–89.

Wang, H., B. Lu, J. Li, T. Liu, Y. Xing, C. Lv, D. Cao, J. Li, J. Zhang, E. Hashemi (2021) Risk assessment and mitigation in local path planning for autonomous vehicles with lstm based predictive model, *IEEE Transactions on Automation Science and Engineering*.

Wang, J., Y. Mao, J. Li, Z. Xiong, W.-X. Wang (2015a) Predictability of road traffic and congestion in urban areas, *PloS one*, 10(4), p. e0121825.

Wang, M., S. P. Hoogendoorn, W. Daamen, B. van Arem, R. Happee (2015b) Game theoretic approach for predictive lane-changing and car-following control, *Transportation Research Part C: Emerging Technologies*, 58, pp. 73–92.

Wang, Q., S. R. Kulkarni, S. Verdú (2009) Divergence estimation for multidimensional densities via k-nearest-neighbor distances, *IEEE Transactions on Information Theory*, 55(5), pp. 2392–2405.

Wang, R., D. B. Work, R. Sowers (2016) Multiple model particle filter for traffic estimation and incident detection, *IEEE Transactions on Intelligent Transportation Systems*, 17(12), pp. 3461–3470.

Wang, Y., M. Papageorgiou, A. Messmer (2005) A real-time freeway network traffic surveillance tool, *IEEE Transactions on control systems technology*, 14(1), pp. 18–32.

Wang, Y., M. Papageorgiou, A. Messmer (2006) Renaissance - a unified macroscopic model-based approach to real-time freeway network traffic surveillance, *Transportation Research Part C: Emerging Technologies*, 14(3), pp. 190–212.

Wang, Y., M. Papageorgiou, A. Messmer (2008) Real-time freeway traffic state estimation based on extended kalman filter: Adaptive capabilities and real data testing, *Transportation Research Part A: Policy and Practice*, 42(10), pp. 1340–1358.

Whitham, G. B. (2011) *Linear and nonlinear waves*, John Wiley & Sons.

Williams, B. M., L. A. Hoel (2003) Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results, *Journal of transportation engineering*, 129(6), pp. 664–672.

Wolf, A., J. B. Swift, H. L. Swinney, J. A. Vastano (1985) Determining lyapunov exponents from a time series, *Physica D: nonlinear phenomena*, 16(3), pp. 285–317.

Wu, Z., S. Pan, G. Long, J. Jiang, C. Zhang (2019) Graph wavenet for deep spatial-temporal graph modeling, *arXiv preprint arXiv:1906.00121*.

Xie, G., A. Shangguan, R. Fei, W. Ji, W. Ma, X. Hei (2020) Motion trajectory prediction based on a cnn-lstm sequential model, *Science China Information Sciences*, 63(11), pp. 1–21.

Xie, X., H. van Lint, A. Verbraeck (2018) A generic data assimilation framework for vehicle trajectory reconstruction on signalized urban arterials using particle filters, *Transportation Research Part C: Emerging Technologies*, 92, pp. 364–391.

Xing, X., X. Zhou, H. Hong, W. Huang, K. Bian, K. Xie (2015) Traffic flow decomposition and prediction based on robust principal component analysis, in: *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, IEEE, pp. 2219–2224.

Xiong, W., L. Faes, P. C. Ivanov (2017) Entropy measures, entropy estimators, and their performance in quantifying complex dynamics: Effects of artifacts, nonstationarity, and long-range correlations, *Physical Review E*, 95(6), p. 062114.

Xu, Y., Z. Piao, S. Gao (2018) Encoding crowd interaction with deep neural network for pedestrian trajectory prediction, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5275–5284.

Yoon, S., H. Jeon, D. Kum (2019) Predictive cruise control using radial basis function network-based vehicle motion prediction and chance constrained model predictive control, *IEEE Transactions on Intelligent Transportation Systems*, 20(10), pp. 3832–3843.

Yu, B., Y. Lee, K. Sohn (2020) Forecasting road traffic speeds by considering area-wide spatio-temporal dependencies based on a graph convolutional neural network (gcn), *Transportation Research Part C: Emerging Technologies*, 114, pp. 189–204.

Yu, B., H. Yin, Z. Zhu (2017) Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting, *arXiv preprint arXiv:1709.04875*.

Yu, B., H. Yin, Z. Zhu (2018) Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting, in: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 3634–3640.

Yuan, J., Y. Zheng, X. Xie, G. Sun (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.

Yuan, K., J. Laval, V. L. Knoop, R. Jiang, S. P. Hoogendoorn (2018) A geometric brownian motion car-following model: towards a better understanding of capacity drop, *Transportmetrica B: Transport Dynamics*.

Yuan, Y., Z. Zhang, X. T. Yang, S. Zhe (2021) Macroscopic traffic flow modeling with physics regularized gaussian process: A new insight into machine learning applications in transportation, *Transportation Research Part B: Methodological*, 146, pp. 88–110.

Zhan, W., L. Sun, D. Wang, H. Shi, A. Clausse, M. Naumann, J. Kummerle, H. Konigshof, C. Stiller, A. de La Fortelle, et al. (2019) Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps, *arXiv preprint arXiv:1910.03088*.

Zhang, J., X. Shi, J. Xie, H. Ma, I. King, D. yan Yeung (2018) Gaan: Gated attention networks for learning on large and spatiotemporal graphs, in: *UAI 2018: The Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 339–349.

Zhang, W., Y. Yu, Y. Qi, F. Shu, Y. Wang (2019a) Short-term traffic flow prediction based on spatio-temporal analysis and cnn deep learning, *Transportmetrica A: Transport Science*, 15(2), pp. 1688–1711.

Zhang, Z., M. Li, X. Lin, Y. Wang, F. He (2019b) Multistep speed prediction on traffic networks: A deep learning approach considering spatio-temporal dependencies, *Transportation Research Part C: Emerging Technologies*, 105, pp. 297–322.

Zhao, J., V. L. Knoop, M. Wang (2022) Microscopic traffic modeling inside intersections: Interactions between drivers, *Transportation Science*.

Zhao, T., Y. Xu, M. Monfort, W. Choi, C. Baker, Y. Zhao, Y. Wang, Y. N. Wu (2019) Multi-agent tensor fusion for contextual trajectory prediction, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12126–12134.

Zhao, Z., W. Chen, X. Wu, P. C. Chen, J. Liu (2017) Lstm network: a deep learning approach for short-term traffic forecast, *IET Intelligent Transport Systems*, 11(2), pp. 68–75.

Zheng, W., D.-H. Lee, Q. Shi (2006) Short-term freeway traffic flow prediction: Bayesian combined neural network approach, *Journal of transportation engineering*, 132(2), pp. 114–121.

Zhou, X., D. Wang, P. Krähenbühl (2019) Objects as points, *arXiv preprint arXiv:1904.07850*.

Zhu, X., D. Cheng, Z. Zhang, S. Lin, J. Dai (2019) An empirical study of spatial attention mechanisms in deep networks, in: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6688–6697.

Summary

Observing, modelling, predicting, and understanding the dynamics of traffic systems on different levels is one of the most critical topics in the transport and planning domain. At the macroscopic scale, traffic congestion is the central problem that impacts all aspects of society. Traffic congestion costs valuable travel time, extra fuel consumption, and frustration in daily life. Traffic congestion is not always avoidable but accurate predictions of traffic conditions in a road network are useful for road users. For example, drivers can make faster and safer route choices based on the estimated time of arrival and the predicted congestion evolution. Reliable traffic forecasting also provides essential information for real-time traffic control systems and the development of long-term sustainable mobility systems. On the other hand, on the microscopic level, modelling the interaction between road users and predicting their behaviours is drawing more and more attention due to the increasing popularity of autonomous driving. Accurately anticipating other agents' decisions is indispensable for a safe and smooth autopilot system. This actively-studied domain has become the focus of many scholars from academia and engineers from industry.

Recently, with the fast development of sensing technology, data-driven models, particularly using Artificial Intelligence (AI) techniques, have become one of the most popular approaches in the traffic forecasting domain. The macroscopic traffic state is generally monitored by installed sensors, e,g, loop-detectors, and described by the corresponding derived variables, such as traffic flow, density, and average speed. These data greatly enrich our understanding of large-scale traffic congestion patterns. Microscopic trajectory data collection involves the wide application of drones and in-car sensors, e.g. cameras, GPS, Lidar, and radar. These tools allow perceiving real-time trajectories of surrounding agents and localizing map information. The booming of "big data" greatly stimulates the fast iteration of increasingly accurate AI models. In academia, AI-based traffic forecasting and trajectory prediction model has already become the mainstream. These AI models show impressive performances for those larger-scale and more complex forecasting tasks.

However, this "data+AI" paradigm also exposes its weaknesses in practice. Low interpretability and unsatisfactory reliability are the most concerning issues. The "black-box" property of neural networks and the fragility of correlation-based AI in new scenarios hinder further applications in the real world. Additionally, improving the accuracy of traffic forecasting is becoming more and more difficult. Hence two questions naturally arise: How predictable is traffic on different levels? Are we close enough to the limit of predictability? The answers to these questions can put endless performance benchmarks into perspective and indicate what is the most valuable research direction in the future.

This thesis addresses the issue of the predictability of traffic systems on different levels of scale by proposing a systematic approach for quantifying the predictive uncertainty from different sources in AI models. Additionally, the interpretability and the generalizability of the used deep learning models are also improved. The thesis provides a comprehensive and quantitative evaluation of the predictability of traffic.

The scientific gaps addressed in this thesis are categorized into 4 intertwined parts. First, we improve the interpretability of deep-learning-based traffic forecasting models by designing a novel dynamic module that explicitly learns state-dependent spatial associations among road links. This provides insights into how to combine a data-driven model with explainable traffic flow theory. The obtained principles of model design will be used in the following uncertainty quantification methods. Second, we estimate the average limit of predictability of macroscopic traffic quantities for different types of forecasting models. The predictability here is measured by the lower bound of various predictive errors. We first theoretically prove these lower bounds and then use a numerical scheme to directly estimate them from the given dataset. This part provides a set of model-free tools to evaluate whether increasing model sophistication is worthwhile in terms of additional gains in accuracy. In the third part, based on a set of practical requirements, we propose a method for estimating two types of input-specific predictive uncertainty metrics in macroscopic traffic state forecasting. The two types of uncertainty metrics represent respectively the inherent randomness in traffic dynamics and the rareness of congestion patterns. The results provide insights into how inherently predictable traffic is, how frequently rare patterns happen, and why macroscopic predictability is limited. These conclusions are essential for continuous data collection. The final part extends the previous uncertainty quantification method to the microscopic motion prediction problem. We demonstrate that embedding causal effects significantly enhances the robustness of the motion prediction model in new interaction scenarios. We also found that exposing intended directions to other road users is the key to reducing predictive uncertainty and domain knowledge of driving behaviours is indispensable for generalizable AI models. The 4 parts will be discussed individually below.

Learn dynamic spatial associations by designed deep learning models

The spatial association between links in a road network is a key factor in classical traffic flow models. For example, the discrete form of a spatial operator describes how the current state of one link influences the future state of another link. This "influence" is generally state-dependent (dynamic), localized (there is a limit of the spreading speed), and asymmetric (influences from upstream and downstream links are different). However, deep neural networks assimilate the spatial association from the dataset in an implicit and global way, which hinders us from seeking interpretations. To address this issue, we design a dynamic module that aims to explicitly learn the spatial dependencies and implement it in a deep learning framework. The experiments on real-world highway networks show that the proposed method can indeed give both accurate short-term predictions and explainable spatial associations that are consistent with traffic flow theory. We especially demonstrate that this spatial association is causation-like. Deeper neural networks with improper hyperparameters cannot give clear propagating stop-and-go waves due to causal confusion. These conclusions are

important for the model design in uncertainty quantification.

Estimate the average limit of predictability for macroscopic traffic variables

Modelling is not the only important respect of macroscopic traffic forecasting. The available amount and types of data also restrict how accurately the traffic can be predicted. Due to the limited observability of those critical factors that potentially drive the change of traffic condition, e.g. detailed traffic demand patterns, route choices, driving behaviours around on-ramps, and weather, traffic dynamics shows significant stochasticity. We need a method to quantitatively link the stochasticity to the limit (lower bound) of predictive accuracy.

This thesis addresses this issue from both theoretical and practical perspectives. We first mathematically prove a lower bound of accuracy for any deterministic forecasting models and any probabilistic forecasting models, respectively. The key concept that we use is the so-called conditional differential entropy and developing a numerical scheme to estimate it is necessary. In practice, directly estimating conditional entropy from a multivariate time series is challenging due to the curse of dimensionality. To mitigate this problem, the spatial-temporal features obtained from the first part of this thesis are used to partition the entire dataset into time-of-day and location-related subsets. By using this technique, the dimensionality is dramatically reduced and directly estimating the limit of predictive accuracy is possible. We show that, for network-level traffic production and traffic speed forecasting, the current deep learning models are approaching the estimated theoretical limit in short term. This approach is also informative of the most uncertain time periods in one day and the most uncertain locations in a highway network. This proposed method provides a tool to evaluate whether the given dataset can potentially fulfil the accuracy requirement, and also insights into the spatial-temporal distribution of predictability. The latter one is informative for traffic control practitioners.

Estimate the inherent stochasticity and the rareness of the current traffic state

In this part, we comprehensively analyze the "predictability" of macroscopic traffic by quantifying both the inherent stochasticity of traffic dynamics and the occurrence of rare congestion patterns. The existing ensemble learning method is applied to traffic speed forecasting problems. The experimental results show that, although rare or new congestion patterns always emerge in the data stream, the predictability of speed is mainly restricted by the inherent stochasticity of traffic dynamics. By directly assimilating the evolution of predicted speed distribution, we further demonstrate that the predictability drops rapidly with the prediction horizon due to the bifurcation of future traffic state. This bi-modality causes high and irreducible uncertainty and it can be explained by the capacity drop in traffic flow theory.

Quantify the uncertainty in motion prediction by a generalizable method

Different from macroscopic traffic forecasting, the output of a motion prediction task is not a scalar but a 2D coordinate. Additionally, the 2D spatial distribution is restricted by arbitrary road layouts. Currently, the major driving force of motion prediction is

the relevant applications in autonomous vehicles. An autonomous vehicle is expected to accurately anticipate other road users' intentions when facing diverse scenarios that have not been seen in the training set. Therefore, the motion prediction model should be able to identify rare behaviours and adapt itself to new scenarios. This is important for safety issues.

Based on these requirements, this thesis extends the scalar uncertainty quantification method to a 2D plane. The output spatial distribution is directly approximated by a non-parametric histogram instead of a simple prior. The causal effect of surrounding agents is considered by a cascade strategy so the model is better generalizable to new scenarios. We demonstrate that this approach can find those rare driving behaviours and the uncertainty of interaction is mainly caused by the unknown intended direction (indicated by the turning signal) and driving styles. We also found that the AI model cannot learn how to adapt low-speed driving behaviours to high-speed cases because the domain knowledge tells us that the heterogeneity of behaviours increases with speed. Based on these observations, we suggest that diversifying the perception is the key to safer and smoother motion prediction and planning.

Conclusion and perspective

In summary, this thesis addresses two major issues: How predictable are traffic dynamics on different levels? How to quantify uncertainty in traffic forecasting, and use this to improve data collection and modelling? We answer these two research questions by proposing a systematic approach for quantifying predictive uncertainty in macroscopic traffic state prediction and microscopic trajectory prediction.

For the first question of predictability, we conclude that for both network-level traffic state prediction and motion prediction, the limit of predictability is mainly restricted by the irreducible randomness caused by the limited perception data types. Currently, AI models are approaching this limit. A significant improvement in predictive accuracy is not going to happen unless we diversify the available data types. For macroscopic traffic, this might be trajectory or demand data. In the context of autonomous vehicles, recognizing the turning signal is important. For the second question, we conclude that most data in traffic systems are high-recurring. This thesis establishes a quantitative method to recognize those rare, high-value samples from large datasets based on estimated uncertainty. The domain knowledge on traffic flow and driving behaviours cannot be ignored in building models and quantifying the uncertainty.

The findings and methods proposed in this thesis have several important implications for practice. One major implication is that understanding predictability limits, supports better decision-making in terms of what to invest in: more model sophistication, more (diverse) data collection, or both (or neither). More specifically, for macroscopic traffic forecasting, speed, flow, and density data must be combined with the trajectory data and the demand data to break the accuracy bottleneck. For motion prediction, learning how to recognize the intention exposed by other drivers or directly communicate with other vehicles via connection is more valuable than developing more sophisticated predictive control algorithms.

From the discussion above, we close the summary by proposing several research directions. First, due to the importance of diversifying data sources, how to fuse different

data types needs more investigation in the traffic domain. Second, how implementing human knowledge represented by causal graphs into AI models is a promising solution that can potentially improve both the interpretability and the robustness of AI models. Third, how to continuously improve a model's performance without re-training after collecting enough rare samples is critical for processing the long-term data streams. We believe that answering these questions can pave the path to more application-oriented traffic prediction systems.

Samenvatting

Het observeren, modelleren, voorspellen en begrijpen van de dynamiek van verkeerssystemen op verschillende niveaus is een van de meest kritische onderwerpen in het vervoers- en planningsdomein. Op macroscopische schaal is verkeersopstopping het centrale probleem dat gevolgen heeft voor alle aspecten van de samenleving. Verkeersopstoppingen kosten kostbare reistijd, extra brandstofverbruik en frustratie in het dagelijks leven. Verkeersopstoppingen zijn niet altijd te vermijden, maar nauwkeurige voorspellingen van de verkeersomstandigheden in een wegennet zijn nuttig voor weggebruikers. Zo kunnen automobilisten snellere en veiligere routekeuzes maken op basis van de geschatte aankomsttijd en de voorspelde congestie-evolutie. Betrouwbare verkeersprognoses bieden ook essentiële informatie voor realtime verkeersregelsystemen en de ontwikkeling van duurzame mobiliteitssystemen op de lange termijn. Aan de andere kant, op microscopisch niveau, krijgt het modelleren van de interactie tussen weggebruikers en het voorspellen van hun gedrag steeds meer aandacht door de toenemende populariteit van autonoom rijden. Nauwkeurig anticiperen op de beslissingen van andere agenten is onmisbaar voor een veilig en soepel stuurautomaatsysteem. Dit actief bestudeerde domein is de focus geworden van veel wetenschappers uit de academische wereld en ingenieurs uit de industrie.

Onlangs, met de snelle ontwikkeling van detectietechnologie, zijn gegevensgestuurde modellen, met name met behulp van kunstmatige intelligentie (AI) -technieken, een van de meest populaire benaderingen geworden op het gebied van verkeersprognoses. De macroscopische verkeerstoestand wordt over het algemeen bewaakt door geïn- stalleerde sensoren, bijv. lusdetectoren, en beschreven door de overeenkomstige afgeleide variabelen, zoals verkeersstroom, dichtheid en gemiddelde snelheid. Deze gegevens verrijken ons begrip van grootschalige verkeerscongestiepatronen enorm. Het verzamelen van microscopische trajectgegevens omvat de brede toepassing van drones en sensoren in de auto, b.v. camera's, GPS, Lidar en radar. Met deze tools kunnen realtime trajecten van omringende agenten worden waargenomen en kaartinformatie worden gelokaliseerd. De opkomst van "big data" stimuleert in hoge mate de snelle iteratie van steeds nauwkeurigere AI-modellen. In de academische wereld zijn op AI gebaseerde verkeersprognoses en trajectvoorspellingsmodellen al de mainstream geworden. Deze AI-modellen laten indrukwekkende prestaties zien voor die grotere en complexere prognosetaken.

Dit "data+AI"-paradigma legt echter ook zijn zwakke punten in de praktijk bloot. Lage interpreteerbaarheid en onbevredigende betrouwbaarheid zijn de meest zorgwekkende problemen. De "black-box"-eigenschap van neurale netwerken en de kwetsbaarheid van op correlatie gebaseerde AI in nieuwe scenario's belemmeren verdere toepassingen in de echte wereld. Bovendien wordt het steeds moeilijker om de nauwkeurigheid

van verkeersprognoses te verbeteren. Daarom rijzen er natuurlijk twee vragen: Hoe voorspelbaar is het verkeer op verschillende niveaus? Zitten we dicht genoeg bij de grens van voorspelbaarheid? De antwoorden op deze vragen kunnen eindeloze prestatiebenchmarks in perspectief plaatsen en aangeven wat de meest waardevolle onderzoeksrichting in de toekomst is.

Dit proefschrift behandelt de kwestie van de voorspelbaarheid van verkeerssystemen op verschillende schaalniveaus door een systematische benadering voor te stellen voor het kwantificeren van de voorspellende onzekerheid van verschillende bronnen in AI-modellen. Daarnaast zijn ook de interpreteerbaarheid en de generaliseerbaarheid van de gebruikte deep learning-modellen verbeterd. Het proefschrift geeft een uitgebreide en kwantitatieve evaluatie van de voorspelbaarheid van verkeer.

De wetenschappelijke hiaten die in dit proefschrift worden behandeld, zijn ondervverdeeld in 4 met elkaar verweven delen. Ten eerste verbeteren we de interpreteerbaarheid van op deep learning gebaseerde verkeersvoorspellingsmodellen door een nieuwe dynamische module te ontwerpen die expliciet toestandsafhankelijke ruimtelijke associaties tussen wegverbindingen leert. Dit geeft inzicht in hoe een datagedreven model te combineren met verklaarbare verkeersstroomtheorie. De verkregen principes van modelontwerp zullen worden gebruikt in de volgende onzekerheidskwantificatiemethoden. Ten tweede schatten we de gemiddelde voorspelbaarheidslimiet van macroscopische verkeershoeveelheden voor verschillende soorten voorspellingsmodellen. De voorspelbaarheid wordt hier gemeten door de ondergrens van verschillende voorspellende fouten. We bewijzen deze ondergrenzen eerst theoretisch en gebruiken vervolgens een numeriek schema om ze rechtstreeks te schatten op basis van de gegeven dataset. Dit deel biedt een set modelvrije tools om te evalueren of het vergroten van de verfijning van het model de moeite waard is in termen van extra winst in nauwkeurigheid. In het derde deel, op basis van een reeks praktische vereisten, stellen we een methode voor voor het schatten van twee soorten input-specifieke voorspellende onzekerheidsmetrieken in macroscopische voorspelling van verkeerstoestanden. De twee soorten onzekerheidsmetrieken vertegenwoordigen respectievelijk de inherente willekeur in verkeersdynamiek en de zeldzaamheid van congestiepatronen. De resultaten geven inzicht in hoe inherent voorspelbaar verkeer is, hoe vaak zeldzame patronen voorkomen en waarom macroscopische voorspelbaarheid beperkt is. Deze conclusies zijn essentieel voor continue gegevensverzameling. Het laatste deel breidt de eerdere onzekerheidskwantificatiemethode uit tot het microscopische bewegingsvoorspellingsprobleem. We laten zien dat het inbedden van causale effecten de robuustheid van het bewegingsvoorspellingsmodel in nieuwe interactiescenario's aanzienlijk verbetert. We ontdekten ook dat het blootleggen van de beoogde aanwijzingen aan andere weggebruikers de sleutel is tot het verminderen van voorspellende onzekerheid en domeinkennis van rijgedrag is onmisbaar voor generaliseerbare AI-modellen. Hieronder worden de 4 delen afzonderlijk besproken.

Leer dynamische ruimtelijke associaties door ontworpen deep learning-modellen

Modellering is niet het enige belangrijke aspect van macroscopische verkeersprognoses. De beschikbare hoeveelheid en soorten gegevens beperken ook hoe nauwkeurig het verkeer kan worden voorspeld. Vanwege de beperkte waarneembaarheid van die kritieke factoren die mogelijk de verandering van verkeerssituatie veroorzaken, b.v. ge-

detailleerde verkeersvraagpatronen, routekeuzes, rijgedrag rond opritten en het weer, verkeersdynamiek vertoont aanzienlijke stochastiek. We hebben een methode nodig om de stochasticiteit kwantitatief te koppelen aan de limiet (ondergrens) van voorspellende nauwkeurigheid.

Dit proefschrift behandelt dit probleem vanuit zowel theoretische als praktische perspectieven. We bewijzen eerst wiskundig een ondergrens van nauwkeurigheid voor respectievelijk deterministische voorspellingsmodellen en alle probabilistische voorspellingsmodellen. Het sleutelconcept dat we gebruiken is de zogenaamde voorwaardelijke differentiële entropie en het ontwikkelen van een numeriek schema om het te schatten is noodzakelijk. In de praktijk is het direct schatten van voorwaardelijke entropie uit een multivariate tijdreeks een uitdaging vanwege de vloek van dimensionaliteit. Om dit probleem op te lossen, worden de ruimtelijk-temporele kenmerken verkregen uit het eerste deel van dit proefschrift gebruikt om de gehele dataset op te delen in tijdgebonden en locatiegerelateerde subsets. Door deze techniek te gebruiken, wordt de dimensionaliteit drastisch verminderd en is het mogelijk om de limiet van voorspellende nauwkeurigheid direct in te schatten. We laten zien dat, voor verkeersproductie op netwerkniveau en voorspelling van verkeerssnelheid, de huidige deep learning-modellen op korte termijn de geschatte theoretische limiet naderen. Deze benadering is ook informatief voor de meest onzekere tijdsperioden op één dag en de meest onzekere locaties in een wegennet. Deze voorgestelde methode biedt een hulpmiddel om te evalueren of de gegeven dataset mogelijk aan de nauwkeurigheidseis kan voldoen, en ook inzicht in de ruimtelijk-temporele verdeling van voorspelbaarheid. De laatste is informatief voor verkeersregelaars.

Schat de gemiddelde voorspelbaarheidslimiet voor macroscopische verkeersvariabelen

Modellering is niet het enige belangrijke aspect van macroscopische verkeersprognoses. De beschikbare hoeveelheid en soorten gegevens beperken ook hoe nauwkeurig het verkeer kan worden voorspeld. Vanwege de beperkte waarneembaarheid van die kritieke factoren die mogelijk de verandering van verkeerssituatie veroorzaken, b.v. gedetailleerde verkeersvraagpatronen, routekeuzes, rijgedrag rond opritten en het weer, verkeersdynamiek vertoont aanzienlijke stochastiek. We hebben een methode nodig om de stochasticiteit kwantitatief te koppelen aan de limiet (ondergrens) van voorspellende nauwkeurigheid.

Dit proefschrift behandelt dit probleem vanuit zowel theoretische als praktische perspectieven. We bewijzen eerst wiskundig een ondergrens van nauwkeurigheid voor respectievelijk deterministische voorspellingsmodellen en alle probabilistische voorspellingsmodellen. Het sleutelconcept dat we gebruiken is de zogenaamde voorwaardelijke differentiële entropie en het ontwikkelen van een numeriek schema om het te schatten is noodzakelijk. In de praktijk is het direct schatten van voorwaardelijke entropie uit een multivariate tijdreeks een uitdaging vanwege de vloek van dimensionaliteit. Om dit probleem op te lossen, worden de ruimtelijk-temporele kenmerken verkregen uit het eerste deel van dit proefschrift gebruikt om de gehele dataset op te delen in tijdgebonden en locatiegerelateerde subsets. Door deze techniek te gebruiken, wordt de dimensionaliteit drastisch verminderd en is het mogelijk om de limiet van voorspellende nauwkeurigheid direct in te schatten. We laten zien dat, voor verkeersproductie op net-

werkniveau en voorspelling van verkeerssnelheid, de huidige deep learning-modellen op korte termijn de geschatte theoretische limiet naderen. Deze benadering is ook informatief voor de meest onzekere tijdsperioden op één dag en de meest onzekere locaties in een wegennet. Deze voorgestelde methode biedt een hulpmiddel om te evalueren of de gegeven dataset mogelijk aan de nauwkeurigheidseis kan voldoen, en ook inzicht in de ruimtelijk-temporele verdeling van voorspelbaarheid. De laatste is informatief voor verkeersregelaars.

Schat de inherente stochastiek en de zeldzaamheid van de huidige verkeerstoestand

In dit deel analyseren we uitgebreid de "voorspelbaarheid" van macroscopisch verkeer door zowel de inherente stochasticiteit van verkeersdynamiek als het optreden van zeldzame congestiepatronen te kwantificeren. De bestaande ensembleleermethode wordt toegepast op problemen met het voorspellen van verkeerssnelheid. De experimentele resultaten laten zien dat, hoewel er altijd zeldzame of nieuwe congestiepatronen in de datastroom naar voren komen, de voorspelbaarheid van snelheid voornamelijk wordt beperkt door de inherente stochasticiteit van de verkeersdynamiek. Door de evolutie van de voorspelde snelheidsverdeling direct te assimileren, tonen we verder aan dat de voorspelbaarheid snel daalt met de voorspellingshorizon als gevolg van de splitsing van de toekomstige verkeerstoestand. Deze bimodaliteit veroorzaakt een hoge en onherleidbare onzekerheid en kan worden verklaard door de capaciteitsdaling in de verkeersstroomtheorie.

Kwantificeer de onzekerheid in bewegingsvoorspelling met een generaliseerbare methode

Anders dan macroscopische verkeersprognoses, is de uitvoer van een bewegingsvoorspellingstaak geen scalaire maar een 2D-coördinaat. Bovendien wordt de ruimtelijke verdeling in 2D beperkt door willekeurige weglay-outs. Momenteel zijn de belangrijkste drijvende kracht achter bewegingsvoorspelling de relevante toepassingen in autonome voertuigen. Van een autonoom voertuig wordt verwacht dat het nauwkeurig anticipeert op de intenties van andere weggebruikers wanneer het geconfronteerd wordt met verschillende scenario's die niet in de trainingsset zijn gezien. Daarom moet het bewegingsvoorspellingsmodel in staat zijn om zeldzaam gedrag te identificeren en zichzelf aan te passen aan nieuwe scenario's. Dit is belangrijk voor veiligheidskwesties.

Op basis van deze vereisten breidt dit proefschrift de methode voor het kwantificeren van scalaire onzekerheid uit naar een 2D-vlak. De ruimtelijke verdeling van de output wordt direct benaderd door een niet-parametrisch histogram in plaats van een eenvoudige prior. Het causale effect van omringende agentia wordt beschouwd door een cascadestrategie, zodat het model beter generaliseerbaar is naar nieuwe scenario's. We laten zien dat deze aanpak die zeldzame rijgedragingen kan vinden en dat de onzekerheid van interactie voornamelijk wordt veroorzaakt door de onbekende beoogde richting (aangegeven door de richtingaanwijzer) en rijstijlen. We ontdekten ook dat het AI-model niet kan leren hoe het rijgedrag op lage snelheid kan worden aangepast aan gevallen met hoge snelheid, omdat de domeinkennis ons vertelt dat de heterogeniteit

van gedrag toeneemt met de snelheid. Op basis van deze observaties stellen we voor dat het diversifiëren van de perceptie de sleutel is tot veiligere en soepelere bewegingsvoorspelling en planning.

Conclusie en perspectief

Samengevat behandelt dit proefschrift twee belangrijke kwesties: Hoe voorspelbaar is de verkeersdynamiek op verschillende niveaus? Hoe de onzekerheid in verkeersprognoses kwantificeren en deze gebruiken om de gegevensverzameling en -modellering te verbeteren? We beantwoorden deze twee onderzoeksvragen door een systematische benadering voor te stellen voor het kwantificeren van voorspellende onzekerheid in macroscopische voorspelling van verkeerstoestanden en microscopische baanvoorspelling.

Voor de eerste vraag naar voorspelbaarheid concluderen we dat voor zowel de voorspelling van de verkeerstoestand op netwerkniveau als de bewegingsvoorspelling, de limiet van voorspelbaarheid voornamelijk wordt beperkt door de onherleidbare willekeur die wordt veroorzaakt door de beperkte perceptiegegevenstypen. Momenteel naderen AI-modellen deze limiet. Een significante verbetering van de voorspellende nauwkeurigheid zal niet plaatsvinden tenzij we de beschikbare gegevenstypen diversifiëren. Voor macroscopisch verkeer kunnen dit traject- of vraaggegevens zijn. In de context van autonome voertuigen is het herkennen van de richtingaanwijzer belangrijk. Voor de tweede vraag concluderen we dat de meeste data in verkeerssystemen hoogrecurrent zijn. Dit proefschrift stelt een kwantitatieve methode vast om die zeldzame, hoogwaardige monsters uit grote datasets te herkennen op basis van geschatte onzekerheid. De domeinkennis over verkeersafwikkeling en rijgedrag kan niet worden genegeerd bij het bouwen van modellen en het kwantificeren van de onzekerheid.

De bevindingen en methoden die in dit proefschrift worden voorgesteld, hebben verschillende belangrijke implicaties voor de praktijk. Een belangrijke implicatie is dat het begrijpen van voorspelbaarheidslimieten een betere besluitvorming ondersteunt in termen van waarin te investeren: meer modelverfijning, meer (diverse) gegevensverzameling, of beide (of geen van beide). Meer specifiek, voor macroscopische verkeersprognoses moeten snelheids-, stroom- en dichtheidsgegevens worden gecombineerd met de trajectgegevens en de vraaggegevens om de nauwkeurigheidsknelpunt te doorbreken. Voor bewegingsvoorspelling is het waardevoller om te leren hoe de intentie van andere bestuurders te herkennen of rechtstreeks met andere voertuigen te communiceren via verbinding, dan het ontwikkelen van meer geavanceerde voorspellende besturingsalgoritmen.

Uit de bovenstaande discussie sluiten we de samenvatting af met een aantal onderzoeksrichtingen. Ten eerste, vanwege het belang van diversificatie van gegevensbronnen, is meer onderzoek nodig in het verkeersdomein hoe verschillende gegevenstypen kunnen worden samengevoegd. Ten tweede, hoe het implementeren van menselijke kennis, weergegeven door causale grafieken in AI-modellen, een veelbelovende oplossing is die zowel de interpreteerbaarheid als de robuustheid van AI-modellen kan verbeteren. Ten derde, hoe de prestaties van een model continu kunnen worden verbeterd zonder opnieuw te trainen nadat voldoende zeldzame monsters zijn verzameld, is van cruciaal belang voor het verwerken van de gegevensstromen op de lange termijn.

Wij zijn van mening dat het beantwoorden van deze vragen de weg kan effenen naar meer toepassingsgerichte verkeersvoorspellingssystemen.

About the Author

Guopeng Li was born in Dandong, China. He had his childhood in this small but beautiful city. In 2011, he moved to Nanjing, a prominent place in Chinese history and culture, and started his studies in Physics at Nanjing University. After spending 4 years full of knowledge and happiness, he completed his bachelor's degree with a thesis about the Raman effect in June 2015.

In July 2015, Guopeng went to France and started his engineer diploma program and master's studies at ENSTA-Paris. Between the studies, he did an internship at CEA (The French Alternative Energies and Atomic Energy Commission). In 2018, he obtained the engineer diploma and master's degree in applied mathematics.

In February 2019, He joined DiTTlab (Data Analytics and Traffic Simulation), as a PhD candidate. His research project, MIRRORS (Multiscale Integrated Traffic Observatory for Large Road Networks), is funded by NDW (National Data Warehouse). His research focuses on data-driven uncertainty quantification in traffic predictions at multiple scales. He combines deep learning models and uncertainty quantification methods to study different types of uncertainty in macroscopic traffic state prediction and microscopic driving behaviour prediction.

Journal Articles

1. Li, G., V. L. Knoop, H. van Lint (2021) Multistep traffic forecasting by dynamic graph convolution: Interpretations of real-time spatial correlations, *Transportation Research Part C: Emerging Technologies*, 128, p. 103185 - published

2. Li, G., V. L. Knoop, H. van Lint (2022a) Estimate the limit of predictability in short-term traffic forecasting: An entropy-based approach, *Transportation Research Part C: Emerging Technologies*, 138, p. 103607 - published

Peer-reviewed Conference Contribution

1. Guopeng, L., V. L. Knoop, H. van Lint (2020) Dynamic graph filters networks: A gray-box model for multistep traffic forecasting, in: *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, pp. 1–6

Articles Under Review

1. Li, G., Z. LI, V. Knoop, H. van Lint (2022b) Uqnet: Quantifying uncertainty in trajectory prediction by a non-parametric and generalizable approach, *Available at SSRN 4241523*

2. Li, G., V. L. Knoop, J.W.C. van Lint (2022) How predictable are macroscopic traffic states: a perspective of uncertainty quantification, *Under review*

TRAIL Thesis Series

The following list contains the most recent dissertations in the TRAIL Thesis Series. For a complete overview of more than 275 titles see the TRAIL website: www.rsTRAIL.nl.

The TRAIL Thesis Series is a series of the Netherlands TRAIL Research School on transport, infrastructure and logistics.

Li, G., *Uncertainty Quantification and Predictability Analysis for Traffic Forecasting at Multiple Scales*, T2023/5, April 2023, TRAIL Thesis Series, the Netherlands

Harter, C., *Vulnerability through Vertical Collaboration in Transportation: A complex networks approach*, T2023/4, March 2023, TRAIL Thesis Series, the Netherlands

Razmi Rad, S., *Design and Evaluation of Dedicated Lanes for Connected and Automated Vehicles*, T2023/3, March 2023, TRAIL Thesis Series, the Netherlands

Eikenbroek, O., *Variations in Urban Traffic*, T2023/2, February 2023, TRAIL Thesis Series, the Netherlands

Wang, S., *Modeling Urban Automated Mobility on-Demand Systems: an Agent-Based Approach*, T2023/1, January 2023, TRAIL Thesis Series, the Netherlands

Szép, T., *Identifying Moral Antecedents of Decision-Making in Discrete Choice Models*, T2022 /18, December 2022, TRAIL Thesis Series, the Netherlands

Zhou, Y. *Ship Behavior in Ports and Waterways: An empirical perspective*, T2022/17, December 2022, TRAIL Thesis Series, the Netherlands

Yan, Y., *Wear Behaviour of A Convex Pattern Surface for Bulk Handling Equipment*, T2022/16, December 2022, TRAIL Thesis Series, the Netherlands

Giudici, A., *Cooperation, Reliability, and Matching in Inland Freight Transport*, T2022/15, December 2022, TRAIL Thesis Series, the Netherlands

Nadi Najafabadi, A., *Data-Driven Modelling of Routing and Scheduling in Freight Transport*, T2022/14, October 2022, TRAIL Thesis Series, the Netherlands

Heuvel, J. van den, *Mind Your Passenger! The passenger capacity of platforms at railway stations in the Netherlands*, T2022/13, October 2022, TRAIL Thesis Series, the Netherlands

Haas, M. de, *Longitudinal Studies in Travel Behaviour Research*, T2022/12, October 2022, TRAIL Thesis Series, the Netherlands

Dixit, M., *Transit Performance Assessment and Route Choice Modelling Using Smart Card Data*, T2022/11, October 2022, TRAIL Thesis Series, the Netherlands

Du, Z., *Cooperative Control of Autonomous Multi-Vessel Systems for Floating Object Manipulation*, T2022/10, September 2022, TRAIL Thesis Series, the Netherlands

Larsen, R.B., *Real-time Co-planning in Synchromodal Transport Networks using Model Predictive Control*, T2022/9, September 2022, TRAIL Thesis Series, the Netherlands

Zeinaly, Y., *Model-based Control of Large-scale Baggage Handling Systems: Leveraging the theory of linear positive systems for robust scalable control design*, T2022/8, June 2022, TRAIL Thesis Series, the Netherlands

Fahim, P.B.M., *The Future of Ports in the Physical Internet*, T2022/7, May 2022, TRAIL Thesis Series, the Netherlands

Huang, B., *Assessing Reference Dependence in Travel Choice Behaviour*, T2022/6, May 2022, TRAIL Thesis Series, the Netherlands

Reggiani, G., *A Multiscale View on Bikeability of Urban Networks*, T2022/5, May 2022, TRAIL Thesis Series, the Netherlands

Paul, J., *Online Grocery Operations in Omni-channel Retailing: opportunities and challenges*, T2022/4, May 2022, TRAIL Thesis Series, the Netherlands

Liu, M., *Cooperative Urban Driving Strategies at Signalized Intersections*, T2022/3, January 2022, TRAIL Thesis Series, the Netherlands

Feng, Y., *Pedestrian Wayfinding and Evacuation in Virtual Reality*, T2022/2, January 2022, TRAIL Thesis Series, the Netherlands

Scheepmaker, G.M., *Energy-efficient Train Timetabling*, T2022/1, January 2022, TRAIL Thesis Series, the Netherlands

Bhoopalam, A., *Truck Platooning: planning and behaviour*, T2021/32, December 2021, TRAIL Thesis Series, the Netherlands

Hartleb, J., *Public Transport and Passengers: optimization models that consider travel demand*, T2021/31, September 2021, TRAIL Thesis Series, the Netherlands

Azadeh, K., *Robotized Warehouses: design and performance analysis*, T2021/30, February 2021, TRAIL Thesis Series, the Netherlands

Na, C., *Coordination Strategies of Connected and Automated Vehicles near On-ramp Bottlenecks on Motorways*, T2021/29, December 2021, TRAIL Thesis Series, the Netherlands

Onstein, A.T.C., *Factors influencing Physical Distribution Structure Design*, T2021/28, September 2021, TRAIL Thesis Series, the Netherlands

Olde Kalter, M.-J. T., *Dynamics in Mode Choice Behaviour*, T2021/27, November 2021, TRAIL Thesis Series, the Netherlands

Los, J. *Solving Large-Scale Dynamic Collaborative Vehicle Routing Problems: an Auction-Based Multi-Agent Approach*, T2021/26, November 2021, TRAIL Thesis Series, the Netherlands