TAG-F: A TRAFFIC PREDICTIVE ANALYTICS GUIDANCE FRAMEWORK

A thesis submitted to

The University of Manchester for the degree of Doctor of

Philosophy in the Faculty of Humanities

2020

Aniekan Essien

Management Sciences & Marketing Division

Alliance Manchester Business School

List of Contents

Word Count: 48,327

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
ANN	Artificial Neural Networks
API	Application programming Interface
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
ASP	Algorithm Selection Problem
ATCS	Adaptive Traffic Control Systems
ATIS	Advanced Traveller Information Systems
ATMS	Advanced Traffic Management System
AUC	Area under ROC Curve
BN	Bayesian Networks
BPNN	Backpropagation Neural Network
CART	Classification and Regression Trees
CNN	Convolutional Neural Network
DBN	Deep Belief Network
DC	Data Context
DCM	Data Collection Method
DGS	Decision Guidance System
DL	Deep Learning
DSR	Design Science Research
DynaMIT	Dynamic Network Assignment for the Management of Information to Travelers
DynaSMART-X	Dynamic Network Assignment-Simulation Model for Advanced Roadway Telematics
FCD	Floating Car Data
FFNN	Feedforward Neural Network
FURIA	Fuzzy unordered rule induction algorithm

GPS	Global Positioning System
HTML	Hypertext Markup Language
IBL	Instance-based Learning
ICT	Information and Communication Technologies
ILD	Inductive Loop Device
IREP	Incremental Reduced Error Pruning
ISR	Information Systems Research
IT	Information Technology
ITS	Intelligent Transportation Systems
KF	Kalman Filter
k-NN	k- Nearest Neighbours
LBD	Literature-based Discovery
LDA	Linear Discriminant Analysis
LR	Linear Regression
LSTM	Long Short-term Memory
MAC	Media Access Control
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MATSIM	Multi-Agent Transport simulation toolkit
ML	Machine Learning
MNIST	Modified National Institute of Standards and Technology
NARXNN	Nonlinear Autoregressive Model with Exogenous input Neural Network
NFL	No free lunch principle/theorem
NLP	Natural Language Processing
PAM	Predictive Analytical Method
PRESIMM	Prediction Simulation-based Traffic Management Modell
RADAR	Radio Detection and Ranging
RBF	Radial Basis Function

RIPPER	Repeated Incremental Prunning to Produce Error Reduction
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
ROC	Receiver Operator Characteristics
RW	Random Walk
SAE	Stacked Autoencoder
SARIMA	Seasonal ARIMA
SGD	Stochastic Gradient Descent
SUMO	Simulation of Urban Mobility
SVM	Support Vector Machine
SVR	Support Vector Regression
TAG-F	Traffic Predictive Analytics Guidance Framework
TCC	Traffic Control Centre
TPA	Traffic Predictive Analytics
VAR	Vector Auto Regression

Abstract

The provision of accurate and timely information to traffic analysts and road users are critical components for the successful implementation of intelligent transportation systems (ITSs) around the world. This is typically achieved via the application of predictive analytics on historical data to make forecasts about traffic parameters. However, given the broad spectrum of data sources, data collection methods and traffic predictive models at the disposal of traffic data scientists, making accurate predictions becomes challenging for some reasons. Firstly, the complexity of the traffic domain makes traffic predictive analytics (TPA) problem description complicated. Secondly, the plethora of available predictive models makes the choice of which model to be applied in each TPA scenario difficult. Thirdly, there is not yet a predictive method that works well over time and in all scenarios (Joyce and Herrmann, 2018; Zhang et al., 2019). Due to these limitations, there is a need for the provision of guidance to traffic data scientists performing data-driven traffic prediction.

Traffic Predictive Analytics Guidance Framework (TAG-F) is a guidance framework that aims at bridging this gap. The framework delineates data-driven traffic prediction as a set of three dimensions, thereby providing a structured collection of analytical decision points that can serve as a roadmap to enable the traffic data scientist traverse from the traffic problem space to the analytical solution space, culminating in an action/outcome, usually prediction. TAG-F – via the tool – can also be used to provide decision support for traffic data scientists by providing guidance in the choice of predictive analytical method (PAM), given the data context specifications. The framework and tool were evaluated using real-world traffic prediction scenarios in an urban arterial in Greater Manchester, United Kingdom.

The contributions made through the study include a novel end-to-end guidance mechanism for TPA using a framework that fosters a structured definition of the TPA solution development process. In addition, the identification of a set of key dimensions and parameters that influence TPA. A prototype support tool is also presented, which complements the framework by providing semi-automated guidance by suggesting alternative predictive models to given TPA scenarios. The framework and tool can foster productivity in the TPA process by encouraging adaptability, reuse, and shared domain knowledge about TPA. Results from empirical analysis support the value of the proposed framework and support tool towards the provision of guidance to traffic data scientists in

TPA, however with some limitations. Finally, in this thesis, suggestions about furthering the study, addressing the identified limitations, and refining the framework and tool are articulated.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

Acknowledgements

My first and foremost thanks go to my supervisors, Dr Ilias Petrounias and Dr Pedro Sampaio for their guidance and support, which made this thesis possible. Over the last four years of this research study, they have consistently ensured that I stayed on track, improved my academic writing and research thinking by contributing their time, constructive feedback and reading through conference and journal papers. I highly commend their desire and push for high-quality research.

Special thanks also go to Prof Nikolay Mehandjiev and Dr Babis Theodoulidis, whose insightful feedback and contributions during the annual reviews have contributed to improving this research work. The content, quality, and extent of the research are a result of their insight and encouragement. I would also like to thank my boss, Dr Cinzia Giannetti, for her patience, understanding, and kindness towards developing my early research career. I also wish to give thanks to Dr Sandra Sampaio, who gave up her time to contribute towards the project and for providing the traffic datasets used within this research study.

I wish to especially thank my lovely, patient, supportive, and understanding wife, Chidinma, for her encouragement, prayer, support, proof-reading, continued support and enduring my conversations about the PhD regularly for many years. I also give thanks to my friend, Godwin Chukwukelu, for his contributions, prayers and support and continually believing in me.

Finally, my greatest thanks are given to my family. Their love, help, understanding and support through these years is far beyond any words and something for which I am eternally grateful.

List of Associated Publications

1. Essien, A., Petrounias, I., Sampaio, P., Sampaio, S. **Deep-PRESIMM: Integrating Deep Learning with Microsimulation for Traffic Prediction**. 2019 48th IEEE International Conference on Systems, Man and Cybernetics (SMC), Bari, Italy. Accepted/In Press.
2. Essien, A., Petrounias, I., Sampaio, P., Sampaio, S. **A Deep-Learning Model for Urban Traffic Flow Prediction with Traffic Events Mined from Twitter**. (2019) World Wide Web Journal. Springer Verlag. Under Review.
3. Essien, A., Giannetti, C., **A Deep Learning Framework for Univariate Time Series Prediction using Convolutional LSTM Stacked Autoencoders**. 2019 8th IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA), Sofia, Bulgaria. IEEE Xplore | Pages 1-6. | Best paper award.
4. Essien, A., Petrounias, I., Sampaio, P., Sampaio, S. **Improving Urban Traffic Speed Prediction Using Data Source Fusion and Deep Learning**. 2019 6th IEEE International Conference on Big Data and Smart Computing (BigComp), Kyoto, Japan. IEEE Xplore. | Pages 1-8. | Selected as one of best papers to be extended for journal submission.
5. Essien, A., Petrounias, I., Sampaio, P., Sampaio, S. **The Impact of Rainfall and Temperature on Peak and Off-Peak Urban Traffic**. 2018 29th International Conference on Database and Expert Systems Applications (DEXA), Regensburg, Germany. Proceedings in Lecture Notes in Computer Science (vol. 11030), Springer Verlag. | Pages 399-407.

Chapter 1 Introduction

1.1 Motivation and Problem Statement

Reducing traffic congestion is an essential priority for cities of the world and has, therefore, received significant research interest over the past decades (Vlahogianni et al., 2014). Consequently, a considerable amount of services that aim to minimise the negative consequences of increased traffic such as air pollution, traffic congestion, and noise (Falcocchio and Levinson, 2015) have been developed in the past few years, and this number will continue to grow in line with advances in information and communication technologies (ICT). Traffic congestion has a negative economic impact, resulting from increased fuel consumption, decreased productivity and increased cost of infrastructure use (Barth and Boriboonsomsin, 2008; Levy et al., 2010).

In this context, research in Intelligent Transportation Systems (ITSs) has emerged, aided with advanced traffic prediction techniques, based on historical and real-time traffic data capable of providing reliable information to road network users and traffic analysts. Current traffic management and control practices are dominated by the application of ITSs, resulting from technological advancement, data ubiquity and the multitude of statistical and machine learning predictive algorithms. ITSs can be defined as systems that make use of advanced technology and telecommunication concepts to develop and improve transportation systems (Dimitrakopoulos and Demestichas, 2010). A critical component for the success of ITSs across the world is the provision of accurate and reliable information to road users, businesses, and traffic authorities. This is realised by the application of *predictive analytics*, mainly using historical traffic data collected from one or more traffic data collection sensors. Predictive analytics is a subset of data analytics that involves the use of statistical, data-mining, and machine learning techniques to find patterns in data to make predictions about unknown future events. The application of predictive analytics in the domain of traffic management is one that has yielded tremendous benefit over the past decades. However, achieving accurate data-driven traffic prediction is difficult for at least three key reasons.

First, traffic management/control is a 'wicked problem' (Churchman, 1967) – one where both the solution and the means (of achieving it) are unknown, ambiguous, uncertain, or where a change in attributes causes a change in the entire problem understanding/description. Traffic flow involves interactions between individual agents (i.e. road users), infrastructure such as traffic controls and other transport media, affected by dynamically-changing traffic variables (such as traffic density, occupancy, and intensity) and exogenous factors such as calendar (time of day, day of week, etc.), weather (rainfall, temperature, etc.), events, road works, and accidents. For this reason, the field of traffic predictive analytics (TPA) is much more challenging. A typical challenge encountered in TPA involves defining the predictive analytical problem space, which is a prerequisite to the development of a corresponding analytical solution capable of solving the problem. For instance, what are the key factors that influence the TPA problem description space? How can a traffic data scientist select/develop an appropriate methodological approach towards performing TPA? Before developing an effective TPA solution to a traffic prediction problem, these questions need to be answered.

Secondly, there is a plethora of traffic predictive algorithms/models in use today, reflecting the highly interdisciplinary nature of the subject with contributions from engineering, computer science, mathematics, and operations management. Table 1-1 presents a list of some machine learning (ML) predictive algorithms with each column representing a distinct family of predictive algorithms, grouped by their respective similarities in terms of their functioning/method of operation. As can be seen, it may be difficult for a data scientist to possess expert knowledge about each and every one of the algorithms in Table 1-1. To a *traffic data scientist*, the different *Predictive Analytical Methods* (PAMs), (such as those presented in Table 1-1) should provide a broad portfolio of tools and techniques that can be employed towards solving a given traffic prediction problem. However, the choice regarding which PAM or analytical approach to adopt in a given traffic predictive analytics scenario is one that is highly complex and prone to uncertainties (Vlahogianni et al., 2014). Therefore, it is common to find traffic data scientists choosing PAMs either because they have limited understanding about alternatives, or do not understand the underlying assumptions for the selected algorithms or predictive methods. In reality, there is a limited number of studies that provide meta-knowledge about traffic prediction methods/algorithms, their advantages and disadvantages, or in what particular traffic prediction problem scenario(s) a given PAM is most appropriate. This shortage can be rationalised by such studies (Barros et al., 2015; Ermagun and Levinson, 2018; Lana et al., 2018; Vlahogianni et al., 2014, 2004; Xiaofeng Wang et al., 2009) quickly becoming superseded due to the rapidly evolving research area of TPA.

Thirdly, the no free lunch (NFL) principle states that averaged across all predictive (optimisation) problems, all algorithms perform equally well (Wolpert and Macready, 1997). In other words, there is no single best predictive algorithm that can be used in all situations. This has been experimentally (Joyce and Herrmann, 2018; Wolpert and Macready, 1997) and theoretically established (Brazdil, 2003; Goodfellow and

Bengio, 2015). Therefore, it is often the case that a predictive algorithm or PAM that tends to perform very well under a given set of conditions or scenarios may likely perform poorly in other scenarios or conditions.

Table 1-1: A Sample of Machine Learning Prediction Algorithms

Deep Learning	**Ensemble**	**Artificial Neural Networks (ANN)**	**Regression**	**Bayesian**	**Decision Trees**	**Instance-Based**	**Rule System**
Deep Boltzmann Machine (DBM)	Random Forest (RF)	Radial Basis Function Neural Network (RBF-NN)	Linear Regression (LR)	Naïve Bayes	Classification and Regression Trees (CART)	*k*-Nearest Neighbour (*k*-NN)	Cubist
Convolutional Neural Networks (CNN)	Gradient Boosting Machine (GBM)	Multi-Layer Perceptron (MLP)	Ordinary Least Squares (OLS) Regression	Average One-Dependence Estimators (AODE)	C 4.5	Learning Vector Quantization (LVQ)	One Rule (OneR)
Stacked Auto Encoder (SAE)	Boosting	Backpropagation Neural Network (BPNN)	Stepwise Regression	Bayesian Belief Network (BBN)	C 5.0	Self-Organizing Map (SOM)	Zero Rule (ZeroR)
Recurrent Neural Network (RNN)	Bootstrap Aggregation (Bagging)	Wavelet Transform Neural Network	Multivariate Adaptive Regression Splines (MARS)	Gaussian Naïve Bayes	Chi-Square Automatic Interaction Detection (CHAID)	Locally-Weighted Learning (LWL)	Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
Gated Recurrent Units (GRU)	AdaBoost		Locally Estimated Scatterplot Smoothing (LOESS)	Bayesian Network	Conditional Decision Trees		
Long Short-Term Memory (LSTM)	Stacked Generalization (Blending)		Logistic Regression	Multinomial Naïve Bayes			
	Gradient Boosted Regression Trees (GBRT)						

From the foregoing, given the challenges encountered in (traffic) predictive analytics, there is a need for the provision of guidance to traffic data scientists performing TPA. While there are existing traffic systems and approaches that provide guidance to TPA end-users (i.e. road users/travellers) in the form of route guidance (Liang and Wakahara, 2014; Mahmassani, 2001) and traffic state information using Advanced Traveller Information Systems (ATIS) – Google Maps, Waze, and Garmin, etc., there is a need for systems or approaches that provide guidance to traffic data scientists in the quest for the development of analytical solutions to the variety of traffic prediction problems they face.

This research, therefore, proposes TAG-F, – A Traffic Predictive Analytics Guidance Framework – which intends to provide guidance to traffic data scientists in the execution of TPA. The goal is the provision of directional guidance via an organized, analytical, TPA problem-definition structure, and the recommendation of a potential choice of predictive models (PAMs) for the given traffic prediction scenario. In this research study, there is an argument for the use of predictive algorithm meta-modelling and knowledge representation approaches to describe and define the TPA problem space and consequently provide guidance in the process of developing analytical solutions capable of solving traffic prediction problems. In particular, the study aims to:

i. Develop a guidance framework structure that supports traffic data scientists towards developing a suitable analytical solution to a given traffic prediction problem.

ii. Enable the critical analysis of each PAM to identify the advantages, disadvantages, assumptions and generalisations, which will lead to the most appropriate PAM choice in a given TPA scenario.

iii. Develop a meta-level (i.e. information about a given TPA scenario and individual PAM) knowledge base about traffic predictive analytical methods.

iv. Develop a meta-learning method for providing alternative PAM(s) for given TPA scenarios

The novelty of the approach described in this thesis lies in the provision of systematic guidance to traffic data scientists in the TPA process via the structured characterization of the analytical problem space and the subsequent decision support provided via a suggestion of alternative(s) for the choice of the predictive model(s). The adopted approach towards model suggestion is based on a combination of predictive model meta-knowledge extraction derived from a literature-based discovery process (Ruch, 2010), and a subsequent instance-based learning (and rule induction) inference methodology.

In this study, two terms are defined – the traffic analyst and traffic data scientist, respectively. These terms are distinctively used within this thesis and should not be confused one for the other. Although both job functions work together to achieve a common goal (traffic management and control), their approaches, skills, and technical prerequisites are diverse. Within the existing literature, there is not a clear distinction about who the primary stakeholder(s) is(are) in TPA. It can be argued that the primary beneficiaries of TPA are the road users since the end goal of a TPA task is reducing, controlling, or mitigating traffic congestion. On the other hand, traffic analysts and control personnel also benefit from TPA by receiving accurate and timely traffic forecasts, which will enable better decision-making.

Table 1-2 presents a characterisation of the three (3) main stakeholders in TPA – The *traffic data scientist*, *traffic analyst*, and *road (end) user*. The traffic analyst is mainly involved in the traffic system planning and engineering process and is directly responsible for traffic network control and management (Taylor and Bonsall, 2017). A typical example involves traffic control centre (TCC) analysts and operatives. In congested situations, traffic analysts can manually, automatically or semi-automatically alter the traffic parameters, such as traffic light signals, lane closure/opening, and speed limit alteration. The traffic analyst role, therefore, requires expertise in the traffic domain. This role differs from the traffic data scientist, who is a data science and predictive analytics domain expert. Thus, the traffic analyst may (or may not) possess data science and predictive analytics domain knowledge, for instance, about traffic predictive models and data-driven analytics.

Table 1-2: Characterisation of the different stakeholders in TPA

Categories	**Traffic Data Scientist**	**Traffic Analyst**	**Road User**
Primary Role/Duties	• Development and application of predictive analytical models on traffic data. • Using historical and real-time traffic data to extract patterns that can be used to infer future traffic conditions and/or describe current traffic network.	• Traffic system planning and engineering process design. • Traffic network control and management (for instance, TCC personnel). • Manually or automatically alter traffic signals, lane closure/opening, speed limit alteration (e.g. in smart motorways)	• Safely and efficiently utilise the traffic network to traverse from an origin to a destination.
Main Stakeholder	• Traffic Analyst • Road User • ITSs	• Road User	• Self • Other road users • Environment
Job/Role Prerequisite(s)	• Expertise in predictive analytics, data science, machine learning, and basic understanding of traffic flow modelling. • Understanding the application of quantitative and qualitative methods of forecasting	• Expertise in the traffic design, control, and engineering domain.	• Appropriate navigation skills (licenses, certifications, etc.)
Guidance Needs	• Development of TPA solution	• Traffic Network Simulation	• Route Guidance

For this study, the main stakeholders are the traffic data scientists (shaded column in Table 1-2). Therefore, the framework proposed in this study aims to provide guidance to the traffic data scientists in developing a suitable TPA solution capable of solving a traffic prediction problem. However, the proposed framework and methodology can be extended to provide of guidance to traffic analysts (traffic control personnel), but is outside the scope of this study and has been highlighted for future research in Section 8.7.2.

1.2 Research Question Statement

As previously stated, cumulative research contributions in traffic data science and Intelligent Transportation Systems (ITSs) have resulted in significant innovations and techniques for traffic control and management (Barros et al., 2015). According to Vlahogianni et al. (2004), traffic prediction can be defined as the process of estimating the anticipated traffic conditions given historical and present traffic conditions. Traffic prediction is a critical component of ITSs in use around the world today. Due to geographic, economic, and environmental constraints, the increase in the use of ITSs for traffic management and control has become further highlighted due to the inability of government and town/city planning authorities to continually construct new roads or expand existing ones in a bid to increase the traffic capacity to improve traffic flow. It, therefore, goes without reasoning that the integration of better predictive analytical techniques in ITSs will result in better traffic management systems. In the United Kingdom, local authorities like Transport for London have recognised the need for short term prediction of traffic parameters and the integration of such to existing ITSs in order to develop improved functionalities that allow the systems to automatically adjust traffic signals based on short-term forecasts of traffic conditions (Goves et al., 2016).

Section 1.1 summarised the challenges traffic data scientists face in TPA as (i) the complexity of the traffic management and control domain, (ii) the multitude of traffic predictive algorithms, and (iii) the absence of a single best algorithm that performs optimally in all prediction scenarios. Based on these challenges, the research study presented in this thesis is an attempt at addressing these challenges. This precipitated some research questions. The primary research question explored in this study is:

Can a predictive analytics guidance framework be designed to facilitate traffic data scientists in exploring the analytical decision space of TPA tasks?

To adequately address this primary research question, some sub-research questions have been generated, which are:

1. *What are the key (critical) dimensions of data-driven traffic prediction problems?*
2. *What are the analytical decision parameters within each key traffic analytical problem dimension required to explore the decision space of TPA tasks?*
3. *Given a set of analytical parameters and a decision space, can guidance be provided regarding alternate prediction modelling techniques/algorithms?*

The answers to the research sub-questions will together provide an answer to the primary research question. In the context of TPA, given the challenges as summarised in Section 1.1, it is important to have a grounded understanding of the key factors that affect TPA. Some studies have presented efforts at identifying key parameters affecting traffic prediction (Vlahogianni et al., 2004), spatiotemporal traffic prediction (Ermagun and Levinson, 2018), and predictive algorithms for real-time short-term traffic prediction (Barros et al., 2015). It, therefore, goes to show that the identification of the key parameters impacting TPA will enable the provision of a guidance mechanism, which will be beneficial to traffic data scientists. Within this thesis, Chapter 4 presents detailed discussions about the primary and sub-research questions.

1.3 Research Aims and Objectives

The aim of this study is the proposal and validation of a framework that provides directional guidance to traffic data scientists via an organised, analytical problem definition structure and recommending a choice of suitable predictive models for the given TPA problem scenario. In order to achieve this overall aim, three objectives have been stipulated, which are developed to provide answers to the set of research sub-questions (see Section 1.2). Table 1-3 presents a summary of the research objectives, as well as the corresponding mappings to the respective chapters within this thesis.

RO1: The first objective of this research seeks to investigate the *key* dimensions that describe the TPA problem space. This is important given that TPA involves the use of algorithms or models trained on input data – typically in the form of time-series – for parameter prediction. The prior identification of these dimensions, therefore, becomes an essential requirement for the development of a structure that is capable of providing analytical guidance.

RO2: Within the identified (key) dimensions, it is essential to investigate and identify the analytical decision parameters, which contribute to holistically describing the analytical dimension space. This will aid the traffic data scientists to arrive at quicker and more logical conclusions in the analytical development process.

Table 1-3: Mapping of research objectives to thesis chapters

Research Objectives		Chapters
RO1	To investigate the key dimensions that describe the TPA problem space.	Chapter 2 – Section 2.7.1 & Chapter 5 – Section 5.2
RO2	To investigate and characterise the analytical decision parameters, which contribute to the definition of the analytical dimension space.	Chapter 2 – Section 2.3
RO3	To develop a method based on predictive model meta-learning that can infer the choice of alternate PAM(s), given a set of pre-determined analytical decision parameters in a given TPA scenario	Chapter 6 – Section 6.4 and 6.5

RO3: Review studies (Barros et al., 2015; Ermagun and Levinson, 2018; Lana et al., 2018; Vlahogianni et al., 2014) have shown that the multitude of existing data-driven predictive models each have their individual strengths and weaknesses, hypotheses on the nature of data, generalisations, as well as appropriate scenario for which the application of such a model would provide more value (Xiaofeng Wang et al., 2009). Therefore, given a set of identified analytical decision points and dimensions, it can be possible to determine which predictive model/algorithm is appropriate for the TPA task/scenario. Achieving these three objectives led to the design and development of the traffic analytics guidance framework proposed in this research (see Chapter 5).

1.4 Research Method

This research study is guided by the design science research (DSR) methodology (Hevner et al., 2004). DSR can be seen as a set of analytical techniques or perspectives that are used to perform Information Systems Research (ISR) involving the development of artefacts that aim to explain, understand, and/or improve some aspects of information systems (Hevner and Chatterjee, 2010). The choice of DSR methodology for this present study owes to the fact that this thesis aims at producing an artefact for solving/improving an organisational problem/process. DSR methodology begins with problem awareness, where initial suggestions for solving the problem are drawn from extant theories and domain knowledge before the development of an artefact (based on the suggested domain knowledge and theories) is undertaken (Baskerville et al., 2018). The next stage involves the evaluation of the proposed solution. Details about the research methodology and the implications to the actualisation of this study are presented in Chapter 4.

1.5 Research Contributions

The main contributions of this research are:

1. A guidance mechanism for TPA using a framework that fosters a structured definition of the TPA solution development process. The proposed framework offers structured guidance to traffic data scientists in the traffic analytical solution development process for executing traffic prediction (see Section 5.2).
2. The identification of a set of key dimensions that influence TPA as well as the dimension elements. Within this thesis, Chapter 5 presents details about the identified traffic analytics dimensions and elements.
3. A prototype support tool, which complements the proposed framework by providing semi-automated traffic analytics guidance and fosters productivity in the traffic data analytics process by encouraging adaptability, reuse, and shared domain knowledge about TPA. The tool is driven by a literature-driven, meta-learning, instance-based learning algorithm for traffic predictive model suggestion and is presented in Section 6.3.

4. A characterisation of the TPA stakeholders, including their respective characteristics, role(s) or functions, input and outputs, as well as their guidance needs. The clear identification and distinction of these stakeholders improve the understanding of the TPA process and the requisite solution development process. In this thesis, this characterisation is presented in Section 7.5.

1.6 Thesis Structure

This thesis is organised in eight chapters, which closely follow the design science research phases suggested in (Peffers et al., 2007) and is graphically represented in Figure 1-1. As can be seen, the stages are problem awareness and objective definition, design and development, evaluation, and conclusion. Details about the respective stages and overall research design are presented in Chapter 4.

As can be seen from Figure 1-1, the research process begun with problem awareness and objective definition, which is presented in Chapters 1-3. The problem awareness description is supported by the literature review chapter (Chapter 2), where a review of existing related studies led to the identification of research opportunities that are actualised in this study.

The second phase involves the design and development of the Information Technology (IT) artefact, which is the TPA guidance framework presented within this thesis. Within this thesis, the research strategy is presented in Chapter 4, while the developmental process is discussed in Chapters 5 and 6, corresponding to the framework and support tool, respectively. The next phase presents the evaluation of the framework, which is elucidated in Section 6.6, where the framework and tool are demonstrated using three (3) case scenarios. This is followed by a discussion of the findings in Chapter 7, including the characterisation of TPA stakeholders as well as the guidance that can be provided to the respective roles. The study is concluded in Chapter 8, where a reflection and synthesis of the research process is presented.

A concise overview of the remaining chapters in this thesis are as follows:

i. Chapter 2: Literature Review

In this chapter, a review of existing literature about TPA, the characterisation of TPA, and the challenges encountered in TPA is presented. It begins with a brief overview of data analytics and a review of existing related studies that led to the classification of TPA into the key factors or parameters.

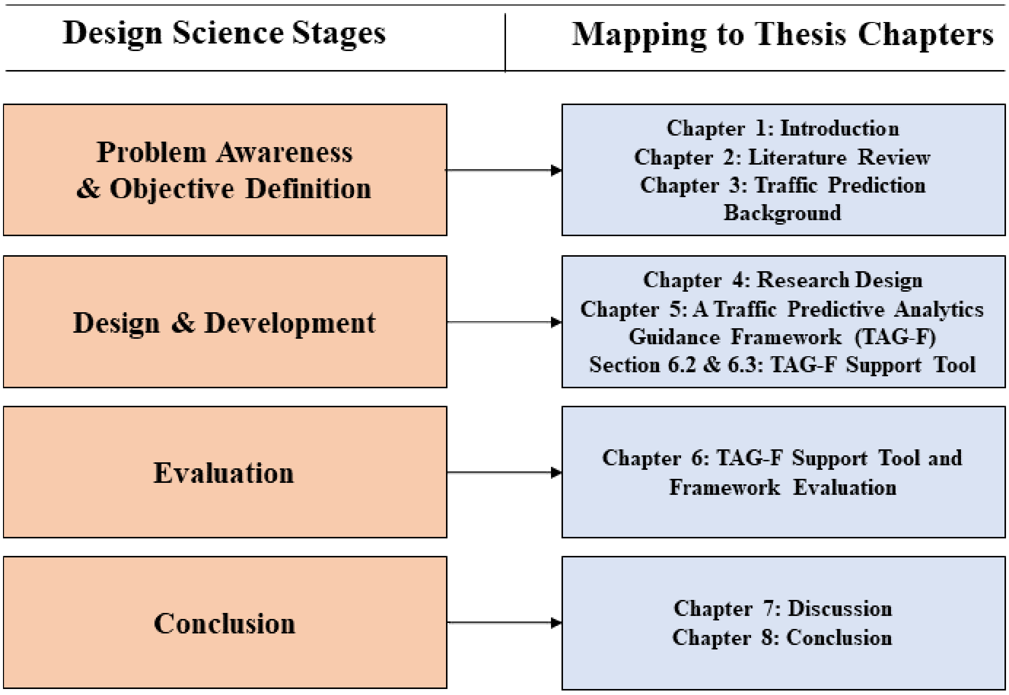

Design Science Stages
Mapping to Thesis Chapters
Problem Awareness & Objective Definition
Chapter 1: Introduction
Chapter 2: Literature Review
Chapter 3: Traffic Prediction Background
Design & Development
Chapter 4: Research Design
Chapter 5: A Traffic Predictive Analytics Guidance Framework (TAG-F)
Section 6.2 & 6.3: TAG-F Support Tool
Evaluation
Chapter 6: TAG-F Support Tool and Framework Evaluation
Conclusion
Chapter 7: Discussion
Chapter 8: Conclusion

Figure 1-1: Thesis structure

ii. Chapter 3: Traffic Prediction Background

This chapter provided technical background to traffic flow theory, traffic data collection methods, and prediction and analysed existing relevant studies that fall into the category of data-driven traffic parameter forecasting methods.

iii. Chapter 4: Research Design

Chapter 4 presented discussions about the adopted research method, research design, as well as the research strategy used in this study. It presented detailed articulations about the research choices made within this study. In this study, a deductive research approach and quantitative research methodology were adopted in the research strategy. The chapter also discussed the justification for the individual research design choices made.

iv. Chapter 5: A Traffic Predictive Analytics Guidance Framework

The preceding chapters presented identified opportunities for research, which strengthened the argument of the need for a structured traffic prediction guidance framework as a means of providing decision support for traffic analysts towards delivering better traffic congestion management and control. In Chapter 5, discussions about the proposed TAG-F framework, its dimensions, and the underlying logic are presented.

v. Chapter 6: TAG-F Support Tool and Framework Evaluation

In order to quantitatively evaluate and validate the proposed framework, a prototype (software) tool was developed. This tool, known as the TAG-F support tool, is presented in this chapter including details about the design, development, and implementation. The support tool was developed to provide semi-automated guidance via predictive model suggestion for traffic analytics. In the chapter also, the prototype tool was used to evaluate the framework using three (3) case scenarios from sensor collected data in Greater Manchester, United Kingdom.

vi. Chapter 7: Discussion

Chapter 7 presented discussions of the findings from the study. It begun with a recap of the identification of key factors that affect TPA. It proceeded further to highlight the key affecting factors, before presenting a discussion about the main stakeholders and actors in a TPA 'ecosystem'. In Section 7.5, a characterization of the main TPA stakeholders was presented, detailing the guidance requirements for each one of them.

vii. Chapter 8: Conclusions and Future Work

This chapter concluded the research study, focusing on the findings derived from the study, as well as articulating the contributions of the study, including the theoretical, methodological, and practical contributions. A reflective synthesis of the research findings with the research questions, aims and objectives are also presented in this chapter, which highlighted the proposed offering of the study.

Chapter 2 Literature Review

2.1 Introduction

There is increasing academic and industry interest in the field of *data science*, *predictive analytics*, and *big data* for traffic management and control. For instance, in the *MIS Quarterly*, a special issue on data science and predictive analytics for traffic management exists (Chen et al., 2012). Consequently, there are a plethora of studies revolving around the topical context of predictive analytics. More specifically, in the field of TPA, many studies abound that apply predictive analytical techniques towards traffic parameter forecasting. This chapter will provide a review of the extant literature on TPA, the characterisation of TPA, and the challenges encountered in TPA. A brief overview of the concept of data analytics is presented in Section 2.2. In Section 2.5, a conceptual guidance model for TPA is presented, which is a TPA-specific enhancement of the conceptual guidance model in Ceneda et al., (2017). The guidance framework presented in Ceneda et al. (2017) provided a foundational basis for the development of a guidance mechanism for TPA. The model shown in Section 2.5 accounts for identified research opportunities or challenges in TPA, input and output stages of the TPA process, and the degree/manner of guidance provided to the traffic data scientist. In Section 2.7, a review of existing related studies led to the delineation of TPA into the key factors or parameters affecting TPA, as well as the challenges encountered in TPA. The chapter is concluded in Section 2.9.

2.2 Data Analytics

The concept of analytics is one that is mainly used to define advanced computational analysis of data to infer knowledge or insight. A more formal definition of analytics can be found in (Davenport and Harris, 2017). The authors define analytics as *“the extensive use of data, statistical, or quantitative analysis, exploratory and predictive models to drive decisions and actions*”. Three main categories of analytics exist, which are:

i. Descriptive

ii. Prescriptive

iii. Predictive

Descriptive analytics uses statistical, data-mining, machine learning, or artificial intelligence techniques and algorithms to provide a descriptive context about data such as trending information, answering questions about what has happened or what is happening. More specifically, descriptive analytics tends to (through statistical, mathematical, machine learning, data mining techniques, etc.) categorise, describe, classify, and fuse data to convert it into useful information or intelligence (Evans and Lindner, 2012). Descriptive analytics is the most commonly used type of analytics.

Prescriptive analytics uses optimisation techniques, artificial intelligence, simulation, and case-based reasoning in order to identify the best alternatives towards minimising or maximising a particular objective or set of objectives. A typical application of prescriptive analytics in commercial organisations is the determination of the optimal pricing plan and advertising strategy to maximise profits. Therefore, the mathematical or statistical techniques applied in prescriptive analytics can be combined with decisions from optimisation algorithms or mechanisms to make decisions that accommodate uncertainty in the dataset (Evans and Lindner, 2012). Prescriptive analytics is mainly applied towards recommending one or more courses of action, with the likely outcome of each decision, especially in business scenarios.

Finally, predictive analytics involves the use of advanced statistical techniques, machine learning, artificial intelligence, and predictive models to provide forecasts of future events, occurrences, or values. In other words, predictive analytics is the analysis of historical occurrences in an effort to predict future happenings. It typically involves analysing historical datasets, observing patterns, and extrapolating these to future projections.

2.3 Defining guidance

The term, guidance, is generic and open to diverse understandings. Going by the dictionary definition, the term guidance refers to "advice or information aimed at resolving a problem or difficulty" (Dictionary, 2018). Another perspective relates to the field of decision support and decision guidance systems (DGS). Brodsky and Wang (2008) define DGS as systems that "*provide guidance via actionable recommendations based on prescribed analytic models or techniques*". A more generic description of the term 'guidance' can be thought of as the provision of *help* or *assistance* to a user towards a task that the user faces challenges (e.g. does not know how to use a tool or perform analytical tasks) (Brodsky and Luo, 2015).

To provide a more precise context of the concept of guidance, consider the following hypothetical illustration. Imagine a TPA guidance system supporting a traffic data scientist in the process of performing traffic parameter prediction. If the traffic data scientist is highly knowledgeable about the analytical solution development process, then the system may provide minimal guidance, for instance, in the form of highlighting details about the predictive analytical solution development process (i.e. as he traverses from the complex traffic prediction problem to an analytical solution). This could assume the form of visualisations of details about the dataset (i.e.

descriptive statistics, displays, etc.), predictive model, prediction accuracy, etc. In this case, therefore, the system provides minimal analytics guidance.

Consider another scenario where the traffic data scientist is less knowledgeable about the TPA solution development process, then the system can provide a higher level of guidance, for instance, about what type of prediction model to adopt, hyper-parameter values to utilise, etc., which can be determined on the basis of the given prediction problem. In this scenario, although the system provides a higher level of guidance, the traffic data scientist still, however, has the responsibility of applying the suggested modelling approach towards making the required predictions, as well as performing hyperparameter optimisation, train-test split, etc.

Finally, in an advanced guidance provision scenario, where the traffic data scientist has minimal knowledge about data analytics and traffic prediction models, the system provides the highest level of guidance, which can include data pre-processing, feature engineering/dimensionality reduction, prediction model/algorithm selection, model training, hyper-parameter optimisation, prediction, and visualisation of the predicted results.

The three scenarios described above give an indication of the levels of support or guidance that an analytics guidance system can provide to a traffic data scientist. This is related to a set of essential questions relating to guidance, as pointed out in Ceneda et al., (2017):

(i) *What are the human needs (what is the knowledge gap)?*

(ii) *How much guidance is to be provided to the user/analyst?*

(iii) *On what data/information is the guidance generated?*

In the next subsection, using the enhanced version of the conceptual framework presented in Ceneda et al., (2017), the above-listed questions are addressed in detail, with emphasis on TPA.

2.4 Goals of Guidance in TPA

This subsection discusses the goals of guidance in relation to TPA. In order to proceed, the broad question to answer is: *what makes a perfect analytics guidance system?* The answer to this question will depend on the particular domain under consideration, data type, and skill or analytics knowledge level of the user. According to Collins et al. (2018), the goal of an analytics guidance system is to provide knowledge about a dataset to answer questions about the analytical process accurately. In the case of TPA, the dataset and end goal are known. However, there is a challenge (i.e. knowledge gap) in selecting the methodological/predictive modelling approach towards achieving this goal (i.e. traffic parameter prediction). Therefore, in clear terms, the goals of guidance in TPA include:

i. **Information:** this is the main aim of guidance, especially in predictive analytics. A guidance system should, therefore, provide information about the development of the analytical solution to a traffic prediction problem.

ii. **Reduction of cognitive load:** with the advancement in computational efficiency and artificial intelligence come guidance systems capable of ‘learning’ from previous suggestions or guidance offerings in order to improve upon the initial state. This way, the guidance system is able to keep track of previous analyses, which can be applied to future prediction problems.

iii. **To inform:** According to Collins et al. (2018), a guidance system should be able to grow the user's knowledge of an unknown dataset. In the case of TPA, a guidance system should be able to inform the user (i.e. traffic data analyst) about improving the user's knowledge of the underlying dataset.

2.5 Conceptualising Guidance in TPA

Ceneda et al. (2017) summarised the three main characteristics of guidance as (i) the knowledge gap, (ii) input and output, and (iii) the degree of guidance. Adapting the guidance framework in Ceneda et al. (2017), this research study presented an enhanced traffic predictive analytics guidance model, which is depicted in Figure 2-1, showing the guidance offered by the framework (proposed in this present study). Guidance can be characterised as a function of the knowledge gap and input (i.e. dataset, analyst knowledge level, etc.) to provide an output, which is prediction (in traffic predictive analytics) (Ceneda et al., 2017). We represent this as:

$$guidance(knowledge\ gap, input) \rightarrow output$$

The knowledge gap attempts to answer the question: *what does the user need to know to make progress?* In the context of TPA, this involves the identification of the path towards the development of a suitable analytical solution to a traffic prediction problem. More specifically, the aim of providing guidance in traffic predictive analytics is the identification of a pathway with which the traffic data scientist traverses from a broad and complex problem space to a narrow and well-defined analytical solution space (see Figure 2-1).

The second dimension within the conceptual guidance model is related to the domain. Five domains are identified (Ceneda et al., 2017): (i) Data, (ii) Tasks, (iii) Methods, (iv) Users, and (v) Infrastructure, which relate to the domain(s) that describe the input perspective. In the context of traffic predictive analytics, the guidance domain pertains to infrastructure use. More specifically, this can be described as a situation where the user (traffic data scientist) is unsure about which analytical approach to adopt in the form of a predictive model. This implies that guidance provided to the user can be in the form of prediction model suggestions.

The 'input and output' dimension helps provide answers to the question: *how is the guidance generated, and how is it presented?* The input defines what characterises the foundational component for the guidance system. In traffic predictive analytics, this can be seen as the domain knowledge about the analytical/prediction techniques. This can be achieved via a number of methods such as expert system generation, meta-knowledge, or machine learning/AI methods. The output specifies how the guidance will be offered/presented to the user. Two possible output media are identified in Ceneda et al. (2017): (i) a *means* and (ii) an *answer*. The goal of a means is the stimulation of an impulse that triggers further exploratory options. In the context of TPA, this is the provision of a means to achieving the desired end goal, which is traffic parameter prediction.

Finally, the degree of guidance assists the traffic data scientist in answering the critical question of how much guidance is/should be provided to the user? In the context of TPA, guidance can be provided at all levels as described in the previous section (i.e. the prescription, direction, and orienting). A guidance system providing mere orientation constitutes the low end of analytical guidance (Ceneda et al., 2017). As portrayed in the example of a traffic analytics guidance system, this can be provided via graphical visualisations of the dataset, potential target variables, results, etc. Directional guidance constitutes medium-level guidance. Unlike orienting, this level of guidance highlights a certain choice of preference for the analytics solution development process. In other words, directional guidance presents the user with a set of alternative options (in this study, prediction models) in order to achieve the desired objective (Ceneda et al., 2017). Finally, prescriptive guidance represents the highest level of guidance offered by an analytics system and relates to a situation where the system unanimously and independently makes decisions on the analytics solution developmental process. The degree of guidance to be provided to the user is a continuum that should be in tandem with the needs of the user. In our proposed framework, the focus is on the provision of medium level guidance – providing directional guidance.

From Figure 2-1, it can be seen that the key identified issues in TPA relate to the choice of one (or more) PAMs from a plethora of available PAMs. Therefore, a possible guidance approach can offer a ranked list of potentially accurate PAMs from a possible list of candidate PAMs, which can be applied towards traffic prediction. Although our framework is streamlined to provide directional (medium level) guidance, it must, however, be mentioned here that it is an achievable task to modify the framework to provide prescriptive (highest level) guidance. For instance, via an automated predict-visualise framework in which the suggested models (from the framework) are passed to a prediction-simulation model. However, given the constraint of resources available to the writer, the complete integration of this component and module has been identified for future research work.

Figure 2-1 presents a conceptual model of the guidance offered in this research study. The overall aim is to guide or assist the traffic data scientist to traverse or navigate from a state in which the solution path is unknown or unclear (i.e. ambiguous, complex or vague) to a point wherein this path towards achieving the solution is clearly defined. In Figure 2-1, this is represented using the bottom arrow (i.e. from bottom left to bottom right).

The left-hand side of Figure 2-1 represents the inputs to the TPA guidance process. The problem specification refers to the specific TPA requirements, which are typically obtained via requirements gathering, business analysis, etc. Secondly, the domain knowledge refers to the knowledge base developed from what is available from the literature (refer to Section 6.3). Thirdly, the TPA expertise of the user constitutes the input to the guidance process. Finally, the degree of guidance refers to one of the three levels of guidance – prescriptive, directional, and orientative (refer to Section 2.5).

The guidance (middle partition of Figure 2-1) is generated by providing an answer to the question ‘what does the user need to know in other to make progress?’ and is achieved by providing solutions to the key issues that arise from the understanding of each respective TPA dimension (three circles). The output (i.e. right-hand side of Figure 2-1) refers to the guidance offered by the guidance system to the user. In this study, these include the characterisation of the TPA problem space using the key dimensions identified in the TAG-F framework, as well as the suggestion of an ordered list of alternative PAMs that can be applied towards the given TPA scenario.

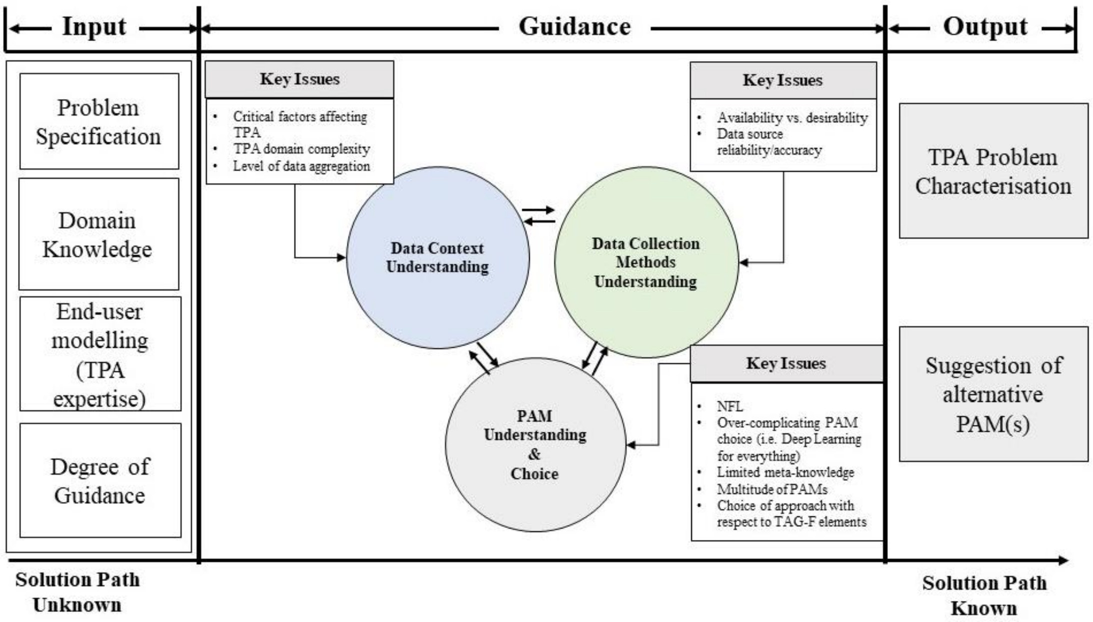

Input
Guidance
Output
Problem Specification
Domain Knowledge
End-user modelling (TPA expertise)
Degree of Guidance
Key Issues
• Critical factors affecting TPA
• TPA domain complexity
• Level of data aggregation
Data Context Understanding
Data Collection Methods Understanding
Key Issues
• Availability vs. desirability
• Data source reliability/accuracy
PAM Understanding & Choice
Key Issues
• NFL
• Over-complicating PAM choice (i.e. Deep Learning for everything)
• Limited meta-knowledge
• Multitude of PAMs
• Choice of approach with respect to TAG-F elements
TPA Problem Characterisation
Suggestion of alternative PAM(s)
Solution Path Unknown
Solution Path Known

Figure 2-1: A conceptual model of guidance in TPA

2.6 Guidance, Analytics and Business/Organisational Decisions

The previous sections have discussed the definition of guidance, its goals and conceptualised guidance in the context of TPA. This section links the concepts together by discussing the need for guidance in TPA. The definition and characterisation of guidance presented in the preceding sections make clear the fact that analytics and guidance both relate to the provision of advice to users for properly understanding, making sense, or achieving a target action – traffic parameter prediction.

The authors in Edwards and Taborda (2016) presented a conceptual model that relates analytical techniques, data, and human knowledge in the process of linking analytics to business/organisational proceedings. The authors highlight the importance of analytics and guidance in order to achieve the set objectives of business requirements. Relating this to traffic management authorities/organisations, it demonstrates that there is a need for guidance in the analytical solution development process, as this will contribute to improving the overall efficiency and effectiveness of the analytical approach towards solving the traffic predictive analytical problem. The benefits of TPA to business organisations are numerous. For instance, businesses that are into logistics and supply chain management will benefit from a congestion-free road network, fostering prompt and reliable deliveries to their customers.

2.7 Characterising Traffic Predictive Analytics (TPA)

Going by the definition of *predictive analytics* in Section 2.2 above, TPA refers to the application of advanced statistical techniques, machine learning, artificial intelligence and predictive models to perform traffic parameter prediction. In the literature, the terms *traffic prediction* or *forecasting* are typically used and refer to TPA. Consequently, few studies exist that present an attempt towards describing or characterising the TPA domain. Although the studies each propose varying descriptive characteristics, factors or parameters affecting TPA, a number of central themes can be gathered. Table 2-1 presents a summary of key TPA factors or parameters, as shown in existing relevant studies. The studies presented in the table are review or survey papers, which are relevant to the field of TPA. The subsequent paragraphs present discussions about the key emerging factors affecting TPA, as well as the challenges encountered in TPA.

2.7.1 Key factors or parameters affecting TPA

Predictive model

The model refers to the predictive methodological approach to be used for the analytical modelling process, which comprises a broad portfolio of algorithms and techniques, such as Autoregressive Integrated Moving Average (ARIMA), state-space models, Artificial Neural Networks (ANN), Support Vector Regression (SVR), and deep learning models. The predictive modelling technique/algorithm comprises the analytical component of data-driven traffic parameter prediction. It is the engine that drives the entire process, thereby making the choice of modelling approach to follow a critical decision. Given the multidisciplinary nature of TPA, brought about by the abundance of available traffic data at high resolutions and aggregations, traffic prediction has consequently been addressed from a number of perspectives: as a time series analysis problem (Min and Wynter, 2011; Moayedi and MA Masnadi-Shirazi, 2008; Qiao et al., 2013), regression and function approximation (Dunne and Ghosh, 2011), and pattern recognition (Jia et al., 2017a), to mention a few.

According to recent TPA studies, there is a shortage of a clear view of the modelling requirements of TPA (Lana et al., 2018, Vlahogianni et al., 2014). TPA modelling approaches are classified into *parametric* and *nonparametric* techniques, with the former class mainly comprising statistical and classical algorithms/models such as ARIMA, historical averaging, and smoothing techniques. In recent times, multivariate, state-space models or *spatiotemporal* models are finding popularity in the field of TPA, due to their multivariate nature and their ability to capture both the spatial and temporal dimension of transportation data (Vlahogianni et al., 2004). Vlahogianni et al. (2014) identified the selection of an appropriate model as a key factor impacting TPA. The authors suggest that the norm is the selection of the model that provides the most accurate predictions based on a collected dataset, ignoring the underlying data characteristics.

In a more recent study, Lana et al. (2018), there is a summary of the efforts made in TPA studies, and the subsequent articulation of the main criteria as well as challenges encountered in TPA. The review study adopted a taxonomy of existing studies using key criteria most commonly found in the studies. The requirements are related to the *predictive method or model, horizon, scale, and output variables*. According to the authors, the predictive model is a relevant part of the TPA process, although the task of evaluating the predictive performance is more challenging.

Table 2-1: Summary of key parameters affecting TPA from related studies

Source	Key Parameters
(Karlaftis and Vlahogianni, 2011; Poonia et al., 2018; Vlahogianni et al., 2004)	• Modelling methodology (type of output and input, data quality, etc.) • Data Scope determination • Conceptual output specification (i.e. traffic parameters, data resolution)
(Bengio et al., 2013; Buch et al., 2011; Lana et al., 2018)	• Prediction method (i.e. predictive model) • Prediction horizon • Prediction scale (i.e. single location or road segment) • Predictive context (i.e. urban or non-urban) • Data sources (i.e. traffic sensors, GPS, cellular, etc.) • Exogenous factors (i.e. calendar, time of day, weather, events, accidents, road works, etc.) • Predicted variables (i.e. flow, speed, time, etc.) • Application scope (i.e. ATMS or ATIS) • Stream mining (i.e. real-time or not)
(Barros et al., 2015; Ermagun and Levinson, 2018)	• Real-time data collection • Prediction metrics and targets • Measuring prediction accuracy
(Oh et al., 2015)	• Prediction range • Accuracy • Efficiency • Applicability • Robustness
(Vlahogianni et al., 2014)	• Implementation area • Traffic input parameter(s) • Prediction range (i.e. steps and horizon) • Data collection method • Methodology (i.e. model type, state-space, optimisation, approach)

Similarly, in Barros et al., (2015), the focus of the analysis was on the dichotomy of model-driven and data-driven short-term prediction. Model-driven prediction refers to the computational modelling of the road network typically via simulation and visualisation to analyse the performance and behaviour of the road users (Bacchiani et al., 2019). This differs from data-driven prediction, which is the focus of this present research study, that makes predictions using historical and real-time (and/or historical) data features. Barros et al. (2015) performed a critical analysis to allow an understanding of the advantages, disadvantages, trade-offs of some predictive models (*k*-NN, hidden Markov, particle filter, Bayesian combined neural network, particle filtering with non-explicit state-transition model, adaptive Kalman filter, ARIMA, and state-space ARIMA), which can provide useful insight for future model development. In another study, Oh et al. (2015), a review of data-driven prediction in highways is presented, with focus on critically analysing the common predictive models in use within the field of TPA. In the study, five (5) main perspectives (or parameters) are identified, which include *prediction range, accuracy, efficiency, applicability,* and *robustness.* The study considered four (4) predictive models – ARIMA, ANN, *k*-NN, and Kalman Filter using the identified perspectives or parameters listed above. The findings from the study identified the strengths and potential weaknesses amongst the reviewed predictive models, specifically emphasising the efficiency (i.e. fast computation) of parametric models, which use well-defined theoretical foundations for parameter prediction.

On the other hand, this class of predictive models is known to have problems relating to adaptation, limited performance on network-wide prediction, and vulnerabilities when applied on non-linear data (Oh et al., 2015). ANNs, in the other class of predictive models, – nonparametric or machine learning – perform well in complicated, non-linear datasets by learning the underlying patterns obtainable from the training datasets. In recent times, deep learning models have shown promise in predicting complex and non-linear traffic parameters (Goodfellow and Bengio, 2015).

Although the accuracies obtained from artificial intelligence (AI) models (ANN and deep learning) are sometimes exceptional, there are a number of challenges and demerits for adopting these models. One of such has to do with the interpretability or traceability of the predictions. For instance, what features affect the predictive response of the model? Also, how can one rationalise the predictive outcome of a neural network? The answers to these questions have birthed a new field of research known as *interpretable AI* (Hall et al., 2017).

Input Data sources

In this context, *data source* refers to the source of the traffic data to be used in the TPA process. With rapid technological advancement and data ubiquity comes a multitude of traffic data sources, such as inductive loop devices (ILDs), Bluetooth sensors, infrared, video camera, global positioning systems (GPS), floating car data (FCD), to mention a few. It is typical to find predictive models built on single-source traffic data, thereby limiting the generalisability of the resultant model (Lana et al., 2018). In terms of the exogenous factors, it is essential to account for which non-traffic inputs would impact the traffic prediction space. This constitutes a key influencer of the potential accuracy that is obtainable (especially in long-term TPA) because some non-traffic factors influence the traffic state and its stochasticity but are not reflected in the seasonal behaviour (Lana et al., 2018).

In recent times, the more common exogenous factors adopted in the TPA process include weather-related (Essien et al., 2019a; Tsapakis et al., 2013), social media (Lu Lin et al., 2018; Ni et al., 2014) and accidents (Kumar et al., 2015; Lu Lin et al., 2018). Over the years, it has been observed from studies that harsh weather considerably affects traffic flow. Furthermore, many articles report a significant relationship between rainfall and traffic accidents (Peng et al., 2018; Smith et al., 2004; Tsapakis et al., 2013). Similarly, Qiu and Nixon (2008) showed that rainfall increased crash and injury rates by 71% and 49% respectively. Therefore, rain conditions reduce traffic capacity and operating speeds, thereby increasing congestion. For these reasons, traffic data scientists and engineers are seeking ways to incorporate weather-related data into traffic planning and operations, because this can improve traffic prediction and modelling. A comprehensive analysis of weather effects on urban transport networks is essential for understanding traffic network performance (Koetse and Rietveld, 2009). The absence of a clear understanding of the direct impact on traffic by weather conditions minimises the potential for transportation stakeholders and policymakers to capitalise on additional intelligence provided by weather-related data sources to develop improved traffic management strategies (Agarwal et al., 2005).

The writer of this thesis conducted a study investigating the impact of weather data (temperature and precipitation) on urban traffic flow characteristics (Essien et al., 2018). The research was carried out to establish and quantify the effects of weather on traffic characteristics – speed, volume, and density. More specifically, historical weather data obtained from the Centre for Meteorological Services (CMS) at the University of Manchester in addition to historical traffic data (average speed, travel time, density, and flow) from an urban arterial, Chester Road A56, within the Greater Manchester region of the United Kingdom were used. Findings from the study revealed that rainfall did affect urban traffic, but was dependent on the intensity and time of day. Furthermore, light rain had no impact on peak urban traffic, while average speed reduced with moderate and heavy rainfall by 2.6% and 9.7% respectively. The findings also revealed that light, moderate, and heavy rain decreased peak traffic flow by 2.5%, 1.3%, and 5.2% respectively. However, a different set of results were obtained at off-peak periods, as light rainfall reduced average off-peak speed by 4.9%, while moderate rainfall reduced off-peak speed by 5.5%. However, in off-peak periods, heavy rainfall increased average speed by 11.4%. The study also performed a detailed analysis of the quantification of the temperature effects on traffic flow parameters. The conclusion was that atmospheric temperature differently affected peak and off-peak traffic flow parameters. For instance, *cold*, *normal*, *warm*, and *hot* temperatures increased average off-peak speeds by 5.1%, 13.9%, 19.2% and 18.7% respectively. Also, during peak periods, *cold, normal, warm,* and *hot* temperatures caused reductions in average speed by 4.1%, 18.7%, 28.2% and 26.7% respectively. This is in agreement with many prior studies that show that weather-related data sources can affect traffic status, thereby indicating the importance of the inclusion of weather-related data in TPA solution development.

Traffic Scope/Area of Implementation

The traffic scope or area of implementation for traffic prediction is typically categorised into highway/motorway (or freeways) and urban/arterial roads. The striking distinction between the two groups is the presence of controlled intersections, which are found within urban arterial roads (Van-Lint et al., 2005). Another marked difference relates to the nature of the traffic flow, with urban traffic having a more dynamic and complex flow pattern compared to highway traffic (Vlahogianni et al., 2004). Furthermore, in urban road networks, the spatiotemporal characteristics are more complex to model given the presence of many adjacent links, upstream, and downstream traffic.

As observed in the literature (Barros et al., 2015; Lana et al., 2018; Vlahogianni et al., 2014, 2004), the vast majority of traffic prediction studies are implemented in freeway or highway traffic conditions. This can be rationalised by the greater variance in the traffic parameters in highways or motorways when compared to urban traffic scenarios, which have lower speed limits (typically 30 mph). However, some studies have performed traffic prediction in urban conditions (Min *et al.*, 2009; Alajali, Wen and Zhou, 2017; Essien *et al.*, 2019b). Consequently, there is a need for the development of urban traffic predictive models, as this would be more useful in urban traffic congestion management and control, thereby resulting in the provision of reliable, accurate information to road users and transport network managers.

In Vlahogianni et al. (2004), the determination of scope covers the area and type of implementation. The type of implementation is defined by the nature of the application of the implementation. Two main systems are identified – Advanced Traffic Management Systems (ATMS), and Advanced Travel Information Systems (ATIS). Both systems, however, can be affected by the reliability and accuracy of the real-time information about the evolution of the traffic network with time. ATMSs are typically used to control and manage the traffic network on the basis of real-time traffic data, which is obtained using traffic sensors. ATISs, on the other hand, are mainly used for providing real-time traffic information to road users and traffic control personnel/authorities (e.g. Google Maps, Waze, and Garmin).

The conceptual output specification considers the data resolution, which relates to the *prediction horizon* and *time step*. The *prediction horizon* is the extent of the time ahead to which the forecasting or prediction is executed (Vlahogianni et al., 2004). This is different from the predictive time step, which refers to the time interval, and is a function of the frequency of predictions made in the prediction horizon. For instance, a predictive model predicts over a 30-min prediction horizon in 10-min intervals or time steps. The relationship between the prediction horizon and predictive model accuracy can be intuitively understood – the larger the prediction horizon, the less accurate the predictions. This has been experimentally proven in Ishak and Al-Deek (2002). The conceptual output specification also considers the traffic parameters, which comprise the input and output feature space. Therefore, there lies the distinction between *univariate* (single) and *multivariate* traffic prediction. The typical traffic parameters considered include traffic *flow, speed,* and *density*.

In terms of *prediction mode* (i.e. real-time or offline), the authors in Barros et al., (2015) argue that traffic analysts typically follow a reactive 'analytics ⇒ response' approach, which increases operating cost via 24/7 monitoring. In order to implement proactive traffic management, the application of real-time traffic parameter prediction becomes necessary. This, therefore, makes prediction mode a key influencer of TPA process lifecycle, as it has the potential to impact the nature of the TPA solution development. There is a consequent increase in the number of studies concerning data-driven methods for real-time traffic prediction. For instance, Zhang et al. (2013) presented a traffic prediction model based on k-Nearest Neighbours (k-NN) for predicting traffic parameters in 5-min time steps. Zheng, Li, and Chi (2006) also presented a hybrid machine learning approach to traffic prediction using backpropagation and radial basis function (RBF) neural networks, resulting in an overarching Bayesian combined model (BCNN). The model was tested on real-time data in Singapore on a 15-min prediction horizon and deducted that the BCNN model outperformed the other models in terms of prediction accuracy.

Data quality

Garbage in, garbage out (GIGO) is a famous phrase in computer science that is used to state the concept that the quality of the output is a linear function of the input data.

This comprises another critical dimension of data-driven traffic parameter forecasting. In today's data era, the quantum and ubiquitous availability of data accelerate the risk of low-quality data being used within the TPA process. Large historical datasets are particularly advantageous when it comes to data-driven, nonparametric modelling approaches. For instance, in a real-time traffic prediction scenario, the input data quality can be significantly impacted due to equipment failure/malfunction, noise, or missing values. In a situation where the prediction model cannot accommodate such circumstances, then the onus lies with the traffic data scientist in terms of ensuring that the quality of the input data is kept at optimal levels at all times.

Spatial and Temporal considerations

According to Ermagun and Levinson (2018), the inclusion of temporal and spatial relationships of input data is used to improve the prediction accuracy of traffic prediction models. As is known, traffic data is seasonal at daily and weekly levels (Barros et al., 2015), implying that the inclusion of traffic data from, say upstream locations, can improve the predictive accuracy of prediction models aiming at making downstream traffic predictions. The first use of spatiotemporal characteristics for traffic prediction was recorded in 1984, where the authors considered spatial information from upstream feeder links (Okutani and Stephanedes, 1984). A popular space-time prediction model, the Space-Time Auto-Regressive Integrated Moving Average (STARIMA) has been in use in many traffic prediction studies and shows promising signs in traffic forecasting (Ding et al., 2011; Duan et al., 2016; Min et al., 2009).

Table 2-2: Summary of TPA challenges from related studies

Source	Traffic Category	Key Challenges
(Vlahogianni et al., 2004)	Short-term TPA	• Appropriate model selection
(Lana et al., 2018)	Short-term TPA	• Stochastic nature of traffic • Prediction context • Urban traffic • Model selection • Performance evaluation metrics • Model hybridisation
(Barros et al., 2015)	Short-term TPA	• Input data selection • Performance evaluation metrics • Predictive model selection
(Oh et al., 2015)	Short-term TPA	• Appropriate model selection • Lack of (current) shared knowledge about TPA predictive models/algorithms
(Vlahogianni et al., 2014)	Short-term TPA	• Developing responsive algorithms • Freeway/motorway prediction • Short-term predictions: from volume to time • Data resolution, aggregation and quality • Using new technologies for collecting and fusing data • Temporal characteristics and spatial dependence • Model selection and testing • Compare models or combine forecasts? • Explanatory power, associations and causality • Realising the full potential of AI

2.7.2 Challenges in TPA

Several themes and challenges can be extracted from the prior TPA review studies. In this sub-section, a discussion about the identified challenges encountered in performing TPA is presented. In the end, a summary of the key challenges identified is presented, which forms the direction of the research study presented in this thesis. Table 2-2 summarises the key challenges encountered in TPA, from the main studies discussed in this chapter. According to Barros et al., (2015), the challenges obtainable in data-driven traffic parameter prediction are identified as *input parameter selection*, *model performance evaluation metrics*, and *predictive model selection*. The predictive model selection is a recurring theme in many TPA review studies (Karlaftis and Vlahogianni, 2011; Lana et al., 2018; Oh et al., 2015; Vlahogianni et al., 2004), which enables the conclusion that identifying the appropriate modelling approach constitutes a major challenge in TPA.

Model Evaluation

Similarly, in Vlahogianni et al., (2014), another challenge refers to *model evaluation* or *testing*. In the literature, there is a common practice where many studies have emphasised the discussion of findings obtained, thereby neglecting the need to account for the quality of the proposed model using popular statistical diagnostics (Vlahogianni et al., 2014). As previously stated, the typical error metrics include Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), and although it is good practice to assess the performance of predictive models on these statistical metrics, it is even better to identify and discuss the presence of any ''strong'' properties in the error metrics, – including correlation, volatility, etc. – which may indicate a generalisation or model bias that can be attributed to variables or misspecification of the functional form (in parametric models). According to Vlahogianni et al. (2014), AI approaches rarely incorporate any testing of the properties of the error and the model specification. In terms of the challenges encountered in TPA, Lana et al. (2018) identified a number of emerging themes in TPA literature, which centre around the *stochasticity of traffic*, the *network-less application*, *applicability and model selection*, *performance evaluation metrics*, and *hybridisation of methods* (see Table 2-2). Traffic prediction is typically challenging due to the randomness of the events that can alter the traffic situation, as well as the impacts the predictions have on the road users' decisions and habits (Lana et al., 2018).

Predictive Model Selection and meta-learning

Research studies have empirically and theoretically proven that there is no single algorithm that performs optimally in all prediction scenarios (Ferrari and De-Castro, 2015; Smith-Miles, 2009; Wolpert and Macready, 1997). In 1976, Jordan Rice proposed the *Algorithm Selection Problem (ASP)*, which stated the relationship between the features or characteristics of a prediction problem/scenario and the performance of the optimal algorithm that can be used for solving it (Rice, 1976). Ever since then, there has been an increase in research studies that attempt to solve the (NP-Hard) ASP problem (Ferrari and De-Castro, 2015; Pappa et al., 2014). As can be imagined, there are apparent limitations when systems that are based on human judgement are applied towards solving these problems. For instance, such methods will be static and require significant work to update it based on previous experiments. In addition, there is also the susceptibility to human error (Xiaofeng Wang et al., 2009).

In the literature, predictive algorithm model selection has typically adopted one of two approaches:

(i) An extensive review of various approaches towards traffic prediction using data-driven models and the subsequent use of (human) expert knowledge in providing guidelines for model selection.

(ii) Extensively reviewing results obtained from empirical studies to estimate a relationship between the model, data features/attributes and model performance.

The first approach has been extensively covered within the last two decades of traffic prediction, so it is common to find studies reviewing traffic predictive methods in the existing literature (Barros et al., 2015; Davis and Nihan, 2007; Karlaftis and Vlahogianni, 2011; Lana et al., 2018; Poonia et al., 2018; Vlahogianni et al., 2004) to develop specific guidelines for predictive model selection.

The second approach involves extracting meta-knowledge about the algorithms to infer in what predictive scenarios they are most effective. This is referred to as *meta-learning.* Meta-learning typically focuses on selecting a predictive algorithm or set of hyperparameters by learning about the characteristics of predictive algorithms (meta-features) that characterise a given dataset (Smith, Mitchell and Giraud-Carrier, 2014). A meta-learning algorithm, therefore, aims to determine which prediction scenario's characteristic(s) contribute towards improving the performance of one algorithm in comparison to others and utilises this knowledge (meta-knowledge) for selecting the most suitable prediction algorithm for the given problem scenario (Reif et al., 2014). According to Ferrari and De-Castro (2015), meta-knowledge can be categorised either as *meta-attributes*, (i.e. the set of characteristics or features that are within the prediction problem/scenario), or *meta-target* (referring to the particular target variable for the meta-learning algorithm).

There has been an increased interest in the use of expert systems for aiding in predictive algorithm choice for predictive analytics (Laud and Ibrahim, 1995; Piironen and Vehtari, 2017). A widely adopted approach is a *rule-based* forecasting method in Collopy and Armstrong (1992), where the proposed 'expert' system formalised model selection using rules extracted from a meta-learning process. The authors in Collopy and Armstrong (1992) obtained 99 rules (from statistical experts), which were applied to evaluate four different models to infer which scenario is optimal for a given model. However, the potential drawbacks that can be obtainable by adopting such an approach, especially subjectivity and human error, have been circumvented by the use of machine learning algorithms, which can automatically acquire meta-knowledge for model selection. Makridakis, Hibon, and Moser (1979) represents an early study proposing the notion that data features can provide useful knowledge that can be utilised for predictive algorithm selection.

The ASP problem described in the previous section can be applied to the TPA problem domain. As pointed out in prior studies (Brodley, 1993; Laud and Ibrahim, 1995; Piironen and Vehtari, 2017), a critical issue in traffic prediction relates to the selection of an appropriate predictive model (or predictive analytical method). Some existing studies have attempted to solve the problem by adopting selective multi-model approaches. The following paragraphs present brief discussions of these studies that have attempted to solve the TPA predictive model selection problem stated in Rice (1976).

The authors in Asencio-Cortés (2016) presented a methodology incorporating ensemble learning for urban traffic prediction. The proposed methodology comprised seven (7) machine learning algorithms – *k*-Nearest Neighbours (*k*-NN), C4.5 decision trees, Artificial Neural Networks (ANN), stochastic gradient descent (SGD) optimisation, fuzzy unordered rule induction algorithm (FURIA), Bayesian Networks (BN), and Support Vector Machines (SVM). The framework was tested using sensor collected data in the urban Spanish city of Seville. Prediction accuracies reaching 83% were realised from the study, and the authors were able to show that turning the traffic prediction problem into a binary classification task showed great potential.

In another study, Zhou et al. (2019), the authors presented a learning-based multi-model integrated framework for online dynamic traffic parameter forecasting. The proposed model incorporated a deep-learning architecture, combining predictive algorithms that have been widely adopted in the TPA domain. The framework adopted stacked autoencoders (SAE) for extracting the relationships hidden within the traffic data, before passing the prediction to a probability-based model integration module. The authors, however, pointed out a limitation of their framework with respect to the shallow nature of the framework, which displays vulnerabilities when exploring real-time datasets.

Table 2-3: Summary of challenges in TPA model selection

S/No	Challenge Description	Source(s)	Potential Solution
1.	Model ranking is based on actual/physical execution using the dataset (either using grid search, genetic algorithm, or other optimisation techniques)	(Hall et al., 2017; Kotthoff et al., 2016; Chris Thornton et al., 2013; Xiaofeng Wang et al., 2009)	An inference model trained using meta-knowledge obtained from existing empirical studies to estimate a relationship between the data features, attributes and model performance
2.	Existing multi-model approaches are not scalable and adaptive and consequently quickly get superseded due to the rapidly-evolving field of data science	(Brazdil, 2003; Lu Lin et al., 2018; Lindauer et al., 2017; Xiaofeng Wang et al., 2009)	A scalable meta-knowledge model selection method.
3.	Subjectivity due to human-in-the-loop ‘expert’ judgment	(Mendoza et al., 2016; Xiaofeng Wang et al., 2009)	An automated (self-learning) approach to TPA model suggestion.

From the foregoing, although multi-model approaches towards traffic prediction have attracted interest in recent studies, it – however – is still deficient in some respects.

Table 2-3 summarises the key challenges encountered in multi-model approaches for appropriate model selection in TPA. First, current model selection algorithms typically provide suggestions based on the actual execution of the dataset (i.e. the rankings of the model performance are a by-product of the predictive analytics task). Given the large amount of data available today, and the long training time required in the complex (deep learning) models, there is little benefit in this approach. Secondly, there is no scalable model selection algorithm that can be extended to accommodate the rapidly-evolving research field of TPA. The common trend is the existence of model selection algorithms that are developed using a fixed set of models, which are not extendable to be able to accommodate the additional models. Thirdly, some model selection algorithms adopt ‘expert’ (human) judgment in the rule induction process, which has severe limitations as pointed out in the preceding section (Xiaofeng Wang et al., 2009).

2.8 Discussion

A prerequisite for effective traffic management is the accurate and timely provision of information to road users. This typically assumes the form of data-driven, short to medium-term, traffic parameter predictions. In reality, short-term traffic forecasting is mostly used in developing and testing traffic predictive algorithms, mainly due to the abundance of historical traffic datasets (Lana et al., 2018). However, achieving accurate data-driven traffic parameter prediction is a difficult task due to the complexity of the traffic domain and the difficulty in selecting a suitable PAM due to the plethora of such predictors. This leaves a number of research opportunities, which this study aims to address.

Firstly, given the wide spectrum of data sources and traffic prediction algorithms at the disposal of traffic data scientists, the process of developing an analytics effort capable of solving a traffic prediction problem becomes challenging. The number of factors to consider and the dimensions and attributes for each dataset being analysed further contribute to the complexity. For this reason, there is a need for a structured approach towards developing an analytical solution capable of providing support to the traffic data scientists, in order to traverse from a complex traffic problem space to a well-defined analytical solution space, culminating in an action or outcome, which is usually prediction. To provide a solution to this, there is a need for a framework comprising structured decision points that can serve as a roadmap for traffic data analytics.

Secondly, due to the plethora of predictive algorithms available to traffic data scientists today, there is an added difficulty associated with the choice of predictive model (Vlahogianni et al., 2004). Research into short-term traffic forecasting has attracted a lot of interest in the past decade and has, therefore, seen the development of many predictive algorithms. Regardless of the problem categorisation, the approach adopted towards short-term traffic forecasting should be centred on the choice of a model that provides the most accurate predictions on the basis of the available dataset. Furthermore, due to recent big data explosions, realised by sensors, interconnected vehicles, smart cities, Internet of Things (IoT), etc., there is an inherent need for the use of deep nonlinear architectures to successfully and effectively analyse large datasets. In order to get this ‘best’ predictive model, there should be a comparison between a set of baseline models via series of tests and rigorous comparisons of modelling specifications and prediction results. However, in reality, the possibility of evaluating every available predictive algorithm in all possible prediction problem scenarios (i.e. brute force approach) is a task that may run into many years even on the most powerful supercomputer. Therefore, an alternative to this would be the availability of a model or framework approach towards disseminating shared knowledge about prediction algorithms, given a set of TPA data specifications.

Thirdly, it has been identified that most short-term traffic predictors were developed and tested on freeway/highway/motorway traffic networks. Urban traffic tends to receive minimal attention in existing research studies, and this can be attributed to a number of reasons. First, urban traffic represents a more complex problem scenario that freeway traffic due to many factors such as reduced operating speed, steep fluctuations due to signalised intersections, pedestrian crossings, and bus stops. Therefore, urban traffic constitutes a major point of concern for policymakers, given that a greater percentage of traffic congestions occur in urban traffic settings. Secondly, urban traffic links are shorter in length than motorway or freeway links, thereby giving more aggregated data readings from the data collection sensors.

In an attempt to address the above-listed research opportunities, this research is motivated to explore an approach towards the development of a framework and tool capable of providing guidance to traffic data scientists in performing TPA. The result is a novel framework, referred to as The *Traffic Analytics Guidance Framework* (TAG-F). Details of this framework are presented in Chapter 5.

2.9 Chapter Summary

The objectives of this chapter were to build upon the background presented in Section 1.1 and provide a review of related literature about key concepts, trends, and identified opportunities for research in TPA. The key concepts reviewed formed the foundation upon which this thesis is built. The next paragraphs briefly summarise the arguments made and findings realised within this chapter.

Firstly, in Section 2.3, the definition of guidance, which relates to the provision of a set of alternatives or advice to a user in order to achieve a set objective, is presented. The chapter also introduced an enhanced traffic guidance conceptual framework, building on a visual analytics guidance framework (Ceneda et al., 2017), customised to the TPA solution development space. In Section 2.6, a discussion is presented, which links guidance to analytics and business/organisational decision-making, stressing the importance of guidance and analytics in the decision-making process of business organisations, especially in the field of traffic data analytics which is complicated, dynamic and stochastic.

Furthermore, a brief background of data analytics was presented, including the three (3) sub-classes of analytics – *predictive*, *prescriptive,* and *descriptive analytics* – which form the core of business intelligence and data-driven analytical practice. Secondly, using prior review studies about TPA, a characterisation of the TPA problem space was established, delineating TPA into critical factors that impact or affect the execution of TPA. In addition, the challenges encountered in TPA were also listed in Section 2.7.2. This resulted in the identification of the core challenges encountered in TPA, which include *predictive model selection*, *model performance evaluation metrics*, and over-reliance on *freeway/traffic implementation area*. This chapter highlighted the key elements affecting TPA and how external data sources (weather-related) affect traffic flow prediction and modelling (Section 2.7.1). It concluded by identifying research opportunities within the literature – the lack of shared knowledge about existent traffic prediction algorithms, the major focus of traffic prediction studies on freeway or motorway traffic networks, and the difficulty in selecting an appropriate predictive model to solve a complex traffic prediction problem.

In summary, as a potential solution to the identified research shortcomings, the development of a robust predictive analytics guidance framework that can serve as a roadmap for traffic data scientists may prove useful. However, providing guidance to traffic data scientists requires an understanding of the traffic prediction problem. In the next chapter, a brief background about technical concepts such as traffic flow theory, traffic data collection methods and predictive modelling techniques is presented.

Chapter 3 Traffic Prediction Background

3.1 Introduction

This chapter will provide a technical background to traffic prediction and will analyse existing relevant studies that fall into the category of data-driven traffic parameter forecasting methods. The chapter begins with a brief overview of mobility and the provision of traffic information. In Section 3.3, an overview of the fundamentals of traffic flow theory is presented. Prior to the successful application of this theory, traffic-related data needs to be collected, which is why the chapter also presents an overview of traffic data collection methods, highlighting the individual strengths and weaknesses of some traffic data collection methods (Section 3.4). In Section 3.5, discussions about some traffic prediction algorithms are presented. The chapter is concluded in Section 3.6.

3.2 Traffic mobility and information provision

In recent history, the value of mobility and information has rapidly grown, and this trend is expected to continue. The need for easy access to goods and services has precipitated an increase in the demand for transport and mobility the world over. According to the Department for Transport (DfT) in the United Kingdom, over 80% of passenger journeys were by car, van, or taxi (GOV.UK, 2018). This implies that road mobility has become an essential part of economic and social development, with the most utilised travel mode being car, van, or taxi (see Figure 3-1). This rapid increase in *mobility* – defined as the ability to move or be moved freely – cannot be over-emphasised. Figure 3-1 presents a graphic of the composition of passenger journeys, grouped by the various travel modes. As can be seen, there is a year-on-year increase in the number of trips completed by car, van, or taxi, implying the significant contribution of this travel mode to traffic congestion, environmental pollution and degradation.

In the past, several approaches have been devised to control traffic congestion, such as the expansion of road infrastructure to meet predicted travel demand. However, with the exponentially-growing population of the world, road expansion no longer constitutes an optimal solution. In this context, ITSs have proven to be a successful alternative for effective traffic management and control. As a result, advanced traffic management systems (ATMS) incorporating ITSs have been developed in the past decades to efficiently manage the existing road network capacity. These systems are able to deliver such services by providing a steady stream of information about the entire traffic network and environment at real-time. The data is obtained via different data collection methods such as Bluetooth sensors, inductive loops, cameras, radar, and floating-car data (FCD), to mention a few. A fundamental characteristic of the information provided by an ITS is the dynamic nature of the data. This, therefore, suggests that accurate and timely forecasts need to be in place to ensure the success of an ITS for effective traffic control and management.

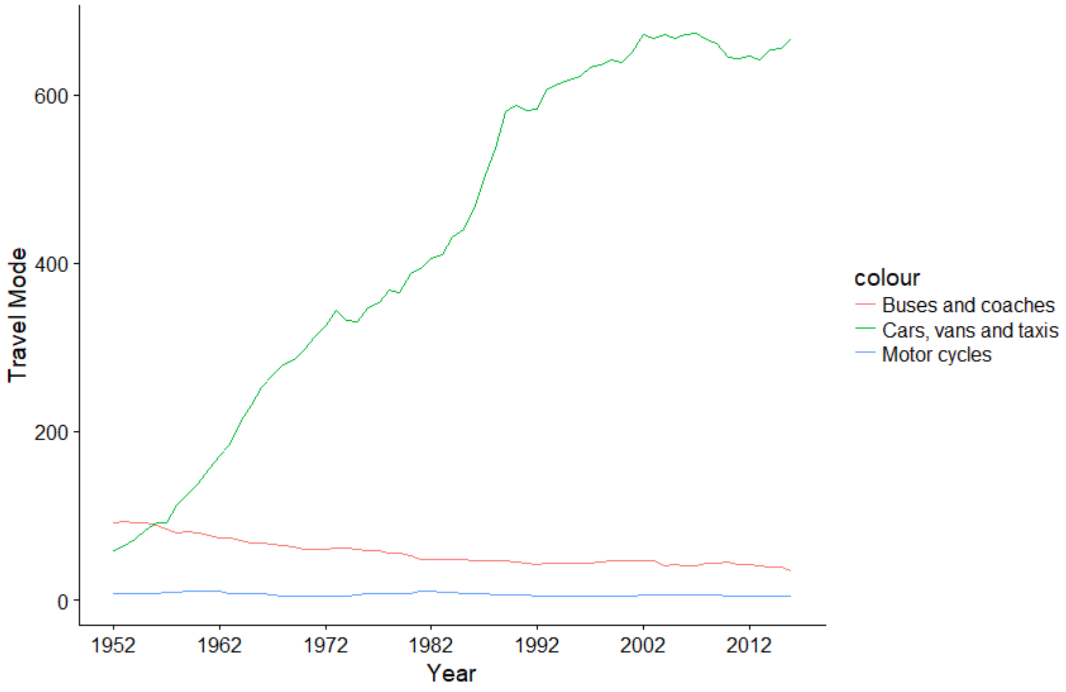


Figure 3-1: Passenger kilometres by travel mode, Great Britain. [Source: www.gov.uk]

Consequently, in recent years, there has been growing interest in the development of a variety of data-driven, statistical and machine learning prediction methods using different configurations to model traffic data and subsequently produce short-term forecasts (Vlahogianni et al., 2004). To provide a temporal context in terms of the evolution of this field of research, Figure 3-2 presents a life-cycle graphical plot of the number of studies relating to short-term prediction. As can be seen from the figure, there has been a significant increase in published research about traffic analytics and forecasting over the past two decades, with the first article published in 1958 (Brokke and Mertz, 1958).

However, achieving accurate data-driven traffic parameter prediction is difficult for three main reasons, as initially summarised in Section 1.1. To reiterate, the traffic domain is a complex one comprising individual actors in the form of road users, affected by dynamically-changing traffic variables and exogenous factors such as rainfall, temperature, events, road works, and accidents. Secondly, selecting a suitable predictive analytical method (PAM) or algorithm can be challenging due to the plethora of available traffic predictive algorithms. Finally, there is no single best predictive algorithm that works best in all scenarios and analytical situations. Therefore, a predictive model that performs optimally in a given predictive scenario tends to perform poorly when exposed to a different problem scenario (Wolpert and Macready, 1997). For these reasons, therefore, it is common to find traffic data scientists mention the complexity involved in planning and organising the data analytics effort and deciding about which approach to follow.

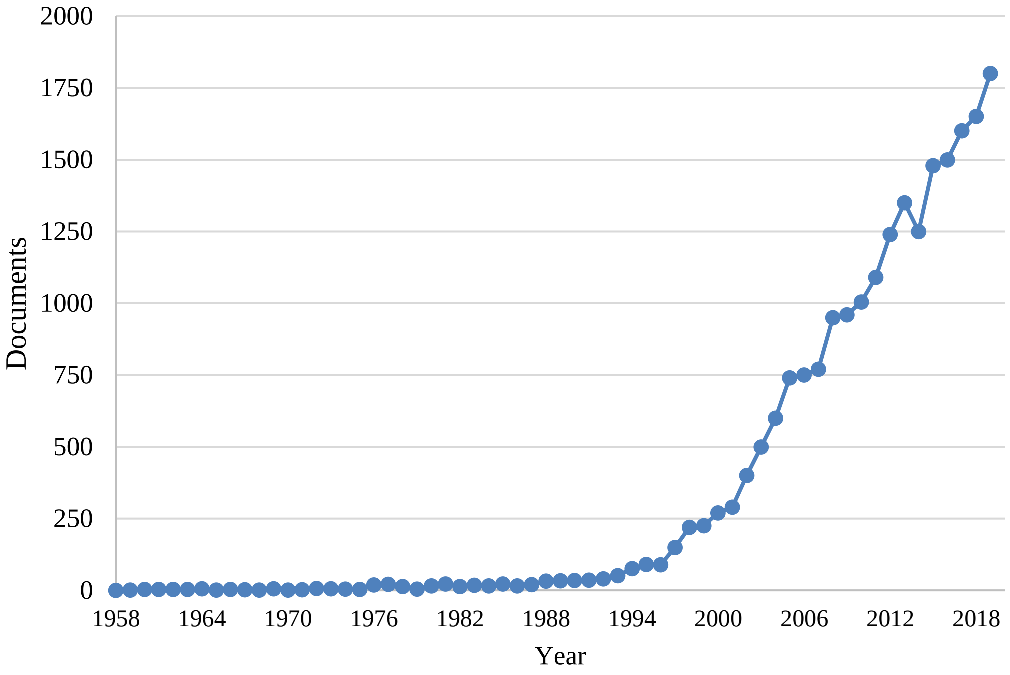


Figure 3-2**:** The traffic prediction life-cycle (source: Scopus 2018)

3.3 Fundamentals of Traffic Flow Theory

Traffic flow involves the movement of discrete units, either in the form of liquids, particles, electrons, vehicles, or people, from one place to another within a transportation system. Generally speaking, these units tend to move independently (as with individual road users), or sometimes interact (as in convoy). According to (Taylor and Bonsall, 2017), three main components of road traffic systems exist. These are:

i. the driver,

ii. the vehicle, and

iii. the environment.

These components all interact with each other. Adequate knowledge and understanding of traffic stream characteristics is a prerequisite to effective traffic management and control as this enables the means of understanding traffic flow characteristics in situations like queuing, car following, lane changing, crossing, to name a few. These characteristics, in addition to the fundamental traffic flow parameters including traffic *speed*, *volume*, and *density*, combine to form the main determinants of road capacity in every traffic environment (Taylor and Bonsall, 2017). In this context, *traffic flow theory* relates to the operations stage within the spectrum of transportation analysis. In more specific terms, an interesting definition of traffic flow theory is presented in Elefteriadou, (2014) “as the aspect of transportation that concerns road capacity and traffic operation quality”. Therefore, traffic flow theory mainly aims to monitor and evaluate the quality of a given traffic stream in a particular set of conditions.

3.3.1 Fundamental Relationships of Traffic Flow Theory

Traffic flow, speed, and volume constitute the primary characteristics of traffic flow and are used to describe the critical aspects of traffic operations. In describing traffic operations, the focus is typically on a group of moving vehicles (i.e. microscopic) or the holistic (macroscopic) traffic stream. Therefore, traffic flow is usually described as either macroscopic (i.e. holistic view of the traffic stream) or microscopic (i.e. focusing on the individual vehicles).

Traffic Flow(q): this is defined as the rate of vehicles travelling through a particular point or road segment. It is expressed in units of traffic per unit time. The standard units of traffic flow are veh/hr, veh/day, or veh/s.

Speed: traffic speed, or velocity, v, is defined as the distance travelled per unit of time. Typical units of speed in traffic engineering are Miles per hour, (mph), or Kilometres per hour (km/h). Speed can be obtained in one of two ways. The first method refers to a situation where the observer measures the instantaneous speed at a location or point. This method, based on instantaneous speeds, produces *time-mean speed* at a given location. The average speed in this method can be computed using the equation below:

$$v_{avg-time} = \frac{\sum_1^n v_i}{n} \tag{3-1}$$

Where v_i represents the instantaneous speed, and n is the total number of instant speed samples.

Another method is where the observer measures the travel time of each vehicle between two locations and then obtains the speed of each vehicle as the inverse of the travel time. This method measures *space-mean speed*. The space-mean speed can be computed using the equation below:

$$v_{avg-space} = \frac{d}{\sum_1^n \frac{t_i}{n}} \tag{3-2}$$

Where d represents the distance, t_i is the observed travel time, and n represents the total number of observations made.

Density: this is otherwise referred to as concentration and is used to express the number of vehicles per unit length of a lane or road segment per given instance of time. It is expressed in units of traffic per unit distance. Typical units for density are vehicles per kilometre (veh/km), Vehicles per mile (veh/m), or Vehicles per mile per lane (v/mpl).

3.3.2 Traffic Stream Models

The three fundamental traffic flow characteristics described above – flow, speed, and density – are related to each other through an equation referred to as the continuity of flow equation (Taylor and Bonsall, 2017), which is calculated as:

$$q = k\bar{v}_{avg-space} \tag{3-3}$$

Where $\bar{v}_{avg-space}$ is the mean space speed and k is the density.

From the inception of this theory, researchers have attempted to model the relationship between these characteristics, as represented in the equation above. This type of modelling can be used to provide forecasting and evaluation of the performance of a road facility. For instance, a traffic analyst can estimate how a road segment will operate if the traffic flow rate is 500 veh/hr. This analysis is used to develop traffic stream models.

Over the years, many such models have been developed. An early traffic stream model is *Greenshield's* model, which was developed using traffic field data (Greenshields, 1935). Figure 3-3 presents the fundamental diagrams of the model. As can be seen, the model describes speed as being inversely related to density, as shown in Figure 3.3(a). In the diagram, v_t is the free-flow speed when density is zero, and this reduces with increase in density until it reaches a minimum point, where the density is at maximum point d_j. This situation is referred to as *traffic jam*, a point where the speed becomes zero (i.e. vehicles are at a standstill).

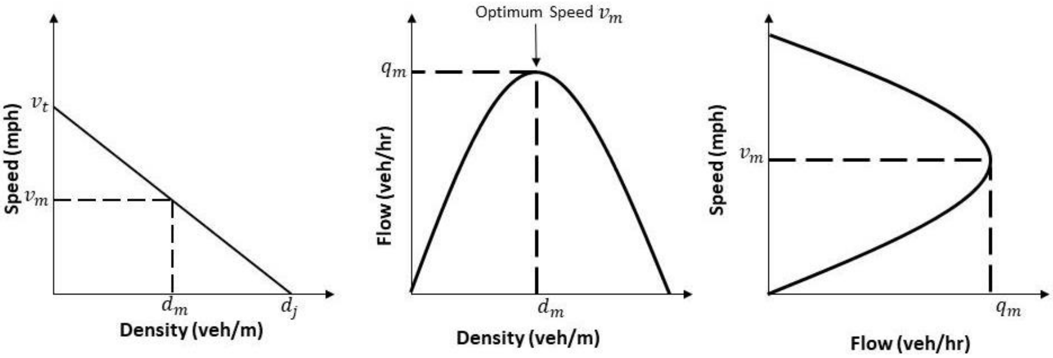


Figure 3-3: Greenshield's Model Diagrams (Greenshields et al., 1935) (a) Speed Density plot, (b) Flow Density plot, (c) Speed Flow plot

Similarly, Figure 3.3(b) presents the *flow* vs *density* plot. As can be seen, there is a hyperbolic relationship between the flow and density, implying that the flow increases with density, until it reaches a peak, at which the traffic network operates at maximum efficiency. Beyond this point, the flow begins to decay, until it reaches a minimum when the density is at maximum. From the *speed-flow* plot (Figure 3.3(c)), commencing from the left of the plot, when the flow (x-axis) is zero, the average speed is also zero, and causes an increase in the flow. However, as the flow increases, the speed reduces, and the highest point (rightmost) represents one where the network is operating at capacity.

3.4 Traffic Data Collection Methods

Numerous traffic data collection methods exist today, reflecting a trend that is likely to increase in the near future. Bennett, Solminihac, and Chamorro (2006) present an overview of the various traffic data collection methods. The report discusses data collection techniques for three different types of data: volume, vehicle classification, and truck weights, although other traffic characteristics such as vehicle speed, occupancies, journey details, etc. can be measured. A contrast is made between the two major categories of sensors used in traffic data collection equipment: intrusive and non-intrusive sensors. While intrusive sensors are those that involve placement of the sensors on or in the road to be monitored (example, inductive loop detectors, pneumatic rubber road tubes, piezo-electric detectors, etc.), non-intrusive sensors are not directly interfering with the traffic flow. Examples of non-intrusive sensors are passive acoustic sensors, cameras, infra-red radars, etc. Besides these data sources, in-vehicle data systems known as Floating Car Data (FCD) can serve as a source of traffic data, for instance, vehicles that have Global Positioning System (GPS) installed, or mobile (cellular) phones. Each of these various data collection methods has its own benefits and demerits, which makes their use more appropriate or not in particular traffic variable measurement. Table 3-1 summarises the various traffic data collection sensor types and their properties. The individual sensor types are analysed by the traffic flow characteristics (parameters) they measure – volume, speed, and occupancy.

Table 3-1: Traffic Sensor types and their properties

Sensor Type	Traffic Volume	Vehicle speed	Occupancy
Inductive Loop Detector	✓	✓	✓
Microwave Radar	✓	✓	✓
Infrared	✓	✓	X
Bluetooth	✓	✓	✓
Floating Car Data	X	✓	X
CCTV	✓	✓	✓

3.4.1 Inductive Loop Detectors (ILD)

The inductive loop detector is about the most common intrusive traffic data collection method used in many traffic networks today. It basically observes vehicles utilising the principle of induction. Banks (2002) explains the working principle of the inductive loop detector. The system has three major parts: the *wire loop*, which is mounted on the roadway, the *detector device*, and the *computer system*. It uses an insulated wire loop, which acts as the inductive component of the circuit, buried in the road such that it forms a closed oscillatory circuit. When a vehicle passes over the detection zone of the sensor, it affects the magnetic field in the loop, thereby reducing the inductance of the circuit. In addition to the circuit, a loop detector unit has the primary function of energising and monitoring the loop. This detector unit responds to the decrease in circuit inductance, which then sends a signal to the controller unit (See Figure 3-4). The Inductive Loop Device is majorly used to detect vehicle passage and presence.

There are two types of ILDs, *single-loop* and *dual loop* detectors. Single loop detectors are usually applied in combination with traffic lights to control traffic, while dual loop detectors have the advantage of being able to compute the vehicle speed and the direction of travel of the vehicle. In a dual loop detector, since the distance between the two loops is constant, the detection of vehicle speed is easy to achieve unlike the single loop detector, where an estimation model (obtained from historical data analysis) needs to be applied.

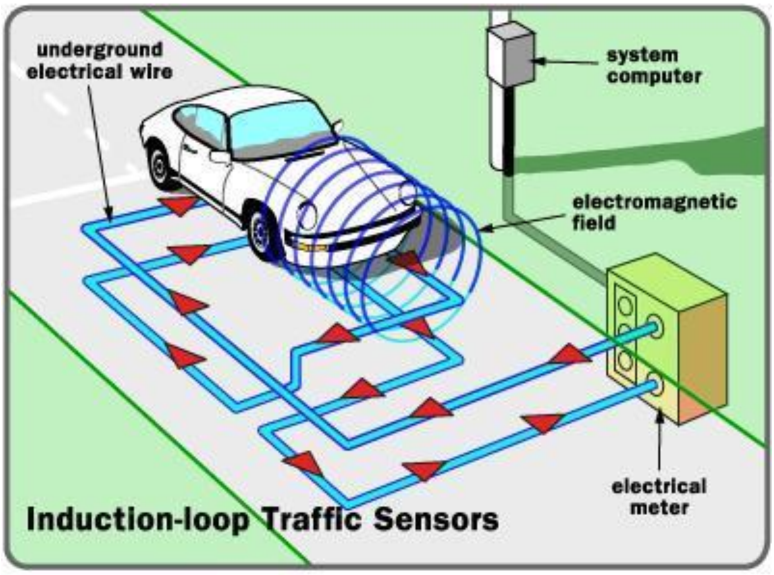


Figure 3-4: Working of the ILD (Source: www.iwatchsystems.com)

3.4.2 Microwave Radar

Microwave radar technology can detect average vehicle speed, traffic volume counts, and can also perform vehicle classification. Radar is an acronym for ***RA***dio ***D***etection ***A***nd ***R***anging and is applied by focusing high-frequency radio waves, which are transmitted by a source device, to the road segment that is to be monitored, and the calculation of the time delay for the returning signal, thereby enabling the calculation of the distance of the vehicle. The distance is then used to calculate the vehicle speed. The advantage of radar detectors is that they are not sensitive to weather disruptions and can provide day and night operation (Taylor and Bonsall, 2017). The distance can be computed with the formula as shown below:

$$R = \frac{ct}{2} \tag{3-4}$$

Where c is the speed of light $(3 \times 10^8\ m/s)$, t is the measured time in seconds and R is the distance between the detector and the object.

3.4.3 Bluetooth

Bluetooth technology is a wireless data exchange technology that uses the 2.4GHz radio frequency band and can be used to exchange data over short distances. Class 1 Bluetooth transceivers typically have a transmission range of 100 meters, while class 2 devices can transmit up to about 10 meters. Class 3 chips can exchange data over a distance of about 1meter.

Bluetooth technology used for traffic data collection is based on the measurement of travel time. Typically, the detectors are installed on the roadway, and the distance between them is known. All Bluetooth devices detected by the sensors are stored with a timestamp, and since each Bluetooth device has a unique Media Access Control (MAC) address, it is possible to quickly identify addresses that have appeared in more than one sensor. In this way, it is easy to determine the time taken to get from the location of one detector to another, and this can be used to compute the travel time. A simple calculation can also be done on the travel time so that the vehicle speed is obtained. According to Abbas et al. (2013), a major limitation of this data collection method is the fact that only vehicles installed with Bluetooth devices can be detected. Although this method is useful in measuring the travel time and speed, it becomes less effective in providing accurate information about traffic flow and density, since not all vehicles are installed with Bluetooth devices (Bhaskar and Chung, 2013).

3.4.4 Floating Car Data (FCD)

As mentioned above, it is possible to get data collected from vehicles in the form of real-time data or offline data. This type of data is known as floating car data (FCD). The particular advantage of this type of data is that it provides the exact route details followed by the car in addition to travel speed, and FCD provides the possibility to provide real-time data. One type of data collection method for the FCD can get data based on a GPS module installed in the car. Another way of getting FCD is by mobile phones since there is a high likelihood of one or more of the vehicle occupants to possess a cellular phone. The advantage of this mode is that the vehicle does not need to be installed with a GPS module, as the GPS locations can easily be gotten from the cellular phone by the triangulation between cell antennae the mobile phone will connect to as they traverse the journey route.

3.5 Short-term traffic prediction methods

As stated in Section 2.3, traffic prediction is categorised as two distinct types, namely *model-driven* and *data-driven* prediction methods (Barros et al., 2015). Model-driven methods rely on simulations of the traffic system, including the traffic flow, signal, intersections, and signal control plans in order to forecast future states of the traffic network under consideration. This class of traffic prediction typically adopts macroscopic traffic prediction models. Macroscopic traffic prediction models focus on the prediction of a traffic stream by seeing traffic flow as an analogy of fluid and gas dynamics (Van-Lint et al., 2005), as opposed to microscopic traffic prediction models that tend to focus on the individual vehicle trajectories affected by driver behaviour within the network (Ben-Akiva et al., 1998).

According to Barros et al. (2015), model-driven predictive models allow the inclusion and visualisation of traffic control factors such as ramp metering, routing, traffic light control, etc. within the prediction process, thereby providing a holistic view of the traffic network. However, these models are disadvantageous when the computational complexities for the operationalisation of such systems are taken into account (Barros et al., 2015). Similarly, the predictive accuracy obtained from model-driven prediction models depend on the quality of the estimated traffic demand via the origin-destination (O-D) matrices. Detailed discussions about the comparison of model-based and data-driven prediction models are outside the scope of this research study but can be found in (Algers et al., 1997; Barros et al., 2015; Hinsbergen and Sanders, 2007).

However, in this thesis, the focus is on the use of data-driven prediction methods for traffic prediction. The main advantage of data-driven traffic prediction is the (relatively) lower computational requirement in comparison to model-driven predictive models. Another advantage is the tendency to yield more accurate predictions due to the ability of data-driven predictive models to quickly adapt to traffic data (Barros et al., 2015). Data-driven traffic prediction is typically categorised as *short*, *medium*, and *long-term*. Short-term prediction, which usually sees prediction horizons of 5-min to 1-hr predictive windows, is the most common prediction regime in studies. Short-term traffic prediction techniques are broadly categorised into (i) *parametric* approaches, (ii) *nonparametric* approaches, and (iii) *hybrid* approaches. Details of the respective categories are presented in the subsequent sections.

3.5.1 Parametric approaches to short-term traffic prediction

According to Russell and Norvig (2012), a parametric model is one that estimates data using a set of parameters of fixed size by simplifying the input function to a known form. This implies that once a parametric model has been trained, no matter how much more new data is introduced to the model, the number of parameters as well as the function remains the same. Parametric functions are sometimes referred to as *model-based* prediction methods due to the fact that the model structure is predetermined using computed model parameters on empirical data. This subsection presents discussions about some parametric prediction models in use for traffic prediction.

For data-driven traffic parameter prediction, the input data is typically a time-series. In order to simplify the definition, let each traffic input time series be univariate. Thus, let $[\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3 \dots, \mathrm{x}_\mathrm{t}]$ denote a historical time series input vector representing observations of the given traffic speed. Also, let the observation data stream be collected in batches of b most recent observations within each time step. Thirdly, let $\mathrm{y}_{\mathrm{t}+1}$ be mapped from previous values with independent and identically distributed (*i.i.d*) Gaussian white noise represented as $\in_{\mathrm{t}+1}$. In this case, the function $f(x)$ is unknown. The fundamental objective of every prediction model is to approximate the function $f(x)$ to a known form.

In this section, detailed mathematical definitions of some traffic prediction models are presented. For the remainder of this section, let the following be assumed:

$$\boldsymbol{X}_t = [x_{t-1}, x_{t-2}, \dots, x_t] \in \mathbb{R}^w \tag{3-5}$$

$$\boldsymbol{Y}_t = [y_{t+1}, y_{t+2}, \dots, y_{t+b}] \in \mathbb{R}^b \tag{3-6}$$

Each training data pair, (X, Y) represents the input and target values for the training data, such that each input observation incorporates a batch of w historical observations, with the target variable represented as Y. Similarly, let the following set of equations represent the dimensions or variables of X and Y, respectively.

Auto-Regressive Integrated Moving Average (ARIMA) model

ARIMA is one of the widely-adopted parametric time series prediction method for time series (traffic) prediction. The model implements a statistical methodology to identify patterns from historical observations of time series input data. The forecast is made using a combination of historical and current observations. ARIMA is the most general class of models for performing time series forecasting, by making the time series stationary either by performing differencing, or the introduction of non-linear transformations such as logging (Ho et al., 2002). ARIMA forecasting equation for a stationary time series can be defined as a linear equation in which the predictors are lags (i.e. previous time values) of the dependent variable and/or lags of the forecast errors (known as residuals). If the function consists of only lagged values, then this is an autoregressive (or AR) model. However, if some of the predictors are the lagged values of the residuals, then the model is an Auto-Regressive Moving Average (ARMA) model.

Mathematically, ARIMA models can be described using equation (3-7):

$$\hat{Y}_t = \mu + \phi_1 y_{t-1} + \ \ldots + \phi_p y_{t-p} - \theta_1 e_{t-1} \ldots - \theta_q e_{t-q} \tag{3-7}$$

where:

y	represents a general time series
$\hat{\mathrm{Y}}_{\mathrm{t}}$	denotes the forecast of the time series at time t
$\mathrm{y}_{\mathrm{t}-1} \ldots \mathrm{y}_{\mathrm{t}-\mathrm{p}}$	represents the previous p values of the time series, which forms the auto-regression term
$\phi_1 \ldots \phi_{\mathrm{p}}$	slope coefficients to be determined by model fitting
$\mathrm{e}_{\mathrm{t}} \ldots \mathrm{e}_{\mathrm{t}-\mathrm{q}}$	zero mean white noise
$\theta_1 \ldots \theta_{\mathrm{t}-\mathrm{q}}$	Moving average coefficients to be determined by model fitting

p	number of auto-regression terms
d	number of difference terms
q	number of moving-average terms
μ	constant representing average difference in Y

The variables in equation (3-11), p and q are integers greater than or equal to zero and represent the autoregressive and moving average components, respectively. It is a requirement for the input time series to be stationary for the successful application of ARIMA (p, d, q) model. For this reason, differencing is often applied to induce stationarity of the dataset, which involves the consecutive differences between the observations. Thus, the third parameter, the difference (d) such that if $d = 0; y_t = y_t$ and if $d = 1; y_t - y_{t-1}$, etc.

The main assumption made by this model is that the input data is represented by a stationary process, implying the stationarity of the mean, variance, and auto-correlation. Many studies have proposed ARIMA-based models for traffic prediction (Hillmer and Tiao, 1982; Ho, Xie and Goh, 2002; Moayedi and Masnadi-Shirazi, 2008; Kumar and Vanajakshi, 2015; Yu *et al.*, 2015). Furthermore, a variant of the generic ARIMA model is the Seasonal ARIMA or SARIMA model, which can accommodate seasonal time series (the majority of traffic time series are seasonal). The variation includes the inclusion of seasonal terms, thereby producing the SARIMA(*p*,*d*,*q*)(*P*,*D*,*Q*) model. Studies that have used the SARIMA model for prediction conclude about its superiority in capturing the seasonal components of traffic data, especially when compared to its ARIMA counterpart (Kumar and Vanajakshi, 2015; Williams et al., 1998; Yu et al., 2015).

The main advantage of ARIMA-based models is their relative ease of implementation. Furthermore, ARIMA models are built on well-estimated statistical and theoretical backgrounds, thereby making them easy to interpret and reproduce. Another advantage of this class of models is related to their computational efficiency due to simple model structure. However, a significant drawback of the use of ARIMA for traffic prediction is the tendency of such models to focus on the means, thereby missing the extreme values, which are particularly prevalent within traffic datasets (Smith et al., 2002). Traffic datasets tend to exhibit peaks, especially at rush hours, as well as the influx of rapid fluctuations during incidents or accidents. Therefore, ARIMA based models became weak when applied to traffic forecasting. Another disadvantage of the application of ARIMA for traffic prediction concerns the determination of the optimal model. The process of determining the optimal model structure (*p, d, q* parameters) in ARIMA models is more of an art than a science, which makes it challenging to implement.

<u>Linear Regression (LR)</u>

Linear regression is a method used to model the relationship between two or more variables. Linear regression models make the assumption that there exists a linear correlation between the time lag and the observed for all readings. Fitting a linear regression model to time series data is solved by determining the line that minimises the *sum of squares* of the residuals (prediction error) or deviation from the ground truth. This is known as computing the least squares regression line or line of best fit. In terms of traffic prediction, studies that have applied linear regression models include (Davis and Nihan, 2007; Kwon et al., 2000; Oh et al., 2015; Smith et al., 2002; Smith and Demetsky, 1997; Sun et al., 2003; Tebaldi et al., 2002). A linear regression line is defined by equation (3-8) below:

$$y = \alpha + \beta x \tag{3-8}$$

where x is the explanatory variable and y is the dependent or predictor variable and α and β are the intercept and slope respectively.

<u>Kalman Filter (KF)</u>

The Kalman filter algorithm is a widely used predictive algorithm in short-term traffic prediction. It was first introduced by Kalman in Kalman (1960). The Kalman filter is an optimal estimator that infers parameters of interest from uncertain and inaccurate observations. It is an optimal recursive algorithm, which makes it very useful in real-time online traffic prediction.

The Kalman filter is a parametric prediction algorithm that continually updates its prediction for a given variable of interest based on explicit models that measure the physical process of the system. Therefore, the KF algorithm successively updates the parameters while the prediction is going on, and as new input data sources are fed in. The theory basically comprises two sets of equations: the process and measurement equations represented by equations (3-9) and (3-10) respectively.

$$x_t = F_{t,t-1}x_{t-1} + w_t, x_t, x_{t-1} \in \mathbb{R}^n \tag{3-9}$$

$$y_t = H_t x_t + v_t, y_t \in \mathbb{R}^m \tag{3-10}$$

where $\mathbb{R}^{\mathrm{n}}$ and $\mathbb{R}^{\mathrm{m}}$ represent n and m dimensional real variable domains respectively, $\mathrm{x_t}$ and $\mathrm{x_{t-1}}$ are state vectors at steps t and $\mathrm{t}-1$ respectively. $\mathrm{y_t}$ is the observed measurement at time step t. The transition matrix is represented as $\mathrm{F_{t,t-1}}$, while vectors $\mathrm{w_t}$ and $\mathrm{v_t}$ represent the process and measurement noises respectively. The generic algorithm for a loop or iteration of the KF is depicted in Figure 3-5 below, while the summary of the recursive KF algorithm is presented in Table 3-2.

The Kalman filter uses a *predictor-corrector* algorithm to estimate the x_b such that an initial tentative estimate is calculated, which is subsequently refined or filtered using the measurement or actual value y_b. The method has been used within academic studies for the purpose of traffic prediction, such as in (J. Guo et al., 2014; Julier and Uhlmann, 1997; Okutani and Stephanedes, 1984; Qiao et al., 2013) and is very efficient in real-time prediction due to its recursive nature and computational efficiency.

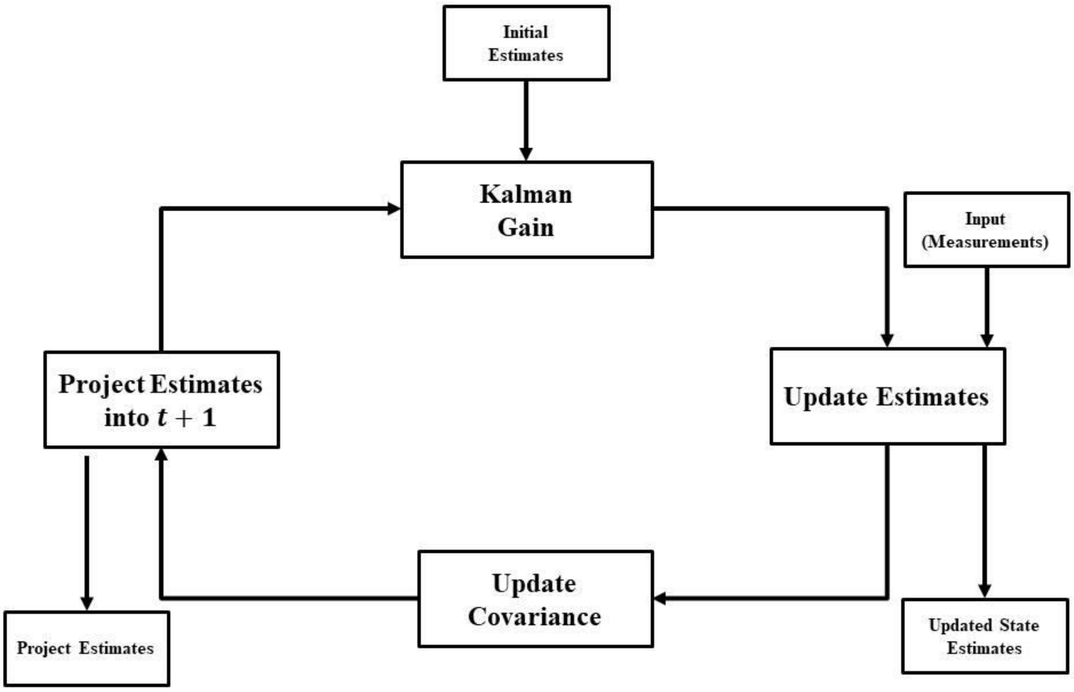


Figure 3-5: KF Algorithm [adapted from: (Thacker and AJ Lacey, 1996)]

The KF is a multivariate state-space model that can allow the use of both traffic and non-traffic input data, which has been proven to improve prediction accuracy. In addition to that, the recursive nature of the KF makes it very good for real-time online traffic prediction. However, a major drawback of the KF is the reliance on the assumption that the system and measurement noises are white and Gaussian distributed, respectively, which leads to significant limitations in the practical use of the algorithm (Barros et al., 2015).

Table 3-2: Summary of Equations for KF Iterative Algorithm

No.	Description	Equation
1	Kalman Gain	$K_t = P'_t H^T (H P'_t H^T + R)^{-1}$
2	Update Estimates	$\hat{x}_t = \hat{x}'_t + K_t(y_t - H\hat{x}'_t)$
3	Update Covariance	$P_t = (I - K_t H P'_t)$
4	Project into $t+1$	$\hat{x}_{t+1} = A\hat{x}_t$ $P_{t+1} = A P_t A^T + Q$

3.5.2 Nonparametric approaches to traffic prediction

In this class of prediction models, the model structure, as well as the parameters, are not predetermined or fixed. Therefore, algorithms that learn from the data, or do not make strong assumptions about the mapping function are referred to as *nonparametric* or non-linear models (Russel and Norvig, 2012). The main advantage of such models is their ability to learn from any type of dataset provided. Nonparametric models tend to select the function that best fits the training dataset, meaning they can fit a large number of functions.

k-Nearest Neighbour (k-NN)

This is a widely used nonparametric prediction model for traffic prediction (Gong and Wang, 2003; Qiao et al., 2013; Zhang et al., 2013). It is a 'lazy' learning method (i.e. does not require any model prior construction) and is referred to as an *instance-based* learning model (Mitchell, 2006). The model thrives in being a simplistic machine learning model, capable of making accurate predictions without any defined model structure. It is sometimes referred to as undergoing a *lazy* learning process because, during the training, all samples of the training dataset are stored (Mitchell, 2006). The underlying assumption of the algorithm is that if k most similar observations in a feature space are categorised, then the observed sample will likely belong to this category. More specifically, the model searches for the matching *nearest neighbours* in historical and current observations, based on specific parameters and similarities. Then, the searched nearest neighbours are used for prediction. The generic algorithmic flow for a typical *k*-NN based prediction method is represented in Figure 3-6. As the figure reveals, the approach searches for a set of nearest neighbours (i.e. from the historical input data observations similar to the current observation). The nearest neighbours, therefore, serve as inputs to the predictive step, which is used to calculate the predicted value.

The main parameters of the model are (i) distance metric, (ii) the number of nearest neighbour *k*, and (iii) predictive algorithm.

i. *Distance Metric:* this is used for the determination of the distance between the feature vector (i.e. observations) and the historical observations. The most common distance metrics are *Euclidean* distance and the *Minkowski* distance metric (Van-de-Geer, 1995). Table 3-3 summarises the equations of the abovementioned distance metrics.

Table 3-3: Distance metric equations

No.	Description	Equation
1	Euclidean Distance	$d_{(p,q)} = \sqrt{(q_1 - p_1)^2 + \ (q_2 - p_2)^2 + \cdots (q_n - p_n)^2}$
2	Minkowski Distance	$d_{(p,q)} = \left(\sum_{i=1}^{n} \lvert p_i - q_i \rvert^s\right)^{\frac{1}{s}}$

		for $s \geq 1$

ii. *Number of k nearest neighbours:* this determines the number of nearest neighbours that will be chosen from the historical data. Accordingly, if $k = 5$, then the top 5 historical observations having the closest observations to the input vector (current observation) will be used during the prediction process.

iii. *Predictive function:* this is the driving function for the prediction method. The prediction algorithm describes how the searched nearest neighbour groups are used in the prediction of the state vector in the next time step.

This learning algorithm is advantageous in being simple to implement, having a fast training process (i.e. simply computing Euclidean or Minkowski distance between training samples), as well as the fact that it can learn complex, non-linear functions. However, it is severely disadvantaged in the vast amount of memory it occupies (since it stores the entire training set), resulting in slower prediction time. Also, the algorithm can easily be swayed or ‘fooled’ by irrelevant/noise attributes of training data values.

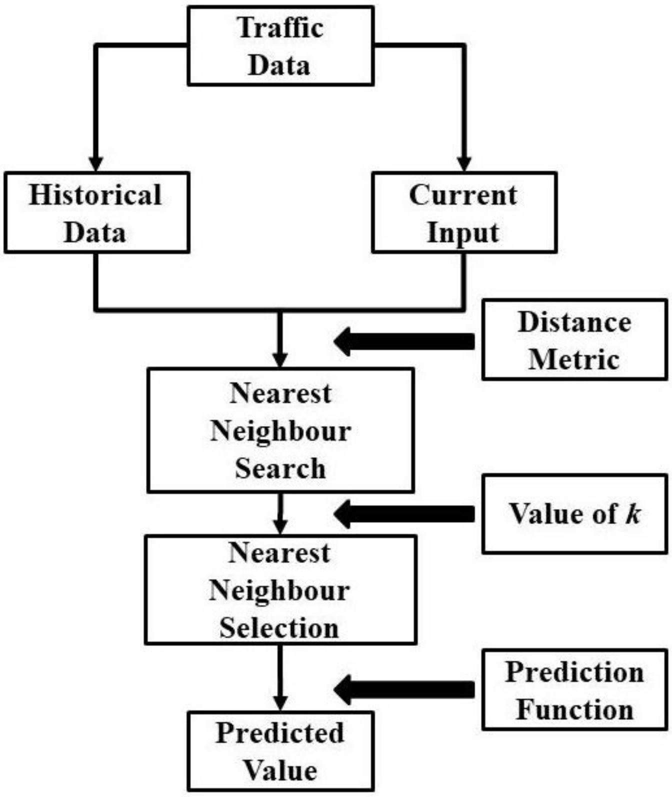


Figure 3-6: Algorithmic Structure of k-NN method

Support Vector Regression (SVR)

Support Vector Regression (SVR) was proposed in Cortes and Vapnik (1995), on the basis of statistical learning theory. The model implements an objective function using a *structural risk minimisation* equation adopted from computational learning theory. The base regression model is defined as:

$$f(\boldsymbol{X}_t) = \phi(\boldsymbol{X}_t)^T w + c \qquad (3\text{-}11)$$

where ϕ represents a user-defined function that maps the traffic speed in the memory window to the features within the higher dimension, and c represents a bias.

Suppose we have a training dataset $\{(\mathrm{x}_1, \mathrm{y}_1), (\mathrm{x}_2, \mathrm{y}_2), \ldots, (\mathrm{x}_\mathrm{n}, \mathrm{y}_\mathrm{n})$, then the ε-SV regression is the computation of a function f(x) that has a maximum deviation of ε from the actual values of y_i for the entire training set (Cortes and Vapnik, 1995). In other words, all other errors are not important so long as the deviation is not greater than ε. The objective function for a SVR model at time $\hat{\mathrm{t}} + \mathrm{i}$ is defined using the structural risk minimisation framework (Cortes and Vapnik, 1995) represented as:

$$\min_{w,c} \sum_{t=t-\widehat{b-\hat{T}}+1}^{\hat{t}-b} L_{\delta}(f(\boldsymbol{X}_t - y_{t+k}) + \lambda||w||^2 \qquad (3\text{-}12)$$

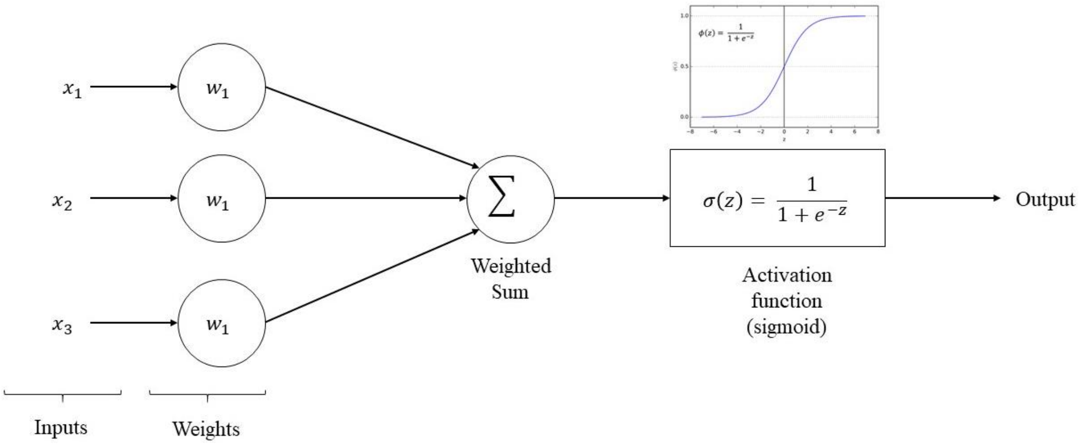
where $\mathrm{L}_{\delta}(\mathrm{f}(\mathrm{X}_{\mathrm{t}} - \mathrm{y}_{\mathrm{t+k}})$ represents a term known as empirical loss, which needs to be minimised via the training dataset and $\lambda||\mathrm{w}||^2$ is a control measure used to prevent overfitting. The problem can be written as a convex optimisation problem.

Support vector machines (SVM) are efficient and dynamic predictive algorithms based on artificial intelligence and have been extensively studied in the last decades. SVM maps data into a feature space using a non-linear relationship and then performs linear regression in this space. This class of nonparametric algorithms have been proven to outperform their regression-based and time series counterparts. For instance, (Wu et al., 2004) presented an SVM-based model to predict traffic and show that the model can perform satisfactorily for traffic data analysis. Similarly, (Ma and Perkins, 2014) predicted bus arrival time in China using SVM. (Hsu and Lin, 2002) perform support vector regression for travel time prediction and compared the results to benchmark models and prove that SVM outperformed the other models.

<u>Artificial Neural Networks (ANN)</u>

Artificial Neural Networks (ANN) are considered as complex predictive models, due to their inherent ability to deal with multi-dimensional data, non-linearity, and adept learning ability and generalisation (Goodfellow and Bengio, 2015). The basic framework of a neural network comprises four atomic elements, namely: (i) nodes, (ii) connection, (iii) layers, and (iv) transfer/activation function. The nodes within a neural network represent the neurons, which are simple processing units. The atomic structure for a neural network is the multilayer perceptron (MLP). ANN models reduce error by employing *optimising* algorithms, such as backpropagation (Hecht-Nielsen, 1988; Rumelhart et al., 1988).

A set of nodes are connected by weighted connections, which represent the connecting interactions. The optimal weights of each connection between a set of layers are calculated during each backward pass of a training dataset, which is also used for weight optimisation using the derivatives obtained from the input and predicted values of the training data. The layers represent the network topology, representing neurons interconnected. Within the network, the transfer function or activation function represents the transfer function or state of each neuron. The basic process in a single neuron is presented in Figure 3-7.



$\phi(z) = \frac{1}{1+e^{-z}}$
x_1
w_1
x_2
w_1
x_3
w_1
$\sum$
Weighted Sum
$\sigma(z) = \dfrac{1}{1+e^{-z}}$
Activation function (sigmoid)
Output
Inputs
Weights

Figure 3-7: Single Neuron Process for Neural Network

A particular variation of neural networks is the feed-forward neural network. This is widely used within traffic prediction, and the generic architecture is depicted in Figure 3-8. As the figure shows, the elementary model structure comprises three layers – the input, hidden, and output layers respectively. In Feedforward Neural Networks

(FFNN), each individual neuron is interconnected to the output of each unit within the next layer.

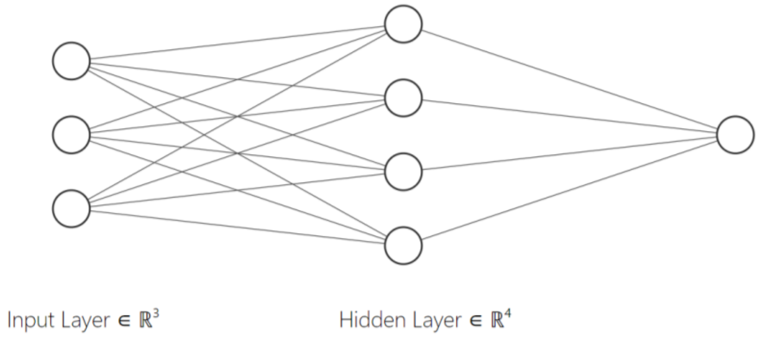


Figure 3-8: Generic Architecture of Feed-Forward Networks

Research has revealed that neural network (NN) models consistently outperform their ARIMA-based counterparts by adequately capturing the dynamic flow observed in traffic prediction (Lana et al., 2018; Vlahogianni et al., 2004). The main advantage of neural networks is their adept learning ability in capturing traffic patterns from large historical traffic datasets. However, a demerit of this class of models is the extensive model training and retraining time. Furthermore, neural networks are susceptible to noisy datasets, which makes it quite inconsistent for traffic datasets.

However, although these models made accurate traffic flow predictions, they used shallow learning networks and were still somewhat weak and vulnerable in many aspects. For instance, the BPNNs had the problem of the diminishing gradient (Hochreiter, 1998), which is a situation where the model can shut down in instances where the gradient or the error function (used to update the model to reflect the actual outcome) becomes too small for the model to carry on 'learning' (Hochreiter et al., 2001). The aforementioned ANN-based models also neglect the temporal or sequential element of time-series data inputs, which led to the development of a variant that considers the time dependency of time series data. This class of neural networks are referred to as *recurrent neural networks* (RNNs) (Rodriguez et al., 1999). This class of models typically adopt deep learning approaches.

Deep Learning (DL) is an advanced machine learning technology that is made of stacks of multiple NN processing layers capable of learning data via multiple levels of data abstraction (LeCun and Bengio, 1995). DL has already been successfully applied in classification, regression-based, natural language processing (NLP), computer vision, and many other applications too broad for the scope of this study. According to Hochreiter et al. (2001), traditional RNNs have issues, yet to be addressed. Firstly, traditional RNNs are unable to learn from time series having long time lags, which is in reality very common to traffic datasets. Secondly, the models majorly rely on predetermined time lags prior to learning the temporal sequence, but this is, however, a complicated process to automatically identify the optimal time window size.

Long Short-Term Memory Neural Networks (LSTM)

The demerits of traditional RNNs led to the development of Long Short-Term Memory Neural Networks (LSTM), originally proposed by German engineers – Hochreiter and Schmidhuber (1997) – having the objectives of modelling long-term time dependencies of time series data by determining the optimal time lag for the problem. The basic architecture of the LSTM having one memory block is depicted in Figure 3-9. As can be seen, the model is built around the memory block (instead of the neuron node in traditional ANNs). It can be observed from Figure 3-9 that each memory block contains input, output, and forget gates, which respectively can be analogous to write, read, and reset functions on each cell. The multiplicative gates allow the model to store information over long periods of time, thereby eliminating the vanishing gradient problem commonly observed in traditional neural network models (Hochreiter, 1998).

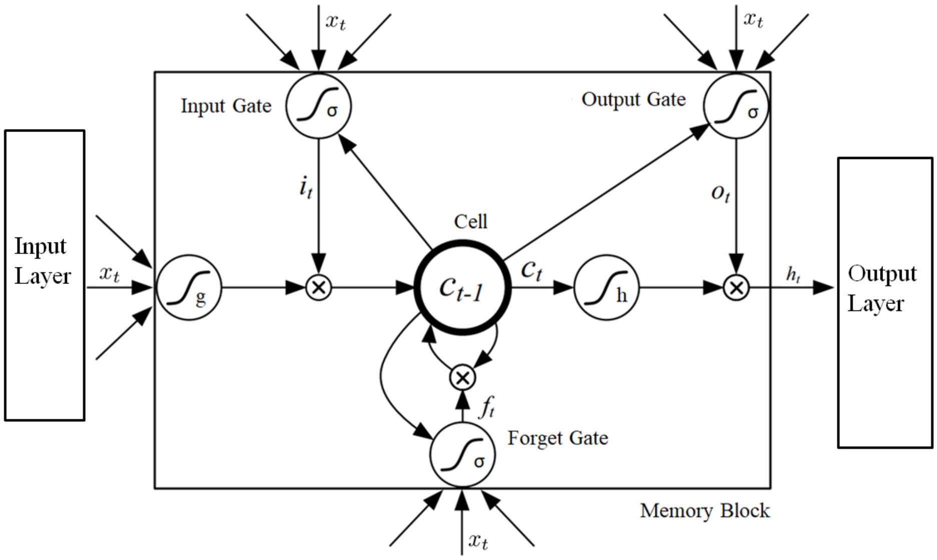


Figure 3-9: LSTM RNN model architecture having one memory block

Each block contains *input*, *output*, and *forget* gates, which respectively can be analogous to write, read, and reset functions on each cell. The multiplicative gates allow the model to store information over long periods of time, thereby eliminating the vanishing gradient problem. Consider a univariate time-series input sequence denoted as $x(t) = \{x_1, x_2, x_3 \dots\ x_t\}$, and a corresponding output sequence of $y(t) = \{y_1, y_2, y_3, \dots y_k\}$, where k is the prediction horizon and t refers to the length of the input timeseries. The LSTM computes the predicted output in the next time step using the historical information supplied, without being told how many backward time steps should be traced. The following set of equations is performed by the model and enables the model to predict the output variable:

$$\mathrm{i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i)} \tag{3-13}$$

$$\mathrm{f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f)} \tag{3-14}$$

$$\mathrm{c_t = f_tc_{t-1} + i_tg(W_{xc}x_t + W_{hc}h_{t-1} + b_c)} \tag{3-15}$$

$$\mathrm{o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o)} \tag{3-16}$$

$$\mathrm{h_t = o_th(c_t)} \tag{3-17}$$

Where *W* and *b* represent the weight matrix and bias vector respectively and σ(.) denotes a standard logistic sigmoid function defined as:

$$\mathrm{\sigma(x)} = \frac{1}{1+\mathrm{e}^{-\mathrm{x}}} \tag{3-18}$$

$$\mathrm{g(x)} = \frac{4}{1+\mathrm{e}^{-\mathrm{x}}} - 2 \tag{3-19}$$

$$\mathrm{h(x)} = \frac{2}{1+\mathrm{e}^{-\mathrm{x}}} - 1 \tag{3-20}$$

Where *g*(.) and *h*(.) are the respective transformations of the sigmoid function above. The variables *i,f,o,* and *c* are the input gate, forget gate, output gate, and cell activation vector respectively. Studies have focused on using LSTM for traffic prediction, for instance (Ma et al., 2015) present a LSTM model for traffic speed prediction using microwave sensor data. They use a prediction horizon of 2-min and apply a traffic dataset spanning 1-month, as well as compare the results to other nonparametric algorithms (SVM, Kalman Filter, and ARIMA), and conclude about the superiority in performance of the LSTM model. Similarly, Tian and Pan (2015) present a LSTM model for predicting traffic flow. The comparison against benchmark models like SVM, feed-forward neural networks (FFNN), and stacked auto-encoders (SAE) revealed that the proposed model achieved greater accuracy and generalisation.

In terms of data source fusion, Jia et al. (2017a) present an LSTM and deep belief network (DBN) model to predict short-term traffic speed using traffic and rainfall data in Beijing, China. The results of the experiment revealed that fusing weather and traffic data sources improved the prediction performance of the models, and that the LSTM outperforms the DBN in capturing time-series characteristics of traffic speed data. Also, Jia, Wu, and Xu (2017b) investigated the impact of fusing weather data with traffic data for predicting traffic flow. The study incorporated a DBN model, and the results obtained showed that the combination of data sources yielded superior prediction accuracy.

3.6 Summary

This chapter has presented technical background about the key concepts presented in this thesis. It began with a background to the domain of traffic control and management, including the fundamentals of traffic flow theory, relationships between traffic parameters/characteristics, and traffic stream models. In addition, Section 3.4 presented an overview of some traffic data collection methods, including inductive loops, microwave, Bluetooth, and floating car data. Brief technical background about some short-term TPA predictive models was presented in Section 3.5.

The next chapter will present the research methodology that was adopted towards the development of a traffic predictive analytics guidance system. It will also highlight the epistemological, research design, outcomes, and expectations supporting the realisation of this study.

Chapter 4 Research Design

4.1 Introduction

The previous two chapters presented comprehensive reviews of key concepts in TPA and characterised TPA, including the key parameters and challenges encountered in TPA (Chapter 2). In Chapter 3, background about traffic flow theory, traffic data collection methods, and short-term traffic prediction models was presented. In this chapter, the research methodology adopted for this study is presented, described, and justified. Also highlighted within the chapter are the epistemological, empirical, and methodological expectations underpinning the actualisation of this study. After a thorough review process, owing to the nature of this study, the *design science research* methodology (Hevner et al., 2004) was adopted.

The first section of this chapter introduces the concept of the design science research methodology, which focuses on the development of an artefact for the purpose of solving a real-world or organisational problem (Hevner and Chatterjee, 2010). It will then go on to present the chosen research design strategy for the study, including a justification for the choice of research strategy.

4.2 Design Science Research (DSR) Methodology

Design science research methodology is an appropriate research paradigm for information systems research (ISR) (Hevner and Chatterjee, 2010). DSR can also be seen as another analytical technique or perspective that is used to perform ISR involving the development of artefacts that aim to explain, understand, and/or improve some aspects of information systems. It, therefore, involves the creation of insightful knowledge, theories, and philosophies brought about by the development of Information Technology (IT) artefacts such as algorithms, prototype systems, or technological infrastructures, in order to analyse the use and performance of such systems.

According to Hevner et al. (2004), two broad paradigms characterise IS research: (i) *behavioural* (or natural) sciences and (ii) *design science*. The need to understand the distinction between both paradigms is essential to achieving an effective ISR process. Behavioural or natural science research mainly aims at developing and justifying theories that are capable of explaining or predicting human/social behaviour. This research paradigm can be seen as a '*reactive'* research paradigm (Hevner and Chatterjee, 2010). On the other hand, design science attempts to cover the boundaries of human and social competences by the development of contemporary and innovative IT artefacts, which can affect human, social, and organisational behaviour. In other words, the design science paradigm mainly aims at creating innovative artefacts that are able to address real-world problems. Behavioural science research can be seen as a 'problem understanding' paradigm, while design science can be seen as a 'problem-solving' paradigm (Hevner and Chatterjee, 2010).

Design science has become a relevant and critical research archetype in IT research, which has emerged as a new research direction that has yielded tremendous benefits in the field of information systems research (Denning, 1997; Hevner and Chatterjee, 2010). Design science research is rooted in the engineering and systems development discipline, otherwise known as research of the '*artificial'* (Simon, 1996). Design science, also referred to as the *research of the artificial*, is defined as a design research paradigm or body of knowledge that produces artificial/man-made artefacts or objects that aim at achieving certain predefined organisational or social objectives. According to Orlikowski and Iacono (2001), IT artefacts can be defined as '*bundles of materials and cultural properties packaged in a socially recognisable form as software/hardware*'.

Design science research provides researchers with guidelines that will enable them to develop and evaluate artefacts that are capable of solving organisational problems (Hevner and Chatterjee, 2010; March and Smith, 1995). Therefore, it is now common to find many IS researchers adopting design science research, especially the studies that aim to develop IT software, hardware, or artefacts for the purpose of solving organisational/real-world problems (Dresch et al., 2015; Prat et al., 2014; Venable et al., 2016).

4.3 Design Science Research Framework

Design science research is typically performed in iterations of defined stages. Although there is not a consensus on the definition and description of these phases or stages, there are similarities in the various attempts made by scholars to categorise or enlist the phases of design science research (Gregor and Hevner, 2013; Peffers et al., 2007). For instance, Sein et al. (2011) suggest that basic design science research comprises three stages: (i) identification of the need or problem, (ii) developing an artefact to solve the need, and (iii) evaluation of the artefact. Likewise, Hevner et al. (2004) suggest that design science research comprises three stages, iterations, or cycles: (i) relevance, (ii) design, and (iii) rigor cycles respectively. They present these stages in a framework for design science research which can be used as a template for the development, understanding, and evaluation of design science research in IT (see Figure 4-1). As can be seen from Figure 4-1, the research process begins in the relevance cycle, which comprises the initiation phase of the IS project, beginning with business requirements gathering within the environmental domain. This can pose itself in the form of opportunities, business requirements, or problems. The end product of the relevance cycle, which defines the success or failure of the project, is the actual implementation of the project. This is measured by the use of the appropriate evaluation method that measures the impact of the artefact in the environmental domain.

The second cycle, the rigour cycle, involves constant interaction between the knowledge domain (scientific body of knowledge) and the research process via the provision of existing knowledge in the form of theories or models as a contribution to the present research work. According to Hevner et al. (2004), it is the responsibility of the researcher to thoroughly research and reference the particular knowledge domain in order to ensure that the artefacts produced are contributions to the knowledge base. Therefore, within the rigour cycle, it is critical for the researcher to perform detailed and systematic research in order to produce a rich selection of appropriate theories or models relevant to the development of the research artefact.

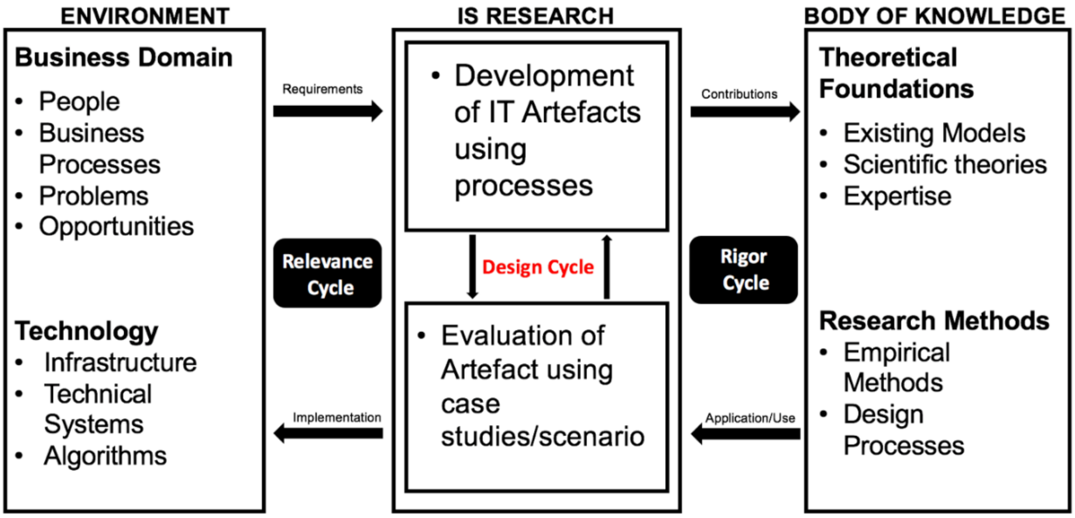


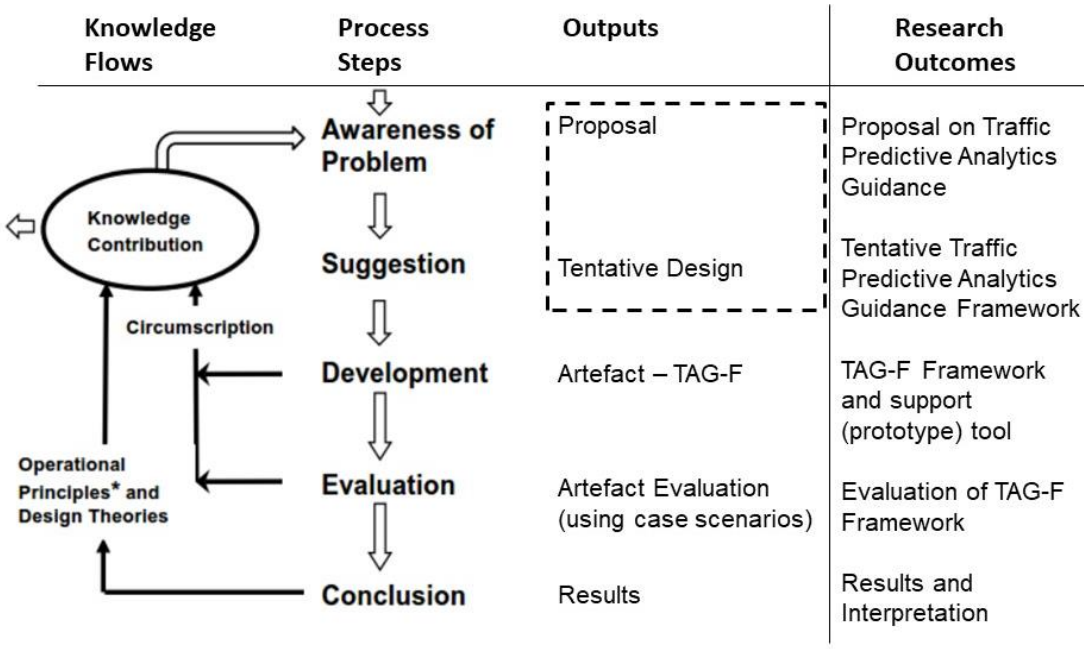
Figure 4-1: DSR Framework [Adapted from Hevner et al. (2004)]

The third cycle is the design cycle. This phase comprises the core element of the design science research process. It sits within the research domain and involves iterations of the development and evaluation of the IT artefact. As Figure 4-1 shows, the artefacts are developed using processes, theories, or models, obtained via a thorough research process from the knowledge domain. Similarly, the developed artefact provides contributions to extant theories or models within the knowledge base or body of knowledge. To put these in context, the relevance cycle identifies and defines the requirements in the form of problems or opportunities. Then, relevant existing theories and models from the body of knowledge or research domain space are drawn upon within the rigour cycle in order to develop the artefact in the design cycle, which forms the core of every design science research project (Hevner et al., 2004).

Vaishnavi and Kuechler (2004) presented a design science framework for reasoning knowledge flow in the design cycle, which is depicted in Figure 4-2. Their framework comprises five distinct stages: (i) problem awareness, (ii) suggestion, (iii) development, (iv) evaluation, and (v) conclusion. The framework suggests that design science research begins with problem awareness. The suggestion phase gives the researchers alternative problem-solving approaches, for instance, literature review, focus groups, etc. The concluding phase is where the project results are evaluated, and the contributions to the body of knowledge are applied. The Vaishnavi and Kuechler framework (see Figure 4.2), which (as explained above) enhances the Hevner et al. (2004) framework is more appropriate to the TAG-F framework introduced in this thesis as the process steps, and outputs can be directly related to the framework proposed within this study.

Table 4-1 presents a summary of the research methodology, mapping the various process steps from the Vaishnavi and Kuechler framework (Vaishnavi and Kuechler,

2004) to their respective sections within the thesis where they are discussed and addressed. As the table shows, the research work presented in this thesis begun with a problem awareness process, which involved defining/describing the process towards the development of a guidance approach for traffic data analytics.



Knowledge Flows
Process Steps
Outputs
Research Outcomes
Awareness of Problem
Suggestion
Development
Evaluation
Conclusion
Knowledge Contribution
Circumscription
Operational Principles* and Design Theories
Proposal
Tentative Design
Artefact – TAG-F
Artefact Evaluation (using case scenarios)
Results
Proposal on Traffic Predictive Analytics Guidance
Tentative Traffic Predictive Analytics Guidance Framework
TAG-F Framework and support (prototype) tool
Evaluation of TAG-F Framework
Results and Interpretation

Figure 4-2: Framework for reasoning in DSR [adapted from: (Vaishnavi and Kuechler, 2004)]

The aims and objectives of the research presented in this thesis, as well as background and motivation, can be found in Sections 1.1 - 1.4 and Chapters 2-3, respectively. The suggestion phase forms the foundation of the research process, which built on the aims and objectives of the research project in order to fully understand the process for the development of an approach capable of providing analytical guidance for traffic predictive analytics.

Table 4-1: Summary of Research Methodology Process

No.	Process Step	Output	Section(s) in Thesis
1.	Awareness of the problem	Proposal for the understanding of the research problem.	Sections 1.1 to 1.4, Section 2.3
2.	Suggestion	Tentative understanding of guidance in traffic data analytics and unrefined approach towards guidance in traffic data analytics	Sections 5.2 to 5.5, Chapter 2, and Chapter 3.
3.	Development	Framework for the provision of guidance for traffic data analytics as well as a support tool for predictive model choice	Section 5.6, Section 6.3 to 6.5
4.	Evaluation	Quantitative evaluation of the proposed framework and support tool using case scenarios from sensor-collected data in Greater Manchester, UK	Section 6.6
5.	Conclusion	Presentation of results, conclusions, and future work	Chapter 7, and Chapter 8

The next step of the research process focuses on the developmental process, with the output for this stage being the artefact – a guidance framework for traffic predictive analytics, as well as a support tool for predictive model choice. These can be found in Sections 5.6, 6.3 to 6.5. The evaluation stage involves a quantitative assessment of the proposed framework and support tool. This was achieved by three case scenarios involving traffic predictive analytics tasks with sensor collected traffic data and weather data from a road section in Greater Manchester, United Kingdom. The research process culminates with a conclusion, which involves an articulation of the research contributions (both theoretical and practical), and future work. Within this thesis, this can be found in Chapter 7, which presents major findings, contributions, and discusses future work.

4.4 TAG-F research strategy

The design of a research strategy for IS research can be challenging, as stressed in existing studies (Pozzebon and Pinsonneault, 2005; Wohlin and Aurum, 2015). Wohlin and Aurum (2015) presented a research strategy decision framework, which was adopted to articulate the strategy decision points used in describing the research design strategy for this study. The framework basically comprises three stages, which are (i) strategy, (ii) tactical, and (iii) operational stages. The decision points within the strategy stage of the framework are: research outcome, research logic, purpose, and approach. The tactical stage has decision points about the research process and research methodology, which has been described in the preceding section. The final step within the framework is the operational stage, which has the data collection and analysis methods as the decision points. The framework presented in Wohlin and Aurum (2015), which was adopted for the research design strategy in this study is shown in Figure 4-3. The graphic represents an adaptation of the framework, with the selected decision points for this research study highlighted in blue.

As the figure shows, the research strategy design process begins with a set of identified research questions and follows through using the arrow indicators, as depicted. According to Wohlin and Aurum (2015), the choice of decision points depends on the nature of the research as well as researcher experience. Details about the reasoning behind the decisions made, as well as the justifications, are elaborated in the following sub-sections.

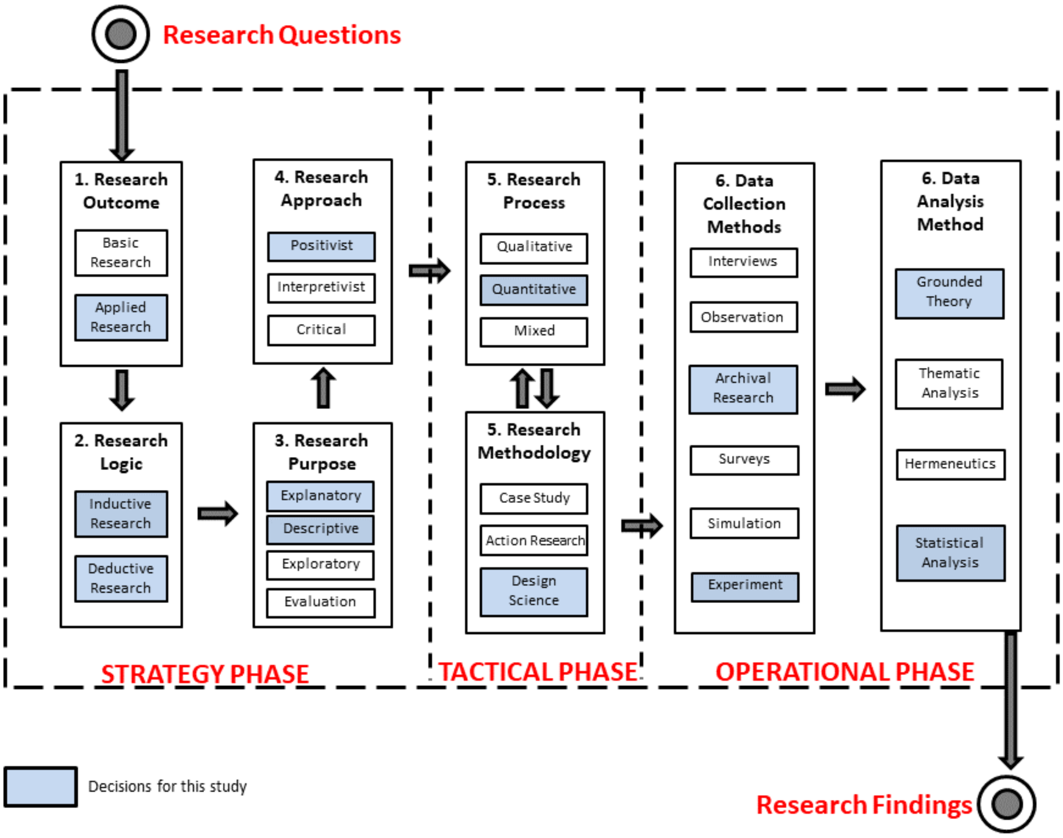


Figure 4-3: Research strategy decision-making framework [Source: Wohlin and Aurum (2015)]

4.4.1 Research Questions

From Figure 4-3, it can be seen that the research strategy design commences with identification of one or more research questions (top of the box in Figure 4-3). According to Chen and Hirschheim (2004), research questions are the major determinants for the choice of data analysis method, data collection method, and research methodology.

<u>Primary Research Question</u>

The core research question for this study was:

Can a predictive analytics guidance framework be designed to facilitate traffic data scientists in exploring the analytical decision space of TPA tasks?

The approach adopted towards answering this core research question has been elucidated within the introductory and literature review chapters, which contribute to making TPA challenging. To reiterate, these challenges include the difficulty in traffic management, which can be likened to a *wicked problem* (Churchman, 1967). Secondly, the plethora of data-driven traffic predictive models and algorithms. Thirdly, the *no free lunch* (NFL) principle (Wolpert and Macready, 1997), which states that there is no single best predictive algorithm that can be used in all situations (Brazdil, 2003).

In Chapter 2, the argument has been presented to show that the interplay of the above-listed factors combine to complicate traffic prediction, which is a vital component for effective traffic control and management using ITSs. In this context, the primary research question that formed the core composition of this study is to investigate if the above arguments are valid within an urban traffic prediction setting, supported by empirical analyses using case studies or scenarios. In order to answer the primary research question, several sub-research questions were raised, which contributed to the investigation and achievement of the primary objective of the study. These are presented in Section 1.2.

4.4.2 Research Outcomes

According to Wohlin and Aurum (2015), the research outcome can be classified either as *basic* or *applied* research. The former refers to a situation where the research is applied to a given problem for the main purpose of understanding and developing solutions to the problem. In this type of research, the main contribution of the research is the knowledge derived from the study. On the other hand, applied research refers to a research study that has the main objective of solving a given problem by the application of knowledge from existing bodies of knowledge (Collis and Hussey, 2013; Nunamaker Jr et al., 1990).

Given the nature of this study, which aims to provide a solution to an existing problem by understanding the problem, the research outcome is applied research. More specifically, the research presented in this study aims at providing a structured decision-making framework and tool for aiding traffic data scientists in data-driven traffic prediction.

4.4.3 Research Logic

According to Collis and Hussey (2013), research logic refers to the direction in which research proceeds. This can either be *inductive* research (i.e. research that is based on inductive arguments and transcends from a specific research standpoint to general research standpoint) or *deductive* (i.e. which refers to research that works from a general point to more specific one). This present study applies a mix of both inductive and deductive research logic, given that the proposed framework adopted a bottom-up approach based on the induction of knowledge from theories obtained from rigorous literature review. This required the knowledge about traffic modelling theories, traffic data collection methods, predictive models involving artificial intelligence, machine learning, statistical, and hybrid methods. On the other hand, a deductive research approach is applied in the framework evaluation using the support tool, which is a result of statistical analysis on the meta-knowledge base.

4.4.4 Research Purpose

The research purpose is either *descriptive*, *exploratory*, *explanatory*, or *evaluative* (Collis and Hussey, 2013). In descriptive research, the main aim is the provision of a description of phenomena or concept of a problem. Descriptive research questions usually begin with *"what"* or *"how"*, given that it aims to explain a given phenomenon. Exploratory research, on the other hand, is usually applied to areas of research having limited information, where the researcher aims to gather more information about the research area or problem. Explanatory research is applied where there is a need to explain the nature of relationships between the elements of a given problem area. Typical research questions in explanatory research begin with *"why",* given that this type of research aims at providing an explanation about a particular problem or phenomenon. Evaluation research mainly aims at determining the impact of concepts, methods, tools, or frameworks on the given research area of the phenomenon.

This research is aligned towards an explanatory and descriptive research because it aims to explain the nature of relationships between elements of the framework and traffic predictive analytics. In addition, the framework aims to offer guidance to traffic data scientists by describing or characterising the TPA problem space, thereby enabling the development of an appropriate solution.

4.4.5 Research Design

Research Approach

Research can be broadly classified as *ontological* (i.e. research that can be understood), *epistemological* (i.e. how the research is understood for instance, via empirical or experimental analysis), or *methodological* (i.e. the method in which the understanding or knowledge is acquired). Research approaches have the tendency to either be *positivist* (where the researcher is separate from the reality), *interpretivist* (an approach to explaining human behaviour from a given context), and *critical* research (critical evaluation of systems) (Wohlin and Aurum, 2015). This present research study follows a positivist approach, as it aims to use knowledge in a particular field to produce an artefact that can be used by traffic data scientists in order to make their work easier.

Research Methodology

Research process or methodology is often classified as *qualitative*, *quantitative*, or *mixed* methods, based on the data collection method, analysis method, and evaluation method. This research follows a quantitative research process, given that the research study involves the collection of quantitative data in the form of historic traffic parameter and weather data and codifying published articles using the TAG-F framework, applying these on given set of mathematical or statistical models or systems in order to provide explanations to relationships between the identified research parameters. In this study, the data collection, evaluation, and analysis are performed using quantitative methods, which informs our decision for the research approach adopted.

Research Method

The research method is a critical part of every research process. As previously stated, this research study adopted design science research, given that the research involved the development of an IT artefact that aims at solving an organisational problem (Hevner et al., 2004). Section 4.5 provides a detailed justification of the choice of design science research adopted for this study.

Data Collection Method

The data collection method for the evaluation process of this research was the archival data collection method. This is because the research made use of historical data archived by someone other than the researcher (i.e. relevant studies about traffic prediction and Transport for Greater Manchester and/or Centre for Atmospheric Studies at The University of Manchester). The dataset provided was in the form of archived or historical datasets retrieved from the respective databases. In addition, there is some experimental data collection performed via the LBD data collection process used to populate the base-level meta-knowledge base (see Section 6.3).

Data Analysis Method

The data analysis method adopted for this study was *quantitative* and *statistical* analysis and required technical analysis involving techniques such as statistical and mathematical modelling. For this research, the data analysis method was based on *grounded theory* (Glaser and Strauss, 2017). Grounded theory involves the analysis of data by coding, categorising, and comparing data in order to build theories and interrelated hypothesis. Similarly, a deductive data analysis method is also utilised in the framework evaluation, where the instance-based learning algorithm is used to provide model suggestions to traffic data scientists. This research also adopted grounded theory for the analysis of data collected by interpreting, coding, categorising data (obtained from existing relevant studies) about the elements within the framework in order to develop and synthesise knowledge about the traffic data analytics problem space.

4.5 Justification for research design strategy

Design science research methodology specifies a set of available perceptions and approaches for conducting information systems research (Hevner et al., 2004). Contrary to qualitative or action research design that applies extant knowledge to interpreting organisational problems, design science is used to develop knowledge that can be used to solve important (and unsolved) organisational problems by the development of innovative IT artefacts.

Although some aspects of this research can be interpreted as socio-technical (the development of a framework to be utilised by traffic data scientists), the study mainly involves the development of an IT artefact to be used towards solving an organisational problem. For instance, the framework will utilise test scenarios for the evaluation. This will involve quantitative evaluation methods, which closely align the research to design science. More importantly, the development of an IT artefact – the analytics guidance framework and supporting tool – point towards the design science research method.

4.6 Chapter Summary

This chapter detailed the decision points realised in the selection of the research design strategy. The chapter also presented the justification of the choice of the adopted research methodology, as well as the decision points towards achieving the research design. The chapter provided a detailed description of how the research design strategy was created, as well as how it mapped to the design science research methodology. The chapter serves as the building block upon which the following chapters concerned with the development of the TAG-F framework are actualised.

Chapter 5 A Traffic Predictive Analytics Guidance Framework

5.1 Introduction

The previous chapters highlighted research opportunities in the existing literature and discussed data-driven traffic prediction methods. For instance, due to the rapidly-evolving nature of the traffic prediction research area, there is a dearth of studies about meta-knowledge from traffic prediction models, which can be rationalised by such studies, for instance Vlahogianni, Golias and Karlaftis (2004); Lana et al. (2018); Vlahogianni, Karlaftis and Golias (2014); Barros, Araujo and Rossetti (2015), etc., quickly becoming superseded. This is a recurring issue due to the growing interest in short-term traffic prediction, resulting in constant research in the development of prediction algorithms.

The argument supporting the need for a structured guidance mechanism as a means of decision support for traffic data scientists towards making better and more effective traffic predictions has also been highlighted in Chapter 2 and 3. This chapter, therefore, provides detailed discussions about the proposed TAG-F framework, its dimensions (and dimension parameters), as well as the logic driving the framework process. The chapter begins with a concise definition of the concept of guidance using a non-technical illustration (see Section 2.3). In Section 5.2, an overview of the proposed TAG-F framework, providing a comprehensive discussion on the identified dimensions of the framework, as well as their respective elements.

5.2 TAG-F Framework Overview

The main hypothesis of the TAG-F framework is that a structured, well-defined description of a traffic prediction task, as well as meta-knowledge about extant traffic prediction algorithms, can be leveraged to improve upon data-driven traffic prediction. However, in order for this approach to be effective, there has to be a shared knowledge-base containing meta-knowledge about predictive algorithms. To achieve this, a *literature-based* discovery process (Bruza and Weeber, 2008) can be useful to serve as a reference point for stakeholders when undertaking traffic prediction tasks. Details about literature-based discovery are presented in Section 6.2.

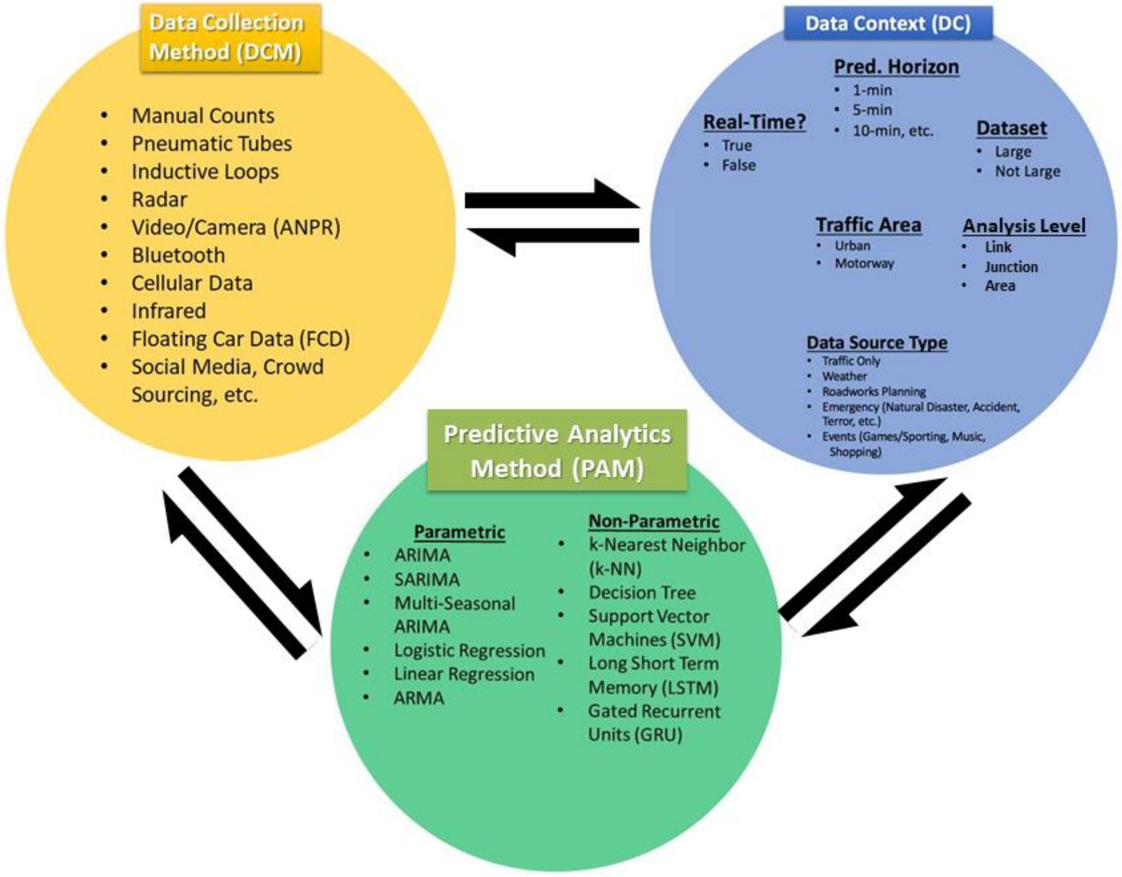


Figure 5-1: The TAG-F Framework

The thinking behind the design of TAG-F is pivoted on structured analytical solution design guidance, decision support, and shared knowledge. The framework leverages on an analytical solution decision support approach through identified dimensions, thereby improving reusability and sharing of knowledge (between traffic data scientists) for traffic prediction. By using the framework, a shared knowledge base can be developed, which can be reused by a given organisation for future projects. The use of shared knowledge ensures that meta-knowledge about traffic predictive algorithms is searchable, shareable, and updateable. The framework, which is depicted in Figure 5-1, delineates data-driven traffic parameter prediction into three (3) dimensions: (i) data context (DC), (ii) data collection method (DCM), and (iii) predictive analytic/modelling method (PAM). Further details of these dimensions and the framework are presented in subsequent sections. The novelty of our approach lies in the structured description of the analytical solution space and the subsequent guidance or support in the choice of prediction method adoption.

The argument is that the three dimensions presented in the framework interact with each other within each data-driven traffic prediction cycle. More specifically, the combination of elements within the DC and DCM dimension interacts with the chosen modelling approach. Giving it a more thorough examination, it becomes clear that interactions between the data context (i.e. problem-analytical solution mapping) and the data collection method can enable a suggestion of an appropriate PAM. For instance, conceptually speaking, the selection of prediction horizon and prediction type (i.e. real-time or not) interacts with the determination of the type of modelling approach to be adopted. Therefore, meta-knowledge about the prediction algorithms, such as what the various assumptions, generalisations, etc. are, can tell a traffic analyst what prediction algorithm is ideally optimal for a given prediction scenario. For instance, in the previous section, we identified the disadvantage of the *k*-NNs as being memory intensive and easily affected by noisy datasets. For massive datasets, it is intuitively determined that the k-NNs would be an algorithm to avoid for this prediction problem. According to Brazdil et al. (2003), the use of meta-knowledge can be very useful for model type selection.

5.2.1 Data Context (DC) Dimension

The DC dimension comprises factors related to the data scope or context including the traffic implementation area which distinguishes, for instance, urban (intra-city) networks from inter-city networks, for instance, freeway/highway networks. Another parameter within this dimension is the prediction horizon, which specifies the number of future prediction time steps. The dataset size is considered as another dimension element. It is evident that the dataset size can affect the suitability of an applicable PAM. For example, data-intensive *k*-NN or deep learning models may not prove suitable for prediction problems where only a limited or small dataset is available. Another element within this dimension is the data source(s) used, which identifies the external sources of data which may directly or indirectly affect traffic, e.g., weather, accident, road works, city events, etc. (Tsapakis et al., 2013). In today's data era, the option of fusing external data sources with traffic data is one that should not be ignored when performing traffic flow prediction (Essien et al., 2018). Finally, the prediction type is defined, which details if the prediction is to be done online/real-time or not. The next sub-sections present discussions about the respective elements within the data context dimension of TAG-F and how they can affect the choice of the predictive algorithm.

Traffic Scope/Area of Implementation

In Chapter 2, it was stated that the vast majority of short-term traffic prediction studies have been implemented on freeways/highways and motorways. This can be attributed to the fact that there is greater variety on the extent at which the impact of traffic characteristics can be felt, which can affect the degree of prediction accuracy (Kirby et al., 1997). Alternatively, it can also be the case that urban traffic is more stochastic and dynamic due to the interplay of exogenous factors such as pedestrian crossings, bus lanes and bus stops, etc. Therefore, in urban short-term traffic prediction, the forecasting task becomes more complex and specialised, having to incorporate other factors such as intersection control, traffic signals, and network optimisation (Vlahogianni et al., 2004).

For this reason, the traffic scope or area of implementation can be an indicator of the choice of the predictive model adopted in a given TPA task, due to the complexity and dynamism of urban traffic characteristics. Recent and advanced prediction models have been proven to outperform older and parametric predictive algorithms (Ermagun and Levinson, 2018; Jia et al., 2016; Smith et al., 2002; Vlahogianni et al., 2014) in modelling complex, dynamic traffic flow data, giving them a slight advantage over their parametric modelling counterparts. Therefore, if all other dimension elements kept constant, in modelling situations involving urban traffic settings, it is more likely for non-parametric models, such as neural networks, to outperform (i.e. with respect to predictive accuracy alone) parametric methods (for instance, ARIMA and regression-based models). A caveat to this statement, however, is not taking into account other factors such as dataset size, computational availability and requirements, model self-learning, and result interpretability/traceability.

Prediction Horizon

The prediction horizon represents the extent of time into the future for which the prediction is made. It can also be described as the time interval, which the forecasts are made and also represents the frequency of future predictions made (Vlahogianni et al., 2004). For instance, a prediction algorithm can be said to have a prediction horizon of 15-min ahead but achieve this in three (3) five-minute forward time intervals or time steps. The relationship between prediction accuracy and prediction horizon is directly proportional and has been empirically tested (Ishak and Al-Deek, 2002). The findings from the study showed that: the larger the prediction horizon, the less accurate the predictions become. The shorter the prediction horizon, the greater the potential of achieving more accurate predictions.

Due to the dynamic nature of traffic conditions, longer prediction horizons yield lower prediction accuracies. The results from Ishak and Al-Deek (2002) indicated that prediction error increased with the length of the prediction horizon. Defining the prediction horizon of a predictive analytics task is critical to accurate data modelling, especially in data-driven traffic prediction because it impacts upon the model's ability to capture the underlying interrelationships between the data points. The Highway capacity manual recommends 15-min intervals as the optimum prediction horizon because traffic tends to exhibit fluctuations in shorter prediction horizons. Similarly, other studies claim that prediction accuracy declines when using prediction horizons greater than 10-min (Smith and Demetsky, 1997; Vythoulkas, 1993). Similarly, it has been empirically shown that higher levels of data aggregation (i.e. prediction horizon) improved prediction accuracy by resulting in lower prediction errors (Abdulhai et al., 2002). This can be attributed to the loss of dynamic errors observed in the measuring sensors, which led to the recommendation in (Abdulhai et al., 2002) that the level of data resolution should be the same as the prediction horizon, as this would lead to better prediction results.

Studies such as Guo et al. (2017) have shown that Support Vector Regression (SVR) models tend to exhibit higher stability with increasing prediction horizon, compared to other models. This can be explained by the discussion in Kirby, Watson and Dougherty, (1997) that the use of more granular data aggregation can lead to reduced data fluctuations, making statistical models more efficient. However, for real-time adaptive traffic prediction, it is better to reduce the prediction horizon and time steps as low as possible, for example, 1-min.

In terms of prediction accuracy, a number of studies have investigated the impact of prediction horizon on prediction accuracy. For instance, Tian and Pan (2015) performed empirical evaluations to investigate the impact of prediction horizon on five (5) prediction models. The candidate models evaluated were: Random Walk (RW), SVR, Feed Forward NN (FFNN), SAE, and LSTMs. Each of these models was tested using 15, 30, 45, and 60-min prediction horizons, respectively. Although the results showed a clear superiority in terms of prediction accuracy for LSTM deep networks with an increase in prediction horizon, the SVR model, however, showed relative stability in prediction accuracy as the prediction horizon increased. In another study, the authors in Tian et al. (2018) performed quantitative comparisons between a number of traffic prediction models, including BPNN, SVR, ARIMA, Radial Basis Function Neural Network (RBFNN), Stacked Auto Encoders (SAE), and LSTM NN. The prediction horizon was varied between 15-min and 60-min traffic time steps. The results of the study revealed that, in comparison to a 15-min prediction horizon, the ARIMA, SVR, and RBFNN were most negatively affected by an increase in the prediction horizon (to 60-min), with the deep neural networks showing the smallest sensitivity to increased prediction horizon.

Dataset Size

The size of the available dataset to be used for data-driven traffic parameter prediction is a critical component that can impact on predictive model accuracy, training time, and computational demand. For the purpose of this study, let the dataset size refer to the data shape (matrix dimensions) of the dataset. For instance, consider a dataset that records 1,000 observations of 10 (features/variables) attributes, then the dataset size will be $1{,}000 \times 10 = 10{,}000$ data points. This is different from the actual size (in bytes) of the dataset. The dataset size in the TAG-F framework is categorised as large or small, with dataset sizes greater than 1,000,000 records considered as large datasets. A detailed discussion on the effect of dataset size, feature selection, and metrics on prediction is presented in Kumar and Vanajakshi (2015). According to the authors, prediction accuracy improves with an increase in training dataset size, especially in parametric models. In terms of the impact of dataset size on prediction accuracy, Barry-Straume (2018) performed an evaluative study on the impact of training size on validation accuracy using a popular dataset – the Fashion-MNIST (Modified National Institute of Standards and Technology) dataset of handwritten images (LeCun et al., 2018). The study findings suggested that validation accuracy increases with training dataset set, with a 40% reduction in training size resulting in an accuracy loss of about eight (8) percentage points. Similarly, Ajiboye et al. (2015) conducted an empirical analysis on backpropagation neural networks to investigate the impact of dataset size on model prediction accuracy. The findings showed that the model having the largest training dataset size outperformed the other models in terms of prediction accuracy.

However, there is a trade-off in this relationship (i.e. prediction accuracy and dataset size) when one considers an increase in dimensionality. This leads to a problem often referred to as Bellman's *curse of dimensionality* (Bellman, 1954). This refers to an exponential increase in time and space required to compute a solution to a problem as the dimension (i.e. number of variables) increases. On the other hand, an increase in the training dataset size provides more learning examples for learning algorithms (in the case of supervised learning), thereby improving learning ability. This, however, can lead to a situation where the algorithm learns the training dataset 'almost perfectly', resulting in poor performance with new/test/validation datasets because the learning algorithm has 'learned' both the data and noise in the training process, and therefore poorly performs in an unseen dataset. This condition is known as overfitting (Hawkins, 2004). In machine learning, a learning algorithm $h \in H$ is said to overfit a training dataset S if there exists a $h' \in H$ that has a significantly lower training error than a testing dataset error.

Data Source

Due to the highly stochastic, dynamic, and non-linear nature of urban traffic data, making accurate predictions becomes difficult and challenging. Empirical results from numerous studies (Essien et al., 2019a; Jia et al., 2017a, 2017b) have shown that the inclusion of non-traffic input data sources resulted in improved traffic prediction accuracy. In terms of traffic prediction, a number of challenges arise when fusing traffic data with non-traffic input data (Lin *et al.*, 2018). One of such challenges concerns the heterogeneity of the multi-source data, given that the data sources have diverse properties, and are perhaps on varying scales.

A typical example is the fusion of traffic and weather data, where values of traffic flows may range from 300 to 1,000 vehicles per hour (veh/hr). An attempt at fusing this with a weather dataset having a temperature range of say -10°C to 40°C may be challenging due to the differing scales. Another challenge of data source fusion has to do with the granularity of the individual datasets, for instance, synchronising time stamps of various features. This is a challenge in time series analysis, as different readings from various features might make the dataset fusion challenging. A typical example refers to a condition where traffic data is obtained from an ITS in a minute-wise manner. On the other hand, weather data, or accident data might be recorded in hourly (or even daily) intervals. In addition, there is the issue of structured vs unstructured datasets, which also needs to be accounted for when considering dataset fusion.

In terms of model type suitability, deep learning models are known to perform better with multi-source input data but, as earlier mentioned, are susceptible to overfitting (i.e. when the model captures the noise with the actual signal), thereby making them perform poorly when making predictions on new datasets (Hawkins, 2004). Given that deep neural networks have a large number of parameters and multiple non-linear hidden layers, they can, therefore, learn complicated relationships between the input and outputs. To overcome this demerit, many methods have been developed. For instance, it is common to 'intuitively' stop the training as soon as the performance of a validation subset starts to degenerate, introduce weight penalties – known as regularisation such as L1 and L2 – (Nowlan and Hinton, 1992), and apply a technique known as dropout (Srivastava et al., 2014). The technique of dropout refers to 'dropping out' units in a neural network. Dropping out means that the unit, alongside its incoming and outgoing connections, are temporarily removed from the network during training (Srivastava et al., 2014).

5.2.2 Data Collection Method (DCM) Dimension

The DCM dimension within the TAG-F framework enables the articulation of the method of traffic data collection. In line with rapid advances in technology, there exist a number of traffic data collection methods in use today. According to Leduc (2008), these methods are broadly categorised as intrusive and non-intrusive methods. Intrusive data collection methods like ILDs refer to such methods where the sensors are buried within the road segment being measured. Conversely, obtrusive or non-intrusive data collection methods refer to the methods that involve sensors that are external to the road segment being observed. The application of the different data collection methods can impact the outcome of a traffic prediction activity, as different collection methods are able to record traffic information at varying levels of granularity, aggregation, detail and accuracy. Intrusive devices – specifically ILDs – show the smallest average percentage error when compared to non-intrusive methods, such as video cameras, RADAR devices, etc. (Banger and Adriano, 2015). For instance, Bluetooth sensors incorporate a lot of noise in traffic observation given that mobile phones do have Bluetooth devices and so the sensor is unable to decipher the difference between a pedestrian with a (Bluetooth-enabled) mobile phone or a slow-moving vehicle. This impacts the accuracy of the measurement obtained from Bluetooth devices. Furthermore, video cameras and other non-intrusive measurement methods are cheap and inexpensive to implement but may show vulnerabilities when applied in congested situations.

This goes to show that the predictive accuracy of a model can be impacted by the choice of traffic data collection method. Accurate collection of traffic parameters will be of immense benefit to the traffic data scientist in ensuring that the dynamic and complex non-linear deep learning models are not also fitting the ‘noise’ or reading error (from the traffic collection method) within their modelling function. Going by the aforementioned, it is therefore reasoned that for non-intrusive traffic data collection methods (e.g. RADAR, video camera, infrared, etc.), which are less accurate (Banger and Adriano, 2015), simpler predictive algorithms, such as SVR, *k*-NN, and KF should be considered ahead of the deep learning models, which might result in less accurate predictions (Goodfellow and Bengio, 2015). This is because the more complex and advanced the learning algorithm, the more likely it is to capture both the underlying data patterns and the noise within the provided training dataset.

In summary, the previous sub-sections have discussed the effects of DC and DCM dimensions with respect to analytical modelling algorithm (i.e. PAM) selection. Meta-knowledge about the individual predictive models, deduced from the literature, can be used to enable the development of an inference model for providing guidance about traffic analytical model (i.e. PAM) choice or selection in a given TPA scenario.

5.2.3 Predictive Analytical Method (PAM)

The PAM dimension comprises the portfolio of predictive algorithms that are incorporated. For this study, we constrained the number of models to include seven (7) predictive/forecasting algorithms. The models were chosen on the basis of 'popularity' (or availability of literature), coverage of the solution space (i.e. applicability), and ensuring a fair reflection of each family of a predictive model is selected. Traffic prediction approaches are broadly classified into two categories: *parametric* and *nonparametric* methods. According to Russel and Norvig (2012), a parametric model is one that summarises data with a set of parameters of fixed size by simplifying the input function to a known form. This implies that once a parametric model has determined the model function or 'made up its mind' about the prediction, no matter how much more new data is introduced to the model, the number of parameters as well as the function will remain unchanged. Parametric functions are sometimes referred to as linear models, due to their inherent property of fitting a line to prediction functions, or in other words assuming linearity of prediction functions. Some examples of parametric PAMs are time-series analysis models (ARIMA based models), linear regression, naïve Bayes, and *linear discriminant analysis* (LDA). A major drawback of parametric methods, such as ARIMA and linear regression, is their tendency to focus on the means of the available training data, thereby missing the extreme values, which are very common to traffic datasets. Traffic data tends to exhibit peaks, especially at rush hours, as well as the influx of rapid fluctuations during incidents or accidents. For this main reason, ARIMA based models have the tendency to show serious vulnerabilities when applied to traffic forecasting.

On the other hand, algorithms that are able to learn from the training data, or do not make any assumptions about the mapping or prediction function, are referred to as nonparametric or non-linear models. The main advantage of such models is their ability to learn from any type of training dataset. Nonparametric models select the function that best fits the training dataset, thereby treating the prediction problem as an optimisation problem, meaning they are able to fit a large number of functions. An early example of a nonparametric model is the k-nearest Neighbour (*k*-NN) (Keller et al., 1985). Other examples of nonparametric models are decision trees (Classification and Regression Trees –CART- and C5 classification Trees), support vector regression (SVR), Artificial Neural Networks, etc. In the literature, Vlahogianni et al. (2004) present a number of advantages of using nonparametric traffic prediction methods. For instance, nonparametric models can produce accurate forecasts, are able to accurately model non-linear relationships in multivariate conditions and are relatively successful in structured or unstructured data-driven traffic prediction.

In this context, Barros et al. (2015) reviewed short-term, real-time prediction methods and provided a distinction between model-driven and data-driven short-term, real-time traffic prediction models. They suggest that model-driven traffic prediction approaches, such as Dynamic Network Assignment for the Management of Information to Travelers (DynaMIT) (Ben-Akiva et al., 1998) and Dynamic Network Assignment-

Simulation Model for Advanced Roadway Telematics (DynaSMART-X) (Mahmassani, 2001) execute operations via virtual representations (simulation), while data-driven models only take into account data collected via traffic sensors. The argument is that model-driven approaches are mostly used to analyse traffic as a whole, or macroscopically, which is why they are applicable in analysing driver behaviour and lane changing behaviour. The article concluded by presenting a framework that can provide guidance in the selection of a predictive approach in terms of data-driven or model-driven choices.

Due to increased interest in short-term traffic prediction, there is a rapid increase in the number of research studies. In terms of parametric prediction methods, Yu et al. (2015) present a real-time prediction model for predicting unoccupied parking space using a simple time series model (ARIMA). The authors show that the model is able to predict traffic parameters over a 15-min prediction horizon. Similarly, Tebaldi et al. (2002) and Kwon et al. (2000) present linear regression models capable of predicting traffic flow and travel time respectively over a 1-min rolling prediction horizon and 20-min prediction horizon. More so, Min and Wynter (2011) presented a spatiotemporal traffic flow prediction model based on a linear Space-time ARIMA (STARIMA) model and predicted traffic flow over a 5-min prediction horizon. Recent advancements in computing power and technology led to the development of robust, complex, and non-linear prediction models based on AI and ML. These set of models mostly use deep learning architectures, which means that stacks of neural network layers (representing the model depth, hence the term deep learning) are interconnected to transform data to effectively capture the stochastic nature of traffic flow (Goodfellow and Bengio, 2015).

The rationale for PAM selection in TAG-F

The TAG-F framework comprises a set of seven (7) candidate predictive algorithms or PAMs, including time-series modelling, instance-based learning methods, machine learning, and deep learning. The candidate models are Auto-Regressive Integrated Moving Average (ARIMA), Linear Regression (LR), Kalman Filter (KF), *k*-nearest neighbours (*k*-NN), Support Vector Regression (SVR), artificial neural network (ANN), and Long Short-Term Memory (LSTM) deep neural networks. The selected algorithms are chosen based on three criteria.

First, the models are chosen based on the availability of existing studies about the model architecture as well as relevant studies applying the model in TPA. Thus, the 'popularity' or effectiveness of the approach within the literature is a contributing factor to the choice of selected PAMs. For instance, there are more papers on the application of auto-regressive models (i.e. ARIMA) for traffic prediction than there are papers about the application of deep learning for traffic prediction. Consequently, more papers about the application and results obtained from this experimental process can lead to increased understanding of the underlying model architecture and design process.

Secondly, the models were chosen to reflect a balanced composition of parametric and nonparametric families of predictive algorithms. A representative sample of some machine learning algorithms have been presented in Table 1.1. In that table, each column represents a unique family of predictive models (for instance, regression based, rule-based, etc.). However, these families are broadly classified as parametric (i.e. models that assume the presence of a relationship between the input and target variables) and nonparametric (i.e. models that learn from the data) models. In this study, the seven (7) models have been selected to include a balanced inclusion of the parametric and nonparametric families of predictive models.

Thirdly, there was also a consideration and inclination towards the coverage of the solution space by the chosen models. For instance, the choice to exclude classifiers (such as Random Forest, Logistic Regression, etc.) given that the task of TPA is mainly more a regression (i.e. absolute or real number prediction) problem than a classification (i.e. binary or multi-class classification) problem. It is important to mention here that although the framework presently comprises only seven candidate models, this number can be extended to accommodate additional predictive algorithms. Extending the knowledge base to include additional predictive algorithms will require the inductive process (i.e. Literature-based Discovery [LBD] from the literature) about the PAM(s), as well as the critical analysis of the model in order to understand the advantages, disadvantages, generalisations, underlying assumptions, and areas of applicability.

Our approach towards model selection is based on a combination of theoretical meta-knowledge, which is knowledge about the predictive algorithms derived from a literature-based discovery process, and an instance-based learning approach. The concept of the use of meta-knowledge about learning algorithms is not new within the existing literature. For instance, Brodley (1993) captured expert knowledge about the applicability of prediction algorithms in order to provide inference about the suitability of prediction algorithms in a given prediction context.

5.3 Chapter Summary

This chapter presented the proposed traffic data analytics guidance framework (TAG-F). It built upon the foundation laid in Chapter 2 about data-driven traffic prediction and prediction algorithm choice. Detailed discussions about the conceptualisation of guidance in TPA, as well as the underlying logic for the framework, the identification of the framework dimensions, dimension elements, as well as the respective effects of the data context and traffic data collection method dimensions on the data analytical modelling dimension were also presented (see Section 5.2). The proposed framework is able to provide directional guidance towards data-driven traffic parameter prediction via the articulation of traffic analytical decision points, as well as providing guidance in terms of traffic prediction algorithm (i.e. PAM) selection.

In order to facilitate the practicality of the framework, a support tool was developed – the TAG-F support tool – which offers semi-automated guidance to traffic data scientists for prediction algorithm selection. The tool, its architecture, driving logic, and implementation details, are presented in the next chapter.

Chapter 6 TAG-F Support Tool and Framework Evaluation

6.1 Introduction

The previous chapter presented the proposed TPA guidance framework, the TAG-F framework, which has the main objective of providing systematic guidance to traffic data scientists for developing predictive analytics solutions to solve traffic prediction problems. The framework presented in Section 5.2 presented a guidance mechanism that can be used by traffic data scientists to perform TPA. According to the DSR methodology (see Figure 4-2 in Section 4.3), the outputs of the development (or design stage) and the evaluation are an artefact and the (artefact) evaluation, respectively. For this study, the artefact is the TAG-F framework.

In order to quantitatively evaluate/demonstrate the applicability of the framework, a prototype (software) tool – another artefact – was developed, which enables semi-automated directional guidance to traffic data scientists by suggesting alternative predictive models to given TPA problem scenarios. The tool, known as the *TAG-F support tool*, which complements the framework, takes as input a set of TAG-F dimension values (provided by the user) and returns a ranked list of predictions (suggestions) in terms of alternative predictive models. The TAG-F support tool has at its heart of operation an instance-based learning algorithm that is trained using a meta-learning framework (discussed in Section 6.5.1) via a meta-knowledge base that is populated using a *literature-based discovery* (LBD) process (see Section 6.4). Details (and justification) of the selected predictive models are presented in Section 5.2.3.

This chapter begins with a brief overview of the concept of *literature-based discovery* (LBD) in Section 6.2. Section 6.3 presents a systematic LBD-based process for extracting meta-knowledge about traffic predictive models using the TAG-F dimension parameters and existing literature. In Section 6.4, a discussion about the design of the support tool, including its architecture, implementation and underlying concepts, is presented. Section 6.5 presents an inference model for TPA predictive model selection, which is developed using a meta-learning framework for predictive model selection. The framework presented in Section 6.5.1 is an enhanced version of Vilalta's meta-learning framework for data mining (Vilalta et al., 2004). The TAG-F framework is evaluated using the developed prototype TAG-F support (software) tool, which is presented in Section 6.4. This evaluation process, presented in Section 6.6, was achieved using three (3) use case scenarios that a traffic data scientist may face. The three scenarios were selected to demonstrate different aspects of the TAG-F framework and support tool.

6.2 Literature-based discovery (LBD)

Literature-based discovery refers to the synthesis or extraction of existing articles or databases to discover new, hypothetically relevant relations between a given concept of interest and/or other concepts, by mining existing databases or knowledge bases (Bruza and Weeber, 2008; Hristovski et al., 2003). In other words, LBD can be seen as a knowledge extraction technique that uses existing databases or articles – the literature – to discover relationships between existing knowledge. This form of knowledge discovery aligns the process to what has been published, rather than what is known, by discovering valuable latent connections (of knowledge) between disparate studies (Sebastian et al., 2017). LBD differs from conventional empirical studies by generating knowledge via hypotheses that seek to connect knowledge from existing publications, rather than using empirical/experimental methods. LBD has mostly been applied in the field of bio-sciences and applied sciences.

Early research in the field of LBD was a study by Don Swanson (Swanson, 1988), which discovered (from academic publications) eleven (11) connections between Magnesium and Migraine that the 'standalone' articles did not know existed. He hypothesised that the combination of prior premises from two published articles, each positing that "*A* causes *B*" and "*B* causes *C*" respectively, leads to the hypothesis that there is a relationship between *A* and *C* (Ruch, 2010). LBD can be used for both *open* and *closed* knowledge discovery. Open discovery takes a concept *A* and seeks to identify a set of concepts that can be linked to the original concept. On the other hand, closed discovery assumes a potential relationship between two concepts and attempts to identify an intermediary concept linking the two concepts (Bruza and Weeber, 2008). In the biomedical sciences field, automated LBD techniques exist, which generate new knowledge by combining concepts in academic publications (Korhonen et al., 2014). Although LBD has been widely adopted in the field of biomedical sciences, its uptake in the general computing and data analytics community has, however, been limited (Ermagun and Levinson, 2018). This limitation can be attributed to the 'shallow' nature of the existing LBD methodology (Korhonen et al., 2014).

However, the application of LBD towards meta-level knowledge extraction is a technique that can yield benefit, owing to the multitude of predictive algorithms and existing literature (about the algorithms) in use today. Within this study, LBD is applied to existing relevant articles for the purpose of extracting meta-knowledge about few (seven) traffic predictive models in order to establish the relationship between the factors affecting TPA and the predicted ‘suitability’ of a given predictive model. A meta-learning framework, which aims at providing directional guidance to traffic data scientists using the structure provided by TAG-F (i.e. the three dimensions and the dimension parameters) enables the suggestion of alternative predictive models in a given scenario. This meta-learning framework is presented in Section 6.5.1.

6.3 Using TAG-F for developing/evolving a TPA meta-knowledge base

The TAG-F guidance framework presented in Section 5.2 described TPA using three dimensions. The proposed framework aimed at providing a guidance mechanism to traffic data scientists in performing TPA. The guidance to traffic data scientists proposed in this study (see Figure 2-1 in Section 2.5) is two-fold. Firstly, the TAG-F framework offers to traffic data scientists a mechanism for structurally defining the TPA problem space, which is fundamental to the development of a corresponding solution. Secondly, via the support tool, guidance or decision support is provided in the suggestion of possible predictive model alternative(s) to traffic data scientists. To provide guidance to traffic data scientists, there is, therefore, need for a robust, scalable, meta-learning, predictive model selection algorithm for TPA. This section presents a practical application of TAG-F towards developing a meta-knowledge base that can be used to provide predictive model guidance to traffic data scientists. Although the list of possible PAMs available to traffic data scientists is large, this research constrained the number of PAMs to seven (7), selecting these seven models using a number of criteria that include popularity, availability in the literature, and coverage (all criteria are listed in Section 5.2).

6.3.1 Systematic Literature Review process

The literature review is a critical element of academic research, which can serve many purposes and provide answers to research questions. According to Booth et al., (2016), reviewing literature *systematically* helps to eliminate bias, thereby resulting in a clear presentation of results that represent the objectivity of the research methods applied. Mulrow (1997) also stated that a systematic review represents the search for the *whole truth*, which can be classified as a scientific activity. Therefore, in this present study, adopting a systematic review framework similar to (Harris et al., 2014), the following steps were taken: (i) identifying the unanswered answerable question, (ii) specifying the inclusion and exclusion criteria, (iii) extraction and analysis of study data, and (v) summary of findings.

Recent research advancement has resulted in the development of models, frameworks, and algorithms for traffic parameter prediction (Barros et al., 2015; Jia et al., 2016; Lai et al., 2016; Lu Lin et al., 2018; Lv et al., 2009; Poonia et al., 2018). For the study reported in this thesis, the systematic review process examined a broad range of articles, identified by carrying out web searches on Scopus[1] electronic library. The results obtained included a diverse range (with dates commencing from 1958) of articles, with the search process extended further from search words in abstracts, titles/keywords, to include in-text search. This ensured that a reasonable amount (of studies in the literature) was gathered. The search keywords were: *"Traffic forecast", "Traffic prediction", "Traffic forecasting".*

[1] Scopus can be found at https://www.scopus.com/search/form.uri?display=basic

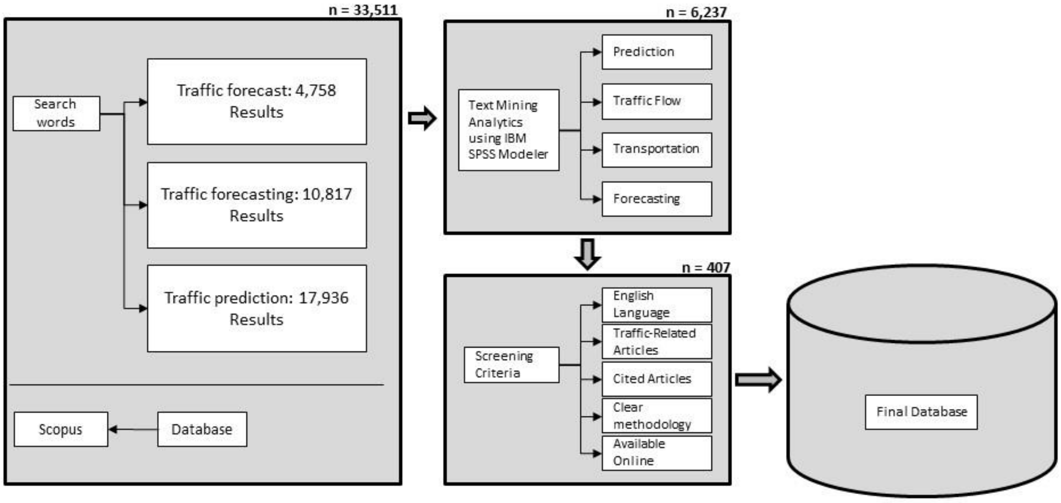


Figure 6-1: Flowchart of Systematic Review approach

6.3.2 Screening Criteria

To summarise the study inclusion and exclusion criteria, a systematic review flowchart comprising four steps (see Figure 6-1) was used. As can be seen from the diagram, the process began with an online search through the database to extract all articles related to the keywords searched. This returned a large number of results files containing about 33K records. The next step was data pre-processing, including duplicates elimination and text mining on the resultant database. This step was carried out using IBM SPSS Modeller, where text mining analytics was performed on the initial result set to prune the unwanted results. It is, however, important to mention here that although the IBM SPSS Modeller tool provided guidance in the entire process, the results were verified and reviewed by the author of this thesis manually. This first filtering and pre-processing step resulted in a smaller dataset comprising 6,237 records.

However, the results still included a few ‘irrelevant’ articles (for instance, studies about mobile/internet traffic prediction, or water current traffic flow prediction, etc.).

For this reason, a further filtering process was undertaken to retain studies that were of interest to “road traffic flow”, “prediction”, “forecasting”, and “transportation”. The next step applied pre-defined screening criteria to reduce the number of articles further. The screening criteria were: (1) Articles in English Language only, (2) Road Traffic-relevant articles (i.e. exclude web traffic, air traffic, mobile traffic, etc.), (3) clearly stated methodology. The justification for the screening criteria mainly bordered around the objectives of the study reported in this thesis. The first two criteria were applied out of necessity rather than of choice to narrow the results obtained in terms of relevance to this present study. Secondly, traffic flow is observed, studied, and analysed in various sectors such as network traffic for communications (i.e. mobile, internet, etc.), air traffic, sea and water traffic, etc. This precipitated the need to streamline the search results to road traffic studies alone, which constitutes the focus of this research study. Thirdly, for every academic research, including this one, there needs to be a clear methodology that should also be relevant to the research presented in this thesis, which justifies the third criteria. This resulted in a final/complete dataset containing 407 articles.

Within this study, of the 407 papers reviewed, 150 used parametric data modelling techniques, 234 used nonparametric, and 23 used hybrid modelling approaches. However, some schools of thought tend rather to classify modelling techniques as either *statistical* or *machine learning*. Statistical methods basically provide a model that delivers intelligence from the provided dataset by inference and/or estimation (Smith et al., 2002). Conversely, machine learning methods produce predictions via a process of ‘learning’ from the provided dataset using a feedback or activation function to update the learning process. Given that these terms are used interchangeably, we refer to *parametric* and *nonparametric* methods for the remainder of this study.

6.3.3 PAM Characterisation

Table 6-1 presents a summary of characteristics for the candidate prediction models within the TAG-F framework support tool, which can be used to infer the choice of a particular model in a given scenario. As can be seen, the characteristics focus on the assumption of the data about the temporal dependency of the dataset (for instance, time-series, sequential data, etc.), multivariate modelling, input data required, prediction nature, as well as disadvantages and advantages. The table also corresponds to the elements within the dimensions of the proposed framework, thereby making it possible to obtain inferences about PAM approaches to be employed in given scenarios.

Some studies have made attempts at comparing the performance in terms of predictive accuracy between statistical and machine learning models. For instance, Zhu, Wang, Zhang, and Song (2016) claim no significant difference in performance between the Bayesian network (BN) and ARIMA model. However, when the prediction horizon was varied, it was observed that the BN outperformed the ARIMA model in longer prediction time steps. Jiang et al. (2016) also provided a comparative study involving five machine learning methods – BPNN, non-linear AR model with exogenous input neural network (NARXNN), SVM with linear function (SVM-LIN), SVM with radial basis function (SVM-RBF), and multilinear regression involving three statistical methods: ARIMA, Vector Auto Regression (VAR), and space-time (ST) model. The results from the study revealed that the machine learning models significantly outperformed two of the statistical models and that with an increase in the prediction horizon, the prediction accuracy of the ST model reduced. Smith, Williams and Oswald (2002) compared the predictive performance of three traffic flow prediction methods (ARIMA, ANN, and spatiotemporal ARIMA) in varying traffic conditions.

Table 6-1: Summary Characteristics of PAMs in this study

	Parametric PAMs			Nonparametric PAMs			
Hypotheses	***Linear Regression***	***ARIMA***	***KF***	***SVR***	***k-NN***	***ANN***	***LSTM***
Hypothesis on time-dependency of data	Assumes a time-dependent input	Requires a stationary time-series data	None	None	None	None	None
Hypothesis on data nature	Deterministic/Reproducible	Random	Random	Random	Random, dependent on nearest neighbours	None	None
Multivariate modelling	Complicated	Complicated	Easy	Easy	Easy	Very easy	Very easy
Linearity	Data must be linear	Linearity must be predefined	Must be predefined	Must be predefined	Linearity must be predefined	Not required	Not required
Input Dataset size	Can work with small size datasets	Requires large time series data	Can work with medium size time series	Can work with medium-size time series	Requires large dataset size	Data-intensive, may not work well with small datasets	Data-intensive, may not work well with small datasets
Accuracy	Performs well in linear datasets, poorly in non-linear datasets	Performs well in linear datasets, poorly in non-linear datasets	Performs better than ARIMA/LR in non-linear datasets.	Relatively (i.e. compared to LR, ARIMA) accurate in non-linear datasets	Relatively (i.e. compared to LR, ARIMA) accurate in non-linear datasets such as traffic flow	Highly accurate in non-linear datasets	Highly accurate in non-linear datasets, especially in long predictive horizons.
Prediction Horizon	Performance degrades as PH increases	Performance significantly degrades as PH increases	Performance slightly degrades as PH increases	Performance degrades as PH increases	Performance degrades as PH increases since the predicted value is addedto training data	Performance slightly degrades with PH	Smallest (compared to other models in this table) degradation with PH.
Prediction Nature	Linear/static	Static	Static	Dynamic	Dynamic	Dynamic	Dynamic
Advantages	• Not data intensive • Works well in short prediction horizons • Easy to implement • Implements a purely statistical model, which is reproducible and explainable	• Well-established theoretical background • Relatively easy to implement • Works very well with short time series • Can compete with the best models in short-run high frequency predictions • Simple model structure	• Statistics-based model • Ability to perform multivariate prediction • Ideal for continually-changing systems • Computationally light and fast • Best in online real-time prediction	• SVRs have regularisation parameters, which implies that the user always has to think about overfitting • It is kernel-based, implying that it can build in expert knowledge. • It is defined by a convex optimisation curve, eliminating the problem of local optimums	• Easy to implement • Flexible to retrain • Supports multivariate prediction • Provides better accuracy than ARIMA and KF	• Non-stationary, non-linear model • Very easy to incorporate multivariate prediction • Does not require advanced statistical knowledge to implement	• Non-stationary, non-linear model • Very easy to incorporate multivariate prediction • Does not require advanced statistical knowledge to implement • Best in capturing time dependency of time series data • Can work perfectly well with unstructured datasets • Auto feature optimisation
Disadvantages	• Very poor results in non-linear prediction problems like traffic • Oversimplifies prediction problems, leading to massive errors in prediction of non-linear functions	• Model selection is more of an art than a science • No automatic updating or retraining (i.e. new data requires retraining) • Very unstable	• Assumes linearity of the data variables and planes • Assumes a Gaussian distribution for the belief	• The theory only covers the determination of the model parameters • Can be overly sensitive to over-fitting the model selection criterion.	• Data intensive • Requires storage of the entire training dataset, which is also memory-intensive • Large search problem to find the nearest neighbours • Computationally intensive, especially for large datasets • Highly susceptible to noisy datasets	• Results are not traceable (i.e. black box approach) • Extremely data-intensive • Computationally expensive • Requires extensive processing time for training • Complex internal structure • Susceptible to over-fitting • Mainly empirical	• Results are not traceable (i.e. black box approach) • Extremely data-intensive • Computationally expensive • Requires extensive processing time for training • Complex internal structure • Susceptible to over-fitting • Mainly empirical

The result from the study revealed that as prediction horizon increased, machine learning methods produced superior outputs in comparison to the statistical models. This result is consistent with a number of research studies, such as (Smith et al., 2002; Vlahogianni et al., 2004; Zhu et al., 2016).

These set of results reveals a trend in terms of traffic prediction and model selection. From the trajectory observed, it is imminent that the superiority of machine learning predictive methods is truly manifest in longer prediction horizons; thus, the increased interest in recent years. However, before a conclusive decision is reached, many more comparative studies need to be carried out.

6.4 Design of the TAG-F Support Tool

A prototype support tool, known as TAG-F tool, was developed as a practical implementation of the framework which was proposed in Chapter 5. The three (3) main components of the TAG-F tool, which complements the framework, are the *knowledge base*, the *inference engine*, which ranks prediction models from a set of pre-defined candidate models and the *user interface*. The ranking of predictive models is performed by an instance-based ML algorithm built using the meta-knowledge extracted using a literature-based knowledge discovery process described in Section 6.4 above. Figure 6-2 presents the architecture of the TAG-F tool. The user interface performs the interaction with the end-user, thereby controlling the guidance process. As can be seen from the figure, the input decisions are passed from the user interface to the control program and vice versa.

6.4.1 TAG-F Support Tool Architecture

Figure 6-2 represents the TAG-F support tool architecture. The three (3) main components of the tool are the *knowledge base*, *user interface* and the *inference engine*. The knowledge base houses insights about the predictive models, obtained via a literature-based discovery process (Bruza and Weeber, 2008) and meta-knowledge (Vilalta et al., 2004) extraction. The concept of meta-knowledge for automatic model selection is not new to the field of prediction (Brodley, 1993; Pappa et al., 2014; Vilalta et al., 2004; Xiaofeng Wang et al., 2009).

Meta-knowledge extraction is based on the fact that predictive algorithms tend to be 'biased', therefore leading to superior results in one scenario, and poor results in other scenarios. This bias can be attributed to the generalisations, assumptions, and approximations made during the predictive algorithm learning. A typical example of such generalisations is the assumed stationarity of the mean and variance in auto-regressive integrated moving average (ARIMA) models (Smith et al., 2002). Due to this assumption, one can intuitively gather that an ARIMA predictive model would ideally perform well when trained on datasets having small or no variance, and may consequently perform poorly when exposed to datasets with high variance (for example, urban traffic which has spikes at peak periods or during incidents).

The user interface provides a graphical representation of the underlying program logic, which allows users to interact accordingly. The control program includes the dimension elements parser, which was developed using Hypertext Markup Language (HTML). The entire source code for this thesis can be found online[2], but code snippets for the TAG-F support tool are included in Appendix b. The graphical user interface

[2] https://github.com/nakessien/tagf_evaluation.git

of the TAG-F Support Tool is presented in Figure 6-3. As can be seen, the user is presented with options to select parameter values in accordance with the TAG-F framework parameters. The output screen of the support tool interface is shown in the bottom half of Figure 6-3

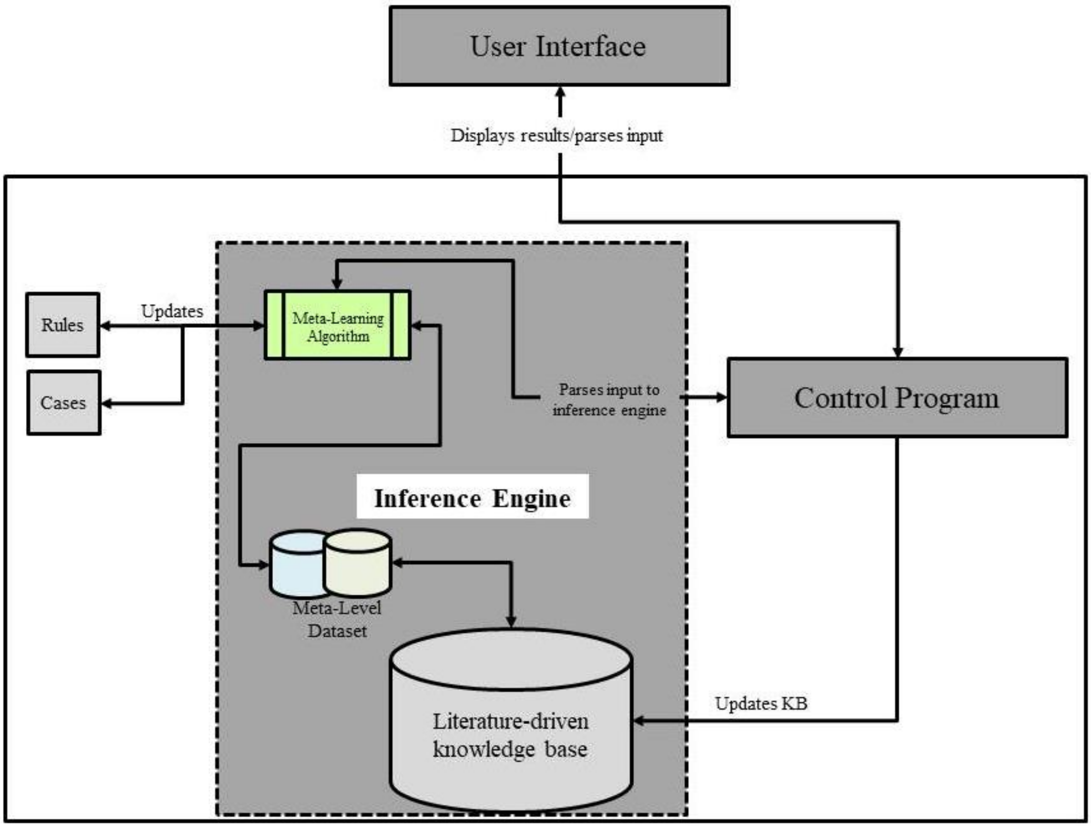
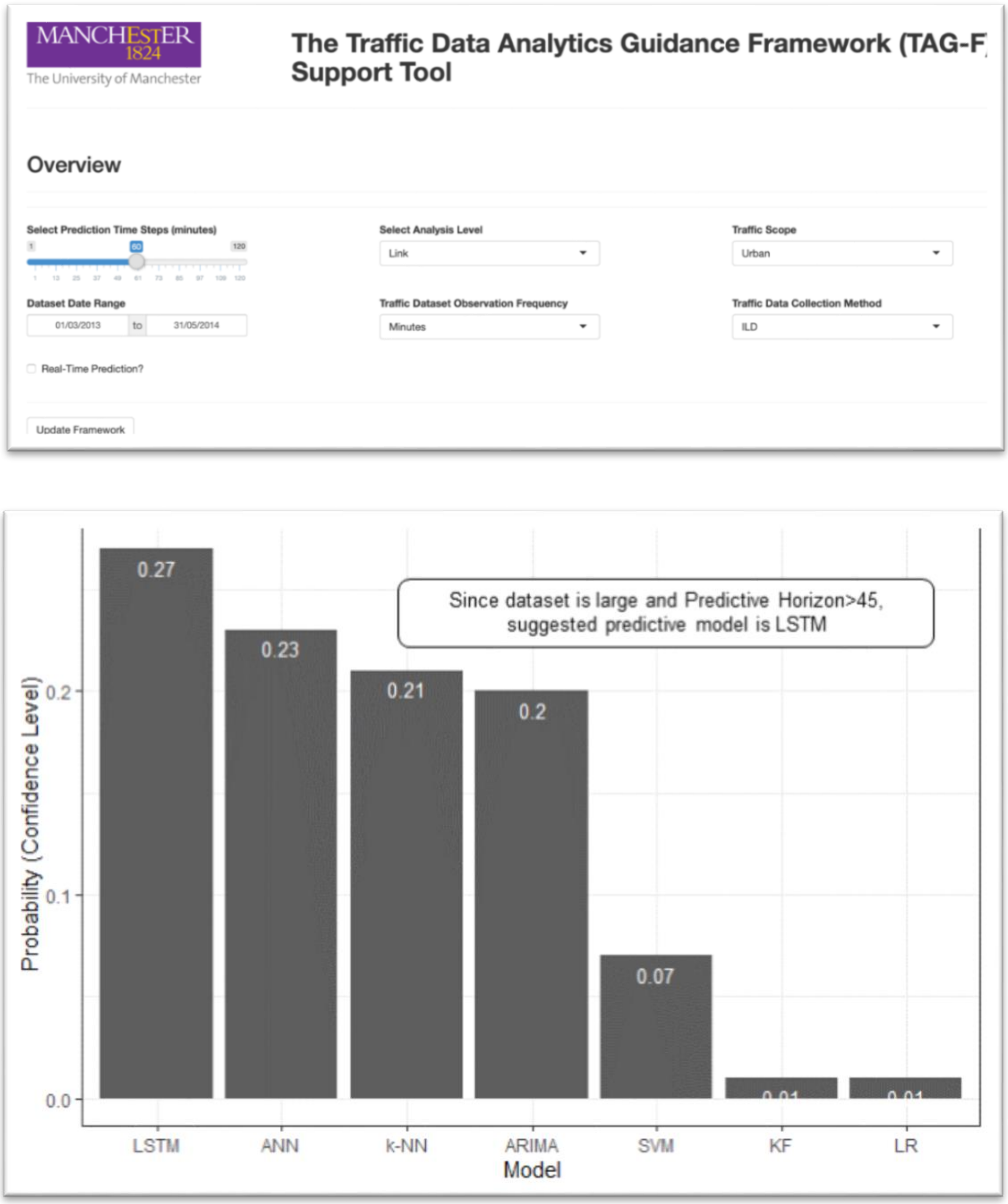


Figure 6-2: TAG-F Tool Architecture

At the core of the prototype tool is the inference engine, which comprises a machine learning algorithm that learns from the input meta-dataset built from the literature-based discovery process described in Section 6.4. The implementation details about the support tool, including the inference engine, are discussed in the next section.



The Traffic Data Analytics Guidance Framework (TAG-F
Support Tool

Overview

Select Prediction Time Steps (minutes): 60

Select Analysis Level: Link

Traffic Scope: Urban

Dataset Date Range: 01/03/2013 to 31/05/2014

Traffic Dataset Observation Frequency: Minutes

Traffic Data Collection Method: ILD

□ Real-Time Prediction?

Since dataset is large and Predictive Horizon>45, suggested predictive model is LSTM

Figure 6-3: TAG-F Support Tool User Interface

6.4.2 TAG-F Support Tool Implementation

The development of the user interface for the TAG-F support tool is realised using *'shiny'* package (RStudio, 2016), which is a web application development interface for R[3]. The inference engine is implemented using an instance-based learning classification algorithm, which is discussed in the next subsection.

6.5 A PAM Suggestion model for Traffic Prediction using Meta-learning

As summarised in Section 2.8, due to the combination of the NFL theorem (Wolpert and Macready, 1997) and the plethora of learning algorithms, the optimal combination of predictive analytical strategies is dependent on the specific problem at hand. It will, therefore, be useful to provide traffic data scientists with an approach that can provide suggestions about the optimal predictive model to adopt, and also rank them in order of their potential or predicted suitability. The support tool presented in this chapter aims to provide this support functionality, focusing on a single aspect of the TPA guidance offering proposed in this research study – traffic predictive analytical method (PAM).

More specifically, this phase of the TAG-F tool architecture (see Figure 6-2) provides a ranked list of potential candidate traffic predictive algorithms to be adopted, given a set of input parameters corresponding to an encoded meta-level dataset. It must, however, be stated here that the onus lies with the data scientist to perform the requisite model hyperparameter optimisation, data wrangling, cleaning and pre-processing of the input data prior to performing the predictive analytics task. This happens outside of the TAG-F framework and support tool.

Models have *parameters* and *hyperparameters*. It is important to distinguish between the two (often confused) terms. A model parameter is an internal variable that is updated/estimated from the data during the learning process. Typical model

[3] R is a statistical programming framework that can be found at: https://www.r-project.org/

parameters include variables, such as the mean μ and standard deviation σ in a Gaussian distribution, support vectors in SVMs, and weights in artificial neural networks (Kuhn and Johnson, 2013). This differs from a model's hyperparameter, which is a configuration or variable that is external to the model, and therefore is not updated automatically by the model during the learning process. Model hyperparameters are often a set of heuristics that are '*tuned'*, affect model performance, and are context-specific. Typical examples of hyperparameters include the number of nearest neighbours k in a k-NN model, layer size, dropout, and layer width in deep learning models, and the C and σ hyperparameters in SVMs (Kuhn and Johnson, 2013).

In the TAG-F tool, the selection and ranking of the algorithms were performed using an instance-based learning (IBL) algorithm. The choice of this learning algorithm is due to the fact that, in meta-learning, the input dataset and problem description metadata (i.e. dataset description, TAG-F dimension element metadata, etc.) is typically small. For this reason, adopting general models such as decision trees and rule-based algorithms that generate crisp assumptions and generalisations may not result in an optimal result set (Lindauer et al., 2017). Furthermore, IBL algorithms are known to be advantageous in that they can incorporate self-learning (i.e. once a new input and result become available, it can easily be added to the model training data without any need for additional model re-training) (Brazdil, 2003). This is particularly advantageous because, as with this study, the knowledge base comprises only a small meta-dataset (407 observations in this case) and potentially will increase with time (refer to Appendix a for dataset or online for the entire code on GitHub[4]). The entire

[4] https://github.com/nakessien/tagf_evaluation.git

dataset used for training the TAG-F support tool has been included in Appendix b. The data segments utilise simple text-to-integer encoding, such that the class ‘ARIMA’ is represented as ‘0’, while ‘Neural Network’ is represented as ‘1’, etc. To achieve this, a simple algorithm, the *k*-NN ML algorithm was adopted. The algorithm uses a distance function, which is based on the set of encoded input variables (from the TAG-F characterisation) to compute the most similar neighbours. The recommended ranking is then constructed by computing the probability distribution of the algorithms on the selected dataset.

6.5.1 A Meta-learning framework for predictive model selection using TAG-F

As previously stated in the introductory chapter of this thesis, one of the challenges encountered in TPA is the presence of a plethora of predictive models/algorithms with the shortage of available guidelines to select the right method given the nature of the given TPA problem. The research field of *meta-learning* is primarily concerned with producing such ‘guidelines’ by understanding the interaction between the mechanism (and generalisations/assumptions) of learning in individual predictive models, and the given contexts/scenarios in which the model or mechanism is applicable (Vilalta et al., 2004; Xiaozhe Wang et al., 2009). The process of meta-knowledge capturing for predictive model selection involves capturing certain relationships between the available characteristics and the performance of the algorithms. Several studies (Brazdil, 2003; Pappa et al., 2014; Vilalta et al., 2004; Xiaozhe Wang et al., 2009) have demonstrated the plausibility and potential of meta-learning for model type selection.

From a theoretical standpoint, meta-learning can solve important problems in relation to the application of machine learning, data analytics and TPA. Firstly, without the provision of some manner of guidance or assistance, predictive model selection can quickly become an obstacle to a traffic data scientist and a data scientist (in general). For instance, data scientists are short of the required time and computational resources to advance a trial-and-error (i.e. brute force) regime for all existing predictive models in the literature, which the TAG-F framework presented in this study aims to solve.

From the foregoing, this study proposed a meta-learning approach towards predictive model selection from a dictionary of seven (7) candidate models (see Section 5.2). Therefore, in addition to providing a structured description or definition of the analytical problem space, TAG-F – through its support tool – can provide suggestions with regard to a predictive algorithm to be adopted. This study adopted a meta-learning approach for data mining adapted from Vilalta's meta-learning framework for data mining tasks (Vilalta et al., 2004). The enhanced version of Vilalta's framework is presented in Figure 6-4. The meta-learning approach comprises three core components, namely:

(i) Predictive algorithm evaluation

(ii) TPA Characteristics extraction, and

(iii) Rule induction/instance-based learning

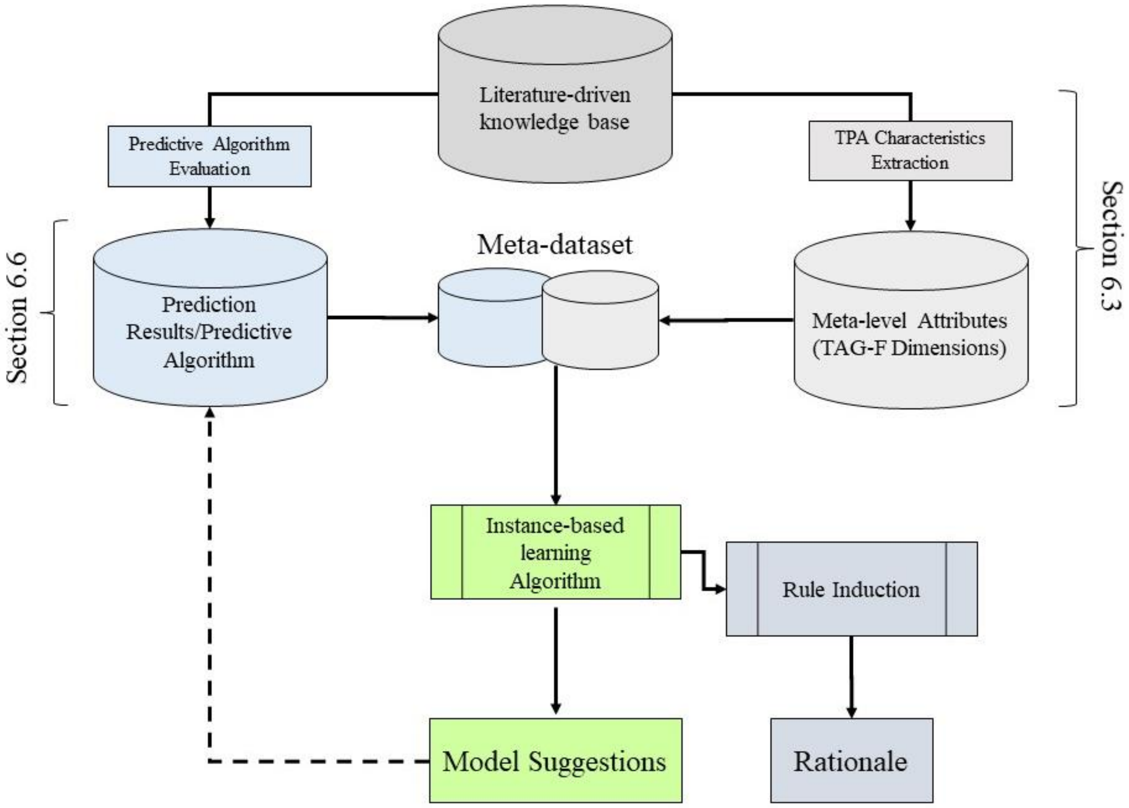

Literature-driven knowledge base
Predictive Algorithm Evaluation
TPA Characteristics Extraction
Section 6.3
Section 6.6
Prediction Results/Predictive Algorithm
Meta-dataset
Meta-level Attributes (TAG-F Dimensions)
Instance-based learning Algorithm
Rule Induction
Model Suggestions
Rationale

Figure 6-4: A literature-driven meta-learning framework for predictive algorithm selection

The framework commences with the development of a knowledge base. In this present study, this is achieved using the literature-based discovery (LBD) process, which is described in Section 6.4 via the TAG-F framework. The result is a characterisation of the final literature-driven knowledge base into meta-level attributes from the TAG-F framework (i.e. prediction horizon, data collection method, dataset size, etc.). The predictive model evaluation (dashed line in Figure 6-4) is presented in Section 6.6, where the framework – via the support tool – is quantitatively evaluated using real-world TPA case scenarios. The output of the combination of the prediction results evaluation and data dimension characteristics extraction is a *meta-level* dataset, which comprises meta-level attributes from the knowledge base and the evaluation metrics from the literature-driven knowledge base. The meta-level dataset is subsequently used to train a machine learning algorithm (the TAG-F instance-based learning algorithm presented in Section 6.5.2) for the purpose of discovering the relationships between the forecasting methods (predictive methods/algorithms) and the data dimension characteristics (TAG-F dimensions).

The meta-learning model proposed in Figure 6-4 has the objective of providing predictive model suggestions based on a set of meta-level attributes obtained from the TAG-F framework categorisation. The suggestion of the best model comes from the literature, but the evaluation is performed using actual runs of the data in each of the scenarios. The evaluation metrics (i.e. RMSE, MAE, sMAPE, etc.) are standard and come from the literature. However, the choice of each PAM suggested to the analyst in the evaluation of a particular scenario is based on actual runs based on what is suggested from the literature (tool). In addition, the model provides the data scientist with a rationale/justification of the inference or prediction of the most likely PAM in the given scenario. The explanation or justification of the specific inference is realised using a rule-based learner – Repeated Incremental Pruning to Produce Error Reduction (RIPPER) – (Cohen, 1995). Brief background about this rule-based learning algorithm is presented in the next subsection. A machine learning algorithm is then trained on the resultant meta-level dataset in order to infer patterns in the relationship between predictive algorithms and TAG-F dimension characteristics/elements. This framework can, therefore, be used to provide directional guidance to traffic data scientists by suggesting alternative predictive model(s) in the TPA solution development process.

6.5.2 The TAG-F Tool Instance-Based Learning (IBL) Algorithm

The inference engine of the TAG-F support tool is a k-NN learning algorithm that, given a new set of traffic predictive analytics problem parameters (categorised using the TAG-F), generates a ranked list of the seven (7) candidate algorithms presented in

Section 5.2. The main advantage of adopting this technique (i.e. meta-level learning) compared to existing AutoML techniques (for instance, H2O driverless AI[5] (Hall et al., 2017), Auto-WEKA (Thornton *et al.*, 2013) and Auto-WEKA 2.0 (Kotthoff et al., 2016)) is as follows. In this study, each of the PAMs in the TAG-F support tool are not physically trained (using the input dataset) to develop the ranked list in each TPA scenario (for instance Scenario 1, etc.), which would require extensive computational power and significant training time. For instance, when handling large datasets (Scenario 3) and using deep learning algorithms, such as LSTM.

The meta-dataset developed using LBD and TAG-F framework (Section 6.3), which categorises the information relating to case-specific traffic prediction problem → predictive algorithm is used as the training dataset for the IBL algorithm. The algorithm was trained to minimise the distance function (i.e. relative similarity of the input dataset to the predicted observation). The distance function used to measure this similarity was the unweighted L_1 norm (Atkeson et al., 1997). The distance is computed using the following equation:

$$dist(x_i, y_j\ ) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p} \tag{6-1}$$

Where x_i and y_j are datasets corresponding to the target and label variables (see Appendix), and p is a value that can be manipulated to calculate the distance using one of three ways. In equation 6-1, if $p = 1$, then the distance is known as Manhattan. If $p = 2$, then the distance metric is the Euclidean distance. If $p = \infty$, then Chebychev distance metric is used.

[5] Can be found at: https://www.h2o.ai/products/h2o-driverless-ai/

6.5.3 The RIPPER rule-based learner for rule-induction meta-learning

Meta-learning for time series data has been explored in the data mining community and studies, such as (Bradley and Fayyad, 1998; Halkidi et al., 2001; Kalpakis et al., 2001). Many rule-based learning algorithms are in existence today, such as 1R (one rule), prism, and decision trees. For the purpose of generating rules on how to select the most appropriate predictive model for time series, Wand, Smith-Miles and Hyndman (2009) proposed a characteristic-based meta decision tree for rule induction using the C4.5 algorithm. However, a technique known as incremental reduced error pruning (IREP) is mainly the driving engine behind rule-based learning algorithms. These algorithms operate by growing rules one at a time, thereby building rules for binary or multi-class problems. In 1995, William Cohen implemented a propositional rule learner known as Repeated Incremental Pruning to Produce Error Reduction (RIPPER) in Cohen (1995). The RIPPER algorithm functions as follows.

First, the classes are ordered according to increasing prevalence $(C_1,, C_2, C_3, \dots . C_k)$ where C_1 is the least prevalent class and C_k is the most prevalent. Next, a rule set (using IREP) is applied which separates $C_1$from other classes, such that $IREP(Pos = C_1, Neg = C_2, \dots C_k)$. After this, all the instances covered in the learned rule set are removed from the dataset, and this is repeated until a single class C_k remains, which will be used as the default class. Within R studio, the RIPPER learning algorithm is implemented using the $RWeka$[6] package. In this present study, the JRip algorithm is applied towards meta-learning rule induction on the meta-level dataset developed from an LBD process described in Section 6.3.

[6] https://www.rdocumentation.org/packages/RWeka/versions/0.4-40/topics/Weka_classifier_rules

In this research, the rule-induction system was applied on the set of seven (7) PAMs in the TAG-F framework and trained on the meta-level dataset. The JRip algorithm was trained on the dataset using R scripts and different parameter settings, evaluating each output using the F-measure (f-score) in order to obtain the best rules. Therefore, in addition to the suggested PAMs, it is possible to provide a justification or rationale as to why the particular PAM comes as the first suggestion to the user (data scientist).

Table 6-2: Some Rules induced from meta-level dataset

Rule No.	**PAM**	**Description**
1.	ARIMA	IF • Analysis level is link • Traffic Scope is non-urban • DCM is NOT manual • Dataset is not large • Prediction horizon <=15
2.	LR	IF • Dataset is not large • Traffic Scope is non-urban • DCM is manual or Bluetooth or ILD • Prediction horizon between 10 and 15 • Multivariate
3.	*k*-NN	IF • Dataset is large • Traffic Scope is urban • DCM is FCD or Bluetooth or ILD or Microwave • Prediction horizon is large • Analysis level is area
4.	SVR	IF • Dataset is not large • DCM is Bluetooth or ILD • Traffic scope is urban
5.	KF	IF • Real-time prediction
6.	ANN	IF • Analysis level is urban • DCM is ILD • Prediction horizon >=15 • Multivariate data • Dataset is large
7.	LSTM	IF

		• Dataset is large • Traffic Scope is non-urban • DCM is manual or Bluetooth or ILD • Prediction horizon >30

Table 6-2 presents a summary of some rules induced by the output of the JRip algorithm. It must be mentioned here that the rules presented in the table are by no means an exhaustive list of rules that are generated from the JRip algorithm in the TAG-F support tool following the principles mentioned in this subsection. The JRip algorithm generated in excess of 100 rules, many of which may not be human-interpretable. Therefore, what is shown in Table 6-2 is a high-level rule set to the traffic data scientist about why a particular PAM was ranked first by the support tool. However, the table can be treated as set of heuristics that can be used by traffic data scientists for PAM suggestion in TPA.

6.5.4 Data Description

The summary of the data used in training the inference engine for the TAG-F tool is presented in Table 6-3. The dataset contained 407 related studies that were encoded using the TAG-F framework (see Section 5.2). The entire dataset used for training the TAG-F support tool has been included in Appendix b. The data segments utilise simple text-to-integer encoding, such that the class 'ARIMA' is represented as '0', while 'Neural Network' is represented as '1', etc. Furthermore, the entire code for this thesis can be found online at GitHub[7]. The dataset target variable contained seven (7) classes, corresponding to the set of candidate models used within this study (see Chapter 5). The learning algorithm was, therefore, trained to – given a set of input parameters – provide suggestions about the predictive models to adopt.

[7] https://github.com/nakessien/tagf_evaluation.git

Table 6-3: Inference Engine Training Data description

Class Details	Train	Test	Total
Neural Network	71	13	84
ARIMA	53	11	64
Kalman Filter	54	8	62
Linear Regression	51	8	59
SVR	43	7	50
LSTM	39	7	46
k-NN	36	6	42
		TOTAL	**407**

6.5.5 Evaluation Metrics

In order to evaluate the performance of classifiers, many methods exist that can be used for discriminative evaluation of the optimal classification model. Typically, a confusion matrix is used to visualise the discriminative evaluation of the classifier output vector. Table 6-4 shows a sample confusion matrix, where the rows represent the predicted class, and the columns the actual class. In the table, tp and tn represent the true positive and true negative values, respectively, while fp and fn represent the false positive and negative classes, respectively.

Several metrics can be used to evaluate the performance of classifiers. For instance, studies such as (Bekkar et al., 2013; Hossin et al., 2011) advocate that accuracy is a balanced indication of binary or multi-class classification problems. However, *accuracy* or *error-rate* is limited in that it produces less discriminable values and does not penalise false positive prediction (Huang and Ling, 2007). For this reason, the *F-measure* or *F-score* (i.e., the harmonic mean of the recall and precision values) and

Geometric Mean (GM) (a metric that maximizes the tp and tn rate) are better suited towards discriminating fp and fn predictions (Huang and Ling, 2005). Similarly, the *Area under the Receiver Operator Characteristics* (ROC) *Curve* (AUC) is a reflection of the overall ranking performance of a classifier, which is different from other performance evaluation metrics that consider thresholds and probability spread (Huang and Ling, 2005). Going by those mentioned above, this study adopted the F-measure (FM) as the classifier performance evaluation metric for the TAG-F tool.

Table 6-4: Confusion Matrix for Binary Classification

	Actual Positive	**Actual Negative**
Predicted Positive	tp	fn
Predicted Negative	fp	tn

The FM is the harmonic mean of the recall and precision of a classifier, described as:

$$FM = \frac{2 \times p \times r}{p + r} \tag{6.2}$$

Where p is the precision or positive predictive value defined as $p = \frac{tp}{tp+fp}$ and r represents the recall or true positive rate defined as $r = \frac{tp}{tp+fn}$.

6.5.6 Performance Evaluation of TAG-F tool inference Engine

To evaluate the performance of the TAG-F tool inference engine, three-fold cross-validation was performed on the meta-data knowledge base. To achieve this, the dataset was split using a 70:30 ratio for train and test partitions respectively. Table 6-5 presents the results of the average values of the evaluation metrics. As the table shows, the learning algorithm achieved high accuracy (evident in the high evaluation metrics

– precision, recall and f-measure) in predicting the respective classes. Figure 6-5 presents the plotted confusion matrix for the TAG-F tool inference engine. The y-axis represents the actual values, while the x-axis represents the predicted values. The confusion matrix is colour-coded, with the darkness of the shading corresponding to the number of 'true' predictions. As can be seen, the model performed well in the majority of the class predictions provided.

Table 6-5: Performance Evaluation Metrics for TAG-F tool IBL algorithm

Class Label	**Precision (%)**	**Recall (%)**	**f-measure (%)**	**Support**
Neural Network	50	50	50	6
ARIMA	100	100	100	8
Kalman Filter	89	73	80	11
Linear Regression	100	100	100	8
SVR	100	100	100	13
LSTM	75	86	80	7
k-NN	88	100	93	7

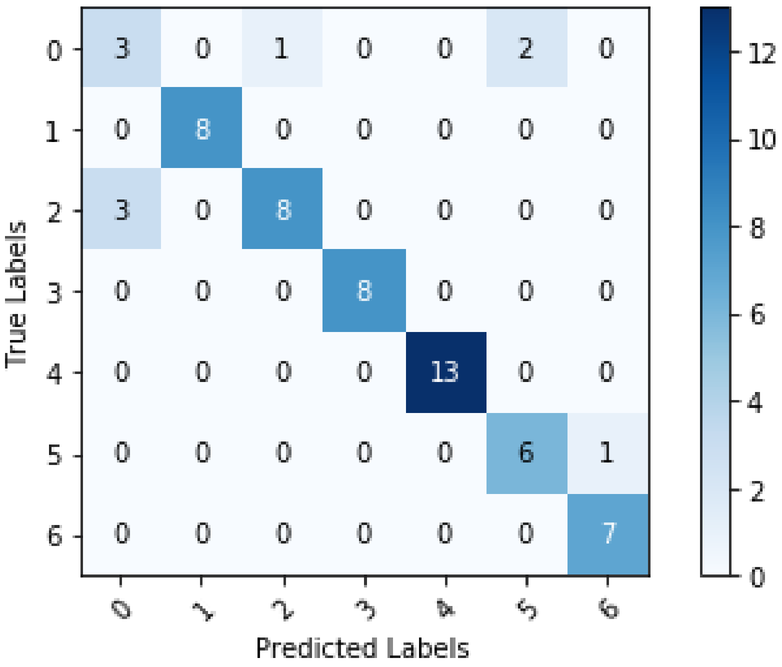


Figure 6-5: Confusion Matrix for TAG-F tool inference engine

6.6 Evaluation of the TAG-F framework using the Support Tool

The TAG-F framework and support tool were both developed to provide structured decision making to traffic data scientists for developing optimal analytical solutions to solve traffic data-driven predictions. In order to evaluate the performance of TAG-F framework and tool in providing TPA guidance empirically, three (3) scenarios were developed. In each of these scenarios, the problem goal is traffic parameter prediction.

These three scenarios have been carefully chosen to demonstrate the applicability of the tool and framework in providing guidance to traffic data scientists in performing TPA. Although the TAG-F tool presented here is a prototype or *proof of concept*, the scenarios can still demonstrate the range of benefits that the framework and tool can offer. For instance, the scenarios are selected to include times where real-time (online) prediction is required, varied prediction horizons, univariate and multivariate traffic prediction, instances of small, medium, and large input datasets respectively. The possibility of using the TAG-F framework and tool to obtain a ranked list of predictive models based on the respective problem scenarios constitutes directional guidance and can be beneficial to organisations and traffic authorities. In the subsequent subsections, the top three (3) PAM suggestions obtained from the TAG-F support tool are quantitatively evaluated by implementing the predictions iteratively. The top three (3) PAMs are evaluated due to the fact that it may not be profitable to evaluate the entire list (i.e. all seven PAMs) since the rules extracted as the meta-level dataset have provided insight on the identification of suitable PAMs. In addition to the PAM suggestion, the support tool also provides a textual justification of the most likely PAM to enable the data scientist to understand the justification or rationale behind the PAM inference. This is presented as a text box inside the PAM suggestion histogram plots, as will be discussed in the various scenarios.

6.6.1 Experimental Setup

We evaluated the TAG-F framework and support tool using traffic datasets about a chosen study area in Greater Manchester, UK. The traffic data used within this study was provided by the Transport for Greater Manchester (TfGM). The provided traffic database comprised per minute observations of traffic flow characteristics (average speed, flow, density), collected using inductive loop sensors. The study area comprised 10 traffic measurement sensors, each of which was 0.3 miles apart on the arterial road. The study area is an urban arterial road (Chester Road - A56) in Stretford, Greater Manchester, UK, between coordinates of longitude and latitude between (53.46281, -2.28398) and (53.43822, -2.31394) as depicted by the pinpoint markers on the map in Figure 6-6. This represents an ideal characteristic of serving as a conduit from a residential area to the city centre. Landmark locations around are the Manchester United Football Stadium – Old Trafford – in addition to other leisure points such as shopping malls (Stretford Mall), clubs, restaurants, etc. Although it is an ‘A’ road, implying that it should be a motorway or freeway/highway, the section under consideration has a reduced speed limit of 30mph due to it being a busy segment, having many pedestrian crossings, business places, and stores.

The weather data obtained during the study period comprised hourly observations of temperature (Celsius) and precipitation (measured in millimetres). The weather data was obtained from the Centre for Atmospheric Studies (CAS), University of Manchester. The weather stations are located within a 3-mile radius of the study area. The workdays were further sub-categorized into peak and off-peak hours.

For each of the three scenarios, no data pre-processing was done. The same input data was applied to all three models in each scenario. The only data manipulation was in the supervised learning models, where the input data was converted from multivariate to supervised learning format using sliding window method and has been described in Section 6.6.3

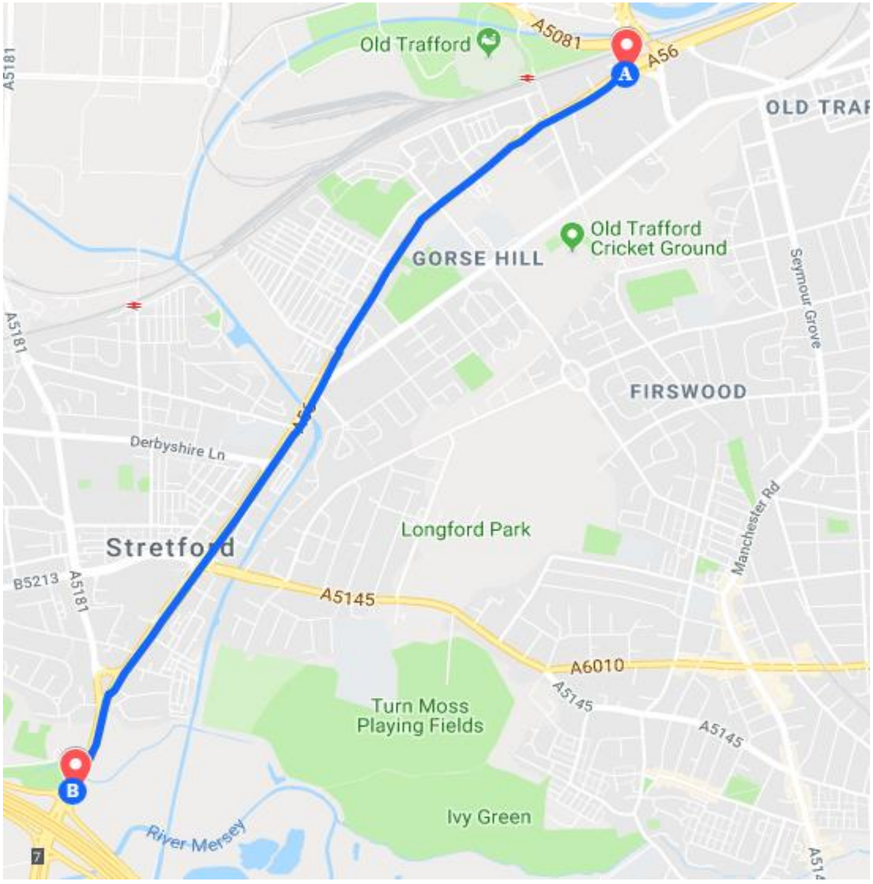


Figure 6-6: Study Area

All experiments were run on a single machine in Windows 10 Operating System with Intel® Core ™ i7-6800K CPU @3.40 GHz, 32-GB Memory, and NVIDIA Quadro K420 GPU. The development environment included Python 3.7, Tensorflow 1.12.0 and R version 3.5.1.

6.6.2 Model Evaluation Metrics

Many performance measures exist, which can be used to evaluate prediction models by computing the discrepancies between the actual (A_t) and predicted (P_t) values. For the purpose of this evaluative study, we adopted two (2) error indexes: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), defined by the equations below.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |e_i| \tag{6.3}$$

$$RMSE(x, y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - y_i)^2} \tag{6.4}$$

Where $\mathrm{e_i}, \mathrm{i} = 1, 2, \dots \mathrm{n}$ represents n samples of modal errors,$\mathrm{x_i}$ and $\mathrm{y_i}$ respectively represent the input and output values.

6.6.3 Converting the univariate time series to a supervised learning format

For some of the suggested models in the presented scenarios, the input time series needed to be converted into a supervised learning format. Nonparametric models such as k-NN, ANN, and SVR model require the data to be in a supervised learning manner prior to model training (Mitchell, 2006). Therefore, the input time series sequence $X_{Speed} = [x_s, x_{s+1}, x_{s+2} \dots x_{s+n}]$ and target variable sequence $Y_t = [x_{s+1}, x_{s+2}, x_{s+3} \dots x_{s+m}]$ must be framed to a format that enables an algorithm to learn the mapping function from the input to the output, which is represented as:

$$y_t = f(X_t) \tag{6-5}$$

Many techniques exist that can enable this data manipulation. The most widely adopted technique refers to the ‘sliding window’ approach (Goodfellow and Bengio, 2015). In this method, a portion of the training data sequence (window) is reframed to be used as input features. The main goal of this process is to enable the calculation of the function that best fits the input data such that, given a set of new/unseen input data X_t, an output vector Y_t can be predicted.

Sliding window data manipulation is a process that involves restructuring a given input sequence of numbers from a time series dataset into a supervised learning manner. This can be achieved by using previous time steps (observations) as input variables (features) and the next time step (i.e. Y_{t+1}) as the output variable. A contrived example is shown below. Consider a simple univariate 5-step input time series vector presented in Table 6-6:

Table 6-6: Sample univariate time series

Index	$\boldsymbol{X_1}$
1	100
2	110
3	120
4	130
5	140

The input vector can be transformed into a supervised learning manner by restructuring it (for instance, using the previous time step to predict the next time step) by reorganizing the data into a supervised learning format (see Table 6-7).

As can be seen from Table 6-7, the previous time step observation is used as the input, while the next time step is the output in the supervised learning format. Another point to note here is that the sequential order is preserved in this transformation, which is key for every time series problem (Weigend, 2018). However, for index 1, there is no previous (known) value of the time series, which explains the N/A, which is the same for index 6 (thus ‘Unknown’). For this contrived example, the step movement stride is single (i.e. equals to 1), and the window size also equals to one, since we are considering only the single prior time step observation. Therefore, the number of previous time steps refers to the *window width/size*. This method of restructuring a time series problem to a supervised learning format is known as a *sliding window method*. In statistics and time series analysis, it is otherwise known as lag observations method. Although the example above used a step-size and sliding window size of one (1) each, this can be increased to include more prior time steps, as well as on multivariate input dataset (as the example showed).

Table 6-7: Transformed univariate time series

Index	$\boldsymbol{X_1}$	$\boldsymbol{Y_t}$
1	N/A	100
2	100	110
3	110	120
4	120	130
5	130	140
6	140	Unknown

Therefore, for this scenario, which involved a single input and output vectors (like the example above), we used a sliding window step size of 1, and sliding window size of 3, in order to agree with the lag observations of the ARIMA(p, d, q) model. Given that the input time series is aggregated in 5-minute observations, it, therefore, implies that our models are trained using the observations of the previous 25-minutes (i.e. prior 5 time steps which refers to the sliding window size) in order to predict the class in the next 5-minutes recursively.

6.6.4 Scenario 1:

Traffic data analysis is applied for controlling or managing the present traffic network status but is also used for environmental and town/city planning. This typically involves microscopic traffic analysis, which can be used to understand, optimize, or analyse local aspects of road traffic networks such as changes in priority regulation at single pedestrian crossings or bus intersections (i.e. lane merge) (Esser and

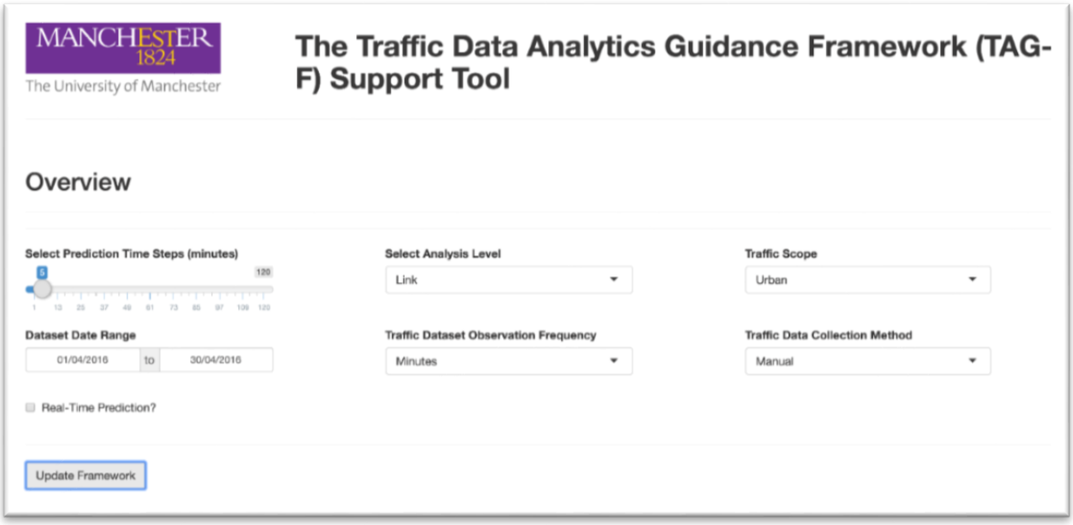
Schreckenberg, 1997) across urban arterials. In order to develop effective systems to optimize these local traffic flow aspects, it is typical for traffic planning/control authorities to obtain short-term predictions (i.e. up to 5-min) of the traffic flow parameter – typically traffic flow. This is mainly done using historical data applied to learning algorithms, and (due to the dynamic and rapidly changing nature of urban traffic) typically requires a medium-to-large and consistent traffic time series data (Hamed et al., 1995; Lana et al., 2018). A typical example of this situation was the planning involved prior to the commencement of road works on Chester Road in August 2018, which would cause a major disruption to an already-busy road segment (MEN, 2018).

This scenario, therefore, presents an attempt at performing TPA for short-term, daily, multi-step ahead, *traffic flow* forecasting. The available training dataset was one month's worth of traffic flow collected using inductive loop devices (ILDs) along the study area. The entire code used for the thesis has been placed on GitHub. In addition, portions of the dataset used have been uploaded on the GitHub repository to enable reproducibility or verification of the results. In this scenario, the prediction problem is offline (i.e. not real-time) short-term traffic speed prediction with a small univariate training dataset. Performing TPA for this scenario presents a challenge to the traffic data scientist in terms of requirements gathering as well as the TPA problem description, which is a prerequisite to the solution developmental process. As stated in Taylor and Bonsall, (2017), the path from problem to solution in TPA is significantly impacted by the understanding and proper description of the problem. Therefore, a structured mechanism for describing the TPA problem space can improve the TPA solution outcome. This is what TAG-F framework offers. Furthermore, prior to performing TPA, several data analytical stages or techniques need to be followed (for instance, exploratory data analysis, problem categorisation, etc.) in order to ascertain the appropriate PAM to be adopted. However, the use of TAG-F and support tool have eased this process.

TAG-F has presented a framework that described the TPA process using three dimensions, as well as some dimension parameters, which have been discussed in detail in Section 5.2. Using the TAG-F framework to describe the TPA problem, the following requirements were identified with respect to the problem definition and mapping to the analytical solution space, presented in Table 6-8. As can be seen from the table, a list of parameters, in line with the TAG-F framework, have been identified. The next stage of the guidance is offered by the TAG-F support tool. This involved manually entering the parameters from Table 6-8 into the TAG-F support tool interface (see Figure 6-7). Upon clicking the 'update framework' button, the ranked list of the top three (3) PAMs from the TAG-F tool (in descending order of inferred suitability) were ARIMA, ANN, and k-NN respectively (see Figure 6-8).

Table 6-8: Scenario 1 TPA requirements

S/No	TAG-F Parameter	Value
1.	Traffic Area (of implementation)	Urban
2.	Data Source Type	Univariate (traffic flow)
3.	Dataset Size	Small
4.	Level of analysis	Link
5.	Prediction Horizon	5-min
6.	Real-time prediction	False
7.	Data Collection Method	ILD



The Traffic Data Analytics Guidance Framework (TAG-F) Support Tool

Overview

Select Prediction Time Steps (minutes): 5

Select Analysis Level: Link

Traffic Scope: Urban

Dataset Date Range: 01/04/2016 to 30/04/2016

Traffic Dataset Observation Frequency: Minutes

Traffic Data Collection Method: Manual

□ Real-Time Prediction?

Figure 6-7: Scenario 1 TAG-F tool Input Entries

Rationalising the suggested PAMs in scenario 1

The TAG-F framework – coupled with the support tool – represents a generalizable and reusable guidance and decision support mechanism to traffic data scientists for performing TPA. In scenario 1, without the guidance offered by the framework and tool, the traffic data scientist faces the demanding task of problem description and understanding, requirements gathering, experimental data analysis, and a number of statistical tests on the actual data in order to infer what PAM may be suitable for the given scenario. Figure 6-8 presents the PAM suggestions obtained from the support tool output. As can be seen, the textual justification of the PAM suggestion is presented, which can provide the data scientist insight about the choice of predictive model to adopt. Given the specific scenario, where a small dataset and prediction horizon is available, then the PAM of choice will be ARIMA.

From the summary of the chosen PAMs presented in Table 6-1, it can be seen that a key advantage of the ARIMA model is its effectiveness in performing short-term prediction over a stationary dataset (i.e. stationarity of mean and variance). This has also been confirmed in studies that have applied ARIMA towards short-term traffic prediction (Kirby et al., 1997; Qiao et al., 2013; Williams et al., 1998). The input dataset in this scenario is one that contains trend and seasonality, typical traffic flow patterns observed in rush/peak hours, and weekends. For this reason, the seasonal variation of the ARIMA model – SARIMA – was configured in this example, with model parameters as shown in Table 6-9.

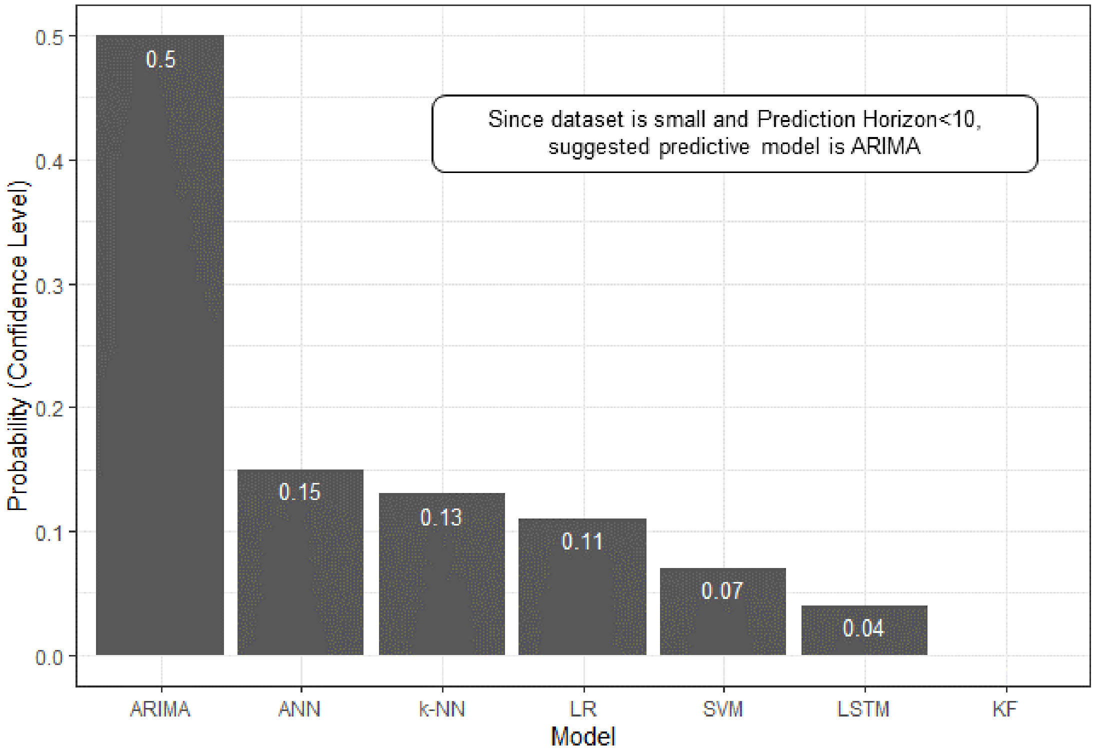


Figure 6-8: Scenario 1 TAG-F tool PAM suggestions

Similarly, the IBL algorithm is a nonparametric, lazy learning algorithm. The term ‘lazy’ means that the model does not make any generalisations from the training dataset, thereby implying a minimal training phase (Mitchell, 2006). A k-NN model is particularly advantageous in instances where there is little or no prior knowledge about the distribution data. In this scenario, owing to the size of the data, and the patterned nature of the data (given the data granularity in 5-min windows), the model seems to be a suitable fit for the predictive scenario. The third model suggested was the ANN model. From Table 6-1, it can be seen that ANNs are advantageous when a relatively large data-to-feature ratio is obtainable. The results from Table 6-9 show that the ANN model marginally performed in comparison to the ARIMA model. However, from Table 6-1, a major drawback of ANN is the fact that the results are not traceable and cannot be rationalised. In addition to this is the extensive computation time requirement, as well as its data-intensive learning process. Given that the scenario was one in which a small dataset was available, as well as the prediction over a short period of time, the ANN, therefore, may not be a recommended choice of PAM. The scenario experiments were conducted using Python scripts using identical datasets, lookback (i.e. lag observations), train-test split, and data pre-processing regimes for each model evaluation execution. For this scenario, the evaluative runs for ANN is repeated thirty (30) times and the average value taken to ensure that this represents the mean value for the non-deterministic (i.e. stochastic) model. The optimal model hyperparameters were obtained by grid search, and these include: learning rate (for optimizer) as 1×10^{-6}, number of epochs for model training as 500, and no dropout was introduced. Furthermore, the loss (cost) function adopted was the RMSE. The optimal ANN configuration was a neural network comprising two hidden layers with 10 neurons per layer (see Table 6-9).

Table 6-9: Scenario 1: Summary of Results

Model	Parameter(s)	RMSE	MAE
ARIMA	SARIMA(1,1,7)(1,1,7,7)[12]	**36.388**	**28.660**
ANN	Hidden layers = 2 Hidden neurons = 10 per layer	41.601	31.2601
k-NN	*k* = 5	43.558	33.139

6.6.5 Scenario 2:

Urban traffic networks are often disrupted by unplanned events, such as accidents, vehicle breakdown, traffic signal malfunction, etc. However, in other instances, this disruption can be planned and announced beforehand. The most common example is a case of events, such as football matches, or road construction works, where notices are provided well in advance prior to the commencement of the construction works. Significant road works projects sometimes result in lane (or road) closures, which greatly impact traffic status. Therefore, there is a constant need for monitoring, evaluation, analysing, and controlling the traffic situation in these conditions to mitigate traffic jam situations.

This scenario presents an attempt at performing TPA in a period (on the chosen study area) where road construction works were being carried out. The construction, which lasted four (4) days, resulted in a lane closure. The consequence of the lane closure was severe traffic congestion, especially at peak/rush traffic hours. In order to manage the traffic congestion situation and execute appropriate control schemes, measures or interventions, the traffic analyst (or ITS) would need the projected build-up of the traffic situation in real-time. For this reason, the traffic data scientist is tasked to perform TPA using the available dataset and make multi-step real-time traffic speed prediction. In this scenario, multivariate time series data was available, which included traffic, weather-related, and tweet messages. As with Scenario 1, the problem goal is short-term traffic speed prediction at a road segment – between the lane closure.

Background to traffic prediction using non-traffic input

Traffic data science has advanced over the years by expanding the number of data sources used to train predictive models (Guo et al., 2014). Existing studies have identified the importance of weather data on traffic flow parameters by influencing driving behaviour (Peng et al., 2018), travel demand, travel mode, road safety, and traffic flow characteristics (Essien et al., 2018; Heilman et al., 2002; Tsapakis et al., 2013). Furthermore, research has – over the years – shown that rainfall reduces traffic capacity and operating speeds, thereby increasing congestion and road network productivity loss.

Recent studies have, therefore, investigated the impact of the inclusion of non-traffic input datasets for urban traffic parameter prediction, many (if not all) of which have yielded improved prediction accuracies (Essien et al., 2019a; Jia et al., 2017b, 2017a). For instance, a deep bi-directional LSTM model was proposed in (Essien et al., 2019a), which was trained using rainfall and temperature datasets in addition to traffic flow characteristics. The findings from the study revealed an improvement in prediction accuracy when compared to baseline traffic-only datasets. Similar results were obtained in studies including non-traffic input data for model training (Jia et al., 2017b, 2017a). This can be explained by the fact that traffic parameter prediction relies on machine learning techniques applied towards data, which is structured in a supervised manner from historical observations to extract patterns that can be used to predict future observations. This has been effective due to the mostly recurring/cyclical nature of urban traffic data. For instance, morning and evening rush hour peaks are easily predictable, and can, therefore, be anticipated. A model that is, therefore, adept at extracting/learning these patterns from the historical dataset will be skilful in predicting future observations.

However, in unusual or non-recurring situations, such as events or incidents that cannot be inferred from historical observations, even the most accurate predictive models will exhibit poor predictive performance. Typical examples of non-recurring or stochastic events/incidents include accidents, road construction works, lane closures, sporting, and public events. Given that such events are sudden, unexpected and rare, the need for developing robust predictive models to enable accurate traffic prediction in these circumstances becomes necessary. For this reason, short-term traffic prediction has sought to incorporate relevant and reliable sources of information about non-recurring or sudden events that may impact traffic status. For instance, road users can be frustrated, when stuck in traffic congestion, and sometimes choose to vent out the frustration in the form of tweets about the traffic situation on their respective timelines, which indirectly serve as information to upstream or future road users.

Social media, as an online discussion platform, has seen a remarkable explosion in the last few years. Examples include Facebook[8], Twitter[9], Instagram[10], and Snapchat[11]. These services are widely employed for communication, news reporting, and advertising events. Each of these social media platforms provides application programming interfaces (APIs) that enable data retrieval in real-time. Twitter is a public social media platform popular for short messages (up to 280 characters), thereby resulting in data streams with high velocity and timely dissemination of information concerning real-world events. Given the enormity and variance of information obtainable on twitter due to the large user base, numerous studies have sought to harness this online data repository for various data mining purposes, such as stock market prices (Bollen et al., 2011; Nguyen et al., 2015), crime rate prediction (Wang et al., 2012), and traffic prediction (Abidin et al., 2015; Goh et al., 2018; Wongcharoen and Senivongse, 2016).

[8] https://facebook.com/
[9] https://twitter.com/
[10] https://www.instagram.com/
[11] https://www.snapchat.com/

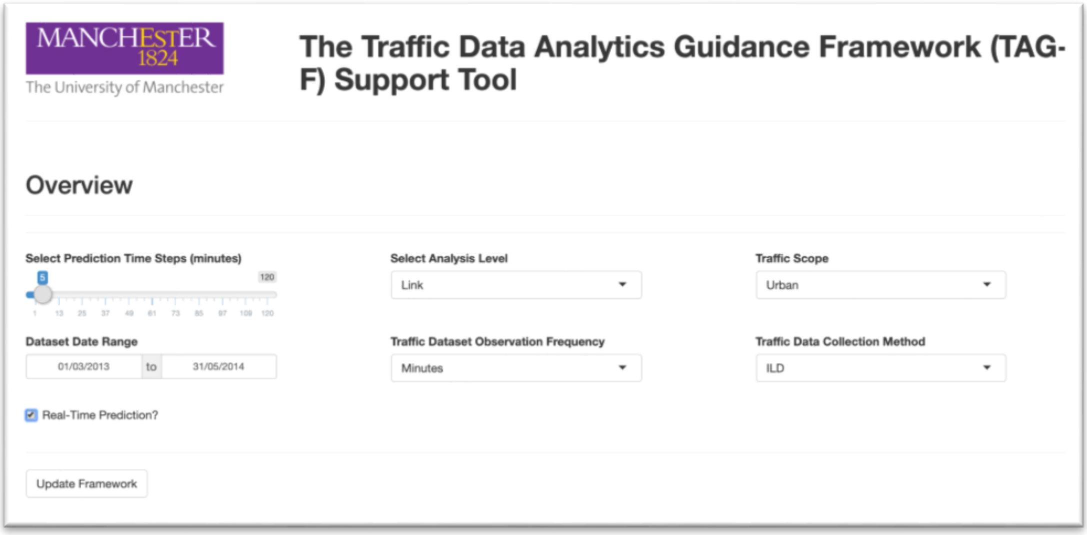
Advanced Traveller Information Systems (ATIS), such as Waze and TomTom, already capitalise on crowd-informed social media data to improve their traffic navigation and route guidance system. In general, many twitter accounts report current traffic conditions, which can be used by road users to infer future traffic conditions and inform the choice of travel mode. For instance, in Northern England, Highways North West England (*@HighwaysNWEST*), Traffic for Greater Manchester (TfGM *@OfficialTfGM*), *@nwtrafficnews*, and Waze *(@WazeTrafficMAN*) are typical examples of such Twitter accounts that provide road-traffic condition information. In addition to tweets posted by major organizations in the transportation sector, road users can also tweet on their respective timelines to broadcast (to their followers) current road traffic conditions, which can be mined to infer future traffic conditions.

The study period in this scenario spanned from 1 April 2016 to 16 April 2017. For this scenario, a long training history is required. This is because of the seasonality, trends, and cycles observed with the different times of the year (for instance, during school holidays, peak periods, etc.). The input data was aggregated into 2-min time steps, in order to enable multi-step prediction, which would enable a near real-time evaluation and visualization of the traffic situation by the traffic analyst. Thus, the input data comprised of 262, 800 observations of six (6) variables/features of traffic speed, volume, rainfall, temperature, and the tweet dataset. A significant challenge encountered in the inclusion of tweets in traffic prediction is the process of determining the level of authenticity, veracity and filtering high levels of noise in the unstructured datasets (Goh et al., 2018). To account for this, the tweets utilised for this scenario included tweets from road-traffic organisation Twitter accounts – specifically Transport for Greater Manchester (*@OfficialTfGM*) and Waze (*@WazeTrafficMAN*).

Similar to the case in scenario 1, the TAG-F framework and support tool are together used to provide guidance to the user (i.e. traffic data scientist) both in describing the TPA problem space, and suggesting PAMs that may be suitable for the given scenario. Therefore, using TAG-F framework dimension and parameters to characterize the TPA problem space, the configuration presented in Table 6-10 was realised with respect to the problem definition and subsequent mapping to the analytical solution space.

Table 6-10: Scenario 2 TPA requirements

S/No	TAG-F Parameter	Value
1.	Traffic Area (of implementation)	Urban
2.	Data Source Type	Multivariate (traffic flow, speed, density, rainfall, temperature, tweet messages)
3.	Dataset Size	Large
4.	Level of analysis	Link
5.	Prediction Horizon	5-min
6.	Real-time prediction	True
7.	Data Collection Method	Bluetooth


MANCHESTER 1824
The University of Manchester

The Traffic Data Analytics Guidance Framework (TAG-F) Support Tool

Overview

Select Prediction Time Steps (minutes): 5

Dataset Date Range: 01/03/2013 to 31/05/2014

☑ Real-Time Prediction?

Select Analysis Level: Link

Traffic Dataset Observation Frequency: Minutes

Traffic Scope: Urban

Traffic Data Collection Method: ILD

Figure 6-9: Scenario 2 TAG-F tool Input Entries

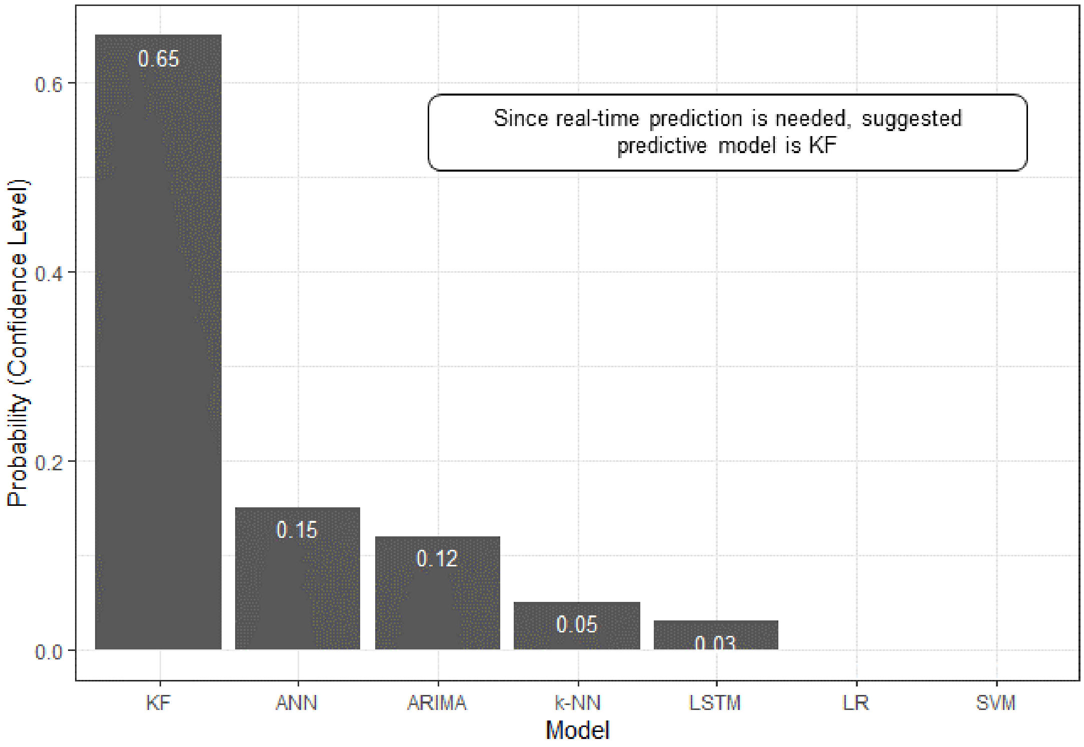


Figure 6-10: Scenario 2 TAG-F tool PAM Suggestions

Rationalising the suggested PAMs in Scenario 2

The second scenario presented an attempt at remodelling a period where road construction works were being carried out on a section of the study area. For this reason, real-time traffic prediction was required by the traffic analyst in order to provide traffic congestion control measures. Once again, as can be seen, the TAG-F framework enables a clear articulation of the TPA problem space. By identifying the factors that need to be considered in TPA, the traffic data scientist can know what parameters or values are to be used in order to define the problem space properly. The accurate requirements gathering and a requisite problem description is a prerequisite for the solution development process. Secondly, through the support tool, the parameters obtained from the TPA problem description stage can enable the traffic data scientist to know which model can be applied in the given scenario.

In this scenario, manually entering the parameters from Table 6-10 into the support tool (see Figure 6-9), the top three (3) suggested PAMs were KF, ANN, and ARIMA respectively (see Figure 6-10). The following paragraphs present a rationalisation about the guidance provided in the form of PAM suggestion. From Figure 6-10, the rationale behind the suggestion of the KF algorithm is since a real-time prediction is required. A few key parameter values can be observed from the TPA requirements identification articulated in Table 6-10. First, the required prediction is in real-time, which means that the model to be used is required to be fast, efficient, and scalable. Secondly, a multi-step or recursive prediction is required in short time steps. From Table 6-1, the Kalman filter, being a recursive estimator, would constitute the model of choice for real-time prediction. Secondly, an advantage of the Kalman Filter is its ability to take into account quantities that are partially or outright rejected in other algorithms (Welch and Bishop, 1995). In addition, due to the large dataset size, (i.e. 280K × 6 variables), the application of the KF algorithm will be a recommended choice. Thirdly, the KF is computationally 'lightweight' and can perform multivariate time series prediction.

Table 6-11: Scenario 2: Summary of Results

Model	Parameter(s)	RMSE	MAE
ARIMA	N/A	N/A	N/A
KF	Sig_act = 0.1, pred_sig = 0.3, mu = 0, sig = 1000	**19.691**	**16.327**
ANN	Hidden layers = 8 Hidden neurons = 10 per layer	33.063	25.250
	Hidden layers = 1 Hidden neurons = 10 per layer	34.281	25.600
	Hidden layers = 1 Hidden neurons = 20 per layer	34.565	25.798
	Hidden layers = 1 Hidden neurons = 30 per layer	34.345	25.576
	Hidden layers = 2 Hidden neurons = 10 per layer	34.361	25.584

	Hidden layers = 2 Hidden neurons = 20 per layer	34.560	25.761
	Hidden layers = 2 Hidden neurons = 30 per layer	34.446	25.664

For these reasons, as well as the predict-assess-update cycle obtainable in the KF, it would be a suitable choice of algorithm for this scenario, evident in the results obtained, as shown in Table 6-11. The second model suggested by the support tool was the ANN. Given that the prediction is required in real-time, deep learning neural network models and ANNs may not be suitable due to the large training time required, which would defeat the essence of real-time predictive requirement. For this scenario, the hyper-parameters for the ANN were obtained using a grid search framework that was developed by the author and is published online[12]. For the model, the optimiser adopted was a stochastic gradient descent-based adaptive algorithm referred to as Adam (Kingma & Ba, 2014). The learning rate was 1×10^{-3}. In the ANN, each hidden layer had the Rectified Linear Unit (ReLU) activation function, while the output layer had a linear activation function (given that the problem is a regression problem). Although the TAG-F tool suggested ARIMA as the third alternative, it is, however, inappropriate as ARIMA on its own cannot perform multivariate prediction. In order to use ARIMA for multivariate prediction, the vector ARIMA model can be used (Smith et al., 2002). Therefore, in this instance, the support tool provided a wrong PAM suggestion. This error in PAM inference can be mitigated by increasing the meta-dataset size, which will result in a more robust and accurate inference engine for the support tool. This has been included as future work, presented in Section 8.7.

[12] https://github.com/nakessien/tagf_evaluation.git

6.6.6 Scenario 3:

Urban traffic networks are significantly impacted by road accidents or other incidents that impede road capacity (C. Wang et al., 2009). Some accident situations cause entire road closure, as a safety measure prior to the arrival of medical, police, or emergency services. This significantly impacts the traffic situation, especially at peak periods. This scenario re-models a historical road accident along a segment of the study area that resulted in lane closure and severe congestion. The accident happened on May 2, 2014, at 13:04 hrs. Accident data were obtained from an online[13] database, which made available a report that contained details about the incident/accident, which include the date and time, geo-coordinates of the concerned road, cause of incident, lane(s) closed, etc. Figure 6-11 shows a snippet of the specific accident location along the road segment.

In order to facilitate congestion management, the traffic analyst may want to see hourly step-ahead predictions of traffic flow parameters (traffic speed in this case). This can be rationalized by the fact that the analyst would require a medium-term prediction (i.e. 60-minutes) due to the time for which the police, ambulance, fire or emergency services may arrive the scene. It may not be useful for the traffic analyst to require short-term predictions, as this may not adequately enable effective congestion management and control in the given circumstance. For this scenario, the traffic data were aggregated into 5-min intervals to match the requirement of the traffic data scientist. Furthermore, the dataset contained 245,376 observations of the five (5) variables – speed, flow, density, rainfall, and temperature, which is 852 days' worth of input data. Similar to scenarios 1 and 2 above, characterising the TPA problem according to the TAG-F dimension elements, the following analytical mapping was

[13] https://www.crashmap.co.uk/Search

actualised with respect to the problem definition (see Table 6-12). The TAG-F support tool was used to provide guidance in a similar manner to the previous scenarios. Entering the dimension parameters into the support tool interface as shown in Figure 6-12 resulted in a ranked list of PAMs (see Figure 6-13).

Table 6-12: Scenario 3 TPA requirements

S/No	TAG-F Parameter	Value
1.	Traffic Area (of implementation)	Urban
2.	Data Source Type	Multivariate (traffic flow, speed, density, rainfall, temperature)
3.	Dataset Size	Large
4.	Level of analysis	Link
5.	Prediction Horizon	60-min
6.	Real-time prediction	False
7.	Data Collection Method	Bluetooth

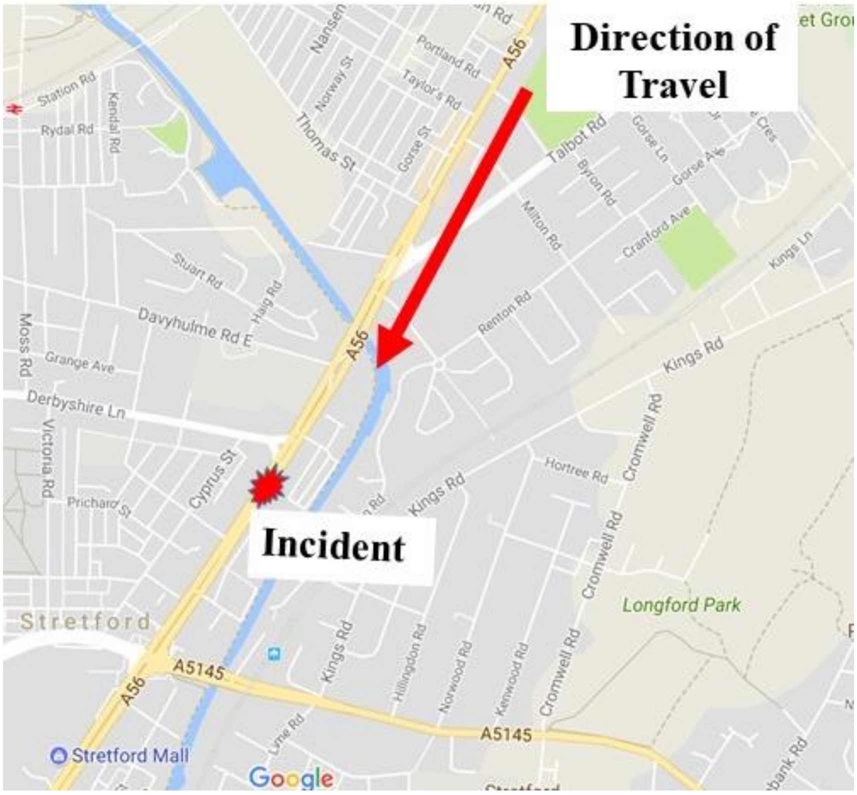
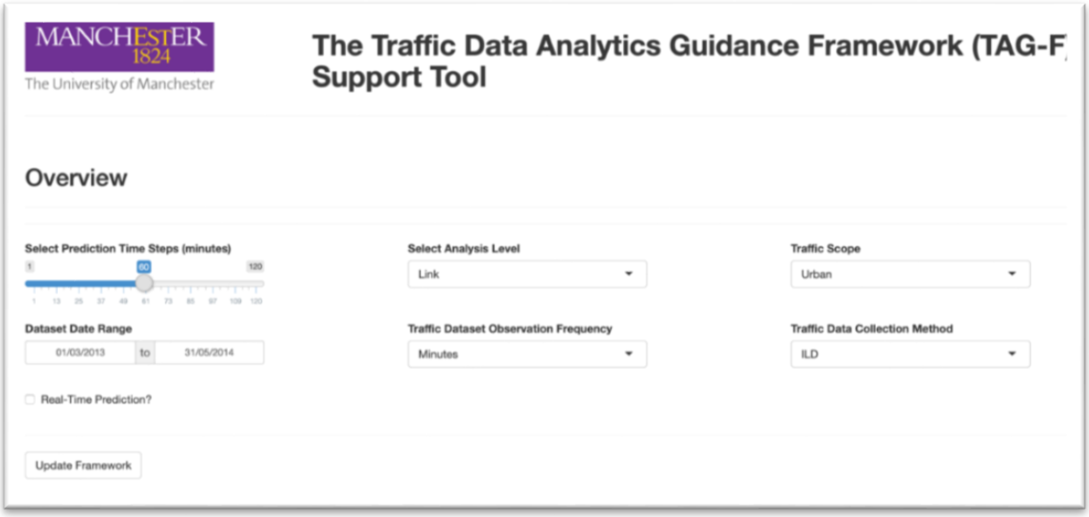


Figure 6-11: Study Area showing the incident location

<u>Rationalising the suggested PAMs in Scenario 3</u>

In this scenario, the suggested models were LSTM, ANN, and *k*-NN, respectively. The results of the scenario evaluation in Table 6-13 showed that the LSTM model significantly outperformed the other models. This is rationalised by the fact that the LSTM models are able to adequately capture and account for the sequential input of time series data, as has been proven in many TPA studies (Essien et al., 2019a; Jia et al., 2017a; Ma et al., 2015). The TPA problem description in scenario three is one that involves a large prediction horizon, large dataset size, and offline (i.e. not real-time) prediction. From Table 6-1, it can intuitively be gathered that one of the advantages of nonparametric models, such as deep learning models and ANN-based nonlinear models, is their effectiveness when making predictions over long prediction horizons (Barros et al., 2015; Ishak and Al-Deek, 2002).


The Traffic Data Analytics Guidance Framework (TAG-F
Support Tool

Overview

Select Prediction Time Steps (minutes): 60 (1–120)

Select Analysis Level: Link

Traffic Scope: Urban

Dataset Date Range: 01/03/2013 to 31/05/2014

Traffic Dataset Observation Frequency: Minutes

Traffic Data Collection Method: ILD

□ Real-Time Prediction?

Figure 6-12: Scenario 3 TAG-F tool Input Entries

Figure 6-13 presents the PAM suggestions for scenario 3 from the support tool. As can be seen, the justification of the LSTM suggestion is due to the fact that a large dataset and long prediction horizon is required. Specifically, given the nature of the TPA problem in scenario 3, a model that can capture the complex, nonlinear, dynamic nature of traffic prediction over a long prediction horizon is required. The main advantage of LSTMs is their ability to accurately capture dynamic and stochastic temporal dependencies in time series data over long time intervals. Therefore, the LSTM will be the model of choice in this scenario.

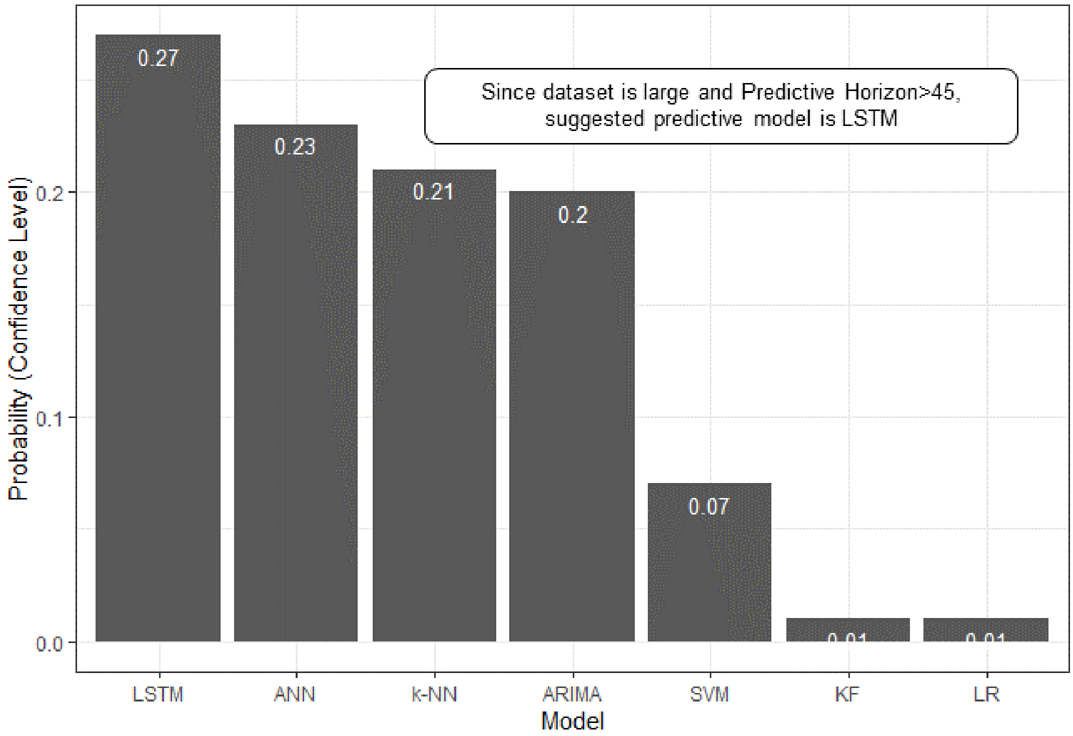


Figure 6-13: Scenario 3 TAG-F tool Model Suggestions

The results of the empirical analysis conducted in this scenario are presented in Table 6-13. As can be seen, the LSTM network significantly outperformed the other models in terms of prediction accuracy. The use of LSTM-neural networks for traffic prediction is becoming very popular in academic studies. A number of research studies have shown that deep learning algorithms have the ability to make accurate traffic flow predictions when compared to their linear/parametric or shallow learning counterparts, especially over long prediction horizons (Essien et al., 2019a; Essien and

Giannetti, 2019; Poonia et al., 2018). A major advantage of this class of models is its inherent ability to perform multivariate data modelling, as well as accepting multi-dimensional data. However, a demerit of this class of prediction models is the extensive training time due to complex model structure. Given the complexity of the internal model structure, the process of retraining the model is time-consuming, which may be a drawback for making predictions over a short prediction horizon. However, given that the prediction is not required in real-time, and the data context specifications, it is appropriate to use this model.

Table 6-13: Performance Evaluation of Scenario 3

Model	Parameter(s)	RMSE	MAE
LSTM	Model depth (layer size) = 5 Model width (number of units) = 1024 Number of epochs = 300 Batch size = 8 Dropout rate = 0.2 Optimizer = Adam Learning rate = 1×10^{-3} Lookback = 12	**0.342**	**0.117**
ANN	Hidden layers = 2 Hidden neurons = 10 per layer	3.2273	2.4302
k-NN	*k* = 14	3.1477	1.556

6.6.7 Extending the guidance provided by the TAG-F support tool

In Scenario 3 (Section 6.6.6), the PAM suggestions provided by the TAG-F tool are beneficial to the traffic data scientist, and not very much to the traffic analyst. Although the two job functions work together towards effective traffic congestion management and control, their paths vary in some ways, as pointed out in Section 1.1. Therefore, in this scenario, while the traffic data scientist understands and appreciates the suggestions in terms of PAMs (i.e. LSTM, ANN, *k*-NN, etc.), the traffic analyst is only interested in decision support towards effective traffic congestion management, and may, therefore, have little knowledge about the methods and data analytics.

However, there is the potential to extend the guidance offered by the TAG-F framework and support tool to the benefit of the traffic analyst.

In order to extend the functionality of the guidance mechanism to benefit the traffic analyst, a proactive traffic visualization model can be applied. This model referred to as *Prediction Simulation-based Traffic Management Model* (PRESIMM) in (Essien et al., 2019b), integrates traffic prediction and microsimulation to provide a visualization of the future traffic state. The argument that supports this model is the introduction of some proactive level into the traffic prediction/control cycle. The PRESIMM model is a novel two-stage model for proactive traffic management. The model comprises two stages:

i. A Prediction stage

ii. A traffic microsimulation stage via a traffic microsimulation tool (Simulation of Urban Mobility [SUMO]), which allows for the simulation of future traffic states in scenarios that involve changes in traffic conditions.

According to Vlahogianni et al., (2014), the literature on short-term traffic forecasting is extensive and has typically considered single data points, mainly employing univariate predictive models. Early research in short-term prediction mainly applied statistical methods like Auto-Regressive Integrated Moving Average (ARIMA), which ignored the spatial dependency of traffic (Karlaftis and Vlahogianni, 2011). The second generation of prediction methods saw the rise of non-linear predictive models such as neural networks, and kernel-based algorithms, which exposed the inherent vulnerabilities in classical predictive models. Simulation-based methods for traffic prediction have been explored in the literature within the last two decades

(Abdelghany et al., 2000; Dombalyan et al., 2017; Nafi et al., 2015; Xu and Dailey, 1995). The success of such approaches has resulted in increased research interest towards simulation-based traffic prediction models. For instance, online traffic network simulation models have been integrated with real-time decision support systems for integrated corridor management (ICM) (Hashemi and Abdelghany, 2016).

However, many studies involving simulation-based traffic prediction focus on traffic parameter prediction at an individual or microscopic level, rather than at the network (macroscopic) level (Hashemi and Abdelghany, 2016). Table 6-14 presents a summary of existing studies in simulation-based traffic prediction. As the table shows, only two models have decision-support functionalities. The two simulation-based real-time traffic management systems (DYNASMART-X (Abdelghany et al., 2000) and DynaMIT (Ben-Akiva et al., 1998) that provide real-time traffic short-term prediction and have decision support capabilities for traffic management, however, sacrifice traffic prediction (i.e. network state estimation) accuracy for simulation latency, thereby ignoring non-traffic input factors capable of significantly affecting traffic state, such as rainfall and temperature (Essien et al., 2018). Furthermore, a limited effort is put into the development of real-time proactive simulation-based traffic systems with decision-support capabilities that incorporate robust and accurate non-linear predictive algorithms, such as deep learning networks, due to the computational and data demands of such models. The intent of providing a degree of proactive-ness involves creating accurate representations of the future traffic state, which is not adequately represented by a simulated current network state. PRESIMM, therefore, is a system that accurately captures the present traffic state using traffic and non-traffic input data sources on short-term traffic parameter prediction of the network state.

Table 6-14: Summary of Model-based traffic prediction studies in the literature

Paper	Area[14]	Decision Support[15]	Predictive Model	Spatio-temporal	Real-Time?
(Dombalyan et al., 2017)	M	N	Entropy Maximization	N	N
(Fountoulakis et al., 2017)	M	N	Kalman Filter	Y	Y
(Zhou et al., 2017)	M	N	RNN	N	N
(Abid and Hussain, 2017)	U	N	Fast Simulation	N	F
(Abdelghany et al., 2000)	U	Y	Kalman Filter	Y	Y
(Lai et al., 2016)	U	N	Neural Network	Y	Y
(Zhu et al., 2016)	U	N	ARIMA	N	Y
(Ben-Akiva et al., 1998)	M	Y	Fast Simulation	Y	Y

The simulation stage within the PRESIMM model adopted an open-source traffic microsimulation tool known as SUMO[16]. Chen and Cheng (2010) present two main open-source agent-based traffic simulation software: Multi-Agent Transport Simulation Toolkit (MATSIM) (Horni et al., 2016), a toolbox for implementing largescale agent-based simulations, and Simulation of Urban Mobility (SUMO) (Behrisch et al., 2011), a portable microscopic road traffic multi-agent simulation package designed to handle large road networks. Apart from the two mentioned above, a number of multi-agent simulation packages exist, such as VISSIM, VISUM, etc. After careful consideration, this study decided to adopt SUMO as a traffic simulation tool. This stems from SUMO being an open-source tool, with space continuous and time discrete capabilities, providing individual routes for vehicles start and end times and positions, and most importantly, SUMO allows for the import of maps or networks from OpenStreetMap[17], an open-source online map.

[14] M: Motorway/Highway, U: Urban
[15] Y: Yes, N: No
[16] Can be found online at: https://www.dlr.de/ts/en/desktopdefault.aspx/tabid-9883/16931_read-41000/
[17] OpenStreetMap can be found at: https://www.openstreetmap.org/

The first task of the simulation stage (in PRESIMM) using SUMO involves the generation of the road network, which was downloaded from OpenStreetMap. Once imported to SUMO, the map, alongside the predicted speeds and volumes are used to create the simulation of the road network in SUMO using a built-in application package *DFROUTER* - which uses inductive loop values to compute respective vehicle routes using a variant of the classic car-following model (Gipps, 1981) developed by German transport company DLR ("DLR - Institute of Transportation Systems - SUMO – Simulation of Urban MObility," 2019). SUMO allows for customizations to be performed on the virtual road network such as lane closure/opening, traffic light signal alteration, etc. which would offer the possibility of performing detailed analyses on the future or predicted traffic network state.

Figure 6-14 shows the conceptual control loop showing the traffic analyst, data scientist, and how PRESIMM model can be extended to provide guidance in TPA and traffic congestion control. In the diagram, the traffic analyst is represented with a green icon, while the traffic data scientist is represented using the blue icon. The conceptual model can be visualised by traversing from the top right (using the arrows), where the TAG-F framework and support tool is used to provide directional guidance to the traffic data scientist. After the TPA problem description (i.e. articulation of the TPA problem space) and the suggestion of alternate PAMs for the TPA process using the TAG-F support tool. As depicted in Figure 6-14, the choice of the 'final' intervention/control measure to be applied by the control personnel is made easier by the iterative feedback control cycle in PRESIMM, where the control personnel can have a number of 'what-if' situations, before arriving at the intervention that is most effective. In this way, the guidance provided by the TAG-F framework and support tool can be extended to the traffic analyst by incorporating the outputs from the TPA process as inputs to the traffic management system, where the traffic analyst operates.

The TAG-F framework can be integrated with PRESIMM to provide proactive traffic management guidance to traffic analysts and data scientists. It must, however, be mentioned here that this functionality has yet to be incorporated into the TAG-F tool, due to constraints of resources. However, this will form the future research direction for this study.

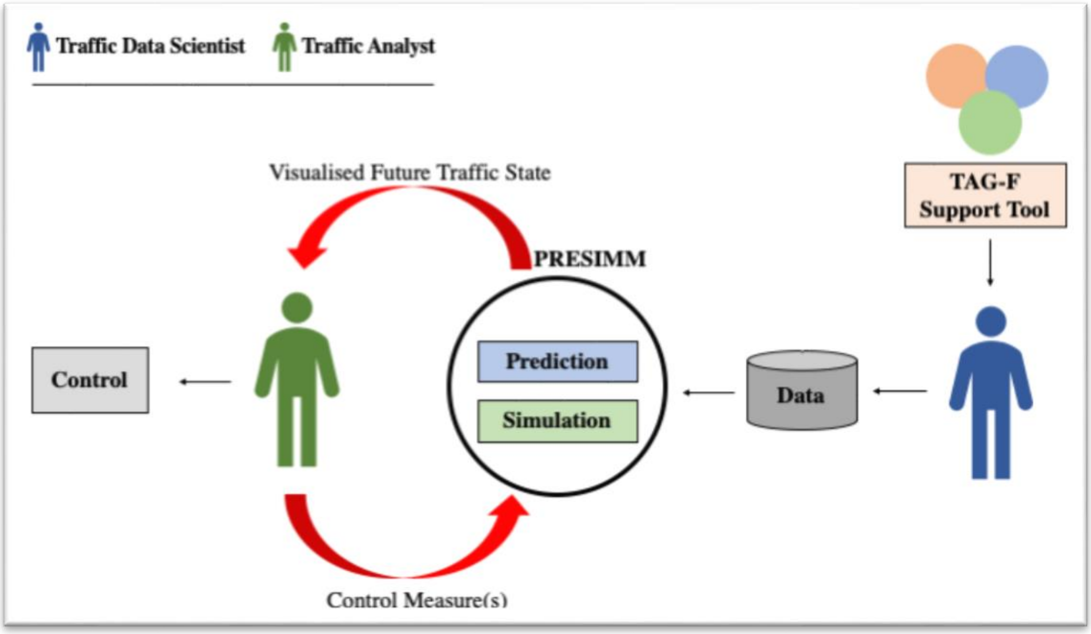


Figure 6-14: Conceptual model of PRESIMM

In providing guidance to the traffic analyst, the author of this thesis presented a study that demonstrated the applicability of PRESIMM towards urban traffic management (Essien et al., 2019b). In the study, it was assumed that as a way of controlling the traffic congestion brought about by the accident, the traffic analyst was able to choose from three available traffic control measures:

i. Add an extra lane (i.e. free up the bus lane),

ii. Alter the signal priority for the junction traffic light, and

iii. Divert upstream traffic to a link road.

In the study, visualisations of the consequences of the applied control measures were developed using PRESIMM. The results of the respective control measures are discussed in the subsequent paragraphs. First, the predicted traffic parameters (1-hour ahead) were passed as inputs into SUMO and visualised (see Figure 6-15).

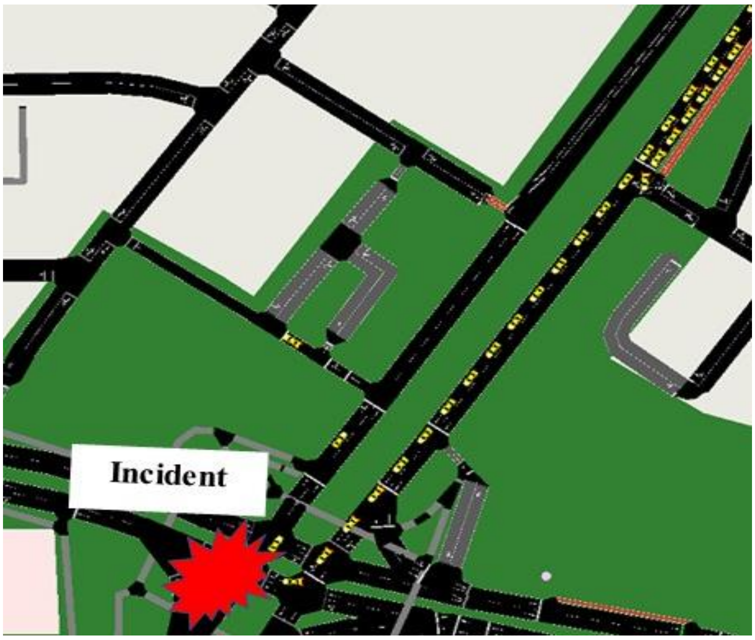


Figure 6-15: Congested Junction as a result of the incident (1-hour after)

<u>Traffic Control (TC) 1: Adding an extra lane</u>

Most cities introduce bus lanes, also known as Bus Rapid Transit (Levinson et al., 2002), as a means of speeding up public transport to control congestion and encourage public transport usage. By this, a dedicated lane is set aside for buses and taxis or emergency services. In this scenario, the control analyst may decide to free up the bus lane (i.e. open it up for use by private vehicles). In the PRESIMM virtual network, this is equivalent to adding an extra lane. The simulation run for TC-1 showed an increase in the traffic flow (i.e. vehicles/hour), indicating that the network reaches its overload or saturation point quicker than the in original 1-hour-ahead simulation (see Figure

6-15). This implies that TC-1 resulted in a more congested network 1-hour after its implementation (see Figure 6-16).

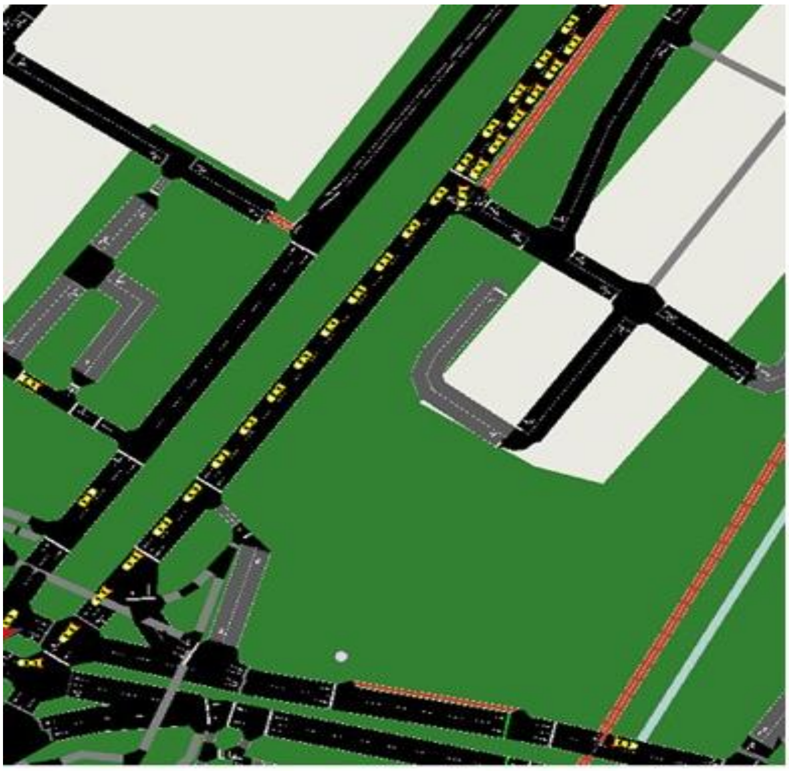


Figure 6-16: Consequence of opening up a bus lane (1-hour later)

<u>TC 2: Signal Alteration</u>

Adaptive Traffic Control Systems (ATCS) are used for traffic control by automatically adjusting traffic signals based on traffic conditions. The major aim of ATCS is to maximize road network throughput (He et al., 2012). SUMO allows for signal alteration, using its interactive interface – NETEDIT. The simulation run for TC-2 is presented in Figure 6-17, with the consequence (i.e. 1-hour later) shown in Figure 6-18. The simulation showed a network that experiences free flow for a short period, accompanied by a rapid increase in traffic flow, which resulted in network overload and congestion in the adjacent road link.

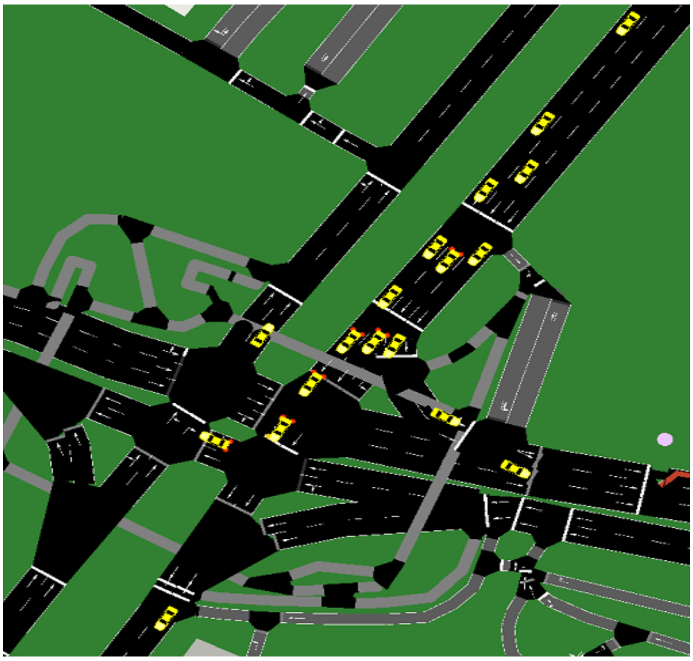


Figure 6-17: Traffic Signal Alteration

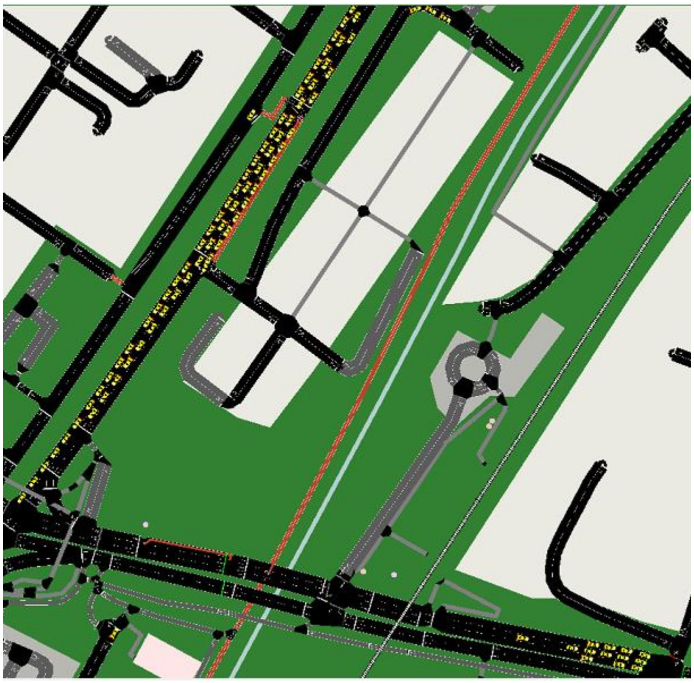


Figure 6-18: Signal Alteration Causes congestion at the adjacent road (1-hour later)

<u>TC 3: Road Diversion</u>

In very congested traffic situations caused by significant reductions in road capacity (brought about by accidents or road construction works), road diversions are employed to free up the network. It is advantageous in that it frees up the congested road but has the disadvantage of ‘transferring’ the congestion to the link road. The simulation run for TC-3 is presented in Figure 6-19 and Figure 6-20. The visualisation of the consequence of this action can be seen as the affected road is freed up once all traffic is diverted to the link road. TC 3 appears to be the most viable control measure, even from a non-technical ‘common-sense’ standpoint. This is because the accident caused a reduction in the road capacity (by 50% due to the closure of a lane), and the intuitive optimal solution would be to divert the upstream traffic to a link road.

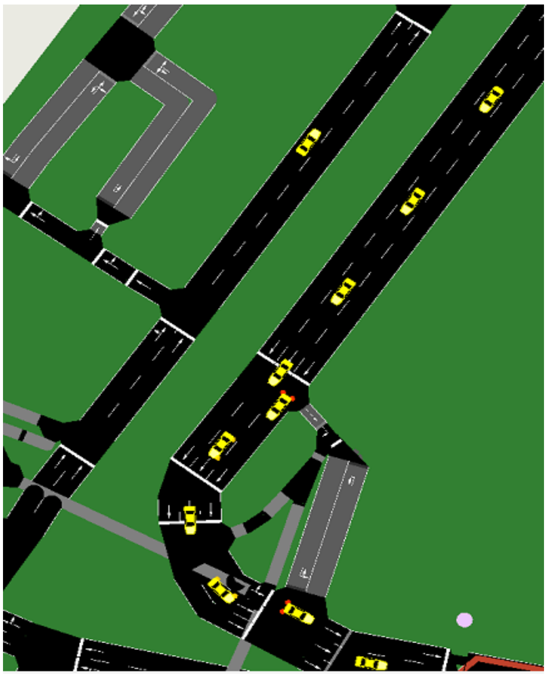


Figure 6-19: Traffic Diverted to link road

6.7 Chapter Summary

This chapter presented the evaluation of the TAG-F framework using the support tool for traffic predictive analytics guidance. The chapter began by presenting a brief discussion about literature-based discovery, a process of extracting knowledge from existing articles. In Section 6.3, an LBD process and the TAG-F framework were used to populate a meta-knowledge base about seven (7) predictive models popularly used in TPA research studies. A brief discussion of the literature-based meta-learning approach adopted within the TAG-F framework was also presented. The process commenced with the extraction of relevant data characteristics and base-level prediction algorithms/models from a set of relevant published research articles, aligning them with the TAG-F dimensions and dimension elements, and finally feeding the combined dataset into an instance-based learning inference algorithm for model suggestion.

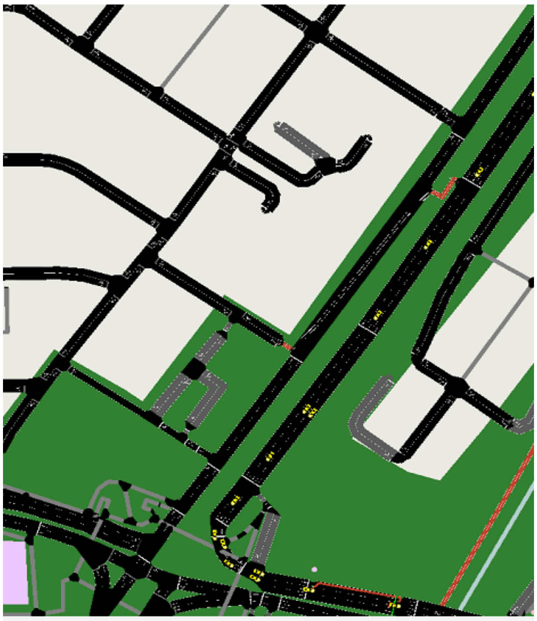


Figure 6-20: Diversion reduces traffic congestion (1-hour later)

The TAG-F support tool was presented in Section 6.4. The tool, which provides semi-automated predictive model/algorithm suggestions for traffic prediction, was discussed, including its design, architecture and implementation. The tool presented in Section 6.4 represents a general approach towards predictive model guidance in TPA. Coupled with the TAG-F framework for describing the TPA problem space, the approach is generic enough to be utilised in diverse TPA problem situations. Although the framework and tool are restricted to seven (7) predictive algorithms, in reality, this number can be extended to accommodate more PAMs. The procedure for updating, evolving, or extending the framework is similar to the work presented in this thesis –

systematic literature review, meta-knowledge extraction about additional PAMs, and model training/re-training.

Furthermore, the support tool was evaluated using three scenarios of traffic prediction problems using sensor collected data from a given road network in Stretford, Greater Manchester, United Kingdom, as presented in Sections 6.6.4, 6.6.5, and 6.6.6 respectively. The results from the scenarios provided an indication of the potential value of the guidance offered by the framework. In each of the three scenarios, the traffic data scientist benefited from the TPA problem description and articulation via the TAG-F framework, as well as the suggestion of alternate PAMs that can be used in the given TPA problem.

Chapter 7 Discussion

7.1 Introduction

This chapter discusses the key findings obtained from the proposed framework and its evaluation. The chapter begins by discussing the findings obtained from the framework evaluation (using the support tool) via the three case scenarios in Section 6.6. In Section 7.5, a characterisation of the TPA stakeholders is presented, listing the road users, traffic analysts, and the traffic data scientists, including their guidance requirements or needs. A conceptualisation of the TPA process for the various stakeholders is presented in a graphical format. This chapter also discusses the summary of findings of the research questions proposed in Section 1.2 and concludes in Section 7.10.

7.2 Discussion of empirical evaluation

Section 6.4 presented a prototype tool, called the TAG-F tool, which was developed to demonstrate the practicality of the guidance that can be provided to traffic data scientists in TPA. It presented three practical scenarios involving traffic prediction. The scenarios demonstrated the application of the TAG-F framework and tool to provide predictive analytics guidance, which can be used by traffic data scientists for solving traffic prediction problems. All the scenarios used data collected from a road segment in A56 (Chester Road) in Stretford, Greater Manchester.

The TAG-F framework presented in Chapter 5 has three (3) dimensions, each of which has elements within. As one would appreciate, a ‘brute force’ attempt at each possible combination of the three dimensions within the framework to develop a robust knowledge base would be a highly demanding task that could run into many years' worth of effort. In addition, it is almost impossible to perform a like-for-like comparison of each of the prediction outcomes suggested by TAG-F tool. For instance, there is a degree of subjectivity introduced when one considers the different outputs of traffic data scientists in performing data pre-processing, hyperparameter optimisation, etc. Therefore, beginning with these three case scenarios, an extendable knowledge base can be developed, which can eventually contain more instances or scenarios of the operationalisation of the TAG-F framework alongside the results obtained. Further exploration of the various combinations of DC, DCM, and support from the tool would expand the rapidly evolving knowledge base of data-driven traffic prediction. This can serve as a foundation for improving knowledge about traffic predictive analytics, which can culminate in the development of traffic prediction ontologies, thereby saving traffic data scientists a reasonable time for predictive analytics.

7.3 Identifying the key dimensions of TPA

Within this thesis, in Section 1.1, the terms traffic analyst and traffic data scientist are defined. The traffic analyst or control personnel is involved with the traffic network or system planning process and is directly responsible for traffic network control and management. This role differs from the traffic data scientist, who is a data scientist and predictive analytics domain expert. Within the existing literature, it is unclear who is/are the main stakeholder(s) of a TPA process (Taylor and Bonsall, 2017), and the findings from this study have enabled a distinction between the two main roles.

This thesis demonstrates that the provision of guidance to traffic data scientists performing TPA via a structured, well-defined description of the TPA task, as well as meta-knowledge about predictive models, results in improvements in the TPA solution development process. In order to achieve this, it is essential to delineate TPA into key dimensions. Data-driven prediction adopts algorithms that fit predictive models to training data, thereby making forecasts. The findings from this study identified and presented three core (key) dimensions of TPA, which are: Data Context (DC), Data Collection Method (DCM), and Predictive Analytical Method (PAM).

7.3.1 Data Context

The DC dimension comprises elements that include the traffic implementation area, prediction horizon, data source(s), prediction type, and training dataset size $(n_{columns} \times m_{rows})$. The traffic implementation area distinguishes the TPA implementation scope or area. For instance, urban (intra-city) networks, which differ from inter-city networks, such as, freeway/highway networks. Many short-term TPA studies have been implemented on freeways/highways and motorways, for reasons that have been earlier discussed. This study, however, focused on the provision of guidance to traffic data scientists in performing urban TPA. The reason for focusing on urban traffic is because this has a direct impact on day-to-day life and constitutes most of the traffic congestion. The prediction horizon represents the time interval or frequency for which the forecasts are made. In certain scenarios, shorter prediction horizons may not suffice. For instance, for road development/planning purposes, daily, annual, or longer prediction horizons may be required, as opposed to short-term (i.e. 5-min to 1-hr) typical prediction horizons that are applicable in day-to-day traffic management schemes. Therefore, the process of defining the prediction horizon in a TPA task is critical to accurate data modelling, given that it impacts on the predictive capability of a predictive model. The size of the available dataset to be used for data-driven traffic parameter prediction is a critical component that can impact on predictive model accuracy, training time, and computational demand, as has been pointed out in Section 5.6.1. Furthermore, the data source(s) used for the TPA process can determine the nature of the predictive problem, for instance, if it is a univariate or multivariate prediction problem.

7.3.2 Data Collection Method (DCM) Dimension

Another dimension that was identified as impacting TPA is the traffic data collection method. Given the rapid advancement in sensor technology and electronics, the number of traffic data collection methods in use today has tremendously increased. The application of the various data collection methods can impact the outcome of a TPA task, relating to the sensor reading capability of the various collection methods. This variation can be attributed to the difference in the level of granularity, aggregation, detail and accuracy obtainable within each individual sensor/device. In this study, the DCM dimension was identified as having an impact on the PAM selection, thereby prompting the inclusion in the support tool development process.

7.3.3 Predictive Analytical Method (PAM)

The PAM dimension contains the collection of predictive algorithms that are available to a traffic data scientist in each TPA solution development process. This study, however, restricted this portfolio of models to include seven (7) predictive algorithms (see Section 5.6.3). The choice of a suitable predictive model can impact greatly on the quality of the TPA solution to a prediction problem. Given that there is no best predictive algorithm that performs optimally in all situations, it is useful to provide guidance to traffic data scientists in the choice of the adequate PAM, which can be applied to a traffic prediction scenario. The support tool presented in Chapter 6 provided suggestions about the seven PAMs used in this study. The hypothesis of this study that given an articulated list of input parameters, a suggestion of alternatives of PAMs can be made, was verified and evaluated in Chapter 6 using the stipulated case scenarios (see Section 6.6).

7.4 Guidance in TPA

Section 5.4 related guidance to TPA and provided a discussion about the goals of guidance in TPA. The main goal of an analytics guidance system is the provision of knowledge about a dataset that can enable the user to answer questions about the dataset and analytical process (Collins et al., 2018). In clear terms, the goals of guidance in TPA were presented to include the accurate and timely dissemination of information to the user, which is the traffic data scientist in this case. A conceptual model of TPA guidance is presented in Figure 2-1 in Section 2.5. The framework, which is an enhanced and TPA-specific and enhanced version of the visual analytics framework presented in Ceneda et al. (2017), characterises guidance using three main characteristics. These include the knowledge gap, the input/output, and the degree of guidance. The knowledge gap in a TPA task has been identified within this study as the identification of the path towards the development of a suitable analytical solution to a TPA task.

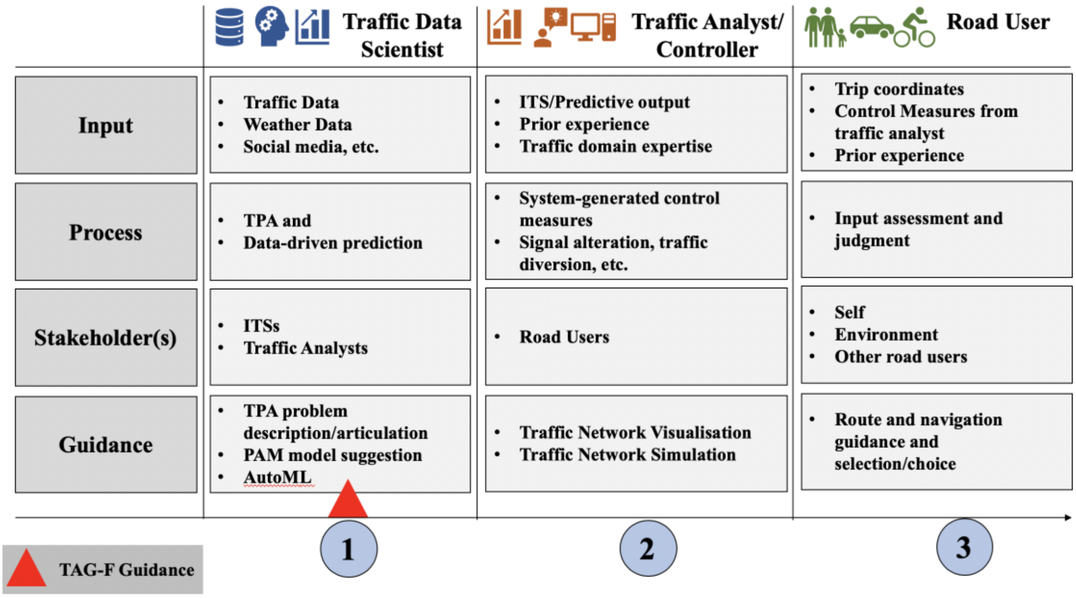
Therefore, the aim of providing TPA guidance in this study is the identification of a pathway with which the traffic data scientist can traverse from a broad and complex problem space to a narrow and well-defined analytical solution space (see Figure 2-1). The 'input/output' dimension describes how the guidance is to be generated and presented to the user. In TPA, this can be achieved via a number of approaches such as expert-system, meta-knowledge, or machine learning/AI methods. The output specifies how the guidance will be offered/presented to the user. Two possible output media are identified in Ceneda et al., (2017), which are a *means* and an *answer*. The *means* is the provision of an impulse that triggers further exploratory options. In the context of TPA, this is the provision of a means of achieving the desired goal, which is data-driven traffic prediction.

Finally, the degree of guidance represents how much guidance should be provided to the user. In TPA, guidance can be provided at all levels – orienting, directing, or prescribing, as described in Section 2.5, corresponding to minimum, medium, and maximum guidance levels, respectively. The degree of guidance to be provided to the user is a continuum that corresponds to the needs of the user. This study was streamlined to provide directional (medium level) guidance to traffic data scientists via the development of a TPA solution using a structured framework that enables apt TPA problem description, and (via the support tool) predictive model suggestion. Although this study focuses on directional guidance, it must, however, be mentioned here that it is possible to modify the framework to provide orienting or prescriptive guidance levels.

7.5 Characterising the TPA stakeholders

As stated in Section 1.1, there is not a clear distinction and description – in the existing literature – of the main stakeholder(s) in a TPA process. The two terms – traffic analyst and traffic data scientist – have been clearly defined in this thesis (Section 1.1). Given the varying nature of their individual inputs, processes, outputs, and stakeholders, each of these TPA 'actors' require varying levels of guidance.

Figure 7-1 presents a graphical summary of the characterisation of the main actors involved in TPA. In the figure, the x-axis represents the three main TPA actors – the traffic data scientist, the traffic analyst, and the road user. The y-axis, on the other hand, categorises and elucidates the characteristics of each actor. As can be seen, the characteristics, which include, input, process, stakeholder(s), and guidance requirement vary across the actors, implying that a single, one-size-fits-it-all guidance structure/mechanism will be complex and intricate. It is obvious that both the data scientist and analyst can also be road users (at least at one point or the other), thereby indirectly helping themselves (i.e. traffic data scientists and analysts) in the TPA and congestion control processes.



	Traffic Data Scientist	Traffic Analyst/ Controller	Road User
Input	Traffic Data Weather Data Social media, etc.	ITS/Predictive output Prior experience Traffic domain expertise	Trip coordinates Control Measures from traffic analyst Prior experience
Process	TPA and Data-driven prediction	System-generated control measures Signal alteration, traffic diversion, etc.	Input assessment and judgment
Stakeholder(s)	ITSs Traffic Analysts	Road Users	Self Environment Other road users
Guidance	TPA problem description/articulation PAM model suggestion AutoML	Traffic Network Visualisation Traffic Network Simulation	Route and navigation guidance and selection/choice
	▲ 1	2	3

▲ TAG-F Guidance

Figure 7-1: Characterisation of Road Traffic Main Actors

This current study aimed at providing guidance to address the first category of TPA actors – the traffic data scientists (depicted in a red triangle). As the graphic shows, the main inputs for the traffic data scientist job function is the input datasets and the suite of existing and custom predictive algorithms. The job function of the data scientist involves the development of novel (or application of existing) predictive algorithms, applied on the available datasets for the purpose of performing data-driven traffic parameter prediction. The output of the process mainly benefits the second actor (number 2), which is the traffic analyst. This means that the predictions from the traffic data scientist enable the traffic analyst to make informed decisions about what control interventions need to be applied to mitigate, control, or manage traffic congestion. The guidance needs for the traffic data scientist have been discussed and articulated in the previous chapters of this thesis. They include but are not limited to, predictive model choice, TPA problem articulation and description, and AutoML mechanisms for TPA processes. This study has addressed the guidance needs of the traffic data scientist by providing a framework (TAG-F framework – Section 5.2), which is a reusable framework that can enable TPA problem characterisation and PAM decision support via the TAG-F tool (see Section 6.4), which is trained using data obtained from a LBD process, and applied on an IBL learning algorithm (see Section 6.5.2).

The second actor in the TPA process is the traffic analyst. The analyst serves as the middleman between the abstract (i.e. data-related) component of TPA and the road users. As Figure 7-1 shows, the guidance requirement of the traffic analyst is a mechanism with which the traffic network (past, present, and/or future) can be visualised, simulated, and optimised. Therefore, a simulation-based traffic visualisation system would prove beneficial to traffic analysts, and is, for this reason, being implemented in many TCCs across the world. However, in order to introduce a 'proactive' component to the traffic analyst traffic management cycle, the simulation, visualisation, and optimisation of the future (i.e. predicted) traffic network are essential. Scenario 3 (see Section 6.6.6) presented a potential method of addressing this guidance need. This involved integrating the output from the TAG-F framework to a two-stage prediction-simulation model (PRESIMM) for traffic management. The model simulates the predicted (future) traffic network state using SUMO, thereby enabling the traffic analysts to visualise the impact of their chosen traffic intervention(s). In this way, the guidance provided by TAG-F can be extended to cater for the traffic analysts. This integration has been listed in the future work section of this thesis.

Finally, the third actor in the TPA process is the road user. The guidance requirement of the road user involves route guidance and navigation guidance. Typical ATISs like Google Maps, Waze, Garmin, etc. provide this level of guidance to the road users. In summary, the entire traffic TPA ‘ecosystem’ can be summarised using the analogy: the *road users* benefit from TPA via reduced traffic congestion, which is a by-product of the interventions of *traffic analyst/control personnel*. However, in order for the traffic analyst to make informed decisions, the *traffic data scientist* needs to make forecasts/predictions, which are achieved via the application of TPA methods on historical and real-time traffic (and non-traffic) datasets. This research study focused on the provision of guidance to the traffic data scientist but can be enhanced to accommodate the traffic analysts and – possibly – the road users (this will be discussed in the future work section 8.7). Therefore, the output of this research study (TAG-F framework and support tool) is a structured guidance mechanism for traffic data scientists performing TPA.

7.6 Chapter Summary

In this chapter, a discussion about the findings of this research study is presented. A recap of the dimensions of the TAG-F framework is presented in Section 7.3, which are the data context, data collection method, and predictive analytical method. In Section 7.4, a discussion about guidance in TPA is presented, while the characterisation of the TPA stakeholders is presented in Section 7.5. There is a clear explanation and distinction between the three TPA stakeholders – traffic analysts, data scientists, and road users. It was stated in Section 7.5 that the traffic analyst is involved with the overall traffic network or system planning process, directly addressing the needs of the road users (i.e. congestion-free road networks), while the (traffic) data scientist is responsible for ensuring accurate and timely predictions are provided to the traffic analyst, ITSs or other data/expert system to assist in the traffic planning and congestion control purposes.

The next chapter will conclude the study, articulate the research contributions – theoretical, methodological, and practical – as well as discuss the implications of this research study to the academic and industrial sectors. A reflection and synthesis of the research process is also presented, including the articulation of the research objectives, questions, and respective contributions to answering the research questions. In addition, the research limitations and future work will also be presented.

Chapter 8 Conclusions and Future Work

8.1 Recap of the research problem

There is an increase in the use of ITSs for traffic management and control, brought about by the rising need for easy and efficient mobility and road transportation. ITSs provide traffic status information to road users, mainly by performing *predictive analytics* on historical (and/or real-time) traffic data collected via one or more traffic data collection sensors. Predictive analytics refers to the use of statistical, data-mining, and machine learning techniques to find patterns in data, which are subsequently used to make forecasts about future events (Shmueli and Koppius, 2011). Although TPA has been beneficial to traffic management and control, achieving accurate data-driven traffic parameter prediction is difficult and convoluted. This can be attributed to three major reasons, which were summarised in Section 1.1.

First, urban traffic control and management can be likened to a *wicked problem* (Churchman, 1967) – one where both the solution and the means of achieving it are unknown, ambiguous, or uncertain. The interactions between road users, infrastructure, and exogenous factors such as calendar or time of day, rainfall, temperature, events, road works, and accidents further account for the dynamic and stochastic nature of traffic flow. For this reason, TPA is much more intricate. Achieving accurate and effective TPA is fraught with challenges, such as the precise definition of the TPA problem space. In order to develop an effective TPA solution, a number of questions need to be answered. For instance, what are the key factors that influence TPA? How can a traffic data scientist develop an appropriate analytical method for performing TPA?

Secondly, there are many traffic predictive algorithms/models in use today due to increased research interest in the field of TPA. The suite of existing/available predictive models should serve as a broad portfolio of ‘tools’ that can be employed towards performing TPA. However, it is difficult for a traffic data scientist to possess knowledge about each and every one of the existing predictive algorithms. Studies critically analysing traffic predictive methods/algorithms, their benefits and demerits, or in what particular traffic prediction problem scenario(s) a particular predictive method is (are) most appropriate abound (Ermagun and Levinson, 2018; Lana et al., 2018; Vlahogianni et al., 2014, 2004), but have quickly become superseded due to the rapidly evolving research area of traffic prediction.

Thirdly, as of yet, there is no single best predictive algorithm that can be used in all situations, as has been experimentally and theoretically deduced (Goodfellow and Bengio, 2015). Therefore, it is the case that a predictive model that performs well in a given condition or scenario may likely perform poorly in other scenarios. Traffic prediction takes different forms based on the problem definition, which makes it challenging to decide upon which predictive methodological approach should be adopted towards solving the traffic prediction problem. Road TPA requires different strategies for various scenarios or conditions, and a particular predictive model may only be suitable for only one – or more, but not every – TPA problem or scenario. For instance, the predictive analytics strategy to be applied when performing TPA for regular/recurring peak traffic prediction may not be suitable for use in an accident or severe congestion resulting in lane closure, as shown in the scenarios in the previous chapter. The provision of a ranked list of predictive algorithms to a traffic data scientist is a useful guidance and decision support mechanism for traffic data scientists in performing TPA.

The above-listed issues highlight the need for the provision of guidance to traffic data scientists in performing TPA. It was hypothesised in this thesis that the provision of this guidance will result in improvements in the overall quality of the TPA solution.

8.2 Summary of research findings

This research study was founded on a set of research questions, which were initially presented in Chapter 1, further elaborated in Section 4.4.1 and validated/answered within the course of the study. The primary research question prompted three (3) research sub-questions, which have been answered in this thesis. It must be mentioned here that although only three sub-questions were prompted, it is in no way a claim by the researcher that this represents a complete set of questions relating to the subject of guidance in TPA. Rather, the set of research questions proposed in this study were sufficient in providing the requisite answer to the primary research question. In this section, a summary of the findings within the study with respect to the individual questions is discussed.

8.2.1 Primary Research Question (PRQ)

Can a predictive analytics guidance framework be designed to facilitate traffic data scientists in exploring the analytical decision space of TPA tasks?

The primary research question sought to investigate the viability of providing guidance to traffic data scientists in the quest for traffic analytical solution and to validate if it improves the overall quality of the traffic analytics process. The results from the evaluation section support for the argument and validates the fact that the provision of a structured decision-making framework can improve the overall quality of the traffic analytics process.

The primary research question formulated a number of sub-research questions, which contribute to the investigation and achievement of the primary objective of the study, which is the development of a traffic data analytics guidance framework. These are presented in the subsequent sub-sections. The evaluation section (see Section 6.6) presented a quantitative evaluation of the framework using a support tool. Three real-world scenarios were presented, and the guidance obtained from the application of the framework and tool was discussed. The findings obtained supported the hypothesis that a TPA guidance framework can be used to aid data scientists in the process of performing TPA. By articulating the problem space and adequately providing a description of the problem space using TPA factors (identified by the framework), the traffic data scientist can develop a TPA solution quickly. In addition, the suggestion of the PAM to be used in each scenario constitutes additional decision support to the traffic data scientists in performing the TPA.

8.2.2 Research Sub-Question 1

What are the key (critical) dimensions of data-driven traffic prediction problems?

The objective of the first sub-research question was the apt identification of the dimensions that describe data-driven traffic prediction. This was argued to be critical to the development of a guidance structure capable of providing guidance to traffic data scientists in traffic prediction. After a detailed and systematic literature review process, and a set of scenario examples, we were able to identify three dimensions that comprise of data-driven traffic prediction. The evaluation of the framework and support tool quantitatively show the interaction between the dimensions and dimension elements within the traffic data analytics solution space. To reiterate, from Section 5.2, the identified dimensions relate to (i) the method of traffic data collection,

(ii) the data/analytical scope, and (iii) the predictive analytical method or predictive algorithm. These three dimensions altogether contribute to form the core underlying foundation of every data-driven traffic parameter prediction.

8.2.3 Research Sub-Question 2

What are the analytical decision points within each dimension that can support traffic data scientists in TPA?

The main objective of the second sub-research question was the investigation into the availability (if any) of analytical decision points within the traffic analytics solution development process, which will provide directional guidance towards the structured problem definition, as well as a pathway towards solving the prediction problem. The answer to this question was realised by a systematic and rigorous literature review in Chapter 2. The identified analytical decision points within the framework dimensions, which are discussed in detail in Section 5.2.1, are *prediction horizon*, *dataset size*, *traffic scope* (urban or non-urban), *real-time prediction* (or not), and *training data composition* (univariate or multivariate data sources). The knowledge and insight obtained from the academic articles contributed to the articulation of critical analytical decision points that ultimately formed the composition of the proposed framework.

8.2.4 Research Sub-Question 3

What are the analytical decision parameters within each key traffic analytical problem dimension required to explore the decision space of TPA tasks?

Research sub-question 3 aimed at investigating the effect of the set of critical analytical points (described in Research Sub-Question 2) towards the inference of the predictive model. The review of existing relevant studies revealed that each of the data-driven prediction models has its own individual strengths and weaknesses, assumptions, and generalisations, that indicates in which scenario it is likely to perform better. In order to answer this question, the development of the TAG-F support tool was required. The support tool was used for the quantitative evaluation of the framework, which provided the needed answers to this sub-question. The results presented in Section 6.6, where the support tool and framework were demonstrated, showed that there is a possibility of providing guidance in terms of appropriate prediction models based on a set of articulated analytical decision points.

8.3 Research Contributions

According to Hevner et al. (2004), the key output from a design science research is a contribution to knowledge, which either improves the understanding of the problem or provides a design/solution to improve the artefact design. This section presents the main contributions of this research to the TPA body of knowledge.

The main contribution of this research study is a guidance framework for traffic predictive analytics, which contributes to furthering knowledge in the field of TPA (see Section 5.2). This includes the design, development, and introduction of a novel traffic predictive analytics guidance framework, which enables a structured definition of the traffic data analytical solution space. The framework was developed to address the gaps identified following an extensive and systematic literature review in the area of traffic predictive analytics (see Section 6.3). TAG-F serves as a framework that can provide traffic TPA guidance by providing structured decision points throughout the traffic analytics development process.

Secondly, this research study proposed a meta-learning model for predictive model suggestion in TPA. The model, which is presented in Section 6.5, is developed using a meta-dataset obtained using a literature-based discovery process described in Section 6.4. The meta-learning model has the objective of providing predictive model suggestions based on a set of meta-level attributes obtained using the TAG-F framework dimensions and dimension parameters. The proposed model differs from existing AutoML frameworks such as Auto-WEKA and H2O by adopting a meta-learning approach in contrast to the other frameworks that actually perform model training on the input data. The model can be extended, updated, and reused for meta-modelling and traffic prediction model selection. This model can be used to provide directional guidance to traffic data scientists by suggesting ranked alternative predictive model(s) in the TPA solution development process.

Thirdly, this framework and support tool represented a practical demonstration of the DSR framework, whereby an artefact was developed in order to solve a real-world organisational problem. In this case, the contribution of the framework and support tool is directly beneficial to the traffic data scientists but can be extended to provide guidance to traffic analysts and possibly road users.

A final contribution is a clear distinction and characterisation of the TPA stakeholders, including their respective characteristics, role(s) or functions, input and outputs, as well as respective guidance needs. The clear identification and distinction of these stakeholders can improve understanding of the TPA process and the requisite solution development process. In this thesis, this categorisation is presented in Section 7.5 (see Figure 7.1).

8.4 Implications for Research and Practice

The study has implications for practice and research by improving the overall quality of the TPA process. The TAG-F framework and support tool was theorised for the purpose of providing guidance to traffic data scientists in the traffic analytics problem formulation and execution. The identified dimensions within the framework contribute to the delineation of the traffic predictive analytics process, which can serve as a foundation stone for future research into improving traffic data analytics.

In terms of business/organisational implications, the literature-driven knowledge base, framework and meta-learning support tool can be commercialised, adapted, customised, extended, and reused in industrial traffic predictive analytics within the domain of practice. Traffic data scientists will especially find the support tool useful in the traffic analytics process, given the plethora of existing traffic predictive models available and in use today. Finally, although the framework is developed and tested within the traffic setting in Greater Manchester, United Kingdom, there is no barrier to using the same set of findings, methodology, framework, and knowledge base in other geographical regions.

8.5 Reflection on the research process

In the preceding seven chapters, this thesis presented a research study on the provision of guidance for traffic predictive analytics. The main aim of the research was the development of a traffic predictive analytics guidance framework, which can provide directional guidance to traffic data scientists for the execution of traffic prediction. To answer the research questions posed in Section 1.2, three research objectives were explored, as presented in Section 1.3. In this section, a reflection or synthesis of the research process is presented, which is tabulated in Table 8-1.

After a rigorous literature review in Chapter 2, a number of research opportunities were identified, which the study aimed to address. Firstly, it was identified that there is a lack of shared knowledge about existing traffic prediction algorithms. This was attributed to the rapidly-evolving field of short-term traffic prediction, leading to increased development of predictive algorithms. Secondly, it was also identified that there is a difficulty in selecting an appropriate and optimal predictive model for urban traffic prediction, due to the multitude of predictive algorithms, and the major focus of the application of these traffic prediction models on freeway/motorway traffic area/scope. Thirdly, in line with the NFL theorems, there is no single best algorithm that performs best in all situations.

As part of conducting this research, research questions and objectives were formulated, which are summarised in Table 8-1. In order to answer the questions, the research study adopted the design science research methodology (Hevner et al., 2004), undertaking the study according to the design science research stages as presented in (Peffers et al., 2007), which are: problem awareness, objective definition, design and development of IT artefact, demonstration, and evaluation (see Figure 4-2 in Section 4.3). The proposed solution is a guidance framework for TPA guidance, which has been presented in Chapter 5. The framework, identified as TAG-F, delineates traffic prediction into three dimensions – Data Context (DC), Data Collection Method (DCM), and Predictive Analytical Method (PAM). The three dimensions, alongside the respective dimension elements, provide a structured guidance mechanism that enables traffic data scientists to develop appropriate analytical solutions to complex traffic prediction problems.

The framework is complemented by a tool, the TAG-F Support Tool (see Section 6.4), which provides predictive model selection guidance towards the traffic predictive analytics development process. Within this study, a quantitative evaluation of the practicality of the proposed framework via the support tool has been carried out using three case scenarios, which is presented in Sections 6.6.4, 6.6.5, and 6.6.6 respectively.

The research limitations, weaknesses, and future work opportunities are highlighted in this chapter.

Table 8-1: Summary of research questions, objectives and contributions

	Research Question	Research Objective	Research Contribution	Evaluation
PRQ	Can a predictive analytics guidance framework be designed to facilitate traffic data scientists in exploring the analytical decision space of TPA tasks?	**RO1:** The first objective seeks to investigate the key dimensions that describe a typical TPA problem space. The prior identification of these dimensions becomes an essential requirement for the development of a structure that is capable of providing analytical guidance.	**RC1:** the development of the TAG-F framework that delineates TPA into three key dimensions **(Chapter 5).**	Using the support tool presented in Section 6.4, the TAG-F framework was evaluated in Section 6.6 using three real-world case scenarios.
RQ-1	What are the analytical decision points within each dimension that can support traffic data scientists in TPA?	**RO2:** Within the identified (key) dimensions, it is essential to investigate and identify the analytical decision parameters, which contribute to holistically describing the analytical dimension space. This will aid the traffic data scientists to arrive at quicker and more logical conclusions in the analytical development process.	**RC2:** The identified TPA dimensions and dimension parameters (see Sections 5.2.1, 5.2.2, and 5.2.3) **RC3:** A characterisation of the key TPA stakeholders, which was not present in the existing literature (Section 7.5)	A meta-learning model for TPA predictive model suggestion using TAG-F (Section 6.5.1). The provision of directional guidance to traffic data scientists (Section 6.6)
RQ-2	What are the analytical decision points within each key dimension, which can support traffic data scientists in performing TPA tasks?			
RQ-3	Given a set of analytical decision points, can a set of alternative prediction modelling techniques/algorithms be made?	**RO3:** To develop a method whereby, given a set of identified analytical decision points and dimension, it can be possible to determine which predictive model/algorithm is appropriate for the given TPA task/scenario.	**RC3:** a prototype support tool that complements the TAG-F framework **(Chapter 6).**	Quantitative evaluation of the TAG-F support tool in Chapter 6.

Research Objective 1: The first objective of this research study was to investigate the key dimensions that describe a typical TPA problem space. This objective was set out to provide an answer to the primary research question (see 4.4.1). The prior identification of these dimensions becomes an essential requirement for the development of a structure that is capable of providing analytical guidance. The literature was reviewed systematically in order to identify the factors that affect TPA, using a literature review process presented in Chapter 2. The identification of the dimensions that characterise the TPA solution development process include predictive model, input data sources, traffic scope/area of implementation, data quality, and spatiotemporal considerations (see Section 2.7.1). The contribution from the TPA characterisation enabled the development of the TAG-F framework, which is presented in Chapter 5. In order to evaluate the proposed framework, a support tool was developed, which will enable the practical evaluation of the TPA guidance framework. The TAG-F support tool was presented in Section 6.4, with the practical evaluation of the framework and tool presented in Section 6.6.

Research Objective 2: The second objective of this research study was to investigate and identify the analytical decision parameters within each TPA dimension identified (i.e. the output of research objective 1), which contribute to holistically describing the TPA dimension space. In addressing this objective, Chapter 2 and 3 presented characterisation and traffic prediction background, respectively. It was identified that the data context dimension of the proposed framework contained dimension parameters, including the prediction horizon, dataset size, traffic scope, level of analysis, and data source type(s). The identification of the analytical decision parameters within the TPA dimensions will provide guidance in the TPA problem description process, thereby aiding traffic data scientists to arrive at quicker and more logical conclusions in the analytical development process. In order to evaluate the output of this research objective, a meta-learning model for TPA predictive model suggestion is presented in Section 6.5.1.

Research Objective 3: The third objective of this research study was to develop a method whereby, given a set of identified analytical decision points and dimension, the determination of which predictive model/algorithm may be appropriate. The output of this research objective is the artefact – the TAG-F support tool, presented in Section 6.4. The model adopted an instance-based learning approach, trained on meta-level dataset obtained via a literature-based discovery process, presented in Section 6.4. Similarly, a rule-based learning algorithm was also applied on the meta-dataset in order to provide a set of rules that can be used to provide justification or rationale for the suggested likely PAM (see Section 6.5.3 for details about the rule-based algorithm). The evaluation of this objective was articulated in the use case scenarios presented in Sections 6.6.4, 6.6.5, and 6.6.6, respectively.

8.6 Research Limitations

This research presented an attempt at providing guidance to traffic data scientists in the development of traffic analytical solutions to traffic prediction problems. A traffic analytics guidance framework and support tool have been presented, discussed, and evaluated quantitatively and qualitatively. However, in a study of this magnitude, limitations and opportunities for improvements are inevitable. These limitations constitute the basis of the recommendations for further studies presented in Section 8.7. The limitations are summarized here.

The main limitation of the proposed guidance methodology and support tool is the non-existence of a self-learning or adaptive functionality. Due to the rapidly evolving research field of TPA, and the plethora of existing predictive algorithms, a ‘static’ guidance knowledge base (and tool) will easily be outdated. In this present study, the writer of this thesis performed manual searching and update of the knowledge base, using the process as described in Section 6.3. However, in order to include a feedback or adaptive module, further research work needs to be performed. For instance, automatic literature searches and meta-knowledge extraction is in another field of research (i.e. Natural Language Processing – NLP), which is outside the scope of this study. This has, however, been included for consideration in the future research section of this thesis (see Section 8.7).

Furthermore, another limitation is the absence of an ‘auto-predict’ or the provision of prescriptive guidance by the support tool. The need for guidance in traffic analytics was identified, and although the framework has enabled a structured problem definition, as well as predictive model selection via the support tool, an important aspect of guidance will be provision of a mechanism that not only suggests the appropriate prediction model to adopt, but take the input dataset(s), perform the necessary pre-processing, hyperparameter optimisation, and execute the predictions based on the suggested algorithms.

Thirdly, constraining the support tool to seven (7) candidate models is relatively small, considering the large number of existing predictive algorithms that are available within the spectrum of data-driven traffic prediction. This can be mitigated by extending the support tool further to accommodate more predictive models, which would build on the foundation of the proposed framework and support tool presented herein. However, this extension has been considered for future work and discussed in Section 8.7.

Finally, the support tool evaluated the individual model performances using predictive model accuracy alone (i.e. error metrics – MAE and MSE). However, there are other factors that can be considered when evaluating the performance of prediction models, for instance, processing time, model complexity, traceability, and computational demand. The area of prediction traceability and interpretable artificial intelligence is a research area that is currently receiving a lot of attention. This has presented an opportunity for improvement, where the user can specify the level of model (or prediction) interpretability that is required, or what performance yardstick/metric to be used for the performance evaluation of the predictive models that have been suggested by the framework. For instance, in real-time prediction scenarios, it is critical to consider model training and inference time in the choice of a predictive model to adopt.

8.7 Future Work

This section presents a number of refinements and modifications that can be applied to the proposed traffic analytics guidance approach that will strengthen the claims made within the study. This research introduced the first predictive analytics guidance framework for traffic prediction in the literature. As such, there are some limitations, which were identified in Section 8.6 above. Further work could facilitate improvements.

8.7.1 An adaptive feedback loop for the support tool

As identified in the research limitations, a critical function of the support tool is a self-learning loop. This is emphasised due to the plethora of predictive algorithms, a list that continually and rapidly grows in line with advances in mathematical, technology and computing research. In order to incorporate this functionality, some form of natural language processing (NLP) needs to be integrated into the entire LBD process. In this way, the systematic review, meta-knowledge extraction, and model re-training can be performed in real-time and as new models are developed. However, a criterion for model or PAM inclusion into the knowledge base will be based on the evidence (from the literature) that the 'new' model is better than the current models in the knowledge base. For instance, in image processing, the state-of-the-art deep learning model was the Convolutional Neural Network (CNN) by Yann Lecun and Bengio (LeCun and Bengio, 1995). However, in 2017, Geoffrey Hinton's capsule network was experimentally and theoretically proven to significantly outperform the CNN in image recognition and classification (Sabour et al., 2017). If this was proven to be the case, therefore, the knowledge-base can be (automatically) updated with meta-knowledge about capsule networks instead of CNNs.

8.7.2 Extending the guidance offered to traffic analysts

Figure 7-1 in Section 7.6 presented a characterisation of the main stakeholders/actors in a TPA process, as well as their input(s), processes, and guidance requirements. The graphic there depicts three main actors – (1) traffic data scientists, (2) analysts, and (3) road users, respectively. The research presented in this thesis was restricted to the provision of directional guidance to cater for traffic data scientists. Scenario 3 in Section 6.3.5 presented an illustration of how the extended guidance can be realised. It represents the integration of the TAG-F framework and support tool with a proactive traffic management model – PRESIMM (Essien et al., 2019b) – which is a two-stage model that can enable proactive visualisation of future (and past) traffic network status. In this way, the traffic analysts can benefit from the guidance offered in problem articulation (using TAG-F framework), PAM model choice (using the support tool), and the visualisation and simulation of the predicted traffic situation, and re-simulation of the interventions (or control measures) applied by the traffic analysts towards effective traffic congestion management and control. The integration with PRESIMM can be automated by the development of an ‘auto-predict’ function to eliminate the human-in-the-loop mechanism, where the TAG-F framework and tool is used to assist the traffic data scientist in making the requisite parameter predictions using the PAM suggestions from the support tool. In this way, the predictions are then passed as input to the PRESIMM visualisation module, where the traffic analyst can see the consequence(s) of his/her chosen control intervention in a simulated or virtual network. If this integration is done, the outcome can indirectly benefit the road users since the traffic analysts will make informed decisions about the suitable choice of action to apply/deploy on the traffic network.

8.7.3 Extending the number of candidate models

The framework and support tool can be extended to accommodate more predictive models than the seven presented in this study. This will strengthen the provision of guidance to analysts in the traffic prediction execution. As stated, the framework attempted to include a predictive model from time series analysis, instance-based learning, machine learning, deep learning, statistical, recursive (KF) and kernel methods. However, there are many more predictive models from those listed above, modifications which become new models (for instance, Seasonal ARIMA or SARIMA), and altogether new prediction model types which can be incorporated into the framework for improved overall performance. Incorporating additional PAMs into the support tool will be beneficial, especially if they are of a different class/family of PAMs. For instance, some scenarios may require clustering algorithms, explainable predictions, and interactive guidance. The inclusion of these models for consideration in the support tool will make the support tool more robust, and the guidance provided will be more useful to the data scientists. However, the data scientist should always be provided with a limited set of ranked model alternatives because a very large set of PAM suggestions to the data scientist will be confusing and the essence of the guidance offering will be defeated.

Despite the need for future improvements on the proposed TAG-F framework and support tool, which have been highlighted in this chapter, the framework and tool can in this initial state be beneficial to traffic organisations and/or businesses seeking to improve the overall traffic analytics process. This can be realised in a number of ways. Firstly, the proposed framework and tool can be used as a template for enhanced guidance in executing traffic data analytics. Secondly, the knowledge base can facilitate learning, sharing of domain knowledge about traffic analytics throughout the traffic analytics solution development process. Thirdly, they provide guidance in traffic analytics and aid in the semi-automated and knowledge-driven (via a literature-based process) reuse of domain knowledge to improve traffic predictive analytics.

References

Abbas, M., Rajasekhar, L., Gharat, A., Dunning, J.P., 2013. Microscopic modeling of control delay at signalized intersections based on Bluetooth data. J. Intell. Transp. Syst. 17, 110–122.

Abdelghany, A., Abdelghany, K., Mahmassani, H., Murray, P., 2000. Dynamic Traffic Assignment in Design and Evaluation of High-Occupancy Toll Lanes. J. Transp. Res. Rec. 1733, 39–48. https://doi.org/10.3141/1733-06

Abdulhai, B., Porwal, H., Recker, W., 2002. Short-term traffic flow prediction using neuro-genetic algorithms. ITS Journal-Intelligent Transp. Syst. J. 7, 3–41.

Abid, N.M., Hussain, S.S., 2017. Transportation network planning using simulation: Case study\ Al Mansour city, in: 2017 2nd IEEE International Conference on Intelligent Transportation Engineering, ICITE 2017. IEEE Xplore, pp. 272–279. https://doi.org/10.1109/ICITE.2017.8056923

Abidin, A.F., Kolberg, M., Hussain, A., 2015. Integrating Twitter Traffic Information with Kalman Filter Models for Public Transportation Vehicle Arrival Time Prediction, in: Big-Data Analytics and Cloud Computing. Springer International Publishing, Cham, pp. 67–82. https://doi.org/10.1007/978-3-319-25313-8_5

Agarwal, M., Maze, T., R Souleyrette, 2005. Impacts of weather on urban freeway traffic flow characteristics and facility capacity, in: Proceedings of the 2005 Mid-Continent Transportation Research Symposium. pp. 18–19.

Ajiboye, A.R., Abdullah-Arshah, R., Hongwu, Q., 2015. Evaluating the effect of dataset size on predictive model using supervised learning technique. Int. J. Softw. Eng. Comput. Sci. 1, 75–84. https://doi.org/http://dx.doi.org/10.15282/ijsecs.1.2015.6.0006

Alajali, W., Wen, S., W Zhou, 2017. On-street car parking prediction in smart city: A multi-source data analysis in sensor-cloud environment, in: International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage. Springer, Cham., pp. 641–652.

Algers, S., Bernauer, E., Boero, M., Breheret, L., Di Taranto, C., Dougherty, M., Fox, K., Gabard, J.-F., 1997. Review of micro-simulation models. Rev. Rep. SMARTEST Proj.

Asencio-Cortés, G., Florido, E., Troncoso, A., Martínez-Álvarez, F., 2016. A novel methodology to predict urban traffic congestion with ensemble learning. Soft Comput. 20, 4205–4216. https://doi.org/10.1007/s00500-016-2288-6

Atkeson, C.G., Moore, A.W., Schaal, S., 1997. Locally Weighted Learning, in: Lazy Learning. Springer, Dordecht, pp. 75–113. https://doi.org/10.1023/A:1006559212014

Bacchiani, G., Molinari, D., Patander, M., 2019. Microscopic Traffic Simulation by Cooperative Multi-agent Deep Reinforcement Learning, in: Proceedings of the

18th International Conference on Autonomous Agents and MultiAgent Systems. International Foundation for Autonomous Agents and Multiagent Systems, pp. 1547–1555.

Banger, K., Adriano, N., 2015. Performance Evaluation of Non-Intrusive Methods for Traffic Data Collection, in: TAC 2015: Geting You There Safely - 2015 Conference and Exhibition of the Transportation Association of Canada//ATC.

Banks, J., 2002. Introduction to transportation engineering, 21st ed. McGraw-Hill, New York.

Barros, J., Araujo, M., Rossetti, R.J.F., 2015. Short-term real-time traffic prediction methods: A survey, in: 2015 International Conference on Models and Technologies for Intelligent Transportation Systems, MT-ITS. pp. 132–139. https://doi.org/10.1109/MTITS.2015.7223248

Barry-Straume, J., Tschannen, A., Engels, D.W., Fine, E., 2018. An Evaluation of Training Size Impact on Validation Accuracy for Optimized Convolutional Neural Networks. SMU Data Sci. Rev. 1, 12.

Barth, M., Boriboonsomsin, K., 2008. Real-world carbon dioxide impacts of traffic congestion. Transp. Res. Rec. 2058, 163–171.

Baskerville, R., Baiyere, A., Gregor, S., Hevner, A., Rossi, M., 2018. Design science research contributions: finding a balance between artifact and theory. J. Assoc. Inf. Syst. 19, 358–376.

Behrisch, M., Bieker, L., Erdmann, J., Krajzewicz, D., 2011. SUMO–simulation of urban mobility: an overview, in: Proceedings of SIMUL, The Third International Conference on Advances in System Simulation. ThinkMind.

Bekkar, M., Djemaa, H.K., Alitouche, T.A., 2013. Evaluation measures for models assessment over imbalanced data sets. J. Inf. Eng. Appl. 3.

Bellman, R., 1954. The Theory of Dynamic Programming. Bull. Am. Math. Soc. 60, 503–515. https://doi.org/10.1090/S0002-9904-1954-09848-8

Ben-Akiva, M., Bierlaire, M., Koutsopoulos, H., Mishalani, R., 1998. DynaMIT: a simulation-based system for traffic prediction, in: DACCORD Short Term Forecasting Workshop. Delft The Netherlands, pp. 1–12. https://doi.org/10.1002/elan.200603552

Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: A review and new perspectives. IEEE Trans. Pattern Anal. Mach. Learn. 35, 1798–1828.

Bennett, C., Solminihac, H. De, Chamorro, A., 2006. Data collection technologies for road management. Data Collect. Technol. Road Manag.

Bhaskar, A., Chung, E., 2013. Fundamental understanding on the use of Bluetooth scanner as a complementary transport data. Transp. Res. Part C Emerg. Technol. 37, 42–72.

Bollen, J., Mao, H., Zeng, X., 2011. Twitter mood predicts the stock market. J. Comput. Sci. 2, 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

Booth, A., Sutton, A., Papaioannou, D., 2016. Systematic approaches to a successful literature review. Sage.

Bradley, P.S., Fayyad, U.M., 1998. Refining Initial Points for K-Means Clustering., in: ICML. Citeseer, pp. 91–99.

Brazdil, P.B., 2003. Ranking Learning Algorithms: Using IBL and Meta-Learning on Accuracy and Time Results. Mach. Learn. 50, 251–277.

Brodley, C., 1993. Addressing the selective superiority problem: Automatic algorithm/model class selection, in: Proceedings of the Tenth International Conference on Machine Learning. pp. 17–24.

Brodsky, A., Luo, J., 2015. Decision Guidance Analytics Language (DGAL), in: Proceedings of the 17th International Conference on Enterprise Information Systems-Volume 1. SCITEPRESS-Science and Technology Publications, Lda, pp. 67–78.

Brodsky, A., Wang, S.X., 2008. Decision-Guidance Management Systems (DGMS): Seamless integration of data acquisition, learning, prediction, and optimization, in: Proceedings of the Annual Hawaii International Conference on System Sciences. IEEE, pp. 71–71. https://doi.org/10.1109/HICSS.2008.114

Brokke, G., Mertz, W., 1958. Evaluating trip forecasting methods with an electronic computer. Highw. Res. Board Bull. 203.

Bruza, P.D., Weeber, M., 2008. Literature-based discovery. Springer.

Buch, N., Velastin, S.A., Orwell, J., 2011. A review of computer vision techniques for the analysis of urban traffic. IEEE Trans. Intell. Transp. Syst. 12, 920–939.

Ceneda, D., Gschwandtner, T., May, T., Miksch, S., Schulz, H., Streit, M., Tominski, C., 2017. Characterizing Guidance in Visual Analytics. IEEE Trans. Vis. Comput. Graph. https://doi.org/10.1109/TVCG.2016.2598468

Chen, B., Cheng, H.H., 2010. A review of the applications of agent technology in traffic and transportation systems. IEEE Trans. Intell. Transp. Syst. 11, 485–497. https://doi.org/10.1109/TITS.2010.2048313

Chen, H., Chiang, R.H.L., Storey, V.C., 2012. Business intelligence and analytics: From big data to big impact. MIS Q. 36.

Chen, W., Hirschheim, R., 2004. A paradigmatic and methodological examination of information systems research from 1991 to 2001. Inf. Syst. J. 14, 197–235.

Churchman, C.W., 1967. Guest Editorial: Wicked Problems. Manage. Sci. https://doi.org/10.1366/000370209787169876

Cohen, W.W., 1995. Fast Effective Rule Induction, in: Machine Learning Proceedings 1995. https://doi.org/10.1016/B978-1-55860-377-6.50023-2

Collins, C., Andrienko, N., Schreck, T., Yang, J., Choo, J., Engelke, U., Jena, A., Dwyer, T., 2018. Guidance in the human–machine analytics process. Vis. Informatics 2, 166–180. https://doi.org/10.1016/j.visinf.2018.09.003

Collis, J., Hussey, R., 2013. Business research: A practical guide for undergraduate and postgraduate students. Macmillan International Higher Education.

Collopy, F., Armstrong, J.S., 1992. Rule-Based Forecasting: Development and Validation of an Expert Systems Approach to Combining Time Series Extrapolations. Manage. Sci. 38, 1394–1414. https://doi.org/10.1287/mnsc.38.10.1394

Cortes, C., Vapnik, V., 1995. Support vector machine. Mach. Learn. 20, 273–297.

Davenport, T., Harris, J., 2017. Competing on Analytics: Updated, with a New Introduction: The New Science of Winning. Harvard Business Press.

Davis, G.A., Nihan, N.L., 2007. Nonparametric Regression and Short-Term Freeway Traffic Forecasting. J. Transp. Eng. https://doi.org/10.1061/(asce)0733-947x(1991)117:2(178)

Denning, P.J., 1997. A new social contract for research. Commun. ACM 40, 132–134.

Dictionary, O.E., 2018. Oxford English Dictionary.

Dimitrakopoulos, G., Demestichas, P., 2010. Intelligent transportation systems. IEEE Veh. Technol. Mag. 5, 77–84.

Ding, Q.Y., Wang, X.F., Zhang, X.Y., 2011. Forecasting traffic volume with space-time ARIMA model. Adv. Mater. Res. 156, 979–983.

DLR - Institute of Transportation Systems - SUMO – Simulation of Urban MObility [WWW Document], 2019. URL https://www.dlr.de/ts/en/desktopdefault.aspx/tabid-9883/16931_read-41000/ (accessed 3.19.19).

Dombalyan, A., Kocherga, V., Semchugova, E., Negrov, N., 2017. Traffic Forecasting Model for a Road Section. Transp. Res. Procedia 20, 159–165. https://doi.org/10.1016/j.trpro.2017.01.040

Dresch, A., Lacerda, D.P., Antunes, J.A.V., 2015. Design science research, in: Design Science Research. Springer, pp. 67–102.

Duan, P., Mao, G., Zhang, C., 2016. STARIMA-based traffic prediction with time-varying lags, in: 19th International Conference on Intelligent Transportation Systems. IEEE, pp. 1610–1615.

Dunne, S., Ghosh, B., 2011. Regime-based short-term multivariate traffic condition forecasting algorithm. J. Transp. Eng. 138, 455–466.

Edwards, J.S., Taborda, E.R., 2016. Using knowledge management to give context to analytics and big data and reduce strategic risk. Procedia Comput. Sci. 99, 36–49.

Elefteriadou, L., 2014. An introduction to traffic flow theory. Springer, New York.

Ermagun, A., Levinson, D., 2018. Spatiotemporal traffic forecasting: review and proposed directions. Transp. Rev. 38, 786–814.

https://doi.org/10.1080/01441647.2018.1442887

Esser, J., Schreckenberg, M., 1997. Microscopic simulation of urban traffic based on cellular automata. Int. J. Mod. Phys. C 8, 1025–1036.

Essien, A., Giannetti, C., 2019. A Deep Learning Framework for Univariate Time Series Prediction Using Convolutional LSTM Stacked Autoencoders, in: 2019 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA). IEEE, Sofia, pp. 1–6.

Essien, A., Petrounias, I., Sampaio, P., Sampaio, S., 2019a. Improving Urban Traffic Speed Prediction Using Data Source Fusion and Deep Learning, in: 2019 IEEE International Conference on Big Data and Smart Computing (BigComp). IEEE, pp. 1–8.

Essien, A., Petrounias, I., Sampaio, P., Sampaio, S., 2019b. Deep-PRESIMM: Integrating Deep Learning with Microsimulation for Traffic Prediction, in: IEEE International Conference on Systems, Man, and Cybernetics. IEEE Xplore, pp. 1–6.

Essien, A., Petrounias, I., Sampaio, P., Sampaio, S., 2018. The impact of rainfall and temperature on peak and off-peak urban traffic, in: International Conference on Database and Expert Systems Applications. Springer, Cham., pp. 399–407. https://doi.org/10.1007/978-3-319-98812-2_36

Evans, J.R., Lindner, C.H., 2012. Business Analytics: The Next Frontier for Decision Sciences. Decis. Line 43, 4–6. https://doi.org/10.1007/978-1-4614-6080-0

Falcocchio, J.C., Levinson, H.S., 2015. The Costs and Other Consequences of Traffic Congestion, in: Road Traffic Congestion: A Concise Guide. pp. 159–182. https://doi.org/10.1007/978-3-319-15165-6_13

Ferrari, D., De-Castro, L., 2015. Clustering algorithm selection by meta-learning systems: A new distance-based problem characterization and ranking combination methods. Inf. Sci. (Ny). 301, 181–194.

Fountoulakis, M., Bekiaris-Liberis, N., Roncoli, C., Papamichail, I., Papageorgiou, M., 2017. Highway traffic state estimation with mixed connected and conventional vehicles: Microscopic simulation-based testing. Transp. Res. Part C Emerg. Technol. 78, 13–33. https://doi.org/10.1016/j.trc.2017.02.015

Gipps, P.G., 1981. A Behavioural Car Following Model for Computer Simulation.pdf. Transp. Res. Part B Methodol.

Glaser, B.G., Strauss, A.L., 2017. Discovery of grounded theory: Strategies for qualitative research. Routledge.

Goh, G., Koh, J., Zhang, Y., 2018. Twitter-Informed Crowd Flow Prediction, in: 2018 IEEE Conference on Data Mining Workshops (ICDMW). pp. 624–631.

Gong, X., Wang, F., 2003. Three improvements on KNN-NPR for traffic flow forecasting, in: IEEE 5th International Conference on Intelligent Transportation Systems. IEEE Xplore, pp. 736–740.

Goodfellow, I., Bengio, Y., 2015. Deep learning. MIT Press. https://doi.org/10.1038/nmeth.3707

GOV.UK, 2018. Transport Statistics Great Britain - GOV.UK [WWW Document]. Transp. Stat. Gt. Britain. URL https://www.gov.uk/government/collections/transport-statistics-great-britain (accessed 5.26.19).

Goves, C., North, R., Johnston, R., Fletcher, G., 2016. Short term traffic prediction on the UK motorway network using neural networks. Transp. Res. Procedia 1, 184–195.

Greenshields, B., 1935. A study in highway capacity. Highw. Res. Board Proc. 1935, 448–477.

Greenshields, B., Channing, W., H Miller, 1935. A study of traffic capacity. Highw. Res. Board Proc. 1935.

Gregor, S., Hevner, A.R., 2013. Positioning and presenting design science research for maximum impact. MIS Q. 337–355.

Guo, C., Jensen, C.S., Yang, B., 2014. Towards Total Traffic Awareness. ACM SIGMOD Rec. 43, 18–23. https://doi.org/10.1145/2694428.2694432

Guo, J., Huang, W., Williams, B., 2014. Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification. Transp. Res. Part C Emerg. Technol. 43, 50–64.

Guo, J., Liu, Z., Huang, W., Wei, Y., Cao, J., 2017. Short-term traffic flow prediction using fuzzy information granulation approach under different time intervals. IET Intell. Transp. Syst. 12, 143–150.

Halkidi, M., Batistakis, Y., Vazirgiannis, M., 2001. On clustering validation techniques. J. Intell. Inf. Syst. 17, 107–145.

Hall, P., Gill, N., Kurka, M., Phan, W., 2017. Machine Learning Interpretability with H2O Driverless AI.

Hamed, M.M., Al-Masaeid, H.R., Said, Z.M.B., 1995. Short-term prediction of traffic volume in urban arterials. J. Transp. Eng. 121, 249–254.

Harris, J.D., Quatman, C.E., Manring, M.M., Siston, R.A., Flanigan, D.C., 2014. How to Write a Systematic Review. Am. J. Sports Med. 42, 2761–2768. https://doi.org/10.1177/0363546513497567

Hashemi, H., Abdelghany, K.F., 2016. Real-time traffic network state estimation and prediction with decision support capabilities: Application to integrated corridor management. Transp. Res. Part C Emerg. Technol. 73, 128–146. https://doi.org/10.1016/j.trc.2016.10.012

Hawkins, D.M., 2004. The Problem of Overfitting. J. Chem. Inf. Comput. Sci. https://doi.org/10.1021/ci0342472

He, Q., Head, K.L., Ding, J., 2012. PAMSCOD: Platoon-based arterial multi-modal signal control with online data. Transp. Res. Part C Emerg. Technol. 20, 164–184. https://doi.org/10.1016/j.trc.2011.05.007

Hecht-Nielsen, R., 1988. Theory of the backpropagation neural network. Neural Networks. https://doi.org/10.1016/0893-6080(88)90469-8

Heilman, W.E., Potter, B.E., Charney, J.J., Bian, X., 2002. Analysis of weather impacts on traffic flow in metropolitan Washington DC. Growth (Lakeland) 12–17. https://doi.org/10.1029/2002JD002184.Woo

Hevner, A., Chatterjee, S., 2010. Design Science Research in Information Systems, in: Design Science Research in Information Systems: Theory and Practice. pp. 9–22. https://doi.org/10.1007/978-1-4419-5653-8_2

Hevner, March, Park, Ram, 2004. Design Science in Information Systems Research. MIS Q. https://doi.org/10.2307/25148625

Hillmer, S.C., Tiao, G.C., 1982. An ARIMA-Model-Based Approach to Seasonal Adjustment. J. Am. Stat. Assoc. 77, 63–70. https://doi.org/10.1080/01621459.1982.10477767

Hinsbergen, J. van, Sanders, F., 2007. Short Term Traffic Prediction Models. ITS World Congr. Beijing, China. https://doi.org/10.1037/11584-022

Ho, S.L., Xie, M., Goh, T.N., 2002. A comparative study of neural network and Box-Jenkins ARIMA modeling in time series prediction, in: Computers and Industrial Engineering. https://doi.org/10.1016/S0360-8352(02)00036-0

Hochreiter, S., 1998. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. Int. J. Uncertainty, Fuzziness Knowledge-Based Syst. 6, 107–116. https://doi.org/10.1142/S0218488598000094

Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J., 2001. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. A F. Guid. to Dyn. Recurr. Neural Networks.

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Comput. 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Horni, A., Nagel, K., Axhausen, K.W., 2016. The Multi-Agent Transport Simulation MATSim. Ubiquity Press, London. https://doi.org/10.5334/baw

Hossin, M., Sulaiman, M.N., Mustapha, A., Mustapha, N., Rahmat, R.W., 2011. A hybrid evaluation metric for optimizing classifier, in: Conference on Data Mining and Optimization. pp. 165–170. https://doi.org/10.1109/DMO.2011.5976522

Hristovski, D., Peterlin, B., Mitchell, J.A., Humphrey, S.M., Sitbon, L., Turner, I., 2003. Improving literature based discovery support by genetic knowledge integration. Stud Heal. Technol Inf. 95.

Hsu, C.W., Lin, C.J., 2002. A comparison of methods for multiclass support vector machines. IEEE Trans. Neural Networks. https://doi.org/10.1109/72.991427

Huang, J., Ling, C.X., 2007. Constructing new and better evaluation measures for machine learning, in: IJCAI International Joint Conference on Artificial Intelligence. pp. 859–864.

Huang, J., Ling, C.X., 2005. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans. Knowl. Data Eng. 17, 299–310. https://doi.org/10.1109/TKDE.2005.50

Ishak, S., Al-Deek, H., 2002. Performance Evaluation of Short-Term Time-Series Traffic Prediction Model. J. Transp. Eng. https://doi.org/10.1061/(ASCE)0733-947X(2002)128:6(490)

Jia, Y., Wu, J., Ben-Akiva, M., Seshadri, R., Du, Y., 2017a. Rainfall-integrated traffic speed prediction using deep learning method. IET Intell. Transp. Syst. 11, 531–536. https://doi.org/10.1049/iet-its.2016.0257

Jia, Y., Wu, J., Du, Y., 2016. Traffic speed prediction using deep learning method, in: IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC. https://doi.org/10.1109/ITSC.2016.7795712

Jia, Y., Wu, J., Xu, M., 2017b. Traffic flow prediction with rainfall impact using a deep learning method. J. Adv. Transp. 2017. https://doi.org/10.1155/2017/6575947

Jiang, H., Zou, Y., Zhang, S., Tang, J., Wang, Y., 2016. No Title. Math. Probl. Eng.

Joyce, T., Herrmann, J.M., 2018. A review of no free lunch theorems, and their implications for metaheuristic optimisation, in: Nature-Inspired Algorithms and Applied Optimization. Springer, pp. 27–51.

Julier, S.J., Uhlmann, J.K., 1997. New extension of the Kalman filter to nonlinear systems 182. https://doi.org/10.1117/12.280797

Kalman, R., 1960. A New Approach to Linear Filtering and Prediction Problems. J. Basic Eng. 82, 35–45.

Kalpakis, K., Gada, D., Puttagunta, V., 2001. Distance measures for effective clustering of ARIMA time-series, in: Proceedings 2001 IEEE International Conference on Data Mining. IEEE, pp. 273–280.

Karlaftis, M.G., Vlahogianni, E.I., 2011. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. Transp. Res. Part C Emerg. Technol. 19, 387–399. https://doi.org/10.1016/j.trc.2010.10.004

Keller, J.M., Gray, M.R., Givens, J.A., 1985. A fuzzy k-nearest neighbor algorithm. IEEE Trans. Syst. Man. Cybern. 4, 580–585.

Kirby, H.R., Watson, S.M., Dougherty, M.S., 1997. Should we use neural networks or statistical models for short-term motorway traffic forecasting? Int. J. Forecast. https://doi.org/10.1016/S0169-2070(96)00699-1

Koetse, M.J., Rietveld, P., 2009. The impact of climate change and weather on transport: An overview of empirical findings. Transp. Res. Part D Transp. Environ. 14, 205–221.

Korhonen, A., Guo, Y., Baker, S., Yetisgen-Yildiz, M., Stenius, U., Narita, M., Liò, P., 2014. Improving literature-based discovery with advanced text mining, in: International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics. Springer, pp. 89–98.

Kotthoff, L., Thornton, C., Hoos, H.H., Hutter, F., Leyton-Brown, K., 2016. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. J. Mach. Learn. Res. https://doi.org/10.1016/0022-1694(93)90238-5

Kuhn, M., Johnson, K., 2013. Applied predictive modeling. Springer.

Kumar, K., Parida, M., Katiyar, V.K., 2015. Short term traffic flow prediction in heterogeneous condition using artificial neural network. Transport. https://doi.org/10.3846/16484142.2013.818057

Kumar, S.V., Vanajakshi, L., 2015. Short-term traffic flow prediction using seasonal ARIMA model with limited input data. Eur. Transp. Res. Rev. 7, 21. https://doi.org/10.1007/s12544-015-0170-8

Kwon, J., Coifman, B., Bickel, P., 2000. Day-to-Day Travel-Time Trends and Travel-Time Prediction from Loop-Detector Data. Transp. Res. Rec. J. Transp. Res. Board. https://doi.org/10.3141/1717-15

Lai, W.-K., Kuo, T.-H., Chen, C.-H., 2016. Vehicle Speed Estimation and Forecasting Methods Based on Cellular Floating Vehicle Data. Appl. Sci. 6, 47. https://doi.org/10.3390/app6020047

Lana, I., Del-Ser, J., Velez, M., Vlahogianni, E.I., 2018. Road Traffic Forecasting: Recent Advances and New Challenges. IEEE Intell. Transp. Syst. Mag. 10, 93–109. https://doi.org/10.1109/MITS.2018.2806634

Laud, P.W., Ibrahim, J.G., 1995. Predictive Model Selection. J. R. Stat. Soc. Ser. B 57, 247–262. https://doi.org/10.1111/j.2517-6161.1995.tb02028.x

LeCun, Y., Bengio, Y., 1995. Convolutional networks for images, speech, and time series. Handb. brain theory neural networks 3361. https://doi.org/10.1590/S0102-09352009000200027

LeCun, Y., Corinna, C., Chris, B., 2018. MNIST handwritten digit database [WWW Document]. New York Univ.

Leduc, G., 2008. Road traffic data: Collection methods and applications. Work. Pap. Clim. Transp. Clim. Chang. 1, 55.

Levinson, H.S., Zimmerman, S., Clinger, J., Rutherford, G., 2002. Bus Rapid Transit: An Overview. J. Public Transp. 5, 1. https://doi.org/10.1038/scientificamerican1209-53

Levy, J.I., Buonocore, J.J., Von Stackelberg, K., 2010. Evaluation of the public health impacts of traffic congestion: a health risk assessment. Environ. Heal. 9, 65.

Liang, Z., Wakahara, Y., 2014. Real-time urban traffic amount prediction models for dynamic route guidance systems. EURASIP J. Wirel. Commun. Netw. https://doi.org/10.1186/1687-1499-2014-85

Lin, Lu, Li, J., Chen, F., Ye, J., Huai, J., 2018. Road Traffic Speed Prediction: A Probabilistic Model Fusing Multi-Source Data. IEEE Trans. Knowl. Data Eng. https://doi.org/10.1109/TKDE.2017.2718525

Lin, L, Li, J., Chen, F., Ye, J., J Huai, 2018. Road traffic speed prediction: a probabilistic model fusing multi-source data. IEEE Trans. Knowl. Data Eng. 30.7, 1310–1323.

Lindauer, M., Hutter, F., Hoos, H.H., Schaub, T., 2017. AutoFolio: An automatically configured algorithm selector. IJCAI Int. Jt. Conf. Artif. Intell. 53, 5025–5029. https://doi.org/10.1613/jair.4726

Lv, Y., Tang, S., Zhao, H., 2009. Real-time highway traffic accident prediction based on the k-nearest neighbor method, in: 2009 International Conference on Measuring Technology and Mechatronics Automation. pp. 547–550.

Ma, J., Perkins, S., 2014. Time-series novelty detection using one-class support vector machines. Proc. Int. Jt. Conf. Neural Networks, 2003. 3, 1741–1745. https://doi.org/10.1109/IJCNN.2003.1223670

Ma, X., Tao, Z., Wang, Yinhai, Yu, H., Wang, Yunpeng, 2015. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transp. Res. Part C Emerg. Technol. 54, 187–197. https://doi.org/10.1016/j.trc.2015.03.014

Mahmassani, H.S., 2001. Dynamic Network Traffic Assignment and Simulation Methodology for Advanced System Management Applications. Networks Spat. Econ. 1, 267–292. https://doi.org/10.1023/A:1012831808926

Makridakis, S., Hibon, M., Moser, C., 1979. Accuracy of Forecasting: An Empirical Investigation. J. R. Stat. Soc. Ser. A 142, 97. https://doi.org/10.2307/2345077

March, S.T., Smith, G.F., 1995. Design and natural science research on information technology. Decis. Support Syst. 15, 251–266.

MEN, 2018. Why are there only 40 people working on Manchester’s massive roadworks and why do they stop at 5pm? - Manchester Evening News [WWW Document]. URL https://www.manchestereveningnews.co.uk/news/greater-manchester-news/regent-road-mancunian-way-roadworks-15258493 (accessed 9.7.19).

Mendoza, H., Klein, A., Feurer, M., Springenberg, J.T., Hutter, F., 2016. Towards Automatically-Tuned Neural Networks, in: Proceedings of the Workshop on Automatic Machine Learning.

Min, W., Wynter, L., 2011. Real-time road traffic prediction with spatio-temporal correlations. Transp. Res. Part C Emerg. Technol. 19, 606–616. https://doi.org/10.1016/j.trc.2010.10.002

Min, X., Hu, J., Chen, Q., T Zhang, 2009. Short-term traffic flow forecasting of urban network based on dynamic STARIMA model, in: 12th International IEEE Conference on Intelligent Transportation Systems. IEEE, pp. 1–6.

Mitchell, T., 2006. The discipline of machine learning.

Moayedi, H., MA Masnadi-Shirazi, 2008. Arima model for network traffic prediction and anomaly detection. Int. Symp. Inf. Technol. 4, 1–6.

Mulrow, C.D., Cook, D.J., Davidoff, F., 1997. Systematic reviews: critical links in the great chain of evidence. Ann. Intern. Med. 126, 389–391.

Nafi, N.S., Khan, R.H., Khan, J.Y., Gregory, M., 2015. A predictive road traffic management system based on vehicular ad-hoc network, in: 2014 Australasian Telecommunication Networks and Applications Conference, ATNAC 2014. https://doi.org/10.1109/ATNAC.2014.7020887

Nguyen, T.H., Shirai, K., Velcin, J., 2015. Sentiment analysis on social media for stock movement prediction. Expert Syst. Appl. 42, 9603–9611. https://doi.org/10.1016/j.eswa.2015.07.052

Ni, M., He, Q., Gao, J., 2014. Using social media to predict traffic flow under special event conditions, in: The 93rd Annual Meeting of Transportation Research Board.

Nowlan, S.J., Hinton, G.E., 1992. Simplifying Neural Networks by Soft Weight Sharing. Neural Comput. 4, 473–493. https://doi.org/10.1201/9780429492525-13

Nunamaker Jr, J.F., Chen, M., Purdin, T.D.M., 1990. Systems development in information systems research. J. Manag. Inf. Syst. 7, 89–106.

Oh, S., Byon, Y.J., Jang, K., Yeo, H., 2015. Short-term Travel-time Prediction on Highway: A Review of the Data-driven Approach. Transp. Rev. 35, 4–32. https://doi.org/10.1080/01441647.2014.992496

Okutani, I., Stephanedes, Y.J., 1984. Dynamic prediction of traffic volume through Kalman filtering theory. - Transp. Res. Part B Methodol. 18, 1–11.

Orlikowski, W.J., Iacono, C.S., 2001. Research commentary: Desperately seeking the “IT” in IT research—A call to theorizing the IT artifact. Inf. Syst. Res. 12, 121–134.

Pappa, G.L., Ochoa, G., Hyde, M.R., Freitas, A.A., Woodward, J., Swan, J., 2014. Contrasting meta-learning and hyper-heuristic research: the role of evolutionary algorithms. Genet. Program. Evolvable Mach. 15, 3–35. https://doi.org/10.1007/s10710-013-9186-9

Peffers, K., Tuunanen, T., Rothenberger, M.A., Chatterjee, S., 2007. A Design Science Research Methodology for Information Systems Research. J. Manag. Inf. Syst. 24, 45–77. https://doi.org/10.2753/MIS0742-1222240302

Peng, Y., Jiang, Y., Lu, J., Zou, Y., 2018. Examining the effect of adverse weather on road transportation using weather and traffic sensors. PLoS One. https://doi.org/10.1371/journal.pone.0205409

Piironen, J., Vehtari, A., 2017. Comparison of Bayesian predictive methods for model selection. Stat. Comput. 27, 711–735. https://doi.org/10.1007/s11222-016-9649-y

Poonia, P., Jain, V.K., Kumar, A., 2018. Short Term Traffic Flow Prediction Methodologies: A Review. Mody Univ. Int. J. Comput. Eng. Res. 2, 37–39.

Pozzebon, M., Pinsonneault, A., 2005. Challenges in conducting empirical work using structuration theory: Learning from IT research. Organ. Stud. 26, 1353–1376.

Prat, N., Comyn-Wattiau, I., Akoka, J., 2014. Artifact Evaluation in Information Systems Design-Science Research-a Holistic View., in: PACIS. p. 23.

Qiao, W., Haghani, A., Hamedi, M., 2013. A nonparametric model for short-term travel time prediction using bluetooth data. J. Intell. Transp. Syst. Technol. Planning, Oper. https://doi.org/10.1080/15472450.2012.748555

Qiu, L., Nixon, W.A., 2008. Effects of Adverse Weather on Traffic Crashes. Transp. Res. Rec. J. Transp. Res. Board 2055, 139–146. https://doi.org/10.3141/2055-16

Reif, M., Shafait, F., Goldstein, M., Breuel, T., Dengel, A., 2014. Automatic classifier selection for non-experts. Pattern Anal. Appl. 17, 83–96. https://doi.org/10.1007/s10044-012-0280-z

Rice, J., 1976. The algorithm selection problem. Adv. Comput. 15, 65–118.

Rodriguez, P., Wiles, J., Elman, J.L., 1999. A Recurrent Neural Network that Learns to Count. Conn. Sci. 11, 5–40. https://doi.org/10.1080/095400999116340

RStudio, 2016. Shiny. R Cheat Sheet. https://doi.org/10.1080/14786419.2013.841686

Ruch, P., 2010. Literature-based Discovery. J. Am. Soc. Inf. Sci. Technol. https://doi.org/10.1002/asi.21236

Rumelhart, D., Hinton, G., Williams, R., 1988. Learning representations by back-propagating errors, 1st ed, Cognitive Modelling.

Russel, S., Norvig, P., 2012. Artificial intelligence—a modern approach 3rd Edition, The Knowledge Engineering Review. https://doi.org/10.1017/S0269888900007724

Sabour, S., Frosst, N., Hinton, G.E., 2017. Dynamic routing between capsules, in: Advances in Neural Information Processing Systems. pp. 3856–3866.

Sebastian, Y., Siew, E.-G., Orimaye, S.O., 2017. Emerging approaches in literature-based discovery: Techniques and performance review. Knowl. Eng. Rev. 32.

Sein, M., Henfridsson, O., Purao, S., Rossi, M., Lindgren, R., 2011. Action design research.

Shmueli, G., Koppius, O.R., 2011. Predictive analytics in information systems research. MIS Q. 553–572.

Simon, H.A., 1996. The sciences of the artificial. MIT press.

Smith-Miles, K., 2009. Cross-disciplinary perspectives on meta-learning for algorithm selection. ACM Comput. Surv. 41, 6.

Smith, B., Byrne, K., Copperman, R., Hennessy, S., Goodall, N., 2004. An investigation into the impact of rainfall on freeway traffic flow. 83rd Annu. Meet. Transp. Res. Board, Washingt. DC.

Smith, B.L., Demetsky, M., 1997. Traffic Flow Forecasting: Comparison of Modeling Approaches. J. Transp. Eng. https://doi.org/10.1061/(ASCE)0733-947X(1997)123:4(261)

Smith, B.L., Williams, B.M., Keith Oswald, R., 2002. Comparison of parametric and nonparametric models for traffic flow forecasting. Transp. Res. Part C Emerg. Technol. 10, 303–321. https://doi.org/10.1016/S0968-090X(02)00009-8

Smith, M., Mitchell, L., C Giraud-Carrier, 2014. Recommending learning algorithms and their associated hyperparameters. arxiv.org 1407.

Srivastava, N., Hinton, G., Krizhevsky, A., Salakhutdinov, R., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research.

Sun, H., Liu, H., Xiao, H., He, R., B Ran, 2003. Short term traffic forecasting using the local linear regression model, in: 82nd Annual Meeting of the Transportation Research Board. Washington, DC.

Swanson, D.R., 1988. Migraine and magnesium: eleven neglected connections. Perspect. Biol. Med. 31, 526–557.

Taylor, M., Bonsall, P., 2017. Understanding traffic systems: data analysis and presentation.

Tebaldi, C., West, M., Karr, A.F., 2002. Statistical analyses of freeway traffic flows. J. Forecast. https://doi.org/10.1002/for.804

Thacker, N., AJ Lacey, 1996. Tutorial: The likelihood interpretation of the kalman filter. TINA Memos Adv. Appl. Stat. 2, 1–11.

Thornton, Chris, Hutter, F., Hoos, H.H., Leyton-Brown, K., 2013. Auto-WEKA. Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. https://doi.org/10.1145/2487575.2487629

Thornton, C, Hutter, F., Hoos, H.H., Leyton-Brown, K., 2013. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms, in: 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 847–855.

Tian, Y., Pan, L., 2015. Predicting short-term traffic flow by long short-term memory recurrent neural network, in: 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity). IEEE, pp. 153–158.

Tian, Y., Zhang, K., Li, J., Lin, X., Yang, B., 2018. LSTM-based traffic flow prediction with missing data. Neurocomputing 318, 297–305.

Tsapakis, I., Cheng, T., Bolbol, A., 2013. Impact of weather conditions on macroscopic urban travel times. J. Transp. Geogr. https://doi.org/10.1016/j.jtrangeo.2012.11.003

Vaishnavi, V., Kuechler, W., 2004. Design research in information systems.

Van-de-Geer, J., 1995. Some aspects of Minkowski distance.

Van-Lint, J., Hoogendoorn, S., Van-Zuylen, H., 2005. Accurate freeway travel time prediction with state-space neural networks under missing data. Transp. Res. Part C Emerg. Technol. 13, 347–369.

Venable, J., Pries-Heje, J., Baskerville, R., 2016. FEDS: a framework for evaluation in design science research. Eur. J. Inf. Syst. 25, 77–89.

Vilalta, R., Giraud-Carrier, C., Brazdil, P., Soares, C., 2004. Using Meta-Learning to Support Data Mining. IJCSA 1, 31–45.

Vlahogianni, E.I., Golias, J.C., Karlaftis, M.G., 2004. Short-term traffic forecasting: Overview of objectives and methods. Transp. Rev. 24, 533–557. https://doi.org/10.1080/0144164042000195072

Vlahogianni, E.I., Karlaftis, M.G., Golias, J.C., 2014. Short-term traffic forecasting: Where we are and where we're going. Transp. Res. Part C Emerg. Technol. 43, 3–19. https://doi.org/10.1016/j.trc.2014.01.005

Vythoulkas, P., 1993. Alternative approaches to short term traffic forecasting for use in driver information systems. Transp. traffic theory 12, 485–506.

Wang, C., Quddus, M., Ison, S., 2009. Impact of traffic congestion on road accidents: A spatial analysis of the M25 motorway in England. Accid. Anal. Prev. 41, 798–808.

Wang, X., Gerber, M.S., Brown, D.E., 2012. Automatic crime prediction using events extracted from twitter posts, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, Berlin, pp. 231–238. https://doi.org/10.1007/978-3-642-29047-3_28

Wang, Xiaofeng, Smith-Miles, K., Hyndman, R., 2009. Rule induction for forecasting method selection: Meta-learning the characteristics of univariate time series. Neurocomputing 72, 10–12.

Wang, Xiaozhe, Smith-Miles, K., Hyndman, R., 2009. Rule induction for forecasting method selection: Meta-learning the characteristics of univariate time series. Neurocomputing 72, 2581–2594.

Weigend, A.S., 2018. Time series prediction: forecasting the future and understanding the past. Routledge.

Welch, G., Bishop, G., 1995. An introduction to the Kalman filter.

Williams, B., Durvasula, P., Brown, D., 1998. Urban Freeway Traffic Flow Prediction: Application of Seasonal Autoregressive Integrated Moving Average and Exponential Smoothing Models. Transp. Res. Rec. J. Transp. Res. Board. https://doi.org/10.3141/1644-14

Wohlin, C., Aurum, A., 2015. Towards a decision-making structure for selecting a research design in empirical software engineering. Empir. Softw. Eng. 20, 1427–1455.

Wolpert, D.H., Macready, W.G., 1997. No free lunch theorems for optimization. IEEE Trans. Evol. Comput. https://doi.org/10.1109/4235.585893

Wongcharoen, S., Senivongse, T., 2016. Twitter analysis of road traffic congestion severity estimation, in: 2016 13th International Joint Conference on Computer Science and Software Engineering, JCSSE 2016. IEEE, pp. 1–6. https://doi.org/10.1109/JCSSE.2016.7748850

Wu, C.-H., Ho, J.-M., Lee, D.-T., 2004. Travel time prediction with support vector regression, in: IEEE Conference on Intelligent Transportation Systems. pp. 276–281. https://doi.org/10.1109/ITSC.2003.1252721

Xu, H., Dailey, D.J., 1995. Real time highway traffic simulation and prediction using inductance loop data, in: Vehicle Nagivation and Information Systems Conference Proceedings. IEEE Xplore, pp. 194–199.

Yu, F., Guo, J., Zhu, X., Shi, G., 2015. Real time prediction of unoccupied parking space using time series model, in: ICTIS 2015 - 3rd International Conference on Transportation Information and Safety, Proceedings. pp. 370–374. https://doi.org/10.1109/ICTIS.2015.7232145

Zhang, H., Li, Z., Shahriar, H., Tao, L., Bhattacharya, P., Qian, Y., 2019. Improving Prediction Accuracy for Logistic Regression on Imbalanced Datasets, in: 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC). IEEE, pp. 918–919.

Zhang, L., Liu, Q., Yang, W., Wei, N., Dong, D., 2013. An improved k-nearest neighbor model for short-term traffic flow prediction. Procedia-Social Behav. Sci. 96, 653–662.

Zheng, W., Lee, D.-H., Shi, Q., 2006. Short-Term Freeway Traffic Flow Prediction: Bayesian Combined Neural Network Approach. J. Transp. Eng. 132, 114–121. https://doi.org/10.1061/(ASCE)0733-947X(2006)132:2(114)

Zhou, M., Qu, X., Li, X., 2017. A recurrent neural network based microscopic car following model to predict traffic oscillation. Transp. Res. Part C Emerg. Technol. 84, 245–264. https://doi.org/10.1016/j.trc.2017.08.027

Zhou, T., Han, G., Xu, X., Han, C., Huang, Y., J Qin, 2019. A Learning-Based Multimodel Integrated Framework for Dynamic Traffic Flow Forecasting. Neural Process. Lett. 1–24.

Zhu, G., Wang, L., Zhang, P., Song, K., 2016. A Kind of Urban Road Travel Time Forecasting Model with Loop Detectors. Int. J. Distrib. Sens. Networks 12. https://doi.org/10.1155/2016/9043835

Appendix

a. Meta-level dataset used within study

No.	Location	Analysis Level	Traffic Scope	Data Source(s)	Univariate?	DCM	Dataset Size	PH	Real-Time?	DAM
1	China	Area	Urban	Pollution Data	T	Manual	Not Large	1	F	ANN
2	Netherlands	Link	Motorway	Traffic	T	ILD	Large	1	T	ANN
3	Germany	Link	Motorway	Traffic + Accident	F	Microwave /Radar	Not Large	525600	F	ANN
4	N/A	Link	Motorway	Traffic	T	Manual	Large	525600	F	ANN
5	Hong Kong	Link	Motorway	Traffic + Accident	F	ILD	Not Large	525600	F	ANN
6	Iran	Link	Motorway	Traffic Noise	T	Manual	Not Large	525600	F	ANN
7	India	Link	Motorway	Traffic + Environmental	F	ILD	Not Large	525600	F	ANN
8	China	Link	Motorway	Traffic	T	FCD	Not Large	15	F	ANN
9	Ethopia	Junction	Motorway	Traffic	T	Manual	Not Large	15	F	ANN
10	Australia	Area	Urban	Traffic	T	ILD	Large	15	F	ANN
11	USA	Link	Motorway	Traffic	T	Camera	Large	15	F	ANN
12	China	Link	Urban	Traffic	T	ILD	Not Large	15	F	ANN
13	N/A	Link	Motorway	Traffic	T	ILD	Not Large	15	F	ANN
14	Australia	Link	Motorway	Traffic	T	ILD	Not Large	15	F	ANN
15	China	Link	Motorway	Traffic	T	ILD	Not Large	2	F	ANN
16	Hong Kong	Link	Urban	Traffic	T	Manual	Not Large	2	T	ANN
17	N/A	Link	Motorway	Traffic + Noise Level	F	ILD	Not Large	20	F	ANN
18	USA	Link	Motorway	Traffic + Accident	F	ILD	Not Large	1440	F	ANN
19	UAE	Area	Urban	Environmental	T	Manual	Not Large	1440	F	ANN
20	N/A	Area	Urban	Traffic + Environmental	F	Manual	Large	1440	F	ANN
21	India	Link	Motorway	Traffic + Noise Level	F	ILD	Not Large	1440	F	ANN
22	Spain	Area	Urban	Traffic + Meteorological	F	ILD	Large	1440	F	ANN
23	Cyprus	Area	Urban	Pollution Data	T	Manual	Not Large	1440	F	ANN
24	China	Area	Urban	Traffic + Weather + Web Crawler + Events	F	FCD	Not Large	30	F	ANN
25	China	Area	Urban	Traffic	T	FCD	Large	30	F	ANN
26	USA	Link	Motorway	Traffic	T	ILD	Not Large	35	F	ANN
27	Netherlands	Link	Motorway	Traffic	T	ILD	Large	4	F	ANN
28	Hungary	Area	Urban	Environmental	T	Manual	Not Large	2880	F	ANN
29	China	Link	Motorway	Traffic	T	ILD	Large	5	F	ANN
30	USA	Link	Motorway	Traffic + Accident	F	ILD	Large	5	F	ANN
31	China	Link	Motorway	Traffic	T	ILD	Large	5	F	ANN
32	USA	Link	Motorway	Traffic	T	ILD	Not Large	5	F	ANN
33	China	Link	Motorway	Traffic	T	Microwave /Radar	Not Large	5	F	ANN
34	Italy	Link	Urban	Traffic	T	Manual	Large	60	F	ANN

35	China	Area	Urban	Pollution Data	T	Manual	Not Large	1	F	ANN
36	Netherlands	Link	Motorway	Traffic	T	ILD	Large	1	T	ANN
37	Germany	Link	Motorway	Traffic + Accident	F	Microwave/Radar	Not Large	525600	F	ANN
38	N/A	Link	Motorway	Traffic	T	Manual	Large	525600	F	ANN
39	Hong Kong	Link	Motorway	Traffic + Accident	F	ILD	Not Large	525600	F	ANN
40	Iran	Link	Motorway	Traffic Noise	T	Manual	Not Large	525600	F	ANN
41	India	Link	Motorway	Traffic + Environmental	F	ILD	Not Large	525600	F	ANN
42	China	Link	Motorway	Traffic	T	FCD	Not Large	15	F	ANN
43	Ethopia	Junction	Motorway	Traffic	T	Manual	Not Large	15	F	ANN
44	Australia	Area	Urban	Traffic	T	ILD	Large	15	F	ANN
45	USA	Link	Motorway	Traffic	T	Camera	Large	15	F	ANN
46	China	Link	Urban	Traffic	T	ILD	Not Large	15	F	ANN
47	N/A	Link	Motorway	Traffic	T	ILD	Not Large	15	F	ANN
48	Australia	Link	Motorway	Traffic	T	ILD	Not Large	15	F	ANN
49	China	Link	Motorway	Traffic	T	ILD	Not Large	2	F	ANN
50	Hong Kong	Link	Urban	Traffic	T	Manual	Not Large	2	T	ANN
51	N/A	Link	Motorway	Traffic + Noise Level	F	ILD	Not Large	20	F	ANN
52	USA	Link	Motorway	Traffic + Accident	F	ILD	Not Large	1440	F	ANN
53	UAE	Area	Urban	Environmental	T	Manual	Not Large	1440	F	ANN
54	N/A	Area	Urban	Traffic + Environmental	F	Manual	Large	1440	F	ANN
55	India	Link	Motorway	Traffic + Noise Level	F	ILD	Not Large	1440	F	ANN
56	Spain	Area	Urban	Traffic + Meteorological	F	ILD	Large	1440	F	ANN
57	Cyprus	Area	Urban	Pollution Data	T	Manual	Not Large	1440	F	ANN
58	China	Area	Urban	Traffic + Weather + Web Crawler + Events	F	FCD	Not Large	30	F	ANN
59	China	Area	Urban	Traffic	T	FCD	Large	30	F	ANN
60	USA	Link	Motorway	Traffic	T	ILD	Not Large	35	F	ANN
61	Netherlands	Link	Motorway	Traffic	T	ILD	Large	4	F	ANN
62	Hungary	Area	Urban	Environmental	T	Manual	Not Large	2880	F	ANN
63	China	Link	Motorway	Traffic	T	ILD	Large	5	F	ANN
64	USA	Link	Motorway	Traffic + Accident	F	ILD	Large	5	F	ANN
65	China	Link	Motorway	Traffic	T	ILD	Large	5	F	ANN
66	USA	Link	Motorway	Traffic	T	ILD	Not Large	5	F	ANN
67	China	Link	Motorway	Traffic	T	Microwave/Radar	Not Large	5	F	ANN
68	Italy	Link	Urban	Traffic	T	Manual	Large	60	F	ANN
69	China	Area	Urban	Pollution Data	T	Manual	Not Large	1	F	ANN
70	Netherlands	Link	Motorway	Traffic	T	ILD	Large	1	T	ANN
71	Germany	Link	Motorway	Traffic + Accident	F	Microwave/Radar	Not Large	525600	F	ANN
72	N/A	Link	Motorway	Traffic	T	Manual	Large	525600	F	ANN
73	Hong Kong	Link	Motorway	Traffic + Accident	F	ILD	Not Large	525600	F	ANN
74	Iran	Link	Motorway	Traffic Noise	T	Manual	Not Large	525600	F	ANN

75	India	Link	Motorway	Traffic + Environmental	F	ILD	Not Large	525600	F	ANN
76	China	Link	Motorway	Traffic	T	FCD	Not Large	15	F	ANN
77	Ethopia	Junction	Motorway	Traffic	T	Manual	Not Large	15	F	ANN
78	Australia	Area	Urban	Traffic	T	ILD	Large	15	F	ANN
79	USA	Link	Motorway	Traffic	T	Camera	Large	15	F	ANN
80	China	Link	Urban	Traffic	T	ILD	Not Large	15	F	ANN
81	N/A	Link	Motorway	Traffic	T	ILD	Not Large	15	F	ANN
82	Australia	Area	Motorway	Traffic Only	T	ILD	Not Large	15	F	ANN
83	China	Area	Urban	Parking	T	Manual	Large	15	T	ANN
84	Malaysia	Link	Urban	Traffic	T	ILD	Not Large	1440	F	ANN
85	Greece	Link	Urban	Traffic	T	ILD	Not Large	3	F	ARIMA
86	Netherlands	Area	Motorway	Traffic	T	ILD	Large	5	T	ARIMA
87	Canada	Link	Motorway	Traffic	T	Bluetooth	Large	5	F	ARIMA
88	Germany	Area	Urban	Traffic	T	ILD	Not Large	5	F	ARIMA
89		Link	Motorway	Traffic	T	ILD	Large	5	F	ARIMA
90	N/A	Area	Urban	Traffic	T	ILD	Not Large	5	T	ARIMA
91	China	Link	Urban	Traffic	T	ILD	Large	5	F	ARIMA
92	China	Link	Urban	Traffic	T	ILD	Large	5	F	ARIMA
93	USA	Link	Motorway	Traffic	T	ILD	Not Large	5	F	ARIMA
94	USA	Link	Motorway	Traffic	T	Manual	Not Large	60	F	ARIMA
95	Australia	Area	Motorway	Traffic Only	T	ILD	Not Large	15	F	ARIMA
96	China	Area	Urban	Parking	T	Manual	Large	15	T	ARIMA
97	Malaysia	Link	Urban	Traffic	T	ILD	Not Large	1440	F	ARIMA
98	Greece	Link	Urban	Traffic	T	ILD	Not Large	3	F	ARIMA
99	Netherlands	Area	Motorway	Traffic	T	ILD	Large	5	T	ARIMA
100	Canada	Link	Motorway	Traffic	T	Bluetooth	Large	5	F	ARIMA
101	Germany	Area	Urban	Traffic	T	ILD	Not Large	5	F	ARIMA
102		Link	Motorway	Traffic	T	ILD	Large	5	F	ARIMA
103	N/A	Area	Urban	Traffic	T	ILD	Not Large	5	T	ARIMA
104	China	Link	Urban	Traffic	T	ILD	Large	5	F	ARIMA
105	China	Link	Urban	Traffic	T	ILD	Large	5	F	ARIMA
106	USA	Link	Motorway	Traffic	T	ILD	Not Large	5	F	ARIMA
107	USA	Link	Motorway	Traffic	T	Manual	Not Large	60	F	ARIMA
108	Australia	Area	Motorway	Traffic Only	T	ILD	Not Large	15	F	ARIMA
109	China	Area	Urban	Parking	T	Manual	Large	15	T	ARIMA
110	Malaysia	Link	Urban	Traffic	T	ILD	Not Large	1440	F	ARIMA
111	Greece	Link	Urban	Traffic	T	ILD	Not Large	3	F	ARIMA
112	Netherlands	Area	Motorway	Traffic	T	ILD	Large	5	T	ARIMA
113	Canada	Link	Motorway	Traffic	T	Bluetooth	Large	5	F	ARIMA
114	Germany	Area	Urban	Traffic	T	ILD	Not Large	5	F	ARIMA
115		Link	Motorway	Traffic	T	ILD	Large	5	F	ARIMA
116	N/A	Area	Urban	Traffic	T	ILD	Not Large	5	T	ARIMA

117	China	Link	Urban	Traffic	T	ILD	Large	5	F	ARIMA
118	China	Link	Urban	Traffic	T	ILD	Large	5	F	ARIMA
119	USA	Link	Motorway	Traffic	T	ILD	Not Large	5	F	ARIMA
120	USA	Link	Motorway	Traffic	T	Manual	Not Large	60	F	ARIMA
121	Australia	Area	Motorway	Traffic Only	T	ILD	Not Large	15	F	ARIMA
122	China	Area	Urban	Parking	T	Manual	Large	15	T	ARIMA
123	Malaysia	Link	Urban	Traffic	T	ILD	Not Large	1440	F	ARIMA
124	Greece	Link	Urban	Traffic	T	ILD	Not Large	3	F	ARIMA
125	Netherlands	Area	Motorway	Traffic	T	ILD	Large	5	T	ARIMA
126	Canada	Link	Motorway	Traffic	T	Bluetooth	Large	5	F	ARIMA
127	Germany	Area	Urban	Traffic	T	ILD	Not Large	5	F	ARIMA
128		Link	Motorway	Traffic	T	ILD	Large	5	F	ARIMA
129	N/A	Area	Urban	Traffic	T	ILD	Not Large	5	T	ARIMA
130	China	Link	Urban	Traffic	T	ILD	Large	5	F	ARIMA
131	China	Link	Urban	Traffic	T	ILD	Large	5	F	ARIMA
132	USA	Link	Motorway	Traffic	T	ILD	Not Large	5	F	ARIMA
133	USA	Link	Motorway	Traffic	T	Manual	Not Large	60	F	ARIMA
134	France	Link	Motorway	Traffic	T	ILD	Not Large	1	T	ARIMA
135	USA	Area	Urban	Traffic	T	ILD	Large	1	T	ARIMA
136	USA	Junction	Motorway	Traffic	T	ILD	Not Large	15	T	ARIMA
137	Belgium	Junction	Motorway	Traffic	T	ILD	Not Large	15	T	ARIMA
138	USA	Link	Motorway	Traffic	T	Manual	Large	5	T	ARIMA
139	Japan	Junction	Urban	Traffic	T	ILD	Not Large	5	T	ARIMA
140	USA	Area	Urban	Traffic	T	FCD	Large	5	T	ARIMA
141	France	Link	Motorway	Traffic	T	ILD	Not Large	1	T	ARIMA
142	USA	Area	Urban	Traffic	T	ILD	Large	1	T	ARIMA
143	USA	Junction	Motorway	Traffic	T	ILD	Not Large	15	T	ARIMA
144	Belgium	Junction	Motorway	Traffic	T	ILD	Not Large	15	T	ARIMA
145	USA	Link	Motorway	Traffic	T	Manual	Large	5	T	ARIMA
146	Japan	Junction	Urban	Traffic	T	ILD	Not Large	5	T	ARIMA
147	USA	Area	Urban	Traffic	T	FCD	Large	5	T	ARIMA
148	France	Link	Motorway	Traffic	T	ILD	Not Large	1	T	ARIMA
149	USA	Area	Urban	Traffic	T	ILD	Large	1	T	KF
150	USA	Junction	Motorway	Traffic	T	ILD	Not Large	15	T	KF
151	Belgium	Junction	Motorway	Traffic	T	ILD	Not Large	15	T	KF
152	USA	Link	Motorway	Traffic	T	Manual	Large	5	T	KF
153	Japan	Junction	Urban	Traffic	T	ILD	Not Large	5	T	KF
154	USA	Area	Urban	Traffic	T	FCD	Large	5	T	KF
155	France	Link	Motorway	Traffic	T	ILD	Not Large	1	T	KF
156	USA	Area	Urban	Traffic	T	ILD	Large	1	T	KF
157	USA	Junction	Motorway	Traffic	T	ILD	Not Large	15	T	KF
158	Belgium	Junction	Motorway	Traffic	T	ILD	Not Large	15	T	KF

159	USA	Link	Motorway	Traffic	T	Manual	Large	5	T	KF
160	Japan	Junction	Urban	Traffic	T	ILD	Not Large	5	T	KF
161	USA	Area	Urban	Traffic	T	FCD	Large	5	T	KF
162	USA	Link	Motorway	Traffic	T	ILD	Not Large	5	T	KF
163	France	Link	Motorway	Traffic	T	ILD	Not Large	1	T	KF
164	USA	Area	Urban	Traffic	T	ILD	Large	1	T	KF
165	USA	Junction	Motorway	Traffic	T	ILD	Not Large	15	T	KF
166	Belgium	Junction	Motorway	Traffic	T	ILD	Not Large	15	T	KF
167	USA	Link	Motorway	Traffic	T	Manual	Large	5	T	KF
168	Japan	Junction	Urban	Traffic	T	ILD	Not Large	5	T	KF
169	USA	Area	Urban	Traffic	T	FCD	Large	5	T	KF
170	France	Link	Motorway	Traffic	T	ILD	Not Large	1	T	KF
171	USA	Area	Urban	Traffic	T	ILD	Large	1	T	KF
172	USA	Junction	Motorway	Traffic	T	ILD	Not Large	15	T	KF
173	Belgium	Junction	Motorway	Traffic	T	ILD	Not Large	15	T	KF
174	USA	Link	Motorway	Traffic	T	Manual	Large	5	T	KF
175	Japan	Junction	Urban	Traffic	T	ILD	Not Large	5	T	KF
176	USA	Area	Urban	Traffic	T	FCD	Large	5	T	KF
177	France	Link	Motorway	Traffic	T	ILD	Not Large	1	T	KF
178	USA	Area	Urban	Traffic	T	ILD	Large	1	T	KF
179	USA	Junction	Motorway	Traffic	T	ILD	Not Large	15	T	KF
180	Belgium	Junction	Motorway	Traffic	T	ILD	Not Large	15	T	KF
181	USA	Link	Motorway	Traffic	T	Manual	Large	5	T	KF
182	Japan	Junction	Urban	Traffic	T	ILD	Not Large	5	T	KF
183	USA	Area	Urban	Traffic	T	FCD	Large	5	T	KF
184	France	Link	Motorway	Traffic	T	ILD	Not Large	1	T	KF
185	USA	Area	Urban	Traffic	T	ILD	Large	1	T	KF
186	USA	Junction	Motorway	Traffic	T	ILD	Not Large	15	T	KF
187	Belgium	Junction	Motorway	Traffic	T	ILD	Not Large	15	T	KF
188	USA	Link	Motorway	Traffic	T	Manual	Large	5	T	KF
189	Japan	Junction	Urban	Traffic	T	ILD	Not Large	5	T	KF
190	USA	Area	Urban	Traffic	T	FCD	Large	5	T	KF
191	USA	Link	Motorway	Traffic	T	ILD	Not Large	5	T	KF
192	USA	Link	Motorway	Accident	T	Manual	Large	525600	F	KF
193	Mixed	Link	Motorway	Traffic	T	ILD	Large	15	F	KF
194	China	Link	Urban	Traffic	T	FCD	Large	5	T	KF
195	China	Link	Urban	Traffic	T	FCD	Large	5	F	KF
196	USA	Link	Motorway	Traffic	T	ILD	Not Large	5	F	KF
197	China	Link	Urban	Traffic	T	FCD	Large	5	F	KF
198	China	Link	Motorway	Traffic	T	Bluetooth	Large	5	F	KF
199	USA	Link	Motorway	Accident	T	Manual	Large	525600	F	KF
200	Mixed	Link	Motorway	Traffic	T	ILD	Large	15	F	KF

201	China	Link	Urban	Traffic	T	FCD	Large	5	T	KF
202	China	Link	Urban	Traffic	T	FCD	Large	5	F	KF
203	USA	Link	Motorway	Traffic	T	ILD	Not Large	5	F	KF
204	China	Link	Urban	Traffic	T	FCD	Large	5	F	KF
205	China	Link	Motorway	Traffic	T	Bluetooth	Large	5	F	KF
206	USA	Link	Motorway	Accident	T	Manual	Large	525600	F	KF
207	Mixed	Link	Motorway	Traffic	T	ILD	Large	15	F	KF
208	China	Link	Urban	Traffic	T	FCD	Large	5	T	KF
209	China	Link	Urban	Traffic	T	FCD	Large	5	F	KF
210	USA	Link	Motorway	Traffic	T	ILD	Not Large	5	F	KF
211	China	Link	Urban	Traffic	T	FCD	Large	5	F	k-NN
212	China	Link	Motorway	Traffic	T	Bluetooth	Large	5	F	k-NN
213	USA	Link	Motorway	Accident	T	Manual	Large	525600	F	k-NN
214	Mixed	Link	Motorway	Traffic	T	ILD	Large	15	F	k-NN
215	China	Link	Urban	Traffic	T	FCD	Large	5	T	k-NN
216	China	Link	Urban	Traffic	T	FCD	Large	5	F	k-NN
217	USA	Link	Motorway	Traffic	T	ILD	Not Large	5	F	k-NN
218	China	Link	Urban	Traffic	T	FCD	Large	5	F	k-NN
219	China	Link	Motorway	Traffic	T	Bluetooth	Large	5	F	k-NN
220	USA	Link	Motorway	Accident	T	Manual	Large	525600	F	k-NN
221	Mixed	Link	Motorway	Traffic	T	ILD	Large	15	F	k-NN
222	China	Link	Urban	Traffic	T	FCD	Large	5	T	k-NN
223	China	Link	Urban	Traffic	T	FCD	Large	5	F	k-NN
224	USA	Link	Motorway	Traffic	T	ILD	Not Large	5	F	k-NN
225	China	Link	Urban	Traffic	T	FCD	Large	5	F	k-NN
226	China	Link	Motorway	Traffic	T	Bluetooth	Large	5	F	k-NN
227	USA	Link	Motorway	Accident	T	Manual	Large	525600	F	k-NN
228	Mixed	Link	Motorway	Traffic	T	ILD	Large	15	F	k-NN
229	China	Link	Urban	Traffic	T	FCD	Large	5	T	k-NN
230	China	Link	Urban	Traffic	T	FCD	Large	5	F	k-NN
231	USA	Link	Motorway	Traffic	T	ILD	Not Large	5	F	k-NN
232	China	Link	Urban	Traffic	T	FCD	Large	5	F	k-NN
233	China	Link	Motorway	Traffic	T	Bluetooth	Large	5	F	k-NN
234	USA	Link	Motorway	Accident	T	Manual	Large	525600	F	k-NN
235	Mixed	Link	Motorway	Traffic	T	ILD	Large	15	F	k-NN
236	China	Link	Urban	Traffic	T	FCD	Large	5	T	k-NN
237	China	Link	Urban	Traffic	T	FCD	Large	5	F	k-NN
238	USA	Link	Motorway	Traffic	T	ILD	Not Large	5	F	k-NN
239	China	Link	Urban	Traffic	T	FCD	Large	5	F	k-NN
240	China	Link	Motorway	Traffic	T	Bluetooth	Large	5	F	k-NN
241	USA	Link	Motorway	Accident	T	Manual	Large	525600	F	k-NN
242	Mixed	Link	Motorway	Traffic	T	ILD	Large	15	F	k-NN

243	China	Link	Urban	Traffic	T	FCD	Large	5	T	k-NN
244	China	Link	Urban	Traffic	T	FCD	Large	5	F	k-NN
245	USA	Link	Motorway	Traffic	T	ILD	Not Large	5	F	k-NN
246	China	Link	Urban	Traffic	T	FCD	Large	5	F	k-NN
247	China	Link	Motorway	Traffic	T	Bluetooth	Large	5	F	k-NN
248	USA	Link	Motorway	Traffic	T	ILD	Not Large	1	F	k-NN
249	India	Link	Motorway	Traffic + Accident	F	ILD	Not Large	525600	F	k-NN
250	Germany	Area	Motorway	Accident	T	Manual	Not Large	525600	F	k-NN
251	Europe	Junction	Motorway	Pollution Data	T	Manual	Large	525600	F	k-NN
252	Greece	Area	Urban	Environmental	T	Manual	Not Large	525600	F	k-NN
253	Ghana	Link	Motorway	Traffic + Accident	F	Microwave/Radar	Not Large	525600	F	LR
254	UK	Link	Motorway	Traffic	T	ILD	Not Large	10	F	LR
255	USA	Link	Motorway	Accident + Traffic	F	ILD	Not Large	15	F	LR
256	UK	Link	Motorway	Traffic	T	ILD	Not Large	15	F	LR
257	Australia	Area	Motorway	Traffic + Noise Level	F	Manual	Not Large	15	F	LR
258	N/A	Link	Motorway	Traffic	T	FCD	Not Large	20	F	LR
259	USA	Link	Motorway	Traffic	T	ILD	Not Large	5	F	LR
260	Hong Kong	Area	Urban	Traffic + Pedestrian + Car Park	F	Bluetooth	Large	60	F	LR
261	China	Link	Motorway	Traffic + Accident	F	ILD	Not Large	60	F	LR
262	USA	Link	Motorway	Traffic	T	ILD	Not Large	1	F	LR
263	India	Link	Motorway	Traffic + Accident	F	ILD	Not Large	525600	F	LR
264	Germany	Area	Motorway	Accident	T	Manual	Not Large	525600	F	LR
265	Europe	Junction	Motorway	Pollution Data	T	Manual	Large	525600	F	LR
266	Greece	Area	Urban	Environmental	T	Manual	Not Large	525600	F	LR
267	Ghana	Link	Motorway	Traffic + Accident	F	Microwave/Radar	Not Large	525600	F	LR
268	UK	Link	Motorway	Traffic	T	ILD	Not Large	10	F	LR
269	USA	Link	Motorway	Accident + Traffic	F	ILD	Not Large	15	F	LR
270	UK	Link	Motorway	Traffic	T	ILD	Not Large	15	F	LR
271	Australia	Area	Motorway	Traffic + Noise Level	F	Manual	Not Large	15	F	LR
272	N/A	Link	Motorway	Traffic	T	FCD	Not Large	20	F	LR
273	USA	Link	Motorway	Traffic	T	ILD	Not Large	5	F	LR
274	Hong Kong	Area	Urban	Traffic + Pedestrian + Car Park	F	Bluetooth	Large	60	F	LR
275	China	Link	Motorway	Traffic + Accident	F	ILD	Not Large	60	F	LR
276	USA	Link	Motorway	Traffic	T	ILD	Not Large	1	F	LR
277	India	Link	Motorway	Traffic + Accident	F	ILD	Not Large	525600	F	LR
278	Germany	Area	Motorway	Accident	T	Manual	Not Large	525600	F	LR
279	Europe	Junction	Motorway	Pollution Data	T	Manual	Large	525600	F	LR
280	Greece	Area	Urban	Environmental	T	Manual	Not Large	525600	F	LR
281	Ghana	Link	Motorway	Traffic + Accident	F	Microwave/Radar	Not Large	525600	F	LR
282	UK	Link	Motorway	Traffic	T	ILD	Not Large	10	F	LR
283	USA	Link	Motorway	Accident + Traffic	F	ILD	Not Large	15	F	LR

284	UK	Link	Motorway	Traffic	T	ILD	Not Large	15	F	LR
285	Australia	Area	Motorway	Traffic + Noise Level	F	Manual	Not Large	15	F	LR
286	N/A	Link	Motorway	Traffic	T	FCD	Not Large	20	F	LR
287	USA	Link	Motorway	Traffic	T	ILD	Not Large	5	F	LR
288	Hong Kong	Area	Urban	Traffic + Pedestrian + Car Park	F	Bluetooth	Large	60	F	LR
289	China	Link	Motorway	Traffic + Accident	F	ILD	Not Large	60	F	LR
290	USA	Link	Motorway	Traffic	T	ILD	Not Large	1	F	LR
291	India	Link	Motorway	Traffic + Accident	F	ILD	Not Large	525600	F	LR
292	Germany	Area	Motorway	Accident	T	Manual	Not Large	525600	F	LR
293	Europe	Junction	Motorway	Pollution Data	T	Manual	Large	525600	F	LR
294	Greece	Area	Urban	Environmental	T	Manual	Not Large	525600	F	LR
295	Ghana	Link	Motorway	Traffic + Accident	F	Microwave /Radar	Not Large	525600	F	LR
296	UK	Link	Motorway	Traffic	T	ILD	Not Large	10	F	LR
297	USA	Link	Motorway	Accident + Traffic	F	ILD	Not Large	15	F	LR
298	UK	Link	Motorway	Traffic	T	ILD	Not Large	15	F	LR
299	Australia	Area	Motorway	Traffic + Noise Level	F	Manual	Not Large	15	F	LR
300	N/A	Link	Motorway	Traffic	T	FCD	Not Large	20	F	LR
301	USA	Link	Motorway	Traffic	T	ILD	Not Large	5	F	LR
302	Hong Kong	Area	Urban	Traffic + Pedestrian + Car Park	F	Bluetooth	Large	60	F	LR
303	China	Link	Motorway	Traffic + Accident	F	ILD	Not Large	60	F	LR
304	China	Link	Motorway	Traffic	T	Manual	Large	40	F	LR
305	USA	Link	Motorway	Traffic	T	ILD	Large	45	F	LR
306	USA	Link	Motorway	Traffic	T	ILD	Large	30	F	LR
307	`	Link	Motorway	Traffic	T	ILD	Not Large	30	F	LR
308	China	Link	Motorway	Traffic	T	Manual	Large	45	F	LR
309	USA	Link	Motorway	Traffic	T	ILD	Large	45	F	LR
310	USA	Link	Motorway	Traffic	T	ILD	Large	30	F	LR
311	USA	Link	Motorway	Traffic	T	ILD	Not Large	30	F	LR
312	China	Link	Motorway	Traffic	T	Manual	Large	40	F	LSTM
313	USA	Link	Motorway	Traffic	T	ILD	Large	45	F	LSTM
314	USA	Link	Motorway	Traffic	T	ILD	Large	30	F	LSTM
315	USA	Link	Motorway	Traffic	T	ILD	Not Large	30	F	LSTM
316	China	Link	Motorway	Traffic	T	Manual	Large	40	F	LSTM
317	USA	Link	Motorway	Traffic	T	ILD	Large	45	F	LSTM
318	USA	Link	Motorway	Traffic	T	ILD	Large	30	F	LSTM
319	USA	Link	Motorway	Traffic	T	ILD	Not Large	30	F	LSTM
320	China	Link	Motorway	Traffic	T	Manual	Large	40	F	LSTM
321	USA	Link	Motorway	Traffic	T	ILD	Large	45	F	LSTM
322	USA	Link	Motorway	Traffic	T	ILD	Large	30	F	LSTM
323	USA	Link	Motorway	Traffic	T	ILD	Not Large	25	F	LSTM
324	China	Link	Motorway	Traffic	T	Manual	Large	30	F	LSTM

325	USA	Link	Motorway	Traffic	T	ILD	Large	30	F	LSTM
326	USA	Link	Motorway	Traffic	T	ILD	Large	30	F	LSTM
327	USA	Link	Motorway	Traffic	T	ILD	Not Large	30	F	LSTM
328	China	Link	Motorway	Traffic	T	Manual	Large	40	F	LSTM
329	USA	Link	Motorway	Traffic	T	ILD	Large	45	F	LSTM
330	USA	Link	Motorway	Traffic	T	ILD	Large	30	F	LSTM
331	USA	Link	Motorway	Traffic	T	ILD	Not Large	30	F	LSTM
332	China	Link	Motorway	Traffic	T	Manual	Large	40	F	LSTM
333	USA	Link	Motorway	Traffic	T	ILD	Large	45	F	LSTM
334	USA	Link	Motorway	Traffic	T	ILD	Large	30	F	LSTM
335	`	Link	Motorway	Traffic	T	ILD	Not Large	30	F	LSTM
336	China	Link	Motorway	Traffic	T	Manual	Large	45	F	LSTM
337	USA	Link	Motorway	Traffic	T	ILD	Large	45	F	LSTM
338	USA	Link	Motorway	Traffic	T	ILD	Large	30	F	LSTM
339	USA	Link	Motorway	Traffic	T	ILD	Not Large	30	F	LSTM
340	China	Link	Motorway	Traffic	T	Manual	Large	40	F	LSTM
341	USA	Link	Motorway	Traffic	T	ILD	Large	45	F	LSTM
342	USA	Link	Motorway	Traffic	T	ILD	Large	30	F	LSTM
343	USA	Link	Motorway	Traffic	T	ILD	Not Large	30	F	LSTM
344	China	Link	Motorway	Traffic	T	Manual	Large	40	F	LSTM
345	USA	Link	Motorway	Traffic	T	ILD	Large	45	F	LSTM
346	USA	Link	Motorway	Traffic	T	ILD	Large	30	F	LSTM
347	USA	Link	Motorway	Traffic	T	ILD	Not Large	30	F	LSTM
348	China	Link	Motorway	Traffic	T	Manual	Large	40	F	LSTM
349	USA	Link	Motorway	Traffic	T	ILD	Large	45	F	LSTM
350	USA	Link	Motorway	Traffic	T	ILD	Large	30	F	LSTM
351	USA	Link	Motorway	Traffic	T	ILD	Not Large	25	F	LSTM
352	China	Link	Motorway	Traffic	T	Manual	Large	30	F	LSTM
353	USA	Link	Motorway	Traffic	T	ILD	Large	30	F	LSTM
354	USA	Link	Motorway	Traffic	T	ILD	Large	30	F	LSTM
355	USA	Link	Motorway	Traffic	T	ILD	Not Large	30	F	LSTM
356	China	Link	Motorway	Traffic	T	Manual	Large	40	F	LSTM
357	USA	Link	Motorway	Traffic	T	ILD	Large	45	F	LSTM
358	USA	Link	Motorway	Traffic	T	ILD	Large	30	F	SVM
359	USA	Link	Motorway	Traffic	T	ILD	Not Large	30	F	SVM
360	Korea	Link	Motorway	Traffic	T	Bluetooth	Not Large	1	F	SVM
361	China	Area	Motorway	Traffic Safety	T	Manual	Not Large	525600	F	SVM
362	China	Link	Urban	Traffic	T	FCD	Not Large	15	F	SVM
363	China	Link	Motorway	Traffic + Accident	F	ILD	Not Large	5	F	SVM
364	China	Link	Motorway	Traffic	T	Microwave /Radar	Not Large	5	F	SVM
365	USA	Link	Urban	Traffic	T	ILD	Large	60	F	SVM
366	Korea	Link	Motorway	Traffic	T	Bluetooth	Not Large	1	F	SVM

367	China	Area	Motorway	Traffic Safety	T	Manual	Not Large	525600	F	SVM
368	China	Link	Urban	Traffic	T	FCD	Not Large	15	F	SVM
369	China	Link	Motorway	Traffic + Accident	F	ILD	Not Large	5	F	SVM
370	China	Link	Motorway	Traffic	T	Microwave /Radar	Not Large	5	F	SVM
371	USA	Link	Urban	Traffic	T	ILD	Large	60	F	SVM
372	Korea	Link	Motorway	Traffic	T	Bluetooth	Not Large	1	F	SVM
373	China	Area	Motorway	Traffic Safety	T	Manual	Not Large	525600	F	SVM
374	China	Link	Urban	Traffic	T	FCD	Not Large	15	F	SVM
375	China	Link	Motorway	Traffic + Accident	F	ILD	Not Large	5	F	SVM
376	China	Link	Motorway	Traffic	T	Microwave /Radar	Not Large	5	F	SVM
377	USA	Link	Urban	Traffic	T	ILD	Large	60	F	SVM
378	Korea	Link	Motorway	Traffic	T	Bluetooth	Not Large	1	F	SVM
379	China	Area	Motorway	Traffic Safety	T	Manual	Not Large	525600	F	SVM
380	China	Link	Urban	Traffic	T	FCD	Not Large	15	F	SVM
381	China	Link	Motorway	Traffic + Accident	F	ILD	Not Large	5	F	SVM
382	China	Link	Motorway	Traffic	T	Microwave /Radar	Not Large	5	F	SVM
383	USA	Link	Urban	Traffic	T	ILD	Large	60	F	SVM
384	Korea	Link	Motorway	Traffic	T	Bluetooth	Not Large	1	F	SVM
385	China	Area	Motorway	Traffic Safety	T	Manual	Not Large	525600	F	SVM
386	China	Link	Urban	Traffic	T	FCD	Not Large	15	F	SVM
387	China	Link	Motorway	Traffic + Accident	F	ILD	Not Large	5	F	SVM
388	China	Link	Motorway	Traffic	T	Microwave /Radar	Not Large	5	F	SVM
389	USA	Link	Urban	Traffic	T	ILD	Large	60	F	SVM
390	Korea	Link	Motorway	Traffic	T	Bluetooth	Not Large	1	F	SVM
391	China	Area	Motorway	Traffic Safety	T	Manual	Not Large	525600	F	SVM
392	China	Link	Urban	Traffic	T	FCD	Not Large	15	F	SVM
393	China	Link	Motorway	Traffic + Accident	F	ILD	Not Large	5	F	SVM
394	China	Link	Motorway	Traffic	T	Microwave /Radar	Not Large	5	F	SVM
395	USA	Link	Urban	Traffic	T	ILD	Large	60	F	SVM
396	Korea	Link	Motorway	Traffic	T	Bluetooth	Not Large	1	F	SVM
397	China	Area	Motorway	Traffic Safety	T	Manual	Not Large	525600	F	SVM
398	China	Link	Urban	Traffic	T	FCD	Not Large	15	F	SVM
399	China	Link	Motorway	Traffic + Accident	F	ILD	Not Large	5	F	SVM
400	China	Link	Motorway	Traffic	T	Microwave /Radar	Not Large	5	F	SVM
401	USA	Link	Urban	Traffic	T	ILD	Large	60	F	SVM
402	Korea	Link	Motorway	Traffic	T	Bluetooth	Not Large	1	F	SVM
403	China	Area	Motorway	Traffic Safety	T	Manual	Not Large	525600	F	SVM
404	China	Link	Urban	Traffic	T	FCD	Not Large	15	F	SVM
405	China	Link	Motorway	Traffic + Accident	F	ILD	Not Large	5	F	SVM
406	China	Link	Motorway	Traffic	T	Microwave /Radar	Not Large	5	F	SVM
407	USA	Link	Urban	Traffic	T	ILD	Large	60	F	SVM

b. Code snippet of TAG-F support tool

```
library(sh
iny)
       require(class)
       library(shiny)


       ui <- fluidPage(
         fluidRow(
           column(3, tags$img(height=100, src="logo.jpg")),
           column(9, tags$h1(tags$strong("The Traffic Data Analytics
       Guidance Framework (TAG-F) Support Tool")))
         ),

         tags$hr(),
         tags$br(),
         tags$h2("Overview"),
         tags$hr(),
         #tags$p("TAG-F is a traffic data analytics guidance framework
       that delineates data-driven traffic prediction as a set of three
       dimensions: (i) Data Context/Scope (DC), (ii) Data Analytical
       Method (DAM), and (iii) Data Collection Method (DCM). TAG-F
       support tool can serve as a decision support mechanism for traffic
       data scientists by providing guidance in the choice of DAM, given
       the data context specifications. The framework incorporates seven
       (7) candidate models ranging from time series, instance-based
       learning, machine learning, and deep learning models for traffic
       parameter prediction. The tool provides guidance for traffic data
       analytics via prediction model suggestion given a set of traffic
       data parameters. Select the parameters below and click 'Update'
       when completed."),
         tags$hr(),
         fluidRow(
           column(4, sliderInput(inputId = "sliderPH", label = "Select
       Prediction Time Steps (minutes)", value = 5, min = 1, max = 120)),
           column(4, selectInput(inputId = "listAL", label = "Select
       Analysis Level", c("-Select-", "Link", "Junction", "Area"), "-
       Select-", multiple = FALSE)
           ),
           column(4, selectInput(inputId = "listTrafficScope", label =
       "Traffic Scope", c("-Select-", "Urban", "Highway/Motorway"), "-
       Select-", multiple = FALSE))
         ),
         fluidRow(
    column(4, dateRangeInput(inputId = "lblDate", label = "Dataset
Date Range", start = NULL, end = NULL, format = "dd/mm/yyyy",
separator = "to")),
    column(4, selectInput(inputId = "listGranularity", label =
"Traffic Dataset Observation Frequency", c("-Select-", "Daily",
"Hourly", "Half-Hourly", "Minutes", "Seconds"), "-Select-",
multiple = FALSE)),
    column(4, selectInput(inputId = "listDCM", label = "Traffic
Data Collection Method", c("-Select-", "Manual", "ILD",
"Bluetooth", "Microwave/Radar", "FCD"), "-Select-", multiple =
FALSE))),

  fluidRow(
    column(4, checkboxInput(inputId = "cbRealtime", label = "Real-
Time Prediction?", value = FALSE)),
    column(4, checkboxInput(inputId = "size", label = "Large
Dataset?", value = FALSE)),
    column(4, checkboxInput(inputId = "univariate", label =
"Univariate Dataset?", value = FALSE))),



  tags$hr(),
  actionButton(inputId = "go", label = "Update Framework"),
  tags$hr(),
  textOutput(outputId = "txtDataGranularity"),
  tags$hr(),
  textOutput(outputId = "justification"),
  tags$hr(),
  plotOutput("hist")

)


server <- function(input, output, session) {
  data <- eventReactive(input$go, {input$sliderPH})

  output$hist <- renderPlot({
    model_suggest <- function(analysis, urban, univariate, dcm,
large, ph, realtime)
    {

      require(ggplot2)
      require(data.table)
      require(randomForest)
      colnames(train)[2] <- "Urban"
      colnames(train)[3] <- "Univariate"
      colnames(train)[4] <- "DCM"
      colnames(train)[5] <- "Large"
      colnames(train)[7] <- "Realtime"
      colnames(train)[8] <- "DAM"
      gdis<-randomForest(DAM ~ ., data=train, ntree=500,
keep.forest = TRUE,  importance=TRUE,
                         proximity=TRUE)
      pred <- knn(train = train_x, test = new_data,cl = train_y,
k=3)
      summary(pred)
      summary(gdis)
      new_data <- data.frame(Analysis=analysis, Urban=urban,
Univariate=univariate, DCM=dcm, Large=large, PH=ph,
Realtime=realtime)
      pred.model= predict(gdis,new_data,type="prob")
      new_d <- as.data.frame(t(pred.model))
      library(data.table)
      setDT(new_d, keep.rownames = TRUE)[]
      new_d <- new_d[order(new_d$rn, decreasing=T),]
      colnames(new_d)[1] <- "rn"
      colnames(new_d)[2] <- "V1"
      if ((new_data$Realtime=1)){ reason = "Since Realtime
prediction needed, therefore, Suggested DAM is KF"
      } else if ((new_data$Large>=1)&(new_data$PH>=45)){ reason =
"Since dataset is large and PH>45, therefore, Suggested PAM is
LSTM"
      } else if ((new_data$Large=1)&(new_data$PH<=10)){ reason =
"Since dataset is small and PH<10, therefore, Suggested PAM is
SVM"
      } else if ((new_data$Large=0)&(new_data$PH>=10)){ reason =
"Since dataset is small and PH<10, therefore, Suggested PAM is
ARIMA"
      } else if
((new_data$Urban=0)&(new_data$Anlaysis<0)&(new_data$DCM>=1)){reaso
n = "Since traffic scope is non-urban, analysis level is area, DCM
not manual, therefore, Suggested PAM is ARIMA"
      } else if
((new_data$Large=0)&(new_data$PH<=1)&(new_data$Realtime<=0)&(new_d
ata$Analysis>=1)){reason = "Since dataset is small, PH is small,
and non-realtime prediction required, therefore, Suggested PAM is
LR"
      } else if ((new_data$PH>=30)&(new_data$Analysis<=1)){reason
= "Since PH is large, and analysis level is link, therefore,
Suggested PAM is k-NN"
```

```
      } else if
((new_data$PH>=30)&(new_data$Large<=0)&(new_data$Univariate<=0)&(n
ew_data$DCM>=1)){reason = "Since PH is large, and dataset is large
multivariate, therefore, Suggested PAM is ANN (also LSTM)"
      } else {
        print("Still thinking")
      }
      return(ggplot(data = new_d, aes(x=reorder(rn, -V1), y=V1)) +
               geom_bar(stat="identity") +
               geom_text(aes(label=round(V1, digits=4)),
vjust=1.6, color="white", size=3.5)+
               theme_bw()+ labs(x="Model", y="Probability
(Confidence Level)")+
               annotate("label", x = 6, y=0.7, label = reason))

    }
    if(input$listDCM =="Manual"){
      data_dcm <- 0
    }
    else if(input$listDCM == "ILD"){
      data_dcm <- 1
    }
    else if(input$listDCM == "Bluetooth")
    {data_dcm <- 2}
    else if(input$listDCM == "Microwave/Radar")
    {data_dcm <- 3}
    else if(input$listDCM == "FCD")
    {data_dcm <- 4}


    if(input$listAL =="Link"){
      data_al <- 1
    }
    else if(input$listAL == "Area"){
      data_al <- 0
    }
    else if(input$listAL == "Junction")
    {data_al <- 2}


    if(input$listTrafficScope =="Urban"){
      data_ts <- 1
    }
    else if(input$listTrafficScope == "Highway/Motorway"){
      data_ts <- 0
    }
```

```
if(input$cbRealtime == FALSE){
  data_rt <- 0
}
else {
  data_rt <- 1
}


if(input$size == FALSE){
  data_gran <- 0
}
else {
  data_gran <- 1
}



if(input$univariate == FALSE){
  univariate <- 0
}
else {
  univariate <- 1
}

if(input$listAL =="Link"){
  data_al <- 1
}
else if(input$listAL == "Area"){
  data_al <- 0
}
else if(input$listAL == "Junction")
{data_al <- 2}


if(input$listGranularity =="Daily"){
  data_g <- 1
}
else if(input$listGranularity == "Hourly"){
  data_g <- 24
}
else if(input$listGranularity == "Half-Hourly")
{data_g <- 48}
else if(input$listGranularity == "Minutes")
{data_g <- 48}
```

```
    else if(input$listGranularity == "Seconds")
    {data_g <- 48}



    output$txtDataGranularity <- renderText({
      paste("The dataset depth is: ",(data_g*(input$lblDate[2]-
input$lblDate[1]))))})

    output$justification <- renderText({
      paste("Reason: ",(reason))})


model_suggest(data_al,data_ts,1,data_dcm,data_gran,data(),data_rt)
  })



}


shinyApp(ui, server)
```