Enhancing Traffic Prediction Models with Correlation among Traffic Variables

by

Haoyang Chen

Student ID: 1276594

supervised by

Joyce Zhang

Jianzhong Qi

Mingming Gong

Tingjin Chu

A thesis submitted in partial fulfillment for the
degree of Master of Computer Science

in the
Faculty of Engineering and Information Technology
School of Computing and Information Systems
THE UNIVERSITY OF MELBOURNE

October 2023

Abstract

In the past decade, data-driven methods have exhibited remarkable success in the field of traffic prediction. Among the various model architectures available, spatiotemporal neural networks have emerged as a highly effective framework for capturing spatial and temporal dependencies. However, there are three primary limitations within the current body of research. First, the majority of studies have concentrated solely on individual traffic variables, thereby overlooking the potential insights that could be derived from exploring correlations among multiple variables. Second, many studies have predominantly relied on data-driven approaches, neglecting the fundamental physical laws governing traffic systems. In fact, there is a fundamental relationship between traffic speed, flow, and density, which remains largely under-explored. Third, existing predictive models often restrict their forecasts to a fixed time horizon of one hour ahead. This immediate-term prediction is much less applicable in real application scenarios, where long-term traffic forecasting would be preferred.

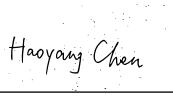
In this thesis, we explore the potential benefits derived from leveraging the correlations between traffic speed and flow. To achieve this, we extended two prominent spatiotemporal neural networks, namely GMAN and DDGCRN, through inductive and learning biases. First, as part of the inductive bias, we expanded the model dimensions to accommodate multiple traffic variables. This expansion allows us to capture a broader range of traffic characteristics. Second, we employed novel physical embedding and physical attention modules to encode the relationships between these variables, ensuring a more comprehensive understanding of their interactions. Third, we introduced a dynamic correlation learning mechanism, enabling the model to adapt to changing traffic conditions effectively. Finally, we introduced a physics-informed loss constraint as the learning bias. This constraint serves as a soft regularisation that facilitates the model's convergence towards the underlying physical principles governing traffic behaviours.

The experimental results on two of the latest real-world traffic datasets demonstrate the effectiveness of utilising correlation between variables for traffic prediction. Specifically, our extended models attain higher predictive accuracy over a range from 12- to 36-step time horizons compared to all baseline models and exhibit robust performance even when trained on scarce data samples.

Declaration of Authorship

I certify that:

- this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text.
- where necessary I have received clearance for this research from the University's Ethics Committee and have submitted all required data to the School.
- the thesis is 25,276 words in length (excluding text in images, tables, bibliographies and appendices.

Signed: 

Date: 30 October 2023

Acknowledgements

First and foremost, I would like to express sincere gratitude to my supervisors, Dr. Joyce Zhang, Dr. Jianzhong Qi, Dr. Mingming Gong and Dr. Tingjin Chu for their continuous support and invaluable suggestions provided throughout my thesis project. It is their immense knowledge and insightful feedback that guided me through this challenging journey.

I would like to thank my parents, without whose encouragement I would not be able to persist through all the stressful times in university. I would like to thank Noah Scotti for his timely help at the beginning of the project, and for helping me settle down in Melbourne Connect. I would also like to thank Xinyu and Yanchuan for their technical support on my coding scripts.

This research was supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative.

Contents

Abstract i

Declaration of Authorship ii

Acknowledgements iii

List of Figures vii

List of Tables viii

Abbreviations ix

Symbols xi

1 Introduction 1
- 1.1 Background and Research Gaps . . . 1
- 1.2 Aims and Objectives . . . 4
- 1.3 Scope . . . 6
- 1.4 Thesis Outline . . . 6

2 Related Works 8
- 2.1 Traffic Data and Analysis Methods . . . 9
  - 2.1.1 Traffic Data . . . 9
  - 2.1.2 Traffic State Estimation . . . 12
- 2.2 Traffic Prediction Models . . . 13
  - 2.2.1 Statistical Models . . . 13
  - 2.2.2 Machine Learning Models . . . 14
  - 2.2.3 Deep Learning Models . . . 16
  - 2.2.4 Meta Learning Models . . . 18
  - 2.2.5 Trends and Limitations . . . 19
- 2.3 Physics of Traffic Flow . . . 20
  - 2.3.1 Traffic Flow Modelling . . . 21
  - 2.3.2 The Fundamental Relationship and Diagram . . . 23
  - 2.3.3 Partition of Traffic Network . . . 26
- 2.4 Physics-informed Neural Network . . . 28
  - 2.4.1 The Fundamental Principles . . . 29

2.4.2 Trends and Methodologies . . . 30
2.4.3 Challenges . . . 32
2.5 Summary . . . 33

3 Preliminaries **37**
3.1 Hybrid Computational Graph . . . 38
3.2 Governing Equations and Automatic Differentiation . . . 38
3.3 Notations and Definitions . . . 39
3.4 Problem Formulation . . . 40

4 Methodology **42**
4.1 Data Collection and Processing . . . 43
4.2 Modelling Prediction Bias . . . 46
4.2.1 Rationale for Bias Selection . . . 46
4.2.2 Inductive Bias for GMAN and DDGCRN . . . 47
4.2.3 Learning Bias for Correlation Learning . . . 48
4.3 Model Dimension Expansion . . . 49
4.4 Inductive Bias . . . 50
4.4.1 Physical Embedding . . . 50
4.4.2 Physical Self-Attention . . . 51
4.4.3 Weighted Attention Fusion . . . 53
4.4.4 Dynamic Correlation Generation . . . 54
4.5 Learning Bias . . . 56
4.5.1 PIDL Framework Overview . . . 56
4.5.2 Traffic Network Parition . . . 57
4.5.3 Formulation using Greenshield's Model . . . 61
4.5.4 Formulation using LWR Conservation Law . . . 62
4.5.5 Curve Fitting . . . 62
4.5.6 Formulation using Fitted Equations . . . 64

5 Experiments **66**
5.1 System and Platforms . . . 67
5.2 Settings and Baselines . . . 67
5.3 Overall Results . . . 70
5.4 Ablation Study on Longer Time Horizons . . . 73
5.5 Different Categories of Physics for Learning Bias . . . 75
5.5.1 Accuracy vs Convergence . . . 75
5.5.2 Choice of Weight of Physics . . . 78
5.6 Performance on Scarse Training Samples . . . 79
5.6.1 Overall Result . . . 79
5.6.2 Accuracy-size Trade-off . . . 81
5.6.3 Accuracy-efficiency Trade-off . . . 82
5.7 Hyperparameter Analysis . . . 84
5.7.1 Number of Physical Attention Heads . . . 84
5.7.2 Physical Embedding Dimension . . . 86
5.8 Discussion . . . 88

6 Conclusions **90**
6.1 Summary 90
6.1.1 Contributions 91
6.1.2 Limitations 92
6.2 Future Works 93
6.2.1 Embedding Observational Biases 93
6.2.2 Alternative Physical Loss Functions 94
6.2.3 Utilisation of More Traffic Variables 94
6.2.4 Benchmarking with More Datasets and Models 95

Bibliography **97**

List of Figures

1.1 Trends of traffic flow and speed of selected road links in Melbourne (left) and California (right) . . . 2

2.1 Traffic signal volume every 15 minutes from 03.01 - 03.14 in Melbourne . 11
2.2 Greenshield's Fundamental Diagram (from Archie J. Huang et al. [36]) . . 24
2.3 Daganzo's Fundamental Diagrams (from Archie J. Huang et al. [36]) . . . 25
2.4 Inverse Lambda Fundamental Diagrams (from Archie J. Huang et al. [36]) 26
2.5 Balance of Data-driven and Model-driven Approaches (from Di et al [18]) 28

3.1 Example Hybrid Computational Graph for Physics-informed Architecture (from Di et al. [18]) . . . 39

4.1 Range of PeMS District 7 . . . 43
4.2 PeMSD7 Station Distribution (left) and Number Counting Clusters (right) 44
4.3 Melbourne Site Distribution and Number Counting Clusters . . . 45
4.4 Spatio-temporal-physical Embedding (STPE) . . . 51
4.5 STP-Attention Block . . . 54
4.6 Dynamic Correlation Generation Process . . . 56
4.7 Flow of PIDL . . . 57
4.8 Overall Framework of PIDL for Spatiotemporal Traffic Prediction Models 58
4.9 Test Region in Melbourne with Color-coded Links based on Clustering Results (obtained from Zhang et al. [71]) . . . 58
4.10 Snake Algorithm (obtained from Mohammadreza et al. [96]) . . . 60
4.11 Example Parameter Estimation for PeMS Freeway 10E . . . 61
4.12 Fitted Curves for Melbourne (left) and PeMS (right) datasets . . . 65

5.1 Comparison of Errors Over Time on PeMS (left) and Melbourne (right) dataset . . . 72
5.2 Error vs Sample Size for PI-models . . . 81
5.3 Accuracy-Time Trade-off . . . 82
5.4 Distribution of Variables with Varying Sample Sizes . . . 83
5.5 Accuracy and Physics Cost vs. Number of Attention Heads . . . 85
5.6 Accuracy vs Embedding Dimension . . . 87

List of Tables

4.1 Dataset Description 46
4.2 PeMS Cluster Results and Parameter Estimation 60
4.3 Melbourne Cluster Results and Parameter Estimation 61

5.1 Overall Results for Short-term Prediction (**bold**: best result and underline: second best result) 71
5.2 Overall Results for Long-term Prediction (**bold**: best result and underline: second best result) 71
5.3 GMAN Performance (**bold**: best result) 74
5.4 DDGCRN Performance (**bold**: best result) 74
5.5 Performance of PIDL with Different Physics Models (**bold**: best result) . 76
5.6 Performance with Different Weights of Physics-Informed Loss (**bold**: best result) 78
5.7 Performance of PI-GMAN on Scarce Training Samples 80
5.8 Performance of PI-DDGCRN on Scarce Training Samples 80

Abbreviations

ARZ	**A**w-**R**ascle-**Z**hang
ARIMA	**A**uto**R**egressive **I**ntegrated **M**oving **Average**
CTM	**C**ell **T**ransmission **M**odel
CNN	**C**onvolutional **Neural Network**
DCRNN	**D**iffusion **C**onvolutional **R**ecurrent **N**eural **N**etwork
DDGCRN	**D**ecomposition **D**ynamic **G**raph **C**onvolutional **R**ecurrent **N**etwork
DNN	**D**eep **Neural Network**
GMAN	**G**raph **M**ulti-**A**ttention **N**etwork
GMM	**G**aussian **M**ixture **M**odel
GC-VAR	**G**ranger-**C**ausality **V**ector **A**uto **R**egressiion
GNN	**G**raph **Neural Network**
GP	**G**aussian **P**rocess
GPS	**G**lobal **P**ositioning **S**ystem
HA	**H**istorial **A**verage
HCG	**H**ybrid **C**omputational **G**raph
ITS	**I**ntelligent **T**ransportation **S**ystem
KF	**K**alman **F**ilter
KNN	**K**- **N**earest **N**eighbour
LSTM	**L**ong **S**hort-**T**erm **M**emory
LWR	**L**ighthill-**W**hitham-**R**ichards
LSSVM	**L**east **Square S**upport **V**ector **M**achine
MAPE	**M**ean **A**bsolute **P**ercentage **E**rror
ML	**M**achine **L**earning
MFD	**M**acrosopic **F**undamental **D**iagram
MSE	**M**ean **S**quared **E**rror

MTGNN	**M**ultivariate **T**ime-series **G**raph **N**eural **N**etworks
NN	**N**eural **N**etwork
PDE	**P**artial **D**ifferential **E**quations
PeMS	**P**erformance **M**easurement **System**
PINN	**P**hysics-**I**nformed **N**eural **N**etworks
PIML	**P**hysics-**I**nformed **M**achine **L**earning
PUNN	**P**hysics-**U**ninformed **N**eural **N**etworks
RMSE	**R**oot **M**ean **S**quared **E**rror
RNN	**R**ecurrent **N**eural **N**etwork
STGCN	**S**patio-**T**emporal **G**raph **C**onvolutional **N**etworks
TSE	**T**raffic **S**tate **Estimation**
VAR	**V**ector **A**uto **R**egression

Symbols

v	speed	kilometers per hour (km/h)
$v_f(v_{free})$	free flow speed	kilometers per hour (km/h)
q	flow rate	vehicles per hour (v/h)
M	accepted threshold for physics-informed loss	
D	degree matrix of a graph	
W	weight matrix of a graph	
I	identity matrix	
ρ	density	vehicles per kilometer (v/km)
$\rho_m(\rho_{max})$	jamming density (maximum density)	
α	weight of losses	

Chapter 1

Introduction

In this section, we will present an overview of the foundational elements of this thesis. The structure will be as follows: first, we will provide a brief background to contextualise the significance of traffic prediction. Second, we will explore the research gaps, shedding light on the limitations of existing studies. Following that, we will propose our main research questions and outline the aim and objectives of this research. We will also briefly address the research scope. Lastly, we will provide an outline of this thesis.

1.1 Background and Research Gaps

As part of the smart city project and Intelligent Transportation System (ITS), traffic forecasting plays a crucial role in urban traffic planning, control and management [2, 73]. The task focuses on predicting future traffic states, such as speed and flow, within specified periods and regions provided with historical traffic information and possibly other external data (weather, accidents, etc.). With the rise of deep learning techniques and the increasing availability of traffic data, studies on traffic forecasting have gradually shifted focus from statistical models to deep learning approaches [112]. Enormous research effort devoted to the area in the past decades focuses on capturing the complex spatiotemporal dependencies within traffic data and has pushed the boundary of the research field [31, 132, 137].

Despite the great success of deep-learning-based methods for traffic prediction, several deficiencies are still present. First, many existing studies concentrate solely on an

individual traffic variable, such as speed or flow. However, there are substantial inter-dependencies among these variables that are under-utilised. Notably, a strong negative correlation exists between two of the most commonly studied traffic parameters: speed and flow. Taking links in Melbourne and California as examples, Figure 1.1 shows the dynamic changes in these variables after standard normalisation over one-day time span, which demonstrates the existence of the relationship. Similar correlations can be observed among other traffic parameters, including traffic density, travel time, and headway. Second, existing approaches rely heavily on sufficient data sources and learnable short-term patterns. The availability of complete and accurate traffic data for deep learning remains scarce, primarily restricted to publicly accessible benchmark datasets like PeMS and METR-LA [55]. Additionally, most real-world traffic applications necessitate long-term forecasting abilities. However, this area remains largely underdeveloped and requires increased attention and innovative approaches. Finally, most existing deep-learning approaches for traffic prediction are purely data-driven, largely ignoring the underlying traffic flow physics that have been studied for decades [29, 58].

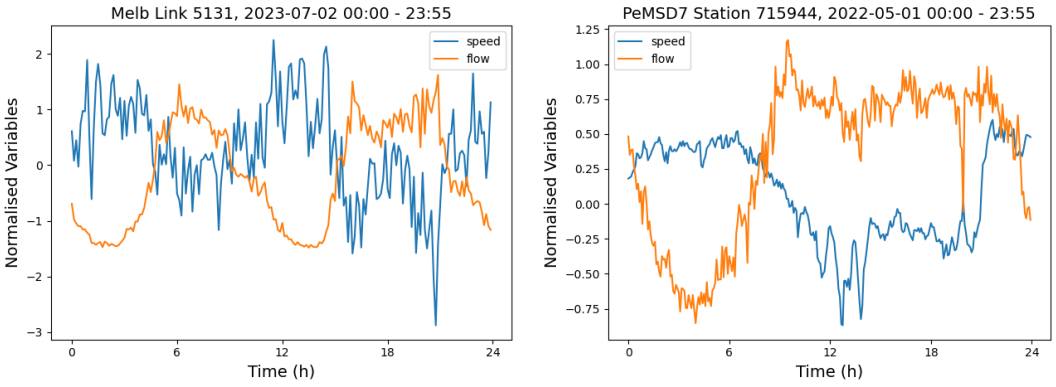


FIGURE 1.1: Trends of traffic flow and speed of selected road links in Melbourne (left) and California (right)

Compared to research efforts in designing spatiotemporal architectures, embedding physics into deep learning models remains under-explored. For the task of traffic prediction, model-driven approaches and data-driven approaches stay largely separated. In recent years, physics-informed neural networks (PINNs) have received widespread attention in various research fields [16, 36, 38, 87]. Researchers have identified three main pathways to embedding physics, including transforming or augmenting the input data, modifying the model structure or enforcing physical laws on loss constraints, known as observational, inductive and learning biases respectively [46]. Despite several attempts with

this structure in traffic prediction [36, 108], few have provided insights into incorporating inductive bias into neural networks. In attempts to embed learning biases, the physical equations are usually calibrated with empirical observations or experimented on synthetic datasets, which cannot accurately describe the real distribution of application scenarios. Furthermore, none of the research has demonstrated a scalable and generalised framework for traffic networks, thus preventing the application of the methods in spatiotemporal domains. Lastly, although data sources tend to fall into an increasing amount of categories (e.g. satellite image data, probe vehicle, loop detectors), few studies have explored physics derived from multiple traffic variables. With single input and output, models can hardly benefit from the correlations. Aside from the strategic deficiencies, the experimental settings of the previous research are also limited. Many contemporary studies have predominantly fixated their gaze upon a confined 12-step forecasting horizon, often neglecting the profound implications that longer-term predictions can unveil within the realm of traffic analysis.

This research aims to exploit correlations among traffic variables through inductive and learning biases to enhance predictive modelling. We will investigate the following key perspectives:

1. Expanding Model Dimensions: To harness the interplay of multiple variables, we will expand both the input and output dimensions of existing models. This expansion will facilitate a more comprehensive exploration of the relationships between these variables and create possibilities for innovative encoding strategies.

2. Incorporating Physics-Informed Techniques to Leverage Correlations: Integrating inductive and learning biases into deep learning architectures for traffic prediction is relatively unexplored but promising. We will introduce a physical attention module designed to capture dynamic correlations between traffic speed and flow at different timestamps and locations. This module will serve as an inductive bias that modifies the model architecture. Furthermore, we will incorporate equations derived from the physical characteristics of the datasets into the loss constraints. This will guide the model toward adhering to underlying physical laws, enhancing its predictive capabilities.

3. Experiments with Scarce Datasets and Extended Time Horizons: To demonstrate the practical utility of our approach, we will assess the model's performance under

conditions of incomplete data and longer time horizons. This experimentation aims to showcase the robustness of physics-informed techniques while highlighting the information gain resulting from correlations between traffic speed and flow.

These techniques may offer various strengths and weaknesses, thereby complementing existing deep learning-based methods. We will start by reviewing existing literature on pure deep-learning techniques and the physics-informed approaches for traffic prediction. Subsequently, we will identify suitable methods for embedding the physics and utilising variable correlations within these models. Next, we will conduct a comparative analysis among different extended models based on our extensive experiments. We will also provide trade-offs for each method adopted. Finally, we will summarise the research findings, main contributions and limitations of this thesis and provide potential future research directions. By covering these steps, we hope to provide insights into leveraging correlation between variables for traffic prediction, as well as facilitating the integration of data-driven and model-driven approaches.

1.2 Aims and Objectives

For traffic prediction, existing deep learning approaches have primarily concentrated on enhancing neural network structures while neglecting the crucial traffic flow theory that underpins traffic physics. Conversely, research that introduces physics-informed techniques for traffic prediction has placed significant emphasis on crafting physics models, often overlooking the importance of dedicated data-driven architectures. Consequently, these two domains have largely remained isolated from each other. Additionally, current methods predominantly rely on a single variable as input, disregarding the potential benefits of incorporating multiple traffic variables to leverage their correlations. Lastly, the existing methods have been fixated on a prediction time horizon of 12 steps (one hour), thus overlooking the performance of long-term forecasting. To bridge the existing gaps in traffic prediction methods, the overall aim of our project is:

Aim: Leveraging Traffic Variable Correlations for Enhanced Traffic Prediction Models

With this aim in mind, our investigation is guided by two key research questions:

1. How can the correlations among traffic variables be harnessed to improve the performance of neural networks in the context of traffic prediction?

2. What are the advantages, disadvantages, and complements of these techniques in terms of accuracy, efficiency and robustness?

This thesis further consists of three objectives, in order to achieve the aim and answer the questions:

Objective 1: Validate the Effectiveness of Leveraging Multiple Traffic Variables

We practically demonstrate it by expanding existing model dimensions and comparing the model performance with those that only accept single input and output. We further investigate the effectiveness of the approach on longer time horizons.

Hypothesis: Leveraging multiple traffic variables makes the model robust with respect to long-term prediction accuracies.

Objective 2: Incorporate Physics-Informed Techniques to Further Enhance the Correlation Gain From Traffic Variables

If the physics-informed techniques are applicable to traffic prediction with deep learning, we evaluate the model's prediction accuracy and robustness to sparse datasets.

Hypothesis: Physics-informed neural networks are more robust than either model-driven or data-driven approaches alone under the condition of sparse training data.

For evaluation purposes, we further propose another hypothesis, which guides the overall aim and objectives:

Hypothesis: Different ways of leveraging correlations among traffic variables have different impacts on the new traffic prediction model, demonstrating its complementarity, and additional merits or demerits.

1.3 Scope

This thesis focuses on the feasibility of leveraging multiple traffic variables. Physics-informed approaches as well as direct modifications in model architectures to reflect relationship properties are explored.

In the context of designing physics-informed loss functions, our strategy entails a focused exploration. Specifically, we will investigate the equation encapsulated by the fitted curve corresponding to the specific traffic speed-flow distribution instance. We will refrain from delving into advanced physics-informed laws, such as the ARZ model [4] or other second-order derivative functions, as they are not directly aligned with the objectives of this thesis.

In the process of incorporating physical attention into existing models, we are not addressing the algorithms related to capturing the spatiotemporal correlations of the variables. A plethora of research has been conducted on problems related to the dynamic relationships [31, 132, 137].

1.4 Thesis Outline

The rest of the thesis is organised into five main chapters, each contributing to a unique perspective of our exploration.

Chapter 2 takes a deep dive into the existing literature, covering topics of traffic data sources and analysis techniques, traffic state estimation, traffic prediction models, the physics of traffic flow, and the application of physics-informed neural networks.

Chapter 3 establishes a solid foundation by introducing essential notations, definitions, and the problem formulation central to our traffic prediction research.

Chapter 4 delves into the methodologies employed, encompassing the intricacies of data collection and processing, model dimension expansion, the incorporation of physical attention as an inductive bias, and the development of physics-informed loss functions as a learning bias.

Chapter 5 first provides insights into our experimental setup, offering a detailed account of the systems and platforms, training configurations, selected baselines and evaluation metrics. Subsequently, the chapter presents our findings and discussions, offering interpretations, comparisons, and valuable insights derived from our research.

Chapter 6 encapsulates the essence of our thesis, providing a comprehensive set of conclusions summarising our key contributions, acknowledging limitations, and outlining potential avenues for future exploration.

Finally, the thesis is supported by a comprehensive reference list.

Chapter 2

Related Works

Traffic prediction is the task of predicting future traffic conditions based on historical information. It can be categorised into a list of sub-tasks, including traffic state prediction, trajectory prediction, estimated time of arrival and map matching [114]. As one of the core technologies of Intelligent Transportation Systems (ITS), traffic state prediction has received widespread attention over the past few decades [74, 104, 112]. The term "state" refers to traffic variables that describe characteristics of traffic on a road segment or a road network, such as speed, flow and density. This research will mainly focus on the task of traffic state prediction. To provide a more detailed review, the following four topics will be covered:

Data sources and analysis methods. This section will focus on the common categories of data sources used for traffic prediction. We will also delve into how these data sources can be effectively represented and leveraged in the context of traffic prediction.

Categories of traffic prediction models. In the second section, traffic prediction models will be categorised into four main groups: mathematical or statistical models, traditional machine learning models, deep learning models, and meta-learning-based models. Each category will be thoroughly explored, including an examination of common algorithms, their strengths, weaknesses, and practical applications within the field.

Physics of traffic flow. In the third section, we will introduce the essential concept of traffic flow theory, which serves as the theoretical underpinning of traffic prediction. Key components, including the fundamental diagram and macroscopic traffic flow models,

will be elucidated, showcasing their vital role in enhancing the accuracy of traffic flow modelling.

Physics-Informed Neural Networks (PINNs). The final section will introduce the concept of physics-informed neural networks within the realm of traffic prediction. We will explore its applications, delineate its limitations, and provide insights into future developmental trends within this approach.

Finally, we will provide a summary of methodologies and trends in previous literature, the research gaps, as well as potential research directions.

2.1 Traffic Data and Analysis Methods

In this section, we will highlight the most common traffic data types and their applications in the field. The encoding strategies of time series and spatial information in the previous literature will also be reviewed. In addition, we will introduce key traffic analysis concepts related to transportation studies.

2.1.1 Traffic Data

Traffic data is crucial for measuring system performance. According to the AusRoads' latest publication [3], survey data can be categorized into three main types: point, linear, and area data. Point data pertains to statistics collected at a point scale, such as vehicles, pedestrians, or cyclists. Common variables falling within this category include traffic flow and speed. Linear data captures properties recorded along a road segment, encompassing parameters like travel time, delay, and queueing conditions. On the other hand, area data provides insights into traffic conditions over a broader area, including origin-destination patterns, parking data, and traffic generation scenarios. Another type of categorisation divides the data into macroscopic and microscopic variables, which distinguishes between point-level and network-scale observations [67, 138]. With this categorisation, speed and flow can fall into both categories. Most transportation studies focus on macroscopic variables to explore network-wide traffic patterns.

Point data is typically collected through loop detectors, probe vehicle detectors, or satellite imagery [80]. On the other hand, linear data can be obtained through GPS

tracking, traffic cameras, or Bluetooth technology [84]. When it comes to gathering data on a larger scale, for area-level analysis, specialised methods are adopted. These methods may involve the use of parking sensors, smartphone apps, or traffic simulation models [85]. Nevertheless, owing to constraints imposed by physical devices or unforeseen events, the data frequently exhibit partial observations or gaps. In addition, restrictions on data storage and detector communication can also introduce imprecision and noise. Furthermore, a significant number of data sources are acquired through an aggregated approach, typically spanning 5 to 15 minutes, thereby worsening the measurement results [136].

One of the most essential steps in transportation studies is the effective representation of the temporal and spatial information of traffic data. An informative and efficient encoding of input not only assists with the model's prediction process but also lowers the training cost acquired.

In the temporal domain, data exhibit clear periodicity and continuity characteristics [7]. The periodicity commonly manifests as daily, weekly, or seasonal patterns. Furthermore, there is a noticeable distinction between traffic flow and speed on weekdays versus weekends. For instance, as depicted in Figure 2.1, a distinct daily traffic flow pattern emerges, with significant variations between day and night attributed to living habits. Moreover, weekend traffic volumes are considerably lower than those on weekdays. To efficiently capture this periodicity, research has predominantly employed two key strategies: periodic positional encoding and time series segmentation. In the first approach, the relative position of a timestep signifies its progress within a designated timeframe. For instance, with a sampling interval of 15 minutes, a day comprises a total of 96 time steps. Consequently, dividing the index of a given timestep by 96 yields the positional value of the instance. This value exhibits periodic patterns similar to the relative index on the subsequent day, or any other days in the entire date range. However, such representation requires multiple distinct position encodings, which may hide the commonalities in the original sequence [7]. To address this issue, existing studies attempted to combine daily and weekly patterns as a hybrid time-series segment to ensure a unified manner of encoding [30].

Similarly, the continuity of time is also concerned with the position of a timestamp within a defined time window, which can be divided into relative continuity and global

continuity. In relative continuity, attention is paid to the relative position of a time step in a short window frame. By contrast, in global continuity, considerations are given to a time step in the context of the whole time span. Reference from the idea of a Transformer [110], studies tend to use the trick of sine and cosine functions to incorporate the relative frequencies into the model, which is a common relative encoding strategy. However, the limitation associated with this approach is that it always assigns different values to the same index that occurs in different local sequences, which may confuse the model's learning process. One way to solve the problem is global positional encoding, where each time step has its distinct index in a global environment. Thus, even if the time step occurs in multiple sequences, the model will not be confused with its location.

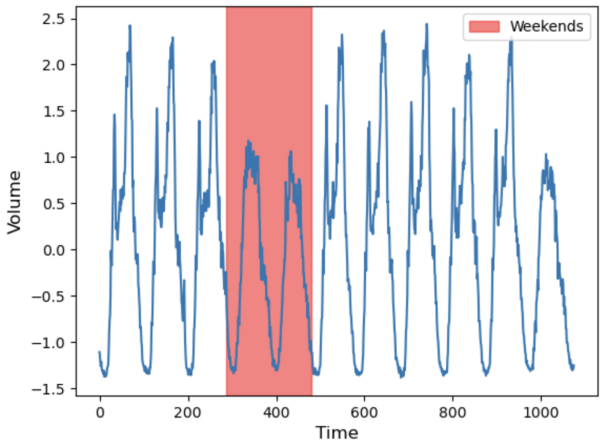


FIGURE 2.1: Traffic signal volume every 15 minutes from 03.01 - 03.14 in Melbourne

On the spatial dimension, traffic sensors in a traffic network can be considered as points in a grid or nodes in a graph. Thus, the modelling of a network is equivalent to finding the connections of the edges and their corresponding distances. Once the basic properties are obtained, the task is to encode the information in a useful way such that the model can learn the correlations between different nodes efficiently.

The most straightforward method is the adjacency matrix representation, where each entry in the matrix signifies a connection between two edges. However, this representation overlooks the crucial distance information, which serves as a vital cue for the model to infer node correlations. To enhance this approach, an additional degree matrix can be introduced. By subtracting the adjacency matrix from its degree matrix, a graph Laplacian can be generated, which has been demonstrated to be more effective and accurate in depicting spatial characteristics within a graph [41].

Nevertheless, for the task of traffic prediction, it is crucial to recognise that spatial correlations among nodes are significantly influenced by time. For instance, the relationship between an upstream node and a downstream node can vary substantially between morning and afternoon due to time-related factors. Therefore, it is imperative to incorporate time-sensitive representations from a dynamic modelling perspective. One widely adopted approach that aligns with this concept is Dynamic Time Warping (DTW), which calculates the similarity between historical and current traffic flows [5]. Despite its successful applications, calculating the DTW for large datasets is computationally expensive, necessitating a trade-off between accuracy and processing time.

2.1.2 Traffic State Estimation

As reflected in section 2.1.1, the primary challenge for understanding traffic conditions lies in efficiently harnessing sparsely sampled data, which may contain noise or outliers. Traffic state estimation typically entails the real-time inference of traffic flow variables, including flows, mean speeds, and densities for roadways, achieving an adequate spatiotemporal resolution by utilizing a limited amount of sensory data [140]. According to the paper from Zhao et al. [140], freeway traffic state estimation consists of four components: traffic measurements, traffic flow modelling, filtering methods and online model parameter estimation.

The first component is the real-time traffic information as reflected in the sensory data. By transitioning to the second component, studies aim to model the data evolution from a spatiotemporal perspective such that the states at a location and timestamp could be estimated. Filtering methods combine observations with the model's state estimation to obtain the "most probable state" [97]. Typical examples include the Kalman Filter based on the state-space model and its variants [102], particle filters that adopt Monte Carlo simulation to represent nonlinear relationships [69] and adaptive smoothing filter that relies on the constant wave speed in each of the free-flowing and congested regimes [106]. Finally, online model parameter estimation refers to the simultaneous estimation of key parameters and the unobserved traffic variables. The approach has demonstrated its potential in TSE but with a limitation in the context [117].

2.2 Traffic Prediction Models

2.2.1 Statistical Models

Early research on traffic prediction tends to focus on model-driven approaches due to insufficient data and the absence of more advanced theory support. Many mathematical models have been proposed and optimised to accurately reproduce the dynamics of traffic flow, which can be divided into hydrodynamic models, kinetic models and microscopic models [11, 58, 90]. The hydrodynamic and kinetic models use the analogy of fluid and gas dynamics respectively to describe the flow of vehicles. Representative models of these categories include the Lighthill-Whitham-Richards (LWR) model and Boltzmann-like models [58, 90]. Both models are easy to implement and can capture the propagation of shock waves in traffic. Microscopic models shift attention from macroscopic observations to individual behaviours. It emphasises vehicle interaction, providing a more detailed way to analyse traffic dynamics. Some common models used in microscopic analysis include the car flowing model [26], lane-changing models [34] and cellular automaton models [72].

However, the intrinsic properties of mathematical models lead to great constraints on their traffic modelling capabilities. Despite the simplicity of the LWR model, it relies heavily on well-defined fundamental diagrams, which can vary based on road conditions. Furthermore, the model assumes that traffic is homogeneous and isotropic, which is not realistic in some situations. Although the kinetic wave describes a partial evolution of the distribution function of vehicles over velocity, it is difficult for the model to derive more realistic interaction rules [13]. Microscopic models generally require high numbers of parameters that must be properly calibrated to match the empirical data [67]. Due to the complexity of traffic systems, such calibrations are hard to obtain, thus potentially worsening the performance of the models.

Traffic prediction can be considered a time-series prediction problem where the future state estimation follows the trend of previous time-series data. Based on this observation, statistical time-series-based models are adopted to solve this task. As early attempts, the Historical Average (HA) and Box-Jenkins Auto-regressive Integrated Moving Average (ARIMA) were applied in traffic prediction [6]. HA works well when the relationship

among timestamps is linear but fails for complex situations. ARIMA can model non-stationary time series through integrations. However, the model is computationally expensive and it can be challenging to determine the optimal parameters. Many variants of ARIMA have been proposed after that to improve the model, among which some typical models are ARIMA with Kalman filter [60], Seasonal ARIMA [120] and VARMA [139]. These models have been proven to be successful for dedicated application scenarios, and have attracted much research attention in the past decade [74, 112].

In summary, as early research attempts, mathematical models provide easy ways to model and estimate traffic dynamics. However, numeric prior assumptions as well as highly accurate parameter calibration are generally required, making the models less applicable to real-world scenarios. Time-series-based statistical models like ARIMA and its variants work well with non-stationary time-series data or instances with seasonal characteristics. Although statistical models are still being extended for different scenarios nowadays [61, 130], their limitations on modelling non-linearity pose great constraints on the model performances.

2.2.2 Machine Learning Models

Machine learning is another promising way for traffic prediction. A Bayesian network approach was proposed in 2005 to model traffic with probabilities [101]. The author explicitly includes information from adjacent road links to analyse the trend of the current connection, which departed from previous forecasting models. In 2011, an adaptive Bayesian network was proposed based on directed acyclic graphical models [88]. It uses mutual information as a learning metric to optimise the network topology for each traffic phase and demonstrate reliable prediction. However, the two approaches share a similar module of the Gaussian Mixture Model (GMM), which can not accurately reflect the distribution of traffic variables in many situations. In 2011, Wang et al. [35] proposed a Least Square Support Vector Machine (LSSVM) regression model to predict passenger flow in Hangzhou but ignored the spatial characteristics of traffic. Similarly, Cong et al. [14] also adopted LSSVM with a Fruit Fly Optimization Algorithm (FOA). The model obtained a slight advantage over a single LSSVM but left spatial properties unaddressed either.

As an essential machine learning algorithm, K-Nearest-Neighbor (KNN) has also been adopted for the task of traffic prediction. In 2016, Xia et al. introduced a map-reduce-based KNN model for traffic flow forecasting [125]. The authors designed offline distributed training and online parallel prediction modules, which alleviate the computation and storage issues associated with the algorithm. However, the model only considers upstream and downstream road segments, which did not make full use of the spatial correlations in the network. Another attempt at the KNN algorithm was carried out by Cai et al. where a spatiotemporal state matrix was incorporated into the vanilla KNN [8]. Nevertheless, the study defines "equivalent distance" as a criterion to determine correlation among road segments, which is unrealistic in many situations.

Despite the feasibility of the machine learning models, each predictor suffers from its own limitations. To utilise the advantages of multiple predictors, many ensemble learning algorithms have been proposed for traffic prediction in recent years [19, 24, 94, 126]. In 2018, a prediction model based on xboost was proposed [19]. The author adopted an ensemble of regression trees on the wavelet-decomposed traffic flow to achieve the forecasting task. In addition, multiple features are selected as input, including both information from the current and the previous lane. In 2019, Xiao et al. [126] proposed an ensemble learning algorithm in a concept drifting environment. In the algorithm, the traditional regression problem was first transformed into a classification problem by shifting the response variable with a small distance and dividing the samples into two groups. Then, a nonlinear kernel and an ensemble framework were adopted to learn the continuous distribution in the time series. In the same year, Rapant [94] proposed an ensemble framework with Kalman filter and differential evolution algorithms. The author used the Cell Transmission Model for the inverse modelling of boundary conditions, which incorporates some physical knowledge into the existing ensemble framework. In 2021, a stacking of learning models was proposed [24], where SVM, CATBOOST and KNN are combined to perform the prediction. The author adopts a simple average detrending algorithm to extract the mean trend of the time series, as well as a meta-regressor to replace the simple weighting mechanism. The model outperforms every single predictor at the task, which validates the effectiveness of ensemble algorithms.

The traditional ML models have achieved improved results compared with statistical or mathematical models on traffic prediction. However, there are still limitations around

these architectures. Firstly, the non-linearity within the time-series data is hard to capture. Although previous research attempted to solve the problem with non-linear kernels or detrending algorithms, the intrinsic constraint of frameworks limits their expressiveness. Secondly, the spatial correlations in the traffic network are under-explored. Most research only considers the upstream and downstream road segments, which ignores the potential long-term dependencies between nodes. Finally, the spatial and temporal features stay largely separated in most ML models. As two main dimensions of traffic data, combining the spatial and temporal features properly is essential for final prediction, which should be paid attention to when building upon existing ML models.

2.2.3 Deep Learning Models

In the past decade, deep learning models have achieved great success in numerous disciplines [89]. To explore the feasibility of neural networks (NNs) in transportation research, efforts have been made to apply single-layered networks to traffic prediction [10, 43]. In 2008, Jin et al. [43] proposed a three-layer neural network that considers the previous and next timestamps as associated tasks for traffic forecasting. Through a shared representation, the author established a multi-task learning structure for the problem. In 2012, Chan et al. employed a hybrid exponential smoothing method to improve the generalisation capabilities of NN training. Although both attempts serve as good starting points for applying NNs to traffic prediction, the structures are too shallow to capture the complex dynamics of traffic systems. In addition, the features were both hand-engineered, which creates unnecessary biases for the models. The first deep-learning-based framework was proposed by Huang et al. [37] in 2014. The paper introduced a deep belief network for unsupervised feature learning and a multitask regression layer for homogeneous task learning. The paper was among the earliest to emphasise the importance of related road segments, which provides directions for future research in both spatial and temporal domains. In the same year, Lv et al. [65] incorporated a Stacked-Auto-Encoder (SAE) approach of feature representation into the deep learning framework for traffic prediction. The author adopted a greedy layer-wise unsupervised learning algorithm in training and gained improved performance compared with previous models. Since then, numerous deep-learning-based architectures have been proposed for traffic prediction, creating diverse directions for future research in transportation research. [103, 135].

A fundamental challenge in traffic prediction is modelling the complex spatiotemporal dependencies in the traffic network. Early solutions use convolutional neural networks (CNN) to model grid-based spatial dependencies and a recurrent neural network (RNN) for temporal pattern learning [132, 137]. Although CNN on traffic networks in grid representations was a good starting point for modelling and extracting the spatial features, its intrinsic principle limits the model's ability to explore non-local node-to-node relationships. In real-world scenarios, correlations among sensors exist in both short- and long-range, which is hard to capture by fixed kernels on a grid. Further research demonstrated that the topological information can better define the propagation of spatial information, which leads to improved model performance [42]. In 2017, Li et al. proposed a diffusion convolutional network (DCRNN) [56]. The authors used a graph convolutional network (GCN) with bidirectional random walks to model spatial dependencies and an encoder-decoder RNN for temporal dependencies. Improved performance on multiple benchmark datasets again proved the superiority of the graph modelling approach.

However, GNN-based models tend to rely heavily on predefined graph structures. In many cases, the relationships can not be represented as ground truth knowledge. In addition, the spatial correlations are highly dynamic concerning time, and fixing the structure is likely to degrade the model performance. To overcome the problems, research efforts have been made in building dynamic graph structures [119, 123, 124]. Graph Wavenet (GWNET) [123] used an adaptive adjacency matrix to learn spatial node embeddings but still requires a predefined graph to perform the best. MTGNN Further improved upon (GWNET) by a graph learning module and a mixed-hop propagation layer. Although both attempts tried to avoid the predefined graph structures, the graph weights are static throughout the training process, which limits the model's performance. To solve the problem, a Decomposition Dynamic Graph Convolution Recurrent Network (DDGCRN) was proposed [119]. The model replaces previous adjacency matrices with a random initial spatial embedding and learns the dynamic graph structure through time. It achieves state-of-the-art results in multiple benchmark datasets, proving the approach's effectiveness. In summary, Graph-based networks can better capture the topological information involved in the traffic network, and less predefined structures are likely to improve the overall performance of the models.

On the temporal dimension, both Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) have shown promising results compared with vanilla RNN [25, 105]. In 2015, Tian et al. adopted an LSTM-RNN that dynamically decides the optimal time lags [105]. In 2016, Fu et al. adopted GRU for traffic prediction and demonstrated the framework's effectiveness [25]. However, there are two major limitations to RNN-based approaches. Firstly, the structure suffers from capturing long-term dependencies. As the time horizon increases, models tend to have large performance drops. In addition, RNN-based models are hard to parallelise due to intrinsic constraints. Since 2017, with the introduction of the Transformers model [110], attention mechanisms have widely been applied to solve the above problems. In 2020, a traffic transformer model was proposed [7]. Four novel positional encoding strategies are introduced and compared to capture the continuity and periodicity of time series. This was the first attempt to apply transformer architecture to traffic prediction. In the same year, a Graph Multi-Attention Network (GMAN) was proposed, which adopted a parallel multi-attention structure for spatial and temporal feature learning. To eliminate the effect of propagation delay in the traffic network, PDFormer adopted a delay-aware transformation module that explicitly models time delay with a spatial dependency matrix [41]. Many other attention-based models have been proposed [31], which greatly pushed the boundary of transportation research.

2.2.4 Meta Learning Models

In the ever-evolving landscape of artificial intelligence and machine learning, the pursuit of models that not only learn but also learn how to learn has emerged as a profound paradigm shift. Meta-learning, often referred to as "learning to learn", represents a revolutionary approach that transcends the boundaries of conventional transfer learning methodologies. Its core objective is to equip machines with the ability to become adept at learning new tasks, rapidly and effectively, with minimal data. A similar and complementary technique is Neural Architecture Search (NAS) [44], where the optimal neural network architecture will be automatically discovered during the training phases. While the NAS strategy primarily focuses on the automated search process, the two techniques can be used together to create highly adaptable and efficient machine learning systems [95].

In recent years, researchers have dedicated their efforts to harnessing the benefits of these technologies for traffic prediction. In 2019, a deep meta-learning-based network was proposed [86] for urban traffic prediction. Unlike the conventional graph learning structure, the author adopts a meta-graph attention network (Meta-GAT), where the network weights are generated from the metaknowledge (embeddings) by the meta-learner. This marked one of the earliest instances of employing meta-learning models for the task of traffic prediction. To reduce the dependency of multi-source data fusion operations on prior expert knowledge, as well as alleviate the semantic gap between different categories of traffic data, Fang et al. [22] proposed a Meta-MSNet that consists of two mete-learning based feature fusion strategy, namely temporal meta-fusion and spatial meta-fusion. Extended upon this model, Auto-MSNet [23] adopts the NAS strategy with adaptive receptive fields to automatically construct the optimal structure for mining spatiotemporal correlations. The two models both attempted to incorporate external data sources, such as weather and time information, into the existing NN frameworks, which provides more insights into traffic prediction with multi-source data. In addition, meta-learning strategies are adopted by both studies, demonstrating the potential of the approach. However, the computing bottleneck of Auto-MSNet lies in the architecture search stage. When the possible parameter space increases, the cost of training can be highly expensive. Therefore, it is crucial to optimise the search space when considering this strategy for traffic prediction.

2.2.5 Trends and Limitations

In the recent decade, the field of traffic prediction has undergone a significant focus shift, transitioning from traditional physics-driven statistical and mathematical models to the realm of deep learning methodologies. This evolution is driven by the increasing recognition of the pivotal role that data plays in enhancing models' predictive accuracy. Indeed, data-driven models have achieved remarkable success on various traffic prediction tasks [74, 112].

Among numerous DL models, spatiotemporal architecture appears to be one of the most successful ways to explore the dynamic spatial and temporal correlations inherent in traffic data. The networks are typically composed of recurrent networks and convolutional

or graph networks. Additionally, the integration of transformer modules, with their attention mechanisms [7, 110], has been a notable breakthrough in traffic prediction and has been widely adopted since its introduction. On the temporal front, the primary objective in the latest decade is to enhance the model's ability to capture periodic and long-term dependencies in traffic patterns. Meanwhile, in the spatial dimension, the focus has shifted towards emphasising the network topological information over grid-based local patterns. Recent studies also suggest that the future direction of traffic prediction models will prioritise adaptability and learning over predefined structures or any other prior information [119], reflecting a dynamic and data-centric approach.

Another school of thought regards "learning to learn" as a critical strategy to be transferred and applied to the task of traffic prediction [86]. By searching through the neural architecture space and automatically constructing the optimal parameter space, the proposed models have demonstrated the potential to push the boundary of DL on the task [23]. However, such an approach requires expensive computation before inference and scales exponentially with the search space, which makes it not applicable to many real-world scenarios.

Compared with early studies, recent success in traffic prediction models lies primarily in the exploration of spatiotemporal correlations, as well as more automatic and dynamic model structures. Moreover, the models' capability to capture long-term dependencies largely improved the prediction performance, which appears to be a promising research avenue. However, compared with the intrinsic complex nature of traffic, the model's abilities are still limited. Most studies only focus on a single traffic variable, ignoring the fact that traffic conditions can be affected simultaneously by multiple factors, such as weather and accidents. Furthermore, data-driven approaches largely ignore the physics of the underlying system. This makes the prediction results less reliable and less interpretable. Considering the black-box nature of DL-based approaches, hybrid approaches combining DL and physics should be a pursuit of future research.

2.3 Physics of Traffic Flow

This section introduces the physics of traffic flow and consists of two interconnected parts, each briefly described as follows:

Subsection 2.3.1 focuses on the modelling of traffic flow characteristics. These characteristics serve as the foundation for developing relationships between traffic variables. In addition, several equilibrium traffic flow models will be introduced. The objective is to solve the temporal-spatial evolution of traffic flow characteristics, considering initial and boundary conditions.

Subsection 2.3.2 introduces the concept of the fundamental diagram, one of the most important concepts to assist with the understanding of the relationship between traffic variables. A review of different forms of the diagram will also be provided.

2.3.1 Traffic Flow Modelling

Traffic flow modelling is of great significance for traffic congestion level estimation and urban traffic planning. To obtain a statistical view of the traffic flow, it is essential to gather enough traffic variables with various sensing techniques. Modern sensing technologies can be classified into mobile sensors, point sensors and space sensors [77], which match the survey types introduced in section 2.1.1.

Mobile sensor data helps to calculate the trajectory, speed and travel time of the vehicles while point sensor data provides count and headways. The definitions of these variables are addressed in detail in the section 3.3. Based on these diagrams, a simple two-dimensional time-space diagram could be formed to describe the traffic flow characteristics [48]. However, such representations do not account for the number of vehicles for a time-space point. In 1971, a three-dimensional representation of traffic flow was introduced [66]. The idea is to interpret the family of space-time trajectory curves as the contours of a three-dimensional surface for which the third dimension is the vehicle number. By drawing tangent curves on the graph, flow and density can be expressed as partial differentials of the surface, which provides useful characteristic modelling for ITS applications [76].

Going beyond the basic 3D representation, traffic flow modelling can be classified into single-regime, multi-regime and other multi-parameter models [48]. One of the earliest single-regime models was Greenshield's model [29], where linear functions are derived for each pair of the speed, flow and density relationships based on observations of vehicles from first-class roadways in the US. Due to its simplicity and elegance, the model has

widely been adopted for traffic flow illustration since its introduction [2, 76]. Inspired by Greenshields's work, numerous models were introduced in subsequent studies to describe speed-density relationships, exhibiting varying degrees of fitting accuracy. By treating the traffic stream as a continuous fluid, Greenberg [28] presented a modified version of the speed-flow-density relationship. The model takes into account the bottleneck situation by handling the dual phenomena separately, which turns out to be a good normalised description of the physical requirements of flow and density. Shortly after, Drake et al. [20] implemented a break-point analysis on the discontinuous regression of speed of flow against density over 1-minute time samples, which agreed well with the visual inspections.

However, a prevalent issue with the single-regime approach is the difficulty in encompassing the full range of densities [78]. To address this challenge, researchers have ventured into a novel approach, adopting the concept of fitting data in a piecewise manner using multiple equations, known as a multi-regime approach. A representative model was proposed by Edie [21], where a mean stream velocity drop-off (discontinuity) was incorporated into the flow-density diagram, reflecting the inherent interference between vehicles. Although the experimental data are limited and the congested flow situation is not explored, the study accurately estimated steady-state-flow-density behaviour. Many similar models have been proposed thereafter, rapidly extending and refining the modelling of the relationships [68].

Despite better fitting qualities of the multi-regime approaches, the piecewise modelling fashion exhibits more observational assumptions. In addition, most of the models involve less than three parameters, which is relatively simple compared with the complex dynamics of traffic. Therefore, efforts are made to incorporate more parameters to better represent the flowing process. Newell [75] introduced a novel car-flowing model with an extra parameter of the slope of the speed-spacing curve. With the parameter's nonlinearity, the model can incorporate the nonlinear phenomena previously obtained from continuum theories, benefiting further investigations on the development of shocks and the spreading of acceleration waves. Similarly, Del et al. [17] used the kinetic wave speed at the jam density and reference flow as extra parameters to be incorporated into the model. The model yields a bilinear fundamental diagram when the shape parameter tends to infinity. However, in contrast to other models discussed in this subsection, which

are derived from corresponding car-following models, Del's model lacks the microscopic counterparts, limiting its applicability to individual behaviours.

To model the nonequilibrium phase transitions and various nonlinear dynamical phenomena, the intelligent driver model [33] introduced safe time headway and acceleration exponent as extensions to previous models. A connection between microscopic and macroscopic models was also established in the work, which helps to simulate freeway sections on both scales. Aside from these, the longitudinal control model [79] is another four-parameter model that uniquely includes the nominal vehicle length, the perception-reaction time and the aggressiveness. The model describes vehicle longitudinal operational control and characterises a traffic flow fundamental diagram, which is suited for transportation applications.

2.3.2 The Fundamental Relationship and Diagram

In traffic flow theory, there exists a fundamental relationship among traffic speed (v), flow (q), and density (ρ), expressed in equation (2.1).

$$q = k \cdot v \tag{2.1}$$

This relationship is visualised through the fundamental diagram, a widely employed concept in transportation studies. Unlike deterministic governing equations that precisely describe the underlying physical laws, the fundamental diagram serves as a tool to enhance our comprehension of driving behaviours, which can vary under different road conditions and individual driving habits. In this review, three frequently used fundamental diagrams will be introduced: the Greenshield's model [29], the Triangular model [15], and the inverse-lambda model [50].

The Greenshield's Model

The Greenshields fundamental diagram stands out as one of the most commonly employed and straightforward models in traffic flow theory. It simplifies the relationship by assuming that the mean velocity exhibits a linear correlation with the density. This relationship among traffic variables is precisely described in Equation 2.2, where ρ_{max} represents the maximum density, also referred to as jam density, and v_{free} signifies the

free-flow speed. The fundamental diagram for this relationship is illustrated in Figure 2.2.

$$
\begin{aligned}
q(\rho) &= v_{free}(1 - \frac{\rho}{\rho_{max}})\rho \\
v(p) &= v_{free}(1 - \frac{\rho}{\rho_{max}})
\end{aligned}
\tag{2.2}
$$

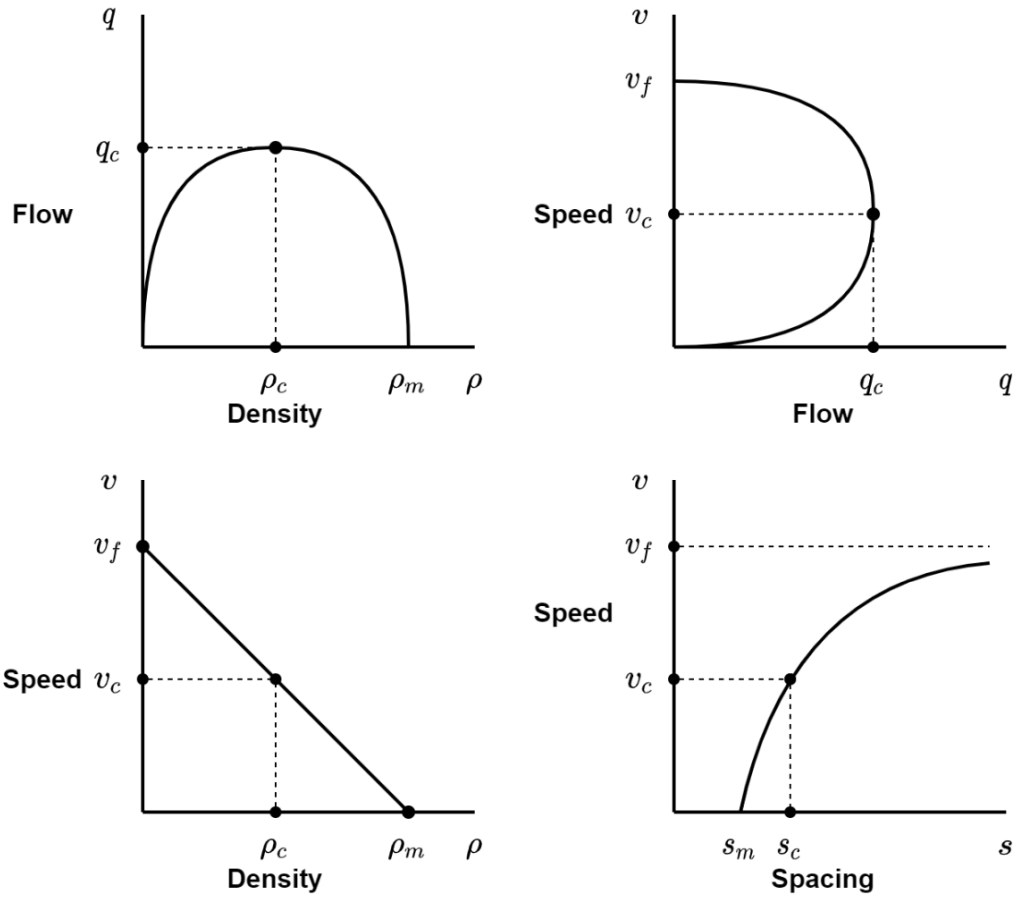


FIGURE 2.2: Greenshield's Fundamental Diagram (from Archie J. Huang et al. [36])

Daganzo's Model

Daganzo's fundamental diagram introduces a *critical capacity*, denoted as q_c, which signifies the point where maximum flow is achieved. In contrast to Greenshields' model, Daganzo's approach employs a triangular shape formed by straight lines to illustrate the relationship between flow and density. This model's analytical solution is represented as a piece-wise function in Equation (2.3), and you can observe the corresponding diagram in Figure 2.3.

$$q(\rho) = \begin{cases} q_c \frac{\rho}{\rho_c} & \text{if } \rho \leq \rho_c \\ q_c(1 - \frac{\rho - \rho_c}{\rho_{max} - \rho_c}) & \text{if } \rho > \rho_c, \end{cases} \tag{2.3}$$

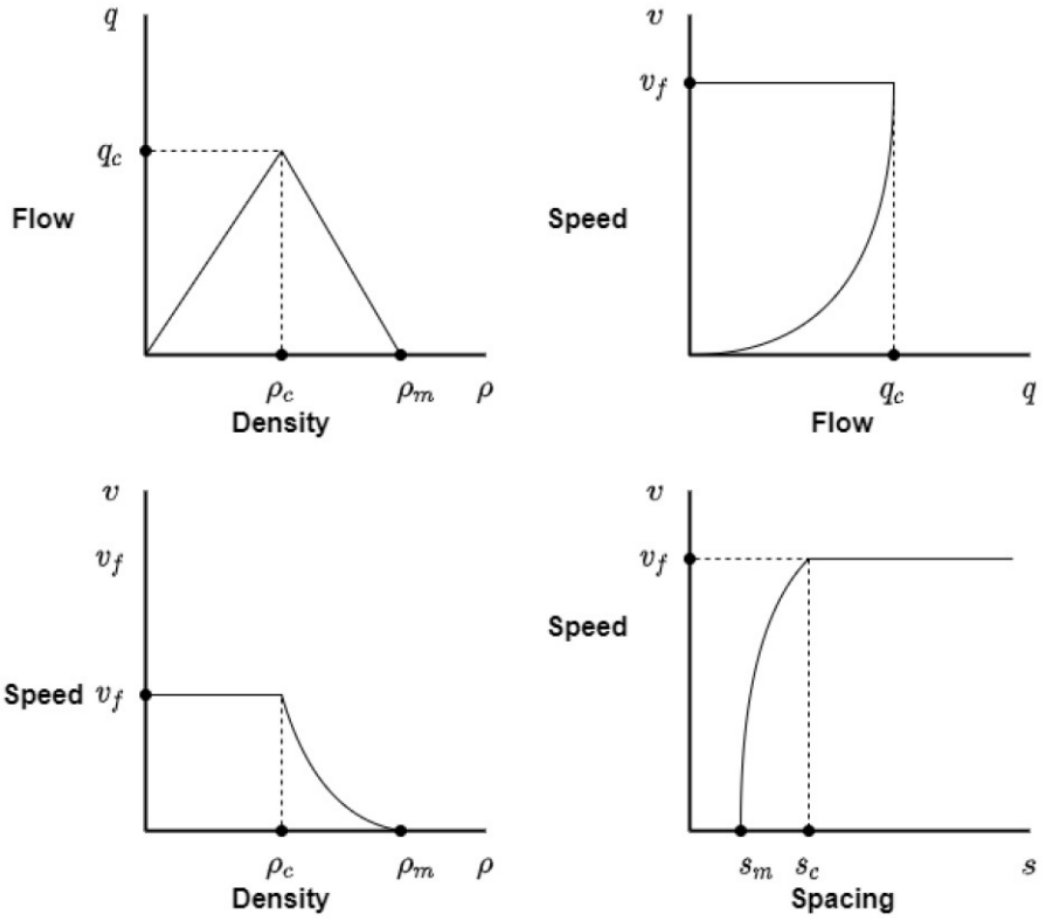


FIGURE 2.3: Daganzo's Fundamental Diagrams (from Archie J. Huang et al. [36])

The Inverse-Lambda Model

Empirical studies have consistently indicated the presence of a capacity drop at freeway-ramp merges [100]. To quantify this phenomenon, the inverse lambda fundamental diagram (FD) has been introduced. The equation for this FD is provided in Equation (2.4), and a visual representation is presented in Figure 2.4. Notably, in this new functional form, the critical capacity is divided into two parameters, q_{c1} and q_{c2}, signifying the conditions before and after the capacity drop.

$$q(\rho) = \begin{cases} q_{c1}\frac{\rho}{\rho_c} & \text{if } \rho \leq \rho_c \\ q_{c2}(1 - \frac{\rho - \rho_c}{\rho_{max} - \rho_c}) & \text{if } \rho > \rho_c, \end{cases} \tag{2.4}$$

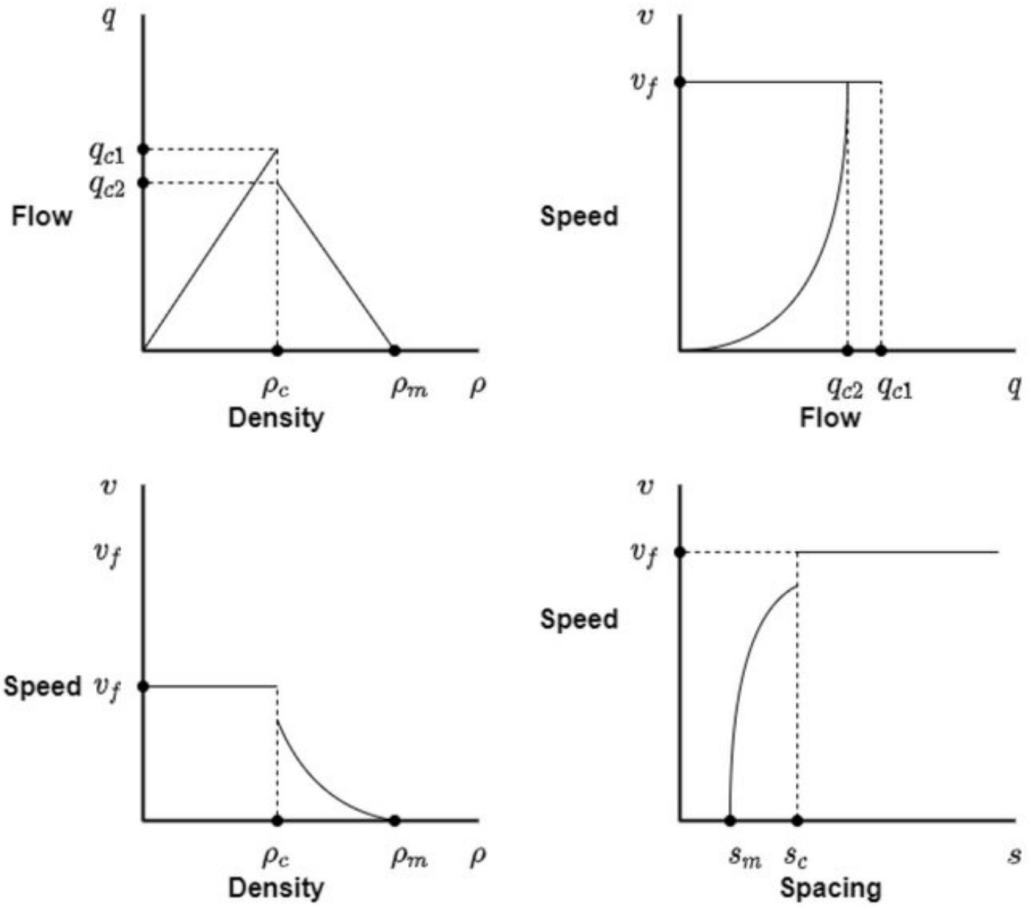


FIGURE 2.4: Inverse Lambda Fundamental Diagrams (from Archie J. Huang et al. [36])

2.3.3 Partition of Traffic Network

While MFDs have been successful in modelling average traffic speed, flow, and density relationships, they often fall short in accurately representing the hierarchical behaviours observed in different regions of the network. This limitation can be attributed to several factors. Firstly, the unpredictability associated with traffic behaviour is a significant challenge. Drivers have diverse driving habits and travel arrangements, making their behaviour highly uncertain and subject to considerable individual-level variations. Secondly, the high complexity involved in physical modelling at a microscopic level poses

additional challenges. With the development of multi-regime approaches, the scales of parameters rapidly increase, leading to significant difficulties in calibration and estimation. Thus, research efforts have been made to the clustering and partitioning of the traffic networks to gain a better understanding of the spatial division of traffic flow characteristics.

During the initial stages of research, the criteria for network partitioning were primarily static and could be broadly categorised into two groups: weighted similarity between links based on traffic variables and spatial compactness indicated by link densities. An example algorithm falling into the first category is the "Snake" similarity clustering method introduced by Saeedmanesh [96]. In this approach, each sequence of road segments referred to as "snakes", is progressively merged based on its similarity to previously added road segments. The similarity is computed as the ratio of the link's density to the average density. Another direction of criterion utilises the spatial compactness and boundary adjustments within clusters for the segmentation of the network [40]. The algorithm aims for small variances of density values within each cluster to align with existing MFDs.

Despite the successful employment of the two partitioning approaches on simulated traffic data, a significant challenge arises as these approaches struggle to capture the dynamic traffic conditions, often characterised by spatiotemporal evolutions. In response to this challenge, Yan et al. [129] proposed a dynamic partitioning scheme that leverages temporal similarity and spatial heterogeneity in link characteristics. This algorithm employs a distinctive iterative "merge and cut" process guided by heuristics, enabling dynamic updates of sub-regions within the network. Similarly, Xing et al. [127] developed a method that uses graph theory and spectral clustering to partition the network by considering both the structural and functional similarities of links. A dynamic updating mechanism was introduced that uses a sliding window technique and an adaptive threshold to update the sub-regions in response to changes in traffic flow. It is important to note that both methods necessitate prior knowledge of the network's structure and function to define the link attributes. However, obtaining such information may be challenging and less accurate in certain cases.

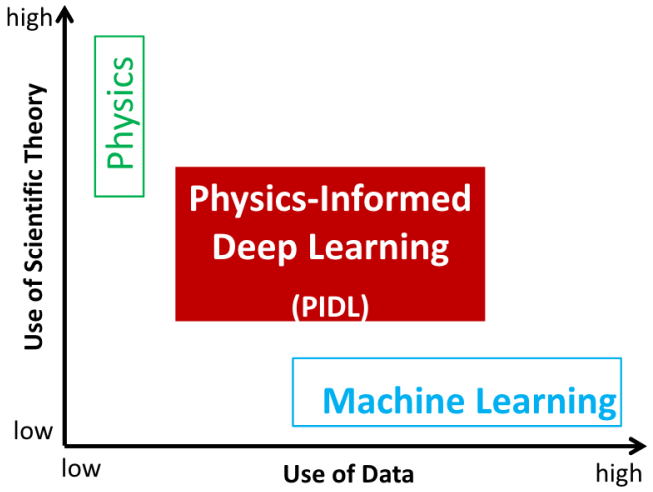


FIGURE 2.5: Balance of Data-driven and Model-driven Approaches (from Di et al [18])

2.4 Physics-informed Neural Network

With the increasing availability of data sources, data-driven approaches have received widespread attention in the past decade. While those neural-network-based approaches have demonstrated remarkable success in modelling and predicting traffic, many of them fail to preserve the underlying physical laws governing the system. To address this critical issue, a paradigm shift is underway in the field of machine learning, emphasising the integration of fundamental physical laws and domain-specific knowledge into the models. Such integration aims to incorporate informative priors that complement the observed data and provide a theoretical basis. This emerging approach is called "Physics-Informed Learning" (PIL), where observational, empirical, physical, or mathematical understandings of the world are leveraged to enhance the performance of the learning algorithms.

At the forefront of this new learning philosophy stands the family of Physics-Informed Neural Networks (PINNs) or Physics-Informed Deep Learning (PIDL), where physical laws are specifically incorporated into deep-learning-based models. A comparison demonstrating the proportion of scientific theory used in data-driven and model-driven paradigms is shown in Figure 2.5.

In this section, we embark on a comprehensive exploration of PINNs, elucidating their fundamental principles, historical evolution and current methodologies. We will also

delve into the challenges within the field and outline potential future directions, thereby illuminating the profound impact of PIL in the landscape of modern deep learning.

2.4.1 The Fundamental Principles

Predictive models are generally constructed with many assumptions, thus sacrificing the generalisation performance of the models. To enhance the generalisation ability of the models, it is crucial to incorporate appropriate biases. In 2021, Karniadakis et al. [46] identified and summarised three main pathways to embed physics into deep learning models: observational bias, inductive bias and learning bias. Observational bias reflects the underlying physics of the system through augmented or transformed data. They are also considered the simplest mode to introduce bias in ML. However, large volumes of data are typically required to reinforce the bias, which is hard to obtain in many real-world scenarios. Inductive bias refers to architectures that implicitly embed prior knowledge in the problem domain. Representative examples are convolutional NNs (CNNs) [51], GNNs [142] and RNNs [1], where the image, graph or sequence patterns are fully utilised to assist with the model prediction process. While the approach has demonstrated remarkable success, it is inherently constrained to tasks where well-defined physical principles can be established. Extending this methodology to more intricate and complex structures presents a great challenge, significantly restricting its applicability across diverse domains. The notion of 'learning bias' is central to a transformative shift in endowing neural networks (NN) with prior knowledge. This paradigmatic change shifts the focus away from specialized architectures, opting instead to impart constraints through soft penalties in the loss function. In this approach, multi-task learning is employed, simultaneously obliging the algorithm to fit observed data and produce predictions that align with predefined physical principles. Although such a soft penalty approach enhances the model's robustness and flexibility, the increased computational demands and the complexity of tuning penalty terms still pose great challenges in the field. In the upcoming section, we will embark on a comprehensive exploration of innovative methodologies specifically tailored to the domain of traffic prediction. Will will illuminate the intricate strategies and cutting-edge techniques that underpin our approach, offering a detailed roadmap into the dynamic landscape of traffic prediction research.

2.4.2 Trends and Methodologies

Observational bias has been demonstrated effective when incorporated into neural network models. In DeepONet [63], the input functions are represented discretely in the input space, based on the universal approximation theorem. The model was employed to simulate various multiphysics systems and was proved to generalise well. An extended PointNet was proposed in 2021 [47] to learn an end-to-end mapping between spatial positions and computational fluid dynamics quantities. The point cloud in the framework simultaneously represents both the geometry of the shape and the space of the flow field, which optimised the computational expenses of the network training. TL-DeepONet [27] extends on the original DeepONet with transfer learning under conditional shift using neural operators. The key idea is to explore the domain-invariant features extracted by the source model, leading to efficient initialisation of target model variables. However, the models all require large-scale data generation, which can lead to expensive experimental costs.

Incorporating an inductive bias into model architectures is essential for capturing the influence of physical constraints, thereby aligning the models with the fundamental laws governing the underlying systems. While Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Graph Neural Networks (GNNs) have gained significant attention, it is important to note that a multitude of specialised frameworks exist beyond these mainstream approaches. One representative category is the kernel method, which analyses data by mapping them to higher dimensional spaces. A common approach that falls into this category is the Gaussian process (GP), which provides predictive prior and posterior distribution towards the observed data. The first attempt at embedding physics into GP was carried out by Raissi et al. [93], where probability distribution was quantified and propagated through time, providing a natural platform for learning from noisy data and computing under uncertainty. However, the cubic scaling concerning the training data points can lead to high computational costs, limiting the model's effectiveness. Other common approaches for solving partial differential equations (PDE) include Bayesian Numerical Homogenisation [82], parametric kernel flows [32], and multi-resolution/grid operator decomposition [83]. The limitation of the Bayesian approach is that the prediction accuracy is sensitive to the initial distribution, thus often requiring optimised prior. In addition, the convergence rate of multigrid-based methods

can be severely affected by the lack of regularity of the coefficients, which remains to be explored. To further explore the physical mechanism of traffic flow dynamics, STDEN [39] was proposed. The model assumes that the traffic flow on the road network is driven by a latent potential energy field (PEF). By incorporating the continuity equation of PEF into a differential equation network (DEN), the model was able to take advantage of both physics-based and data-driven methods.

Compared with modifying the input data or the model structure, learning bias seeks to apply soft penalties as loss constraints such that the convergence of the models is directed towards the underlying physical laws. In recent years, there has been a growing focus on PINNs with such approaches in traffic state estimation and prediction. In 2022, Dahmen et al. introduced a straightforward approach that integrates the continuity equation into a neural network, establishing a mixed loss constraint [16]. Their research showcases superior prediction performance compared to scenarios where no physical loss constraint is imposed. While this method demonstrates the feasibility of physics-informed traffic prediction, it does not delve deeply into the parameters of the fundamental traffic flow diagram. The basic forms of the physical equations cannot fully characterise the complex relationships among traffic variables, which makes the model less applicable to real-world scenarios. In the same year, Huang et al. proposed a deep learning framework that combines the Lighthill-Whitham-Richards (LWR) and Cell Transmission Model (CTM) with neural networks, further affirming the potential of PINNs in traffic prediction [36]. However, it is worth noting that neither of these approaches comprehensively addresses the entirety of a traffic network, ignoring the rich information contained in the road network correlations. Subsequently, Esama et al. introduced a Link-Net framework explicitly tailored for traffic networks [108]. This framework adapts to the unique characteristics of different network links through domain decomposition and effectively integrates individual link information to derive the traffic state across the entire network, which fills the gap of previous research in the spatial domain. However, taking into account each link within the network may result in significant computational expenses for the overall loss, necessitating further optimisation to make it feasible for real-world applications.

Despite the remarkable results achieved by the three physic-informed principles, it is essential to acknowledge that each has distinct advantages and limitations. To explore the synergistic potential of combining these principles, various hybrid approaches have

been proposed, pushing the boundaries of physics-informed ML. The approaches can be roughly categorised into operator learning [57], multi-fidelity strategies [62, 99] and embedding neural networks to numerical methods [131]. For example, by augmenting the existing deep learning framework with an adjoint PDE, DPM [57] seeks to combine both observational and learning bias for out-of-sample realisations of isotropic turbulence. The idea was inherited and extended in a molecular dynamics simulation model [116], where a coarse-graining auto-encoder was proposed and tested. By defining the mapping from an all-atom representation to a reduced representation, augmented data was successfully incorporated as observational bias, reinforcing the inductive bias reflected through the variational auto-encoders.

In conclusion, physics-informed neural networks are effective due to their strong generalisation in small training samples, effective uncertainty quantification and ability to tackle high-dimensionality [46]. Currently, there are no established rules for choosing the correct framework and categories of bias, the ultimate decision depends on the complexity of the problem, as well as the data-physics compatibility.

2.4.3 Challenges

Although PINNs have shown promising results in combining pure data-driven neural network models with the underlying physics governing the system, there are still occasions where the network is hard or even fails to train [115]. Previous research has pointed out that there exists a universal F-principle associated with deep neural networks, where models tend to fit target functions from low to high frequency, leading to poor generalisation performance for high-frequency target functions [128]. Therefore, for physical systems with a rich frequency content, PINNs often suffer from providing accurate results. The study carried out by Cao et al. [9] has provided a comprehensive and rigorous explanation for spectral bias [91] and related it with the neural tangent kernel function. A notable instance of this challenge arises when solving PDEs such as the Poisson-Boltzmann (PB) equation. In such scenarios, the convergence behaviour of DNNs is significantly influenced by factors such as pre-smoothing and other structured mesh components [59]. Another challenge linked to PINNs is the convergence to a global minimum [52]. In physics-informed machine learning models, it's common to train large-scale neural networks with complex loss functions, typically comprising multiple terms

that address different physical properties. The various terms within the loss function may sometimes conflict with each other, leading to an unstable training process.

To address the aforementioned challenges, studies have developed new methodologies to assist with the network training process. XPINNs [45] adopted a generalised space-time domain decomposition approach for solving the PDEs. The model employed a separate neural network with different hyperparameters for each subdomain, which effectively reduced the training cost. Similarly, through a hp-refinement via domain decomposition as h- and p-refinement, hp-VPINN [49] gained improved accuracy compared with normal PINNs and demonstrated better training efficiency. To solve the issues arising from the fact that automatic differentiation does not apply to fractional operators, fPINN [87] introduced numerical differentiation formulas from fractional calculus that encode fractional-order PDEs. By analysing four types of errors related to the convergence of the fractional Poisson problem, the author showed that the model can handle the geometry computational domains in high-dimensional space. However, for some special low-dimensional cases, the NN may still fail to train [46].

Compared to neural networks developed for image and text data, Physics-Informed Neural Networks (PINNs) are still in their early stages of development. Experimental data used in PINNs often comprises synthetic data, and there is a limited availability of benchmark results for comparative analysis. Furthermore, the foundational mathematical underpinnings of this field have not been explored to the same extent. Rigorous numerical analysis will yield more robust and efficient training algorithms, which deserve more research effort.

2.5 Summary

In recent decades, the field of traffic prediction has undergone a significant transformation, shifting from traditional model-driven approaches to data-driven methodologies. This shift has been driven by the growing availability of diverse traffic data sources. Notably, deep learning neural networks have emerged as a highly successful framework for this task, particularly through the adoption of spatiotemporal architectures. These architectures are instrumental in capturing dynamic temporal and spatial correlations within traffic data.

On the temporal dimension, the evolution of time-series modelling has progressed through various stages, including the utilisation of Recurrent Neural Networks (RNNs), Gated Recurrent Units (GRUs), and Long Short-Term Memory networks (LSTMs) [56]. Concurrently, on the spatial dimension, research has demonstrated the effectiveness of incorporating the topological information of the traffic network. This has prompted a methodological transition from Convolutional Neural Networks (CNNs) to Graph Neural Networks (GNNs) [42, 123].

Furthermore, researchers have recognised a significant limitation in the use of pre-defined graph structures. These static structures are unable to effectively capture the evolving dynamic spatial correlations over time. To address this issue, two notable solutions have emerged. The first solution involves the introduction of adaptive adjacency matrices, which represent an early attempt to enhance model adaptability. A further enhancement is the development of dynamic spatial embeddings, which have been empirically validated through experiments [119]. These embeddings introduce self-learning into the model, reducing the reliance on prior knowledge and enabling a more accurate representation of evolving spatial correlations. Furthermore, the introduction of transformer modules [110] has given rise to numerous self-attention-based models. These models have been seamlessly integrated into existing spatiotemporal frameworks, contributing to improved prediction accuracy [7, 31, 141].

Despite the remarkable successes achieved with deep learning approaches, it is important to acknowledge the remaining challenges and shortcomings within these models. From a data perspective, a prevalent issue is that existing studies tend to rely on outdated benchmark datasets, typically collected from 5 to 10 years ago. These approaches largely overlook the ever-changing nature of data resulting from new urban roadway conditions and advancements in sensing technologies. It is essential to prioritise the collection and analysis of new real-world datasets to gain a more comprehensive understanding of current traffic trends. On the model dimension, the increasing complexity of neural network modules has led to a surge in the number of model parameters, often reaching the scale of hundreds of thousands or more. This poses a substantial computational burden, thereby impeding the feasibility of large-scale real-time traffic analysis.

Finally, during the learning phase, it is worth noting that the majority of spatiotemporal models rely exclusively on error metrics such as Mean Squared Error (MSE) or Mean

Absolute Error (MAE). While these metrics are effective for dedicated datasets, they often exhibit limitations in terms of generalization performance. Given the inherent variability in traffic systems across diverse road conditions, it becomes crucial to delve deeper into the exploration of common patterns or underlying laws governing these systems.

The exploration of the physics governing traffic flow constitutes a foundational undertaking for Traffic State Estimation (TSE) and traffic prediction. A pivotal aspect of this exploration lies in traffic flow modelling, which involves utilising equilibrium traffic flow models to elucidate the temporal-spatial evolution of a traffic system. Over the past few decades, various forms of models have been proposed, encompassing single-regime models, multi-regime models, and other models with multiple parameters. While the simplicity of early single-regime approaches had its merits, it became evident that a linear function alone could not capture the intricate evolution of traffic variables comprehensively.

To address this limitation, multi-regime approaches were introduced, employing piecewise linear functions to model distinct segments separately. While this added a degree of complexity, it still exhibited constraints concerning the number of parameters employed. Consequently, more intricate models, often featuring three or more parameters, were developed to surpass existing paradigms. These advanced approaches, along with their corresponding fundamental diagrams, provide valuable guidance for contemporary traffic flow modelling and prove indispensable for a multitude of tasks within the realm of transportation studies. As pointed out by a few studies [48, 113], future advancement of relationship modelling may lie in the stochastic distribution level instead of deterministic approaches.

A promising paradigm that bridges deep learning approaches and the physics of traffic flow is commonly referred to as Physics-Informed Neural Networks (PINNs). Previous research has identified three primary pathways for incorporating physics into neural networks, namely observational bias, inductive bias, and learning bias. Studies focusing on observational bias often involve mapping the underlying physics governing the system into dedicated input spaces through neural network encoding strategies. This approach augments or transforms input features to enable the model to better simulate

the underlying system. However, it is worth noting that such initialisation can lead to computationally expensive processes, limiting its application to large-scale datasets.

Inductive bias, on the other hand, places emphasis on the inner structure of the model. Beyond conventional models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), or Graph Neural Networks (GNNs), whose structures are tailored to specific input patterns (images, sequences, and graphs), various parametric models have demonstrated success in simulating physics. These models include those based on Bayesian methods, kernel flows, and Partial Differential Equations (PDEs). While effective, such approaches can sometimes face challenges in identifying an optimal structure for the system, making it difficult to directly fit architectures to many problem domains.

As a soft approach, learning bias serves as a valuable regularisation factor for neural network models, guiding them towards convergence with the underlying governing equations. Research efforts have been dedicated to incorporating various traffic flow models, often in the form of PDEs, into the loss functions of models. This practice guides the training process and has yielded promising results. However, in traffic prediction, this approach can encounter limitations in capturing information from both spatial and temporal domains. Exploring these domains further is essential to fully harness the characteristics inherent in the traffic system.

Chapter 3

Preliminaries

In this chapter, we provide essential background information and concepts necessary to understand the subsequent chapters of our thesis. This chapter comprises four key sections, each of which plays a vital role in comprehending the intricacies of our research:

Section 3.1 introduces a Hybrid Computational Graph (HCG), which serves as a visual representation of our extended architectures and computational flows. We will introduce common symbols employed in the graph and delineate its structure. Additionally, we will elaborate on the advantages of employing HCGs in the context of traffic prediction.

Section 3.2 introduces and mathematically formulates the governing equations. We will also explain how differential strategies can be incorporated into neural networks to solve these equations.

Section 3.3 provides a formal definition of the traffic variables that will be utilised in this research.

Section 3.4 mathematically formulates the problem of traffic prediction within this thesis. We will also delve into definitions of the road graph, adjacency matrix and traffic feature matrix.

3.1 Hybrid Computational Graph

In deep learning, a hybrid computational graph (HCG) serves as a systematic approach for illustrating both the functional components and the necessary mathematical computations within a model [18]. This visual representation acts as a powerful tool for comprehending neural network architectures. In the context of traffic prediction, where the dynamics of spatiotemporal relationships are paramount, computational graphs are invaluable for understanding the connections between different components within our model. In this thesis, we will utilise the computational graph to illustrate both the correlation-enhanced and physics-informed architectures. The graph is considered "hybrid" as it may encompass multiple categories of models in a combined structure. For example, PINNs often consist of both the physics-informed components and their physics-uninformed counterparts, complementing each other in deep learning tasks [45, 87].

Mathematically, it can be described as a labelled directed graph where the nodes represent both observable and unobservable physical quantities, intermediate values, and target objectives. The directed edges connecting these physical quantities signify the dependency of a target variable on source variables, encompassing a mathematical mapping of source-target quantity. A path from a source to the observable output quantities represents a specific configuration of a model. Establishing such a path within the HCG constitutes a model configuration. In this thesis, we use directed arrows to indicate data flows, round-cornered rectangles for mathematical or ML entities and diamonds for yes or no conditions. An example HCG of physics-informed architecture is illustrated in Figure 3.1.

3.2 Governing Equations and Automatic Differentiation

In the context of traffic flow theory, the governing equations represent the mathematical descriptions of how a system behaves, specifically outlining the physical constraints that traffic flow dynamics must adhere to. To mathematically formulate this process in a multi-dimensional spatiotemporal system, we analyse the governing equations expressed as a set of parameterised partial differential equations (PDEs), as shown in equation 3.1 (adapted from definitions proposed by Chen et al. [12]).

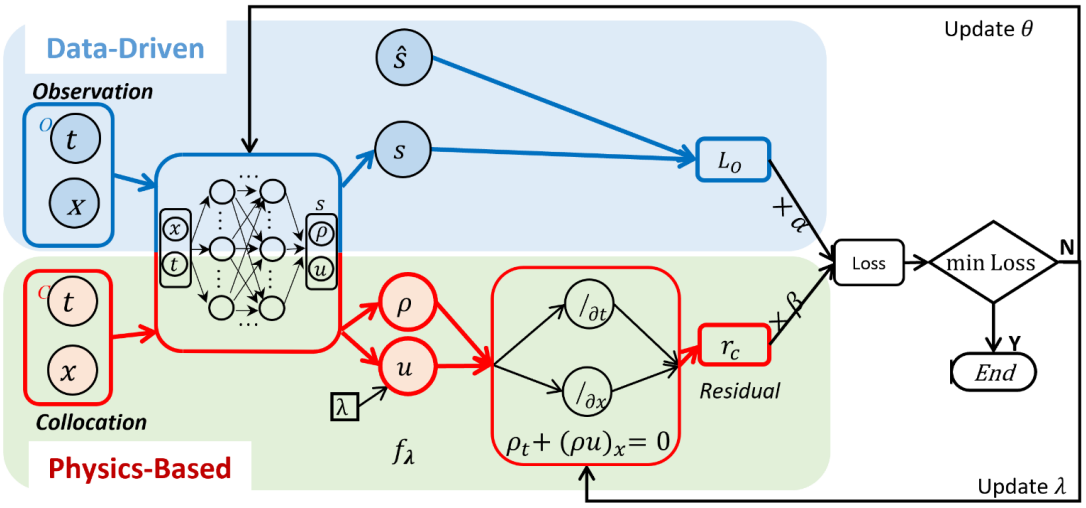


FIGURE 3.1: Example Hybrid Computational Graph for Physics-informed Architecture (from Di et al. [18])

$$u_t + f[u, u^2, ..., \nabla_x u, \nabla_x^2 u; \lambda] = p \tag{3.1}$$

Here, u_t represents the first-order time derivative of the input signal, with t ranging from 0 to T in the temporal space, and x spanning the spatial dimension within Ω. The symbol ∇ denotes the gradient operator, and $p = p(x, t)$ stands for the source input to the system. The primary goal is to derive a closed-form solution for the function f using spatiotemporal measurement data. It is important to note that PDEs are subject to initial and boundary conditions, typically represented as $I[x \in \Omega, t = 0]$ and $B[x \in \partial\Omega]$ respectively.

To solve these equations effectively, automatic differentiation stands out as a valuable strategy. It operates by calculating intermediate numerical values through element-wise partial differentiation. In the context of neural networks, solving the partial differential equation (PDE) is equivalent to minimising the loss represented by the physical constraints. In this thesis, we intend to employ this strategy for the integration of learning bias into neural network models.

3.3 Notations and Definitions

In this section, traffic flow variables introduced in section 2.3.1 will be formally defined:

1. **Density**. Density ρ reflects the number of vehicles per kilometre of road. For a measurement interval at a certain point in time, such as S_1, ρ can be calculated over a road section with ΔX length as:

$$\rho(x_1, t_1, S_1) = \frac{n}{\Delta X}, \tag{3.2}$$

where the index n indicates the number of vehicles at t_1 on the location interval ΔX. For the location S_1, we take the centre of the measurement interval.

2. **Flow Rate** (Flow). The flow rate q represents the number of vehicles that pass a certain cross-section per time unit. For the time interval ΔT at any location x_2, such as the measurement interval S_2, the flow rate is calculated as follows:

$$q(x_2, t_2, S_2) = \frac{m}{\Delta T}, \tag{3.3}$$

where m represents the number of vehicles that passes location x_2 during ΔT at time t_2. This concept is usually exchangeable with *traffic volume*.

3. **Mean Speed** (Speed). Mean speed u is defined as the quotient of the flow rate and the density. The relationship among the density, flow and speed is also called the ***fundamental relationship*** of traffic flow theory:

$$q = k \cdot u. \tag{3.4}$$

4. **Trajectory**. A vehicle's trajectory x_i is defined as the location of the vehicle as a function of time: $x_i = x_i(t)$.

3.4 Problem Formulation

Definition 1 (Road Network Graph). The road network can be represented as a graph $G = (V, E, A)$, where $V = \{v_1, v_2, ..., v_N\}(||V|| = N)$ denotes a set of N nodes, $E_{i,j} = (v_i, v_j), 1 <= i \neq j <= N, E \subseteq V \times V$ is the set of edges in the graph, and $A \in \mathbb{R}^{N \times N}$ represents the adjacency matrix of graph G_r.

Definition 2 (Adjacency Matrix). The adjacency matrix $A \in \mathbb{R}^{N \times N}$ of a graph $G = (V, E)$ denotes the connections of nodes ($||V|| = N$). $A_{i,j} = \epsilon > 0$ if $(v_i, v_j) \in E$ and $A_{i,j} = 0$ otherwise.

Definition 3 (Traffic Feature Matrix). The traffic feature matrix at time t is denoted as $X_t \in \mathbb{R}^{N \times D}$, where N is the number of nodes in the network and D is the dimension of traffic features. For example, if both traffic speed and flow are variables of interest, then $D = 2$. Based on the above definition, we denote the traffic feature matrix spanning across T units of time as $X = [X_1, X_2, ..., X_T] \in \mathbb{R}^{T \times N \times D}$.

Given historical n_1 time steps of traffic condition along with the road network information, traffic prediction aims to find a functional mapping f that maps the given information to the future traffic condition in n_2 time steps. With Section, the problem can be formulated as follows, with a time span of n before and after time t :

$$f([X_{t-n_1}, ..., X_{t-1}, X_t], G_r) = [X_t, X_{t+1}...X_{t+n_2}]$$

To ensure generality, most literature denotes X as a matrix rather than a vector. However, the variable of concern is mainly monotonic, typically speed or flow only. Thus, $X_t \in \mathbb{R}^{m \times n}$ can be indeed regarded as $X_t \in \mathbb{R}^{m \times 1}$, which is a vector of speed/flow at different nodes in the road network. In our study, we aim to incorporate the fundamental relationship between traffic variables so that they may reinforce each other during model prediction. Thus, the above formulation could also be written as:

$$f([X_{t-n1}, ..., X_{t-1}, X_t], [Z_{t-n1}, ..., Z_{t-1}, Z_t], G) = ([X_t, X_{t+1}...X_{t+n1}], [Z_t, Z_{t+1}...Z_{t+n1}])$$

where both $X_t \in \mathbb{R}^{m \times 1}$ and $Z_t \in \mathbb{R}^{m \times 1}$ are vectors, each representing a single traffic feature matrix. The two variables could be any two of density, flow and speed. This form is less general than the first formulation. However, it emphasises the difference between our study and the existing approaches.

Chapter 4

Methodology

In this chapter, we will present a detailed description of the strategies and techniques employed in the development of our correlation-enhanced and physics-informed models. The chapter is divided into five sections, each offering a unique perspective on the methodologies that were employed:

Section 4.1 delineates the procedures employed for data collection and preprocessing. We will introduce the preprocessing steps with a dual focus on both temporal and spatial dimensions. Furthermore, we will underscore the distinctions between our methods and those used in previous studies when processing links.

In section 4.2, our primary objective is to elucidate the rationale behind selecting the two specific types of biases. Subsequently, we will expound upon the intuition of harnessing the correlations among traffic variables by means of these biases.

In section 4.3, the expansion of model dimensions will be discussed, highlighting how increased dimensionality was harnessed to capture intricate patterns in traffic data.

Section 4.4 will explore the incorporation of physical attention mechanisms as inductive biases within the model, demonstrating their role in enhancing its predictive capabilities.

Finally, in section 4.5, the incorporation of physics-informed loss constraints will be delved into, showcasing how learning biases were used to guide the model's training and improve its results.

4.1 Data Collection and Processing

In this thesis, two datasets are collected for experiments. The first dataset is sourced from the PeMS (Caltrans Performance Measurement Systems) data warehouse. The PeMS series, including PeMSD3 to PeMSD8, have been widely adopted as benchmark datasets in traffic prediction studies [41, 56, 119, 134, 141]. Ranging from the 1st of May to the 30th of June in 2022, 200 stations from District 7 were selected in this research. The selection of the nodes is based on the connections and directions of the links, calculated from the station metadata. Regions with dense link connections are prioritised. The range of District 7 is shown in Figure 4.1. The corresponding distribution of the stations and the clusters representing the number of stations are shown in Figure 4.2. As reflected in the statistics, there are 9 freeways (each with two directions) included in this dataset.

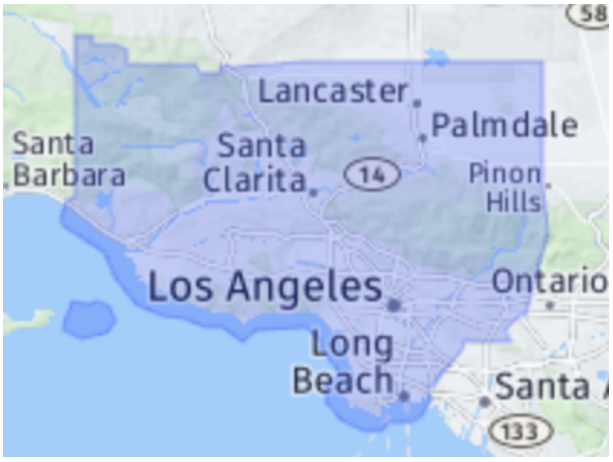


FIGURE 4.1: Range of PeMS District 7

We conduct separate data processing for temporal and spatial data. On the time series side, any timestamps with missing entries in the speed or flow field are excluded. Additionally, common outliers identified through both the Interquartile Range Rule and the z-score method are eliminated. This guarantees that no outliers negatively impact the model's prediction performance. Finally, a frequency transformation is applied to ensure that the time intervals between previous and current timestamps are standardized to five minutes.

In the spatial dimension, to cater to the various input formats required by the models, multiple adjacency matrices are designed. Initially, a distance matrix is computed using

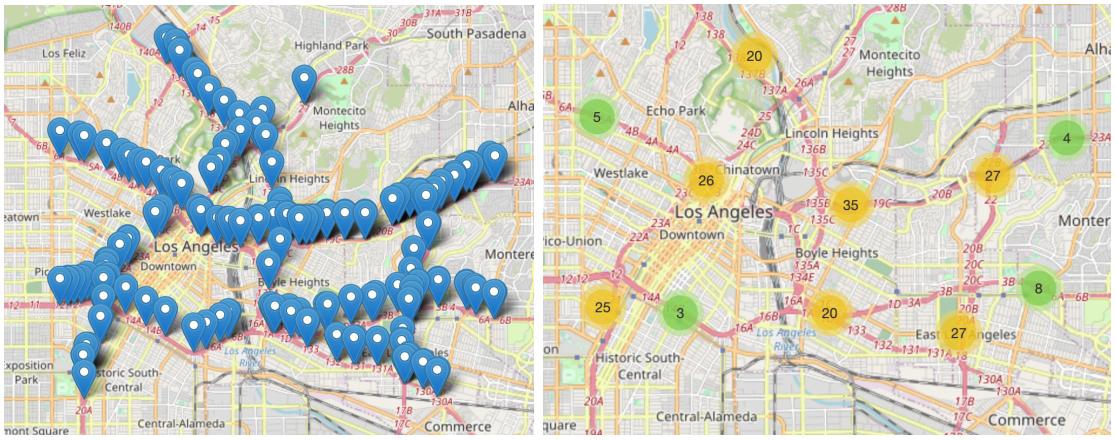


FIGURE 4.2: PeMSD7 Station Distribution (left) and Number Counting Clusters (right)

the Euclidean distances between every pair of stations on the map. Based on this matrix, the following encodings are considered:

1. Threshold Gaussian Kernel. The weight of an edge connecting the vertices i and j is:

$$W_{i,j} = \begin{cases} exp(-\frac{[dist(i,j)]}{2\theta^2}) & \text{if } dist(i,j) \leq k, \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

   where θ and k are parameters, and $Dist(i,j)$ could also be a physical distance between two feature vectors describing i and j.

2. In the context of the Graph Laplacian, an unnormalised version can be formulated as follows:

$$L := D - W, \tag{4.2}$$

   where D represents the degree matrix with each diagonal element being the sum of the weights of edges incident to that element. For a normalised Laplacian, the equation is

$$L := I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \tag{4.3}$$

   This formula incorporates the degree matrix D in a normalised form, where $D^{-\frac{1}{2}}$ represents the square root of the inverse of the degree matrix. To obtain eigenvalues and eigenvectors, eigenvalue decomposition can be performed on this Laplacian matrix. The G smallest eigenvectors are selected to represent the spatial features.

The Melbourne dataset is sourced from two platforms. The traffic volume dataset is downloaded from the DATA VIC [107] and covers 31 days ranging from the 1st of July to the 30th of August in 2023. The traffic speed dataset is collected through the Bluetooth Travel Time API provided by the VicRoads Data Exchange Platform [111]. Due to some server-side breakdown, the data stream was disrupted in the middle. Thus, only four weeks of data have been collected. To align the timestamp frequencies with the volume dataset, every 30 half-minute readings is aggregated into one 15-minute reading by calculating the averages. Given that traffic flows are directional, modelling volumes at an aggregated site level would be inaccurate. As a result, we established correspondences between each site and its respective detectors to obtain traffic flow data at the link level, thereby preserving directional features. When calculating distances between coordinates, without loss of generality, we used the destination site as the geographical location of the link. In this context, two links are deemed adjacent if and only if the origin of one link corresponds to the destination of the other link. Another issue related to the dataset is the link-site-detector correspondence. For each site, there are multiple detectors involved. We aggregated the flow across all detectors for each site so that the final value could reflect the overall traffic volume in the period. The destination site distribution and the clusters are shown in Figure 4.3. Note that two parallel links with opposite directions are treated as unrelated, resulting in a distance value of infinity.

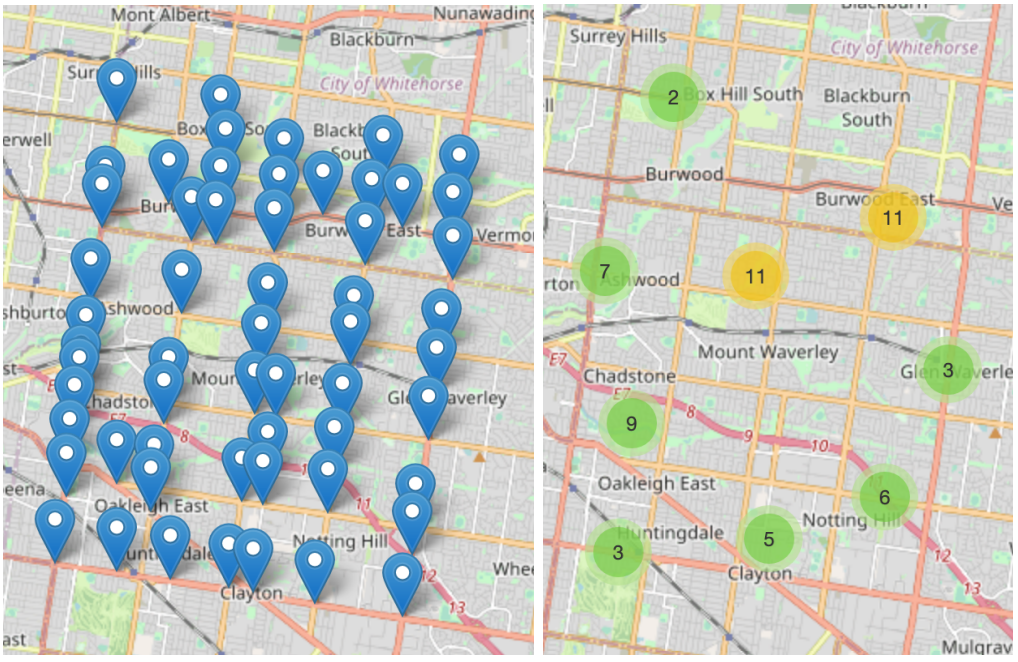


FIGURE 4.3: Melbourne Site Distribution and Number Counting Clusters

Backward linear interpolation is employed to both speed and value datasets to fill in missing values. Frequency transformations are performed to ensure that time differences are maintained to 15 minutes each. To support the spatial file input format for GMAN, the same node2vec algorithm [141] are performed on both Melbourne and PeMS dataset, outputting two spatial embedding matrices. The descriptions of the PeMSD7 and Melbourne datasets are presented in Table 4.1. The attributes included from left to right are the name of the datasets, the number of nodes in the traffic network, the time range, and the time intervals. To avoid confusion with previous studies that used PeMSD7 data from different years, the year attributes are appended after the names of the datasets.

TABLE 4.1: Dataset Description

Name	Node Count	Data Range	Time Interval
PeMSD7_2022	200	1st of May to 30th of June	5 minutes
Melb_2023	193	1st of July to 31st of August	15 minutes

4.2 Modelling Prediction Bias

4.2.1 Rationale for Bias Selection

As mentioned in literature [46], there are three main pathways for embedding physics into an ML model: observational, inductive and learning biases. Each of these approaches possesses distinct strengths and weaknesses to be adopted in traffic prediction tasks.

Observational biases require incorporating underlying physics into the data and learning the physical constraints through data structure or relationships. This is often considered the simplest method for introducing bias. However, given the dynamic nature of the relationships between traffic data entries in both space and time, transforming or augmenting the data to fit a predefined physical model carries a risk of obscuring hidden patterns, potentially leading to worse prediction results. As reflected by the latest traffic prediction models, less prior information and more learning of data structures can generally lead to better performance compared with those with predefined embeddings [119, 124]. Thus, observational bias will not be considered in this research.

Inductive biases impose strict constraints on model architectures. While a custom structure can be effective for specific application scenarios, it may also limit the model's ability

to capture the full variability in traffic data. Therefore, it is crucial to foster the model's self-exploration of patterns that exist in the traffic variables. In my research, the attention modules designed will follow this guideline to leave more space for the learning process.

Compared with the two forms of hard constraints, learning biases try to apply soft penalties during the training phase by including physical equations as a regularisation term in the loss function. Although the governing laws are not enforced under this constraint, the prediction result will still be close enough to the physics with a larger weight parameter. In addition, there is a tradeoff between the computational cost and accuracy. Calculating the physics-informed loss can lead to an increased complexity of the model, but it also gives better results and interoperability. In this study, different portions of physics-informed loss will be experimented with to demonstrate the feasibility and effectiveness of the approach. In the following sections, the proposed architecture and planned methodologies will be presented and described.

4.2.2 Inductive Bias for GMAN and DDGCRN

Besides addressing the observational bias, inductive bias stands out as a valuable approach for incorporating the established physics of a system. Nevertheless, this approach remains relatively underexplored in traffic prediction studies, primarily due to limited support for multiple traffic variables as both inputs and outputs. This limitation can be attributed to the scarcity of comprehensive datasets that encompass more than one traffic variable. Another factor contributing to the limited exploration of inductive bias is that most benchmark datasets primarily rely on accuracy metrics, which assess the disparities between ground truth and predicted values, as the sole criterion for evaluating model performance. Consequently, only a small fraction of research endeavours to align the model's predictions with the fundamental physics governing the system.

The physics of traffic becomes evident through the correlations between various traffic variables. For instance, Greenshield's fundamental diagram [29] establishes a simple relationship between traffic speed, flow, and density, providing a visual representation of the underlying physical principles. Similarly, LWR employs partial differential equations to mathematically describe these relationships, effectively formalizing the physics of traffic. Motivated by these models, infusing inductive bias into existing deep-learning

frameworks involves leveraging these traffic variable correlations. The central challenge lies in crafting a structure capable of recognizing and learning the interactions among physical features, such as speed and flow.

A promising starting point involves integrating attention modules into both the temporal and spatial domains. To investigate the dynamic relationships between different timestamps and locations, a self-attention module can be used to compute pairwise similarity scores. The entries with the highest score significantly influence the final weighted value for the current item. Expanding on this concept, a physical attention module can operate in parallel with temporal and spatial attention, facilitating the modelling of similarities among traffic variables.

Different from GMAN, DDGCRN adopts a recurrent model structure referenced from Gated Recurrent Units (GRUs) [119]. In addition, to generate a dynamic graph matrix, the spatial embeddings are concatenated with the input and fed into each hidden state together. Thus, the final graph structure is learned through each forward evolution in a GRU and backpropagated through time. Based on this idea, our study proposed a novel physical embedding to learn the dynamic correlations between traffic speed and flow. The implementation details of the extended structures for GMAN and DDGCRN are addressed in Section 4.4.

4.2.3 Learning Bias for Correlation Learning

Differing from the inductive bias, the learning bias utilises physical equations to exploit correlations among traffic variables. As elucidated in subsection 2.3.2, traffic flow models delineate traffic flow characteristics by tracing the space-time evolution of speed-flow-density relationships. This implies that physics models inherently encompass correlations between traffic variables by nature. Consequently, by imposing regularisation to align model predictions with the physics models, we effectively imbue the model with predefined variable correlations, steering its convergence toward prior knowledge about the underlying system. We address the detailed process of incorporating learning bias in section 4.5.

4.3 Model Dimension Expansion

Existing traffic prediction models typically involve one input and one output. To capture correlations between different traffic variables, expanding these models to accommodate multiple inputs and outputs is essential. However, this expansion presents an immediate challenge related to dataset representation.

In previous studies, the dataset shape was often 2-dimensional, denoted as $X \in \mathbb{R}^{N \times D}$, where N stands for the number of timestamps within the covered date range, and D represents the number of nodes in the traffic network. To support this new dimension, we must extend the dataset to a 3-dimensional structure, represented as $X \in \mathbb{R}^{N \times D \times F}$. Here, the added dimension F signifies the number of features of interest. In the context of traffic prediction, these features can encompass various aspects such as traffic flow, speed, or other macroscopic variables describing the traffic conditions on a specific road segment. This expansion can also be applied to other multivariate time-series datasets, where the first two dimensions define the spatial and temporal domains, while the last dimension deals with the feature space associated with a particular timestamp and location.

Following the dataset definition, the next step is to encode the training samples into a neural network. Previous research commonly represents such datasets by abstracting and encoding them as 3-channel images, which include width, height, and pixel dimensions. This 3-channel representation aligns with the conventions used in image classification tasks, making it easily compatible with convolutional neural networks. To incorporate the number-of-images parameter, an additional dimension is added in front, resulting in a four-dimensional shape. In the context of traffic prediction, an extra parameter, denoted as T, is introduced during the training phase to indicate the number of time steps to predict in each batch, which is similar to the concept of pixel dimensions in images. As a result, the shape of the training dataset becomes $X \in \mathbb{R}^{N \times T \times D \times F}$. However, for the feature space, this representation may not suffice. To conceptually model this extra dimension, we can expand a 2D image into a 3D representation by introducing a density field. Thus, the final shape of the dataset evolves to $X \in \mathbb{R}^{N \times T \times C \times D \times F}$. This approach has a broad range of applications beyond traffic prediction, including fields such as video frames, computerized tomography, and magnetic resonance imaging (MRI) [70].

In this research, two distinct expansion strategies have been employed for GMAN and DDGCRN to adapt to their respective architectural designs. GMAN operates as a self-attention-based model, wherein it learns correlations using similarity matrices computed across both temporal and spatial dimensions. These matrices require an additional dimension to store the pairwise affinity scores, resulting in data shapes of $X \in \mathbb{R}^{N\times C\times D\times F\times T\times T}$ for temporal dependencies or $X \in \mathbb{R}^{N\times T\times C\times F\times D\times D}$ for spatial dependencies. As mentioned earlier, the shape of the expanded matrix can be conceptualised as a 3D image. Consequently, instead of utilising a fully connected layer that operates on a single dimension or traditional 2D convolutions, 3D convolution is employed. These 3D convolutions not only transform the data but also incorporate nonlinearity, enhancing the model's capability to capture intricate relationships within the data. On the other hand, DDRCGN uses a recurrent architecture that learns information in each time step sequentially. Inspired by stacked autoencoder structures [81], extra layers could be added on top of the bottom layers to learn deeper feature representations, which can be easily incorporated into the existing implementation of DDGCRN.

4.4 Inductive Bias

4.4.1 Physical Embedding

Embedding refers to the encoding of feature spaces into a meaningful input format suitable for neural networks. Numerous types of embeddings have been introduced to accommodate different deep-learning architectures. For instance, a spatial embedding can transform vertices into vectors that preserve the graph structure information [141]. In this thesis, we introduce a physical embedding that encodes the relationship of traffic variables into a vector space. To ensure that the correlation is learned rather than predefined, we initiate a random physical embedding matrix during the data loading phase, denoted as $P \in \mathbb{R}^{F\times F}$. Here, F represents the number of traffic features of interest. This embedding is subsequently concatenated with the spatiotemporal embedding, denoted as $STE \in \mathbb{R}^{B\times T\times D}$, which was originally designed in GMAN [141].

To clarify, the value of the embedding e for variable v_i at time step t_j and location n_k is defined as follows: $e_{v_i,t_j,d_k} = e^F_{v_i} + e^T_{t_j} + e^D_{d_k}$. Here, F represents the number of traffic

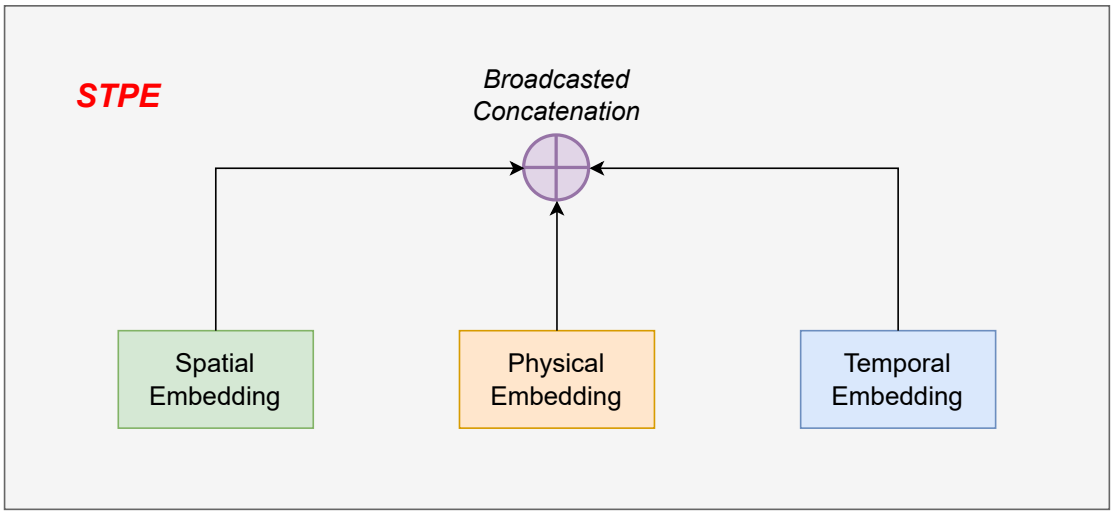


FIGURE 4.4: Spatio-temporal-physical Embedding (STPE)

variables, T is the number of time steps, and D represents the number of nodes in the traffic network. Consequently, the new spatio-temporal-physical embedding (STPE) can be represented as $STPE \in \mathbb{R}^{B \times T \times D \times F}$.

It is worth noting that operations encountering unmatched dimensions are broadcasted to ensure they have the same shape formats. The STPE now encompasses both the original spatiotemporal information and the additional variables' correlations, which will be subsequently utilized in the physical attention module. Visualisation of the structure is illustrated in Figure 4.4.

4.4.2 Physical Self-Attention

As discussed in Section 1.1, there exists a robust negative correlation between traffic speed and flow. Moreover, this correlation is highly dynamic concerning both spatial and temporal dimensions. Drawing inspiration from this concept and the temporal and spatial attention mechanisms employed in various spatiotemporal neural networks, such as those presented in GMAN [141] and the PDFormer [41], this study introduces a physical attention mechanism. The primary objective of this mechanism is to capture the dynamic correlations among traffic variables. The central idea is to assign varying weights to traffic speed and flow at different times and locations, thus enhancing the model's ability to adapt to changing traffic conditions.

For a given traffic variable v_i at time step t_j and location d_k, we calculate a weighted sum that considers all nodes in the network and all traffic variables, as depicted in Equation (4.4):

$$hw^{l}_{v_i,t_j,d_k} = \sum_{v \in V, d \in D} \alpha_{(v_i,v),(d_k,d)} \cdot h^{l-1}_{v,t_j,d}. \tag{4.4}$$

In this equation, V represents the set of traffic variables, D denotes the nodes within the network, and l corresponds to the l^{th} attention block to be computed.

The current traffic conditions at a given time step can be influenced by spatial correlations between nodes in the network. Furthermore, these conditions can vary across different traffic variables. For instance, the traffic flow might be indirectly impacted by the traffic speed at another road intersection. Drawing inspiration from this idea, we compute the attention scores by considering both the graph structure and the correlations among traffic variables. Concretely, we concatenate the hidden states with the STP (Spatio-Temporal-Physical) embedding. Subsequently, we employ the same scaled dot product approach, originally introduced in the Transformer [110], to calculate these attention scores. This approach enables us to capture the intricate interplay between spatial structures and traffic variable relationships in the context of traffic modelling, expressed in Equation (4.5):

$$w_{(v_i,v),(d_k,d)} = \frac{\langle h^{l-1}_{v,t_j,d} || e_{v_i,t_j,d_k}, h^{l-1}_{v,t_j,d} || e_{v_i,t_j,d_k} \rangle}{\sqrt{2D'}} \tag{4.5}$$

In this equation, the symbol $||$ represents the concatenation operation, $\langle\rangle$ denotes the inner product operation, and $2D'$ represents the dimension of the concatenated matrix. To normalise and transform the attention scores into probabilities, a softmax operation is applied, formulated in Equation (4.6).

$$\alpha_{(v_i,v),(d_k,d)} = \frac{\exp w_{(v_i,v),(d_k,d)}}{\sum_{v' \in V, d' \in D} \exp w_{(v_i,v'),(d_k,d')}} \tag{4.6}$$

To maintain consistency with the original implementations of spatial and temporal attention in GMAN, we extend the physical attention to incorporate multiple parallel

modules, known as multi-headed physical attention. This involves several equations shown below.

$$hw^{l}_{v_i,t_j,d_k} = ||_{k=1}^{K}\{\sum_{v\in V,d\in D} \alpha_{(v_i,v),(d_k,d)} \cdot f^{k}_{s,1}(h^{l-1}_{v,t_j,d})\} \tag{4.7}$$

$$w_{(v_i,v),(d_k,d)} = \frac{\langle f^{k}_{s,2}(h^{l-1}_{v,t_j,d}||e_{v_i,t_j,d_k}), f^{k}_{s,3}(h^{l-1}_{v,t_j,d}||e_{v_i,t_j,d_k})\rangle}{\sqrt{2D'}} \tag{4.8}$$

$$\alpha_{(v_i,v),(d_k,d)} = \frac{\exp w^{(k)}_{(v_i,v),(d_k,d)}}{\sum_{v'\in V,d'\in D} \exp w^{(k)}_{(v_i,v'),(d_k,d')}} \tag{4.9}$$

In these equations, the variable k represents the number of attention heads to be utilised. Additionally, the functions $f^{k}_{s,1}$, $f^{k}_{s,2}$, and $f^{k}_{s,3}$ refer to three distinct projectional mappings, each leading to $\frac{2D'}{k}$ outputs. The parallel attention heads help to improve the model's ability to focus on different aspects of the correlations in the data, and is a crucial step in our implementation.

4.4.3 Weighted Attention Fusion

The outputs from the spatial, temporal and physical attention each emphasise the correlations of elements on their own dimension. To effectively learn from these attention outputs, it is necessary to fuse the results based on their importance to the final prediction. Therefore, we design an attention fusion mechanism to encode the outputs from different attention modules.

In the l^{th} block, denoting the outputs from spatial, temporal and physical attentions as $H^{(l)}_S \in \mathbb{R}^{T\times D\times K}$, $H^{(l)}_T \in \mathbb{R}^{T\times D\times K}$ and $H^{(l)}_P \in \mathbb{R}^{T\times D\times K}$, where T is the number of time steps in each instance, D is the number of nodes in the network and K is the dimension size of the attention outputs, the fused results can be calculated a weighted sum:

$$H^{(l)}_{fused} = (x+y)\otimes H^{(l)}_S + (y+z)\otimes H^{(l)}_T + (x+z)\otimes H^{(l)}_P \tag{4.10}$$

where x, y and z are the normalised weights:

$$x = \sigma(H^{(l)}_S W_{z,1} + H^{(l)}_P W_{z,2} + b_{z1}) \tag{4.11}$$

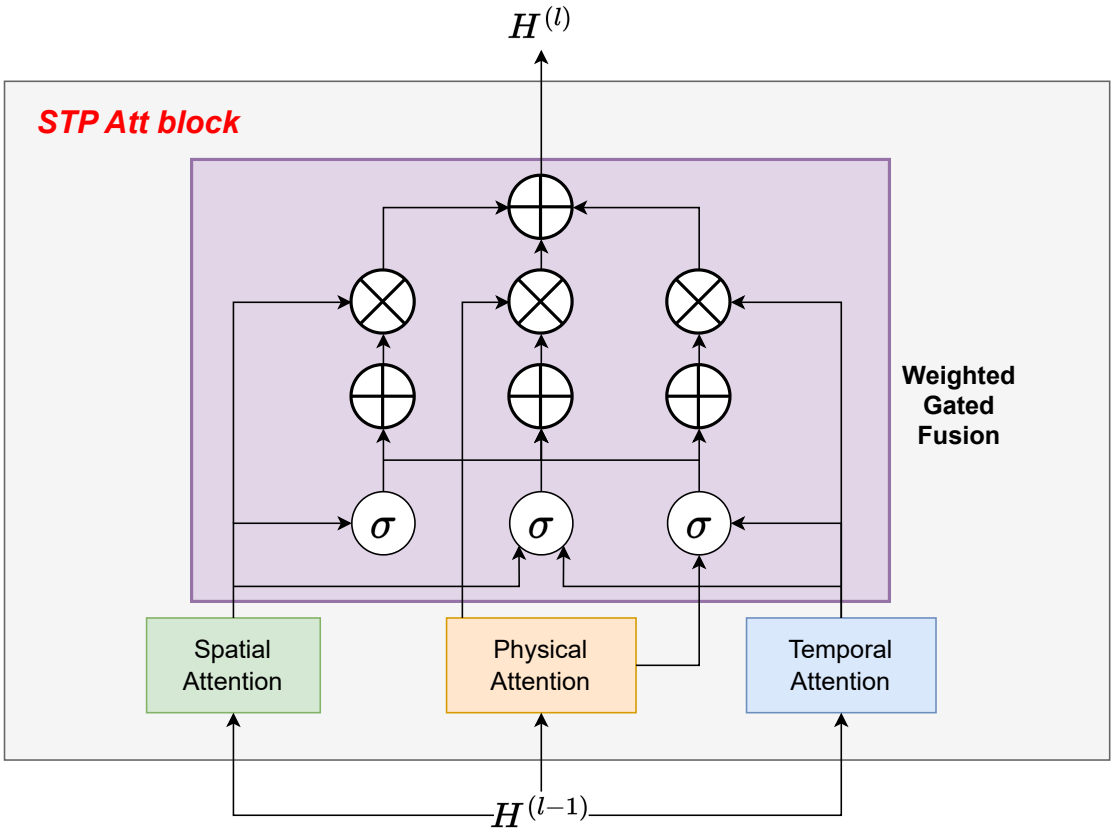


FIGURE 4.5: STP-Attention Block

$$y = \sigma(H_S^{(l)} W_{z,3} + H_T^{(l)} W_{z,4} + b_{z2}) \tag{4.12}$$

$$z = \sigma(H_T^{(l)} W_{z,5} + H_P^{(l)} W_{z,6} + b_{z3}). \tag{4.13}$$

Here, the sigmoid function denoted by σ is utilised as an activation function. The parameters $W_{z,1}$ to $W_{z,6}$ and b_{z1} to b_{z3} are learnable model weights, with shapes of $\mathbb{R}^{D\times D}$ and $\mathbb{R}^{D}$ respectively. The symbol $\otimes$ indicates the element-wise product operation. The attention fusion mechanism ensures that the model effectively combines and weighs the significance of spatial, temporal and physical correlations, which further enhances the model's adaptability. We visualise the fusion process in Figure 4.5.

4.4.4 Dynamic Correlation Generation

Inspired by the dynamic graph structure introduced in DDGCRN [119], we designed a dynamic correlation generation method for the model. This method is tailored to adaptively learn the dependencies between traffic variables at different times and locations.

The generation process starts with a random physical embedding $E_p^{t,d} \in \mathbb{R}^{V \times D}$, where V is the number of traffic variables of interest and D is the embedding size. Denoting the temporal embedding as E_T and spatial embedding as E_D, the new physical spatiotemporal embedding could be formulated as the element-wise product of the three embeddings, where the $\otimes$ denotes the element-wise product:

$$E_p^{st} = E_p \otimes E_T \otimes E_D \tag{4.14}$$

To obtain the dynamic correlation, the following steps are employed. First, we generate the dynamic traffic signal at the current time step t:

$$X_{dynamic} = FCL(X_t) \tag{4.15}$$

In Equation 4.15, FCL stands for Fully Connected Layers, which introduce nonlinearities into the generation process. Subsequently, we perform an element-wise product on the generated dynamic signal $X_{dynamic}$ and E_p to derive the dynamic physical embedding. This embedding is then activated using the $tanh$ function:

$$E'_p = \tanh(X_{dynamic} \otimes E_p^{st}). \tag{4.16}$$

Inspired by the idea that a graph can be constructed based on node similarity [119], we multiply the dynamic physical embedding E'_p with its transpose matrix $E_p^{'T}$ to yield the final dynamic physical correlation matrix. The entire process is demonstrated in Figure 4.6.

There are two main advantages associated with our proposed generation process. Firstly, this approach enhances the model's ability to capture complex and evolving relationships among traffic variables. Traditional adaptive matrices tend to focus solely on time horizons, thus ignoring potential benefits from correlation learning. Secondly, the matrix does not depend on predefined relationships or equations, making it particularly valuable for situations where prior information is scarce or absent. However, it is important to highlight that the dynamic generation method prioritises the model's prediction

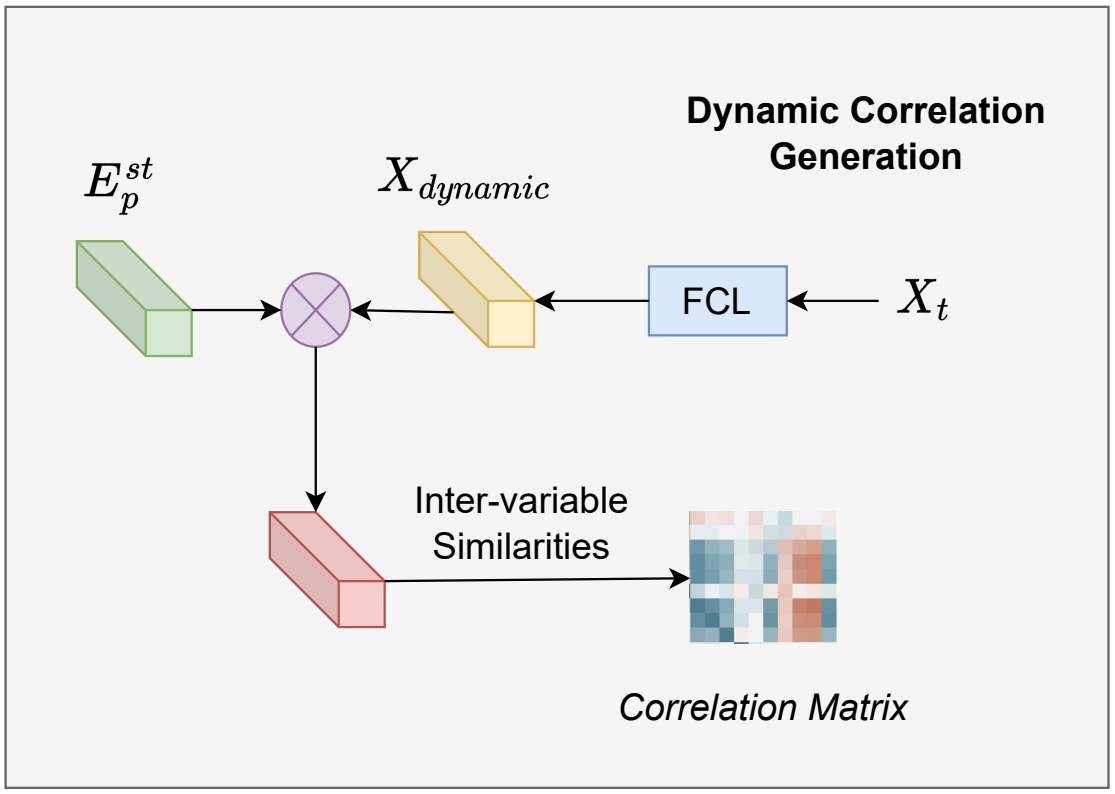


FIGURE 4.6: Dynamic Correlation Generation Process

accuracy over adherence to the actual traffic network structure. Consequently, in situations where two distant nodes lack physical interactions, the algorithm may assign them significant weights to influence the final loss calculation. The delicate equilibrium between prediction performance and empirical patterns is worth investigating in future research.

4.5 Learning Bias

4.5.1 PIDL Framework Overview

Compared with observational and inductive bias, learning bias serves as a softer obeyance to the system's underlying physics. The key idea is to incorporate the non-compliance cost of the physical laws L_{PHY} into the total cost function L_{total}, which is originally designed to include only the deep learning loss L_{DL} [92]. This process is commonly referred to as PIDL (Physics-Informed Deep Learning) rather than PIML, emphasising its focus on loss aspects. During the training iterations of a PIDL approach, the model weights are repeatedly updated until the combined cost of the physics and deep learning

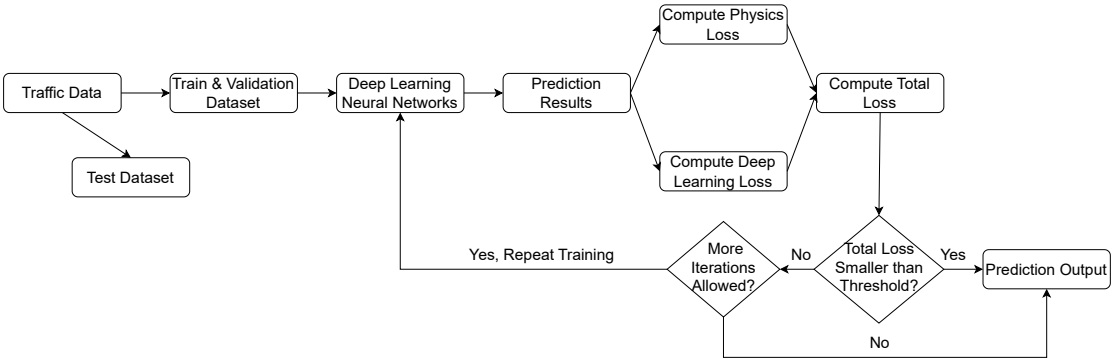


FIGURE 4.7: Flow of PIDL

components falls below a predefined threshold. To prevent the learning process from continuing indefinitely in case of no change in the total cost, a maximum allowed learning iteration, often referred to as the "patience" of the model, is set. The complete flow of the process is illustrated in Figure 4.7.

To control the balance between prediction accuracy and compliance with physics, we introduce an additional parameter, $\alpha \in [0, 1]$, to regulate the weighting of various costs. Consequently, the overall loss can be expressed as the weighted combination of the deep learning loss and the physics-informed loss, as shown in Equation (4.17).

$$L_{total} = \alpha \times L_{DL} + (1 - \alpha) \times L_{PHY} \tag{4.17}$$

Taking the traffic network partition and data calibration of physics models to be addressed in Section 4.5.2 to 4.5.6, the overall framework could be represented in Figure 4.8.

4.5.2 Traffic Network Parition

It is commonly observed that different sub-regions within the traffic network exhibit distinct traffic flow characteristics. For instance, in Figure 4.9, a small network in the Melbourne region displays links that are clustered and colour-coded based on their flow-density relationships.

Previous research commonly computes physical loss constraints based on patterns observed across the entire network. Specifically in the implementation, the key parameters such as the jam density and free flow speed of the traffic are calibrated from

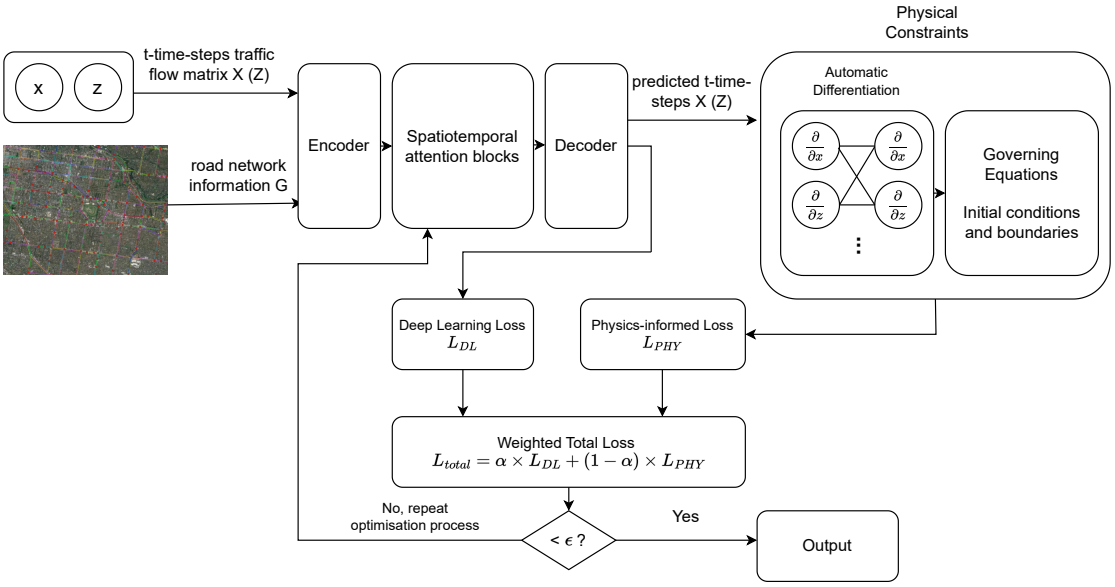


FIGURE 4.8: Overall Framework of PIDL for Spatiotemporal Traffic Prediction Models

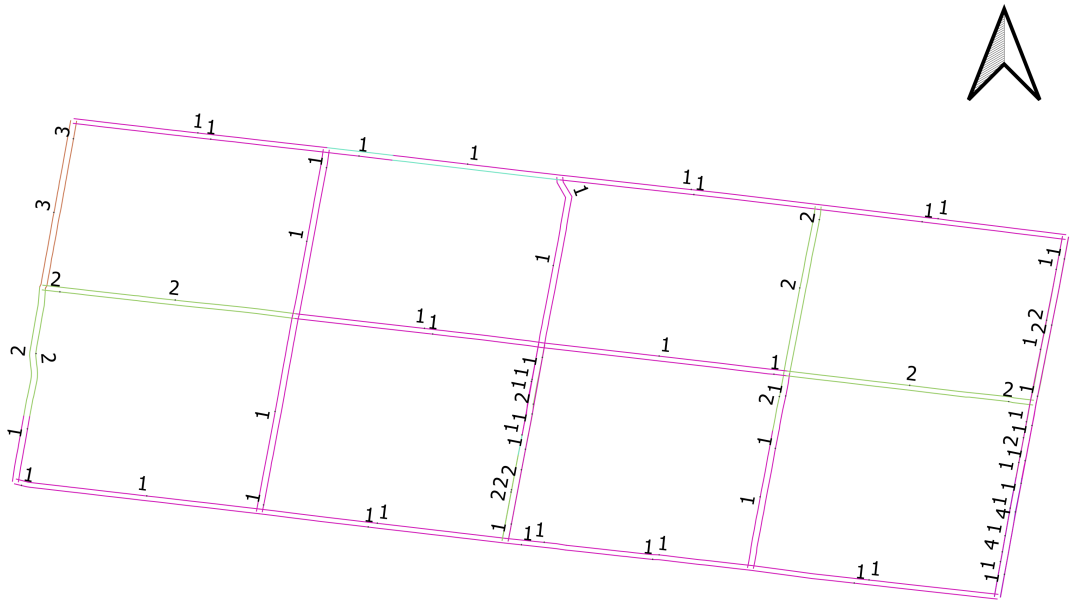


FIGURE 4.9: Test Region in Melbourne with Color-coded Links based on Clustering Results (obtained from Zhang et al. [71])

the entire dataset. However, two noticeable shortcomings are associated with this approach. Firstly, given the uneven and inconsistent distribution of congestion, different road segments may display highly distinct traffic conditions. Consequently, one parameter estimation for the entire network state can result in inaccuracies when calculating the model's physics loss. Secondly, the boundary conditions derived from equations are largely influenced by the properties of a road segment and can vary significantly from one region to another. Consequently, combining these calculations can yield averaged results that fail to represent the unique high or low traffic conditions, further compromising the fidelity to the underlying physics.

In this thesis, we perform clustering on both the PeMS and Melbourne datasets and aggregate the partitioned losses to construct the final physics loss. When clustering based on similarities, we use similar algorithms as proposed by Mohammadreza et al. [96], shown in Figure 4.10. The inputs are the adjacency matrices associated with the PeMS and the Melbourne dataset respectively, while the outputs are the two sets of clusters produced by the algorithm.

For each cluster group, we select two key parameters commonly used in traffic flow modelling to describe the road conditions. The first parameter is the jamming density, denoted as ρ_{jam}. This represents the density at which traffic speed becomes zero, indicating a congested state where traffic flow is blocked, which matches the ρ_m mentioned in Section 2. The second parameter is the free flow speed, represented as v_{free}. This speed reflects the ideal road conditions where traffic flows freely and is typically 5 to 10 times greater than the maximum road speed limit.

To calculate the two parameters for a specific cluster group, we initiate the process by computing the average speed and flow across all nodes within that cluster. This data is then visualised in the form of a scatter plot. Subsequently, we employ a linear curve fitting approach to derive the coefficients for the fitted curves. Assuming that the fitted curve adheres to the equation $y = kx + b$, with the x- and y-axes representing speed and flow, respectively, we can determine ρ_{jam} by setting $x = 0$, resulting in the outcome $\rho_{jam} = b$. Similarly, we can deduce that $v_{free} = -\frac{b}{k}$. An illustrative example of parameter estimation for a freeway segment from the PeMS dataset is depicted in Figure 4.11. Note that the speed axis does not begin at zero, and intercepts are not represented in the

A. Running 'Snakes'
X: set of roads (links)
for $x \in X$ **do**
  $S_x \longleftarrow x$;
  while $Size(S_x) < N$ **do**
    $S' = \text{Adj}(S_x)$; (Neighboring links of the snake)
    $k^* = \{k | \min_{k \in S'} var(S_x \cup k)\}$;
    $S_x \longleftarrow [S_x, k^*]$;
B. Computing similarities
initialize ϕ;
$W \longleftarrow \mathbf{0}_{[N \times N]}$;
for $\forall i, j \in X$ **do**
  $k \longleftarrow 1$;
  while $k \leq N$ **do**
    $w(i, j) = w(i, j) + \phi^{N-k} \times \text{intersect}(S_{ik}, S_{jk})$;
    $k \longleftarrow k + 1$;
C. Symmetric Non-negative Matrix Factorization
$D = diag(d_i)$ **where** $d_i = \sum_{j=1}^{N} w(i, j)$;
$\tilde{W} = D^{-1/2} W D^{1/2}$;
$H^* = \{H | \min_{H \in \mathcal{R}_+^{N \times N_s}} \|\tilde{W} - HH^T\|^2\}$;
for $i \in X$ **do**
  $j^* = \{j | \max_{j \in \{1,\ldots,N_s\}} H(i, j)\}$;
  $i \in A_{j^*}$;
Output: $\{A_1, A_2, \ldots, A_{N_s}\}$ (Set of clusters)

FIGURE 4.10: Snake Algorithm (obtained from Mohammadreza et al. [96])

figure. The clustering results, parameter estimations, and the representative roadway IDs (where most roads are located within that cluster) for the two datasets are presented in Tables 4.2 and 4.3.

Cluster ID	Jam Density	Free Flow Speed	Representative Link ID
1	28.56	73.89	10
2	30.23	77.72	710
3	30.48	76.80	110
4	28.31	79.57	5
5	34.61	73.98	60
6	27.00	80.36	101
7	20.55	77.72	710
8	16.28	71.96	2

TABLE 4.2: PeMS Cluster Results and Parameter Estimation

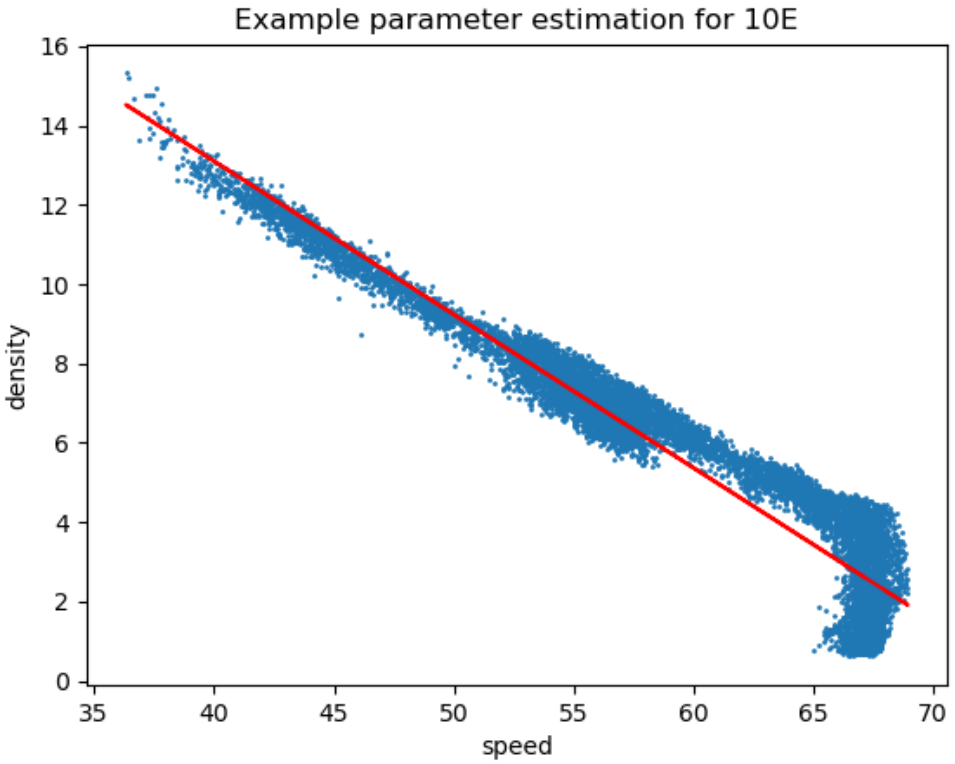


FIGURE 4.11: Example Parameter Estimation for PeMS Freeway 10E

Cluster ID	Jam Density	Free Flow Speed	Representative Freeway ID
1	26.99	55.95	251
2	25.12	57.21	962
3	11.06	105.60	443
4	15.14	62.57	1321
5	6.07	102.49	5148
6	10.36	99.59	4412

TABLE 4.3: Melbourne Cluster Results and Parameter Estimation

4.5.3 Formulation using Greenshield's Model

The most straightforward expression of the physical constraint can be directly obtained from Greenshield's model, as depicted in Equation 4.18. In this context, ρ_m represents the maximum (jam) density, and v_{free} denotes the free flow speed. In the equation, N_c stands for the collocation points to be estimated. Assuming that the evaluation metric used for the physics-uninformed part is a mean squared error and that N_0 is the size of observed points. Similarly, the uninformed deep learning loss is formulated in equation 4.19. The two losses could be combined as indicated in equation 4.17.

$$L_{PHY} = \frac{1}{N_c}\sum_{i=1}^{N_c}\left|\tilde{v}(x_c^i, t_c^i) - (1 - \frac{\tilde{\rho}(x_c^i, t_c^i)}{\rho_m})\right|^2 \tag{4.18}$$

$$L_{DL} = \frac{1}{N_0}\sum_{i=1}^{N_0}\left|v(x_c^i, t_c^i) - \tilde{v}(x_c^i, t_c^i)\right|^2 \tag{4.19}$$

4.5.4 Formulation using LWR Conservation Law

The LWR conservation law could be derived based on the analytical solution to Greenshield's FD (2.2). The derived equation is shown in 4.20, where the traffic speed v is established with respect to both location x and time t.

$$\rho_{max}(1 - \frac{2v(x,t)}{v_{free}})\frac{\partial v(x,t)}{\partial x} - \frac{\rho_{max}}{v_{free}}\frac{\partial v(x,t)}{\partial t} = 0 \tag{4.20}$$

To embed the physical constraints into the loss functions, we could measure the extent of noncompliance of the equation, formulated in 4.21. The uninformed loss L_{DL} is shown in 4.22. The weighted sum of the two items constitutes the total loss, as indicated by equation 4.17.

$$L_{PHY} = \frac{1}{N_c}\sum_{i=1}^{N_c}\left|\rho_{max}(1 - \frac{2\tilde{v}(x_c^i, t_c^i)}{v_{free}})\frac{\partial \tilde{v}(x_c^i, t_c^i)}{\partial x_c^i} - \frac{\rho_{max}}{v_{free}}\frac{\partial \tilde{v}(x_c^i, t_c^i)}{\partial t_c^i}\right|^2 \tag{4.21}$$

$$L_{DL} = \frac{1}{N_0}\sum_{i=1}^{N_0}\left|v(x_c^i, t_c^i) - \tilde{v}(x_c^i, t_c^i)\right|^2 \tag{4.22}$$

4.5.5 Curve Fitting

While Greenshield's model and the LWR conservation law are widely utilised in transportation studies to model traffic flow characteristics, they both hinge on a critical assumption. This assumption posits that the problem under consideration is homogenous, with the fundamental diagram being solely a function of density ρ, expressed as $q(x, t, \rho(x, t)) = Q(\rho(x, t))$. In simpler terms, it implies that both speed and flow are solely dependent on changes in density. However, this assumption often fails to hold true in many real-world scenarios. For instance, in both the PeMS and Melbourne datasets

we collected, a strong negative correlation between traffic speed and flow was evident, suggesting a potential nonlinear functional mapping between these variables. In such situations, the traditional models struggle to accurately depict the underlying traffic data dynamics.

In this thesis, we propose an innovative curve-fitting strategy to address this challenge. Specifically, by fitting an optimised curve to the distribution of speed and flow, we aim to derive a more precise mathematical formulation describing the relationship between these variables. This approach offers two distinct advantages over previous methods. First and foremost, it doesn't rely on assumptions about the dependencies of fundamental diagrams or other fixed relationships. Instead, the coefficients of the physical equation are determined solely by the fitted curve, making the approach adaptable to various datasets. Secondly, the process of curve shape selection can be automated, greatly increasing the likelihood of obtaining a highly optimised equation that accurately captures the underlying physics reflected through the correlations between these variables.

To simplify the fitting process, we make an assumption that the speed and density distributions can be represented by polynomial functions. This means that the speed of any point on the graph can be expressed as a polynomial function of the corresponding density value. To automate the selection of the best polynomial degree, we designed an algorithm that allows the program to search within a specified range. In our implementation, we set the maximum degree to 5. The evaluation metric we used is the Mean Absolute Error (MAE).

The algorithm is presented in Alg. 1, and in this algorithm: v_mean, ρ_mean, and q_mean represent the average speed, density and flow across all nodes in the network respectively, max_d is the maximum degree allowed for the polynomial fitting, $polynomial\_fit$ and $calculate\_error$ are built-in Python functions that we can directly use in the program. By using these elements, we aim to streamline the process of fitting the data to polynomial functions, making it more generalisable to various datasets. To be consistent with the model training process, the curves are fitted on the first 70% of data. It is important to note that obtaining any of the two variables from speed, flow, and density can lead to the determination of the third variable, following the fundamental relationship 2.1. Consequently, fitted curves derived from any of these two variables

can be utilised in the formulation of loss constraints.

Algorithm 1 Selection of the Best Fitted Curve

Require: $n \geq 1$
  $degree \leftarrow 1$
  $best\_degree \leftarrow 1$
  $X \leftarrow v\_mean$
  $Y \leftarrow \rho\_mean$ or q_mean
  $max\_d \leftarrow n$
  $error \leftarrow \inf$
  while $degree \leq max\_d$ **do**
    $\tilde{Y} \leftarrow 0$
    $coef\_lst \leftarrow polynomial\_fit(X, Y, degree)$
    for $i \leftarrow 0$ to $degree$ **do** $\tilde{Y} = \tilde{Y} + X^{degree} \times coef[i]$
    end for
    $cur\_error \leftarrow calculate\_error(Y, \tilde{Y})$
    if $cur\_error < error$ **then**
      $error \leftarrow cur\_error$
      $best\_degree \leftarrow degree$
    end if
    $degree \leftarrow degree + 1$
  end while

The main caveat of using the automatic search algorithm is the increasing time complexity, especially for large-scale data analysis. The time complexity originates from two parts: the iteration space to find the best coefficient degree and the curve-fitting process. For the iteration space, depending on the complexity of traffic variable distribution, the best-fitted degree may not be found in early iterations. For the polynomial curve fitting process, taking matrix inversion as the solution technique, the time complexity is $O(M^3 + NM^2)$, where M is the degree of the polynomial and N is the size of data. Consequently, the computational cost of combining two parts increases sharply as the degree and data size escalate.

The results of the fitted curves for the two datasets are illustrated in Figure 4.12. As indicated by the algorithm, the Melbourne and PeMS datasets are fitted with 3- and 4-degree polynomials respectively.

4.5.6 Formulation using Fitted Equations

One of the advantages of incorporating learning bias into deep learning models, compared to observational and inductive bias, is the introduction of a soft penalty for predictions that do not align with the laws of physics. It serves as a regularisation term for

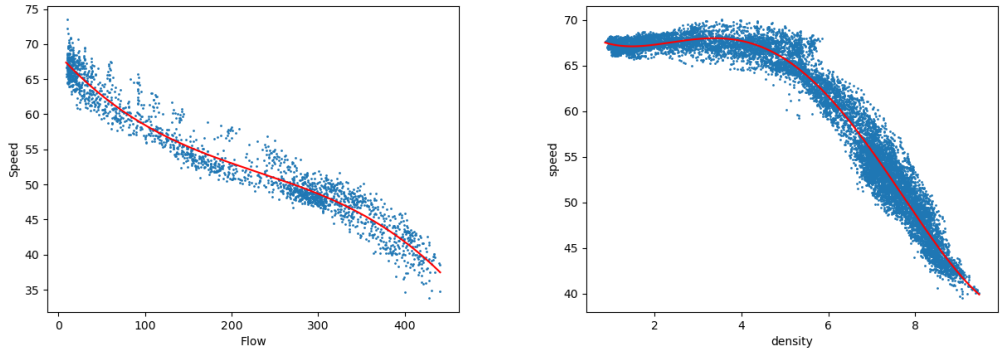


FIGURE 4.12: Fitted Curves for Melbourne (left) and PeMS (right) datasets

the physics-uninformed loss, guiding the model's convergence towards adherence to the underlying physical laws.

In the context of traffic prediction, it is important to recognise that fitted curves provide only approximate estimations of the distribution of variables. Therefore, it is more appropriate to establish an acceptable range for assessing compliance. For instance, we can set a threshold, denoted as m, and consider all data points within this range as good predictions. They do not contribute to the final loss. However, data points falling beyond this threshold are penalised, and the extent of the penalty is proportional to the distance by which they exceed the accepted range. Assuming that the fitted function is f, and the maximum deviation (range) identified is M. Then the loss for the i^{th} instance can be expressed as 4.23:

$$L_i = max(0, abs(\tilde{Y}_i - f(X_i)) - M). \tag{4.23}$$

Accordingly, the total physics loss could be formulated as 4.24, where N is the number of predicted instances.

$$L_{PHY} = \sum_{i=0}^{N} max(0, abs(\tilde{Y}_i - f(X_i)) - M) \tag{4.24}$$

Chapter 5

Experiments

This chapter delves into a comprehensive analysis of the research findings, providing insights into different settings and their impact on the model's performance. This chapter is divided into six sections, each shedding light on a different perspective:

Section 5.1 outlines the systems and platforms utilised in the experiments, offering a detailed overview of the technical infrastructure that underpins our research.

Section 5.2 describes the experimental settings and baselines employed, providing a solid foundation for the subsequent discussions on the results.

Section 5.3 serves as the core of the research analysis, where we present and interpret the outcomes of the experiments, highlighting key findings and their implications.

Section 5.4 investigates the influence of variations in input-output settings and time horizons on the model's performance, offering valuable insights into its superiority in long-term predictions.

Section 5.7 examines the impact of essential hyperparameters within our extended models. Furthermore, we provide an in-depth analysis of the trade-offs between efficiency and accuracy stemming from these choices.

Section 5.6 assesses the model's performance under conditions of limited training data, shedding light on its robustness and reliability.

Finally, Section 5.8 engages in an in-depth discussion that synthesises the research findings, placing them within the broader context of traffic prediction research. We will

critically examine the strengths and weaknesses of the proposed framework and analyse trade-offs that should be considered when adopting it in real-world applications.

These subsections collectively offer a clear roadmap for our experiments, providing a comprehensive analysis of the research approaches and their implications.

5.1 System and Platforms

The experiments are divided into two stages. The first stage involves trial experiments on a small portion of the dataset and model parameter tuning. Google Colab with Nvidia's T4 GPU was used for those experiments. The model expansions and loss function reconstruction were implemented with PyTorch 1.10.1 and Python 3.8.1. In the second stage, computationally expensive tasks such as model training were carried out on Spartan's Nvidia A100 GPUs, with RedHat Enterprise Linux 9, Slurm 23.02.5 and Spectrum Scale 5.1.8.1. All tasks were configured to possess 64 GB RAM and 4 CPU cores.

5.2 Settings and Baselines

To maintain a fair comparison between the existing models, we divided the dataset into training, validation, and test sets with ratios of 7:1:2 for the PeMS dataset and 6:2:2 for the Melbourne dataset. Distinct time horizons are employed for the two datasets. For the PeMS dataset, we utilised data from the past 12 time steps to predict the subsequent 12 time steps, equivalent to a 1-hour time span. In the case of the Melbourne dataset, we employed data from the past 8 time steps to predict the next 8 time steps, covering a 2-hour time span. The choice of 8 steps is attributed to the 15-minute time interval of the Melbourne dataset, which results in three times fewer observations compared to the PeMS dataset. Therefore, a longer time span ensures an adequate number of training samples for each instance. When displaying results, the averaged values ranging from 1 to the max horizon are used. The selection of optimal models is based solely on the validation set. Training epochs and stopping patience parameters are set at 300 and 15, respectively, for both GMAN and DDGCRN. Each experiment was repeated five times and the averaged result will be present.

For our experiments, we implemented two types of extended models: the correlation-enhanced (CE-) models and the physics-informed (PI-) models. In the CE models, we introduce inductive bias through physical attention and dynamic correlation generation, which is detailed in subsection 4.4.2 for GMAN and subsection 4.4.4 for DDGCRN. On the other hand, the PI models are created by incorporating learning bias through additional loss constraints derived from the fitted curve of the prediction distribution, as introduced in Subsection 4.5.6. These two categories of biases provide two distinct perspectives for leveraging the correlations between traffic variables.

When choosing the optimal values for embedding dimensions and the number of attention splits, our selection is informed by a combination of empirical observations from the original model implementations and the theoretical function of these parameters. In terms of the embedding dimension, the difference between the two embedding matrices should encapsulate the distinct characteristics between speed and flow at a specific timestamp and location. This is because the physical embeddings are explored in parallel with the temporal and spatial embeddings, and they are fused through the attention fusion module. Consequently, the relationship between these two variables dynamically evolves concerning spatiotemporal scales. Similarly, the number of splits should reflect the complexity of the correlation involved. We will provide a more in-depth exploration of the impact of these hyperparameters on the model's performance in Section 5.7.

To demonstrate the effectiveness of the extended models, 9 baselines within 3 different categories were selected for comparison. A detailed description of the models and the rationale for selections are explained below.

1. Time-series based models.
    (a) Static prediction. This method simply shifts the historical timestamps to fill the current prediction values without calculation. It serves as the most basic baseline to test if other models can have an improvement in accuracy upon calculations.
    (b) HA [6]. This is one of the simplest time-series models. It calculates the average of the historical values in a specified time range as the prediction for the next step. Therefore, the results will be the same for each time horizon in the future. A 7-day historical period was used for both datasets to infer the next hour's values.

(c) GC-VAR [64]. A multivariate time-series model that takes into account the Granger Causality for network-wide traffic prediction. This model takes into account the spatial information from a multivariate perspective, thus selected to be compared with univariate ones. A 12-step time lag parameter was used.

(d) SARIMA [120]. The model extends the original ARIMA model by including seasonal components, denoted as Seasonal Autoregressive (SAR) and Seasonal Moving Average (SMA), which capture the univariate periodic fluctuations in the data. The model is demonstrated to perform better on traffic data with strong daily and weekly periodic patterns [109]. Seasonal orders for different nodes are determined with the Autoarima module automatically.

2. GNN-based models.

(a) STGCN [134]. The model integrates graph convolution with gated temporal convolution through spatiotemporal convolution blocks and is commonly used as a baseline evaluation for spatiotemporal traffic prediction models.

(b) GWNET [123]. An adaptive dependency matrix was developed in the model to better capture hidden spatial dependencies. The novel encoding of spatial information leads to the model's competitive results in multiple benchmark datasets.

(c) MTGNN [124]. The model adapts GNN to multivariate time-series forecasting by introducing graph learning and convolution layers. The model extends GWNET with more graph learning modules and the concept has been widely inherited for spatial information encoding for traffic prediction.

3. Self-attention-based models and other mixed types.

(a) DCRNN [56]. The model combines bidirectional-random-walk diffusion convolution with RNN for spatial and temporal dependencies. This is one of the earliest models that adopts a hybrid of sequence and image learning modules for spatiotemporal modelling of traffic data.

(b) GMAN [141]. A spatiotemporal traffic prediction model that adopts parallel multi-attention modules to model both spatial and temporal correlations.

(c) DDGCRN [119]. A mixed-type traffic prediction model that uses the Chebyshev polynomial to simulate the graph convolution operations and RNN for time-series modelling.

Aside from the time-series-based models, all other baseline models' source codes are available on their corresponding GitHub repositories [53, 54, 118, 121, 122, 133].

To evaluate and compare various models, and ensure consistency with previous research, three metrics are adopted: mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean squared error (RMSE). The primary goal of traffic prediction is to minimise the difference between predicted future traffic conditions and observed conditions. Thus, the metrics are formulated in Equations (5.1), (5.2), (5.2), where N indicates the number of samples and H represents the number of time steps to predict. Thus, X_t^i indicates the ground truth value of the i^{th} instance at time step t, and $\tilde{X}_t^i$ is the corresponding predicted value. Here, $X \in \mathbb{R}^{N \times H}$ is the averaged value across all the nodes in the network, consisting of only a single traffic variable, speed or flow.

$$MAE = \frac{1}{H}\sum_{t=1}^{H}\frac{1}{N}\sum_{i=1}^{N}|X_t^i - \tilde{X}_t^i| \tag{5.1}$$

$$MAPE = \frac{1}{H}\sum_{t=1}^{H}\frac{1}{N}\sum_{i=1}^{N}|\frac{X_t^i - \tilde{X}_t^i}{X_t^i}| \tag{5.2}$$

$$RMSE = \sqrt{\frac{1}{H}\sum_{t=1}^{H}\frac{1}{N}\sum_{i=1}^{N}|X_t^i - \tilde{X}_t^i|^2} \tag{5.3}$$

5.3 Overall Results

The comparison results for the baselines on the two datasets are calculated for both 1-hour and 3-hour time horizons, which are shown in Tables 5.1 and 5.2. The bold values represent the best results in the column, while the underlined values are the second best.

The accuracy, as directly indicated by the error metrics, clearly shows that the extended models (CE and PI) consistently outperform other baseline models, including their unextended versions. Notably, in comparison to the performance on the 1-hour time horizon, the 3-hour predictions exhibit more significant improvements when leveraging correlations among traffic variables. This observation suggests that the positive impact of correlations becomes more pronounced as the prediction time horizon increases.

Dataset	PeMSD7_2022			Melb_2023		
Metric / Model	MAE	RMSE	MAPE	MAE	RMSE	MAPE
Static	23.03	39.84	13.43%	65.06	111.53	53.75%
HA	36.62	61.71	23.75%	73.77	152.13	72.61%
VAR	18.06	34.14	13.08%	36.40	59.56	35.54%
SARIMA	31.89	37.85	25.95%	83.7	95.8	61.78%
DCRNN	15.68	27.92	11.83%	24.77	38.92	32.53%
STGCN	14.29	26.37	10.53%	23.49	37.65	30.60%
GWNET	11.06	23.15	8.23%	22.98	37.10	27.66%
GMAN	11.10	<u>22.10</u>	7.90%	18.34	**29.54**	29.03%
MTGNN	<u>10.53</u>	22.23	7.66%	22.75	36.62	27.43%
DDGCRN	10.90	22.92	<u>7.26%</u>	20.55	35.54	19.75%
CE-GMAN	11.16	**21.78**	8.53%	<u>18.09</u>	<u>29.93</u>	26.71%
CE-DDGCRN	**10.49**	22.51	**6.63%**	19.27	33.01	**18.48%**
PI-GMAN	11.56	25.46	10.45%	**17.59**	30.06	21.18%
PI-DDGCRN	11.04	23.06	7.92%	18.92	31.78	<u>19.05%</u>

TABLE 5.1: Overall Results for Short-term Prediction (**bold**: best result and <u>underline</u>: second best result)

Dataset	PeMSD7_2022			Melb_2023		
Metric / Model	MAE	RMSE	MAPE	MAE	RMSE	MAPE
Static	54.20	87.64	32.49%	145.37	222.64	198.58%
HA	36.62	61.71	23.75%	73.77	152.13	72.61%
VAR	48.33	75.36	29.04%	97.36	175.48	154.15%
SARIMA	53.60	82.98	42.63%	125.12	186.75	167.94%
DCRNN	23.47	32.93	22.84%	26.92	49.06	37.66%
STGCN	21.84	30.26	23.75%	27.13	42.55	33.64%
GWNET	18.47	27.36	20.15%	24.34	38.63	30.82%
GMAN	15.84	29.29	14.97%	18.06	29.21	28.40%
MTGNN	16.54	30.63	13.82%	21.19	33.06	31.84%
DDGCRN	17.12	32.91	11.61%	24.26	42.61	26.77%
CE-GMAN	**11.84**	**25.68**	9.35%	**16.35**	**28.53**	23.26%
CE-DDGCRN	<u>13.06</u>	27.85	**8.44%**	20.63	35.06	**21.73%**
PI-GMAN	14.05	28.39	10.06%	<u>18.24</u>	<u>30.31</u>	25.82%
PI-DDGCRN	13.29	<u>26.93</u>	<u>9.21%</u>	19.75	32.06	<u>21.93%</u>

TABLE 5.2: Overall Results for Long-term Prediction (**bold**: best result and <u>underline</u>: second best result)

This phenomenon can be attributed to several factors. Firstly, the correlation between traffic speed and flow remains relatively consistent over time. The inductive biases incorporated into the models reflect the relationship between traffic variables, rather than the specific values of these variables, which are inherently time-invariant. Consequently, extending the time horizon does not significantly impact the predictions based on these

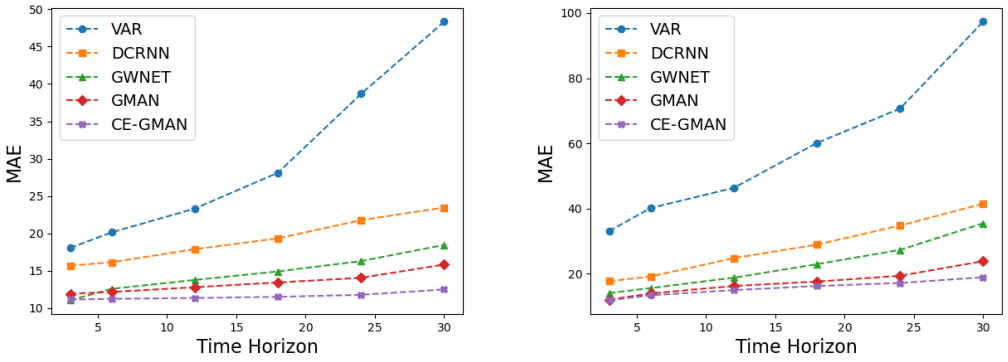


FIGURE 5.1: Comparison of Errors Over Time on PeMS (left) and Melbourne (right) dataset

correlations. Another reason for the improved performance over longer time horizons is the regularisation effect exerted by the physics loss. This regularisation tends to eliminate more outliers as the time horizon extends. With the accepted range of the variable distribution held constant, a consistent number of incorrect or invalid predictions will be removed, which results in a more pronounced improvement in prediction performance. This effect of outlier elimination becomes more apparent as the prediction horizon becomes longer. Lastly, our designed physical attention module is dynamic with respect to time. In our implementation, the attention outputs are fused with the temporal attention module, and this dynamic interaction may enhance the model's ability to comprehend correlations over longer time horizons.

To visually compare the performance of baselines at different time horizons, we plotted the trend of prediction error indicated by the MAE for five different models: VAR, DCRNN, GWNET, GMAN and CE-GMAN. The plots for the PeMS and the Melbourne dataset are shown in Figure 5.1.

In the presented plot, it is evident that time-series-based models exhibit the least robustness as time progresses. The correlation-enhanced model shows only a marginal increment in error throughout the entire time period, which indicates its good adaptability to the time change. It is noteworthy that when examining the Melbourne dataset, the CE-GMAN model exhibits a slightly larger error beginning at a time horizon of 3, in comparison to the original GMAN model. However, as time elapses, the rate of error increment is significantly slower. This observation suggests that the correlation between traffic variables may not have a beneficial impact on short-term predictions. It is possible that the values of different variables interfere with one another, leading to less

accurate estimation of single variables at early stages. However, as time advances, the correlation serves as additional knowledge to be inferred, thereby enhancing the model's performance.

The overall model accuracy on the PeMSD7 dataset surpasses that of the experiments conducted on the Melbourne dataset. This difference can be attributed to two key factors. Firstly, the Melbourne dataset is only half the size of the PeMSD7 dataset, providing insufficient information for neural networks to learn the underlying patterns effectively. Secondly, the Melbourne dataset contains disrupted data streams, while the PeMSD7 dataset is continuous in terms of the date range. Although no studies have demonstrated the negative impact of discontinuity in time-series data, it is worth noting that the encoding of the timestamps' positions does have an impact on the model's performance, as shown in a related study [7], indirectly tied to data completeness.

The variation in dataset characteristics directly influences the model's generalisation performance. Notably, the correlation-enhanced and physics-informed models exhibit much greater robustness in the face of data variation. In contrast, baseline models tend to generalize poorly, with almost a 2-3 times increase in error on the Melbourne dataset compared to PeMSD7. Conversely, the extended models show a lower rate of error increment. For the correlation-enhanced models, this improvement can be attributed to the fact that different traffic systems share similar speed-flow correlations. Similarly, the underlying physics follows comparable trends in Melbourne and California, with only minor variations in the parameters of the physical models.

5.4 Ablation Study on Longer Time Horizons

To further delve into the effectiveness of using multiple variables for traffic prediction, we conducted an ablation study. In this study, we exclusively extended the input dimensions of the models, eliminating any modifications related to inductive bias and learning bias. As a result, the Input-Output (IO) setting became the sole variation between the original and extended models.

Without loss of generality, we focused our investigation on traffic flow as the primary output variable. To assess the model's performance over longer time horizons, we selected three distinct time periods for experimentation: 12, 24, and 36 for the PeMS and 8, 16,

and 24 for the Melbourne dataset. The ranges cover time spans from 1 to 3 hours and 3 to 9 hours, providing insights into mid- to long-term prediction performance. We conducted experiments for both GMAN and DDGCRN models to ensure the generalisability of our observations across different model categories. The results are presented in Tables 5.3 and 5.4. In these tables, bold values highlight better performance in the two types of IO settings.

Dataset	IO Settings		Horizon	Metrics		
	Input	Output		MAE	RMSE	MAPE
PeMS	Flow	Flow	12	12.94	22.74	13.15
			24	14.20	26.08	13.31
			36	15.84	29.29	14.97
	Speed & Flow	Flow	12	**11.16**	**21.78**	**8.53**
			24	**11.99**	**24.79**	**9.19**
			36	**13.29**	**27.12**	**9.92**
Melb	Flow	Flow	8	**18.06**	**29.21**	28.40
			16	17.86	30.48	21.84
			24	18.83	31.26	30.41
	Speed & Flow	Flow	8	18.09	29.93	**26.71**
			16	**17.50**	**29.43**	**21.70**
			24	**17.74**	**29.19**	**25.45**

TABLE 5.3: GMAN Performance (**bold**: best result)

Dataset	IO Settings		Horizon	Metrics		
	Input	Output		MAE	RMSE	MAPE
PeMS	Flow	Flow	12	10.90	22.92	7.26
			24	13.76	27.92	8.80
			36	17.12	32.91	11.61
	Speed & Flow	Flow	12	**10.49**	**22.51**	**6.63**
			24	**12.95**	**27.12**	**7.92**
			36	**14.73**	**30.60**	**9.38**
Melb	Flow	Flow	8	20.55	35.54	19.75
			16	21.56	38.88	21.76
			24	24.26	42.61	26.77
	Speed & Flow	Flow	8	**19.27**	**33.01**	**18.48**
			16	**20.25**	**34.03**	**20.56**
			24	**22.93**	**39.58**	**24.70**

TABLE 5.4: DDGCRN Performance (**bold**: best result)

As depicted in the tables, models with expanded dimensions, incorporating both speed and flow as inputs, consistently outperform those with single input and output. Only two entries at the 8-step horizon for the Melbourne dataset did not achieve better performance. Other than that, the observation holds true for both datasets and both models,

further reinforcing our hypothesis that utilising multiple traffic variables enhances model performance compared to using a single input.

When examining longer prediction horizons, the magnitude of improvement becomes more pronounced as the forecast span extends. For instance, at the 12-step horizon, the expanded IO setting yields only a 0.5 error decrement in MAE for DDGCRN on the PeMS dataset. However, when predicting at the 36-step horizon, this difference increases to 2.39. Similar trends are observed in GMAN for both datasets, further demonstrating the advantages of leveraging correlations between traffic variables for long-term traffic prediction.

5.5 Different Categories of Physics for Learning Bias

5.5.1 Accuracy vs Convergence

One of the pivotal methods for leveraging correlations between traffic variables is the incorporation of learning bias. It aims to embed the physics governing relationships between traffic variables into loss constraints. This is achieved through a weighted combination of the physics-uninformed cost and physics-informed cost, resulting in the regularisation of prediction results and the elimination of outliers that violate physical properties.

To evaluate the performance of the three types of physics models introduced in section 4.5, we conducted experiments for each. We employed specific metrics to assess two vital aspects of these models: accuracy and efficiency. For measuring accuracy, we chose the Mean Absolute Percentage Error (MAPE) as our primary metric. MAPE indicates how closely the model's predictions align with the actual ground truth values, offering a clear assessment of average predictive accuracy. Efficiency, on the other hand, was evaluated through convergence time, which represents the total time required to train each model.

This choice of metric was motivated by two key considerations. Firstly, we sought to determine whether the computational cost of calculating physics parameters for individual sub-regions is a significant factor affecting training efficiency. Additionally, we aimed to understand whether the combined cost, which includes both physics-based and DL components, converges quicker than the sole DL cost. This consideration is particularly

important as it involves the incorporation of additional traffic variables. It is worth noting that inference time was not considered as the evaluation criteria for efficiency, as it does not reflect the impact of physics-based costs during the training phase. The unit of time is measured in minutes. The evaluation results are presented in Table 5.5, with the best-performing values indicated in bold font.

DL Model	Dataset	Physics Model	Criteria	
			MAPE	Convergence Time
GMAN	PeMS	Greenshield	18.05%	**14.3**
		LWR	15.76%	39.5
		Fitted Curve	**10.06%**	20.6
	Melb	Greenshield	41.25%	**9.8**
		LWR	30.06%	27.1
		Fitted Curve	**25.82%**	16.3
DDGCRN	PeMS	Greenshield	16.53%	**18.5**
		LWR	13.88%	41.3
		Fitted Curve	**9.21%**	31.2
	Melb	Greenshield	34.97%	**12.3**
		LWR	**19.06%**	36.1
		Fitted Curve	21.93%	22.4

TABLE 5.5: Performance of PIDL with Different Physics Models (**bold**: best result)

In terms of accuracy, the fitted curve strategy outperforms the other two models in three out of the four model-dataset settings, with the exception being the prediction from DDGCRN on the Melbourne dataset. When compared to the other two model types, the MAPE achieved by the curve-fitting equation is consistently 5 to 10 percent lower. In the context of traffic prediction, this improvement could translate to a reduction in prediction errors equivalent to 5 to 10 cars, showcasing a significant enhancement in accuracy.

Several factors may contribute to this enhancement. Firstly, the distribution of traffic variables appears to be challenging to capture using linear curves derived from Greenshield's model. Despite the complexity of the LWR model, which is based on differential equations, it doesn't align well with the characteristics of this particular dataset. As a result, the fitted curve method seems to provide the best representation of the relationship between traffic variables, thereby gaining a performance advantage.

Another noteworthy factor could be the tolerance inherent in this approach, which is controlled by an accepted range parameter M (4.23). In the process of calculating the

physics loss, the defined threshold acts to mitigate the penalty imposed on deviations from the expected physical behaviour. Consequently, data points that do not strictly adhere to the underlying physics are not entirely eliminated, striking a balance that considers both accuracy and adherence to physics. This tolerance feature allows the fitted curve strategy to be more forgiving of slight variations or anomalies in the data, which might be less forgivingly addressed by other models. It permits a degree of flexibility that can be particularly advantageous in real-world scenarios where traffic behaviour may not always perfectly conform to strict physical models.

On the efficiency dimension, it is evident that Greenshield's model consistently achieves the fastest convergence in all the settings when compared to the other two models. By contrast, the experiments utilising the LWR model as the physics loss consumes the most time to converge. This difference in efficiency can be attributed to the inherent complexity of each model. Greenshield's model relies predominantly on fundamental relationships and employs a linear modelling approach. As a result, calculating the loss in this case can be straightforward, given its reliance on linear equations. On the contrary, the calculation of non-compliance with LWR laws necessitates more intricate computations. It involves the calculation of gradients across multiple dimensions, encompassing time, space, and various traffic variables, which is inherently time-consuming.

Although the fitted curve strategy does not achieve the lowest convergence time, it demonstrates a balance between accuracy and efficiency. When averaged across all the settings, the model employing the curve-fitting strategy achieves a 11% improvement in accuracy compared to Greenshield's model. Furthermore, it incurs only approximately 9 minutes of additional convergence time. This slightly increased computational cost remains well within an acceptable range, especially when considering the substantial gain in performance.

The observations above indicate that although the fitted curve strategy does consume slightly more computational resources, it stands out as the most effective model among the three types in terms of prediction accuracy. While Greenshield's model is simple and suitable for scenarios where fast convergence is crucial, it may not be the ideal choice in application scenarios where high accuracy is the primary concern. Lastly, despite its increased complexity for capturing traffic flow characteristics, the LWR model did not yield a significant advantage in prediction performance. As a result, it may be

less applicable as the physical constraints when weighed against the trade-offs between accuracy and computational cost.

5.5.2 Choice of Weight of Physics

In the computation of the combined loss, we adopted an α (4.17) parameter, which served as a control factor governing the influence of the physics-informed loss. This parameter can be abstracted as an importance score that defines the balance between predictive accuracy and adherence to physical principles. To explore the impact of this weight parameter, we conducted a series of experiments involving variations of its value to 0.3, 0.5, and 0.7. These values were chosen to represent scenarios where varying degrees of emphasis were placed on aligning the predictions with the underlying physics. For instance, a value of 0.3 may reflect a case where relatively less attention is given to physics alignment while 0.7 indicates a greater emphasis on adhering to the physical constraints in the prediction results. To control variables and ensure consistency, we limited our experiments to the PeMS dataset. This dataset was selected due to its superior data completeness compared to the Melbourne dataset. Aside from the three error metrics, we also calculated the physics cost, excluding the weight parameter. The best results are shown in bold in 5.6.

Model	α	MAE	RMSE	MAPE	PHY
GMAN	0.3	19.93	33.23	15.81%	753.62
	0.5	**14.05**	**28.39**	**10.06%**	**121.41**
	0.7	16.98	31.14	13.49%	411.26
DDGCRN	0.3	17.93	31.23	13.92%	581.35
	0.5	**13.29**	**26.93**	**9.21%**	**97.62**
	0.7	16.98	28.14	12.51%	316.33

TABLE 5.6: Performance with Different Weights of Physics-Informed Loss (**bold**: best result)

As demonstrated in the table, a weight proportion of 50% consistently outperforms other parameter values in terms of both accuracy and physics cost. When we increase the α parameter from 0.3 to 0.5, it is clear that the physics cost is being minimised, and the error is reduced. This implies that giving more emphasis to the alignment with physics yields better predictive accuracy when the weight of physics is low.

However, the pattern does not hold when the value progresses from 0.5 to 0.7. In this case, it is intriguing to note that both accuracy and physics costs show an increase. This observation suggests that there might be a specific balance between accuracy and physics adherence that leads to optimal results, and increasing the weight of physics adherence beyond that balance could lead to diminishing returns or even a trade-off between the two.

5.6 Performance on Scarse Training Samples

Despite the growing availability of traffic data facilitated by the development of sensing technologies, acquiring complete and long-term datasets remains a challenge. Consequently, assessing the robustness of traffic prediction models under insufficient data conditions has become a primary focus in many transportation studies. A successful example in this context is the application of physics-informed deep learning (PIDL). The approach has demonstrated its effectiveness in utilising limited input data and incorporating physical principles for traffic state estimation. In this thesis, we assess the robustness of both PI-GMAN and PI-DDGCRN models under the fitting curve loss constraints with limited training samples. In the following subsections, we will first introduce the experimental settings and present the overall performance. Then, we will conduct an in-depth analysis of the trade-offs among accuracy, sample size, and efficiency. By comparing these outcomes across the two datasets, our goal is to offer a comprehensive analysis of the extended models' performances facing limited training samples.

5.6.1 Overall Result

In the experiment, we restricted the training data size ranging from 10% to 30% of its original quantity. These percentages aim to represent a spectrum from extremely scarce to slightly insufficient data. We select the average RMSE from the next 12 time steps as the accuracy measurement. Simultaneously, we record the convergence time for each prediction, measured in seconds. The extended models are compared with their original versions to emphasise the performance differences, with the better results highlighted in bold. Each experiment was repeated five times and the mean values were presented.

The training samples are randomised in each iteration to eliminate order bias.

Model	Dataset	Sample Size	Criteria	
			RMSE	Convergence Time
GMAN	PeMS	10%	43.63	174
		20%	34.69	343
		30%	32.62	548
	Melb	10%	92.92	150
		20%	64.50	306
		30%	**38.94**	348
PI-GMAN	PeMS	10%	**33.89**	**69**
		20%	**31.03**	**125**
		30%	**29.95**	**119**
	Melb	10%	**62.47**	**65**
		20%	**45.29**	**101**
		30%	40.88	**90**

TABLE 5.7: Performance of PI-GMAN on Scarce Training Samples

Model	Dataset	Sample Size	Criteria	
			RMSE	Convergence Time
DDGCRN	PeMS	10%	58.94	196
		20%	47.23	372
		30%	**35.06**	553
	Melb	10%	92.23	164
		20%	59.33	342
		30%	49.41	401
PI-DDGCRN	PeMS	10%	**42.34**	**71**
		20%	**39.15**	**142**
		30%	36.12	**136**
	Melb	10%	**57.64**	**67**
		20%	**43.16**	**98**
		30%	**39.28**	**92**

TABLE 5.8: Performance of PI-DDGCRN on Scarce Training Samples

As shown in Tables 5.7 and 5.8, the physics-informed (PI) versions of the models consistently exhibit superior accuracy when compared to the original models. Notably, PI models showcase robust performance even when trained with only 10% and 20% of the data, demonstrating their effectiveness in scenarios where the training sample is extremely limited. However, we do notice a slight decline in performance when using 30% of the training data for the PI models. We hypothesise that the influence of physics becomes more pronounced when the data proportion is below this threshold and may be less apparent with a larger training size. In terms of convergence time, the PI models

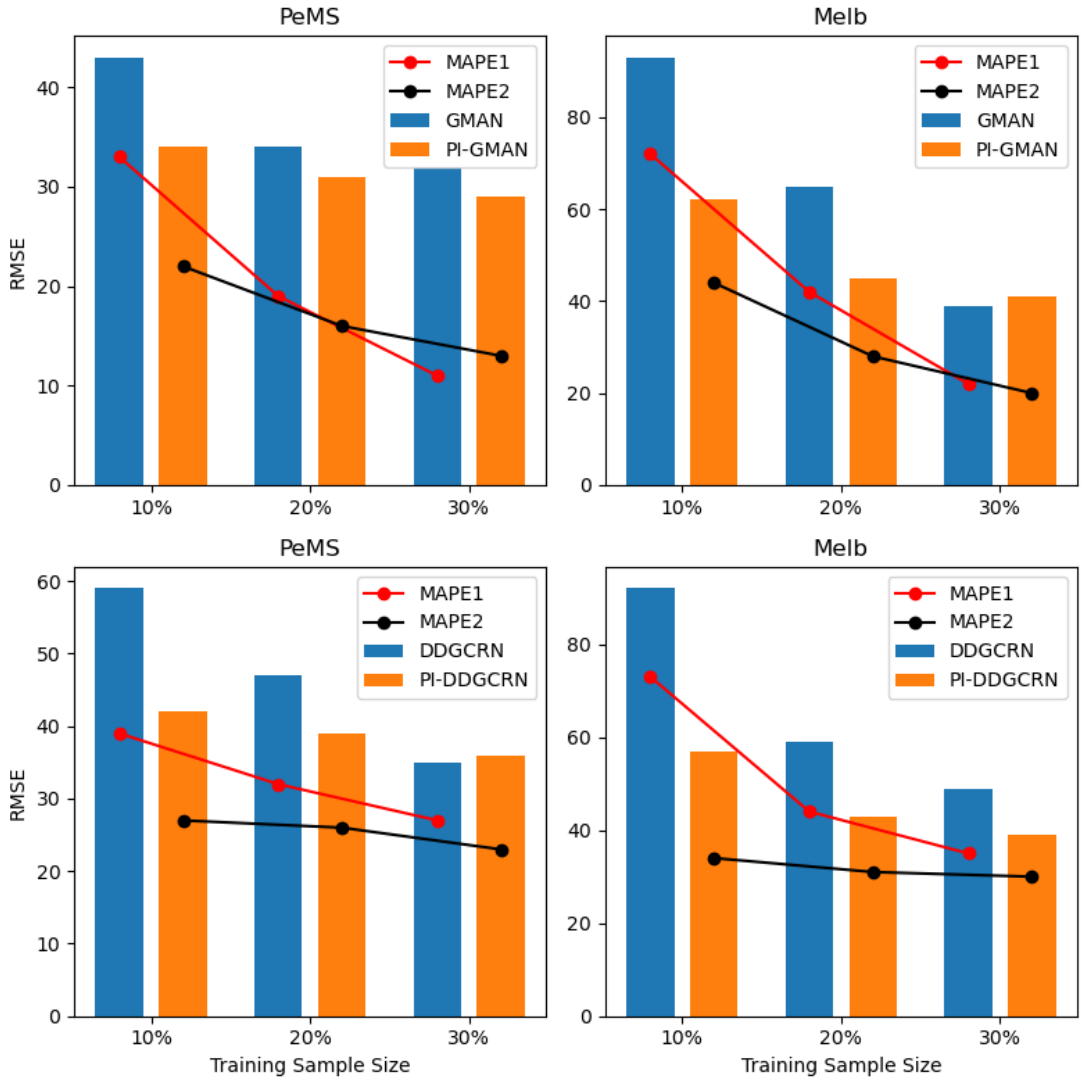


FIGURE 5.2: Error vs Sample Size for PI-models

consistently outperform their uninformed counterparts across all sample sizes. Typically, this results in a significant cost reduction of 2 to 3 times. Such efficiency gains are particularly advantageous for real-world large-scale traffic prediction tasks where operational efficiency is a primary concern.

5.6.2 Accuracy-size Trade-off

To further analyse the error trends, we visualised the statistics from the tables for the two types of extended models and the two datasets in Figure 5.2. We included additional MAPE line plots to provide an intuitive view of the trend of performances.

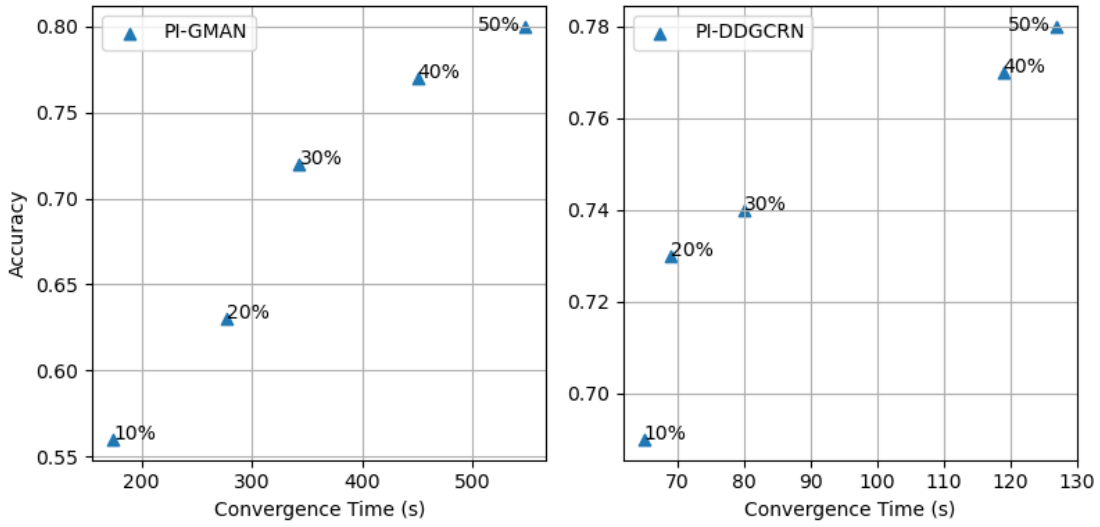


FIGURE 5.3: Accuracy-Time Trade-off

As illustrated in the plots, across both datasets and models, the physics-uninformed models consistently display a steep slope as training sample sizes decrease. This indicates that without utilising the physics or correlations between traffic variables, the models are unstable with respect to data sizes. In contrast, physics-informed models exhibit stable error changes throughout the entire range of limited data. While uninformed models may slightly outperform in high-data scenarios, they show less robustness in low-data conditions. This underscores the potential of leveraging physics between traffic variables for traffic prediction under scarce data conditions.

5.6.3 Accuracy-efficiency Trade-off

The accuracy-efficiency trade-off is a central topic in PIDL. Leveraging prior experimental findings with limited training data, we conduct a comparative analysis of the extended models, taking into account both their accuracy and convergence time under insufficient data conditions. To facilitate this comparison, we have generated an additional plot that aligns time with error metrics, as depicted in Figure 5.3. Note that the convergence time is quantified in seconds, while performance is computed as 1 minus MAPE, which reflects an intuitive view of accuracy.

From the figure, we observe a clear positive correlation between the two variables. Specifically, for PI-GMAN, there's an almost linear relationship between the model's prediction accuracy and convergence time. However, this linear trend does not apply

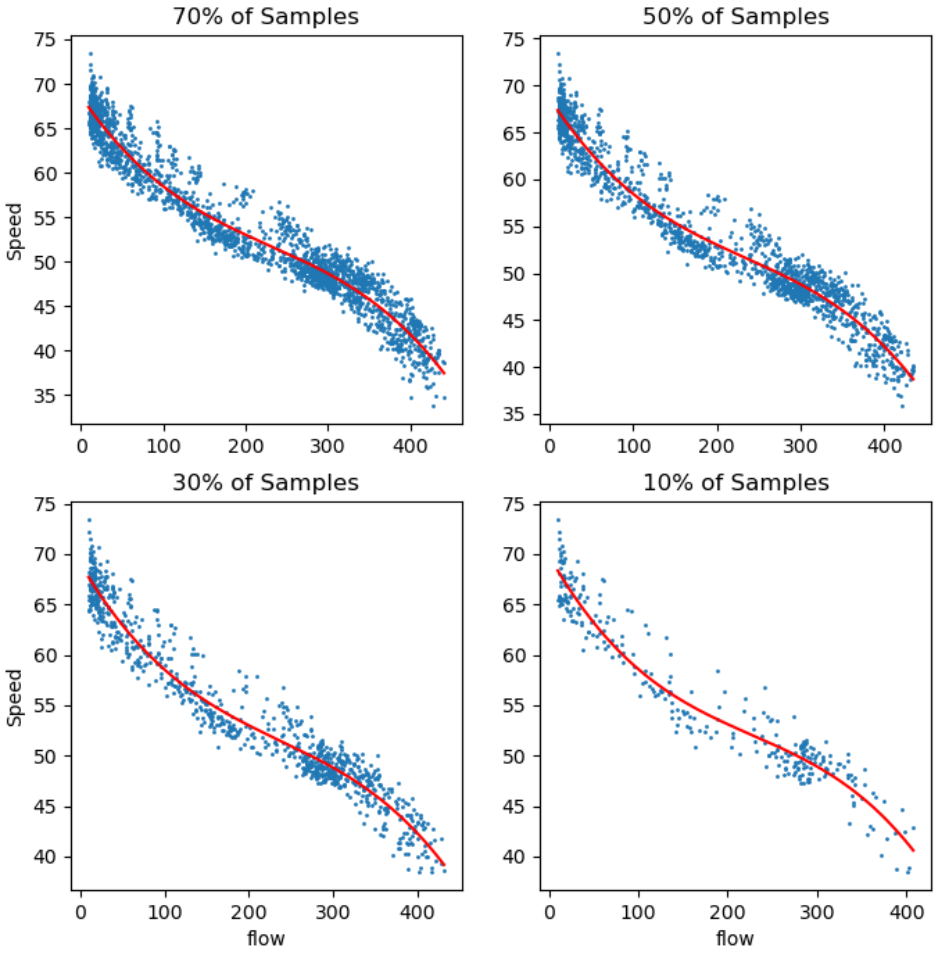


FIGURE 5.4: Distribution of Variables with Varying Sample Sizes

to PI-DDGCRN. In the second plot, we notice distinct phases in the trends. In the condition of lower sample sizes, ranging from 10% to 30%, the accuracy increases from fast to slow as the convergence time lengthens. Conversely, in the range of 30% to 50%, we observe an inverse trend. This observed behaviour can likely be attributed to two reasons. Firstly, with a smaller sample, the model might converge quickly to a suboptimal solution. However, as the sample size increases, the mode has more data to learn from, leading to an increase in global accuracy. Another factor that may be contributing to this behaviour is the interplay of traffic variables. It is possible that the relationship between traffic speed and flow remains relatively stable even as the sample size varies. To explore this hypothesis, we plotted the fitted curve depicting the distribution of speed and flow at 0.7, 0.5, 0.3 and 0.1 of the original size respectively, shown in Figure 5.4.

As demonstrated in the figures, the curves seem to maintain similar shapes when fitted at different sample sizes. This observation in turn implies that the distribution or relationship between traffic variables remains stable irrespective of changes in data size, thereby substantiating our conjecture. This conclusion is also useful for inductive bias and observational bias, as they both leverage the correlations between traffic variables to improve model performances.

5.7 Hyperparameter Analysis

In this section, we embark on a comprehensive analysis of the hyperparameters associated with the proposed physical attention and physical embedding modules. Our objective is to thoroughly investigate and comprehend how various configurations of the inductive bias can influence the model's capability to learn and capture the intricate correlations between traffic variables.

5.7.1 Number of Physical Attention Heads

The number of physical attention heads represents how many split a self-attention module is performed before calculating the importance scores of each individual embedding. Intuitively, this number-of-splits parameter decides how many different patterns or associations the model can learn. However, this is just an indication of the maximum capability. In practice, depending on the complexity of the relationships of the variables being learned, the value may not always be an exact match for the actual intricacies of the patterns. Thus, to optimise the model performance and better understand the complexity of the correlation between variables, it is necessary to find a suitable split value.

In Section 4.4, we designed a novel physical attention module to capture the correlations between traffic variables. Inspired by the Transformer module [110], we adopted a multi-headed attention structure. To investigate the optimal number of splits, we perform experiments on five different configurations of this parameter: 2, 4, 8, 16 and 32. The numbers are all selected as powers of 2 for two reasons. Firstly, most computer systems, especially GPUs, have memory configurations that work optimally with sizes that are powers of 2. This setting minimises the wasted memory during model training and results

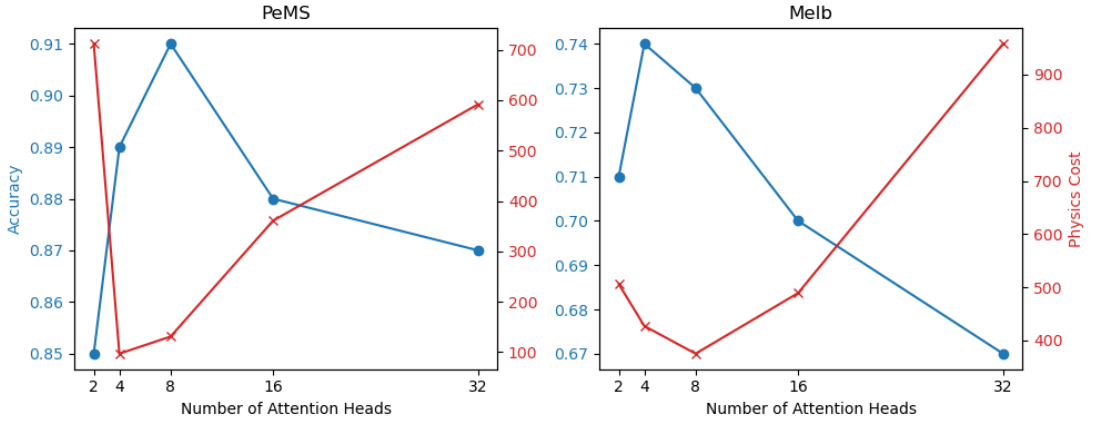


FIGURE 5.5: Accuracy and Physics Cost vs. Number of Attention Heads

in better utilisation of hardware resources. Secondly, this setting simplifies the process of hyperparameter search. By doubling and halving values, potential intermediate values are covered through a combination of numbers within this range, which decreases the searching space from $O(n)$ to $O(\log n)$. Again, we run experiments for both datasets and adopt 1 minus MAPE as an accuracy measurement. Additionally, we calculate and display the physics cost to examine the impact of parameter variation on the prediction's non-compliance with physics. The results are plotted in Figure 5.5.

We observed two distinct phases in the model performance for both datasets concerning the variation in the number of attention heads. In the initial phase, characterized by a low split number, there is a noticeable increase in prediction accuracy. The peak accuracy is achieved in both datasets, representing the optimal number of attention heads. However, beyond this peak, we observe a degradation in model performance as the number of attention heads continues to increase. Specifically, the optimal number of attention heads is 8 for the PeMS dataset and 4 for the Melbourne dataset. This discrepancy can be attributed to the differences in dataset sizes. The Melbourne dataset is approximately three times smaller than the PeMS dataset, providing fewer patterns for the attention heads to capture. Hence, a smaller number is sufficient for learning. In contrast, the PeMS dataset, with its longer time span and larger spatial distribution of nodes, requires more attention heads to adequately capture the underlying complexities.

An interesting pattern is observed from the plots of physics cost. As illustrated, the trend of the cost lines presents a negative correlation with the prediction performance. This indicates that high accuracy does not necessarily correspond to a low physics cost. In

this context, it might indicate a high loss in physics. This phenomenon can be attributed to the function of the physics loss. In traffic prediction models, it acts as a regularization term, enforcing the prediction distribution to conform to underlying physical patterns. Consequently, a higher cost may reflect a stronger penalty for deviations from these patterns, leading to improved performance.

Another unexpected behaviour is observed at the physics cost valley, which does not align with the accuracy peak. Instead, it's consistently the slightly higher cost values that correspond to the model's best performance. We attribute this phenomenon to the regularization effect of the learning bias. When the model attempts to fit noisy data, the cost may not significantly contribute to accuracy. Another possible reason is that the ground truth data itself might not strictly reflect the underlying physics, thus leading to this observation. Finally, the model's accuracy rapidly decreases when surpassing the optimal number of attention heads. It is possible that an excessive number of heads may capture irregular information, thereby diminishing the usefulness of the patterns learned.

5.7.2 Physical Embedding Dimension

As a direct indication of the complexity of vectors, embedding dimensions serve as crucial parameters in various representational learning tasks, similar to the significance of choosing the number of layers in a neural network. In the original DDGCRN implementation [119], the dynamic graph is generated through a randomly initialised spatial embedding with its size ranging from 4 to 6 depending on the number of nodes in the traffic network. Drawing inspiration from this idea, we have designed a dynamic correlation generation process, where a physical embedding is created as the core building block. To investigate the ideal dimension of the physical embedding, we conducted experiments with parameter values spanning from 2 to 15. We chose 2 as the lower limit to align with the simplest one-hot encoding dimension of two variables. For the upper limit, we selected a value equal to the square root of the traffic network's size. This choice is anticipated to encompass all possibilities when multiplied with its transpose vector to produce the correlation matrix. Considering the computational cost associated with the embedding dimension, we additionally recorded the convergence time for each

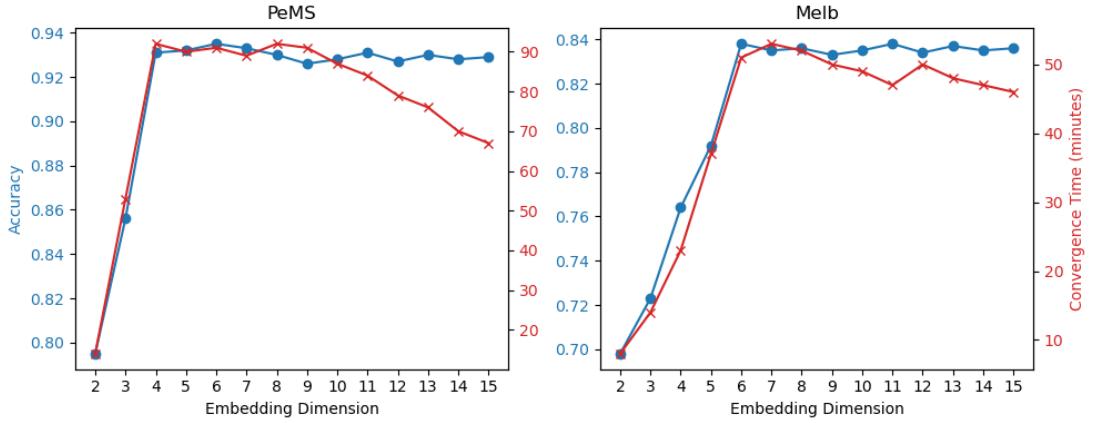


FIGURE 5.6: Accuracy vs Embedding Dimension

configuration. Similar to previous experiments, accuracy is measured as 1 minus MAPE. We visualise the results in Figure 5.6.

In terms of accuracy, we observe similar trends for both datasets: there is a sharp increment before the peak performance is achieved on a dimension size of 6. After reaching this peak, accuracies tend to remain consistent with respect to changes in dimension, with only slight fluctuations. This suggests that at lower dimensions, increasing the parameter value significantly contributes to the model's prediction results. However, once a certain threshold is passed, such a correlation tends to disappear, and the results seem to remain unaffected by the parameter.

Regarding efficiency, the results at lower dimensions exhibit very similar patterns, where the convergence time sharply increases as the embedding dimension increases. Surprisingly, when surpassing this threshold, the convergence time tends to decrease as the dimension increases. This behaviour is unexpected, as we previously hypothesised that the complexity introduced by the embedding dimension would have a more consistent and significant impact on the model's convergence.

One possible explanation for this behaviour may lie in the information gained from correlations among traffic variables. While the computational steps increase with the growth of the embedding size, it is important to note that more information about the relationships between variables is also being captured in each iteration. Consequently, this increased gain of information leads to a reduced time requirement for the model to converge to an optimal set of weights.

5.8 Discussion

We briefly summarise our findings from the experiments. Overall, the correlation-enhanced (CE) models and the physics-informed (PI) models outperform all the baseline models, including their unextended versions, in both short- and long-term scenarios. It is worth noting that the PI models may perform slightly worse than the original models, depending on the compliance of ground truth labels with the underlying physics. This highlights a trade-off between adhering to physical principles and achieving prediction accuracy, which should be carefully tailored to the specific requirements of application scenarios. Regarding long-term predictions, we observed a significant advantage with the CE models. Notably, this advantage becomes more pronounced as the number of time steps increases. We attribute this improvement to the consistent negative correlations between traffic speed and flow over the temporal dimension. This insight suggests a promising avenue for enhancing long-term traffic prediction.

In terms of the choice of learning bias, our proposed fitted curve strategy consistently achieved higher accuracy compared to the conventional Greenshield's and the LWR model. Although our method may require slightly more time to converge than Greenshield's model, the additional computational cost is negligible when weighed against the performance advantage it offers. Furthermore, we conducted an evaluation of the PI models' performance under conditions of limited training data. The results indicated that when the physics of traffic variables is incorporated, the model exhibits remarkable robustness regardless of the dataset's size. Notably, the PI models maintained a respectable level of accuracy even when trained with just 10% of the sample. We attributed this improvement to the stable correlation between traffic variables, regardless of the dataset's size, and we substantiated this hypothesis with experimental evidence. This inherent robustness provides the model with an advantage in addressing real-world traffic prediction tasks with low data availability.

To examine the impact of parameters from inductive bias on the model's performances, we varied the number of attention heads and the physical embedding dimension. Results demonstrated that there are optimal values for both parameters, where either surpassing or reducing the value will not yield the highest accuracy. Notably, while the general trend of convergence time exhibits a negative correlation with the accuracy curves, the peaks and valleys do not consistently align with the same parameter choices. Thus, it

is essential to strike a balance between accuracy and computational efficiency, carefully determining the optimal values that involve trade-offs to enhance the overall performance of the model.

In summary, our research demonstrates the robustness of correlations between traffic variables across varying spatiotemporal scales and sample sizes. In the learning bias dimension, the increased computational complexity is well-balanced with accuracy, making our strategy an efficient extension of the existing data-driven traffic prediction frameworks. For inductive bias, the optimal parameter values should be determined based on the trade-offs between accuracy and convergence to maximise the performance gain.

Chapter 6

Conclusions

6.1 Summary

In this thesis, we extended and implemented enhancements to two prominent spatiotemporal neural networks, specifically GMAN and DDGCRN, to exploit the correlations among traffic variables for traffic prediction. Inspired by the principles outlined in Physics-Informed Machine Learning (PIML) [46], we introduced two categories of biases, namely inductive and learning biases, to embed variable correlations into the models. To establish these biases, we initially expanded the models to accommodate multiple input variables. Subsequently, we devised two dedicated modules for capturing the dynamic relationships between traffic speed and flow within both GMAN and DDGCRN. For GMAN, we introduced a physical embedding alongside a physical attention module, employing multi-headed self-attention mechanisms to capture dynamic correlations effectively. On the other hand, for DDGCRN, inspired by the concept of a dynamic graph [119], we engineered a dynamic correlation generation process, ensuring adaptability in the correlation matrix over time. To integrate the learning bias into the models, we introduced a novel fitted curve strategy and incorporated it as a regularisation term during the loss constraint calculations. For comparison and benchmarking purposes, we also integrated two common traffic flow models, Greenshield and LWR, to assess the physics cost within the loss functions.

We performed extensive experiments on the PeMS and Melbourne datasets to demonstrate the effectiveness of our proposed methods. Experimental results showed that both

the Correlation-Enhanced (CE-) models with inductive bias, and the Physics-Informed (PI-) ones with learning bias outperform the baselines on time horizons exceeding 8 or 12 steps for the two datasets respectively. These findings clearly highlight the superiority of our extended models for long-term traffic predictions. To delve deeper into the value of using multiple traffic variables, we conducted an ablation study where we varied only the input-output settings while eliminating other modifications. We compared models with single input and output variables to those with both speed and flow as inputs and observed a consistent error reduction. This provides further evidence of the advantages of leveraging correlations among multiple variables.

For a comprehensive assessment of the extended models' robustness and efficiency, we conducted additional experiments on limited training samples and evaluated the models' convergence time. Notably, we observed that even when provided with only 10% to 50% of the training samples, the extended models managed to maintain their performance at a decent level of accuracy. In contrast, the unextended models exhibited a rapid increase in errors as more data was omitted. Furthermore, it is noteworthy that physics-informed models exhibit a tendency to converge at a faster rate compared to their uninformed counterparts. These findings underscore the pivotal role of correlations and the influence of physics pertaining to traffic variables. They enhance the robustness and efficiency of models, especially in scenarios where data is limited. Finally, we extended our research by conducting additional experiments to explore the impact of different types of physics to be informed, such as LWR and Greenshields models. Based on this, we made observations and conducted in-depth trade-off analyses. The results suggested that the correlation between variables is stable irrespective of the training sample size, which is a useful finding for studies with insufficient data conditions. We believe that these experiments, in conjunction with the aforementioned findings, can significantly contribute to the integration of traffic variables' correlations into existing deep learning frameworks.

6.1.1 Contributions

Our contributions can be summarised as follows. 1) We introduced a novel physical embedding and physical attention module to extend the capabilities of GMAN, enabling

the capture of dynamic correlations between traffic variables. This serves as an essential inductive bias for deep learning models. 2) Additionally, we devised a dynamic correlation generation process to adapt the physical embedding with time in the learning of DDGCRN, further enhancing the inductive bias. 3) To incorporate traffic flow physics, we developed a fitted curve strategy, which functions as a learning bias to enforce compliance with the governing laws of the traffic system. This strategy serves as a soft penalty, guiding the model towards convergence in accordance with physical principles. 4) We conducted extensive experiments on the correlation-enhanced and physics-informed models to test their accuracy, efficiency and robustness under various conditions, such as longer time horizons and insufficient data samples. Through these experiments, our work complements existing studies on utilising multiple traffic variables for traffic prediction and ultimately, advancing the integration of data-driven and model-driven approaches across various spatiotemporal problem domains.

6.1.2 Limitations

The limitations of our research come from four perspectives. First, the Melbourne dataset used in our research is disrupted and spans only 31 days. The quality, coverage, and completeness of the data sources can significantly influence the model's effectiveness. As a result, the generalisability of our findings to other datasets needs further examination. Second, we exclusively extended two specific spatiotemporal models. It is worth noting that there are many other models with different architectures for traffic prediction. This limited scope may constrain the generalisability of our model to other prediction scenarios. Third, when incorporating learning bias through the fitted curve strategy, we assume a polynomial distribution for the speed-flow-density relationship. Considering the complexity of traffic systems, this assumption may not hold true for some cases, which may affect the method's applicability. Lastly, we did not explore observational bias as a method for integrating traffic variable correlations. This is a potential avenue for further research.

6.2 Future Works

As we wrap up this thesis, we look ahead to promising avenues for future research, specifically centred on harnessing the potential of traffic variable correlations. In this section, we explore aspects that offer researchers the opportunity to enhance our proposed framework and delve into unexplored territories. We believe that by concentrating on these areas, we can maximise the potential of leveraging traffic variable correlations, further advancing the integration of data-driven and model-driven approaches.

6.2.1 Embedding Observational Biases

In this thesis, we have explored both inductive and learning biases as ways to incorporate the physics of traffic variables into deep learning models. The two types of biases focused on the model architecture and loss constraints respectively, but ignored the the information directly gained from the data sources. As one of the simplest modes of introducing physics, observational bias has great potential in transforming and augmenting the data to reflect the underlying traffic laws.

For our extended models, we have already demonstrated that by employing a combination of physical embedding and physical attention, we can significantly enhance the model's long-term prediction performance. In this context, the term "physical embedding" serves as a weak mechanism of observational bias, which imparts a structural foundation for the model to grasp the representation and characteristics of traffic variables. However, this initial approach lacks a comprehensive elucidation of the relationships involved. Drawing inspiration from periodic time embeddings and topological graph embeddings, we propose that integrating the relationships or equations governing these variables into the physical embeddings can be a valuable strategy for exploiting correlations. This assertion is grounded in the fact that neural networks excel at comprehending and contrasting informative vectors when compared to randomly generated ones.

6.2.2 Alternative Physical Loss Functions

By integrating equations derived from the Greenshield's model, the LWR model, and the curve-fitting approach to loss functions, our experiments have provided initial evidence of the benefits associated with learning bias. Nevertheless, it is essential to acknowledge that the models we have employed are relatively simple and conventional. We firmly believe that delving into more advanced physics models represents a promising avenue for research. For example, the Aw-Rascle-Zhang (ARZ) model [4] stands out as a potential choice. It can be interpreted as a generalization of the LWR model, offering a range of fundamental diagram curves, rather than being confined to a single one. This broader perspective can potentially yield deeper insights into traffic behaviour and enhance models' predictive capabilities.

Another viable direction of research lies in the extension of our fitted curve strategy. In this thesis, we made the assumption that the distribution of traffic speed and flow adheres to a polynomial shape, which simplified the implementation process. However, this assumption may not hold in cases where external factors such as traffic accidents or adverse weather conditions substantially disrupt traffic patterns, resulting in irregular behaviours. One solution to address this challenge is the adoption of piecewise functions. When a global solution for the distribution is not evident, we can break the problem down into manageable subproblems and fit each segment with a different function. By doing so, we can more accurately capture the actual relationships between variables, potentially leading to improvements in the model's performance. This approach also allows for greater flexibility and adaptability in modelling other complex real-world traffic scenarios.

6.2.3 Utilisation of More Traffic Variables

In this thesis, we primarily focused on two traffic variables: speed and flow. However, it is important to recognise that there exists a plethora of other variables affecting traffic conditions, such as density, travel time and headway. Each of these variables shares some form of correlation with speed and flow. For instance, within the fundamental relationship, traffic density exhibits a robust positive correlation with traffic flow, which can be leveraged for more effective traffic prediction tasks. By incorporating these additional

variables, we can enhance the model's comprehension of inter-variable relationships, thereby elevating its predictive capabilities.

In pursuit of this goal, one potential direction is to expand the embedding dimension to accommodate the increasing quantity of traffic variables. For example, in our research, we designed a 2-by-2 physical embedding for GMAN. However, this constitutes a relatively small correlation matrix, which may have limitations in carrying useful information. When dealing with a greater number of variables, it becomes essential to employ larger matrices to adequately represent the intricate relationships among them. This scalability in embedding dimensions can also help us capture the richness of data and enhance the model's ability to handle more extensive and diverse traffic scenarios.

Another promising approach involves integrating the key values of these variables into the physical loss constraints. For instance, the free flow speed v_{free} is a commonly utilised parameter in traffic flow modelling. By incorporating such essential parameters into our physical models, it is possible that the physics cost can better reflect the road conditions and the interplay of variables.

6.2.4 Benchmarking with More Datasets and Models

Due to time constraints, we conducted experiments on only two datasets and extended our analysis to two typical spatiotemporal networks. Expanding our benchmarking to include more datasets could provide additional results for verifying and comparing the performance of our extended models. Moreover, diversifying our model selection can enhance the reliability and generalisability of our findings. For instance, the METR-LA dataset has been widely adopted as a benchmark for traffic prediction. If we could obtain variables beyond speed for this dataset, it would allow for additional experiments to validate the feasibility of our proposed approach. When it comes to model selection, we can consider two typical categories: meta-learning-based models, where models 'learn to learn' and adapt their structure for optimal performance, and pretraining-enhanced models like STEP [98], which leverage long-term historical information as prior knowledge for the model.

Performing additional benchmarking can offer valuable insights into understanding how correlations between traffic variables can influence various models and data sources, thus further contributing to the knowledge in this field.

Bibliography

[1] S.-I. Amari. Learning patterns and pattern sequences by self-organizing nets of threshold elements. *IEEE Transactions on Computers*, C-21(11):1197–1206, 1972. doi: 10.1109/T-C.1972.223477.

[2] Sheng-hai An, Byung-Hyug Lee, and Dong-Ryeol Shin. A survey of intelligent transportation systems. In *2011 Third International Conference on Computational Intelligence, Communication Systems and Networks*, pages 332–337, 2011. doi: 10.1109/CICSyN.2011.76.

[3] Austroads. Traffic management training, 2023. URL https://austroads.com.au/network-operations/network-management/guide-to-traffic-management.

[4] A Aw and M Rascle. On a theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 456(1996):437–467, 2000.

[5] Donald J. Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, page 359–370. AAAI Press, 1994. doi: 10.5555/3000850.3000887.

[6] S. M. A. Burney and A. A. Al-Ghamdi. Short-term prediction of traffic volume in urban arterials. *Transportation Research Record*, 1320:47–54, 1991.

[7] Ling Cai, Krzysztof Janowicz, Gengchen Mai, Bo Yan, and Rui Zhu. Traffic transformer: Capturing the continuity and periodicity of time series for traffic forecasting. *Transactions in GIS*, 24(3):736–755, 2020. doi: 10.1111/tgis.12644.

[8] Pinlong Cai, Yunpeng Wang, Guangquan Lu, Peng Chen, Chuan Ding, and Jianping Sun. A spatiotemporal correlative k-nearest neighbor model for short-term traffic multistep forecasting. *Transportation Research Part C: Emerging Technologies*, 62:21–34, Jan 2016. doi: 10.1016/j.trc.2015.11.002.

[9] Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, volume 3, page 2205–2211, Aug 2021. doi: 10.24963/ijcai.2021/304.

[10] Kit Yan Chan, Tharam S. Dillon, Jaipal Singh, and Elizabeth Chang. Neural-network-based models for short-term traffic flow forecasting using a hybrid exponential smoothing and levenberg–marquardt algorithm. *IEEE Transactions on Intelligent Transportation Systems*, 13(2):644–654, 2012. doi: 10.1109/TITS.2011.2174051.

[11] Robert E Chandler, Robert Herman, and Elliott W Montroll. Traffic dynamics: studies in car following. *Operations research*, 6(2):165–184, 1958.

[12] Zhao Chen, Yang Liu, and Hao Sun. Physics-informed learning of governing equations from scarce data. *Nature Communications*, 12(11):6136, 2021. doi: 10.1038/s41467-021-26434-1.

[13] Rinaldo M Colombo and Mauro Garavello. General kinetic models for vehicular traffic flows and monte-carlo methods. *Communications in Applied and Industrial Mathematics*, 1(1):1–16, 2005.

[14] Yuliang Cong, Jianwei Wang, and Xiaolei Li. Traffic Flow Forecasting by a Least Squares Support Vector Machine with a Fruit Fly Optimization Algorithm. *Procedia Engineering*, 137:59–68, 2016. doi: 10.1016/j.proeng.2016.01.234.

[15] Carlos F. Daganzo. *Author Index*, page 328–329. Emerald Group Publishing Limited, Jan 1997. doi: 10.1108/9780585475301-010.

[16] Victoria Dahmen, Allister Loder, Gabriel Tilg, Alexander Kutsch, and Klaus Bogenberger. Traffic state estimation with loss constraint. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, page 1907–1912, Oct 2022. doi: 10.1109/ITSC55140.2022.9921815.

[17] Jose M. del Castillo. Three new models for the flow–density relationship: derivation and testing for freeway and urban data. *Transportmetrica*, 8(6):443–465, Nov 2012. doi: 10.1080/18128602.2011.556680.

[18] Xuan Di, Rongye Shi, Zhaobin Mo, and Yongjie Fu. Physics-informed deep learning for traffic state estimation: A survey and the outlook. *Algorithms*, 16(66):305, 2023. doi: 10.3390/a16060305.

[19] Xuchen Dong, Ting Lei, Shangtai Jin, and Zhongsheng Hou. Short-term traffic flow prediction based on xgboost. In *2018 IEEE 7th Data Driven Control and Learning Systems Conference (DDCLS)*, page 854–859, May 2018. doi: 10.1109/DDCLS.2018.8516114.

[20] Jennifer Drake, Joseph Schofer, and ADJR May. A statistical analysis of speed-density hypotheses. *Traffic Flow and Transportation*, 1965.

[21] Leslie C. Edie. Car-following and steady-state theory for noncongested traffic. *Operations Research*, 9(1):66–76, 1961. doi: 10.1287/opre.9.1.66.

[22] Shen Fang, Xianbing Pan, Shiming Xiang, and Chunhong Pan. Meta-msnet: Meta-learning based multi-source data fusion for traffic flow prediction. *IEEE Signal Processing Letters*, 28:6–10, 2021. doi: 10.1109/LSP.2020.3037527.

[23] Shen Fang, Chunxia Zhang, Shiming Xiang, and Chunhong Pan. Automsnet: Multi-source spatio-temporal network via automatic neural architecture search for traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, 24(3):2827–2841, 2023. doi: 10.1109/TITS.2022.3225553.

[24] Bin Feng, Jianmin Xu, Yonggang Zhang, and Yongjie Lin. Multi-step traffic speed prediction based on ensemble learning on an urban road network. *Applied Sciences*, 11(1010):4423, Jan 2021. doi: 10.3390/app11104423.

[25] Rui Fu, Zuo Zhang, and Li Li. Using lstm and gru neural network methods for traffic flow prediction. In *2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, page 324–328, 2016. doi: 10.1109/YAC.2016.7804912.

[26] PG Gipps. A behavioural car-following model for computer simulation. *Transportation Research Part B: Methodological*, 15(2):105–111, 1981.

[27] Somdatta Goswami, Katiana Kontolati, Michael D. Shields, and George Em Karniadakis. Deep transfer operator learning for partial differential equations under conditional shift. *Nature Machine Intelligence*, 4(1212):1155–1164, Dec 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00569-2.

[28] Harold Greenberg. An analysis of traffic flow. *Operations research*, 7(1):79–85, 1959.

[29] B. D. Greenshields. A study of traffic capacity. In *Proceedings of the highway research board*, volume 14, pages 448–477, 1935.

[30] Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.3301922.

[31] Shengnan Guo, Youfang Lin, Huaiyu Wan, Xiucheng Li, and Gao Cong. Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 34(11):5415–5428, 2022. doi: 10.1109/TKDE.2021.3056502.

[32] Boumediene Hamzi and Houman Owhadi. Learning dynamical systems from data: A simple cross-validation perspective, part i: Parametric kernel flows. *Physica D: Nonlinear Phenomena*, 421:132817, 2021. ISSN 0167-2789. doi: 10.1016/j.physd.2020.132817.

[33] Dirk Helbing, Ansgar Hennecke, Vladimir Shvetsov, and Martin Treiber. Micro- and macro-simulation of freeway traffic. *Mathematical and computer modelling*, 35(5-6):517–547, 2002.

[34] Peter Hidas. Modelling lane changing and merging in microscopic traffic simulation. *Transportation Research Part C: Emerging Technologies*, 10(5-6):351–371, 2002.

[35] Yanrong Hu, Chong Wu, and Hongjiu Liu. Prediction of passenger flow on the highway based on the least square support vector machine. *Transport*, 26(2): 197–203, 2011. doi: 10.3846/16484142.2011.593121.

[36] Archie J. Huang and Shaurya Agarwal. Physics informed deep learning for traffic state estimation: Illustrations with lwr and ctm models. *IEEE Open Journal of Intelligent Transportation Systems*, 3:1–1, Jan 2022. doi: 10.1109/OJITS.2022.3182925.

[37] Wenhao Huang, Guojie Song, Haikun Hong, and Kunqing Xie. Deep architecture for traffic flow prediction: Deep belief networks with multitask learning. *IEEE Transactions on Intelligent Transportation Systems*, 15(5):2191–2201, 2014. doi: 10.1109/TITS.2014.2311123.

[38] Ameya D. Jagtap, Kenji Kawaguchi, and George Em Karniadakis. Locally adaptive activation functions with slope recovery for deep and physics-informed neural networks. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476(2239):20200334, Jul 2020. doi: 10.1098/rspa.2020.0334.

[39] Jiahao Ji, Jingyuan Wang, Zhe Jiang, Jiawei Jiang, and Hu Zhang. Stden: Towards physics-guided neural networks for traffic flow prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(44):4048–4056, Jun 2022. doi: 10.1609/aaai.v36i4.20322.

[40] Yuxuan Ji and Nikolas Geroliminis. On the spatial partitioning of urban transportation networks. *Transportation Research Part B: Methodological*, 46(10):1639–1656, 2012. doi: 10.1016/j.trb.2012.08.005.

[41] Jiawei Jiang, Chengkai Han, Wayne Xin Zhao, and Jingyuan Wang. Pdformer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction. In *AAAI*. AAAI Press, 2023. doi: 10.48550/arXiv.2301.07945.

[42] Weiwei Jiang and Jiayun Luo. Graph neural network for traffic forecasting: A survey. *Expert Systems with Applications*, 207:117921, 2022. doi: 10.1016/j.eswa.2022.117921.

[43] Feng Jin and Shiliang Sun. Neural network multitask learning for traffic flow forecasting. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1897–1901, 2008. doi: 10.1109/IJCNN.2008.4634057.

[44] Jeon-Seong Kang, JinKyu Kang, Jung-Jun Kim, Kwang-Woo Jeon, Hyun-Joon Chung, and Byung-Hoon Park. Neural architecture search survey: A computer vision perspective. *Sensors*, 23(33):1713, Jan 2023. doi: 10.3390/s23031713.

[45] Ameya D. Jagtap Karniadakis and George Em. Extended physics-informed neural networks (xpinns): A generalized space-time domain decomposition based deep learning framework for nonlinear partial differential equations. *Communications in Computational Physics*, 28(5):2002–2041, Jun 2020. ISSN 1815-2406, 1991-7120. doi: 10.4208/cicp.OA-2020-0164.

[46] George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, Jun 2021. doi: 10.1038/s42254-021-00314-5.

[47] Ali Kashefi, Davis Rempe, and Leonidas J. Guibas. A point-cloud deep learning framework for prediction of fluid flow fields on irregular geometries. *Physics of Fluids*, 33(2):027104, Feb 2021. doi: 10.1063/5.0033376.

[48] Femke Kessels. *Traffic Flow Modelling: Introduction to Traffic Flow Theory Through a Genealogy of Models*, volume 8 of *EURO Advanced Tutorials on Operational Research*. Springer, 2019. doi: 10.1007/978-3-319-78695-7.

[49] Ehsan Kharazmi, Zhongqiang Zhang, and George E.M. Karniadakis. hp-vpinns: Variational physics-informed neural networks with domain decomposition. *Computer Methods in Applied Mechanics and Engineering*, 374:113547, Feb 2021. doi: 10.1016/j.cma.2020.113547.

[50] M. Koshi. Some findings and an overview on vehicular flow characteristics. In *Proc. 8th Intl. Symp. on Transp. and Traffic Theory*, 1983.

[51] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

[52] Jason D. Lee, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *29th Annual Conference on Learning Theory*, volume 49, pages 1246–1257. PMLR, 2016. doi: 10.48550/arXiv.1605.00405.

[53] Vinc Lee. Pytorch implementation of gman. https://github.com/VincLee8188/GMAN-PyTorch, 2013.

[54] Yaguang Li. Implementation of diffusion convolutional recurrent neural network in tensorflow. https://github.com/liyaguang/DCRNN, 2018.

[55] Yaguang Li. The traffic data files for los angeles (metr-la) and the bay area (pems-bay). https://drive.google.com/drive/folders/10FOTa6HXPqX8Pf5WRoRwcFnW9BrNZEIX, 2018.

[56] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations (ICLR '18)*, 2018. doi: 10.48550/arXiv.1707.01926.

[57] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Multipole graph neural operator for parametric partial differential equations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. doi: 10.5555/3495724.3496291.

[58] M J Lighthill and G B Whitham. 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):317–345, 1955.

[59] Liu, ZiqiCai, Wei, John Xu, and Zhi-Qin. Multi-scale deep neural network (mscalednn) for solving poisson-boltzmann equation in complex domains. *Communications in Computational Physics*, 28(5):1970–2001, 2020. doi: 10.4208/cicp.OA-2020-0179.

[60] Xingjian Liu, Zhiqiang Wang, Yixiao Li, and Zhenya Liu. A hybrid model of arima and kalman filter for short-term power load forecasting. *Journal of Applied Mathematics*, 2015, 2015. doi: 10.1155/2015/657371.

[61] Xuefeng Liu, Hongke Zhang, Zhiyong Wang, and Yajuan Zhang. Integrating granger causality and vector auto-regression for traffic prediction of large-scale wlans. *KSII Transactions on Internet and Information Systems*, 10(1):314–331, 2016. doi: 10.3837/tiis.2016.01.016.

[62] Lu Lu, Ming Dao, Punit Kumar, Upadrasta Ramamurty, George Em Karniadakis, and Subra Suresh. Extraction of mechanical properties of materials through deep learning from instrumented indentation. *Proceedings of the National Academy of Sciences*, 117(13):7052–7062, Mar 2020. doi: 10.1073/pnas.1922210117.

[63] Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(33):218–229, Mar 2021. doi: 10.1038/s42256-021-00302-5.

[64] Zheng Lu, Chen Zhou, Jing Wu, Hao Jiang, and Songyue Cui and. Integrating granger causality and vector auto-regression for traffic prediction of large-scale wlans. *KSII Transactions on Internet and Information Systems*, 10(1):136–151, 2016. doi: 10.3837/tiis.2016.01.008.

[65] Yisheng Lv, Yanjie Duan, Wenwen Kang, Zhengxi Li, and Fei-Yue Wang. Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873, 2014. doi: 10.1109/TITS.2014.2345663.

[66] Yasuji Makigami, G. F. Newell, and Richard Rothery. Three-dimensional representation of traffic flow. *Transportation Science*, 5(3):302–313, Aug 1971. ISSN 0041-1655. doi: 10.1287/trsc.5.3.302.

[67] Vittorio Marzano and Francesco Viti. *Microscopic and Mesoscopic Traffic Models*, pages 97–131. Springer, 2018.

[68] Adolf Darlington May. *Traffic Flow Fundamentals*. Prentice Hall, 1990.

[69] Lyudmila Mihaylova and René Boel. A particle filter for freeway traffic estimation. In *2004 43rd IEEE Conference on Decision and Control (CDC)(IEEE Cat. No. 04CH37601)*, volume 2, pages 2106–2111. IEEE, 2004.

[70] Sparsh Mittal and Vibhu. A survey of accelerator architectures for 3d convolution neural networks. *J. Syst. Archit.*, 115(C), 2021. doi: 10.1016/j.sysarc.2021.102041.

[71] Juan Esteban Muriel, Lele Zhang, Jiadong Mao, Tingjin Chu, and Timothy Garoni. A short note for partitioning the road network of melbourne's eastern suburbs. Technical report, School of Mathematics and Statistics, the University of Melbourne, 2021.

[72] Kai Nagel and Michael Schreckenberg. A cellular automaton model for freeway traffic. *Journal de physique I*, 2(12):2221–2229, 1992.

[73] Attila M. Nagy and Vilmos Simon. Survey on traffic prediction in smart cities. *Pervasive and Mobile Computing*, 50:148–163, 2018. doi: 10.1016/j.pmcj.2018.07.004.

[74] Attila M. Nagy and Vilmos Simon. Survey on traffic prediction in smart cities. *Pervasive and Mobile Computing*, 50:148–163, 2018. doi: https://doi.org/10.1016/j.pmcj.2018.07.004.

[75] G. F. Newell. Nonlinear effects in the dynamics of car following. *Operations Research*, 9(2):209–229, Apr 1961. ISSN 0030-364X. doi: 10.1287/opre.9.2.209.

[76] Daiheng Ni. Determining traffic-flow characteristics by definition for application in its. *IEEE Transactions on Intelligent Transportation Systems*, 8(2):181–187, 2007.

[77] Daiheng Ni. *Chapter 2 - Traffic Flow Characteristics I*, page 19–35. Butterworth-Heinemann, 2016. doi: https://doi.org/10.1016/B978-0-12-804134-5.00002-7.

[78] Daiheng Ni. *Chapter 4 - Equilibrium Traffic Flow Models*, page 51–71. Butterworth-Heinemann, Jan 2016. doi: 10.1016/B978-0-12-804134-5.00004-0.

[79] Daiheng Ni, John D Leonard, Chaoqun Jia, and Jianqiang Wang. Vehicle longitudinal control and traffic stream modeling. *Transportation Science*, 50(3):1016–1031, 2016.

[80] Seri Oh, Stephen G. Ritchie, and Cheol Oh. Real-time traffic measurement from single loop inductive signatures. *Transportation Research Record*, 1804(1):98–106, Jan 2002. doi: 10.3141/1804-14.

[81] Tiago Prado Oliveira, Jamil Salem Barbar, and Alexsandro Santos Soares. Multilayer perceptron and stacked autoencoder for internet traffic prediction. In *Network and Parallel Computing: 11th IFIP WG 10.3 International Conference, NPC*

2014, Ilan, Taiwan, September 18-20, 2014. Proceedings 11, pages 61–71. Springer, 2014.

[82] Houman Owhadi. Bayesian numerical homogenization. *Multiscale Modeling & Simulation*, 13(3):812–828, Jan 2015. doi: 10.1137/140974596.

[83] Houman Owhadi. Multigrid with rough coefficients and multiresolution operator decomposition from hierarchical information games. *SIAM Review*, 59(1):99–149, Jan 2017. doi: 10.1137/15M1013894.

[84] Michael L. Pack, Brian L. Smith, and William T. Scherer. Automated camera repositioning technique for video image vehicle detection systems: Integrating with freeway closed-circuit television systems. *Transportation Research Record*, 1856(1):25–33, Jan 2003. doi: 10.3141/1856-04.

[85] Vijay Paidi, Hasan Fleyeh, Johan Håkansson, and Roger G Nyberg. Smart parking sensors, technologies and applications for open parking lots: a review. *IET Intelligent Transport Systems*, 12(8):735–741, 2018.

[86] Zheyi Pan, Yuxuan Liang, Weifeng Wang, Yong Yu, Yu Zheng, and Junbo Zhang. Urban traffic prediction from spatio-temporal data using deep meta learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 1720–1730, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 978-1-4503-6201-6. doi: 10.1145/3292500.3330884. event-place: Anchorage, AK, USA.

[87] Guofei Pang, Lu Lu, and George Em Karniadakis. fpinns: Fractional physics-informed neural networks. *SIAM Journal on Scientific Computing*, 41(4): A2603–A2626, Jan 2019. doi: 10.1137/18M1229845.

[88] A. Pascale and M. Nicoli. Adaptive bayesian network for traffic flow prediction. In *2011 IEEE Statistical Signal Processing Workshop (SSP)*, pages 177–180, 2011. doi: 10.1109/SSP.2011.5967651.

[89] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and S. S. Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM Comput. Surv.*, 51(5), sep 2018. doi: 10.1145/3234150.

[90] Ilya Prigogine and Robert Herman. *Kinetic theory of vehicular traffic*. Elsevier, 1971. doi: 10.1016/B978-0-444-10057-9.

[91] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5301–5310. PMLR, 09–15 Jun 2019. doi: 10.48550/arXiv.1806.08734.

[92] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations. *ArXiv*, abs/1711.10561, 2017. doi: 10.48550/arxiv.2307.12008.

[93] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Numerical gaussian processes for time-dependent and nonlinear partial differential equations. *SIAM Journal on Scientific Computing*, 40(1):A172–A198, Jan 2018. doi: 10.1137/17M1120762.

[94] Lukáš Rapant. Traffic speed prediction using ensemble kalman filter and differential evolution. *MATEC Web of Conferences*, 259:02001, Jan 2019. doi: 10.1051/matecconf/201925902001.

[95] Jingtao Rong, Xinyi Yu, Mingyang Zhang, and Linlin Ou. Across-task neural architecture search via meta learning. *International Journal of Machine Learning and Cybernetics*, 14(3):1003–1019, 2023.

[96] Mohammadreza Saeedmanesh and Nikolas Geroliminis. Clustering of heterogeneous networks with directional flows based on "snake" similarities. *Transportation Research Part B: Methodological*, 91:250–269, 2016. doi: 10.1016/j.trb.2016.05.008.

[97] Toru Seo, Alexandre M. Bayen, Takahiko Kusakabe, and Yasuo Asakura. Traffic state estimation on highway: A comprehensive survey. *Annual Reviews in Control*, 43:128–151, Jan 2017. doi: 10.1016/j.arcontrol.2017.03.005.

[98] Zezhi Shao, Zhao Zhang, Fei Wang, and Yongjun Xu. Pre-training enhanced spatial-temporal graph neural network for multivariate time series forecasting. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 1567–1577, New York, NY, USA, 2022. Association for

Computing Machinery. doi: 10.1145/3534678.3539396. URL https://doi.org/10.1145/3534678.3539396.

[99] Justin Sirignano, Jonathan F. MacArt, and Jonathan B. Freund. Dpm: A deep learning pde augmentation method with application to large-eddy simulation. *Journal of Computational Physics*, 423:109811, 2020. doi: 10.1016/j.jcp.2020.109811.

[100] Anupam Srivastava and Nikolas Geroliminis. Empirical observations of capacity drop in freeway merges with ramp control and integration in a first-order model. *Transportation Research Part C: Emerging Technologies*, 30:161–177, 2013. doi: 10.1016/j.trc.2013.02.006.

[101] Shiliang Sun, Changshui Zhang, and Guoqiang Yu. A bayesian network approach to traffic flow forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 7(1):124–132, 2006. doi: 10.1109/TITS.2006.869623.

[102] Chris M.J. Tampere and L. H. Immers. An extended kalman filter application for traffic state estimation using ctm with implicit mode switching and dynamic parameters. In *2007 IEEE Intelligent Transportation Systems Conference*, pages 209–216, 2007. doi: 10.1109/ITSC.2007.4357755.

[103] David Alexander Tedjopurnomo, Zhifeng Bao, Baihua Zheng, Farhana Murtaza Choudhury, and A. K. Qin. A survey on modern deep neural network for traffic prediction: Trends, methods and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 34(4):1544–1561, Apr 2022. doi: 10.1109/TKDE.2020.3001195.

[104] David Alexander Tedjopurnomo, Zhifeng Bao, Baihua Zheng, Farhana Murtaza Choudhury, and A. K. Qin. A Survey on Modern Deep Neural Network for Traffic Prediction: Trends, Methods and Challenges. *IEEE Transactions on Knowledge and Data Engineering*, 34(4):1544–1561, April 2022. doi: 10.1109/TKDE.2020.3001195.

[105] Yongxue Tian and Li Pan. Predicting short-term traffic flow by long short-term memory recurrent neural network. In *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, page 153–158, 2015. doi: 10.1109/SmartCity.2015.63.

[106] Martin Treiber and Dirk Helbing. Reconstructing the spatio-temporal traffic dynamics from stationary detector data. *Cooper@ tive Tr@ nsport@ tion Dyn@ mics*, 1(3):3–1, 2002.

[107] TSVD. Traffic signal volume data, 2023. URL https://discover.data.vic.gov.au/dataset/traffic-signal-volume-data.

[108] Muhammad Usama, Rui Ma, Jason Hart, and Mikaela Wojcik. Physics-informed neural networks (pinns)-based traffic state estimation: An application to traffic network. *Algorithms*, 15(1212):447, 2022. doi: 10.3390/a15120447.

[109] Stylianos I Vagropoulos, GI Chouliaras, Evaggelos G Kardakos, Christos K Simoglou, and Anastasios G Bakirtzis. Comparison of sarimax, sarima, modified sarima and ann-based models for short-term pv generation forecasting. In *2016 IEEE international energy conference (ENERGYCON)*, pages 1–6. IEEE, 2016.

[110] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. doi: 10.5555/3295222.3295349.

[111] VDEP. Bluetooth travel data, 2023. URL https://data-exchange.vicroads.vic.gov.au/.

[112] Eleni Vlahogianni, Matthew Karlaftis, and John Golias. Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C: Emerging Technologies*, 43, 2014. doi: 10.1016/j.trc.2014.01.005.

[113] Haizhong Wang, Daiheng Ni, Qian-Yong Chen, and Jia Li. Stochastic modeling of the equilibrium speed–density relationship, 2013.

[114] Jingyuan Wang, Jiawei Jiang, Wenjun Jiang, Chao Li, and Wayne Xin Zhao. Libcity: An open library for traffic prediction. In *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '21, page 145–148, New York, NY, USA, 2021. Association for Computing Machinery. doi: 10.1145/3474717.3483923.

[115] Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why pinns fail to train: A neural tangent kernel perspective, 2020.

[116] Wujie Wang and Rafael Gómez-Bombarelli. Coarse-graining auto-encoders for molecular dynamics. *npj Computational Materials*, 5(11):1–9, Dec 2019. doi: 10.1038/s41524-019-0261-5.

[117] Y. Wang, M. Papageorgiou, and A. Messmer. An adaptive freeway traffic state estimator and its real-data testing part 1: basic properties. In *Proceedings. 2005 IEEE Intelligent Transportation Systems, 2005.*, pages 531–536, 2005. doi: 10.1109/ITSC.2005.1520104.

[118] Wenchao Weng. Pytorch implementation of ddgcrn. https://github.com/wengwenchao123/DDGCRN, 2023.

[119] Wenchao Weng, Jin Fan, Huifeng Wu, Yujie Hu, Hao Tian, Fu Zhu, and Jia Wu. A decomposition dynamic graph convolutional recurrent network for traffic forecasting. *Pattern Recognition*, 142:109670, 2023. doi: 10.1016/j.patcog.2023.109670.

[120] Billy M Williams and Lester A Hoel. Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results. *Journal of transportation engineering*, 129(6):664–672, 2003.

[121] Zonghan Wu. Pytorch implementation of graph wavenet. https://github.com/nnzhan/Graph-WaveNet, 2019.

[122] Zonghan Wu. Pytorch implementation of mtgnn. https://github.com/nnzhan/MTGNN, 2022.

[123] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, page 1907–1913. AAAI Press, 2019. doi: 10.5555/3367243.3367303.

[124] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on*

Knowledge Discovery & Data Mining, pages 2978–2988, 2020. doi: 10.48550/arXiv.2005.11650.

[125] Dawen Xia, Huaqing Li, Binfeng Wang, Yantao Li, and Zili Zhang. A map reduce-based nearest neighbor approach for big-data-driven traffic flow prediction. *IEEE Access*, 4:2920–2934, 2016. doi: 10.1109/ACCESS.2016.2570021.

[126] Jianhua Xiao, Zhu Xiao, Dong Wang, Jing Bai, Vincent Havyarimana, and Fanzi Zeng. Short-term traffic volume prediction by ensemble learning in concept drifting environments. *Knowledge-Based Systems*, 164:213–225, Jan 2019. doi: 10.1016/j.knosys.2018.10.037.

[127] Yan Xing, Wenqing Li, Weidong Liu, Yachao Li, and Zhe Zhang. A dynamic regional partitioning method for active traffic control. *Sustainability*, 14(1616): 9802, Jan 2022. doi: 10.3390/su14169802.

[128] John Xu, Zhi-QinZhang, YaoyuLuo, TaoXiao, Yanyang, Ma, and Zheng. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, 28(5):1746–1767, 2020. doi: 10.4208/cicp.OA-2020-0085.

[129] Fei Yan, Man Zhang, and Zhongke Shi. Dynamic partitioning of urban traffic network sub-regions with spatiotemporal evolution of traffic flow. *Nonlinear Dynamics*, 105(1):911–929, Jul 2021. doi: 10.1007/s11071-021-06448-6.

[130] Mofeng Yang, Jiaohong Xie, Peipei Mao, Chao Wang, and Zhirui Ye. Application of the arimax model on forecasting freeway traffic flow. *Journal of Advanced Transportation*, 2018:1–10, 2018. doi: 10.1155/2018/4019569.

[131] Yibo Yang, Mohamed Aziz Bhouri, and Paris Perdikaris. Bayesian differential programming for robust systems identification under uncertainty. *Proceedings of the Royal Society A*, 476(2243):20200290, 2020. doi: 10.1098/rspa.2020.0290.

[132] Huaxiu Yao, Fei Wu, Jintao Ke, Xianfeng Tang, Yitian Jia, Siyu Lu, Pinghua Gong, Jieping Ye, Didi Chuxing, and Zhenhui Li. Deep multi-view spatial-temporal network for taxi demand prediction. In *Proceedings of the Thirty-Second AAAI*

Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. doi: 10.5555/3504035.3504351.

[133] Haoteng Yin. Tensorflow implementation of stgcn. https://github.com/VeritasYin/STGCN_IJCAI-18, 2023.

[134] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 3634–3640. AAAI Press, 2018. ISBN 978-0-9992411-2-7.

[135] Haitao Yuan and Guoliang Li. A survey of traffic prediction: from spatio-temporal data to intelligent transportation. *Data Science and Engineering*, 6(1):63–85, Mar 2021. doi: 10.1007/s41019-020-00151-z.

[136] Yufei Yuan, J. W. C. van Lint, R. Eddie Wilson, Femke van Wageningen-Kessels, and Serge P. Hoogendoorn. Real-time lagrangian traffic state estimator for freeways. *IEEE Transactions on Intelligent Transportation Systems*, 13(1):59–70, 2012. doi: 10.1109/TITS.2011.2178837.

[137] Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 1655–1661. AAAI Press, 2017. doi: 10.5555/3298239.3298479.

[138] Yunlong Zhang, Larry E Owen, and James E Clark. Multiregime approach for microscopic traffic simulation. *Transportation Research Record*, 1644(1):103–114, 1998.

[139] Yuxin Zhang, Jian Wang, Qian Zhang, and Xiang Li. Short-term traffic flow forecasting based on varma model. *Journal of Advanced Transportation*, 2017. doi: 10.1155/2017/3035469.

[140] Mingming Zhao, Xianghua Yu, Yonghui Hu, Jingnan Cao, Simon Hu, Lihui Zhang, Jingqiu Guo, and Yibing Wang. Real-time freeway traffic state estimation with fixed and mobile sensing data. In *2020 IEEE 23rd International*

Conference on Intelligent Transportation Systems (ITSC), pages 1–7, 2020. doi: 10.1109/ITSC45102.2020.9294327.

[141] Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. Gman: A graph multi-attention network for traffic prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(0101):1234–1241, 2020. doi: 10.1609/aaai.v34i01.5477.

[142] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020. doi: 10.1016/j.aiopen.2021.01.001.