

Enhanced short and longer term network performance prediction capabilities through data-driven analytics and simulation:

Short-term Traffic Speed Prediction for Perth Roads Using Machine Learning

Submitted to	Main Roads Western Australia & iMOVE CRC
Prepared by	Chris Bartley, Chao Sun, Mark Reynolds, Wei Liu and Sharon Biermann
Submitted by	Chao Sun (chao.sun@uwa.edu.au), Planning and Transport Research Centre (PATREC)
Steering Committee	Kamal Weeratunga, Steve Atkinson, Graham Jacoby and Chao Sun
Project details	iMOVE ITS Project 1-003, Sub-project 1 (Part 1), in part-fulfilment of Milestone 4
Date	16 January 2020
Version	Final

TABLE OF CONTENTS

Executive summary	1
1 Introduction.....	3
1.1 Background	3
1.2 Project brief	3
2 Data	5
2.1 Overview	5
2.2 Datasets	5
Time series data for observed traffic conditions	5
Exogenous variables	7
M-Link characteristics	7
2.3 Data Analysis Pipeline	8
3 Modelling methodology	9
4 Modelling results.....	10
4.1 Performance results.....	10
4.2 Feature selection results	12
5 Conclusions and future opportunities	13
References	14

EXECUTIVE SUMMARY

Big data is driving a rapid digital transformation in the transport industry. The Western Australian Auditor General (2015) pointed out that *'Having consistent, real time information is key to optimising network performance and informing strategic and operational decision-making (p.20).'* However, big data not only brings opportunities but also challenges. Laney (2001) introduced the commonly used three defining dimensions of big data – the 3Vs (volume, variety and velocity). It is not only the *amount* of data, but also the *variation* between datasets and the *speed at which* the data is generated. Others have later extended the concept and added more 'V's but essentially big data is being generated at a *rate* that surpasses human's ability to make sense of it and with levels of *noise* that make interpretation difficult. Therefore, building decision support systems that automate the analytics becomes a necessity.

Data is worthless unless it can be turned into actionable information. For Main Roads Western Australia's (Main Roads) Network Operations Directorate this can be done in two ways: *offline* analysis of historical performance and *online* application to aid real-time operations. The former has been successfully addressed by the Network Performance Reporting System (NetPREs), and the latter is the focus of this report.

This iMOVE sub-project focuses on short-term prediction of average speed for individual road sections in the Perth metropolitan road network up to a horizon of 75 minutes in advance. It aligns well with Network Operations' vision of *'predict in 20 (min), act in 5 (min), change the future'*. Although it will require substantial investment in data infrastructure, good short-term predictions could enable Main Roads to take a proactive approach to network operations, such as stopping gridlocks before they appear and preventing queue spillbacks. It would also enable faster incident detection and recovery.

This investigation used machine learning techniques that *learn from the past to predict the future*. The hypothesis was that machine learning could extract hidden value from the Main Roads datasets that would improve prediction performance over naïve and traditional approaches. The results show that our predictive models are robust and perform well against the benchmarks. The highlight is that the accuracy does not decrease dramatically with an increasing prediction time horizon (how far into the future the model predicts), e.g. during the AM and PM peaks, predicting 15 minutes ahead will produce an average percentage error about 9% while for 75 minutes it is just above 10%. The performance gap with benchmarks becomes more pronounced with increasing timespan (Figure 3 & Figure 4).

Long term, we envisage a traffic *'now-casting'* decision support system for network operations. The concept of *'now-casting'* refers to *'the prediction of the present, the very near future and the very recent past'* (Banbura et al. 2012, p196), a term borrowed from meteorology and economics. *'Predicting the present or the very recent past'* is needed because of the delay in data acquisition

and cleaning; and the dynamic nature of network operations makes predicting the very near future valuable. Knowing what is likely to happen could help traffic operators making more evidence-based decisions.

The project required numerous rounds of data cleaning, manipulation, modelling and experimentation. A data analysis pipeline was developed to ensure this process was repeatable, reliable and reproducible. It allows data transformations and modelling frameworks to be reused for rapid development when new datasets become available in the future. This pipeline will be handed over to Main Roads.

Given that there is no perfect single data source measuring traffic performance, we recommend further improvement on data quality to address consistency and accuracy issues. It is being partially addressed by another PATREC project on data fusion which is funded by Main Roads.

The research presented in this report delivers on Sub-project 1 (Part 1) of a larger research project comprising two sub-projects:

- Sub-project 1: **Data-driven empirical models for short-term traffic prediction (Part 1)** & non-route-based area optimisation of network productivity (Part 2)
- Sub-project 2: Simulating the traffic impact of AVs and CAVs to Perth's freeways and arterial roads

This research is funded by PATREC and the iMOVE CRC and supported by the Cooperative Research Centres program, an Australian Government initiative. The contribution of the steering committee throughout the project, in guiding, monitoring and review, is gratefully acknowledged.

1

INTRODUCTION

1.1

BACKGROUND

The iMOVE project 1-003, *Enhanced short and longer term network performance prediction capabilities through data-driven analytics and simulation*, was co-funded by:

- Planning and Transport Research Centre (PATREC)
- Main Roads Western Australia (Main Roads)
- iMOVE CRC
- The University of Western Australia (UWA)

The project comprised two subprojects:

- Subproject 1: **Data-driven empirical models for short-term traffic prediction (Part 1)** & non-route-based area optimisation of network productivity (Part 2)
- Subproject 2: Simulating the traffic impact of AVs and CAVs to Perth's freeways and arterial roads

This report summarises the findings of the short-term traffic prediction, which is Part 1 of Subproject 1.

The kick-off meeting was held on 12 February 2018 during which the project committee was formed:

- Kamal Weeratunga (Committee Chair) / Manager Network Performance (ACTING), Main Roads
- Graham Jacoby / Network Operations Analysis Manager, Main Roads
- Steve Atkinson / Principal Analyst Strategic Planning, Main Roads
- Chao Sun / Research Fellow (Project Leader), UWA

The project officially started in March 2018, following which the committee met monthly to review the progress and make decisions.

1.2

PROJECT BRIEF

This report summarises the development and findings of *predictive models* for short term *traffic speed* prediction on individual road sections (referred to as “M-Links” in the Network Performance Reporting System – NetPREs) throughout the Main Roads network in metropolitan Perth. The goal is to explore the use of emerging traffic datasets to improve network operations.

Historically, limited data was available for monitoring the status of the road network. Most sensors can only take measurements at particular points of the network, e.g. inductive loops embedded in

the pavement, or traffic counters. The restricted spatial coverage and the lack of ability to differentiate individual vehicles means their data is insufficient for monitoring network performance.

The prevalence of mobile and GPS devices have made several new datasets a reality for traffic monitoring. Currently, Main Roads has access to data generated by GPS and Bluetooth devices from multiple sources. Together with conventional datasets, they enable network-wide performance monitoring leading to rapid improvements in network performance analysis. Currently, most of the analysis is done offline using historical data. The next logical step would be to use near real-time data analysis powered by live streamed data. Although it will require substantial investment in data infrastructure, coupled with good short-term prediction, these data sources could enable Main Roads to take a proactive approach to network operations. For example, this could enable prevention of gridlocks before they appear, prevention of queue spillbacks, and faster incident detection and recovery. As the Western Australian Auditor General (2015) observed: *'having consistent, real time information is key to optimising network performance and informing strategic and operational decision-making (p.20).'*

Data is worthless unless it can be turned into actionable information. For Main Roads Network Operations Directorate this can be done in two ways: offline analysis of historical performance and online application to aid real-time operations. The former has been successfully addressed in NetPReS and the latter is the focus of this report. It addresses the development of predictive models, which are computationally feasible for real-time applications. However, currently the main obstacle to real-time operation is the funding of livestream data acquisition and necessary infrastructure to support real-time data processing and management.

We applied machine learning techniques to harvest the value of Main Roads diverse datasets. The objective of short-term prediction of average speed for individual road sections aligns well with Network Operations' vision of *'predicting in 20 (min), acting in 5 (min), changing the future'*.

The original project goal was to build a predictive model for traffic volume and speed on a given M-Link. However, it became apparent early on that traffic volume is very predictable and there is little value in trying to improve the prediction of it – if changes in volume do not significantly alter speed then Main Roads should not be concerned. Consequently, the team focused solely on predicting *speed* but expanded the timeframe from 30 minutes to projecting forward up to 75 minutes. The main criterion was *prediction accuracy*, with a secondary goal of understanding the relative importance of the features (variables)¹ used for prediction (*feature importance*).

To accomplish these objectives, the following tasks were undertaken:

¹ The terms features and variables are used interchangeably in this report.

1) **Task 1: Data pre-processing**

- a) Data exploration
- b) Data cleaning
- c) Feature engineering of the multiple datasets available for M-Link speed and exogenous variables (weather conditions, incidents etc.).
- d) Selection of the speed measure

2) **Task 2: Modelling at the link level**

- a) Development of appropriate model for speed prediction, where models are created for each M-Link separately
- b) Experimental evaluation of model performance against baseline models
- c) Presentation of results (relative feature importance, model performance)

2 DATA

2.1 OVERVIEW

While big data is driving a rapid digital transformation for the transport industry, it also comes with challenges. Laney (2001) identified the three defining dimensions of big data – the 3Vs (volume, variety and velocity). It is not only about the *amount* of data, but also *variation* within and between multiple datasets and the *speed* at which the data is generated. Others have later extended the concept and added more ‘V’s but essentially big data is being generated at a rate that surpasses human’s ability to make sense of it, making building decision support systems that can automate the analytics a necessity. Another challenge is the noise and inconsistency within the data, as illustrated in Figure 1. Section 2.2 presents the rudimentary approach to data cleaning used in this analysis, but a deeper investigation is underway by the separate PATREC project on data fusion (also funded by Main Roads).

2.2 DATASETS

The project timeline overlapped with the migration of NetPREs to the cloud so multiple datasets from different periods were used due to the ongoing development of NetPREs. Results reported in this report are based on the following sources, which can be broken down into three categories.

TIME SERIES DATA FOR OBSERVED TRAFFIC CONDITIONS

- **Vehicle Detection Stations (VDS, also referred to as NPI in the Main Roads NetPREs dataset) Speed:** Vehicle Detection Stations are only available on the freeways and selected major highways
- **GPS1 Speed:** aggregate GPS tracked vehicle speed data supplied by company 1

- **GPS2 Speed:** aggregate GPS tracked vehicle speed data supplied by company 2
- **VDS volumes:** vehicle counts on freeways
- **SCATS volumes:** arterial volumes reported by SCATS detectors

All the above datasets were aggregated by 15-minute interval.

A number of sources were available for speed data, and it was required to determine a single 'speed' estimate to be used for modelling. The different data sources have significantly different provenance, and as a result have very different 'coverage' of the network. This is summarised in Table 1.

Table 1 Speed data source coverage (1 January 2019 to 30 June 2019)

Route Name	Num_Mlinks		VDS_Speed		GPS2_HarmSpeed		GPS1_HarmSpeed		Bluetooth_Speed	
	Fwy	Art	Fwy	Art	Fwy	Art	Fwy	Art	Fwy	Art
Kwinana Fwy	72	5	90%	79%	69%	93%	96%	97%	90%	100%
Mitchell Fwy	50	19	89%	94%	87%	91%	97%	92%	97%	99%
Roe Hwy	45	6	22%	33%	67%	61%	96%	90%	99%	99%
Tonkin Hwy	43	4	30%	0%	61%	72%	94%	95%	99%	50%
Reid Hwy	24	12	14%	17%	74%	67%	95%	85%	99%	99%
Melville Mandurah Hwy	20	44	0%	0%	27%	38%	89%	90%	50%	92%
Tonkin Hwy and Northlink	10	2	0%	0%	95%	91%	95%	97%	99%	99%
Wanneroo Rd & Indian Ocean Dr	10	62	0%	0%	17%	61%	78%	94%	40%	96%
Great Eastern Hwy	8	71	0%	0%	40%	80%	80%	89%	50%	94%
Toodyay Rd	4	0	0%	-	19%	-	74%	-	0%	-
Albany Hwy	2	81	0%	0%	28%	68%	82%	92%	99%	95%
Armadale Rd / South Western Hwy	2	30	0%	0%	17%	38%	81%	89%	99%	79%
Brookton Hwy	2	0	0%	-	24%	-	83%	-	100%	-
Great Northern Hwy	2	2	0%	0%	28%	37%	91%	90%	0%	99%
Thomas Rd	2	8	0%	0%	33%	35%	92%	90%	100%	99%
Canning Hwy	0	49	-	0%	-	74%	-	91%	-	94%
Cockburn Rd	0	6	-	0%	-	33%	-	86%	-	100%
Graham Farmer Fwy	0	19	-	93%	-	96%	-	97%	-	91%
Guildford Rd	0	32	-	0%	-	76%	-	94%	-	99%
Karriyup-Morley Hwy	0	38	-	0%	-	60%	-	93%	-	83%
Leach Hwy	0	71	-	6%	-	70%	-	85%	-	86%
Marmion Av	0	14	-	0%	-	54%	-	93%	-	99%
Rivervale-Wattle Grove Link	0	28	-	0%	-	76%	-	96%	-	97%
South St	0	31	-	0%	-	68%	-	95%	-	92%
Stirling Hwy	0	36	-	0%	-	70%	-	94%	-	99%
West Coast Hwy	0	22	-	0%	-	43%	-	90%	-	71%

GPS1 and VDS data were chosen as the main sources for this prediction project based on exploratory analysis and a resulting judgement of their quality. For modelling purposes a single 'ModelSpeed' variable was created as follows: the VDS speed data was used unless an M-Link had more than 30% missing values, in which case GPS1 data was used. In effect this meant that VDS was usually used for freeways and GPS1 for arterials.

Nevertheless, there are still inconsistencies between GPS1 and VDS data (Figure 1). The misalignment between other datasets are even more pronounced. It is mostly caused by low sample rates, measurement errors and different natures of the measurements. There is no single perfect data source for traffic monitoring so the data inconsistency problem will be there for the foreseeable future. To partially address this problem, PATREC has developed a data fusion method for Main Roads under a separate project, which is being evaluated by the Network Operations

Analysis team. The pipeline can be easily re-run once the quality of other datasets is improved or confirmed. Our preliminary results also show that the model accuracy improved by running on the fused data.

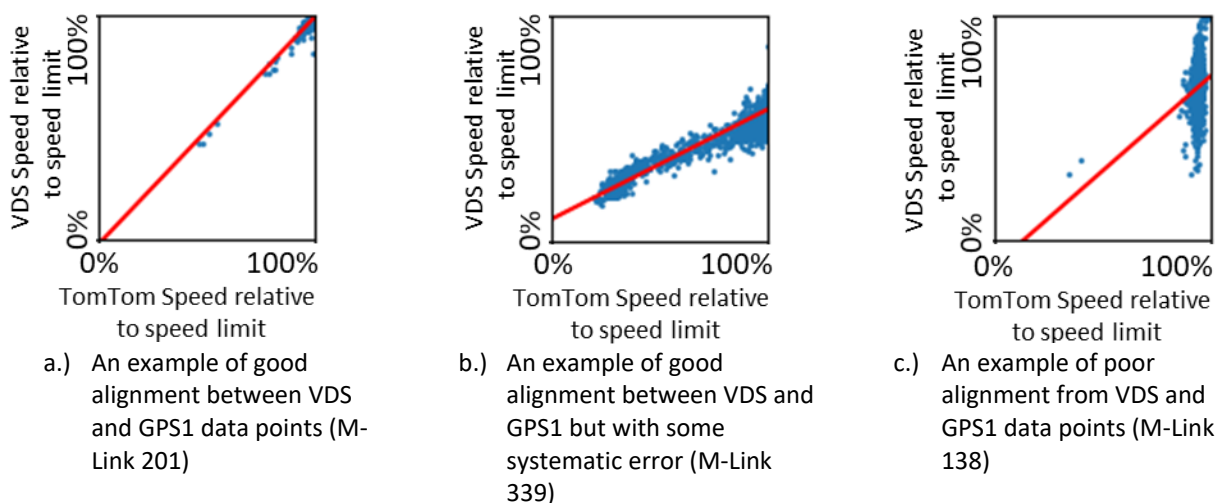


Figure 1 Examples of alignment between VDS and GPS1 data reported on the freeways

Note: The red solid lines in the graphs are regression lines. Ideally the dots need to be close to the line and the line itself needs to be close to 45 degrees, which means the both datasets are reporting the same value at any given time slot.

To address the inconsistency, we applied simple heuristics to remove obvious anomalies, backfill blanks with historical averages and make interpolations when appropriate.

EXOGENOUS VARIABLES

- **Bureau of Meteorology (BOM) Rainfall data:** grid based and translated to M-Links average rainfall.
- **Road Space Bookings:** mapped to M-Links using road name, direction and SLK values.
- **WA School Holidays:** from the Department of Education².
- **Public Holidays:** from the Department of Mines, Industry Regulation and Safety³.

M-LINK CHARACTERISTICS

- **IRIS road link definitions**

The data for the six months from **01/01/2019 to 30/06/2019 (inclusive)** was used for analysis.

² <https://www.education.wa.edu.au/future-term-dates>

³ <https://www.commerce.wa.gov.au/labour-relations/public-holidays-western-australia>

The BOM rainfall data and GPS1 data needed to be mapped from grid references to M-Links, which was performed by Main Roads.

2.3 DATA ANALYSIS PIPELINE

This project required numerous rounds of data cleaning, manipulation, modelling and experiments. To ensure these processes were reliable and reproducible, a flexible data analysis pipeline was developed. It was designed to facilitate repeatable data processing, experimental results, and models and enable any results to be traced all the way back to raw data (through all data transformation, modelling, experimentation and presentation steps). It also allows for a high degree of reusability and flexibility: data pre-processing, data transformations, modelling and experimentation processes can be reused. This analysis infrastructure allows for rapid incorporation of any new datasets that become available in the future.

The analysis pipeline is summarised in Figure 2. It allows for feature sets (collections of variables) to be combined with different models for experimental evaluation, the results of which are shown in the presentation layer. The pipeline will be provided to Main Roads.

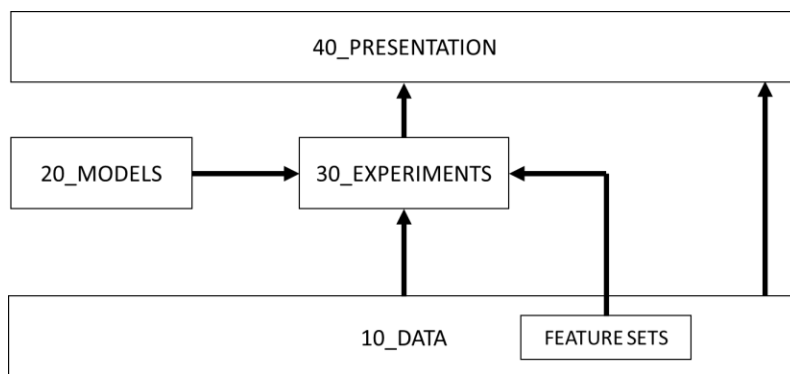


Figure 2 Architecture of the data analysis pipeline

The data set (January 2019 to June 2019) was split into training (75%, approximately four months January-April) and test (25%, approximately May-June). The models for each M-Link were trained on the training partition and evaluated on the test partition, and the results for all M-Links averaged.

Six models were developed for comparison:

Table 2 Six models evaluated in the project

Model	Description	Type
Zero	Always predict the mean speed for the M-Link	Naïve
Time/day mean	Predict the mean speed for that day of week and time of day.	Naïve
Null	Always predict the last available ModelSpeed value (lag 1).	Naïve
RF spd-lag6	A Random Forest regressor ($n_{\text{trees}}=100$, $n_{\text{var}}=6$, $m_{\text{try}}=6$).	Machine Learning
RF all	A Random Forest regressor with all included features ($n_{\text{trees}}=100$, $n_{\text{var}}=26$, $m_{\text{try}}=26$).	Machine Learning
RF featsel	Two stages: (a) Feature selection using Conditional Inference Forest to maximise validation performance (on final 25% of training partition); (b) A final model Random Forest regressor ($n_{\text{trees}}=100$, $m_{\text{try}}=n_{\text{var}}$) with selected features on entire training set.	Machine Learning

The first three models (zero, null, and time/day mean) are so-called ‘naïve’ models for benchmarking purposes. Random Forest (RF) is a classic machine learning model and was used for its robustness and high performance.

For evaluation of the model predictive performance, Symmetric Mean Absolute Percentage Error (**sMAPE**) was used given its popularity in time series analysis:

$$sMAPE = \frac{|pred-actual|}{\frac{|pred|+|actual|}{2}}$$

4

MODELLING RESULTS

4.1

PERFORMANCE RESULTS

The predictive accuracy of the models is summarised in Figure 3, Figure 4 and Figure 5. The first two represent the model performance under the most congested conditions, while the last one represents the average performance across the whole day and in all travel directions.

It is clear from the plots that RF with all features, and RF with feature selection, perform best at all time periods. In fact the feature selected version is very slightly superior across all prediction horizons and subsets, so the sparser feature-selected models certainly do not reduce performance. We have also verified that our RF models outperform ARIMA, which is the most popular time series prediction model. The Auto-ARIMA model performs very similarly to the RF (lag 6) model, but is similarly outperformed by the full RF models (RF all and RF featsel).

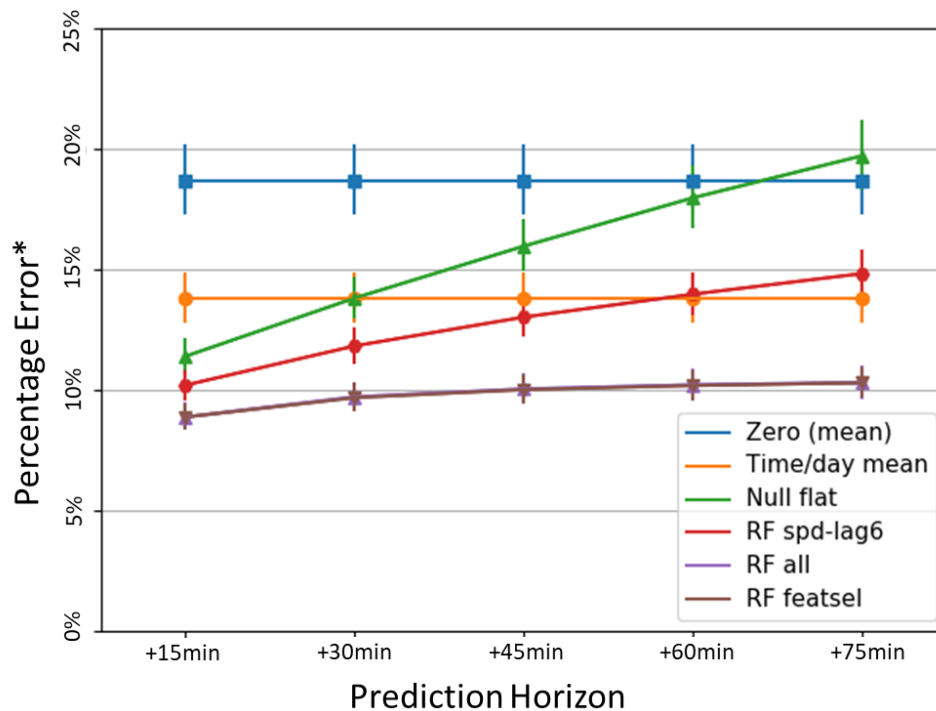


Figure 3 Model Accuracy: AM peaks for Inbound M-Links (RF all is not visible since it overlaps with RF featsel. See Table 2 for more explanation of models used to produce the curves)

**Symmetric Mean Absolute Percentage Error is used to calculate percentage error (see Section 3 for definition)*

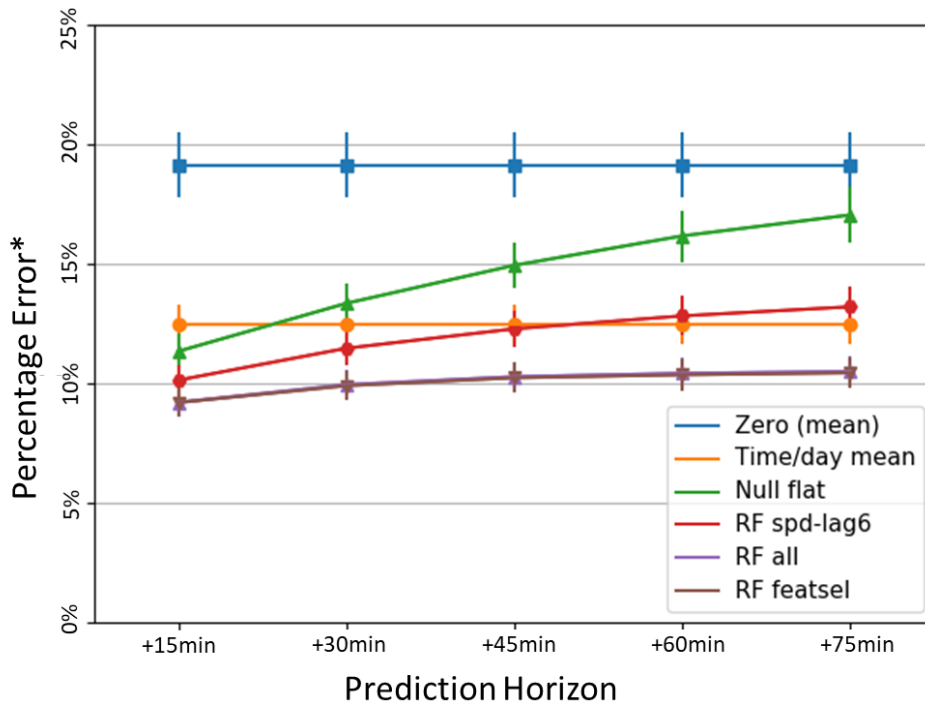


Figure 4 Model Accuracy: PM peaks for Outbound M-Links (RF all is not visible since it overlaps with RF featsel. See Table 2 for more explanation of models used to produce the curves)

*Symmetric Mean Absolute Percentage Error is used to calculate percentage error (see Section 3 for definition)

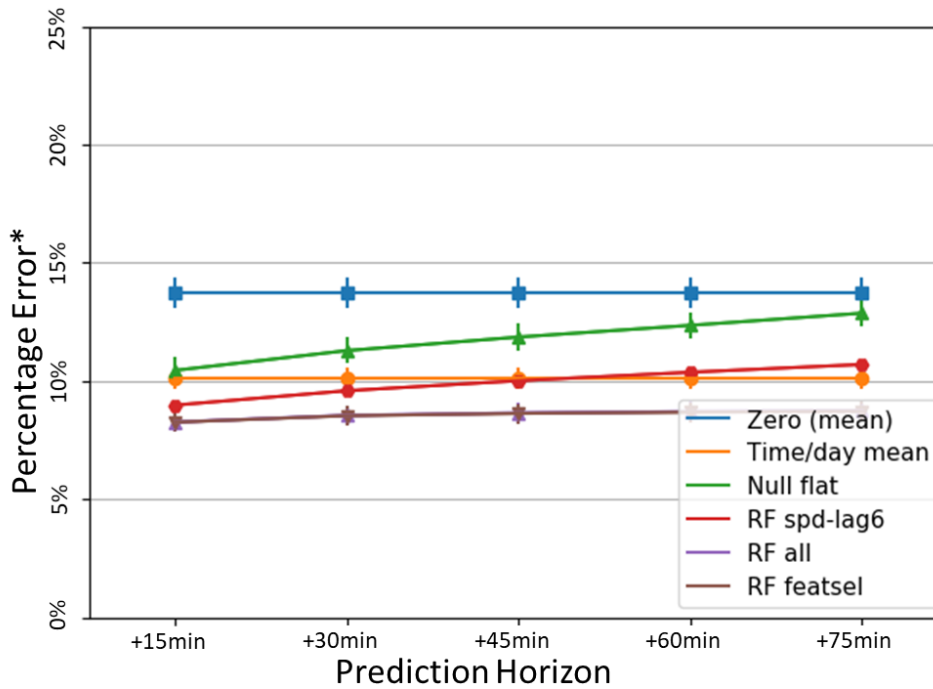


Figure 5 Model Accuracy: All periods for All M-Links (RF all is not visible since it overlaps with RF featsel. See Table 2 for more explanation of models used to produce the curves)

*Symmetric Mean Absolute Percentage Error is used to calculate percentage error (see Section 3 for definition)

4.2 FEATURE SELECTION RESULTS

Not all features/variables are equally important. Feature selection is a process of selecting those that contribute the most to the model performance. The 'RF featsel' model presented in Table 2 and the rest of Section 4.1 was feature selected and it shows slightly better performance than the 'RF all' version that uses all features, although the differences are not visible in the figures. The selection rates for the features is summarised in Table 3.

Table 3 Feature Selection Rates (for RF featsel model, in order of decreasing selection rate)⁴

Variable	Mean Selection Rate
TimeOfDayIndex	97.1%
Speed_lag_1	96.9%
Density_lag_1	96.5%
Density_lag_2	95.0%
Speed_lag_2	93.6%
Speed_lag_1wk	92.2%
Density_lag_3	91.6%
Speed_lag_1day	91.3%
Speed_lag_3	87.5%
Density_lag_4	85.9%
DayOfWeek	82.3%
Density_lag_5	81.1%
Density_lag_6	79.8%
Speed_lag_4	79.7%
Speed_lag_5	72.1%
Speed_lag_6	66.4%
Holiday_long_wkd	51.6%
Holiday_school	50.6%
Holiday_public	50.6%
Road_Space_Booking	31.3%
Rainfall_3hr_lag_1	27.2%
Rainfall_6hr_lag_1	27.2%
Rainfall_1hr_lag_1	24.9%
Rainfall_15mins_lag_1	22.2%

Generally, it can be seen that:

- The most important features generally are the lagged Speed and Density values, and the selection rates for these decay as the lag increases (i.e. the values are more distant in time from the value being predicted).
- TimeOfDayIndex is the most significant time based feature, followed by DayOfWeek.
- Holiday variables (both school and public) are selected as significant for about 50% of the M-Links.

⁴ Note that 'lag' refers to how many 15 minute periods this data is *prior to the first prediction period*. Thus 'lag 1' refers to the 15 minute period immediately before the first predicted period, and 'lag 2' refers to the 15 minute period 15-30 minutes prior to the first predicted period and so on.

- RoadSpaceBooking and Rainfall are selected for 30% or less M-Links. This result is more likely due to the sparsity of these features in the training set (rain is the exception rather than the norm).

5 CONCLUSIONS AND FUTURE OPPORTUNITIES

This project successfully applied machine learning techniques to short-term prediction of average speed for road sections. The proposed Random Forest (RF) models can extract hidden value from the Main Roads' datasets and have been shown to be robust and perform well against the benchmarks (naïve and traditional approaches). Their advantage becomes obvious with an increasing prediction horizon (how far ahead the models predict). Moreover, accuracy does not substantially decrease with an increasing prediction horizon: during the AM and PM peaks, predicting 15 minutes ahead will produce an average percentage error of about 9% while for 75 minutes it is just above 10%.

A data analysis pipeline was developed to ensure all models and results are traceable and the reproducible. It is also reusable and flexible for future development, especially if new data becomes available.

At this stage, the models work from offline historical data. Currently, the main obstacle to real-time application is the funding of livestream data acquisition and necessary data infrastructure. When this occurs, the pipeline could be used by Main Roads as a basis for a prediction engine that can support Network Operations in real-time. Long term, we envisage a traffic '*now-casting*' decision support system for network operations. Such a short-term early warning system aligns well with Main Roads Network Operations' vision of '*predict in 20 (min), act in 5 (min), change the future*', and would help traffic operators make *pre-emptive* decisions to prevent traffic passing the 'point of no return', rather than *reactive* actions after the fact.

A valuable extension to this project would be anomaly detection that focuses on non-recurrent events, not just stopped vehicles or crashes but also in terms of traffic performance. The project team understands that there are several existing systems with some capacities in this area. However, more sophisticated analysis could be done by using the fused data from multiple sources.

The models could also be further improved by:

- Inclusion of more data, particularly an entire year of seasonality.
- Improving data quality, such as further calibrating VDS detectors.
- Fusing multiple speed data sources (using the model speed fusion approach developed in a sister project).
- Development of a deep learning model.

REFERENCES

Bañbura, M, Giannone, D, Modugno, M & Reichlin, L 2013, 'Now-casting and the real-time data flow', In *Handbook of Economic Forecasting*, Vol. 2, pp. 195-237, Elsevier.

Laney, D 2001, '3D data management: Controlling data volume, velocity and variety', *META group Research Note*, 6(70), pp.1.

Western Australian Auditor General 2015, *Main Roads Projects to Address Traffic Congestion*, Office of the Auditor General Western Australia, Perth, Western Australia. Available from: https://audit.wa.gov.au/wp-content/uploads/2015/03/report2015_02-TrafficCongestion.pdf