Traffic volume prediction model

Report for Roads and Maritime Services and other NSW Government stakeholders

1 October 2017

Traffic volume prediction model

Executive summary

- Traffic volume prediction is challenging, but offers a range of benefits to Roads and Maritime Services, including financial savings and improved outcomes in relation to road planning, design and maintenance decisions..
- To build a traffic volume prediction model, Transport for NSW data in relation to daily traffic counts at various traffic stations was combined with weather data (temperature and rainfall) from the Bureau of Meteorology, and demographic data from the Australian Bureau of Statistics.
- This combined dataset was used to predict the daily traffic count, both at existing locations and hypothetical additional locations.
- A Quasipoisson regression model was chosen due to its ability to account for the overdispersion apparent in the data.
- Several methods were attempted to identify the variables which most affected predictive accuracy; the most useful was forward selection with reference to average mean squared error on the testing subset.
- The variables which most significantly explain traffic volume are the distance to the CBD, road function, RMS region, day of week, public holiday/school holiday status and daily rainfall.
- A number of issues were identified in the data itself; to the extent which these are able to be resolved by Roads and Maritime Services, recommendations to that end have been made below.

Recommendations

1. In developing road maintenance schedules and other operational or strategic decisions, Roads and Maritime Services should ensure strategies are appropriately context-specific, noting the significance of location variables to traffic volume variation.

2. To improve the accuracy of traffic volume prediction, Roads and Maritime Services should update its road classification categories, by identifying additional road classification measures. These may include:
    a. a more granular count of lane numbers, lane purposes;
    b. the nature of the route and surrounding area including differentiation between divided, segregated roadways and roads with significant local use;
    c. the nature of roads feeding into or from the route,

3. Subject to the implementation of Recommendation 3, Roads and Maritime Services should focus model application in the first instance on Primary roads, which provide the highest level of predictive accuracy using existing road classifications.

Business understanding

The objective of this project was to develop a model which accurately predicts traffic volume, in a wide range of contexts, with reference to historical traffic data and diverse environmental variables.

Specifically, the model sought to answer the following questions:

1. What is the predicted data volume for any given combination of:
    a. datetime
    b. location, including associated demographics, and
    c. weather conditions?
2. Which NSW locations are most sensitive, as measured by traffic volume variance, to weather conditions?
3. Which variables most affect traffic volume?
4. To what extent are traffic volume predictions transferrable to different road environments, both nearby and further afield?

The uses of such a model for NSW Roads and Maritime Services are numerous, including:

- Reduced need to install expensive traffic counter stations.
- Improved road planning, design and maintenance, through more accurate estimation of local traffic volumes and loads.

In addition, such a model would offer the following benefit to the NSW Department of Planning and Environment:

- Improved urban planning and pollution monitoring, through better understanding of local traffic patterns in areas where permanent traffic counter stations are not installed.

Finally, the model would also likely be of use to commuters more generally, in providing an additional data point for route selection optimisation.

The CRISP-DM methodology, an industry-standard approach to planning and undertaking data analysis, was applied. The phases of that methodology form the basis of this report. RStudio (v1.0.136), running R (v3.3.3), was used for data analysis.

Data understanding

The following datasets were collected and analysed in the design phase:

TABLE 1 - Datasets

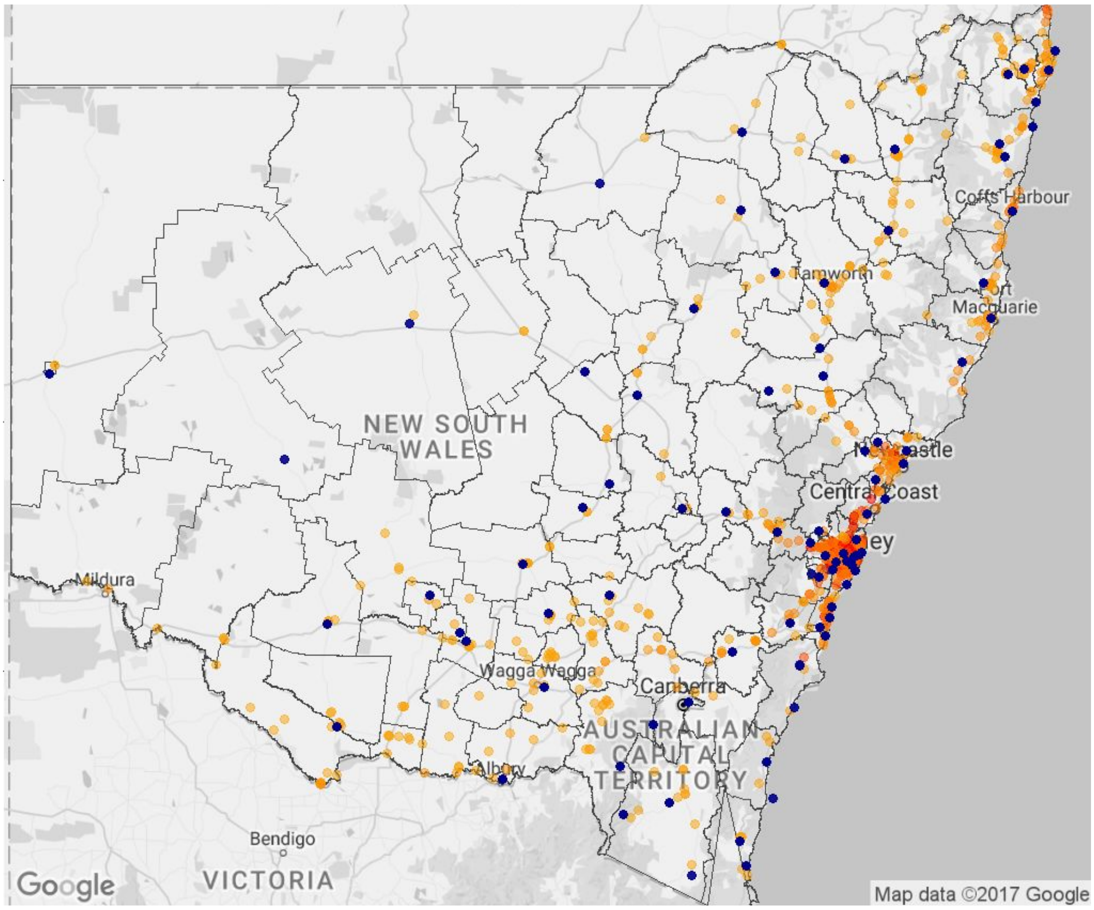
Data	Source	Dates	Description
Traffic Counter Station	OpenData for Transport NSW API	2006-2016	1,818 traffic counter stations from across NSW, consisting of 1,251 sample stations and 567 permanent stations. Variables include: ● Station ID ● Road name (in three different formats) ● Road name base and type ● Intersection and distance to

			intersection ● Road number ● Link number ● Mab way type and number ● Mab identifier ● Road functional hierarchy ● Road on type (ground, bridge, tunnel) ● Lane count ● Road classification type and admin ● RMS region ● LGA ● Suburb ● Postcode ● Device type ● Heavy vehicle checking station status ● Permanent station status ● Vehicle classifier status ● Lambert easting ● Lambert northing ● Latitude ● Longitude ● Direction of traffic flow ● Data quality rating
Traffic Count	OpenData for Transport NSW API	2006-2016	Hourly traffic counts from each traffic counter station around NSW.
Weather	Bureau of Meteorology	2000-2016	Weather conditions for the 115 weather stations around NSW closest to traffic counter stations, in 30 minute increments. Variables include: ● Station number ● Station name ● Locality ● Latitude ● Longitude ● Date (ddmmyyyy format) ● Time (hh:mm format) ● Precipitation in previous 10 minutes (mm) ● Quality of precipitation ● Air temperature (degrees celsius) ● Wind speed (km/hr) ● Wind speed quality ● Wind direction (degrees) ● Wind direction quality ● Year
Demographic	ABS SDMX	2011-2016	ABS data for NSW, by LGA, extracted using the *RSDMX* package . Variables include:

			● LGA ● Population density ● % population working age ● % population school age ● Light vehicle density ● Heavy vehicle density ● Median income
Distance	GoogleMaps API	Real Time	Distance by road calculated between each traffic counter station and Sydney CBD.

Figure 1 below shows the geographical distribution of traffic counters (in orange) and weather stations (in blue) across NSW.

FIGURE 1 - Geographic representation of traffic counter stations and weather stations

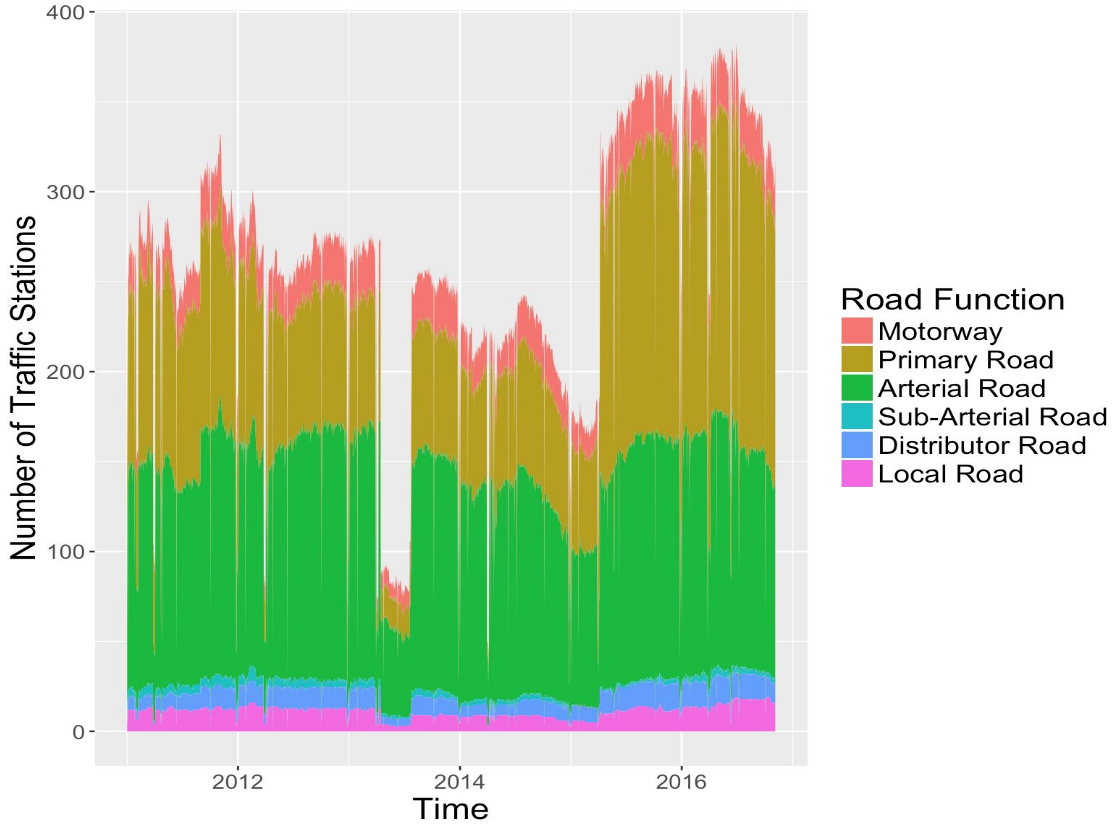


Data issues

Data quality - significant changes in station volume and road type

The traffic data suggests that, in April 2015, 199 permanent traffic counter stations were added to the network in NSW, predominantly on primary roads. This significant change is represented below in Figure 2. The noticeable drop in 2013 is discussed in 'missing data' later in this report..

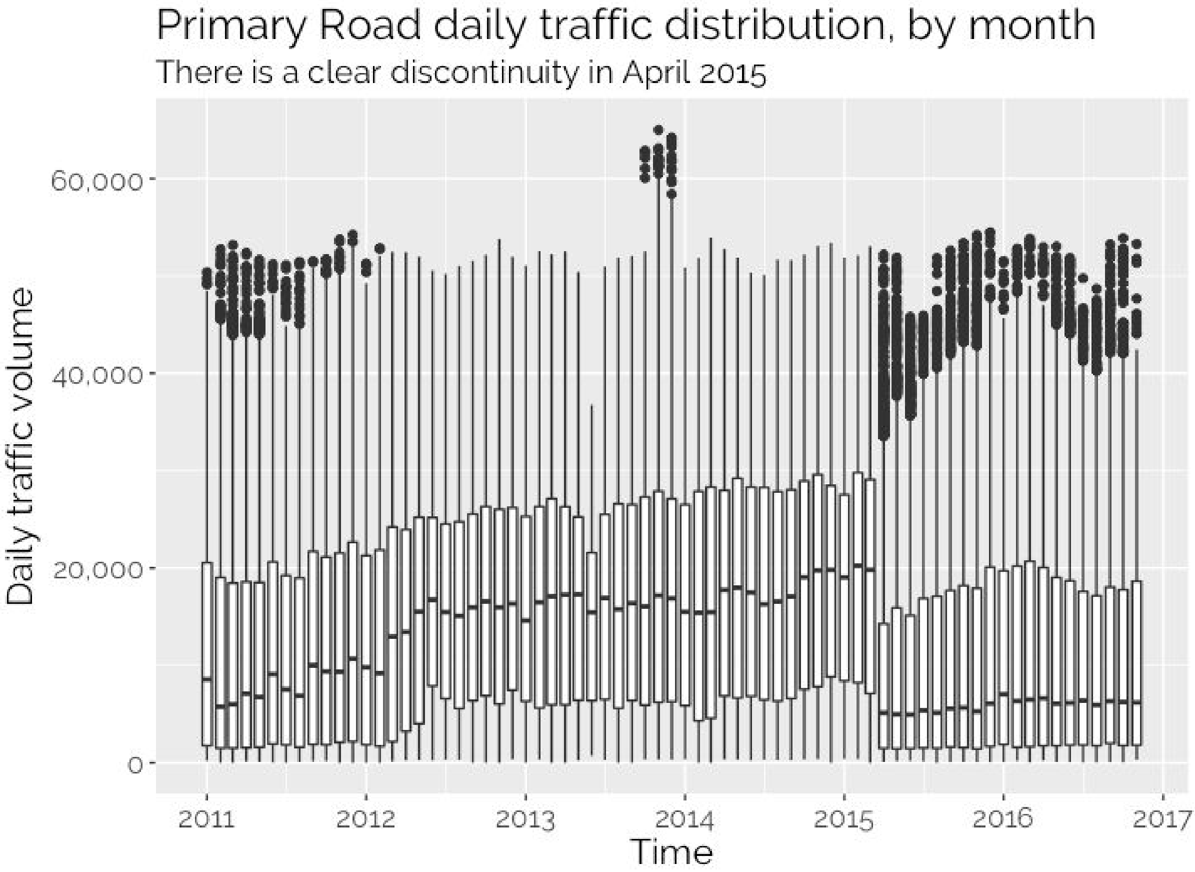
FIGURE 2 - Number of RMS traffic counter stations over time



The introduction of these additional traffic counter stations noticeably impacts the summary statistics for primary roads, with the median daily count, summarised by month, immediately dropping more than 10,000 cars. This adjustment is clearly illustrated in Figure 3 below and suggests an influx of counter stations on roads with relatively smaller traffic volumes. We anticipate that this will have a substantial impact on the ability of the model to allow for the growth in traffic count over time.

This issue was addressed in our modelling to date by working only with data from April 2015 to December 2016. We may be able to extend the modelling to make use of the earlier data, but this would require additional discussions with RMS in order to understand the implications, and further iterations of model training and testing.

FIGURE 3



Data quality - imbalanced data classes

The categories used by Transport for NSW to classify roads are problematic for the current analysis due to the highly imbalanced data classes, represented below in Table 2.

TABLE 2 - Examples of imbalanced data classes

Variable	Classification	% Total variable observations
Road classification	'State'	90.1
Lane count	'Two or more'	98.4
RMS region	'Sydney'	60.9
Road type	'Arterial'	47.1
	'Primary'	34.9

As a result of this classification issue, roads with markedly different characteristics and historical traffic patterns are classified identically; potentially reducing the level of meaningful analysis which can be undertaken of model output.

To illustrate this, the following two traffic stations currently share identical classification in the RMS dataset, but there are some very clear differences between them in reality:

TABLE 3 - Comparison of possibly relevant attributes between two identically classified traffic stations

Attribute	**City West Link** east of James St	**Enmore Road** east of Bailey St
Lane count	3	1-2 depending on clearway
Divided road	yes	no
Side streets	limited	frequent
Dedicated turn lanes	yes	no
Property frontages	0% - noise barriers	100% - retail, restaurants
Street parking	no	yes
Bus stops	no	yes
Inbound traffic feeds from (largest source)	Parramatta Road (3 lanes)	Stanmore Road (2 lanes)
Inbound traffic feeds into	Anzac Bridge (4 lanes)	King St (1 lane plus peak clearway)

This issue has not yet been addressed. We propose that additional road type classifiers be added in the next iteration of the model. We would work with RMS subject matter experts to identify further potential classifiers (those in the above Table are possible candidates) and then add the additional data either sourced from Google or another map service or by visual inspection via Street View.

Data quality - missing data

The dataset collected from the Australian Bureau of Statistics was missing population density data for 2016, for all Local Government Areas. To resolve the issue of this missing data, we assumed no change in those variables between 2015 and 2016. Given the slow change in such socio-economic factors, we believe this broad-brush assumption will not introduce any significant error.

The dataset from RMS also contained an apparent gap in 2013, as shown in Figures 2 & 3. As stated previously, by only working with data from 2015 onwards we have avoided any further issues related to this.

Data quality - outliers

Two outliers were noticeable in the rainfall data for June 2016 collected from the Bureau of Meteorology. These data points were confirmed to be accurate, representing two days of record high rainfall including the June 2016 East Coast Low.

Data acquisition and storage

Significant inconsistencies were noted in API output, apparently related to the time and location from which the database was accessed, as well as the computer and internet connection being used. To resolve this issue and ensure that all team members were analysing a common dataset, key dataframes were periodically updated, saved in *.RDS* format, and shared on Github.

Traffic counter station data, and the associated traffic counts, were sourced from the Transport for NSW Open Data hub. The data was split into three datasets; traffic counter stations, traffic counts from permanent traffic stations and traffic counts from sample traffic stations. Each of these datasets was read into R using the *httr* package, with file size quickly becoming an issue. The Open Data API restricted the amount of data which could be obtained in any single call on the API by limiting the time available to access the API. To work around this restriction, a loop function was written to request each month of count data separately then compile them once in R.

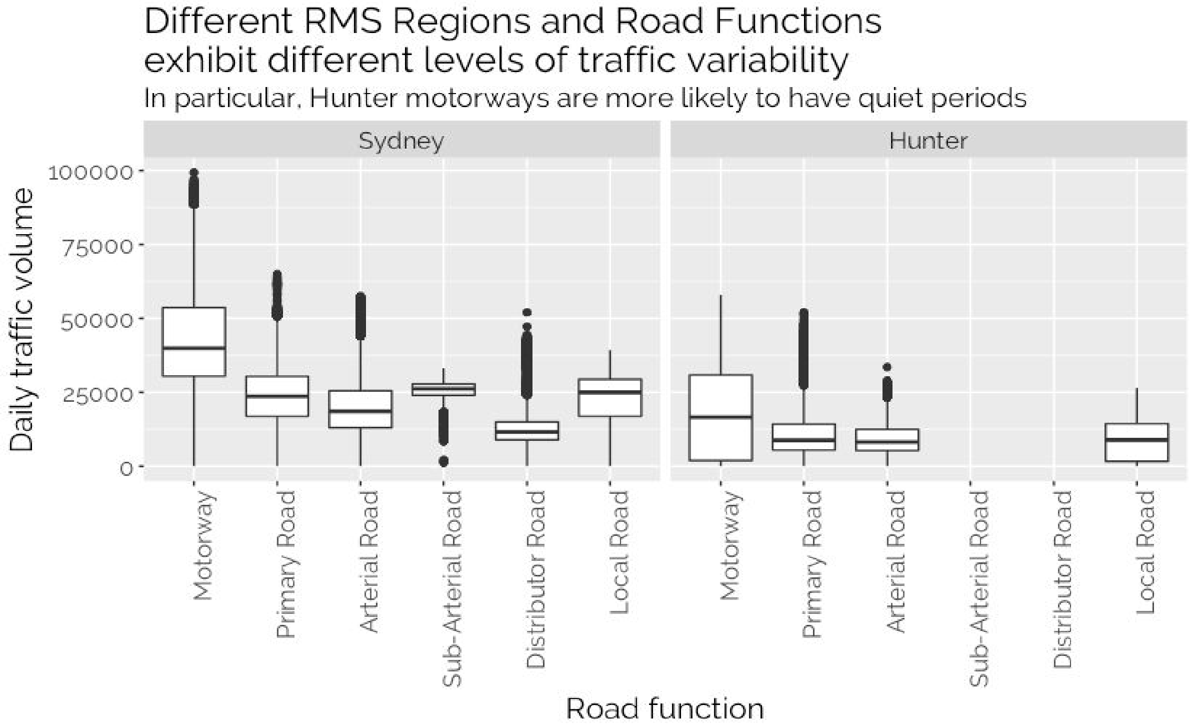
To properly account for Sydney CBD as a population centre and commercial hub, a variable was introduced which measured the distance by road from each traffic counter station to the CBD. This was computed by running each of the coordinates for the traffic counter stations through a function calling the Googlemaps Distance API. Once again, this API call resulted in throttling issues, which were resolved by splitting the request into separate rounds.

The sheer size of the BOM weather dataset (around 28 million records) also required extraction of a relevant subset into Google Drive in *.RDS* format.

Other findings of exploratory analysis

Research undertaken in developing the project proposal indicated that traffic volume is highly context-specific, and findings are difficult to translate between locations (Yang et al. 2017). Exploratory analysis of the dataset collected confirmed this finding. As demonstrated below in Figure 4, the frequency of specific traffic volume observations varies substantially between similar road types in different geographic regions.

FIGURE 4



This finding is supported by simple correlation analysis of key variables against the output variable of traffic volume. The coefficients of each of these variables are summarised below in Table 4. As Table 4 demonstrates, in addition to RMS Region, distance to CBD, population density, vehicle registration density and the percentage working age population are all more closely correlated to traffic volume than the other variables..

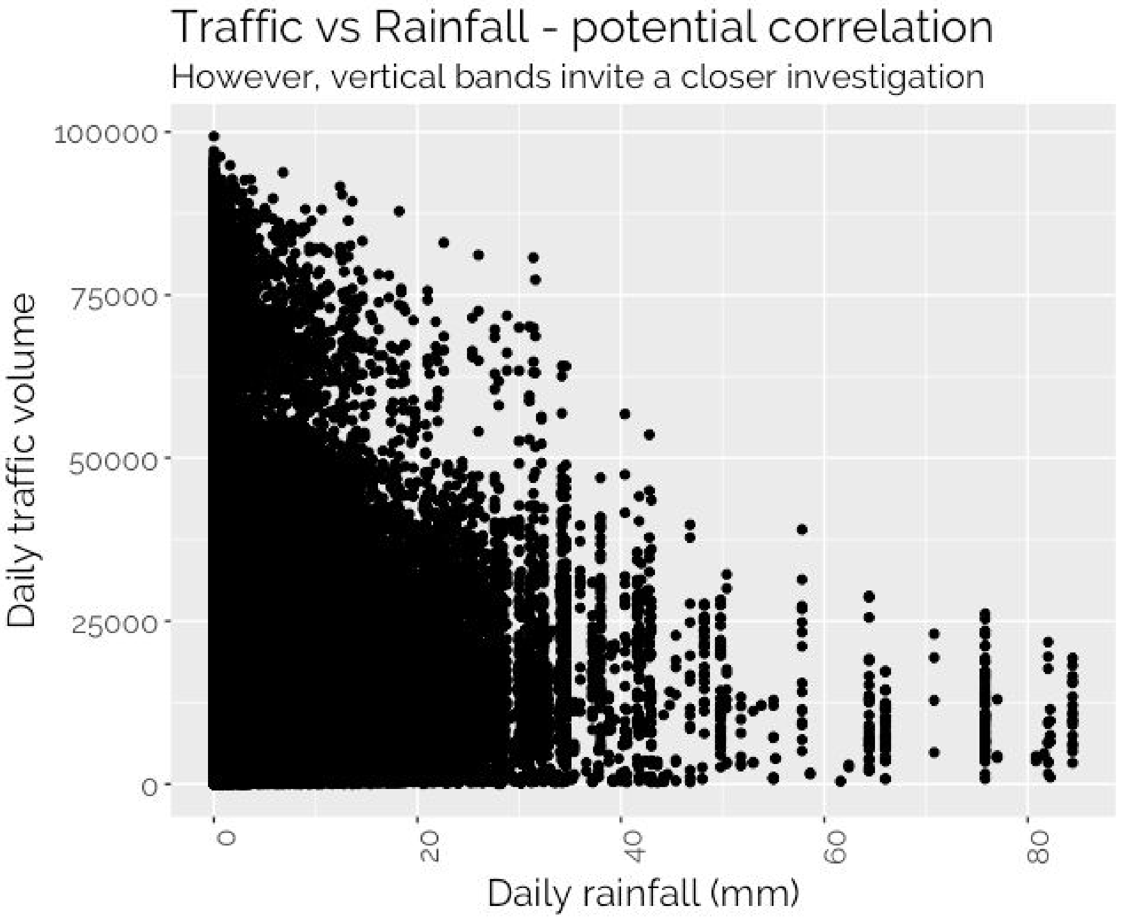
TABLE 4 - Correlation coefficients

Potential explanatory variable	**Correlation with daily traffic volume** (Pearson correlation coefficient)
RMS region	-0.52
LGA population density	+0.49
Distance to CBD	-0.48
Population % of working age	+0.39
Road function	-0.15

Population % of school age	-0.19
Day of week	-0.05
Public holiday	-0.04
School holiday	-0.02
Month of the year	+0.02
Daily rainfall	-0.002

Table 4 suggests that both the percentage school age population, and rainfall, are not closely correlated to traffic volume. However, initial visual inspection suggested a linear relationship between traffic volume and rainfall, as demonstrated below in Figure 5.

FIGURE 5



Upon further examination, the apparent correlation above in Figure 5 was found to be misleading. As shown in Figure 6 below, most of the vertical bands to the right-hand side of the chart represent observations from different traffic stations during a very small number of days of very high rainfall. As mentioned above in our discussion of Data Issues, this included Sunday 5 June 2016, the day of the severe and damaging East Coast Low.

Even with this improved understanding, there still appears to be a possible relationship although it is difficult to be sure from a chart owing to the sheer number of data points. Although there appears to be a possible drop in the frequency of high traffic volume observations as rainfall increases, this could be related to different rainfall patterns in different geographic locations. The true contribution of rainfall data to a predictive multi-variable model is discussed later in the report.

FIGURE 6

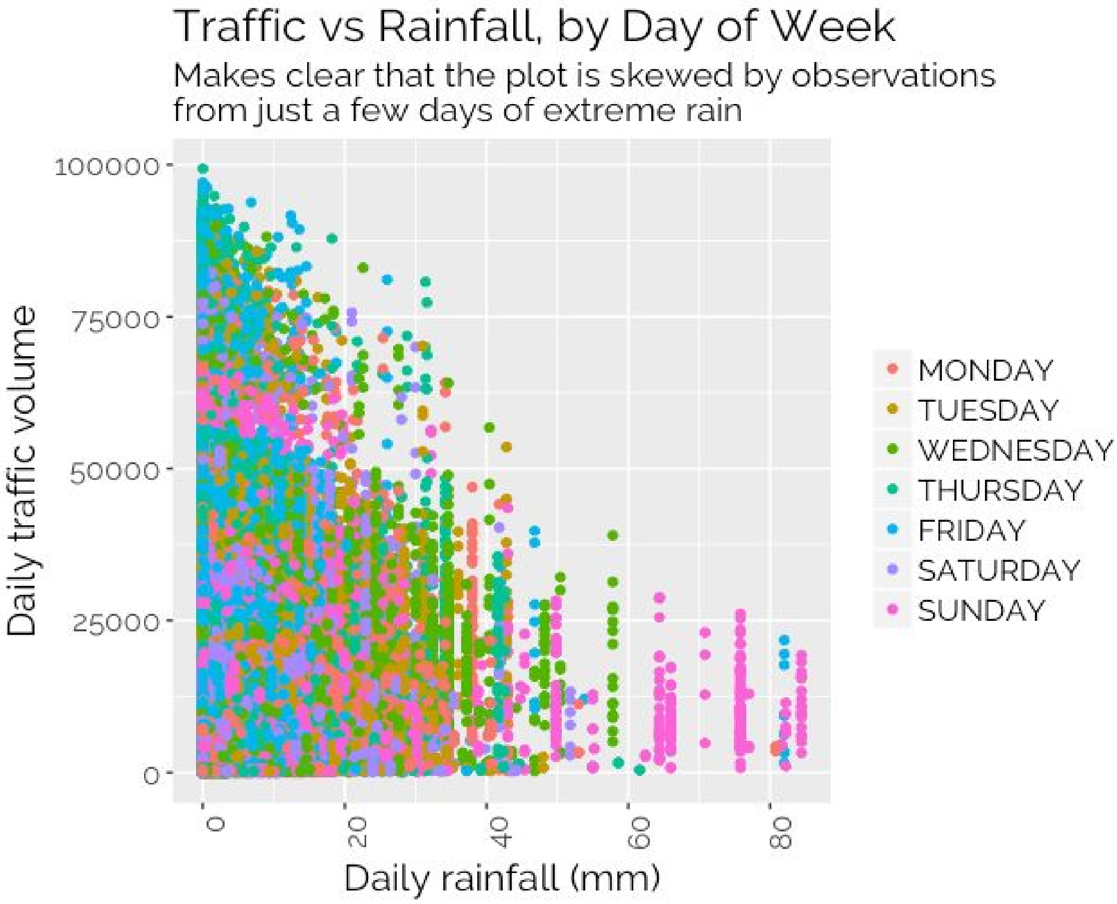
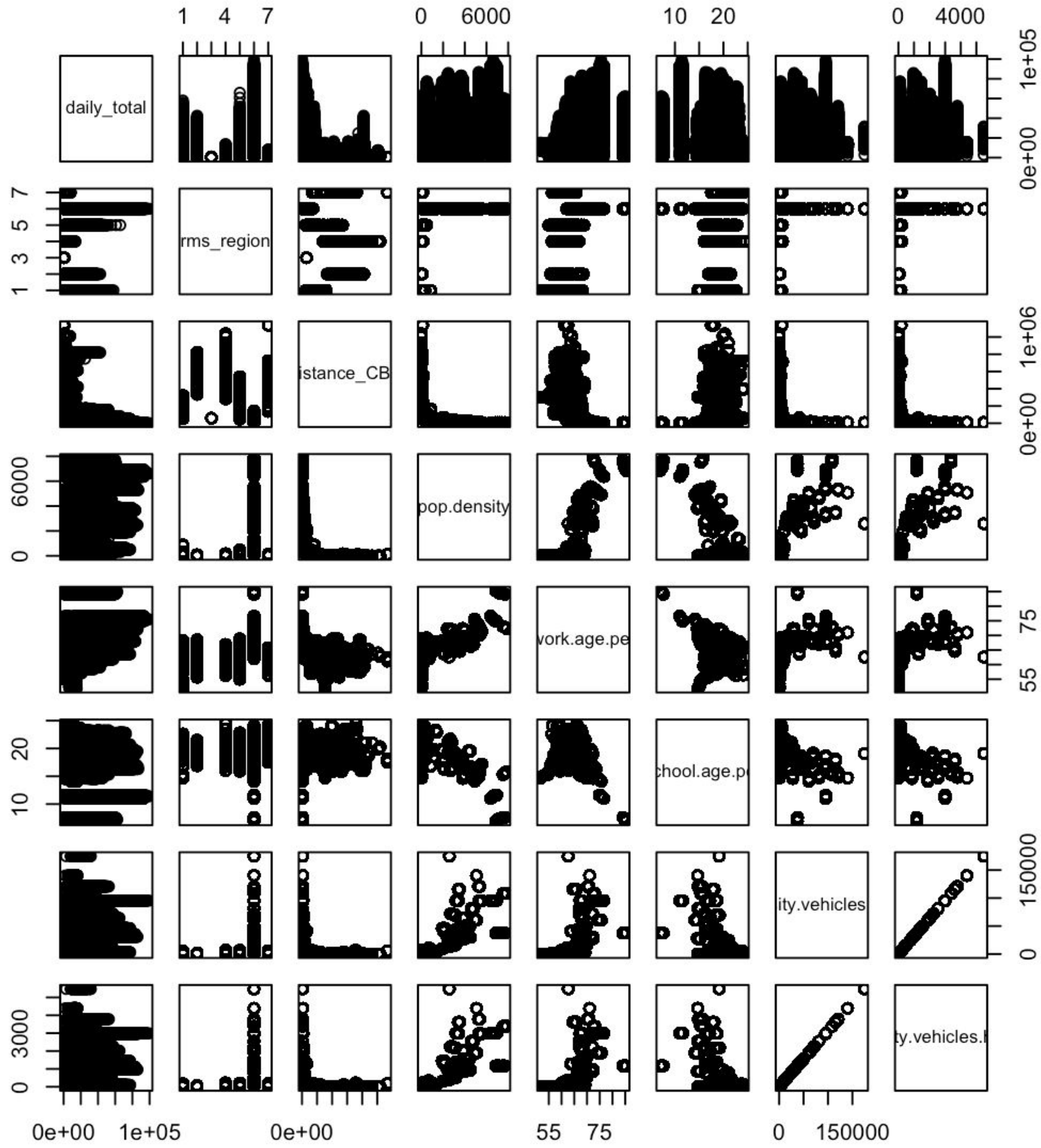


Table 4 also suggests several of the variables with apparent high correlation. Examining these further, in Figure 7, we can see that there is in fact a collinear relationship between many of them. For example, population density and vehicle density are related. Distance to CBD and population density are also related (although it is harder to see in Figure 7 due to the long tail of the plot).

As will be discussed later in the report, the model development shows that these additional variables do not improve model accuracy. Conversely, some of the variables in Table 4 with apparent low correlations to traffic volume do in fact contribute to the predictive power of a multi-variable model

FIGURE 7 - Pairwise scatter plots of ABS variables



Data preparation

Data selection

A number of variables contained in each original dataset were not included in our analysis. In part, this was motivated simply by necessity - the size of the full dataset required more computing power than was available without accessing web services. The decision on which variables to exclude was made as a group on the basis of intuitive expectations of predictive benefit.

From the RMS traffic dataset, we excluded a number of duplicative variables (for example, the multiple variables describing road names, record and station identification numbers, and direction of traffic flow. Similarly, variables considered duplicative or superfluous were excluded from the BOM weather dataset including date and time fields and records in relation to, for example, humidity, vapour levels and cloud. In total, between the RMS and BOM data sets, 33 of the original 89 variables were retained.

Finally, we did not include ABS statistics relating to businesses in the area, language and community group representation and measures of socioeconomic status. We believed that any predictive power of these variables would be adequately allowed for in the more directly-related variables of population and vehicle density, and proximity to the Sydney CBD.

Data cleaning

The following steps were undertaken to ensure variables were in an appropriate and useful format for the analysis at hand:

- Date variables were standardised to ddmmyyyy format.
- School holiday and public holiday variables were converted from a binary format (0/1) to equivalent Boolean values (FALSE/TRUE), to aid interpretability.
- Unnecessary blank space was removed from the road_functional_hierarchy variable using the *trim* function, to aid interpretability of axis label plots in visualisations.
- Classes for continuous and categorical variables which had been misclassified were amended accordingly.
- Duplicates which were introduced in data merging were removed.
- Hourly rainfall data was aggregated daily.

Data merging

The separate datasets collected were merged into a single final dataframe in the following ways:

- Hourly traffic volume counts were combined with detailed traffic counter station information using the unique station key.
- Distances from Sydney CBD to traffic counter stations were merged via the Google API.
- Demographic variables were matched to traffic counter stations by LGA codes.
- Weather stations were linked to the nearest traffic counter stations by location coordinates.

Modelling

Twenty percent of the final dataset was quarantined at the outset for the purpose of model evaluation - referred to in the report as our 'Holdout' dataset. The remaining 80 percent was split into training and testing subsets, of 70 percent and 30 percent respectively, to assist with model selection.

Noting that the target variable is a count (that is, the number of cars past a given point each day) it was assumed to better fit a Poisson distribution than a simple linear regression. A Poisson regression is appropriate where the following assumptions are met:

1. Counts are either zero or a positive integer
2. Events occur independently of one another
3. The rate at which events occur is constant
4. Two events cannot occur at exactly the same instant
5. Variance is equal to mean (dispersion of 1)

The dataset collected meets each of these assumptions, with the exception of number 5. Our data exhibits overdispersion as the variance in counts, driven by the substantial differences between road types and locations, is much higher than the mean.

Accordingly, other models were considered which allow for an overdispersed data set. The first alternative considered was the Quasipoisson model which fits the data in the same fashion but adjusts the standard error to account for the overdispersion in the data. Another alternative is the negative binomial model, which assumes a distribution which resembles an overdispersed Poisson distribution.

While the Quasipoisson model results in slightly more accurate traffic volume predictions on the testing subset, it does not follow a probability distribution and we are therefore not able to determine how well the model fits the data with reference to a standard measure such as AIC. This makes it particularly difficult to compare models for model selection purposes.

Due to this characteristic of the Quasipoisson model, model selection (including variable selection) has been undertaken by comparing the mean squared error on the testing subset.

Evaluation

Feature selection and justification

Forward selection was selected as the preferred method for feature selection to maximise the interpretability of the final model, noting the number of available variables and the broad range of contexts in which the model is intended to apply.

Nearly all of the variables analysed in model iteration obtained an extremely low p-value (of effectively zero), suggesting a high significance to model output. This made the process of variable selection difficult, as variables were near impossible to distinguish on this measure alone. To resolve this issue, the testing subset was used to predict values on each fitted model, with mean squared error used as the test statistic upon which variables were prioritised.

Both the Quasipoisson model and the negative binomial alternative resulted in optimised model accuracy with the following variables included. The order was also consistent across both models:

- Distance to the CBD
- Road function
- RMS region
- Day of week
- Public holiday status
- Daily rainfall total
- School holiday status

The summary output is provided below in Table 5. As discussed above, the p-value of all model coefficients is effectively zero.

TABLE 5 - Model features

Model variable	**Coefficient Estimate**	**Standard Error**	**T value**	**P Value (to 5 significant digits)**
Intercept	9.42415	0.00823	1145.25593	0
Distance to CBD	0	0	-97.35728	0
Road functional hierarchy -> Distributor Road	-0.28592	0.00775	-36.9149	0
Road functional hierarchy -> Local Road	0.20154	0.00765	26.34265	0
Road functional hierarchy -> Motorway	0.73674	0.00462	159.54551	0
Road functional hierarchy -> Primary Road	0.12165	0.00402	30.26565	0

Road functional hierarchy -> Sub-Arterial Road	-0.69829	0.03735	-18.6966	0
RMS region -> Northern	0.92343	0.01559	59.2314	0
RMS region -> South West	-0.21983	0.01975	-11.12799	0
RMS region -> Southern	-0.10686	0.00743	-14.39111	0
RMS region -> Sydney	0.51522	0.00651	79.15561	0
RMS region -> Western	-0.8602	0.01754	-49.03612	0
Day of week -> TUESDAY	0.02524	0.00593	4.25887	2.00E-05
Day of week -> WEDNESDAY	0.04179	0.00588	7.10122	0
Day of week -> THURSDAY	0.06751	0.00586	11.51086	0
Day of week -> FRIDAY	0.11351	0.00586	19.38662	0
Day of week -> SATURDAY	-0.03096	0.00603	-5.13247	0
Day of week -> SUNDAY	-0.15824	0.00624	-25.37566	0
Public holiday -> TRUE	-0.24717	0.01598	-15.4717	0
Daily rainfall	-0.0046	0.00047	-9.84096	0
School holiday -> TRUE	-0.01639	0.00374	-4.37547	1.00E-05

Model limitations, assumptions and mitigation strategies

As noted above, one of the key limitations of the final model is the limited time series against which it has been developed. This could easily be solved through the use of a web services platform which could more easily handle the size of the full dataset. inevitably providing for more accurate incorporation and reflection of any seasonal effects. It is also assumed, for the purposes of this analysis, that the time series chosen is adequately reflective of any longitudinal trends, at least insofar as they might affect predictions in the short to medium-term future.

The overdispersion apparent in the raw data is also a significant limitation. This is likely caused by level of variability between the inputs themselves, in addition to the use of aggregation across a number of those inputs. To some extent, this issue is addressed by the use of the Quasipoisson model or Negative Binominal model as opposed to the straight Poisson model. However, the sheer level of overdispersion in the data means that the accuracy of the model will always be somewhat limited.

A number of other assumptions are also inherent in the model. Firstly, it is assumed that no significant predictors of traffic volume have been omitted. Such an omission would obviously significantly reduce the predictive capacity of the model. However, noting the research undertaken at the proposal stage, no really significant predictors are considered to have been overlooked.

Secondly, it is assumed that there are no major upcoming events (whether environmental, political or social) which might drastically affect traffic volumes. A wholesale ban on a particular vehicle or fuel type, for example, would likely have a significant effect. No such effect was identified in the research undertaken and so, as a hypothetical, this assumption was taken as given.

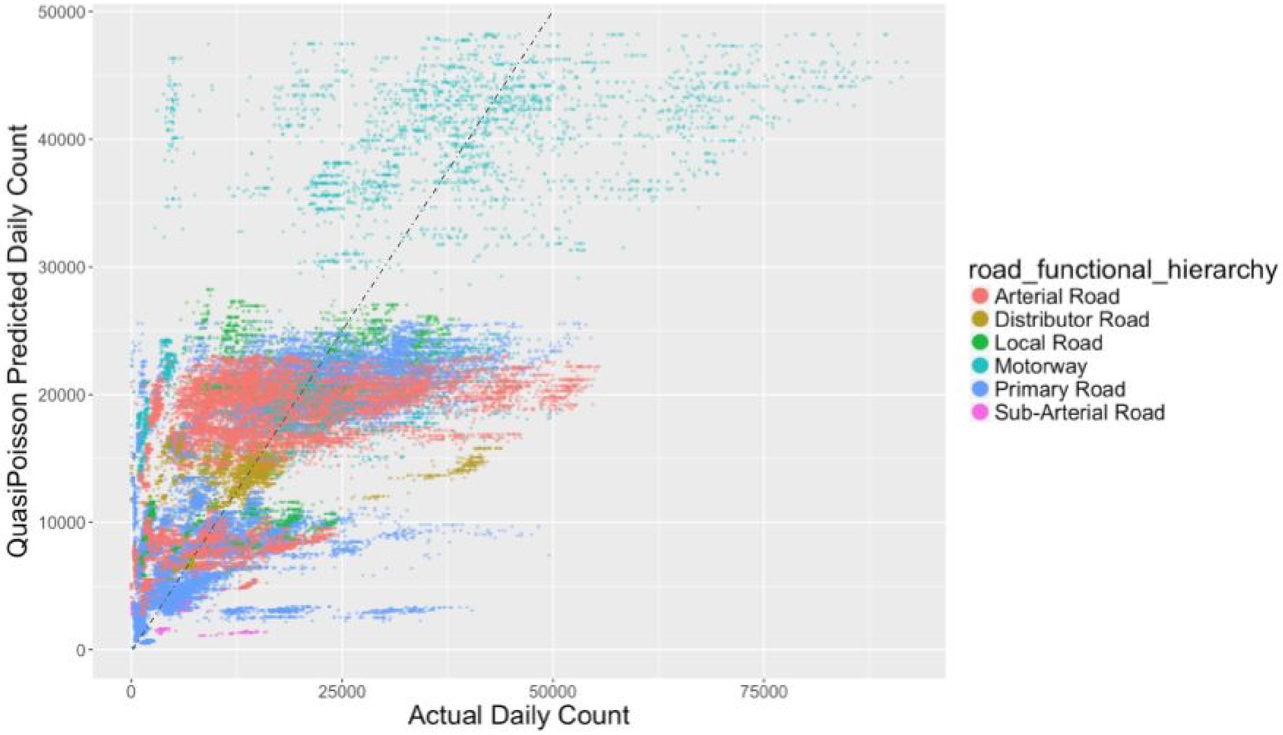
Finally, it is assumed that RMS have both the staffing capacity and technical expertise to implement the findings of this report. Noting the broad context in which findings could be implemented, however, decisions about model and feature selection prioritised interpretability to provide for the broadest possible application.

Deployment

As shown in Figure 8 and Tables 6-10 below, our model is more successful in some areas than others:

- It is better at predicting daily traffic volumes for Primary, Arterial and lesser roads than it is in predicting Motorway traffic. There are large numbers of observed Motorway values in excess of 30,000 daily movements that our model is not yet able to predict.
- There is significant variability in the predictive power of the model for more extreme conditions on non-Motorway roads, for example our model tends not to predict values above 23,000 daily movements on Arterial Roads despite significant number of actual observations between 25,000 and 50,000.
- There are two 'plumes' of Primary Road observations in excess of 25,000 where our model is predicting less than 10,000. These may be examples where the Road Function classifier is inappropriate, or where additional classifiers are required.

FIGURE 8 - Model performance on Holdout dataset - Predicted traffic volume vs actual



Tables 6-10 below illustrate the performance of our model on our holdout data set.

TABLE 6

RMS Region	**Mean Squared Error**
Hunter	6,557 vehicles per day

Northern	6,580 vehicles per day
South West	1,184 vehicles per day
Southern	4,209 vehicles per day
Sydney	7,982 vehicles per day
Western	678 vehicles per day

TABLE 7

Road Function	**Mean Squared Error**
Arterial Road	6,816 vehicles per day
Distributor Road	4,643 vehicles per day
Local Road	7,462 vehicles per day
Motorway	11,806 vehicles per day
Primary Road	4,724 vehicles per day
Sub-Arterial Road	3,303 vehicles per day

TABLE 8

Day of Week	**Mean Squared Error**
MONDAY	6,151 vehicles per day
TUESDAY	6,336 vehicles per day
WEDNESDAY	6,405 vehicles per day
THURSDAY	6,551 vehicles per day
FRIDAY	6,750 vehicles per day
SATURDAY	6,089 vehicles per day
SUNDAY	5,474 vehicles per day

TABLE 9

Daily Rainfall	**Mean Squared Error**
No Rain (0-5mm)	6,235 vehicles per day

Sprinkle (5-10mm)	6,471 vehicles per day
Rain (10-20mm)	6,819 vehicles per day
Heavy Rain (20-50mm)	6,683 vehicles per day
Torrential Rain (>50mm)	7,455 vehicles per day

TABLE 10

Distance to CBD	**Mean Squared Error**
Within 5km	11,528 vehicles per day
5km to 20km	9,112 vehicles per day
20km to 40km	7,281 vehicles per day
40km to 100km	6,287 vehicles per day
100km to 250km	5,509 vehicles per day
Greater than 250km	2,828 vehicles per day

Consequently we recommend that deployment of the model begins with Primary Roads, while the following steps are undertaken in order to be able to deploy more widely:

1. Develop additional road classifiers, as discussed previously in *Data quality - imbalanced data classes* . This is essential if we are to improve the predictive power of the model for Primary and Arterial Roads.
2. Examine Motorway data by specific location or route. Motorways would appear to each have unique characteristics that may be best modelled individually rather than as an overall class of road.
3. Establish a feedback mechanism whereby the results of the model are periodically compared to actual observed traffic volumes in order to assess and review the appropriateness of the model parameters.

At the outset of the project we outlined a number of potential uses for the model. One of these was the application to roads where no physical counter station is in place. As an illustration of this we've prepared some scenarios to show how the model can be used to predict daily traffic volumes in new locations without the need for additional expenditure on traffic counter stations.

Table 11 below outlines a series of hypothetical scenarios, to demonstrate one of the ways in which the model could be deployed. It predicts traffic volume with reference to the key variables noted above, all of which are easily accessible at any given point in time. In doing so, it effectively removes the need for new traffic counter stations to be installed in these areas. At a micro level, it could also be useful in assisting travellers to choose optimum routes, particularly during busy school holiday/public holiday periods.

TABLE 11 - Model predictions for a sample of hypothetical scenarios

Distance to CBD	**Road Function**	**RMS Region**	**Day of Week**	**Public Holiday**	**Daily Rainfall**	**School Holiday**	**Prediction of Daily Count**
10km	Primary Road	Sydney	Monday	No	10mm	No	**21,694**
25km	Arterial Road	Sydney	Sunday	No	30mm	Yes	**14,062**
50km	Motorway	Sydney	Friday	Yes	Nil	No	**32,589**
100km	Primary Road	Hunter	Saturday	Yes	Nil	Yes	**7,707**
200km	Local Road	Western	Tuesday	No	Nil	No	**3,597**

References

Yang, S., Wu, J., Du, Y., He, Y. & Chen, X. 2017, 'Ensemble Learning for Short-Term Traffic Prediction Based on Gradient Boosting Machine', *Journal of Sensors*, pp. 1-15.

David Anker
Rohan Danis-Cox
AJ Duncanson
Jeremy Moon
Jay Radhakrishnan

Gridlock ANALYTICS

Appendix to Assessment Task 2 report

R Code

R Code can be reviewed here: http://rpubs.com/GridlockAnalytics/313831

[Note to UTS Assessors - in the real world we would deliver the code via a more secure channel than RPubs!]

Response to feedback

Feedback on proposal	Response
A Poisson log-linear model would be more appropriate than linear regression, noting traffic volumes cannot take on negative values.	Poisson log-linear models have been considered and a variation used in the final model.
The second paragraph of "modelling techniques to be employed" does not make a lot of sense. Seasonality can be accounted for by including the relevant variables (such as an indicator variable for a public holiday or the day before) in the model.	On reflection we did not feel it was necessary to split our data into test and training sets chronologically.
Multilevel models (also known as mixed effect models, random effect models, longitudinal data, hierarchical models – see Module 2) can account for the correlation induced by multiple measurements within the same area or road. This should be regarded as a later step, after a simpler model has been fitted.	Unsure how to respond to this feedback
Variables which are significant predictors in isolation may not be when treated collectively. In multiple regression, the significance of a variable has to be interpreted as "given other variables already in the model, am I significant"? For example, population density might be a significant predictor when treated alone, but given vehicle registration density it might no longer be important.	Variable significance was considered during variable selection
Forward and backward variable selection are equally appropriate in this scenario.	Forward variable selection included for reasons noted in the report

Further to the clarification provided by Kirsty on CiCAround, feedback received on the presentation has been incorporated into this report but not explicitly listed here.