NACOE
NATIONAL ASSET CENTRE OF EXCELLENCE

FINAL REPORT

R101 – Automated Roadworks Detection using Next Generation Traffic Data (2019–20)

ARRB Project No.: 015300

Author/s: Edward Dann, Will Hore-Lacy, Young Li, Anthony Germanchev

Prepared for: Queensland Department of Transport and Main Roads

June 2020

Version 3

AN INITIATIVE BY:
arrb Queensland Government

SUMMARY

Roadworks cause disruption to the transport network resulting in delay and congestion. Transport and Main Roads Queensland (TMR) maintains a register of planned roadworks events, however in many cases the roadworks might not occur at the exact times listed.

The aim of this project was to conduct real-time detection for roadwork events on the Queensland road network using data sourced from HERE Technologies real-time traffic. The intention is to seek to use probe data sources to validate activation and deactivation times of roadwork sites, to potentially support improved traveller information as well as support auditing.

The research found that data acquisition and matching from the QLDTraffic service and HERE Technologies Traffic API is possible but is a very complex process with significant limitations on accuracy. The process required geospatial selections and map matching across three separate road geometries. Significant limitations in the coverage and quality of both the roadworks and traffic data were identified which should be addressed prior to any further work on this task.

These limitations however did not prevent an anomaly detection method from being developed which in some cases provided good results and provides valuable insights into future work and process improvements. The key findings were:

- The data aggregation method is powerful and extendible to other projects.
- The k-means method tested may be suitable for use with other event types.
- An improved understanding of traffic impacts resulting from the event types is required to further develop an anomaly detection algorithm.

While the data aggregation methods did not overcome the issues of data quality, the map-matching based approach used to associate traffic data with roadworks events represents a successful application of these techniques.

It is expected that the quantity of probes being logged by HERE will continue to improve, which will enable the methods developed to be applied on more roads, in rural areas and on minor roads. Improvements in traffic data quality may also allow anomaly detection to be refined in order to produce more accurate results in future research.

Although the Report is believed to be correct at the time of publication, the Australian Road Research Board, to the extent lawful, excludes all liability for loss (whether arising under contract, tort, statute or otherwise) arising from the contents of the Report or from its use. Where such liability cannot be excluded, it is reduced to the full extent lawful. Without limiting the foregoing, people should apply their own skill and judgement when using the information contained in the Report.

Queensland Department of Transport and Main Roads Disclaimer
While every care has been taken in preparing this publication, the State of Queensland accepts no responsibility for decisions or actions taken as a result of any data, information, statement or advice, expressed or implied, contained within. To the best of our knowledge, the content was correct at the time of publishing.

ACKNOWLEDGEMENTS

The project team would like to acknowledge the prior work and assistance provided by TMR Traffic Engineering, Technology & Systems.

CONTENTS

TABLES

FIGURES

1 INTRODUCTION

Transport and Main Roads Queensland (TMR) rely on accurate and timely data to perform their traffic planning and management responsibilities. The commercial probe traffic data sources that are available today offer network-wide traffic intelligence by providing speed and congestion information at high resolution. This emerging data is sourced from connected navigation systems and mobile telephones within vehicles and offers the potential to better understand traffic flow and disruption.

Roadworks cause disruption to the transport network resulting in delay and congestion. TMR maintains a register and publishes planned roadworks events (via QLDTraffic), however in many cases the roadworks do not occur at the exact times listed. This may be because the roadworks occurred outside of the permitted times or because the works did not use all of the allotted time (some events block out periods of more than 6 months). In both cases an independent and automated system to detect activation of roadworks can assist road managers in improving traffic flow and provide more accurate travel time information. It will also provide information to better manage roadworks and traffic management contractors.

The QLDTraffic web service API (https://qldtraffic.qld.gov.au/) was used to source data from the TMR roadworks register. The service provides planned roadworks events and a mostly structured data feed that provides location, dates, times, conditions and other general information, which are typically sourced from road occupancy permits.

The HERE Traffic API (https://www.here.com/products/traffic-solutions/real-time-traffic-information/) offers real time traffic information across the Australian road network (and the rest of the world). The information is sourced from 'probe' measurements and provides speed information about traffic flow for sections of the road network. It is also a system that is independent from the Queensland roadworks register making it a good candidate data source to be used to detect when roadworks events are in progress.

This project was predicated on the basis that accurate information about the timing (activation and deactivation) of these events will allow road managers to optimise the flow of traffic on their network, providing a more effective information service to road users enabling them to make more informed travel decisions around network impacts, reducing the amount of information and number of notifications sent to road users, and supporting auditing activities.

2 SCOPE

The goal of this project is to conduct real-time roadworks detection for roadwork events occurring on the Queensland road network. For the purpose of this project, roadworks is defined as an expected roadworks type incident in the TMR QLDTraffic API and roadworks detection is the spatial and temporal matching of these incidents to the real-time traffic data.

Two primary aims were outlined for this project, the first was to create a test dataset in which patterns may be observed. The second was to use automated machine learning techniques to detect such patterns within the dataset for a variety of event types and locations. If a sufficiently strong relationship was found the project would attempt to create an automatic roadworks detection algorithm which would flag inconsistencies for TMR.

This project acknowledges that TMR had undertaken initial work to compare planned roadworks against real-time traffic data from HERE Traffic and STREAMS to determine whether the works were in operation (affecting traffic flow). The initial findings of this work indicated that there was a correlation between the speeds obtained from HERE Traffic and STREAMS loop data at two test sites. This project extended on the initial work and investigated more sites to explore the feasibility of a reporting service that identifies and describes discrepancies between planned roadworks and observed traffic. It was intended that the methodology developed in this project would function at both urban and rural locations as well as on major and minor roads.

The approach adopted in this research was as follows. The speed data used to detect roadworks was logged from the HERE Traffic API and stored in a database to allow the detection methods to be developed. To detect roadworks events, the QLDTraffic events were compared and pattern matched against the recorded real-time data.

3 BACKGROUND

3.1 MOTIVATION AND PREVIOUS WORK

TMR use HERE Technologies Traffic products as part of their infrastructure planning, this includes the HERE Traffic API. This web-based API provides real-time traffic data for road segments, and has been used to support real-time operations, primarily relating to identifying unplanned performance issues on the network.

Roadworks information (as planned events) is captured and provided to the public in the interest of enabling informed travel decisions. However, this information is based on permitted times, with visibility of actual activation times (meaning actual impact times) is not always known, making the roadworks information less meaningful to the public.

Roadworks activation detection is a current need for the organisation, and this application was identified for further investigation. Work was conducted by TMR on this application by conducting a case study using the HERE Traffic API. Traffic speed was logged for the duration of a known roadworks event that was located on a major urban road close to a STREAMS induction loop detector. The data from the two sources was closely correlated and showed a drop in speed due to the roadworks activity. This work created several research questions which drove the investigation of this report.

Firstly, is a change in traffic always detectable around roadwork events, what is the distance of impact, and can it be distinguished from transient changes in traffic? Secondly what methods can be used to automatically detect these dips in traffic flow?

3.2 LITERATURE REVIEW

A literature review of anomaly detection, particularly in the domain of traffic flow was conducted. The key findings are summarised and presented below and used to inform the method later used for evaluating the datasets.

Anomaly detection describes a broad range of statistical techniques which can detect outliers based on expected values. In this review anomaly detection techniques are considered which detect any discrete events resulting in an impact on traffic flow. Many of the studies focus on accident detection, an event type with slightly different characteristics. The techniques are transferrable but specific parameters may not be. While the included studies typically restrict their analysis to high volume roads, they still employ a wide variety of detection techniques. Discussion of commercial detection systems is excluded from this review.

Houbraken et al. (2017) applied a heuristic method for road incident detection using real-time floating car data. In the Automated Incident Detection (AID) software, a virtual sensor was placed in each road segment. The average speed for the road segment was updated when a car passed the virtual sensor. If the average speed went below the predefined lower speed threshold, a congestion notificationwas triggered by the sensor. If the average speed went above the predefined upper speed threshold, the sensor reported a free-flow notification. The floating car data was anonymously collected from individually tracked vehicles via a free transport app. This data provided information about geographical position and instantaneous speed of the probe vehicles.

A statistical method was proposed by Pietrobon, Lewis and Heverly-Coulson (2019) to detect the event of road closure. Two models were developed for open roads and road closures, respectively. The road open model contained expected probe activity per time interval which was constrained between the predefined minimum and maximum values, and the expected variance which was equal to expected probe activity. The road closure model for describing residual traffic activities was defined as a normal distribution centred around 1, with a variance of 3, truncated at 0. The road open model was used as the historical model to which real-time probe activity for a road segment was compared. When the collective likelihood of the observations for multiple time intervals was lower than the predetermined threshold (i.e. p-value test), the road open model was not a good description of the observed data, hence the road segment was labelled as

potentially closed. The log-likelihood ratio was also computed to compare the road open and road closure models, and to indicate an on-going road closure. The probe activity data used was from anonymised probe vehicle data, including time stamp, geographical position, speed and heading.

Wang et al. (2017) introduced a two-stage solution for road anomaly detection that consisted of collaborative path inference (stage 1) and road anomaly detection (stage 2) to identify traffic anomalies from Global Position System (GPS) snippet data. The collaborative path inference model incorporated both static and dynamic features into a Conditional Random Field (CRF). GPS snippet data was transformed into road segment-based aggregate data. For road anomaly detection, the Likelihood Ratio Test (LRT) was used to determine whether a road segment was anomalous. Finally, the RN-Scan algorithm searched the whole road network to find all possible anomalies. Two types of traffic anomalies were detected in this research: self-evolving anomaly (i.e. volume deviates from mean values) and context-evolving anomaly (i.e. volume deviates from the values of its connected neighbours in a region). The GPS location data used for this study contained over 29 million measurements from 8,602 taxicabs during May 2012 in Beijing, China. The sampling rate of the GPS data varied from 30 seconds to 10 minutes.

Gakis, Kehagias and Tzovaras (2014) adopted a Support Vector Machines (SVM)-based approach for detecting incidents in traffic networks. In order to improve the detection ability of the SVM classifier, two features were used, which were speed difference between current road segment and next road segment in the same direction as well as average deviation of speed in current road segment. The traffic dataset used for this research was derived from the I-880 Freeway in California. This dataset contained traffic information about speed, occupancy and flow however only speed information was used in this study. The measurements in the dataset were captured by 35 pairs of inductive-loop detectors (18 for northbound and 17 for southbound) across a road section of 14.5 km. The data was collected at a 30-second interval for two time periods each day from 16 February to 19 March 1993. The first period was from 5 am to 10 am and the second from 2 pm to 8 pm. In addition, the dataset provided the information about road incidents, including time, location, severity and the lanes affected. The final dataset was split into training and test subsets to allow for validation for the model.

Ozbayoglu, Kucukayan and Dogdu (2016) proposed a real-time autonomous highway accident detection model based on computational intelligence techniques. Five features were used as model inputs, including average velocity difference, average occupancy difference, average capacity usage, weekday/weekend flag and a rush hour flag. The occurrence of an accident/event was the output feature to be classified by the accident detection models. Three differential computational intelligence models were selected for accident detection analysis. These were nearest neighbour, regression tree, and feedforward neural networks. The data used in this study was obtained from 7 separate Real-Time Monitoring System (RTMS) sensors across 24 km on one major highway in Istanbul, Turkey for the 2015 calendar year. The data included information about vehicle counts, average speed and average occupancy ratio every 2 minutes in each lane. The incident data for the specific road section of the major highway during 2015 was also extracted, which included the time, date, location and description (e.g. direction and type of incident) of the event. While the number of false alarms detected was considerably high, the overall accuracy of the computational intelligence models was mostly over 95%.

A summary of the key methods and data sources identified from the review literature is shown in Table 3.1. The methods used in this review are varied and typically try to take advantage of the data quality as much as possible. Interestingly, Pietrobon et al. (2019) and Wang et al. (2017) both used vehicle probe data indicating a strong similarity to the data used in this project. These approaches diverged from the current projects requirements when considering data quality and coverage. Pietrobon et al. made use of a low coverage but high-quality dataset, enabling accurate insights in small but critical regions. Wang et al. used a dataset closer to the HERE Traffic API but were only able to detect simple traffic patterns such as open and closed roads. Excluding Wang et al. data used for this research was generally of higher quality, but lower coverage than the data available for this project. Many of these studies were primarily focused on high-volume uninterrupted roads while this study was focussed on a range of road types. In a number of studies additional indicators of event occurrence (e.g. volume or occupancy) were not available for this study.

This literature review also suggests that roadwork detection is an unsolved problem, with a lack of consensus on key metrics such as speed, volume and flow. The accuracy of the detection is also clearly related to the quality of the data available. Therefore, methodological decisions will be made based on the characteristics of the test dataset, choosing the method which suits the available traffic and event data.

Table 3.1: Summary of reviewed literature

Literature	Scope	Indicator	Data source	Data coverage	Data capture period	Ground truth
Houbraken et al. (2017)	Real-time traffic monitoring	Speed	Probe data via app	Two different highway sections (one for 19 km, the other for 37 km)	4 weeks	Compared to loop sensor data
Pietrobon, Lewis & Heverly-Coulson (2019)	Road closure detection	Probe vehicle count	Probe data via mobile devices	12 metro areas from small (50,000 people) to large cities (8 million people)	18 months for historical baseline; 2 weeks for closure detection	On-the-ground verification and correlation with other reports
Wang et al. (2017)	Traffic anomaly detection	Volume	GPS snippet data	8602 taxi cabs in four different regions of Beijing, China	1 month	No ground truth data, manually generating anomalies
Gakis, Kehagias & Tzovaras (2014)	Road incident detection	Speed	Inductive loop detectors	14.5 km section of the I-880 Freeway in California, US	32 days	Incident information from the dataset
Ozbayoglu, Kucukayan & Dogdu (2016)	Real-time highway accident detection	Speed, occupancy and road capacity factor	Monitoring sensors	24 km section of a major highway in Istanbul Turkey	2015 calendar year	Accident information from the database of Traffic Department

4 METHODS

The methodology for detecting roadworks based on real-time traffic speed measurement changes is described in the following two sections.

Section 4.1 is chiefly concerned with harmonising event and traffic data into a common structure, in a test dataset. The dataset itself, and its quantitative characteristics are discussed in Section 5. Generating data suitable for anomaly detection is non-trivial, nuances in the structure are discussed in detail, and the impact on potential anomaly detection methods is also considered.

Section 4.2 describes the assessment of suitable anomaly detection methods, project implementation, and the method to compare the real time traffic dataset to historical data. Each of these tasks combined formed the process required to detect roadworks in specific geographic areas using a K-means clustering method.

4.1 GENERATING DATA

4.1.1 ROADWORKS DATA OVERVIEW

The roadworks data was sourced from the QLDTraffic GeoJSON API (Queensland Government 2020) service maintained by TMR. The website provides up-to-date information about roadworks as well as other traffic conditions such as natural hazards. When a request is made via the API, a list of events is returned. The API is used to locate roadworks taking place in real time and study the traffic displacement around the location. Over a two-week period 8 to 22 January 2020, a total of 297 unique events were recorded.

The process to extract and record the data was:

1. Call the QLDTraffic API using a public API key.
2. Check for new roadworks events.
3. Add new events to the database.

4.1.2 ROADWORKS DATA

The purpose of recording and inspecting an existing dataset of roadwork events is twofold. Firstly, it makes it possible to focus the investigation on traffic patterns in areas where events have occurred, thereby speeding up the discovery process. Secondly, listed events can serve as a data source for creating a truth dataset in the context of machine learning methods. The ability to meet these aims is dependent on the strengths, limitations and structure of the event data. Particularly, differences in the representation of physical locations must be considered.

Roadwork event data returned from the QLDTraffic API is returned as a JSON array of listed events. Each listing details an event which will exist for a given period. In the case of unscheduled or dynamic events, the listing can be updated. For each listed event metadata is provided, which can vary in detail and encoding method. For a complete list of available metadata, refer to the QLDTraffic API Specification (Queensland Government 2020). Descriptive characteristics of the event, such as the event type can be used to filter events based on type (for instance if the event is a roadworks event) and other information. At times the metadata provided via the API is feature rich and indicates potential for use via an SVM. Some events have much of their data encoded in a free text field, which cannot be interpreted effectively using automated techniques. All listed events also possess an embedded GeoJSON object, which represents the event's physical location. Table 4.1 summarises the QLDTraffic event keys available.

Table 4.1: QLDTraffic event keys

Key	Description	Guaranteed
Roadwork ID	Unique Identifier of the event	Yes
GeoJSON	A geospatial reference which describes the location of the event, can be a single point, line string or array of line strings	Yes
Duration	The start and end dates of the roadwork events	Yes
Schedule	Scheduled times throughout the day the event should occur. Also includes days of the week the event will occur	No
Description	A text blob with human readable context for the roadwork event, can include expected length of delays, lanes which are closed and more granular information about the timing of the event.	No
Recurrences	List of JSON objects describing periods the event will occur in over a single week. All data is presented via JSON keys enabling easy computer interpretation	No
Impact type	Class of the road restriction, number of lanes closed and direction of the impact.	No

While every listed event contains a spatial reference and start/end dates, granular information on the schedule and class of the event is inconsistently reported. One example is the 'impact' object, which can specify the direction, number of lanes blocked and the expected delay. However, this object can often be missing or embedded in a text blob under the description data. Considering the variability in the presence of some keys, much of the alignment and analysis performed considers only metadata which is present in every event. This includes the roadwork ID, the start/end dates, and the GeoJSON location. Notably, not all events returned by the API are currently active. These events may be listed prior to commencement of works or remain listed after the active period has ended.

4.1.3 TRAFFIC DATA OVERVIEW

The traffic speed data was provided by HERE Technologies via the HERE Traffic API using the 'flow' call (HERE 2020) which provides real-time traffic speed information. This dataset provides a high level of coverage across the state, but it is unclear if the data is representative of true traffic flow when the sampling rate is low. Where insufficient probe data is available the service will return gap fill data to provide information at all times, however this data cannot be differentiated from the actual real-time probe data. Data returned from the HERE Traffic API is the primary tool used to study traffic patterns. Changes in traffic near roadworks events are logged at regular intervals in order to detect any reliable patterns.

This data source requires a licence to access the data which can then be queried with an application code and application key pair. The API was queried every 10 minutes for each roadwork event recorded over the 14-day logging period. The process for extracting the data was:

1. Query the API using a bounding box around roadworks geometry.
2. Select the relevant traffic API geometry using map matching with the roadworks geometry.
3. Add the data to the database.

4.1.4 TRAFFIC DATA

Traffic data can be represented using a variety of data structures. Speeds and positions can be recorded as anonymised GPS probe data, represented as a line of beads along a road segment where each bead contains speed and traffic data. Traffic data can also be stored as recorded volume passing through static points. In this project, traffic data from the HERE Traffic API is recorded as an aggregate value of vehicle data averaged along a road segment. Each road segment or link (TMC) is of variable and significant length. The set of all potential connected links is referred to as a network.

The HERE Traffic API returns traffic data along a proprietary geospatial network of TMC links. Finding out what location or road segment traffic data relates to can be achieved via lookup tables using TMC id codes, or by requesting the geometry through the API. A complete network shapefile of all TMC links and corresponding codes is unavailable, thus geometry was accessed through the API. The underlying network is believed to be a derivative of the SUNA TMC network. It is also possible to request traffic data aligned to the HERE maps network via a 'dynamic links' request, which does not use the TMC network. However, the data is extremely sparse, and only contains road links that are occasionally available. It therefore provides little to no coverage for most of the Queensland road network.

Traffic data requested from the HERE Traffic API is also returned in JSON format. Data can be requested within a defined geographic area, but not for specific road segments. The response is an array of real-time traffic records within the area. Each road segment contains a single traffic data object. It is not possible to access the individual probe data for a given aggregate statistic reported via the API. These aggregate statistics are stored as JSON key-value pairs listed in Table 4.2, which always appear along the road segment. The primary metric extracted from the traffic data was the current speed, and the expected speed. Aggregate and unitless statistics like Traffic Flow and Jam Factor are difficult to compare against neighbouring links without a sophisticated linear referencing system. Physical measurements are also directly comparable between different traffic data sources.

Traffic data does not return volume, nor the number of probes used to estimate the speed. The Traffic API also does not allow for the filtering of gap fill data, which refers to artificial measurements generated via interpolation of historical data. The method, and likelihood of receiving artificial speeds through the Traffic API is also unknown. It is also impossible to know from when the most recent physical measurement was recorded for a given road segment. The lack of transparency in low coverage scenarios poses some challenges to detecting remote roadwork events. The practical use of this traffic data should also consider typical licensing terms for its uses. There is typically an upper limit to the number of requests per month, as such tracking changes in real-time traffic should consider the total number of events, and the interval or epoch requests are made in.

Predicting the total requests per month is calculated using the below formula:

$$\text{formula 1} = D*30.41/((60/E)*24) \qquad \textbf{1}$$

where

D = Average daily events

E = Hourly Epoch, or minute interval at which events are logged

Table 4.2 describes the data attributes available via the HERE Traffic API.

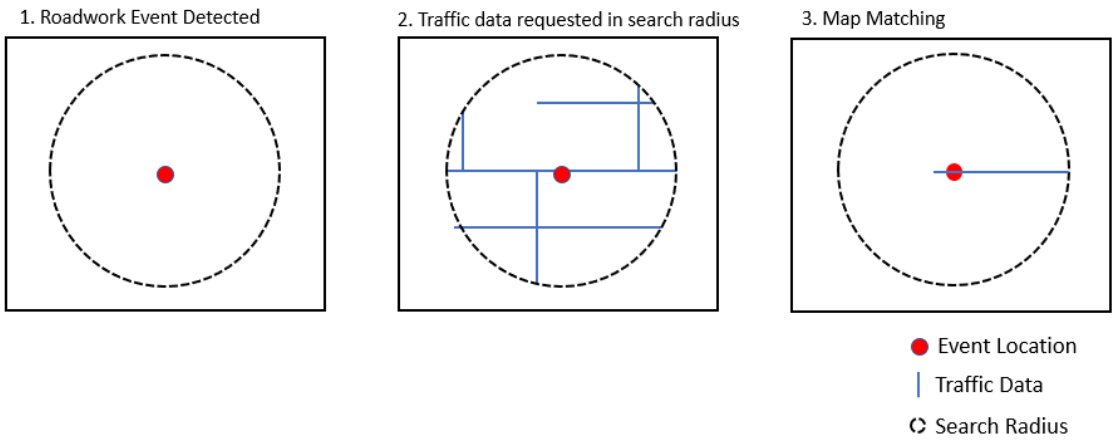
Table 4.2: HERE traffic data keys

Key	Description	Guaranteed
TMC ID	Geometric ID representing the location the traffic data corresponds to	Yes
Link geometry	Array of line strings representing the TMC link	Yes
DE	Text description of the road	No
Current speed	Current estimated speed of traffic on the link	Yes
Free current flow	Aggregate statistic which estimates the flow of traffic along the link	Yes
Speed limit	Speed limit of the link	Yes

4.1.5 DATA LOGGING

Using the HERE Traffic API and QLDTraffic APIs, data was sampled over the two-week period (8 – 22 January 2020) at 10-minute intervals. All QLD traffic events with the 'roadworks' value listed under the 'event_type' tag was recorded. After the metadata was recorded, the event geometry was extracted, and a buffered area was created representing an area of impact the event may have on the surrounding traffic. This buffered area was used as a filter to then request traffic data from the HERE Traffic API. Requesting traffic data indiscriminately would require logging the entire Queensland road network, inflating the cost of the project with marginal benefit to the analysis. Traffic data, and the corresponding geometry was requested only within defined areas surrounding known events. Both the traffic data, and the event data were inserted into separate database tables. If a duplicate event was discovered, the 'last_updated' key was checked against the existing record. Duplicates were only inserted if the event data had changed or been updated. The data was then aligned using a map matching algorithm which determined the physical distance between traffic records and event locations (refer to Figure 4.1 and Figure 4.2). Only traffic data which contained the reported location of the roadwork was included in the analysis.

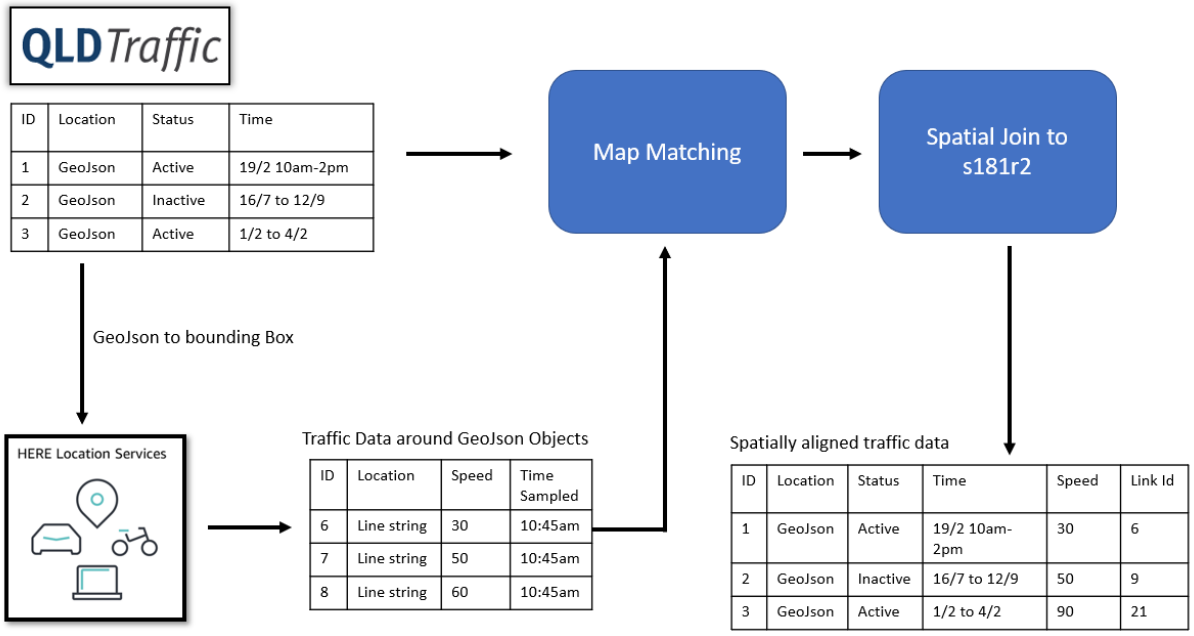
Figure 4.1 Process of selecting HERE traffic data for a roadworks event



4.1.6 COMBINING DATASETS

Data from both sources needs to be linked via a common structure. The most obvious is a tabular format aligning traffic data overlapping with the roadwork events. Importantly, traffic data is specified for a given location which represents a linear section of road, whereas event data may be represented as a single point or set of lines along the road. Any relationship must consider the geographical distance between traffic data and event locations. Traffic data for roads sufficiently distant from the event are likely unimportant and serve only to introduce noise into the analysis. Therefore, a pre-processing step is performed to quantify the physical distances between roadwork events and traffic data geometry. This is referred to as a map-matching problem, where geographic data is snapped to a base network represented as a table of variable length road segments. Combining the event and traffic data via spatial methods (Figure 4.2) creates a flexible data platform thereby enabling an effective analysis and comparison of different machine learning methods.

Figure 4.2 QLDTraffic data logging procedure



4.1.7 MAP MATCHING

Traffic data and event data exist on two separate representations of the physical road network. There is no common referencing system between the structures beyond street names, which are not always present or consistent in either source. Initial research for the project required a GIS user to manually associate specific traffic geometry with event geometry. Quantifying the distance between the physical sources of traffic data and roadworks events is automated in this project through a process of map matching. This module constitutes a step forward in harmonising data from different sources and is highly flexible with potential extensions to other projects.

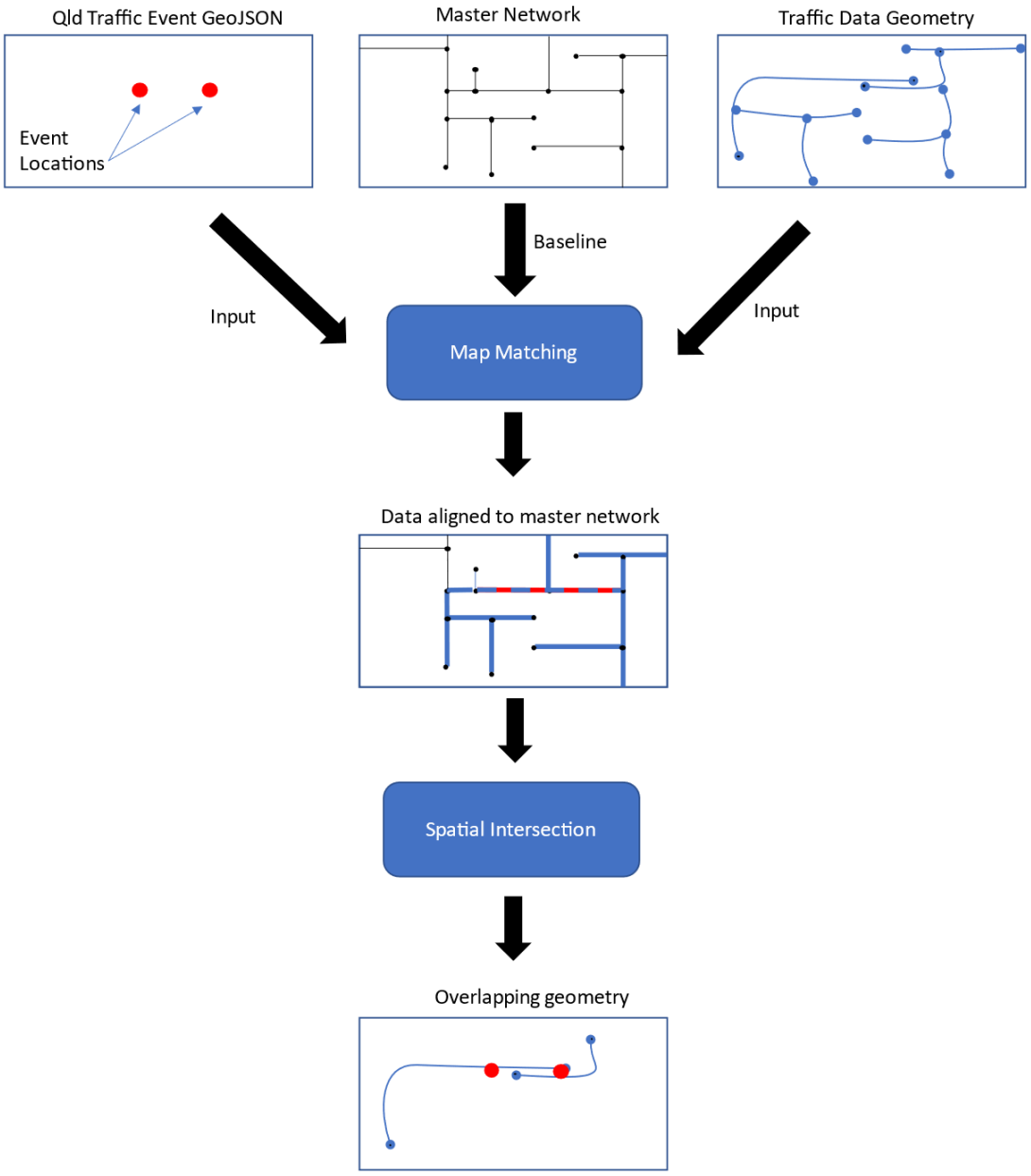
A combination of spatial filtering and map-matching can accurately quantify the distances between geographic objects with high accuracy. Objects with very small distances and high topological similarity are considered comparable, and data is aligned. Instead of comparing geometry from the two datasets directly, a prototype map matching algorithm was used to align both datasets to a master network. This algorithm was an early iteration of the map matching program developed in NACOE project R102. Further documentation of this method can be found in the report for the R102 project which is still in active development at the time of publication, and some limitations to the linear referencing system embedded in the map matcher still exist.

Matching to a master network takes advantage of the invariants in the geographic objects. Both datasets utilize a ‘hidden network’ such that any event or traffic object will be aligned to geometry which exists in a (hidden) master lookup table. Despite these networks being unavailable, it is possible to partially reconstruct them based on duplicated geometry present in API responses, creating an ad hoc network. These ad hoc networks are created iteratively by requesting event and traffic data and isolating the unique geometry. Based on this principle of an ad hoc network, a framework to transform the two datasets to a common spatial network was developed.

Because neither dataset’s geometry is complete, alignments of traffic data to event data, or vice versa will be dependent on the transient state of their respective ad hoc networks. This introduces a potential error source and poor performance in areas with complex geometry but low data coverage. The limited availability of the geometry in both datasets is overcome by referencing a third geometric dataset with complete coverage. The ad hoc networks are snapped to a master network resulting in consistent matches regardless of the ad hoc networks’ current states. This enables spatial intersection operations by determining if any geometry in the master network is shared between two sample traffic and event objects. The result is visualised in Figure 4.3**,** where overlapping geometry from the event and traffic datasets are mapped to a master network. Any

geometry in the master network shared by the traffic and event data is considered an overlap region, and the geometry is considered similar enough that traffic data will capture events at point or road segment locations. In the case of this project, the HERE s181r2 network was used as the common spatial system. All subsequent analysis was performed on traffic data which strictly overlaps the region in which the roadwork event occurred.

Figure 4.3 Traffic data map matching diagram



4.1.8 ANOMALY METHODS ASSESSMENT

Before any attempt at developing an anomaly detection method, the suitability of putative techniques was assessed in the context of the test dataset described in Section 4.1.

Three key factors of the data must be considered when choosing an appropriate anomaly detection method:

1. Minimal roadwork data to serve as ground truth
2. Heterogeneity in the event and traffic data
3. The relatively small window of logged events.

The majority of machine learning classification techniques (e.g. SVM, feedforward neutral network (FNN), regression tree, nearest neighbour) require a large quantity of truth data to allow them to 'learn' the patterns and classify unknown data of the same type. In the context of this project, a truth dataset would be a collection of validated traffic probe recordings at several roadwork sites. The exact timing, nature and impact on the routing around the event would also be recorded. This dataset would ideally consist of hundreds, if not tens of thousands of individual events. Since an accurate truth dataset is neither available, nor feasible to produce, many of these supervised techniques are not suitable for this project. It may be possible to use a subset of the test data itself as training data except this is highly dependent on the quality of the sources. Also, without comparison to a more reliable dataset it would be unwise to use unvalidated truth data in any initial attempts. This caveat excludes most supervised learning approaches from the available options.

Heuristic methods can perform well without a truth dataset, but they require high-quality consistent speed data to set up appropriate triggers, as well as a strong understanding of the dynamics of the system. The presence of unknowns, for example, How much of a change in traffic flow should be expected from a typical roadwork event? Is traffic impact dependent on volume and location? prevents the application of a heuristic method. As an outcome of this project some basic principles of high order traffic behaviour may be better understood, placing subsequent investigations in a better position to use heuristics method.

K-means clustering makes inferences only from input datasets without referring to known outcomes i.e. it is an unsupervised approach. This makes k-means clustering particularly suitable for anomaly detection without ground truth data. In addition, the objective of k-means is just to group similar data points together, so insufficient data has less impact on this method than other alternatives. K-means clustering can perform well in low dimension datasets indicating suitability for this project.

Finally, anomaly detection is conceptually different to feature classification when considering the current problem. When detecting anomalies, the characteristics, key features and dynamics are not entirely understood. Essentially an anomaly is a pattern which is different, but it may not be clear how or why that pattern occurred, the meta-structure of the anomaly classification is therefore poorly understood. As understanding of how a roadwork event manifests as a traffic anomaly is unclear, the method should avoid overfitting or attempting to define the entire domain, leading the investigation away from strict classifiers such as auto-encoders, or unsupervised neural networks which will inherently have high false negative prediction rates. Conveniently, approaching the detection of anomalies by identifying cluster outliers through k-means allows explicitly defining the dynamics of the system to be avoided.

Table 4.3 provides a summary of the suitability of these techniques.

Table 4.3: Summary of anomaly detection techniques

Technique	Functioning with insufficient truth roadwork data	Functioning with insufficient speed data
Support-vector machine (SVM)	✗	✔
Regression tree	✗	✗
Nearest neighbour	✗	✗
Feedforward neutral network (FNN)	✗	✗
Heuristic methods	✓	✗
K-means clustering	✓	✓

4.2 K-MEANS CLUSTERING ANALYSIS

K-means clustering is a method or algorithm used for unsupervised data clustering or partitioning which can be used for classification. It attempts to minimise the sum-of-squares for the number of groups specified using random start points and iterating until relevant 'minimum' criteria are met. The general form of the equation is shown in equation 2 (Weisstein 2020).

$$J = \sum_{j=1}^{K} \sum_{n \in S_j} |x_n - \mu_j|^2 \quad \textbf{2}$$

where

J	=	Sum-of-squares result
x_n	=	Vector of n^{th} data point
μ_j	=	Geometric centroid of data points in S_j
S_j	=	The j^{th} subset of data points
N_j	=	Data points in the j^{th} subset
K	=	Number of subsets

K-means analysis is often demonstrated on x-y datasets however it can cluster using more than two dimensions and can be adapted specifically to be used with longitudinal data to cluster trends. Longitudinal clustering considers the pattern of traffic in a segment of time, as opposed to an instantaneous point in time. As traffic follows predictable patterns over a 24-hour period, the longitudinal approach appeared to fit with the use-case however implementation was unsuccessful.

4.2.1 K-MEANS CLUSTERING IMPLEMENTATION

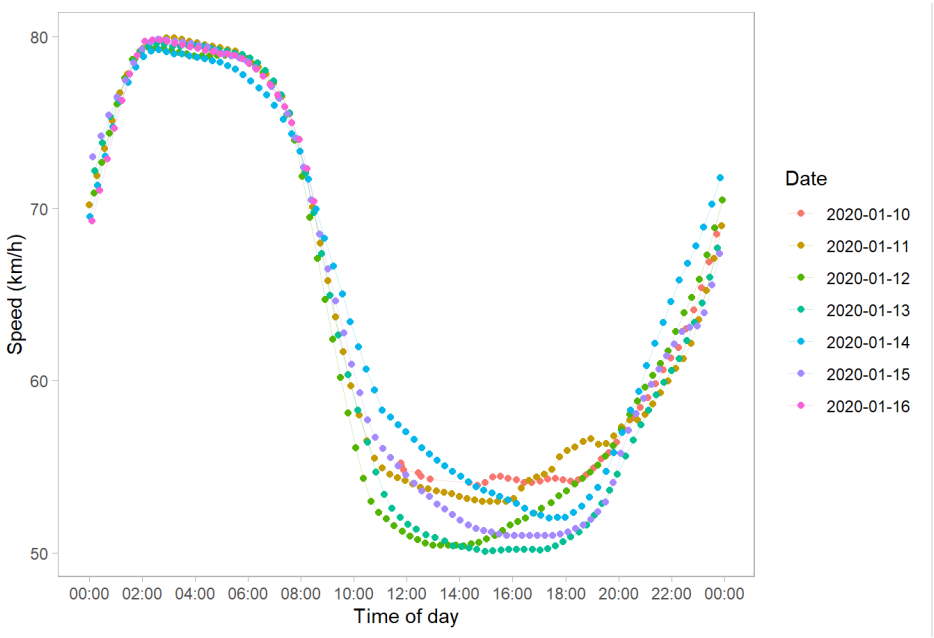
While the clustering algorithm could be run on the entire dataset; it would not provide good information for anomaly detection. The concept for anomaly detection was to overlay the daily traffic pattern and assess the data in 1-hour bins across each 24-hour period.

The steps required to conduct anomaly detection were:

1. Smooth all input data using a 10 to 20-point rolling mean.
2. Overlay the data by time of day to show daily trends.
3. Cut data into 1-hour bins, aggregate by time bin and date, and perform clustering on each 1-hour period to compare the days, this step provides additional smoothing.
4. Use specific criteria for the clusters to determine if specific clusters are flagged as anomalies.

Figure 4.4 shows the daily trends for case study 1.

Figure 4.4 Traffic data aligned by time of day



The k-means clustering process was conducted in the R statistical programming language using the 'kmeans' function from the stats v3.6.2 library (RDocumentation 2020). The additional specified parameters for the functions were:

- centers = 2
  - The specifies clustering into 2 groups
- nstart = 10
  - number of random starts for the iterative process to avoid local minimum convergence

The process for anomaly detection was applied to the clusters in 1 dimension and always returned two clusters for each time bin, regardless of how closely the points are already aligned. To determine if one cluster should be flagged as an anomaly additional logic was required.

A k-means cluster of data points in any 1-hour bin would be flagged as an anomaly if:

1. the difference between the cluster centres was greater than 10%
2. the cluster had the least number of data points.

The logic was quite coarse, but also reasonably robust for the data quality sourced for this project. With better data and potentially more baseline information additional measures that compare the compactness and within-ness of the data (descriptors of k-mean clusters) could be used to determine if the clusters are sufficiently distinct to be classed as an anomaly.

4.2.2 SIMULATING REAL-TIME LOGGING

To simulate real-time anomaly detection a process was set-up that tested the individual traffic data points sequentially and performed the k-means clustering process for each new observation. For this process to work a baseline data set was required, which was set as the first half of the available data. Therefore, the real-time simulation started with the first point in the second half of the data set for a given location. Two slightly different clustering processes were used:

1. Using one-hour aggregate points as per the previous example but performed sequentially, adding an additional point to the dataset each time it ran to assess if that point was an anomaly.

2. The same sequential process but using disaggregate data for the one-hour period in an attempt to provide earlier detection.

4.3 HISTORICAL ANALYSIS

Using historical data as a baseline for typical traffic conditions was considered as one approach to detecting roadworks events however easy (API style) access to this data was not possible so it was not feasible for this project. It is worth noting however that this type of access may be available soon if the R100 project (TMR traffic data access portal) is extended and implemented.

It was expected that the HERE historic data would include more post-processing than the HERE Traffic data so it was important to understand how large these differences were for future project work. Therefore, a comparison of the traffic data and historic data was conducted to determine if this was feasible.

The historic data could not be selected using TMC codes as the download portal did not recognise the TMC codes provided in the HERE Traffic API, therefore the Link ID values identified in Section 4.1.7 were used to download a month of data overlapping with the logged period.

To compare the data both raw and smoothed speed data was plotted together. Additional metrics were considered to compare the speed data however the large differences were obvious on visual inspection, so this was deferred.

4.4 CASE STUDIES

Using data collected in Section 3, four case studies were highlighted to assess the performance of the anomaly detection method. These case studies are four individual roadwork events and the associated traffic data. The events were selected to cover several different scenarios including two likely roadworks anomalies and two locations without anomalies. It also covered two urban and two rural locations.

Event 1 – Brisbane motorway (Figure 4.5)

- Roadworks ID: 263295
- Class: Planned roadworks
- All lanes affected in both directions
- Date: 2020-01-13 20:00 to 2020-01-14 05:00
- Description: Southern Cross Way VMS upgrade project works. The project involves the removal of existing and installation of new variable message (VMS) boards on Southern Cross Way northbound carriageway at Banyo and southbound carriageway at Doomben. Traffic holds and releases up to 15 minutes
- Historical link length: 0.122 km
- Realtime link length: 0.976 km.

Figure 4.5 Brisbane roadworks event 263295 and one site from 266580

Event 2 – Brisbane motorway (Figure 4.5 and Figure 4.6)

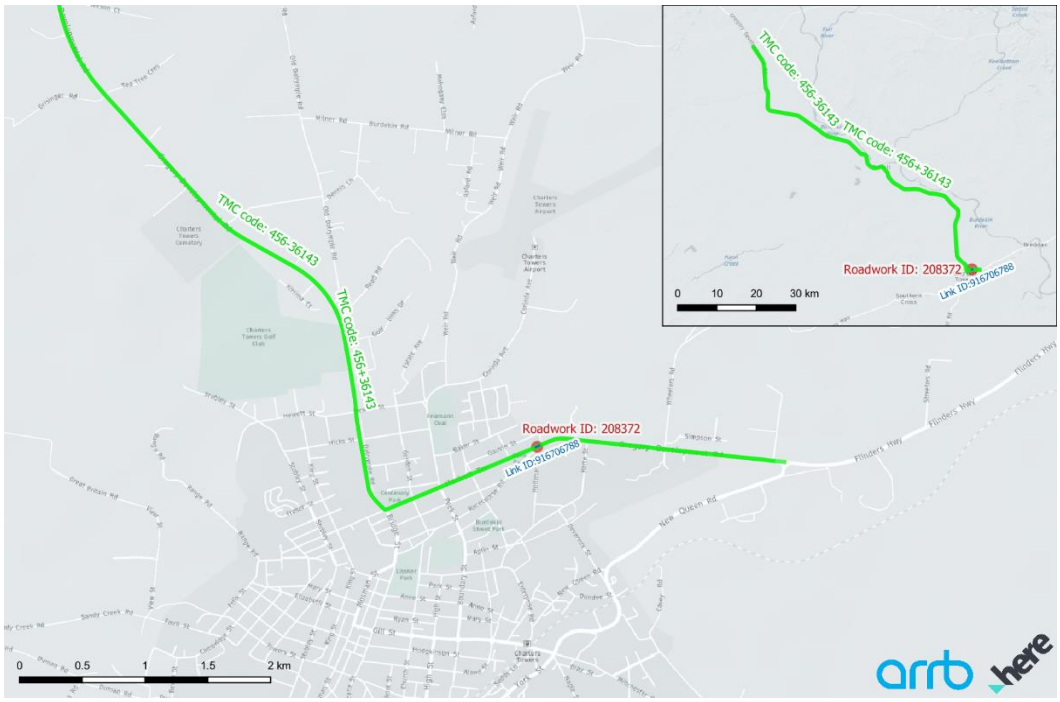
- Roadworks ID: 266580
- Class: Planned roadworks
- Both directions
- Date: 2020-01-15 22:00 to 2020-01-16 03:00
- Description: Short term 15 minute stoppages to remove old VMS board and install new board
- Average Historical link length: 0.674 km
- Average Realtime link length: 0.778 km.

Figure 4.6 One site of roadworks event 266580 in Brisbane

Event 3 – Road in Charters Towers (Figure 4.7)

- Roadworks ID: 208372
- Class: Planned roadworks
- Both directions
- Date: 2019-09-01 06:00 to 2020-07-31 18:00
- Description: trenching out to wells location on main road and install four pits, reinstate and resurface
- Recurrence: Monday to Friday, 6 am to 6 pm
- Average Historical link length: 0.029 km
- Average Realtime link length: 0.778 km.

Figure 4.7 Map of roadworks event 208372 in Charters Towers



Event 4 – Section of road through Chinchilla (Figure 4.8)

- Roadworks ID: 229886
- Class: Planned roadworks
- All lanes affected in both directions
- Date: 2019-07-25 06:30 to 2020-02-13 18:00
- Description: Between Glasson Street and Heeney Street. Impacts during the Holiday Period speed reduced to 40 km/h from Friday 20 December 2019 through to Sunday 5 January 2020. Works are weather permitting. Heavy vehicle operators are advised that width and/or mass restrictions may apply at roadworks
- Recurrence: Monday to Friday, 6:30 am to 6 pm
- Average Historical link length: 0.122 km
- Average Realtime link length: 97.499 km.

Figure 4.8 Map of roadworks event 229886 in Chinchilla

4.4.1 EXAMINATION OF CASE STUDIES

Each of these case studies were the subject of further analysis and evaulation of the anomaly detection methods. The results of the examination of data for these events is presented in Section 5.2.

5 RESULTS

5.1 DATA QUALITY/LOGGING OUTCOMES

This section considers the quality of the dataset in the context of a truth dataset for use in machine learning methods, and general characteristics which impact the approach. In total 297 roadworks events were logged over 14 days. Events were updated on average 2.8 times throughout this period, with the maximum number of updates being 12. An average of 13.2 road segments was recorded for each event. The lengths of each road segment obtained from the HERE Traffic API were significantly longer than the both the roadwork event and the length of links used in the spatial intersection as seen in Figure 5.1 and Figure 5.2.

Figure 5.1 HERE Traffic API geometry lengths

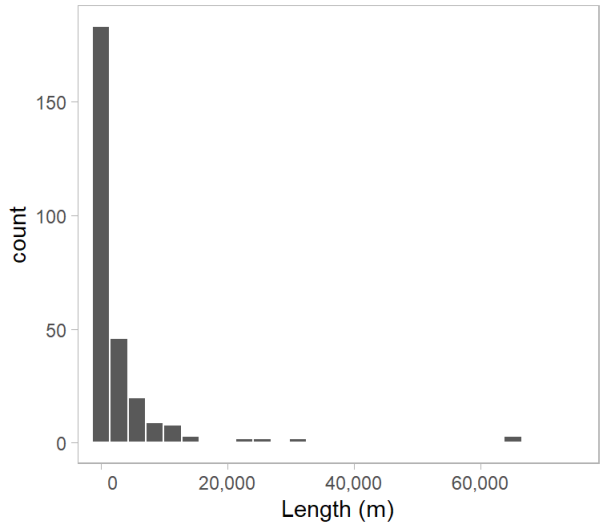
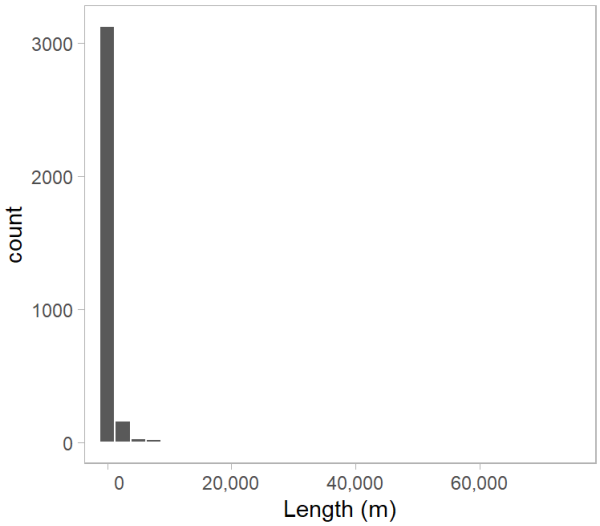


Figure 5.2 HERE link geometry lengths



The average duration of a roadworks event was 212 days (Figure 5.3) of the 297 unique roadworks events, 51% contained no computer readable metadata about the timing or scheduling of the event beyond a 6 month or greater time frame. Any further information was stored in an open text field and could not be analysed. Leveraging the impact type or impact subtype feature of roadwork events for SVM classification was considered but not integrated into the analysis. This was because 77% of roadwork events have the impact type 'Lanes affected', (Table 5.1) which does not have enough discriminatory potential for SVM classification, leaving the bulk of the events unable to be sub-classified.

Figure 5.3 Duration of roadworks events from QLDTraffic

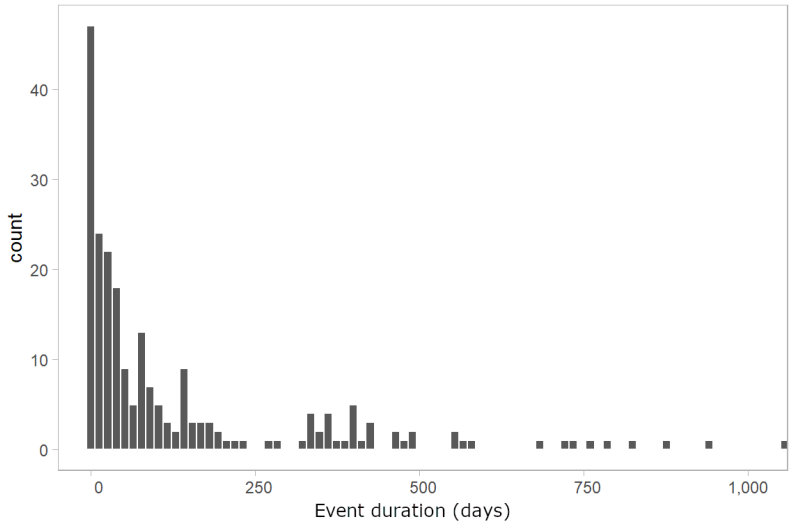


Table 5.1: Proportion of impact type fields in QLDTraffic data

Impact type	Percentage of events
Lanes blocked	1.5%
No blockage	6.3%
N/A	4.3%
Lanes affected	77%
Closures	10.4%

Much of the statistical analysis re-affirms qualitative statements about the two datasets. When directly addressing the aim of creating a truth dataset, several limitations of the logged data become apparent.

Primarily there was a lack of consistent and detailed features which were parse-able via automated methods.

Secondly, the time scale and detail of the schedule for events often spanned multiple months with no detail about days of the week or time of day embedded in the metadata. Roadwork events lasting for several months to a year essentially become semi-permanent and the ability to detect measurable changes on yearly time scales was outside of the scope of a real time anomaly detection system.

Finally, a problem of scale also occurred when map matching TMC geometry with event geometry. The lengths of road over which traffic data were averaged were so long that highly localised reductions in speed were likely to be undetectable when averaged over many kilometres. Roadwork events typically do not span multiple kilometres of a linear road section, but traffic data in all case studies was reported over links with a minimum length of 0.7 km and a maximum of 97 km. As traffic data is reported in an aggregate fashion along a continuous section of road which can be many kilometres in length, probe data within the actual event zone is being reported alongside probe data on road sections before and after the event. This effectively reduces the strength of a traffic signal any event may be generating.

5.2 K-MEANS ANALYSIS OUTCOMES

As discussed previously, there is no accurate ground truth data for this analysis, so results are largely subjective. However, a visual assessment of the k results showed it to be able to detect anomalies likely to be roadworks in the traffic data.

5.2.1 ANOMALY DETECTION – CASE STUDIES

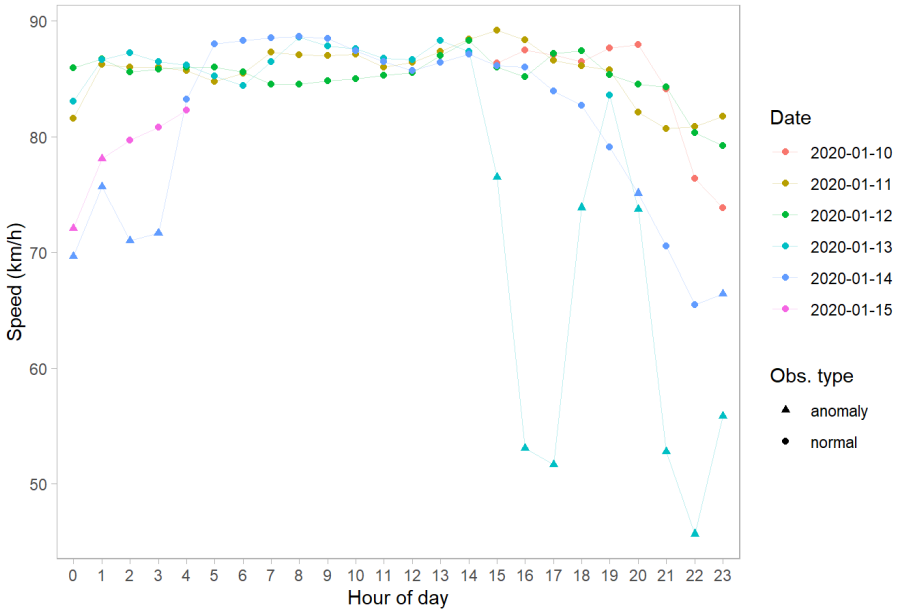
Anomaly detection in each case study showed promising results. For each event, a period of at least six days was able to be logged within the two-week data collection window. This is due to most events occurring after the start of the logging period, and no data could be collected until after the event was recorded in the QLDTraffic database.

Event 1 (Brisbane motorway) – Potential roadworks event

Analysis of this event showed good anomaly detection with potential for further improvement by fine tuning detection logic. The data showed two drops in speed at approximately 2 pm and 9 pm on 13 January 2020. The 2 pm drop fell just outside the roadworks window, but it was correctly identified as an anomaly as shown in **Error! Reference source not found.**.

There were also points flagged as anomalies in the afternoon on 14 January and in the morning on 15 January. These dates also fell outside the roadworks period and a visual assessment suggests they were a borderline case for the anomaly flag. Adjustments to the logic may prevent some of these points from being flagged if required.

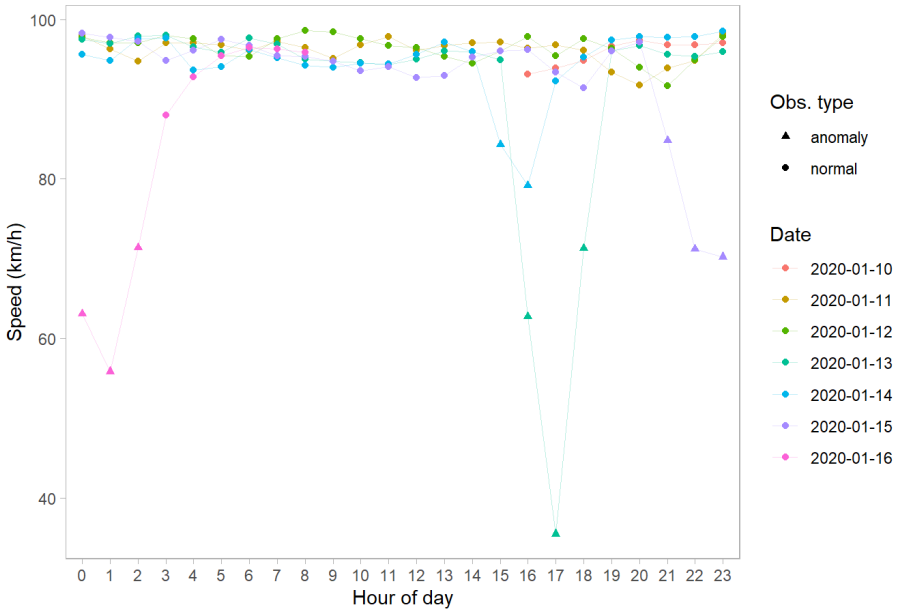
Figure 5.4 Anomaly detection for roadworks event 263295 – 1-hour bins



Event 2 (Brisbane motorway)– Roadworks event plus additional anomaly

This location provided very consistent baseline speed data allowing anomalies to be clearly identified and the results matched with expectation. However, the only roadworks event logged started from 15 January at 10 pm till 16 January at 3 am. The anomalies on 13 and 14 January were outside the roadworks window but were clearly variations from the typical traffic pattern (**Error! Reference source not found.**).

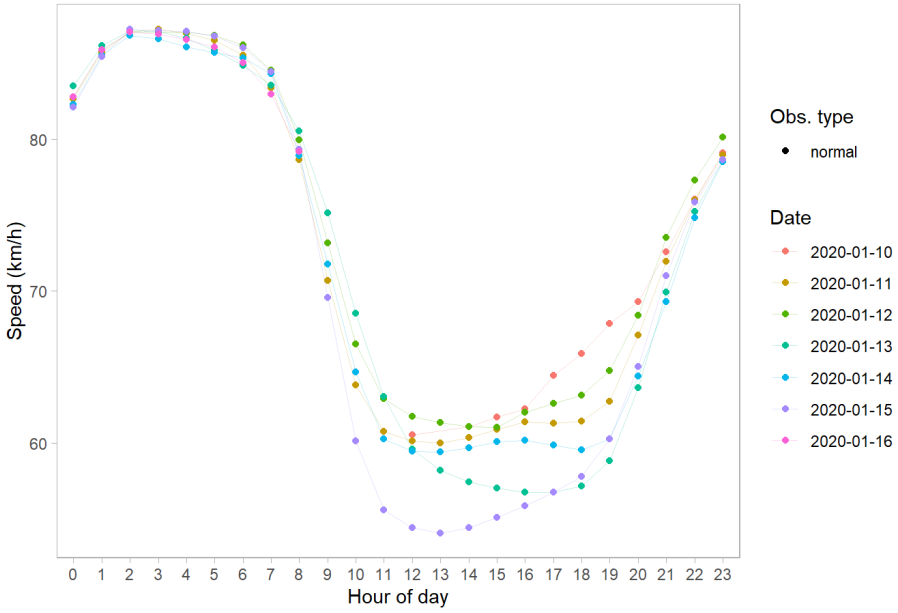
Figure 5.5 Anomaly detection for roadworks event 266580 – 1 hour bins



Event 3 (Charters Towers)– No anomalies

This data fell within the roadwork period however the period was 11 months, so no data was logged for the period outside of the two-week period. In this case no anomalies were detected as all speeds followed a similar pattern for each day the data was logged. It is not known if this was because the roadworks was similar each day or if there were no roadworks. It is worth noting that the speed limit at the location is 60 km/h however the geometry for the traffic speed information extends into a 100 km/h zone (Figure 5.6).

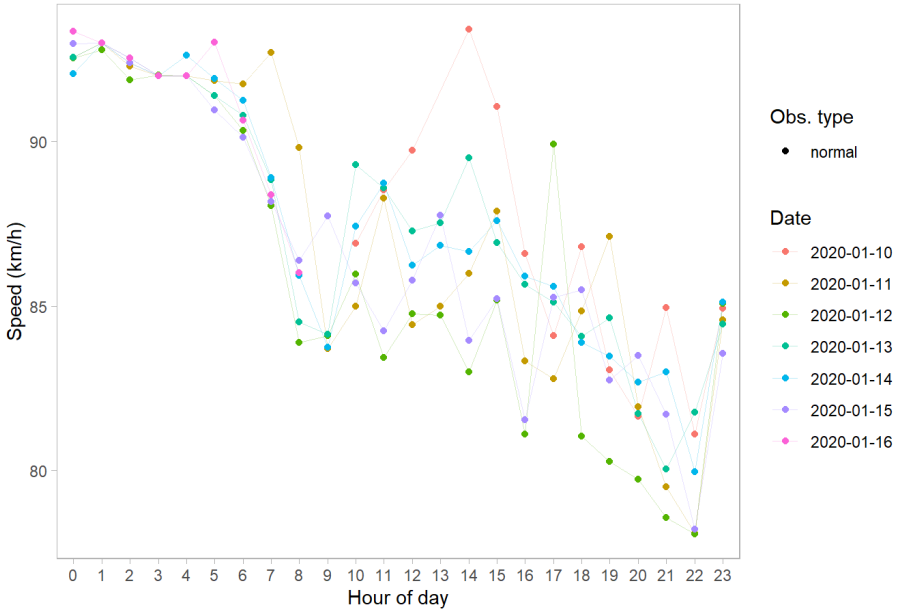
Figure 5.6 Anomaly detection for roadworks event 208372 – 1 hour bins



Event 4 (Chinchilla) – No anomalies

The data in Figure 5.7 fell within the date range for the works, however the text description of the event listed 5 January as the end date indicating the event is suitable to analyse as a null event. All days in this logged data followed a similar pattern. It appears much noisier than the previous example however that is in part due to different y-axis scales. As with the previous case, no roadwork events were detected which fits with the visual assessment as any outliers would be within 10% of the mean speed. The results from this case study may indicate false positives are unlikely when under consistent traffic conditions.

Figure 5.7 Anomaly detection for roadworks event 229886 – 1 hour bins



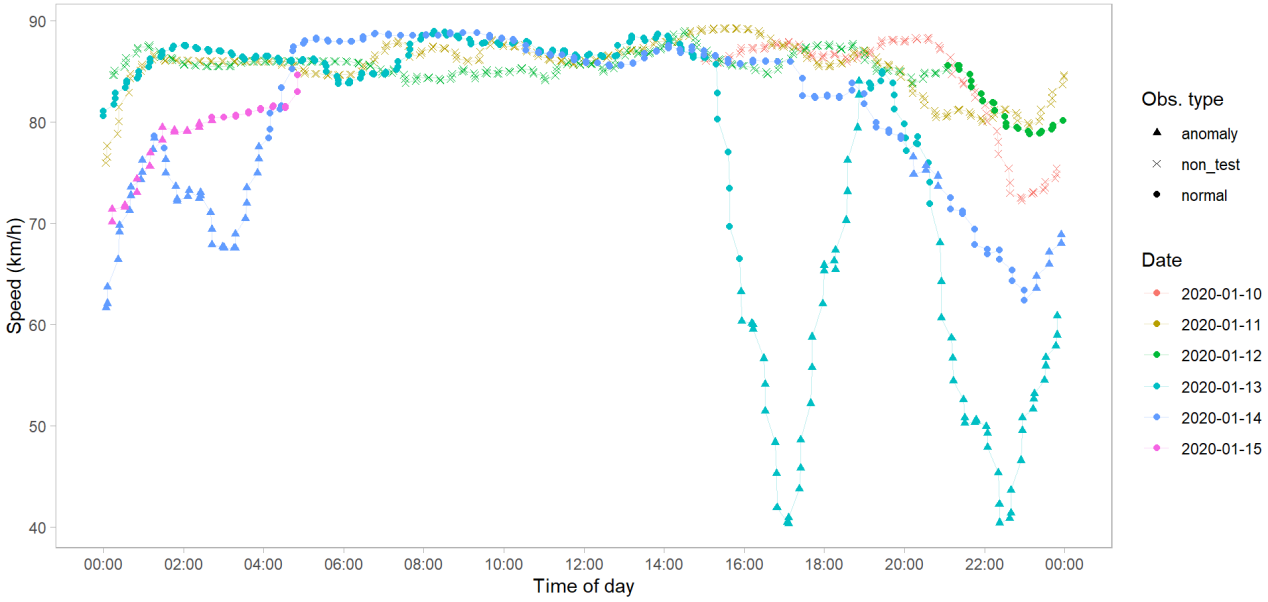
5.2.2 ANOMALY DETECTION – SIMULATED REAL-TIME DATA

As outlined in section 4.2.2, the simulated real-time method runs the k-means clustering anomaly detection on each new data point sequentially to determine if the current data point is an anomaly. Two variations of this method were tested, one that aggregated the data in each 1-hour bin and one that did not. The algorithm was applied to the roadworks events defined in Section 4.4 to determine its effectiveness in detecting anomalies.

Event 1 (Brisbane Motorway) – Potential roadworks event

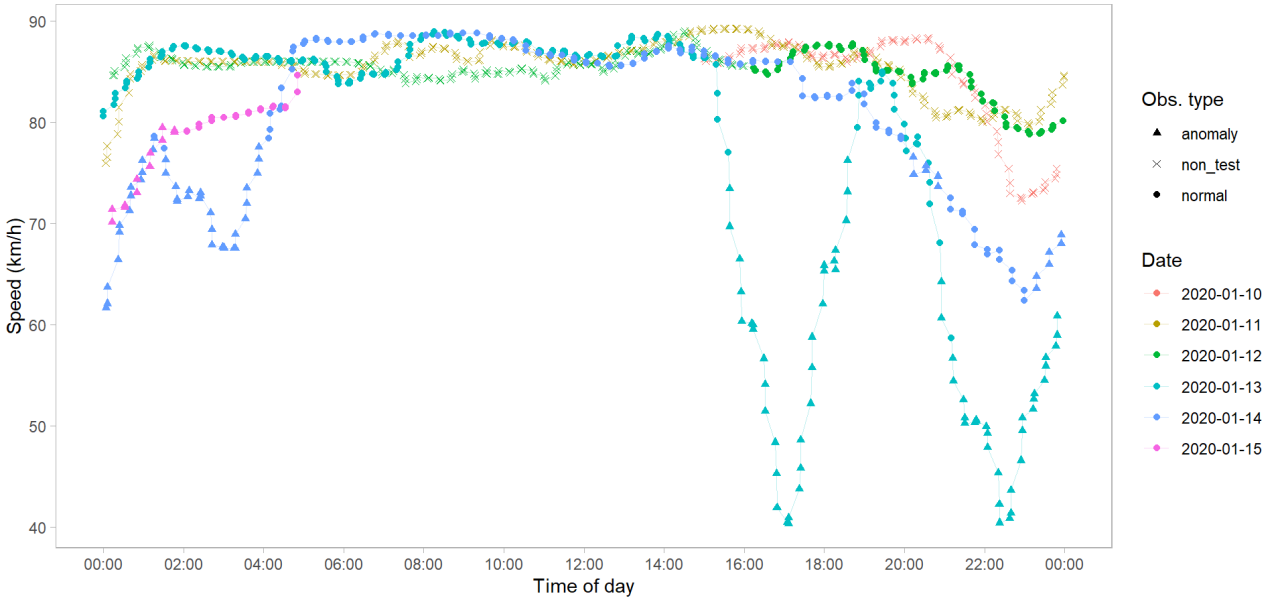
The simulated real-time detection correctly flagged the majority of data points correctly. This method is more susceptible to noise than the full dataset aggregation so additional baseline data would be beneficial. Smoothing and logic for flagging anomalies could also be further refined to detect them earlier.

Figure 5.8 Simulated real-time detection using 1 hour bin aggregation



The sequential 1 hour bin method was slow to detect the initial speed drop on 13 January at approximately 16:00 (4 pm) due to averaging of data points within the 1 hour bin (Figure 5.8). The alternate disaggregated approach is shown in Figure 5.9 which demonstrates that the same anomaly is flagged 30 minutes earlier. It should be noted that this was not true for the second anomaly on 13 January at approximately 20:00 (8 pm) which was detected at the same time for both methods however the disaggregated method also mis-classified a point at approximately 9 pm on 13 January.

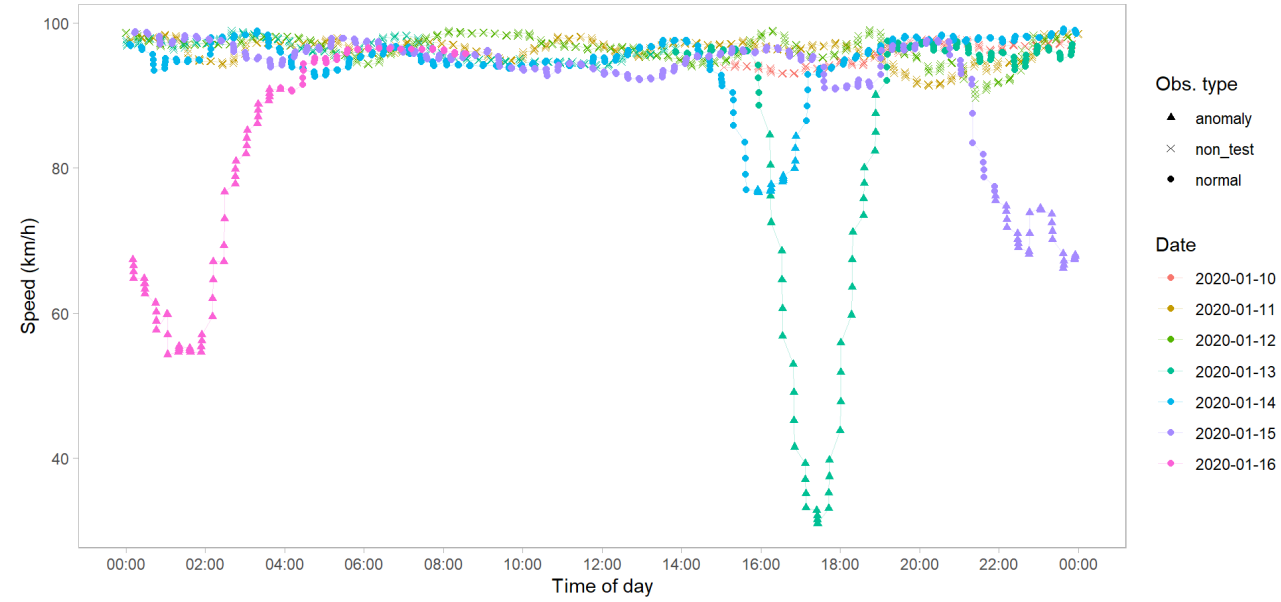
Figure 5.9 Simulated real-time detection without aggregation



Event 2 (Brisbane Motorway) – Roadworks event plus additional anomaly

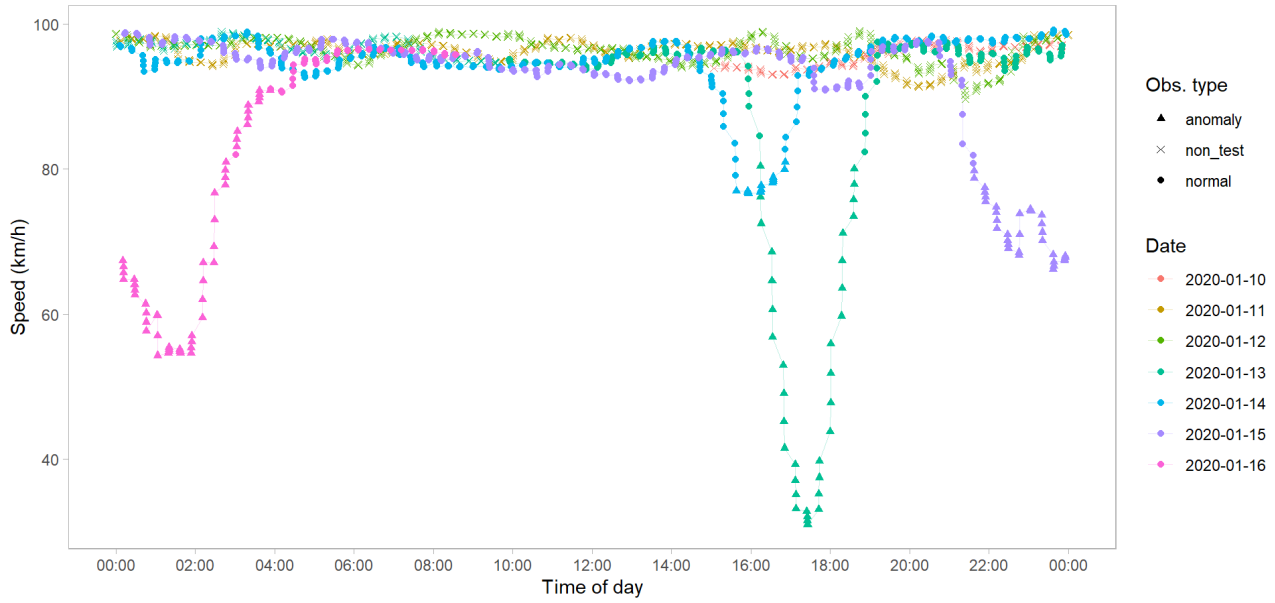
The sequential anomaly detection also functioned well in this case however it was slower to detect an anomaly on 14 January than it was on 13 January potentially due to the earlier anomaly causing additional noise in the data and weakening the baseline data (Figure 5.10).

Figure 5.10 Simulated real-time detection using 1 hour bin aggregation



In this case the anomalies were flagged much the same by the 1 hour binned method and the disaggregate method (Figure 5.11). The disaggregate method generally flagged the points earlier except on 13 January at approximately 4 pm.

Figure 5.11 Simulated real-time detection without aggregation



Event 3 (Charters Towers) – No anomalies

The sequential process did not detect any anomalies with the 1 hour aggregate method (Figure 5.12) or the disaggregate method (Figure 5.13) due to very consistent traffic patterns across the days at this site.

Figure 5.12 Simulated real-time detection using 1 hour bin aggregation

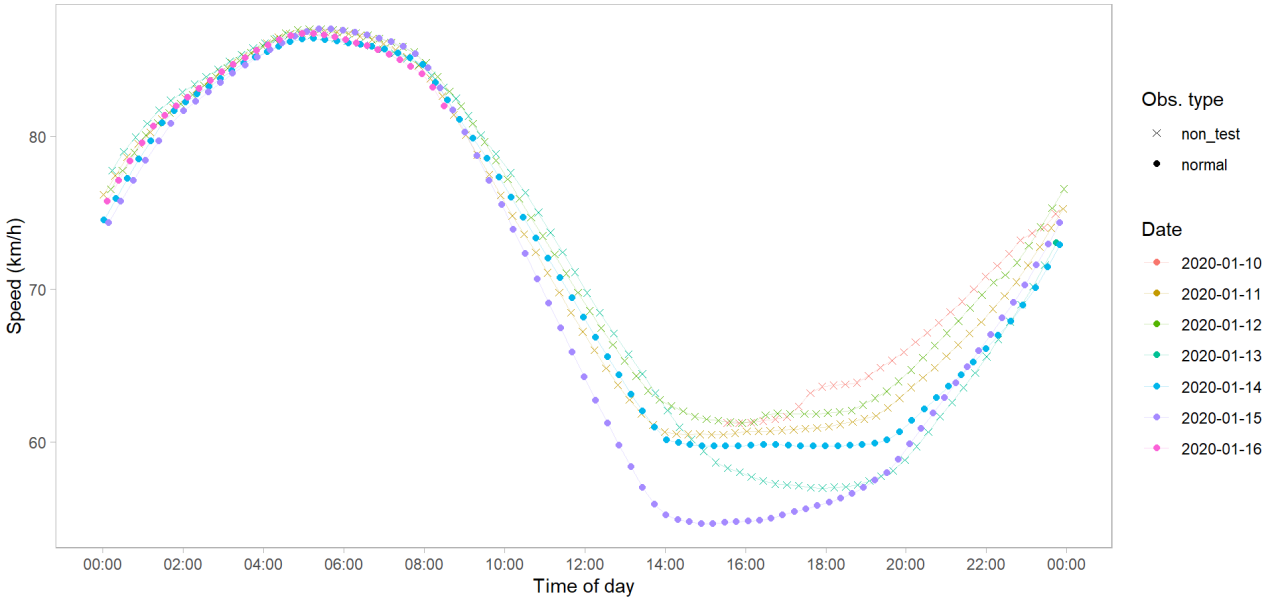
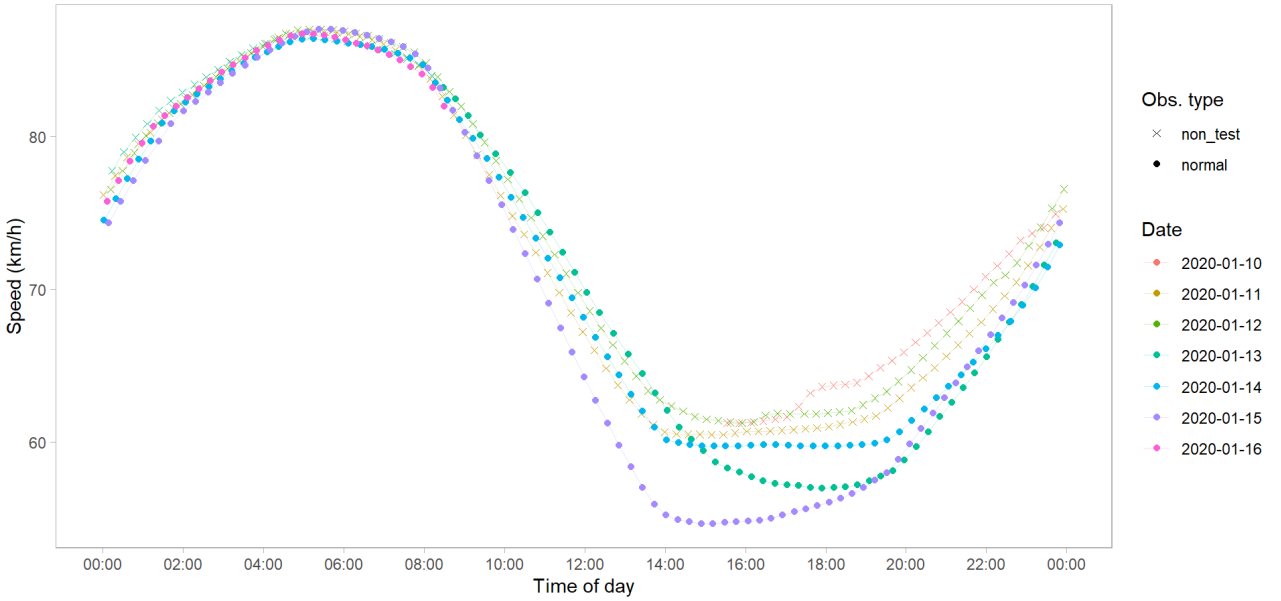


Figure 5.13 Simulated real-time detection without aggregation



Event 4 (Chinchilla) – No anomalies

Without the consistent 1 hour bins the data at this site is hard to plot, however even with the noise no anomalies were detected, likely due to outliers being within 10% of the mean. Both the 1 hour method (Figure 5.14) and the disaggregate method (Figure 5.15) performed in the same way.

Figure 5.14 Simulated real-time detection using 1 hour bin aggregation

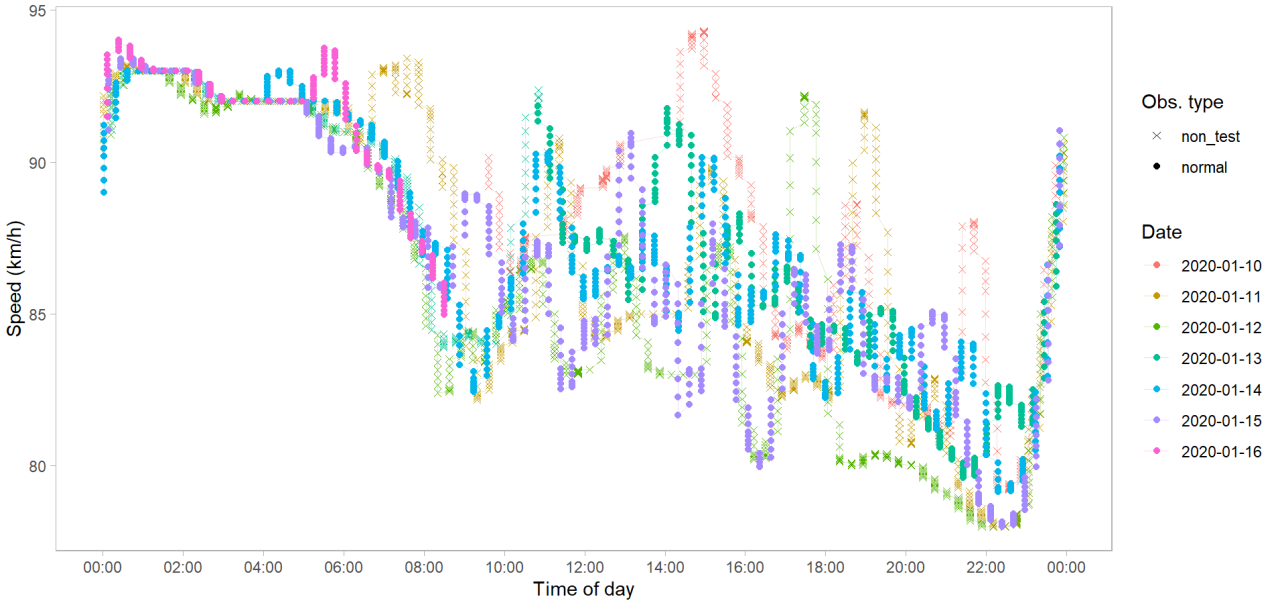
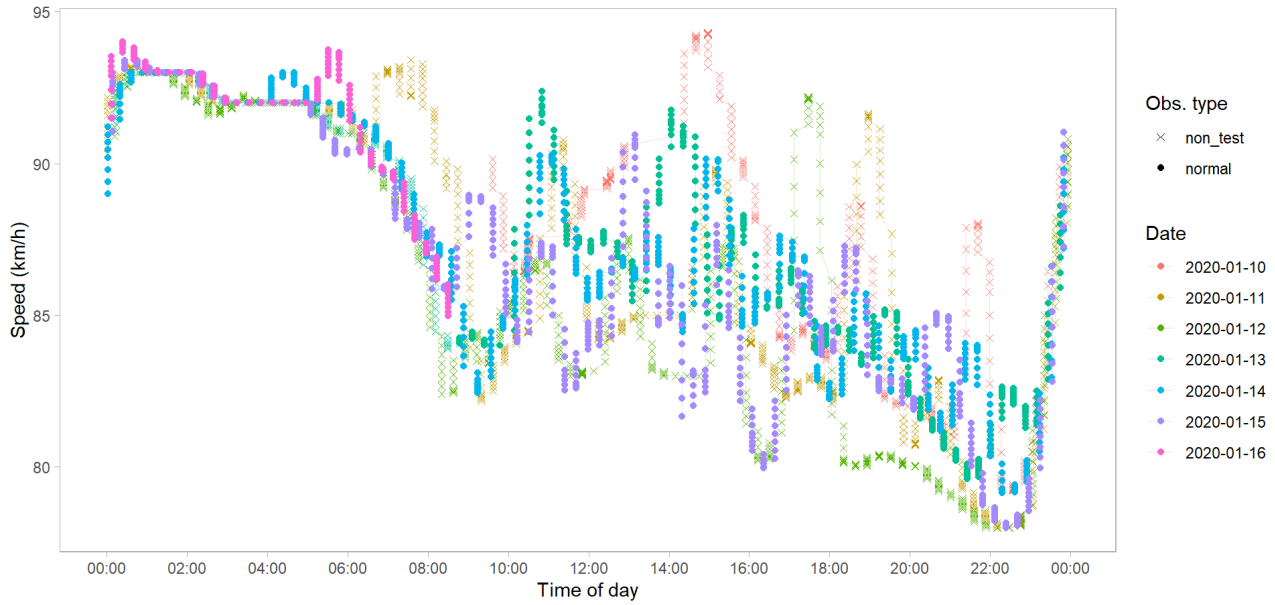


Figure 5.15 Simulated real-time detection without aggregation



The simulated real-time method returned good results and was able to identify anomalies in a similar way to the full dataset method with some minor differences. In some cases, it was able to pick up the anomalies better and in others there were some false positives, potentially due to less smoothing in the time bin early in the time period.

5.3 HISTORICAL COMPARISON DATA

The historical comparison provided an interesting insight into the differences between the logged traffic data and the historical data download. The two key results to come from the comparison were:

1. Realtime traffic data differed from the historic data.
2. There was a large amount of gap fill in the historic data for some locations, therefore, it is likely that the real-time traffic data also contained a lot of gap fill data.

Because of these observations the focus of the comparison was on understanding the differences rather than quantifying them.

5.3.1 SPEED DATA COMPARISON

The speed data comparison revealed large differences between data from the HERE Traffic API and the historic download. Figure 5.16 shows data for a roadworks event in Chinchilla where the mean speeds for the two data sources was approximately 40 km/h different and the historic data had a far more regular daily cycle. Figure 5.17 shows speed data for a roadworks event on the Gateway Motorway where the mean speeds for the two sources was similar however the real-time traffic data source had less regular daily cycles and some significant peaks which were not apparent in the historic data.

Figure 5.16 Example of historic and real-time data with different mean speeds (Chinchilla)

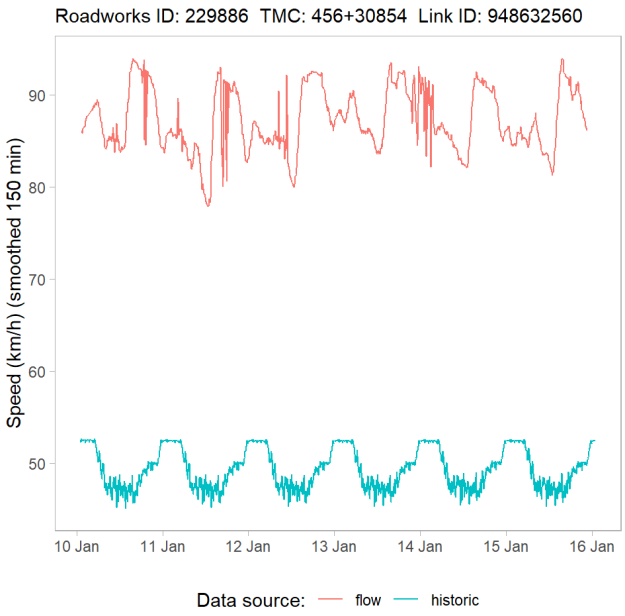
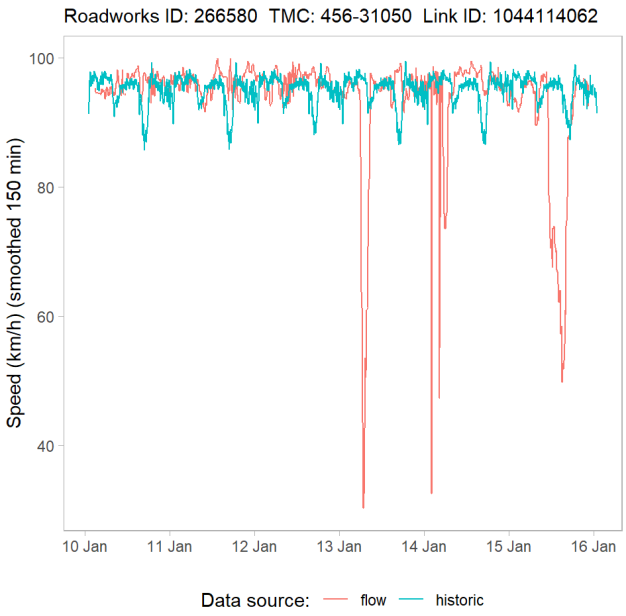


Figure 5.17 Example of historic and real-time data with similar mean speeds (Gateway Motorway)



These results suggest that the two sources were not measuring the same thing so further attempts to quantify the differences were not conducted. The focus of the analysis changed to understanding why the data sources were different.

5.3.2 GAP FILL

Early in this project it was noted that the HERE Traffic API does not flag when a data point is taken from gap fill pattern data, however data download from the historic portal does include this information.

The HERE historical data could be used to determine the level of gap fill applied, although:

- It is not clear if the real-time data is processed in the same way as historical traffic data.
- The flow TMC links are not compatible with the historical traffic service.

5.3.3 GEOMETRY

The investigation into the difference of the flow and historic data revealed that there were some major differences in geometry between the datasets. The HERE Traffic data is only available as TMC geometries which were typically much larger than the HERE link geometries (Figure 4.5 to Figure 4.8).

One example of these geometry differences is the roadworks in Charters Towers which was a point location on a street on the outskirts of town with a speed limit of 60 km/h (Figure 4.7). The HERE link matching this location is 30 m long and is contained entirely within the roadworks location. However, the matching TMC geometry was approximately 97 km long with only 4% as a 60 km/h zone and 94% of the geometry being a 100 km/h zone. In the link lengths in Figure 5.17, which are both approximately 700 metres, the historical data is comparable to the real time data. Differences can be explained in part by aggregation over very long link lengths.

The data comparison may have been closer if link geometries were matched to the full length of the TMC geometries rather than to the roadworks geometries. However, even with more data they may not match without the same aggregation method.

6 DISCUSSION

6.1 EVALUATING ANOMALY PREDICTION PERFORMANCE

As outlined in previous sections, it is not possible to directly evaluate the performance of the anomaly detection algorithm. In some cases, it is readily apparent that a drop in traffic flow was detectable within the scheduled event time. In other cases, the results were unclear. With no distinction between simulated traffic flow and true flow it is possible that the limitations in anomaly detection can be overcome simply by increasing the coverage of traffic data. It may be reasonable to expect that as the number of datapoints increase, the signal of anomalous traffic patterns begins to overcome stochastic variation. This assumes that all events have a detectable impact on traffic, which was not confirmed during this investigation. Partitioning the data by event type or subtype may reveal hidden trends which can explain differences in the detection performance. It does not seem to be possible to detect remote events based on a combination of geographic inequality and data coverage with these data sets. Fortunately, the techniques of data aggregation and analysis used in this project are independent of the data sources themselves and it is likely that a steady increase in data coverage and quality will passively improve the detection system over time.

6.2 ALTERNATIVE TECHNIQUES

Longitudinal k-means clustering could provide an alternative to the methods tested in this report. This technique could either be used to develop a baseline for testing anomalies by using traffic patterns across a year or by using trajectories rather than individual points for anomalies. Using trajectories would utilise a point position relative to adjacent points, however this may also make it slower to trigger an anomaly.

6.3 DATASET CHALLENGES

The project consists of two main tasks: creation of a test dataset; and the subsequent anomaly detection method applied to the dataset. Initially, creating the test dataset was believed to be trivial, but accurate spatial intersections between event and traffic data required more sophisticated methods than expected. Because these modules are interdependent, many of the limitations of the K-means analysis are limitations of the data itself, and the ability to distinguish between poor data quality and poor anomaly detection is also compromised by the lack of truth or reference information about the nature of these events. However, there are some clear issues which stem from this dataset creation process.

The method of combining traffic and event data has shortcomings which are relevant to this project. Namely, no partial map matching was used in this project introducing error to events which span over road segments with complex topology. Data was only logged for 14 days; a greater time period may have yielded a subset of easily detectable anomalies. These shortcomings are overshadowed by the issues of scale in time and road segment present in the data conjugation. The logging period was many times smaller than the average length of a roadwork event, only a handful of case studies were eligible for analysis based on the reported duration of the event. As previously mentioned, roadwork events typically occupy a small section of the road, and the impact on traffic in low volume areas is highly localised to that area. The HERE traffic API reports aggregated traffic data on road segments longer by up to two orders of magnitude. The length of road segments was not foreseen to be an issue prior to commencement. The HERE Traffic API can report traffic data on two networks. The network which was intended to be used had much shorter road segments, however it was discovered that this network uses dynamic links which are unsuitable for this project which required high coverage. The shorter links only exist temporarily along a tiny minority of the road network.

While the data aggregation methods used in this report could not overcome the issues of quality, the map-matching based approach used to associate traffic data with roadworks events represents a successful application of these techniques. The map matching method enabled two datasets from completely different sources, which exist on separate networks to be compared. This procedure is generic and can be applied to

a wide range of data sources enabling unique insights not previously possible. Further exploration of the quality and reliability of traffic data would also rely on a similar spatial intersection of separate sources.

6.4 IMPROVING DATA QUALITY

Improving the roadworks truth datasets would potentially open additional anomaly detection options. This could take the form of a few case studies that can be closely monitored for road occupation times and potentially also alongside STREAMS loop detector data in the same region. It is likely that physical surveying of roadworks compliance would be necessary to create a truth dataset of satisfactory quality. More immediately, partitioning events to those with durations less than one week which are known to have a substantial impact on traffic flow may reveal a stronger relationship. This partitioning would not improve the fundamental method but may outline a sub-domain where unsupervised learning techniques would be viable.

The quantity of probes being logged by HERE and therefore the quality of the various data services is constantly improving. Therefore, the methods developed in this study may be able to be applied on more roads, in rural areas and on minor roads, than they could be at the time of this report. Improvements in speed data quality may also allow anomaly detection to be better tuned to trigger earlier and with less false positives.

7 CONCLUSIONS

Unlocking the potential of emerging big data can be challenging, with complexity arising at the acquisition, storage and combination stages. This project utilised a generic framework for logging time series data from dynamic sources. The novel ad hoc network conjugation step is particularly powerful when dealing with niche data sources that have limited ability to be transformed. Data aggregation steps included in this investigation are extendible to a wide range of business cases where data providers are unable to deliver data in a desirable geospatial structure.

Despite the inconclusive nature of the case studies, this project has inspired several new avenues of investigation for tracking event progression through traffic data. When considering roadwork events, the highest priority is probably categorisation of traffic impact based on the event's properties. A potential investigation into the nature of traffic disruption around events would require a localised approach, studying a handful of representative events and the traffic impact at a high spatial and temporal resolution. Improving the depth and quality of data for a handful of events can help uncover key determinants of traffic impact, providing a foundation for event sub-sampling. Even modest insight into factors which cause an event to have a high impact on traffic can greatly improve the distinguishing power of the k-means clustering method.

Some fundamental questions about the nature of event-based traffic impact need to be answered in a quantitative manner. In time commercial traffic data providers such as HERE will likely provide improved data coverage and quality, enabling a second wave of investigation to better answer these questions.

7.1 LIMITATIONS

A primary research finding was understanding the limitations and suitability of the data sources for roadwork event detection and matching.

Data acquisition from the Queensland Traffic and HERE Traffic API was possible but was a very complex process. The process required geospatial selections and map matching across three separate road geometries. Limitations with both the roadworks and traffic data were identified. The limitations of the roadworks data were:

- a low number of required fields
- additional useful data in non-machine-readable fields
- no confirmation if the works happened at that time.

These limitations were such that the dataset was prevented from serving as an accurate and reliable ground truth.

The limitations of the HERE traffic data were:

- The traffic reports aggregated data on road segments longer than required for this application.
- The logging period was many times smaller than the average length of a roadwork event.
- The quantity of sample traffic events analysed was too low to elucidate event categories.
- Very few case studies were eligible for analysis.

These limitations did not prevent an anomaly detection method from being developed however the results indicate that the data and methodare not suitable for a reliable system in production.

7.2 KEY FINDINGS

- The data sources were found unsuitable for immediate application due to coverage, quality and geometric structure issues.

- Data conjugation does show promise and will likely converge to a suitable level of accuracy in the medium-term future.
- The data aggregation method is powerful and extendible to other projects and the k-means method is similarly extendible to other event types.
- An improved understanding of traffic impacts resulting from the event types is required to further develop an anomaly detection algorithm.

7.3 RECOMMENDATIONS AND FUTURE WORK

This research highlighted the need for good quality and well attributed data for reliable automated pattern detection. Further attempts at automated roadworks detection can be aided by:

- improvements to the way in which QLDTraffic incident data is recorded and quality checked
- improvements in probe data coverage and quality with an expected increase in connected vehicles and devices
- using alternative real-time data sources with road segments of smaller length (non-TMC flow data).

REFERENCES

Gakis, E, Kehagias, D & Tzovaras, D 2014, 'Mining traffic data for road incidents detection', *International IEEE conference on intelligent transportation systems, 17th, 2014, Qingdao, China*, IEEE, Piscataway, NJ, USA, pp. 930-5.

HERE 2020, *Traffic API: flow*, webpage, HERE, Melbourne, Vic, viewed 4 May 2020, <https://developer.here.com/documentation/traffic/dev_guide/topics_v6.1/resource-parameters-flow.html>.

Houbraken, M, Logghe, S, Schreuder, M, Audenaert, P, Colle, D & Pickavet, M 2017, 'Automated incident detection using real-time floating car data', *Journal of Advanced Transportation*, vol. 2017.

Ozbayoglu, M, Kucukayan, G & Dogdu, E 2016, 'A real-time autonomous highway accident detection model based on big data processing and computational intelligence', *IEEE international conference on big data, 2016, Washington, DC*, IEEE, Piscataway, NJ, USA, pp. 1807-13.

Pietrobon, D, Lewis, AP & Heverly-Coulson, GS 2019, 'An algorithm for road closure detection from vehicle probe data', *ACM Transactions on Spatial Algorithms and Systems*, vol. 5, no. 2.

Queensland Government 2020, *QLDTraffic GeoJSON API*, webpage, Queensland Government, Brisbane, Qld, viewed 5 May 2020, <https://www.data.qld.gov.au/dataset/131940-traffic-and-travel-information-geojson-api>.

RDocumentation n.d., *kmeans*, webpage, RDocumentation, viewed 5 May 2020, <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/kmeans>.

Wang, H, Wen, H, Yi, F, Zhu, H & Sun, L 2017, 'Road traffic anomaly detection via collaborative path inference from GPS snippets', *Sensors*, vol. 17, no. 3.

Weisstein, EW 2020, *K-Means clustering algorithm*, webpage, Wolfram, Champaign, IL, USA, viewed 5 May 2020, <http://mathworld.wolfram.com/K-MeansClusteringAlgorithm.html>.