Investigation of current traffic data practices and its application in quantitative decision making

Saif Ali Athyaab

Rachel Cardell-Oliver & Débora Corrêa
Computer Science and Software Engineering
The University of Western Australia

Bruce Ling & TK Kim
CEED Client: Main Roads WA

Abstract

Traffic models are used for simulating traffic behaviour before the physical implementation of infrastructure changes takes place. One of the input data sources for traffic models is survey data. On-site surveys can measure traffic volume, average speed of vehicles, and count different vehicle classes. For the models to be realistic and useful, the data gathered, and traffic model calibrated should represent a "typical" weekday in terms of traffic flow for a road section. To ensure this is the case, we need to be able to identify when a traffic pattern is of an anomalous day. To date, this project approaches the problem of identifying anomalous days by connecting historical traffic flow data with context information from Main Roads Western Australia - data sources that identify accidents, lane closures, rainfall and ongoing construction activities. Clustering results show fair degree of separation between different contexts for seven clusters. The findings of this study will help the Operational Modelling and Visualisation team to estimate the occurrence of "typical" days when making decisions on the collection of survey data, potentially reducing costs in terms of man-hours spent in model calibration.

1. Introduction

Part of the Network Operations Directorate within Main Roads Western Australia (Main Roads), the Operational Modelling and Visualisation (OMV) team is responsible for developing base traffic models. One of the inputs that needs to be gathered for these models is survey data, which is usually outsourced to a survey company. Survey data contains information such as traffic flow (hourly volume of vehicles), average speed of vehicles, and the count by classes of vehicles. To ensure representation of traffic flow during the most critical conditions, peak hours are chosen when collecting the survey data. For these models to be realistic, the data gathered, and model calibrated should represent a "typical" day in terms of traffic flow for a road section. Exploration of dynamic properties and identification of "typical" traffic flow is a serious challenge in transportation engineering (Homburger et al., 1982). Currently, Main Roads has an Operational Modelling Guidelines document which includes general advice on how the traffic modeller should decide when a traffic survey should be conducted. Current survey data collection processes suggest selecting 30th busiest day of the year to represent a typical day. Generally, Mondays, Fridays, weekends, school holidays and public holidays are exempted.

Main Roads data sources contain context information that are equally important but not as widely used in conjunction with survey data. Context information could consist of accidents, lane closures, roadblocks, major social events such as sports and concerts, or ongoing construction activities. This project aims to understand and characterise the “typical” day of traffic flow using this context information and a less complex version of survey data (without lane-by-lane traffic flow bifurcation) which Main Roads stores as historical data. This study aims to utilise the technique of time series clustering to separate out anomalous days from "typical" ones.

Outside of Main Roads, (Li et al., 2021) proposed a clustering framework for traffic flow time series data and concluded the effectiveness of Gaussian Mixture Model (GMM) and BIRCH algorithms in identifying patterns within the Shenzhen city road network. GMM is a time series clustering algorithm based on Gaussian distribution functions, which uplifts the shape of clusters from simple circles through the use of standard deviation. BIRCH is a hierarchical clustering algorithm suitable for large datasets which reduces the input dimensions first into clustering features, and then a clustering tree consisting of several subclusters. Other methods of characterising "typical" traffic conditions include identification of clusters having similar behaviour within a day (Esfahani et al., 2018) and establishing similarity between neighbouring road sections for imputing missing data (Qi et al., 2016). Rainfall, as a form of contextual information, was found to have a notable impact on the changing traffic flow patterns for the city of Brisbane, Australia (Qi et al., 2020). In particular, the amount of traffic flow was found to be increased at a given location on a wet day compared to a dry one.

This project entails evaluating the feasibility of calibrating simulation models through insights obtained about the dynamics of incident impacted traffic flows, as well as to better inform planning and methods of data collection. The outcomes of this project will shed light on the compatibility of some methodologies with the needs of Main Roads and will evaluate the usability of the data currently available. In this project a modular tool capable of characterising the traffic flow (in terms of “typical” or anomalous) of a particular road section will be developed. The possibility of relationships between different contexts and traffic flow profiles will also be explored.

2. Process

The project follows a conventional data science lifecycle. The process comprises of three main stages: data-collection and processing, exploratory data analysis and time series modelling. To gauge the effectiveness of the process, the hypotheses and rules for selecting a "typical" day mentioned in Main Roads' Operational Modelling Guidelines document, shall be critically examined in the analysis.

2.1 Data Collection and Pre-Processing

All of the required data was already collected and maintained at Main Roads. This project involves merging and analysing multiple datasets from different resources. Generally, the data falls into two categories: Traffic Volume and Context Information. Data processing is required to convert the data into a dedicated dataset for this project's analysis and research purposes. The first dataset is the NetPres Traffic Fact, collected from sensors across WA, and conditioned by

the Main Roads Network Analysis team. Relevant variables include VolumeCount and Rainfall_BOM. VolumeCount is the number of vehicles passing across a road segment every fifteen minutes. RainfallBOM is the amount of rain (in millimetres) recorded, on an average, every 15 minutes. The second dataset is WebEOC incidents, obtained from the Emergency Response Team central repository, outlines details of incidents such as weather hazards. incident timestamp and incident type that are important contextual indicators. The Waze incidents dataset is crowd-sourced along with a Reliability indicator (0-10) for each incident reported. While a large number of incidents are reported on Waze, extra care was taken during data pre-processing to filter out duplicate reports of the same incident. Several strategies were explored to de-duplicate, including; simple pruning of incidents having a Reliability less than 7; removal of incidents lasting less than 10 minutes; and spatio-temporal pruning based on same Incident Type.

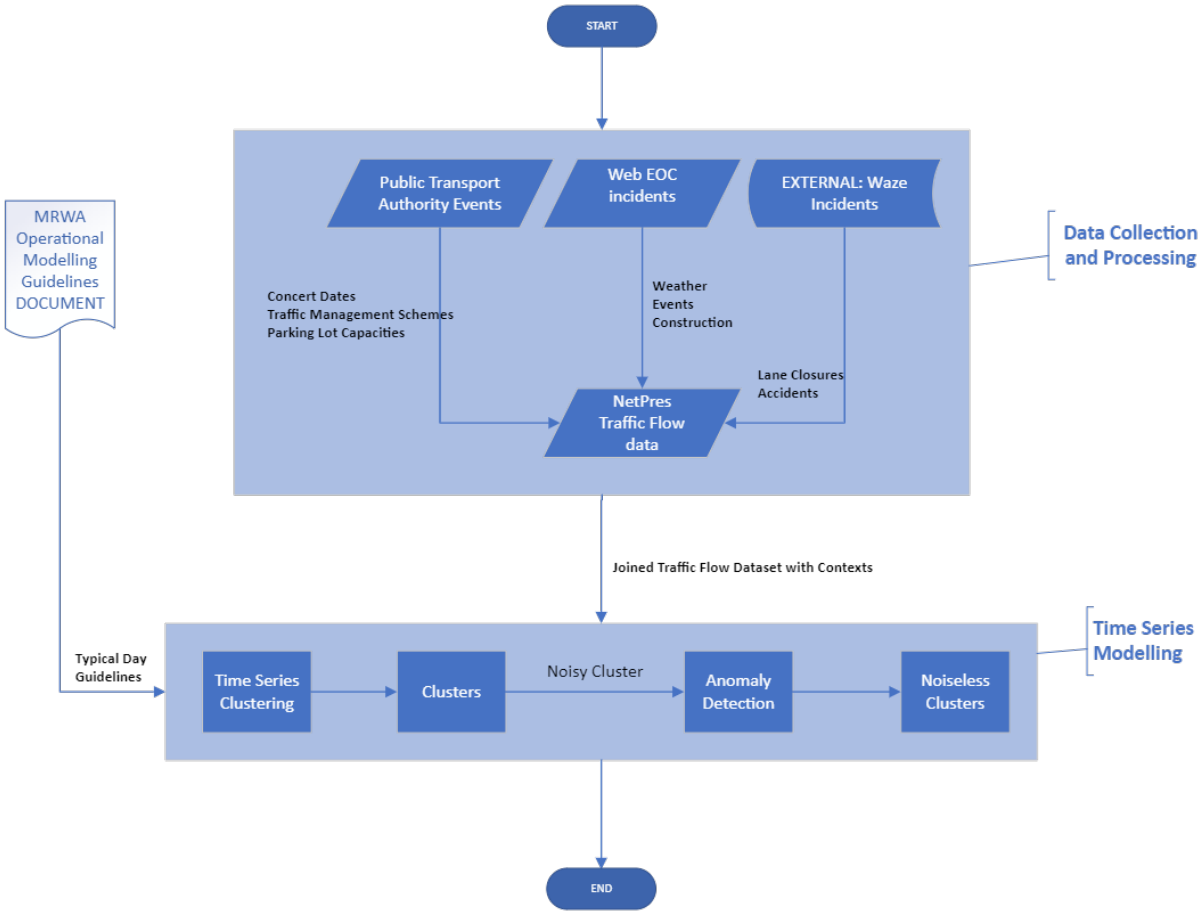

Figure 1 Project Lifecycle process

2.2 Exploratory Data Analysis

This part of the process provided preliminary insights such as the range of volume and speed for each road segment. The main charts of interest include; a line chart plot of volume on y-axis and the time of day on the x-axis; a scatterplot of volume and speed; to identify any correlation, examination of the plots of anomalous days and justification with contexts. A Pearson correlation coefficient is calculated between VolumeCount and SpeedKmh to identify the relationship between them.

2.3 Time Series Modelling

To identify anomalous days in an automated fashion, time series modelling has been employed. The idea is to plot a graph of Volume versus Time of the day. Each graph represents a single time series consisting of 96 points containing volume data points ranging from 12 am to 11:45 pm. By replicating this for each day of the year across a particular road segment, we end up with 365 time series.

The k-means clustering algorithm is deployed to find clusters having common characteristic peaks in the time series graphs. k-means works by calculating distances between different time series and then arranging the least distant ones into the same cluster, while also attempting to maximise distance between different clusters. In our case, the effectiveness of the k-means can be evaluated by how well each cluster only represents a specific class of contexts, and degree of separation between typical day clusters and anomalous ones. Ideally, perfect separation of all contexts would mean good cluster separation. So, if a cluster contains all January non-rainy weekdays without any school, public holidays, without incidents, then we can say the model has good cluster separation.

Experiments were run on a range of 2 to 15 for the number of clusters (k), while also maintaining a set seed value for cluster stability and reproducible results. To complement this, a k Nearest Neighbours (knn) anomaly detection method is used to separate the anomalous days from the noisy typical day clusters. This is particularly useful when typical day clusters have some noise within.

The clustering process went through several iterations. In the first iteration, a single road segment is selected for clustering. We cluster on 365 time series (2023) over a numerical variable such as volume. The expected results are different patterns in each cluster. This gave a good starting point to estimate the number of clusters expected. The cluster labels and corresponding data points were evaluated with Main Roads to then classify each cluster as “typical” or “not typical”. Preliminary contexts such as Day of the Week and weekends were connected to each cluster or day’s time series. The cluster labels were then used to support or refute Main Roads guidelines, such as skipping Mondays and Fridays for data collection.

The second iteration involved filtering out public holidays, school holidays and weekends, since Main Roads doesn't intend to collect data on these days for their transport models. This encourages the model to separate the anomalous days more effectively based on incidents rather than public holidays or weekends. We repeat the process on the same road segment but a smaller number of days. After the clusters are formed, we start by connecting individual stand-out patterns of days with context information such as existence of accidents, and construction on that day. By aggregating these stand-out days with cluster labels, one would expect clusters representing specific patterns to have similar context information data.

3. Results and Discussion

For all values of Volume and Speed of a particular road segment over the year 2023, there is a negative 31% Pearson pair-wise correlation. This can be explained by the simple transport engineering rule that increase in the number of vehicles on the same road segment would decrease the average speed of all the vehicles given capacity of the road segment is constant.

However, the mere 31% doesn't particularly imply causation, and so this gives us more reason to carry out time series modelling with different contexts for inference.

The first iteration of k-means clustering showed a good cluster separation in terms of basic contexts like weekends, public, and school holidays. Typical weekdays appeared in different clusters based on day of the week, month of the year, and to some extent rainfall. However, it required a large k value (>15) to separate anomalous days, but even then, they often appeared inside other clusters. It is important to note that only the time series and not the contexts were fed into the models. This means that our model has separated the input time series into different clusters, which are being inferred from the contexts after obtaining the clusters. Table 1 highlights the cluster separation, for some k values, against types of contexts.

Context	Number of Clusters - k			
	2	5	7	10
Public Holidays	× (2 holidays)	×	×	×
School Holidays	×	×	×	×
Weekends	×	×	✓	×
Day of Week	×	×	✓	×
Month	×	✓	✓ (6 months)	✓
Rainfall	×	×	✓	×
Incidents	×	×	×	×

Table 1 A qualitative ablation study of separation of contexts for different k values in the year 2023. 7 clusters seems to be the most appropriate.

A slight improvement was observed in the second iteration in terms separation of contexts. Upon restriction of only standard Mondays to Fridays as inputs to the clustering model, 2 main clusters were observed – typical January to June weekdays, and typical July to December weekdays. After passing these through the knn anomaly detection method, most of the anomalous days were separated from noisy typical day clusters.

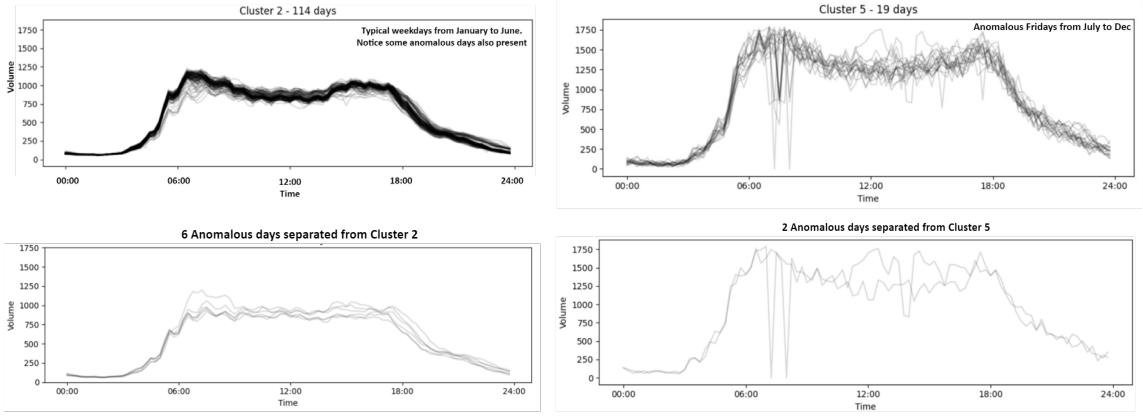

Figure 2 Clustering and Anomaly detection results of all 2023 weekdays excluding public and school holidays. In the top-left we see typical weekdays from Jan to Jun with some anomalous days extracted in bottom-left. On top-right are anomalous Fridays of Jul to Dec along with extreme anomalous days in bottom-right

4. Conclusions and Future Work

A proof of concept has been established for the development of a typical day identification framework. While clustering along with anomaly detection demonstrated fair separation of contexts, the method fails to perfectly separate all anomalous days. Based on the modelling, clustering alone cannot be used as a means for separating typical days. Anomaly detection when used in conjunction yields better results. Further work is needed to obtain stable clustering through the use of hierarchical clustering, by utilising the centroids to feed into the K-means algorithm. Since the current project has been restricted to a single road segment analysis, future work by Main Roads can look at ways to extend the scope to parts of freeways. This would allow for better typical day identification over a wider geographic scope and a narrower temporal scope, as is expected when looking to collect survey data.

5. Acknowledgements

I extend my heart-felt gratitude to everyone involved in the facilitation of this project. My academic supervisors, Prof. Rachel Cardell-Oliver, and Prof. Debora Correa have been paramount in my ability to understand the intricacies of the data. I commend my client mentors TK Kim and Bruce Ling who gave me invaluable guidance and support throughout this project. My appreciation also goes to Bryon Thong and the Data team, for their hands-on domain knowledge explanations and provision of data, respectively. I would also like to extend my thanks to Jeremy Leggoe and Kimberlie Hancock for arranging the CEED program and allowing me to be part of it.

6. References

He, D., Kim, J., Shi, H., & Ruan, B. (2023). Autonomous anomaly detection on traffic flow time series with reinforcement learning. Transportation Research Part C: Emerging Technologies, 150, 104089. https://doi.org/10.1016/j.trc.2023.104089

Homburger, W. S., McGrath, R., & Keefer, L. E. (1982). Transportation and Traffic Engineering Handbook. Second Edition. Prentice-Hall.

Kouhi Esfahani, R., Shahbazi, F., & Akbarzadeh, M. (2018). Three-phase classification of an uninterrupted traffic flow: a k-means clustering study. Transportmetrica B: Transport Dynamics, 7(1), 546–558. https://doi.org/10.1080/21680566.2018.1447409

Li, R., & Yu, J. (2021). Clustering framework based on multi-scale analysis of traffic flow time series. Journal of Physics: Conference Series, 1952(4), 042056. https://doi.org/10.1088/1742-6596/1952/4/042056

Qi, H., Liu, M., Wang, D., & Chen, M. (2016). Spatial-Temporal Congestion Identification Based on Time Series Similarity Considering Missing Data. PloS One, 11(9), e0162043–e0162043. https://doi.org/10.1371/journal.pone.0162043

Qi, Y., Zheng, Z., & Jia, D. (2020). Exploring the Spatial-Temporal Relationship between Rainfall and Traffic Flow: A Case Study of Brisbane, Australia. Sustainability, 12(14), 5596. https://doi.org/10.3390/su12145596

Yan, Y., Zhang, S., Tang, J., & Wang, X. (2017). Understanding characteristics in multivariate traffic flow time series from complex network structure. Physica A, 477, 149–160. https://doi.org/10.1016/j.physa.2017.02.040