FINAL REPORT

PROJECT J3

AUGUST 2022

Identifying and Mitigating Congestion Onset (Phase 1)

George List, Ph.D., North Carolina State University
Billy Williams, Ph.D., North Carolina State University
Michael Hunter, Ph.D., Georgia Institute of Technology
Mohammed Hadi, Ph.D., Florida International University

STRIDE | Southeastern Transportation Research, Innovation, Development and Education Center
UF | Transportation Institute UNIVERSITY of FLORIDA

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

ACKNOWLEDGEMENT OF SPONSORSHIP AND STAKEHOLDERS

This work was sponsored by a contract from the Southeastern Transportation Research, Innovation, Development and Education Center (STRIDE), a Regional University Transportation Center sponsored by a grant from the U.S. Department of Transportation's University Transportation Centers Program.

The authors would also like to acknowledge and thank the following organizations and people for their support and encouragement: North Carolina DOT (Joseph Hummer, Kelly Wells), INRIX (Terri Johnson), the City of White Plains (Thomas Soyk), Monroe County, NY (Jim Pond, Thomas Polech), the Town of Cary, NC (David Spencer, Jerry Jensen, Tom Reilly), Georgia DOT, and Florida DOT.

Funding Agreement Number - 69A3551747104

LIST OF AUTHORS

Lead PI:
George List, Ph.D.
North Carolina State University
gflist@ncsu.edu
ORCID: 0000-0002-7044-1042

Co-PIs:
Billy Williams, Ph.D.
Institute for Transportation Research and Education
North Carolina State University
billy_williams@ncsu.edu
ORCID: 0000-0002-7599-1385

Michael Hunter, Ph.D.
Georgia Institute of Technology
michael.hunter@ce.gatech.edu
ORCID: 0000-0002-3651-6709

Mohammed Hadi, Ph.D.
Florida International University
hadim@fiu.edu
ORCID: 0000-0003-2233-8283

Additional Researchers:
Angshuman Guin, Ph.D.; Georgia Institute of Technology; angshuman.guin@ce.gatech.edu
Ishtiak Ahmed, Ph.D.; Institute for Transportation Research and Education; iahmed2@ncsu.edu
Hector Mata; Florida International University; hmata010@fiu.edu
Ahmad Abdallah; North Carolina State University; amabdal3@ncsu.edu
Nishu Choudhary; Georgia Institute of Technology; nishuchoudhary@gatech.edu
Atika Jabin; Florida International University; ajabi002@fiu.edu

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

This project created tools, empowered by big data, that can identify in real-time the onset of congestion and the occurrence of incidents, for both freeways and arterials. The target audience is highway system managers. The tools all track travel rates and flows, in real-time, to watch for trends that suggest the system's operating status is changing; especially, transitioning from normal to congested operation either due to high traffic demand or an incident. They can use either vehicular data or roadside detector data, with corresponding adjustments to the processing procedure. This report describes what we have done to develop these tools; the ideas we tried; the ones that worked; to some degree, the ones that did not; the data we used; and the current status of tool refinement. Briefly, we used Bluetooth data from Sacramento, CA; probe data from Tampa, FL; and system detector data from Atlanta, GA. We tried about a half-dozen ideas; the three described are the best to date; and we are in the process of refining them. Our anticipation is that these tools will reduce the severity of the impacts from congestion and incidents because their occurrence will be detected sooner, especially for congestion, and more reliably.

The tool developed for the Sacramento case study looks for the onset of Demand Induced Congestion (DIC) by tracking trends in the 5th percentile travel rates (min/mi). Moreover, it watches for the occurrence of Incident Induced Congestion (IIC) by tracking the difference between the 5th and 95th percentile travel rates. Our visual assessment suggests there are no false negatives. The correct positives for DICs in the daytime is 83–86%; for IICs in the daytime it is lower, because of considerable misclassifications (23%–34%). A second algorithm uses deep reinforcement learning to cluster system detector speeds into categories by operating condition and then looks for the occurrence of DIC and IIC on the basis of transitions from one cluster to another. A third uses clustering to categorize probe travel rates in a similar manner.

The case study based on data connected vehicle deployment in Tampa analysis revealed that the traffic states could be successfully classified into groups with six different traffic conditions based on speed, standard deviation of speed between vehicles, standard deviation between points, as well as the deceleration values. The case study also developed a machine learning algorithm for the prediction the breakdown based on connected vehicle data, achieving good accuracy rate and precision rate in the prediction.

The Atlanta case study used roadside detector-based data to estimate different traffic states. The traffic state, characterized by a set of input feature vectors that reflect the lane dynamics and spatiotemporal conditions, was labeled as belonging to one of the two classes, pre-congestion and non-pre-congestion. This problem formulation was tested using a set of generative and discriminative Machine Learning (ML) classifiers. The performance of these classifiers was evaluated using balanced accuracy, recall, and precision scores. Initials results demonstrated superior accuracy performance from tree-based classifiers.

Keywords (up to 5): Incidents, congestion, detection, big data, real-time

EXECUTIVE SUMMARY

This STRIDE project focused on developing tools, fed by "big data," that highway system managers can use to identify the onset of demand induced congestion (DIC) and the occurrence of incident induced congestion (IIC), for both freeways and arterials. It also developed an off-line tool that can be used to assess past performance.

The tools are all based on the same basic idea. They track the distribution of vehicle travel rates (min/mi), either from probes or system detectors, looking for evidence that those distributions have changed from one operating regime to another; specifically, from normal, uncongested operation to either DIC or IIC. They are prepared to do this either in real time or after-the-fact based on historical data. They watch specific metrics and set flags when either DIC or IIC has occurred, or both. One of the tools watches for DIC by tracking the 5^{th} percentile travel rate. If it repeatedly exceeds a threshold for successive samples (0.9 min/mi for the site examined), it sets a flag indicating that DIC has occurred. It can also use that same information to predict the likelihood that DIC will occur within 5, 10, 15… minutes. It also tracks the difference between the 5^{th} and 95^{th} percentile travel rates and sets a flag indicating IIC has occurred if that difference exceeds a threshold value (0.4 min/mi for the site examined). A second algorithm uses deep reinforcement learning to cluster system detector speeds into categories by operating condition and then looks for the occurrence of DIC and IIC on the basis of transitions from one cluster to another. A third uses clustering to categorize probe travel rates in a similar manner. For performance assessment, we created a tool that categorizes probe-based travel rates by TMC segment, time period, and speed range to assess the percentage of users that see specific qualities of service (travel rates).

Our anticipation is that these tools will reduce the severity of the impacts from congestion and incidents because their occurrence will be detected sooner, especially for congestion, and more reliably. Also, the performance assessment tool will help network managers identify where system improvements are needed and defend quantitatively, their benefits.

Our research in Phase 2 will focus on developing similar tools for arterial streets using necessary data from traffic signal plans, detectors, and probes.

1. INTRODUCTION

Degraded network performance due to congestion is a typical problem for urban networks. It is not possible to build enough capacity to eliminate it. Infrastructure investments produce land use changes, more real estate development, and increased destination options. These changes generate more traffic, sharper peaks, and higher congestion levels. The problems are most challenging for urban areas like Raleigh, Atlanta, and Miami where the populations are increasing rapidly. Tools such as peak load pricing may eventually prove useful in mitigating these effects; but tolling is not yet ready for wide-spread implementation in most urban areas. Incidents can be similarly disruptive. They can disturb system operation and create spontaneous congestion. Autonomous vehicles may reduce trajectory management "mistakes" in the future; but, at present, AV technology is not at that level, and its timeline for widespread use is highly uncertain. A useful meaningful strategy, currently, is to get "out in front" of these events, when they happen, and improve the system's performance as much as possible, reducing the severity of the impacts. To do this, however, involves 1) spotting the onset of the degrading performance quickly and reliably, and 2) taking appropriate, effective mitigating action consistent with policy objectives for performance.

Consumer apps like Google Maps and Waze have created an ability to flag these events quickly because of voluntary messages and the user trajectory tracking that is implicit. However, in a way, this is hearsay, it is unverified evidence. Recently, INRIX, an organization that provides connected car services and transportation analytics, developed an algorithm to alert truckers of drastic speed reductions downstream of the road (INRIX, 2022). However, transportation agencies do not have access to these algorithms employed; the apps are carefully guarded, understandably; and the app outputs are maps and text messages; and agencies do not necessarily have carte blanche access to those outputs. With the help of such alert services, transportation agencies could deploy low-cost operational treatments more efficiently and rapidly in response to traffic disruptive events. Such treatments include ramp-metering, traffic diversion via variable message signs, dynamic speed limit, and hard shoulder running.

The apps do demonstrate that "big data" and "crowd sourcing" do provide a way to detect the onset of these conditions; and differentiate between them. List *et al*. (2014), for example, found that disruptive incidents and traffic flow-caused congestion could be detected using Bluetooth data. Based on data for I-5 in Sacramento, they found that the variance in individual vehicle (MacID) travel times (rates) often decreased with the onset of recurring congestion (while the mean was increasing). It was thought that this might be caused by greater consistency in the vehicle trajectories, and as a result, the travel rates (speeds), because of less ability for vehicles to pass each other, change lanes, and thread through traffic. Similarly, but in contrast, during incidents, the travel time rates for the faster traveling vehicles (lower percentiles of the distribution) often increased dramatically, especially when the facility was not already congested. In this case, it was theorized that vehicles which, before the incident,

were able to achieve desired low travel time rates were now being hampered by the incident. Moreover, a drop in the flow rate was observed downstream of the incident. These trends suggested that early detection might be possible if it is informed by a combination of these parameters; as well as identification of the cause. If so, and if this behavior is a more general trend in places besides Sacramento; early detection and differentiation would be possible.

Researchers have found, generally, that the challenge is to develop insights into what the "big data" can tell us about the status of the system; and based on that, identify temporal/spatial "patterns" in the data stream that indicate system performance has dramatically changed (either abruptly or steadily). Once these questions have been answered, it should be possible to create an algorithm; and "train" it to spot these patterns; making sure it can distinguish carefully between false positives and false negatives (i.e., times when it thinks performance is awry, but it is not; or times when performance is deemed "normal", but it is not). This is the objective of this project.

It is recognized that incident detection algorithms exist and have been widely explored by researchers. However, there are still opportunities to explore the use of fused, multiple sources and new parameters and techniques. The work to date also tends to suffer from high rates of false positives, resulting in agencies often ignoring the triggered alarms. For example, in Atlanta, there was an automatic incident detection system, where for a 16 miles section of freeway more than 10,000 slow, congestion, or stop alarms were received over a 3-month period. Unfortunately, only a small number were related to actual, disruptive events. The high percentage of false alarms led to misuse. Here, we have used fused data with machine learning, to reduce the number of false alarms to a much lower percentage. Further, we have focused on using connected vehicle data to predict the occurrence of congestion based on combinations of microscopic and macroscopic traffic parameters.

1.1. OBJECTIVE

This STRIDE project aimed to create tools, empowered by "big data," that highway system managers could use to identify the onset of congestion and the occurrence of incidents, for both freeways and arterials. It had a related emphasis on off-line algorithms that can help them assess past performance.

All the tools use the same basic idea – they process incoming data to create distributions of travel rates and flows, which change in real-time, and watch for indicators of significant change, like the onset of congestion or the occurrence of a disruptive incident. The tools can be fed either probe data or roadside detector data, with differences in the mechanics. The tool for congestion detection watches for a rise in the 5th percentile travel rate (the faster moving vehicles) and predicts how long it will be (minutes) until a certain percentage (e.g., 95%) of the travel rates exceed an acceptable performance criterion (e.g., they will have a travel rate greater than "*y*" min/mi). Its output is a probability that his will happen, ranging from "green" – it is not likely in the next 30 minutes – to "red" – it is likely to happen in the next 5 minutes. The

tool for identifying incidents works similarly. It looks for major, abrupt changes in the spread between the 5th and the 95th percentile travel rates. If the spread increases abruptly, whether during congested or uncongested conditions, the conclusion is that a disruptive incident has occurred. The tool for performance assessment examines travel rate and flow data (by TMC segment -and-5-minute interval, each one being an *instance*) and identifies the number of these instances for which the travel rates exceed certain thresholds. When it is fed probe data (say from Bluetooth sensors), it looks to see what percentage of the probes have a travel rate greater than "*Y*" (a speed less than "*W*"). The reported metric is the percentage of instances that meet or exceed this criterion and where and when they occur (e.g., on I-40 for the TMCs upstream of the junction with I-540 during the AM and PM peak hours). This phase I report describes what we have done to develop these tools; the ideas we tried; those that worked and did not work; the data we used; and the current status of their refinement.

Our anticipation is that these tools will help reduce the severity of the impacts from congestion and incidents because the occurrence of these events will be detected sooner, especially for congestion, and more reliably. Also, the performance assessment tool will help network managers identify actions and system improvements that can mitigate the impacts.

1.2. SCOPE

The scope of the effort encompasses both freeway and arterial networks, although, given the realities of data availability, freeway-based analyses have received more emphasis. The focus is on urban settings, although the tools seem equally applicable to rural settings. The temporal context is both real-time and off-line. That is, we have focused not only on spotting the onset of congestion and the occurrence of disruptive incidents when they occur, but also their detection ex-post-facto and off-line. Within scope is the creation of tools that can spot these occurrences, by using "big data," like vehicular travel rates, and state-of-the-art AI tools like deep reinforcement learning. Not within scope is the identification of the cause – especially for the disruptive incidents – that is, seeking an explanation for why the congestion occurred (a special event) or why the disruptive incident happened, except that, in both instances, we do include a focus within the "big data", on any information about the operating environment which was or is extant at the location and time of the congestion or disruptive incident, including weather conditions, road work, special events, etc.

1.3. REPORT ORGANIZATION

The remainder of the report is organized as follows. Chapter 2 presents the literature review. In Chapters 3, 4, and 5, we conducted case studies for developing a congestion onset identification and performance assessment tool using data from select freeways in three major cities. In each of these three chapters, we describe the data sources, data cleaning process, algorithms for developing the tools, and the results. Chapter 6 presents the overall findings and conclusions based on these efforts and Chapter 7 describes the topics upon which the team will focus during Phase 2.

2. LITERATURE REVIEW

In this research, we investigated the pattern of travel time variation that freeway corridors face in order to develop an algorithm to identify and classify congestion onsets. Hence, in this chapter, we review past studies on travel times and their variability, including the development of a travel time reliability monitoring system, in addition to nuances associated with travel time measurement, statistical modelling, near-future prediction, anomaly detection using travel time data, with emphasis on studies that used Bluetooth technology for travel time measurement. The first section describes the travel time reliability metrics and the data collection techniques in this regard adopted by past studies. The following three sections discuss the real-time modeling and prediction of travel time and its distribution. Next, various filtering techniques for cleaning travel time data are reviewed. A brief review of traffic flow prediction techniques is presented next, followed by a summary of the chapter.

2.1. TRAVEL TIME RELIABILITY

List et al., (2014) developed a guide aimed at helping transportation managers establish a travel time reliability monitoring system (TTRMS). It is a useful document in that it describes what data to collect, and how to study it, so that the system's behavior during various operating conditions can be understood, and then mitigating strategies be developed to ameliorate the impacts. The guide identified five major steps involved in the development of a TTRMS.

The first step is to collect and manage the relevant traffic data from infrastructure-based sources and/or vehicle-based sources. Traffic data from infrastructure-based detectors include count and occupancy data from single point detectors and spot speed data from double point detectors. On the other hand, vehicle-based sources can directly collect information about individual vehicle travel time, either by matching vehicle identification information for a vehicle the passes successive detectors using automatic vehicle identification (AVI) based technologies such as Bluetooth reader (a method which will thoroughly discussed in this document), electronic toll tag reader, and license plate readers, or by tracking the path travelled by the vehicle through a system and consequently obtain travel time between any two points in the area under study, by utilizing automatic vehicle location (AVL) technologies such as GPS.

The second step in developing a TTRMS is to filter and manage the collected data to produce useful segment and route travel time information from the assembled traffic data, starting with filtering erroneous data points and imputing missing sensor values. A series of techniques are presented to identify erroneous data points, malfunctioning sensors, and replace missing data points with imputed values. The report also discusses the limitation of infrastructure-based sensors being unable to report travel time reading for individual vehicles and presents a methodology for synthesizing individual travel time data from the measured average speed values for vehicles passing a sensor. For this task, the main challenges associated with AVI data (particularly Bluetooth) are to choose the most appropriate readings from when a reader records multiple responses for a certain vehicle passing by, in addition to identifying the route

taken by a vehicle while travelling between two points, especially in arterial networks where it is more likely that a trip between two points will have more than one route option. A variety of techniques found in the literature which serve this purpose are discussed in a later section of this review.

The third and fourth steps are to characterize the measured travel times using statistical distributions and identify the operating conditions that pertained at the time the observations were obtained. This includes gathering and studying information about incidents and weather, from supplementary sources, and establishing the prevailing operating condition that existed for each observed travel time. The guide suggests using probability density functions (PDFs) and cumulative density functions (CDFs) to visualize the variation in travel times between vehicles throughout a certain time period and to characterize the reliability and performance of a given segment, and proposed using "regimes" to represent operating conditions and to compare the effects of different events, where each regime identifies a combination of the congestion level at the time and location the TT observations were recorded, along with the simultaneous type of non-recurring event that was taking place, including none. This latter task can be burdensome in cases when the acquired traffic data do not include a specification of the operating conditions at the time of the recorded reading, and challenges in this task include, but are not limited to, insuring that the whole period affected by the non-recurring event captured because, for example, the impact of a weather event or an incident may extend well after the weather event stops or the incident is cleared.

The last step is to understand the impacts of the sources of unreliability, by analyzing the distributions developed in steps 3 and 4, each for a given regime, and comparing between them using statistical parameters and the reliability metrics proposed. This would help in developing inferences, and in managing transportation facilities by being able to understand the sources of unreliability of a system and develop informed decision to tackle these sources and mitigate their effect.

A topic not addressed in the guide is the prediction of travel times and their distributions based on the present conditions, to alert drivers and better manage the system. Various methods for near real time prediction of travel times can be found in the literature and are presented in this review. A research idea in this regard, which has been pursued in this research effort, is to monitor the shape of the travel time PDFs and CDFs and see if they provide leading indicators of upcoming disruptive events.

A little more than a decade ago, Bluetooth technology emerged as a reliable and cost-effective method for acquiring valuable traffic data, such as travel times, origin-destination matrices, route choices, and even vehicle trajectories. The detection capitalized on two facts: 1) each equipped device had a unique media access control (MAC) address and 2) the Bluetooth chip was always in "on" mode looking for devices with which to pair. Hence, Bluetooth scanners, which were relatively inexpensive were "easy" to install on road segments, and by capturing responses from the Bluetooth chips, they were able to track the MAC IDs from one location to

another. Moreover, because each detection had an associated time stamp, based on a GMT-synched clock, travel times between scanners could be ascertained. However, the detections were simply MAC ID *X* at location *Y* at time *T*, without any intervening event information. So, cleaning and filtering were required to extract useful information, like travel times. Bhaskar and Chung, (2013) classified the source of noise into four categories:

1) Bluetooth observations from outside the vehicles, such as pedestrians, cyclists, Bluetooth equipped devices in office building and residences within range of BT scanners. This problem is more pronounced on arterials (Y. Liu et al., 2020).
2) Vehicles reporting trip times rather than travel times. The difference between the two is that trip times could include alternative routes and/or stops, whereas travel time is the required to directly travel between two scanners (List et al., 2014; Mitsakis et al., 2015)
3) Detecting a single MAC address multiple times during a short time interval. This happens when vehicles are detected multiple times by the scanner, due to, for example, driving slowly around a scanner, or when vehicles pass the upstream detector more than once before passing the downstream detector.
4) Missed observations. This is usually the case when vehicles are travelling very fast past a detector or when the scanner's range does not cover the all the lanes in in one direction on a freeway or arterial. This could also happen in the case of malfunctioning scanners.

List et al., (2014) recommend removing travel times in excess of five times the free flow travel time, and described a procedure for imputing travel times for AVI based sensors in the case of missing observation based on the concept of "super-segments", which imputes missing observations using observation from neighboring detectors. Haghani et al., (2010) established a process to filter outliers which included filtering out intervals with less than 3 observations in 5 minutes, and intervals with coefficient of variation higher than 1 in addition to filtering observations beyond 1.5 standard deviations from the mean travel time value of the observations in a certain time interval, where the standard deviation is also that of the observations in the same interval. This, however, is too conservative and may result in failure to capture sudden changes in operating conditions due to an incident or other event. Another approach was proposed by Samandar et al., (2018) which recommends filtering out travel time observations more than 3 standard deviations from the mean of the 30 neighboring data.

Malinovskiy et al., (2010) compared travel times from matched Bluetooth MAC addresses with those from automatic license plate recognition. The study was conducted on a 0.98 mi long portion of a freeway in the State of Washington. The results showed that the average travel times based on the Bluetooth data overestimated the actual travel times. They attributed this to the fact that some vehicles were not detected because they were travelling on lanes out of range for the scanners. Their remedy was to ensure that the detection area was large enough to capture all vehicles passing in the direction of travel towards the subsequent detector at different lanes and speeds.

A problem with this idea is that large detection areas are often a source of inaccuracy when using Bluetooth technology, especially for urban arterials. When the vehicles are queued, or are in a stop and go condition, the detector captures multiple time stamps for individual MAC IDs. Van Boxel et al., (2011) identified these multiple detections as being the main source of Bluetooth travel time data inaccuracy. Saeedi et al., (2013) suggested the accuracy could be improved by using the received signal strength indicator (RSSI) to filter the multiple reads. In their study, a controlled experiment was conducted by collecting data from 2 Bluetooth equipped devices travelling inside a probe vehicle that made 20 runs along a four mile stretch of Highway 99 in Tigard, Oregon. The stretch had a series of signalized intersections with an average segment length of one mile. The Bluetooth scanners recorded multiple hits for each MAC ID while in range. To establish the "ground truth" travel times, manual observations were obtained of passage times at the Bluetooth sensors. The Bluetooth timestamps were then compared with the manually recorded ones. In addition to RSSI, other filtering methods were tested. These were: last to last detection, first to first, and average to average. The field experiment confirmed that travel times recorded using the highest RSSI were the most accurate, followed by last to last. No clear reason was given for why last to last was better. It should also be noted that RSSI may not be available from all Bluetooth sensors.

To further investigate which of the heuristic Bluetooth timestamp matching methods is more accurate, Liu et al., (2020) conducted a study on 2 million records of time-stamped Bluetooth data gathered from 22 weekdays along the Canning Highway in Perth, Australia. They compared these observations with historical GPS data obtained from TomTom datasets. They found that the average-to-average method was more accurate for long arterial segments (distances of one km or longer), while last to last was the most accurate for short segments (distances less than one km). Nonetheless, the authors said the reason for the difference between this study and Saeedi et al., (2013) was unknown, and further research would be needed.

Several other methods have been proposed to identify outliers in travel time data. Robinson and Polak, (2006) introduced an "overtaking rule" method to filter out travel times reported by vehicles that did not follow a direct, non-stop route along a segment. Their method was developed based on travel time data collected through license plate matching. Clark et al., (2002) and H. Liu, (2008) examined the use of statistical methods including the percentile test, deviation test, and traditional z- (or t-) tests to explore the quality of travel time data obtained by license plate matching. The percentile test method defines lower and upper percentile bounds for the acceptable travel times values within an interval and removes observations beyond those limits. The deviation test defines a validity window based on the mean (or median) and the standard deviation of the collected travel time observations in a time period. However, these filtering methods do not perform well in transient conditions and during periods of turbulence such as the onset or dissipation of congestion (Moonam, 2016).

To improve the performance of the outlier detection algorithms in transient periods, Dion and Rakha (2006) introduced an adaptive model which establishes a validity window based on the

trend in the observations during previous time intervals and expands the validity window when 3 outliers are on the same side of the distribution, below or above the accepted bounds. This model is more accurate when compared with standard percentile and deviation test methods, but it can capture extreme outliers. In addition, it is sensitive to starkly changing times during intervals with low sampling rates.

There is also a need for real time outlier detection, in order to use Bluetooth travel time data in real time applications such identification of congestion onset, detection of anomalous operating congestion, and system reliability prediction models prediction models. The time-lag problem is one that has been highlighted by Moghaddam & Hellinga, (2014a) and is particularly pronounced in real time outlier detection as opposed to offline outlier detection. In practice, travel times are usually assigned to the time interval in which the vehicle crossed the upstream detector in the segment under study. Therefore, in real time outlier detection algorithms, some travel time observations that belong to the current time interval or most recent time interval(s) will not be listed until after the vehicle has crossed the downstream detector, which could be during a later time interval. This implies that outlier detection and filtering should be a recursive process, i.e., it should be repeated once a new set of observations is available for the subject time interval.

It can be inferred from above discussion that adaptive travel time outlier filtering algorithms which follow a reactive approach (i.e., depend on trends in recent time intervals) can produce erroneous results in real time filtering. In response to that, Moghaddam & Hellinga, (2014a) proposed an adaptive and proactive outlier detection algorithm that utilizes trends in recent time intervals in addition to similar historical patterns to produce filtering algorithm that performs better specially during traffic fluctuation such as the case of congestion onset and dissipation.

In their paper, Moghaddam & Hellinga, (2014a) used historical travel time data historical as an input in a k nearest neighbor (k nn) model that generates quasi predictions of the expected trends in travel time, which in turn are fused with travel times estimates from the recent time intervals to produce a more accurate travel time validity window for the each time interval. In a study that aimed at detected arterial road incidents in real time using Bluetooth individual vehicle travel times, Yu et al., (2015) computed average travel times for 1 minute intervals and then applied a 10 minute-time moving average technique and filtered individual vehicle travel time values that fell beyond a certain threshold above the smoothing line.

It can be noted that there is a plethora of noise filtering techniques for both real time and offline applications. However, further research is needed to determine the best heuristic Bluetooth MAC address matching technique specially on arterials.

2.2. REAL TIME PREDICTION OF TRAVEL TIME

Prediction of near future travel times is a key element in traffic management strategies, such as advanced traveler information systems, and it is therefore of vital importance for both users

and operators of the transportation system. In a broader sense, Approaches found in literature for travel time prediction can be broadly categorized as model based or data driven (Moghaddam & Hellinga, 2014b).

Many of the models used for travel time prediction are based upon queuing theory, traffic flow theory, and cell transmission concepts (Lo, 2001; Moghaddam & Hellinga, 2014b). However, these models are not easily applied in real time due to their complex nature and dependence on parameters that need continuous calibration. Data driven approaches, on the other hand, are simpler and base their predictions on statistical analyses and pattern recognition. Unfortunately, statistical approaches, such as regression and time series analysis, have not so far performed well during transitional periods, such as the onset of congestion (Moghaddam & Hellinga, 2014b).

For various pattern recognition-based approaches, Moghaddam & Hellinga, (2014b) concluded that, due to its nonparametric and flexible structure, the *k* nn (*k* nearest neighbor) method can be used to construct models that predict travel time in real time. However, the method requires both large-scale historical data sets and real time data

Moghaddam & Hellinga, (2014b) used historical data and the most recent Bluetooth travel time observations to develop a method for predicting near-term travel times based on a *k* nn model. Their model searches the available historical data base to find previous operating conditions like the present time. The ones selected have the smallest Mahalanobis distance from the current condition. The weighted arithmetic mean of those historical instances is then used to calculate the travel time estimate. The performance of the model, with optimized parameter values, was compared against a benchmark method which predicts the travel time for the next time interval based on the average travel time of the previous interval. The optimized *k* nn method exhibited an 18.5 % smaller mean absolute relative error (MARE).

2.3. REAL TIME PREDICTION OF TRAVEL TIME DISTRIBUTIONS

The previous discussion indicates that most travel time prediction methods focus on estimating only a single value of the travel time distribution, such as the mean or a specific percentile, rather than the distribution. Prediction of the travel time distribution is important because many of the reliability performance measures are computed based on the distribution, not a single value. Also, many travel time distributions envelop a range of operating conditions, say for the peak conditions, and help system managers and users make more informed decisions to mitigate or avoid these effects (Taylor, 1999).

Few studies have focused on predicting travel time reliability directly. Van Lint and van Zuylen, (2005) developed an artificial neural network-based model that characterizes the width and skewness of the travel time distribution based on the 10th, 50th, and 90th percentile values. Woodard et al. (2017) present a model that predicts the probability density functions for travel time distributions using mobile phone GPS data. However, the application of this method is

limited due to privacy restrictions associated with utilizing road users' GPS data and as such obtaining the sufficient input data for this method is difficult.

In a study on a 20 mile section of Interstate 5 in Los Angeles, California, and based on data obtained for 65 weekdays along the segment, Chen et al., (2003) concluded that accidents not only increase the median travel times, but increase the variability as well. Furthermore, Tu et al., (2008) studied the relationship between the occurrence of incidents and travel time reliability, by linking a large data set of incident data to estimated travel time data extracted from loop detectors, and concluded that incidents increase the variability in travel times by up to 4 times in comparison to normal operating conditions (without incidents). They found a 3% increase in the 10^{th} percentile travel time, a 15% increase in the median, and a 75% increase in the 90^{th} percentile travel time.

The above discussion emphasizes the need for a model that can predict travel time reliability in the very near future given real time information about both recurring and non-recurring congestion conditions. Samandar, (2019) attempted to do this by developing a statistical model that uses change-point linear quantile regression. It was sensitive to incidents and weather. The reason for using quantile regression was that quantile regression did not assume homogeneity of the variance and it was distribution agnostic. The near-real time travel time reliability estimation model was fitted to travel time, density, incident, and weather data collected from a 12.7-mile section of I-540 in Raleigh, North Carolina. Its application showed that density, the presence of incidents and weather events explained a high proportion of the variations in travel time with a pseudo-R-squared value of up to 0.8. However, the incident and weather-related variables were only introduced as binary variables in the model; it is possible that travel time reliability can be predicted better if other types of operating condition data are included.

2.4. TRAVEL TIME DISTRIBUTION MODELING

Modeling travel time variability using statistical models is an important tool for understanding the characteristics and properties of travel time distributions, and to quantify travel time reliability performance metrics.

To model travel time distributions, Mahmassani et al., (2012) studied the relation between the mean and standard deviation of travel time distributions, and concluded that during congested operation a higher mean travel rate was associated with a higher dispersion of the individual travel rates while during uncongested conditions, a shorter mean travel time was associated with a lower variation in standard deviation. On the other hand, List et al., (2014) , observed that as travel times increased due to the onset of congestion, the variation in travel times decreased. They attributed this to the fact that drivers had less ability to control their speeds.

Mahmassani et al., (2012) also concluded that based on statistical tests and regression models, the relationship between the mean and standard deviation was linear or nearly linear. In a subsequent study, Kim & Mahmassani (2015) extended this relationship by developing a

compound Gamma-Gamma model of individual vehicle travel time distributions that captured both vehicle to vehicle and day to day variabilities.

Delhome et al., (2017), from a set of 11 tested statistical models, concluded that the best ones for fitting ravel time distributions were the 3-parameter Burr distribution and the family of Halphen distributions, followed by the two parameter Weibull and Gamma distributions. The latter are limiting cases of the Burr and Halphen distributions, respectively. The 2 and 3 parameter log-normal distributions performed less well. Both the Burr and Halphen distributions can match a wide variety of distribution shapes, which is helpful, because the travel time distributions seem to have different shapes depending on the operating conditions. Their analysis was performed based on probe travel time data collected using both Bluetooth and License Plate Recognition matching, for over 1000 15-minute study periods.

2.5. ANOMALY DETECTION USING AVI-BASED TRAVEL TIME DATA

Early and accurate detection of unexpected events on the road network (such as incidents) by traffic management centers is of key importance so that fast and relevant actions can be taken to mitigate the impacts, which in turn provides better travel time reliability. Researchers have explored ways to detect anomalous events, using methods ranging from classical statistical techniques to machine learning algorithms. However, only a handful of papers discuss using AVI-based travel time data to detect incidents as will be seen in this section.

Hellinga & Knapp, (2000) presented a statistical method based on observing that vehicle travel times increase more rapidly with the onset of capacity reductions than is the case with increases in traffic flow. That is, a reduction in supply has more impact that an increase in demand. As a result, travel times during normal operating conditions are considered to belong to a different statistical population than those from non-recurring congestion. This means incidents can be spotted because their travel times fall outside the confidence limits for normal operating conditions. However, their study only used simulation, and simulation models are approximations of real-world behavior. Furthermore, the statistical methods used were limited; more advanced methods could have been used.

Other researchers have attempted to detect incidents or anomalous traffic conditions using Bluetooth data. Margreiter, (2016) presents an algorithm that detects incidents using filtered Bluetooth travel time records for individual vehicles. The algorithm is based on three parameters, which are the number of detected Bluetooth matches on a freeway segment in a certain time interval, the mean speed of vehicles travelling the same segment during the same time interval, and the maximum achieved speed in the segment during that time interval. Speeds are obtained by dividing the segment length by the observed travel times. Incident detection is based on comparing the values for these parameters with a set of threshold values. An incident warning is triggered when one of the threshold values is exceeded, and an incident is deemed to have occurred if the number of incident warnings exceeds a second threshold. However, the algorithm is unable to distinguish between different incident severity levels. It is

also unclear how the threshold values are set. The authors do not consider whether these threshold values might vary with hourly, daily, and monthly variations in traffic flow. Furthermore, it is not clear how the parameter values and their thresholds differ between incidents and recurring congestion.

Yu et al., (2015) present another statistical method for detecting incidents on arterials using Bluetooth travel time and volume data. Filtering for noise in travel time data was done by selecting an appropriate window above a time-moving average smoothing line. They observe a pattern of steep increases in the travel times when an incident occurs, which is similar to the observation made by Hellinga & Knapp, (2000). However, they add that an incident flag can be triggered when no vehicle passes the detector during a one-minute time interval. Once the trigger condition is satisfied, they deem that a persistent steep increase in travel times accompanied by a decrease in traffic volume is indicative of an incident. It should be noted that the trigger condition depends on the location of the incident with respect to the detector, which requires the detector to be downstream of the incident location to obtain gaps between detections shortly after an incident. Furthermore, the statistical methods employed do not fully exploit the data; more advanced statistical methods could have been used.

In contrast, Mercader & Haddad (2020) present an investigation of anomalous traffic condition detection using unsupervised learning, they strive to detect incidents and other types of non-recurring congestion without using historic instances to train the algorithm. Unsupervised anomaly detection approaches might be of significant importance because information characterizing a traffic state as normal or anomalous can be inaccurate and linking this information to historical records can be challenging and burdensome.

The unsupervised anomaly detection technique used by Mercader & Haddad (2020) is known as “isolation forest”. The premise of this technique is that anomalous operating conditions are rare in historical data and normal conditions are much more prevalent. Therefore, it is possible to train a detection algorithm using historical traffic data in which most observations are for normal conditions. The method utilizes traffic data obtained from Bluetooth detectors to generate a set of input variables; and based on these variables an anomaly score is calculated and assigned to each instance of data. The input data are time of day, avg speed in a road segment during a 5-minute interval, and spatial and temporal relative changes in speed between the upstream and downstream section, and between the current and previous time interval, respectively. This approach only detects an anomalous operating condition without specifying the anomaly type. They also used the Bluetooth travel times to compute individual vehicle speeds; but they did not use travel time as an input in the model. In this context, future research could be conducted to train unsupervised anomaly detection methods to detect incidents based on anomalous travel time distributions, such as those observed in the case of incidents or other non-recurring congestion sources. Furthermore, the steep slope trend associated with changes in travel times during an incident, which was observed by Hellinga & Knapp, (2000) and Yu et al., (2015) could be further investigated, and combined with the work

presented by List et al., (2014) to identify the onset of recurring and non-recurring congestion based on trends and shapes observed in travel time distributions.

2.6. SHORT-TERM TRAFFIC FLOW PREDICTION USING SPOT DATA

The objective of this part of the study is to identify early indicators of the onset of congestion and subsequently develop a prediction model for congestion events. This would allow traffic management agencies to take preemptive actions that would help minimize or mitigate the impacts of imminent congestion. Previous studies in this regard have primarily explored it in the context of short-term traffic flow or speed prediction. The problem of short-term traffic prediction has been of interest for researchers for more than four decades. (Vlahogianni, Karlaftis, and Golias 2014) provided an extensive review of the progress and existing challenges in the area of short-term traffic flow prediction. Earlier efforts for this research problem mainly used classical time-series based approaches (Moorthy and Ratcliffe 1988); (Lee and Fambro 1999). These efforts typically used Autoregressive models for short-term traffic prediction. The approach, however, was limited in terms of utilization of additional information available from spatially distributed neighboring traffic sensors. These improvements include Seasonal ARIMA (SARIMA) (Williams and Hoel 2003), State-space models (Stathopoulos and Karlaftis 2003), Generalized Autoregressive Conditional Heteroscedasticity (GARCH) (Kamarianakis, Kanas, and Prastacos 2005), among others; all of these studies used fixed point-detector based data for analysis. Further improvements of the ARIMA model were comprehensively reviewed in (Lippi, Bertini, and Frasconi 2013).

Due to the limitations of the assumptions involved in these types of time-series approaches, researchers are increasingly exploring data mining-based approaches. Data-discovery/data-mining based approaches are considered to be more robust than the classical approaches when it comes to unstable traffic flow conditions (Vlahogianni, Karlaftis, and Golias 2014). Early efforts in using data-mining based approaches for traffic flow forecast mainly used Neural Network (NN) (Hua and Faghri 1994); (Dougherty 1995); (Park and Rilett 1999);(Abdulhai, Porwal, and Recker 2002); (Dunne and Ghosh 2012). The NN approach essentially assumes a mapping from input features to the output and tries to approximate this mapping through a series of nonlinear transformations of the input (Goodfellow, Bengio, and Courville 2016). (Karlaftis and Vlahogianni 2011) compared the performance of NNs with time-series approaches and found that NN indeed gave a better prediction accuracy. However, authors have expressed concerns regarding the lack of complete transparency of NNs, especially when it comes to analyzing the nature of errors. (Vlahogianni 2007) employed 'Genetically optimized Probabilistic Neural Networks' for short-term traffic state identification and prediction using detector-based data. Later work in the area of short term traffic prediction, focused on the use of Support Vector Machines (SVM) (Chun-Hsin, Jan-Ming, and Lee 2004); (Castro-Neto et al. 2009). (Wang et al. 2013) used a combination of SVM and time-series approach for automatic identification of incidents. Specifically, they used time-series approach to predict the future and any deviation from this prediction was detected by SVM as an incident. (Xia, Huang, and Guo 2012) used clustering approach to identify the state of the traffic. Both (Xia, Huang, and Guo

2012) and (Wang et al. 2013) have used popularly used California's Freeway Performance Measurement System (PeMS) database for analysis. Studies that have explored other approaches like, Decision Trees and Random Forest, for traffic state identification and/or prediction include (Vasudevan et al. 2016); (Liu and Wu 2017), (Impedovo et al. 2019), and (Filipovska and Mahmassani 2020).

Overall, ML approaches present great potential when it comes to pattern analysis, but studies have expressed concerns about the lack of a good understanding of these techniques, especially in the data requirements, the assumptions involved, and the training of the models, leading to suboptimal applications. Also, the 'black box' nature of some of these techniques makes it difficult to draw insights from the learned parameters and connect the final model to conventional traffic flow theories. Keeping these challenges in mind, this study aims to test the capabilities of these techniques and draw useful insights regarding the applicability and usefulness of these techniques in developing a congestion prediction model.

2.7. APPLICATIONS OF MACHINE LEARNING ALGORITHMS

Machine Learning (ML) algorithms that are typically used to explore state identification problems, formulated as classification problems, can be broadly divided into two categories, generative and discriminative classifiers. The primary difference is that the generative classifiers model the joint probability distribution, p (x, y), of the given inputs and target labels and use Bayes rule to make predictions, whereas the discriminative classifiers model the decision boundary or the posterior distribution, $p(y|x)$, (Vapnik 1995), (Ng and Jordan 2001). Commonly used generative classifiers include the Naïve Bayes classifier (Hand and Yu 2001) and Bayesian Networks (Friedman, Geiger, and Goldszmidt 1997). While, discriminative classifiers include, Logistic Regression, Decision Trees (Breiman et al. 1984), Random Forest (Breiman 2001), and Support Vector Machines (Boser, Guyon, and Vapnik 1992).

Among these, Logistic Regression assumes the posterior distribution to be a logistic sigmoid function acting on an input of linear combination of the features, (Bishop 2006). Unlike Logistic Regression, SVM does not provide the class-probability and instead outputs the classification. This classification is based on the decision boundary estimated by maximizing the margin between classes, (Goodfellow, Bengio, and Courville 2016). SVM uses the formulation given in the equations 2-1 to 2-3 given below (Pedregosa et al. 2011).

This type of formulation for the classifier relies only on points close to the boundary to estimate the shape of the same, hence the name 'Support Vector.' The classifier provides additional flexibility by allowing the transformation of the input data to a higher dimensional space through the usage of a wide variety of kernel functions (Pedregosa et al. 2011). However, the method is susceptible to overfitting and usually requires a lot of computational power (Pedregosa et al. 2011). Additionally, the estimated decision boundary can be difficult to interpret, especially for the case of non-linearly separable data.

$$\min_{w,b} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i \quad \text{(Eq. 2-1)}$$

$$s.t. \;\; y^i\left(w^\top \phi(x^i) + b\right) \geq 1 - \zeta_i, \; \forall i \quad \text{(Eq. 2-2)}$$

$$\zeta_i \geq 0, \forall i \quad \text{(Eq. 2-3)}$$

Where, *w* and *b* represent parameters of the decision boundary (separating hyperplane)

n is the total number of training points

C is the regularization parameter

ζ_i is the i^{th} sample's distance from the correct margin boundary, as can be observed in Figure 2-1

$\phi()$ is the kernel function used

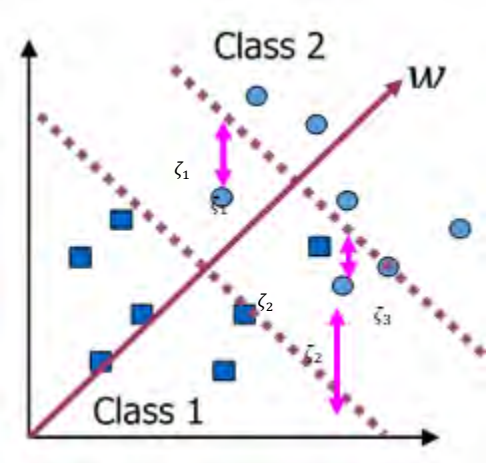


FIGURE 2.1: ITH SAMPLE'S DISTANCE FROM THE CORRECT MARGIN BOUNDARY GIVEN BY Z_I (LAN 2020)

Compared to SVM, the decision boundary for Decision Trees and Random Forest is approximated by dividing the given input feature space into non overlapping sub spaces, with separate parameter for each subspace, (Goodfellow, Bengio, and Courville 2016). The mathematical formulation for decision trees is given in equations 2-4 to 2-7 below (Pedregosa et al. 2011). The tree-like structure, represented by the formulation, for both algorithms allows for an easier interpretation of the decision boundary. However, similar to SVM, the tree-like structure can lead to overfitting. Additionally, decision trees can be sensitive to missing input data points and class imbalance (Pedregosa et al. 2011).

$$\theta^* = argmin_\theta(G(Q_m, \theta)) \quad \text{(Eq. 2-4)}$$

$$G(Q_m, \theta) = \frac{N_m^{left}}{N_m} H\left(Q_m^{left}(\theta)\right) + \frac{N_m^{right}}{N_m} H\left(Q_m^{right}(\theta)\right) \quad \text{(Eq. 2-5)}$$

$$Q_m^{left}(\theta) = \{(x, y) | x_j \leq t_m\} \quad \text{(Eq. 2-6)}$$

$$Q_m^{right}(\theta) = Q_m \backslash Q_m^{left}(\theta) \quad \text{(Eq. 2-7)}$$

Here,

Q_m is the data and N_m is the total number of samples at node m

$\theta = (j, t_m)$ represents parameter for a split for feature j and threshold t_m at node m

$Q_m^{left}(\theta)$ and $Q_m^{right}(\theta)$ are the candidate split at node m in the left and right leaf, respectively

N_m^{left} and N_m^{right} are the number of samples according to the selected split at node m in the left and right leaf, respectively

$G(Q_m, \theta)$ is the impurity measurement criteria at node m

Compared to these methods, the Gaussian Naïve Bayes classifier represents a simpler and computationally cheaper classifier, as can be observed from the formulation in equations 2-8 and 2-9. However, the formulation assumes independence among used predictor variables which can be unrealistic given the highly correlated nature of traffic variables.

$$P(y_k / x_i) = \frac{P(x_i / y_k) P(y_k)}{\sum_k P(x_i / y_k) P(y_k)} \quad \text{(Eq. 2-8)}$$

$$P(x_i / y_k) = \frac{1}{\sqrt{2\pi\sigma_{y_k}^2}} e^{-\left(\frac{(x_i - \mu_k)^2}{2\sigma_{y_k}^2}\right)} \quad \text{(Eq. 2-9)}$$

Here,

$P(y_k / x_i)$ is the probability of belonging to kth class given input features x_i for the ith training example

$P(x_i / y_k)$ is the probability of observing input features x_i given the ith example belongs to kth class

$P(y_k)$ is the probability of observing kth class

σ_{y_k} and μ_k represent the standard deviation and mean, respectively, estimated using given class membership for the training data.

2.8. SUMMARY

This section has presented a review of studies related to travel time reliability and non-recurring congestion identification. Research on the development of a travel time reliability monitoring system, which is a powerful tool that can help both managers and users of transportation systems better understand the impact of different types of operating conditions on the reliability of travel times as been discussed. Furthermore, research on the steps involved

in the development of a TTRMS, such as outlier filtering and cleaning of travel time data (with particular emphasis on Bluetooth-obtained travel time data), and modelling travel time distributions, was highlighted. Moreover, recent research on the near future prediction of travel time key metrics and distributions, and research on detecting incidents and anomalous operating conditions based on AVI traffic data has also been presented.

It can be noted that the different travel time distribution shapes that were associated with different regimes by List et al., (2014), and the steep slope trend associated with changes in travel times during an incident, which was observed by Hellinga and Knapp (2000) and Yu et al., (2015), can be further investigated to explore the possibility of identifying the onset of recurring and non-recurring congestion based on the trends and shapes observed in travel time distribution, which is one of the goals of this project

Moreover, this review shows that there is a lack of an effective tool that identifies the onset of recurring and non-recurring congestion. This is important to improve the reliability of the transportation system by allowing system managers to react in a timely manner to the potential formation of congestion and mitigate its impact, which will be the focus of this project.

3. SACRAMENTO CASE STUDY

Probe data is becoming increasingly more popular as a source of "big data" for monitoring system performance. Bluetooth is now a common technology for collecting this information. In the future, it is likely to be obtained from GPS-based vehicles, as is illustrated, later, by the Tampa-based case study.

This section presents tools that were developed for detecting DIC and IIC, as well as monitoring system performance, from Bluetooth data collected in 2011 on I-5 in Sacramento, CA. The tools 1) detect the onset of DIC, 2) detect the occurrence of IIC, and 3) assess past performance of the system. In this case it is a 5-mile section of freeway. Not only were these tools developed, but a few other ideas were pursued that were eventually set aside. Those tools are described in a set of appendices.

The performance of the tools is assessed by examining metrics like correct positives, false positives, and false negatives. We were especially interested in understanding the reliability of the tools. For example, we wanted to know how often the DIC tool asserted that DIC had occurred when it had not (a false positive). We were also interested in its anticipatory power, the reliability of its ability to predict that DIC was about to arise.

The chapter is organized as follows. Section 3.1 describes the dataset that was employed. Section 3.2 presents the idea aimed at detecting both demand-induced congestion and disruptive incidents simultaneously. Section 3.3 describes the performance of the congestion detection algorithm. Section 3.4 presents the results pertaining to the anticipatory power of the tool. Section 3.5 describes an idea about doing ex-post-facto performance assessment.

3.1. Sacramento Dataset

The Sacramento dataset was obtained using Bluetooth (BT) detectors that were deployed by CalTrans on Interstate 5 (I-5) in Sacramento, California. The data were obtained during project SHRP-2 L02 and reside in a public archive. Segment travel times were computed by identifying and then re-identifying device MAC addresses that passed two or more of the four Bluetooth detectors shown in Figure 3-1. The dataset covers all weekdays between 1/24/2011 and 3/15/2011, except for 2/8/2011 and 2/9/2011 (35 weekdays in total). The northernmost Bluetooth detector is very close to Downtown Sacramento and is located just south of the interchange with US-50. US-50 is a major east-west highway that runs through Sacramento.

Each Bluetooth detector had about a 300-foot detection radius and was installed along the side of the road. Thus, each could easily monitor both north- and southbound lanes (eight lanes total in some locations) of the respective freeway sections. That said, it is likely the detectors also collected undesired samples and background noise, such as from Bluetooth devices on nearby facilities or in the nearby office buildings. The raw data were filtered to exclude background noise, multiple successive detections of the same vehicle by one detector, and very high travel times associated with indirect trips between the upstream and downstream

detectors. To achieve this goal, this study used a travel time filtering algorithm that is described in the following subsection.

3.1.1. Probe Data Filtering Technique

Data aggregation

The first step of the filtering algorithm is to group consecutive observations. The objective is to leverage the similarity in the data points that are temporally proximal. The choice of the number of consecutive data to aggregate, *N,* is driven by the temporal headway between data points, which, in turn, depends on the Bluetooth match rate and the traffic flow rate. For instance, when the traffic flow rate is 2,000 vph and the Bluetooth devices can estimate the travel rate of 4% of vehicles, the average headway will be $= \frac{60}{2{,}000*0.04} = 0.75\ minutes$. Considering the observed Bluetooth match rate and the range of traffic flow rate, we chose $N = 11$. For the above example, each group spans over $0.75 * 11 = 8.25\ minutes$. Note that the span is very high during low-flow conditions (e.g., nighttime), however, traffic condition does not change much during those periods unless a disruptive incident happens.

We aggregated the data with an overlap, *p*%, i.e., in this process, the next group will contain *p*% data from the previous one. This overlap is important to smooth jittery time-series observations. Upon testing different values, we chose $p = 91\%$, i.e., ten out of 11 data in a group come from the previous group.

Apply a lower and upper bound threshold

In this step, travel rate observations that are very low or very high are tagged as outliers. Given that the posted speed limit of the study corridor at the time of data collection was 60 mph, it is highly unlikely that any vehicle would travel at a speed of greater than 85 mph. This speed is equivalent to a 0.7 minutes/mile travel rate and was used as the floor of the travel rate data. We also considered travel rate observations higher than 5 minutes/mile as outliers in order to remove data that represent a trip time. This threshold was selected upon observing the maximum travel rates during periods that are evidently congested.

Detect any abrupt jump in travel rate

The next step is to identify data points in each group that are relatively higher than most data points in that group. To do so, the algorithm arranges the data in each group in ascending order of travel rate. Then, the first differences of travel rate of this sorted data are calculated to highlight any abrupt jump in travel rate. The rank of the travel rate observations (i.e., the cumulative probability) is also calculated in this step. The algorithm checks if the first difference for any observation exceeds a predefined gap threshold, *g*. The filtering process is complete for that group if no such data point is found. If such a jump in travel rate is found, the rank of that data point is checked to see if it belongs to the high end (say, *c*%) of the travel rate distribution for that group. If any first difference value exceeds *g* and its rank is higher than *c*%, that data point and any other with a higher rank are tagged as outliers. We tested the sensitivity of the congestion detection tool to the threshold values by visually assessing the traffic condition from

the probe data, and found that $g = 0.2\ minutes/mile$ and $c = 82\%$ (i.e., rank 9 in the 11 group of data) work best for our purpose.

3.1.2. Exploration of Probe Data

The probe vehicle travel times show demand-induced congestion occurring in the mornings northbound and in the afternoons southbound. The congestion southbound in the PM peak is less severe than that which occurs going northbound in the AM peak.

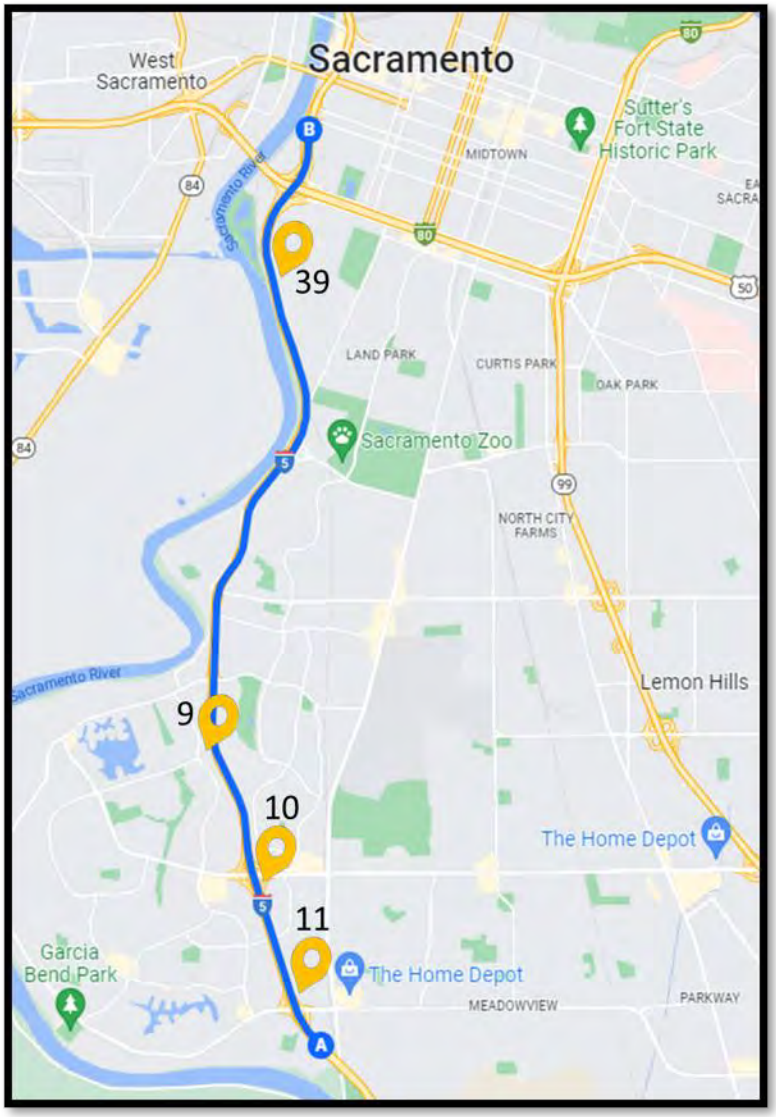


FIGURE 3.1: LOCATION OF THE BLUETOOTH SENSORS ON I-5

At the time the data were collected, it was estimated that 20% of all drivers had Bluetooth devices with them in their vehicles. However, for this same study site, List et al. (2014) compared the number of Bluetooth MAC address detections by Bluetooth detectors on to the

volume of traffic passing loop detectors (described later) embedded in the same location as that of the Bluetooth detectors, and concluded that the percentage of traffic detected by Bluetooth detectors generally ranged between 6% and 10% of that detected by loop detectors. This indicates potential sample size and observation scarcity problems, especially during periods of low traffic volumes such as the nighttime period. Moreover, the percent range mentioned above represents the detection rate, which is the percentage of traffic detected by one BT detector, not the matching rate, which rather represents the percentage of traffic identified by an upstream detector and re-identified by a downstream detector. The matching rate may even be lower than the detection rate since some of the vehicles that are detected by one of the detectors may not be detected by any other. Additionally, some vehicles that were detected at the upstream detector did not travel directly to the downstream detector. These observations gave erroneous section travel time estimates but were removed through the filtering technique applied.

The ideas presented here used a subordinate dataset that contains travel times for probes that were detected traveling between the first and last stations. (The probes might have been observed in-between as well, but those intermediate observations have not been employed.) The dataset contains information for about 150,000 probe "trips" in each direction or slightly less than 300,000 probe trips overall. (This represents about 2400 probe trip observations per day, or 100 per hour, although the peak observation rate reaches about 500 per hour or one every 7 seconds.) For each observation, or record, the dataset has the following data: 1) the MacID, 2) a timestamp for passing the first station, and 3) a timestamp for the second. Moreover, based on the time stamps at each detector and the distance between them, each record shows 4) the travel rate, which is typically 0.8 to 1.4 minutes/mile, and 5) the headway from the previous vehicle, measured at the downstream, exiting location. The headways tend to be shorter when the traffic is heavy, down to about 0.12 minutes (7 seconds); and they are longer when the traffic is light.

Importantly, information about the "ambient" conditions are included in the dataset, for each record. This includes data about 6) the weather conditions, 7) incidents, as recorded in the PeMS (CalTrans) database or discovered using newspaper reports, etc., 8) flow rates based on four system (loop) detector stations that are located along the freeway, and 9) stitched travel times that are based on spot speeds from those same system detectors. This means it is possible to tell the "operating conditions" under which each observation occurred. The SHRP-2 LO2 project referred to these operating conditions as "regimes".

Figure 3-2 presents a snapshot of the probe-based travel rates from the southbound data. The *x*-axis displays the day (in Excel format, starting with day one being 1/1/1900) and time (not demarcated). The *y*-axis shows the probe travel rates in min/mi or headways in minutes. The top line is 4.0 minutes; and the demarcations are every half minute.

The blue dots are the observed travel rates (minutes/mile) for each probe. As can be seen, the values tend to be about 1 min/mile, plus or minus, except during the PM peak when they rise to

1.5 min/mi or more. Also, there are more observations between about 5am and midnight than there are during the early morning hours (say midnight to about 5 am). The green dots show the headways between the probe observations. As can be seen, the headways are larger at night and shorter during the daytime.

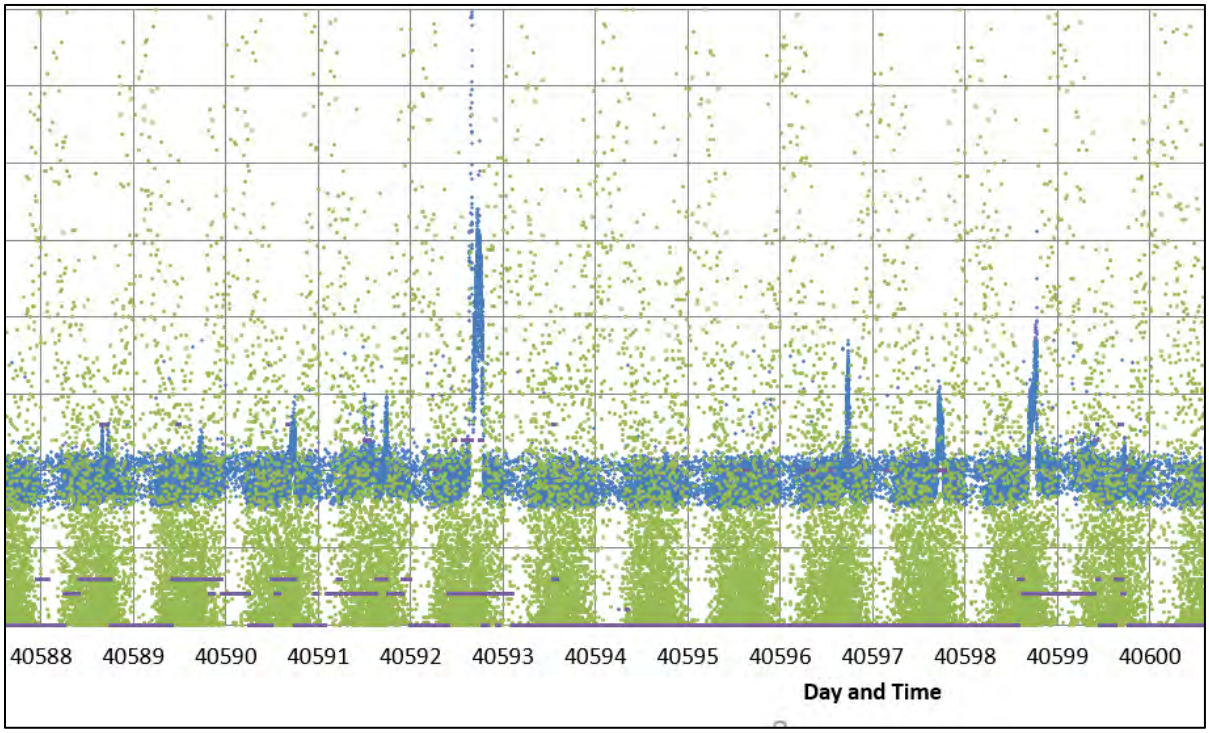


FIGURE 3.2: A SNIPPET OF THE SOUTHBOUND PROBE TRAVEL RATES AND INTERVALS BETWEEN OBSERVATIONS

Figure 3-3 shows these same data for days 40588 and 40592 (2/14 through 2/18/2011), a Monday through Friday. Again, the *x* axis displays date and time, from the beginning of 40588 to the end of 40592 and the *y* axis shows travel rate or headway (0 to 4 minutes in intervals of 0.5 minutes).

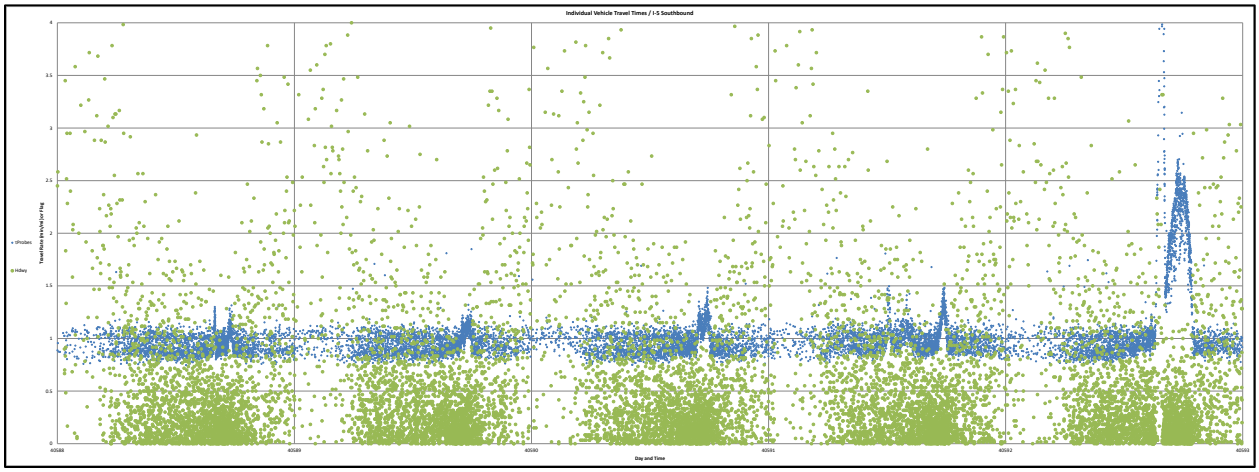


FIGURE 3.3: TRENDS IN THE PROBE TRAVEL RATES AND HEADWAYS FROM FEB 14 TO 18, 2011, MON-FRI

The trends in the probe travel rates and headways are easy to see. Off peak, the travel rates range between about 0.8 and 1.2 min/mi. There is an afternoon peak each day and sometimes there are disruptive incidents; like on Monday, Wednesday, and Friday, when there are incidents that occur just before the peak; the one on Friday is particularly pronounced. On Thursday, a disruptive incident occurs just before noontime.

Figure 3-4 shows examples of probe vehicle travel rates for a typical day for probes traveling on I-5 Southbound. The colors indicate the operating conditions, as in inclement weather and incidents.

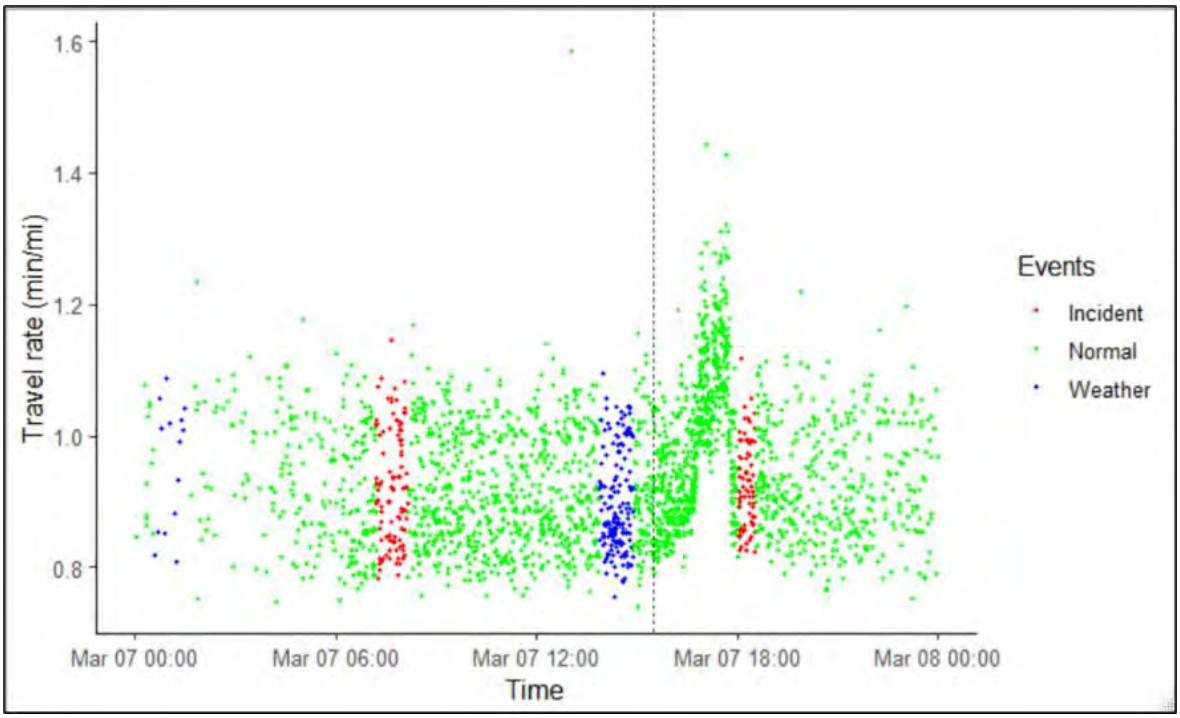


FIGURE 3.4: PROBE VEHICLE TRAVEL RATE AGAINST CLOCK TIME FOR ONE DAY ON I-5 SB, COLORED BY DIFFERENT OPERATING CONDITIONS

On this day, there were two incidents and one inclement weather condition. None of these events caused a significant increase in the travel rates. The travel rates start showing a rising trend at around 4:30 PM and reached their peak at 5:15 PM. The key observation is that the lower part of the travel rate band increases at the knee of the curve before congestion onset (shown by the dashed line), indicating that the faster moving vehicles have to lower their speeds. In contrast, the larger travel rates do not exhibit a similar trend, indicating that the slower moving vehicles are still able to maintain their speeds.

Figure 3-5 shows the same plot for the same corridor but for a different day when, in addition to the recurrent peak, an inclement weather condition caused the travel rate to increase dramatically at around 6 AM. This event could be attributed to an unreported incident as well. The key point is that this rise in travel rate is more dramatic and abrupt compared to that we

see during the recurrent peak hour in this or in the previous plot. Similar observations from other days and from the northbound corridor kindled the idea that when plotted against time, the width of the travel rate band can be used as a signal to detect and classify congestion onset.

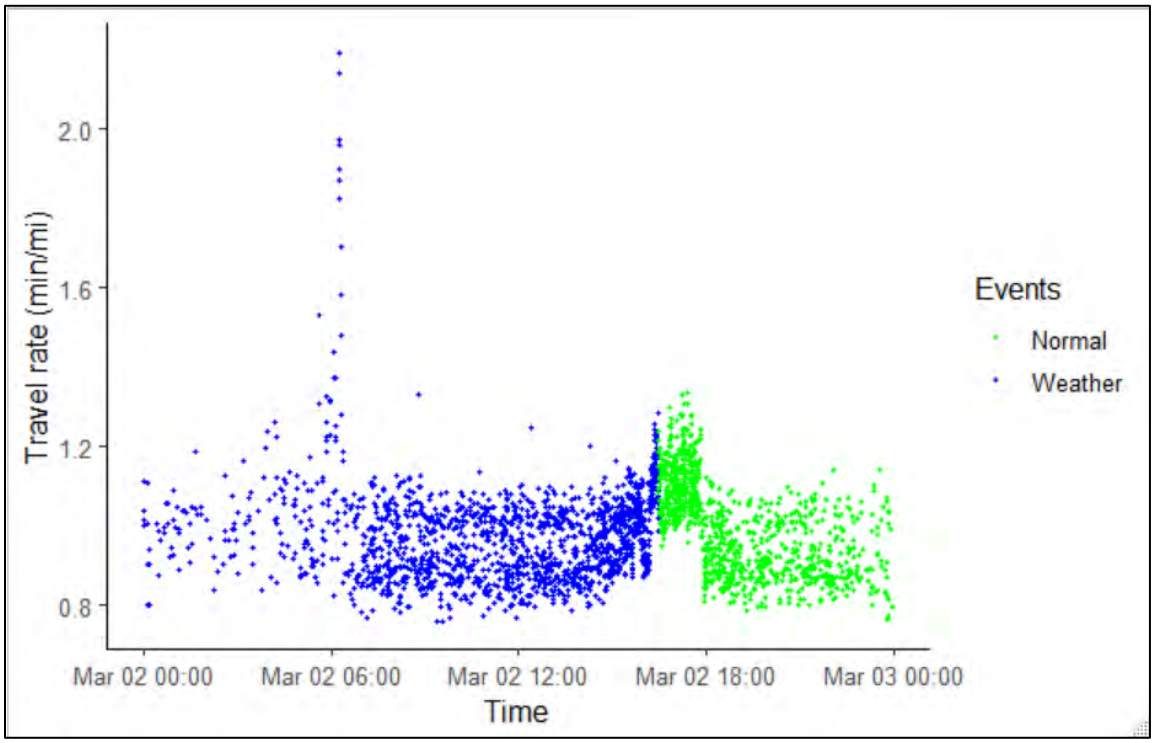


FIGURE 3.5: EFFECT OF INCLEMENT WEATHER ON THE PROBE VEHICLE TRAVEL RATE PATTERN AGAINST TIME

One way to quantify these extremes is to group several successive observations, estimate the extreme travel rate data for each group, and plot that against time for that group. Here, the groups are formed by combining 30 successive observations, and two consecutive groups have 80% overlap between each other. The motivation for creating such fixed-sample aggregates is explained in section 3.2. The 5% and 95% travel rate for each group are estimated and plotted against the latest time stamp for that group. Of course, although seems reasonable, the choice of these number of data points in each group, overlap percentage, and percentiles are somewhat arbitrary.

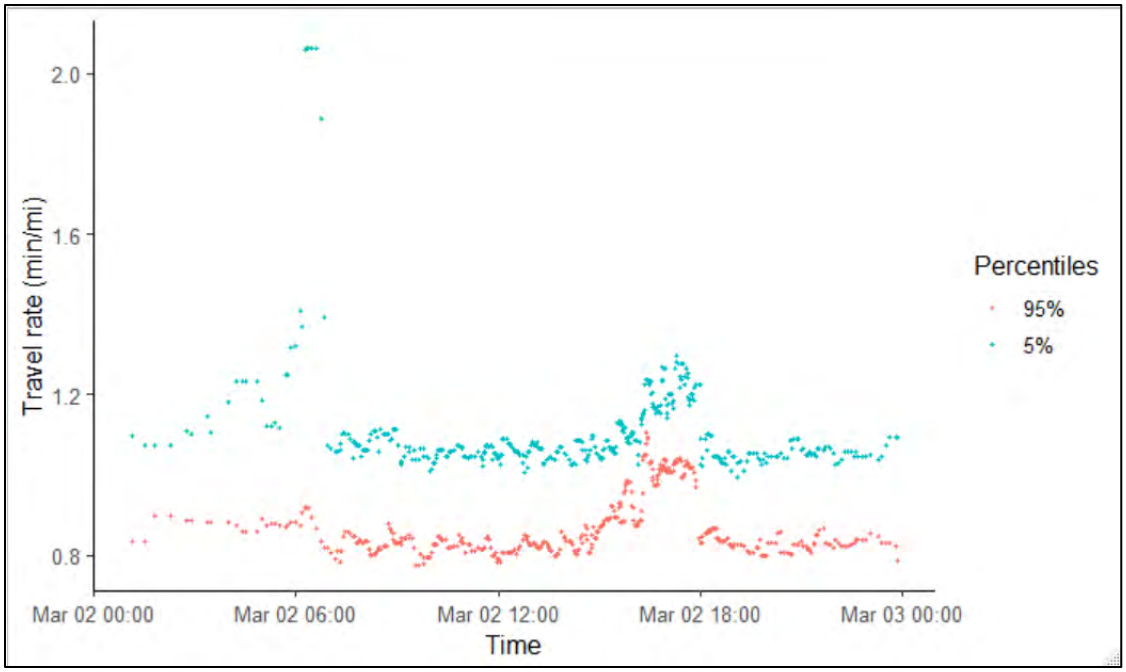


FIGURE 3.6: EFFECT OF INCLEMENT WEATHER ON THE 5% AND 95% TRAVEL RATE FOR EACH GROUP

Figures 3-5 and 3-6 show that just before the afternoon congestion onset time, the 5% travel rate—which is created by the faster vehicles—exhibits a gradual rise. On the other hand, the 95% travel rate, created by the slower moving vehicles, remains stable until reaching the congestion onset time at which point it gradually starts to increase as well. These patterns are completely different when there is an abrupt rise in travel rate in the morning (~6:00 AM). This congestion is possibly attributed to an unreported incident. When this disruptive congestion occurred, the 95% travel rate rose abruptly. The 5% travel rate did not change much possibly because the congestion severity was not the same across all lanes. In all, Figures 3-5 and 3-6 show that the width of the travel rate band and its changing pattern may potentially signal congestion onset and the probable cause of the congestion.

3.2. Algorithm for Detecting and Classifying Congestion

As can be seen in the plots of probe travel rate trends, like those shown in Figures 3-2 through 3-6, it should be straightforward to sense when the facility is operating in demand-induced congested mode and when disruptive incidents occur. However, this is not the case. What we can see by visual inspection and based on a broad-brush examination of the travel rate is not so simple to incorporate into a computer-based algorithm.

The idea presented here had its genesis in a statistical analysis of the probe travel rates. We borrowed ideas presented in SHRP-2 L02 and began looking at distributions of the travel rates. Our preliminary analysis proceeded as follows: a) we grouped the probe observations as described in Section 3.1.1; b) we computed the 5^{th}, 15^{th}, 50^{th}, 85^{th}, and 95^{th} percentile values, using the raw data and interpolation. Then c) we looked for trends in these percentile values.

Figure 3-7 illustrates the outcome of this approach for 5th and 95th travel rates and the spread between these two for a subset of the study period. The *x* axis is again date and time. The *y* axis is the travel rate in minutes/mile. It also includes information about the ambient condition (as reported by PeMS) labeled as "acFlag" and colored light blue. acFlag shows combined information about all the *reported* incidents and weather. Its value is 0.1 if the weather was misty, 0.2 if it was rainy, 0.3 if it was windy, and 0.4 if it was foggy. Further, it was set to 1.0 if an incident had been reported, and to this information the information about the ambient conditions was added. So, 1.1 meant it was also misty at the time of the incident, etc. While it may not be obvious from the figure, the data points occur in stacked sets of four. That is, at each time the 5th percentile value is plotted, the 95th percentile value, spread value for the current set of 30 observations, and the value for the acFlag indicator are also plotted.

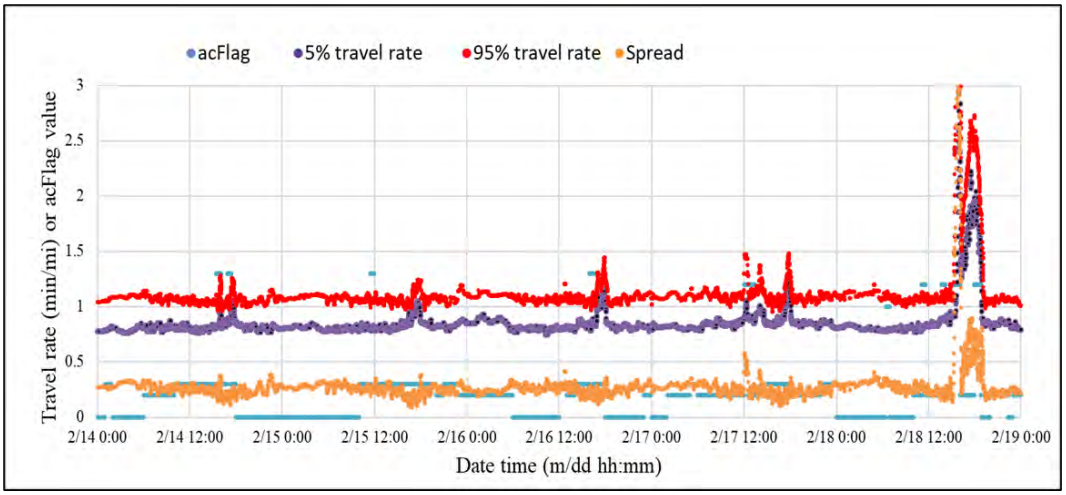


FIGURE 3.7: TRENDS IN THE PERCENTILES OF THE PROBE TRAVEL RATES, FOR 2/14/2011 THROUGH 2/18/2011

While only the results are shown, involving the 5th and 95th percentile values, the spread, and the acFlag values, we also experimented with other ideas, like using the 15th and 85th percentile travel rates, the minimum and maximum travel rates, the standard deviation, higher order moments, values derived by fitting a Burr distribution to the data, etc. Those are described in Appendix A.

In the end, the 5th and 95th percentile values and the spread between them proved to be the most useful inputs regarding the system operating conditions. For these three metrics, we liked three tendencies we saw: 1) the 5th percentile travel rate tended to increase as the traffic conditions trended toward heavy flow rates (congested conditions); 2) this trend did not occur before disruptive incidents (in the time leading up a recorded incident when the acFlag indicator became 1.0 or greater; and 3) sometimes, but not always, when there was an incident recorded by acFlag indicator, the spread between the 5th and 95th percentile travel rates tended to increase to higher-than-normal values suggesting a disruptive event had occurred.

It is possible to see evidence of these trends in Figure 3-7. For example, before the demand induced congested condition arises, say in the PM peak here, the 5th percentile travel rate begins to increase (the faster moving vehicles slow down). The 95th percentile travel rate (the slower moving vehicles) seems to be unaffected, which means the slow-moving vehicles continue to move slowly. Then, as the demand-induced congestion increases, both the 5th and the 95th percentile travel rates increase (all the traffic moves slower), but the spread between them stays constant, their trends seem synchronized. The spread between the 5th and 95th percentile travel rates does not increase dramatically. However, when the travel rates rise abruptly (as would presumably be associated with an incident, whether it is during the peak load conditions or not, the spread between the 5th and 95th percentile travel rates seem to increase. This can be seen at noon (off peak) of February 17 and before the PM peak on February 18 (prior to the demand-induced congestion). We sense that the increase in spread between the 5th and 95th percentile travel rates, is attributed to the fact that, during disruptive incidents, the operating conditions deteriorate, whether the disruptive incident occurs during the peak load conditions or not. Our perception is that the "stable" interactions between and among the vehicles (even though they are traveling at different speeds) become disturbed. While the faster moving vehicles may or may not slow down; the slower moving vehicles (the ones reflected in the 95th percentile, slow down considerably; they seem to be most affected; or, even if it is not exactly the slower moving vehicles that are most effected, some vehicles experience very large travel rates; so the spread between the 5th and 95th percentile travel rates increases.

At first, our goal was to identify all the incidents. But we soon realized that this is not only difficult, but maybe impossible. Based on the probe travel rate data alone, it may not be possible (and probably it is not) to identify incidents that do not disrupt the traffic flow, and hence the operating condition. It may be that only local observations, say from Waze or some other data stream is needed to detect these events (until all vehicles are GPS instrumented and disrupted vehicle trajectories are easy to detect). It is still very important to dispatch "help trucks" and first responders for such events, and a detection scheme is needed, but it is not likely to be based on probe travel rate data. We should be thankful that Waze and other crowd-source input data streams exist.

These findings led us to focus on detecting times when the system's operating condition became degraded; when the travel rates, through demand-induced congestion or disruptive incidents, became unacceptable. As some states and metropolitan areas are doing, setting minimum speed thresholds that are to be achieved during congested conditions, these operating conditions failed to achieve that acceptability requirement. That is, from a policy perspective, the travel rates were too high. This perspective makes sense for the demand-induced congestion conditions. From the TMC operator's perspective, we are asserting that it is their objective to keep the travel rates as low as practical (the speeds as high as possible) by taking actions that offset the effects of heavy traffic, like ramp metering, speed harmonization, changes in toll rates, and alternate route guidance. For incidents, their motivation is not quite

so clear; except that, for incidents that disrupt the traffic flows, the travel rates do degrade, and the TMC operator has a similar objective to take actions that mitigate the impact. That is, in addition to sending emergency response teams to deal with the injuries, etc.; we perceive that the TMC operator will take actions that help improve the travel rate or reduce the duration of unacceptable operation.

Hence, our objective, here, is to create detection algorithms that let the TMC operator know about when demand-induced congestion is about to happen, or when a disruptive incident has just occurred. And we perceived that the sooner the operator is informed the better. In the case of incidents, it is not likely possible to give them a sense that an incident is highly likely. It can be observed as soon as it happens, but foretelling its occurrence is likely to be difficult. But for degraded conditions due to demand-induced congestion, we perceived it ought to be possible to forecast their occurrence and provide a sense of how soon they might occur.

Following this thinking, we sought to identify a method that would 1) identify disruptive incidents when they occur, whether during demand-induced congestion or not, and 2) for degraded operation induced by demand, give the TMC operator a sense of the likelihood that degradation was imminent, and how long it might be until it materialized: a probability that "congested conditions" were going to arise; and how far in the future it would be.

To do this, we first needed to define what we meant by "congested conditions". Of course, we had to make a choice. We found that when the 5th percentile travel rate rose above about 0.9 min/mi, our "eye" told us that the operating condition was congested, as can be seen, at least by inspection, looking back at Figure 3-7. Further, we thought we could test to see if the 5th percentile travel rate was greater than 0.9 min/mi. And if it was, for four of the eight CDFs in a row, then we could assert that the operating conditions were "congested". We are not claiming that this rule is "optimal" yet, or that it is "best", but a by-inspection comparison of the times when this condition is met, and the travel rates look high, suggests it is "good". That is, rarely is there a time when the conditions "look" congested, and this criterion is not met. It is our plan in Phase 2 to refine this rule, "calibrate" its values, and validate its performance.

The facia validity of this test can be illustrated in Figure 3-8. The *x* axis shows date and time for January 25 through January 28 (Tuesday through Friday). The *y* axis is in units of travel rate (min/mi), or the values assigned by the algorithm regarding the induced congestion called "icFlag". The algorithm assigns $icFlag = 0$ when the operating conditions are "normal, uncongested". When it is 2.0, the operating conditions are "disrupted due to an incident". When 2.25, it indicates an "onset of a demand induced congestion". And when 2.5, the operating conditions are demand-induced congestion AND an incident. Displayed in the figure are the 5th percentile travel rate, the 95th percentile travel rate, the spread between these two, and the icFlag value. The flag is mostly zero and occasionally 2.0, 2.25, or 2.5.

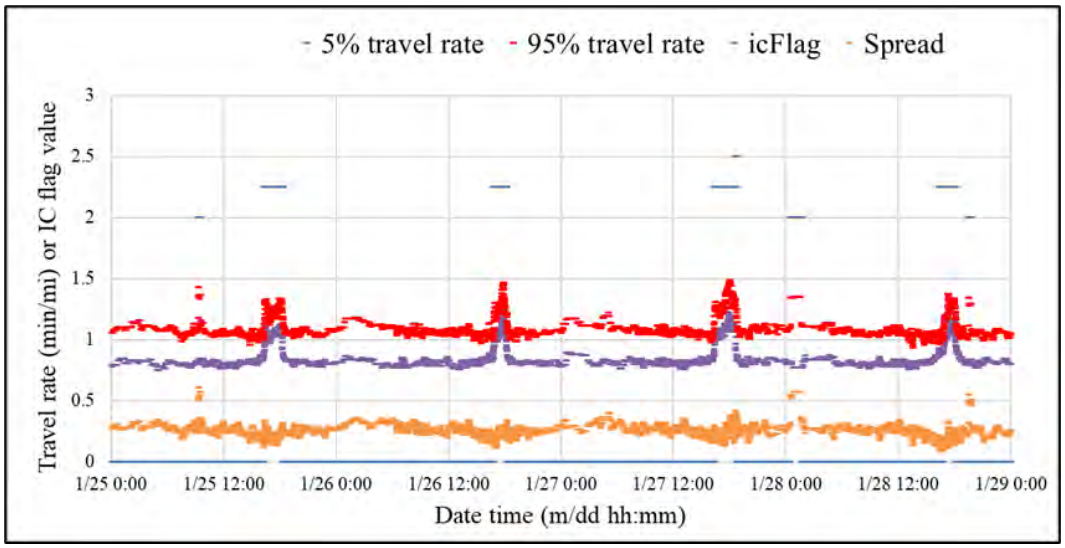


FIGURE 3.8: AN ILLUSTRATION OF OUR DETECTION ALGORITHM'S ABILITY TO SPOT DEMAND-INDUCED CONGESTED OPERATING CONDITIONS

To the "naked eye", it appears that four demand-induced congestions (DICs) occur (one each weekday, Tuesday through Friday). And, as can be seen, for each of these, the icFlag has a value of 2.25. It is also possible to see that the icFlag has spotted three other instances where it thinks an incident-induced congestion (IIC) has occurred.

As explained earlier, we noted, when reviewing the raw probe data, and the sequential CDFs, that the spread in the observations seemed to increase when the flows became disrupted, during either congested or uncongested conditions. It also appeared, "by eye" that when the flow conditions were disrupted, the spread between the 5th and the 95th percentile travel rates were 0.4 min/mi or greater. Remembering that the vertical demarcations in Figure 3-7 are every 0.5 min/mi, the reader can see some evidence that this might be true. We tested 0.35, 0.4, and 0.45 values, and among those three, 0.4 seemed to work best. (Again, a value that will be fine-tuned in Phase 2.)

3.3. Evaluating the Performance of the Detection Algorithm

To check the quality of the detection algorithm, we did a by-hand assessment of false positives, correct positives, false negatives, and miss-classifications. The following metrics were employed:

- Correct positive: The tool signaled an onset of DIC and/or occurrence of IIC and the raw data suggested the same.
- False Positive: The tool signaled an onset of DIC and/or occurrence of IIC and the raw data did not agree.
- False negative: The raw data suggested that a DIC or IIC occurred, and the tool did not indicate anything.

- Misclassification: Both the tool and the raw data suggested that congestion occurred. However, the cause of congestion as appeared from the raw data was different than what the tool indicated.

Table 3-1 shows the results of our assessment for both directions of the corridor. It is divided by time of day to highlight the impact of the headway of vehicle detection (or the temporal span of each group of data) on the accuracy of the algorithm.

TABLE 3-1: EVALUATING THE PERFORMANCE OF THE CONGESTION DETECTION AND CLASSIFICATION TOOL

Direction	Time of day	Congestion type	Number of times the tool alarmed congestion	Correct positive (%)	False positive (%)	Misclassification (%)	False negative (%)
South-bound	Night (10 PM-5:59 AM)	IIC	20	9 (45%)	8 (40%)	3 (15%)	-
		DIC	2	-	2 (100%)	-	-
		DIC+IIC	-	-	-	-	-
	Day (6 AM-9:59 PM)	IIC	31	19 (61%)	5 (16%)	7 (23%)	-
		DIC	51	44 (86%)	7 (14%)	-	-
		DIC+IIC	16	10 (63%)	-	6 (37%)	-
North-bound	Night (10 PM-5:59 AM)	IIC	14	6 (43%)	7 (50%)	1 (7%)	-
		DIC	1	-	1 (100%)	-	-
		DIC+IIC	2	-	1 (50%)	1 (50%)	-
	Day (6 AM-9:59 PM)	IIC	45	21 (47%)	9 (20%)	15 (33%)	-
		DIC	40	35 (88%)	5 (12%)	-	-
		DIC+IIC	33	20 (61%)	2 (6%)	11 (33%)	-

**Values inside the parenthesis are the percentages with respect to the fourth column*

Key observations from this table are listed below.

- The correct positives during the daytime for the DIC flag were 86% southbound and 88% northbound. The false positives were 14% and 12%, respectively. Expectedly, all the DIC flags at night were false positives. The false negative percentage (events missed) was 0%. Undoubtedly, the algorithm can be enhanced and incorporating consideration of the probe headway may be useful.
- The correct positives during the daytime for the IICs were 61% southbound and 47% northbound. To this, we argue we should add the misclassifications. That is, the false positives were 16% and 20%. Again, the false negative percentage (events missed) was 0%.

- Expectedly, the number of DIC+IIC alarms in the nighttime is negligible. The correct positives for this category during the daytime were 63% southbound and 61% northbound. Like IIC, the misclassification percentages in this case were significant as well (37% southbound and 33% northbound), rendering the false positives to a mere 6% northbound and 0% southbound.
- Maybe, expectedly, the algorithm did better in the daytime than at night. We attribute this mainly to the sparsity of observations.
- The percentage of misclassifications, particularly in the IIC and DIC+IIC category is noticeable. The tool seems to be useful for real-time detection of congestion onset given the probe data concentration is adequate (usually in the daytime). However, an individual needs to verify the cause of congestion before taking actions, such as, to deploy an operational treatment to reduce the demand (e.g., ramp-metering) or to alert the first responders about an accident.

3.4. Time Until Congestion Using Probabilities

In many ways, predicting when and if congestion is going to occur is the "ultimate aim" of this part of the investigation. If the algorithm simply indicates it has occurred, that is nice to know, but it does not help with system management. The more important objective is to alert the system manager about impending congestion so actions can be taken to mitigate the impacts.

Here we present some ideas about how to sense that congested operation is about to materialize based on analyzing the 5th percentile of the probe travel rate distributions. The thought is that if the 5th percentile value tends upward as the system works its way toward congested operation, it might be possible to predict that a congested condition lies ahead.

Figure 3-9 shows that this tendency in the 5th percentile is defensible. On the X-axis, for all the congested periods in the northbound direction, is the time until the detection algorithm says that congested operation has started (5th percentile travel rate greater than or equal to 0.9 min/mi for four of the eight CDFs in a row). So, the data have been temporally adjusted so the start of congestion is always at t = 0.

The Y-axis shows the trend in the 5th percentile travel rate before and after congestion commences. Looking to the left of the vertical axis, it is obvious that in some cases there has been an incident leading up to the beginning of congested operation, either "well in advance," as shown by the traces of data points furthest to the left, or closer to the beginning of congestion, as shown by the descending data point traces closer to but to the left of the vertical axis.

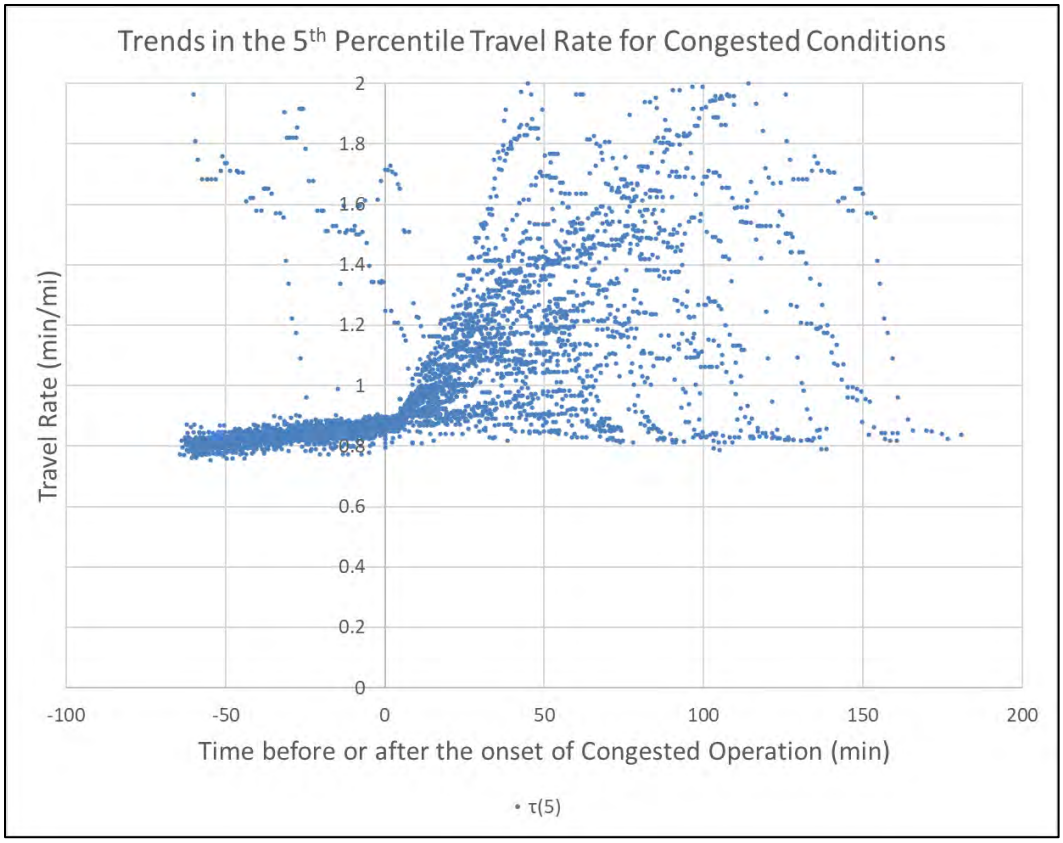
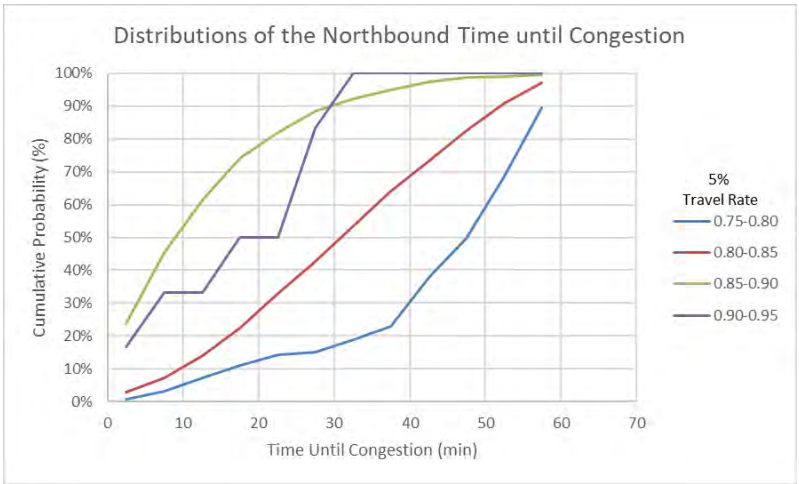


FIGURE 3.9: TRENDS IN THE 5TH PERCENTILE TRAVEL RATE FOR CONGESTED CONDITIONS

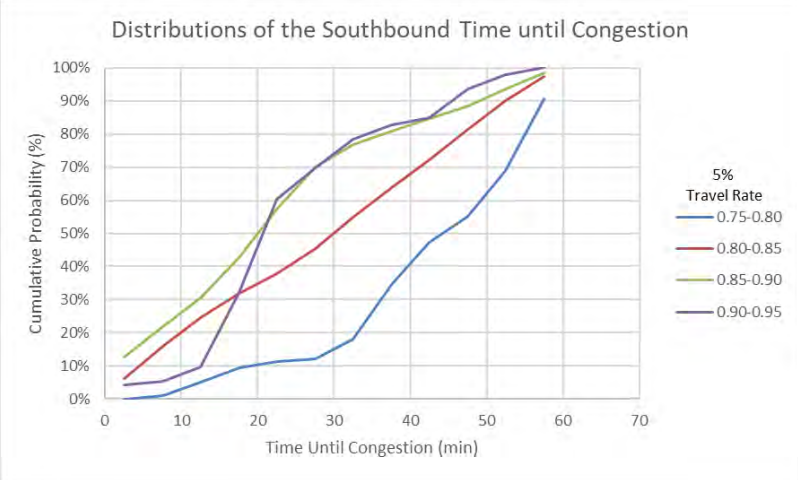
However, the most important observation is that the general trend is for the 5th percentile travel rate to increase constantly, at what appears to be a linear rate as time progresses toward the start of congestion. We captured the 5th percentile travel rates up to an hour before congestion started and the trend is visible that far in advance.

The next question is whether this trend, observable in the aggregate, can be transformed into a useful indicator that congested operation is imminent. Given the stochastic, not deterministic, nature of traffic flows, it seems unreasonable to think that such an indicator will be error-free; but from the perspective of the system operator, if the indicator is "right" most of the time, a leading indication that congested operation lies ahead, is useful.

To explore the ability of the 5th percentile travel rate to predict the onset of congestion we tabulated the value of the 5th percentile travel rates shown in Figure 3-9 and the corresponding time until congestion. We counted the number of times each combination arose. Then we created cumulative distributions for these values. The results are shown in Figure 3-10 for the northbound and southbound directions respectively.



(a)



(b)

FIGURE 3.10: CUMULATIVE PROBABILITY DISTRIBUTION OF TIME UNTIL CONGESTION FOR DIFFERENT RANGES OF 5% TRAVEL RATE VALUES (A) FOR NORTHBOUND (B) FOR SOUTHBOUND

As can be seen, the distribution of the time until congestion involves increasingly smaller (shorter) values as the travel rate rises from 0.75-0.80 min/mi to 0.90-0.95 min/mi. We perceive that the figures can be interpreted as follows. For the northbound direction, if the travel rate is 0.85-0.90 min/mi, there is a 45% chance that congested operation will arise within the next 7 or so minutes. There is a 70% chance that it will arise within the next 15 minutes and an 80% chance that it will occur within the next 20 minutes. Similar interpretations pertain to the distributions for the other travel rate values, with the time until congested operation being longer (further into the future). As can be seen, the distributions for the southbound direction are similar.

3.5. Performance Assessment

This section describes an idea about how to do network performance assessment. It focuses on assessing travel time reliability. This topic has become a major focus for the transportation sector as both researchers and practitioners realize that the service being provided is safe, reliable travel times from one place to another. In fact, reliability has become a key measure of system performance for transportation agencies across the United States (FHWA, 2017a). The Strategic Highway Research Program (SHRP 2) deemed it one of the four factors that should be addressed in capacity expansion decision-making (FHWA, 2017b). In addition, Moving Ahead for Progress in the 21st Century Act (MAP-21), a recent surface transportation law, remanded states to report annually their performance outcome achievements. (FHWA, 2017c, and National Archives and Records Administration, 2017).

We use the instrumented section of I-5 in Sacramento as a case study. We describe the input datasets required, the data fusion process that must occur, how that fusion takes place, the subsequent analysis, and the findings from our analysis.

3.5.1. Idea

Following the lead of SHRP-2 L02, our idea is that reliability can be measured effectively by focusing on the distributions of travel rates that the facility (or system) generates across the span of a "time of interest", which might be a year, a season, a specific week, etc. We focus on distributions of those travel rates, as in the context of probability density functions (PDFs) and cumulative density functions (CDFs). In this case, however, we choose "bins" for the observations that correspond to policies we assume the controlling agency has elected to follow in monitoring the system's performance. For example, it might stipulate that, during periods of demand-induced congestion, the travel speeds (space-mean speeds) will be 45 mph or higher, corresponding to travel rates of 1.333 min/mi or less. List *et al.* (2018) illustrated these ideas in the context of a case study focused on the freeway network in Raleigh, NC.

3.5.2. Methodology

The methodology used by List et al. (2018), which is also employed here, involves eight-steps. They are as follows:

1) *Gather travel rate data*. These are the rates (for example, minutes/mile) at which traffic was traversing the TMC segments during each 5-minute interval for every TMC segment during the study year. In the case of this study network, these data were provided by the Bluetooth probes. Admittedly, the sample is biased in some unknown fashion, because only the probes are included. So, a caveat in the analysis is that the findings are only reflective of the service experienced by these probes.
2) *Gather operating condition data*. Every travel rate observation is produced by an operating condition. The segment of I-5 is subjected to a specific demand (flow rate) for a weather condition plus other influences such as the presence (or absence) of an incident, a work zone, a planned event, lane closures, nearby events, etc. Describing the

operating condition correctly is critical if the cause for the observed performance is to be diagnosed correctly.

3) *Identify Normal / Abnormal Performance*. The next step is to divide the observations into those that have arisen during "normal" operating conditions as to be differentiated from "abnormal" conditions. Normal is seen to be the operating condition when no adverse influences are in play, whether they be weather, incidents, roadwork, special events, or some other activity (e.g., a fire on a nearby property). Put another way, "normal" is the condition that would be "expected" for the segment and time interval to which the observation belongs. Abnormal is everything else. It captures the data collected when the system was being affected by weather, incidents, etc. The reason for separating the two has to do with mitigating actions. Ultimately, the purpose in monitoring performance is to identify actions that make the delivered service better. The analyst wants to identify the causes for the abnormal operating conditions and then identify mitigating actions that can improve that performance and/or eliminate its occurrence.
4) *Label observations with explanations*. This is the process of adding one or more labels to each observation to indicate the operating conditions under which the travel rate was created. This includes flags for adverse weather, incidents (e.g., number of lanes closed), roadwork, etc.
5) *Select / define analysis conditions*. After labeling the observations, the next step is to identify (define) the conditions for which the performance is to be assessed. In the context of the MAP 21 guidelines (FHWA, 2017c), it is the entire year and all operating conditions. Alternately, agencies may be interested more narrowly in the performance under normal PM peaks, AM peaks during adverse weather, AM or PM peaks during incidents, Friday afternoons in the summer when vacation traffic is high, etc. In our case, this time frame is the weekdays between January and March when the Bluetooth data were collected. In any one of these analyses, it is critically important that the analyst and the audience understand these conditions in the same way.
6) *Conduct the analyses*. This step involves developing the distributions that describe the operation of the segment. If a single operating condition is examined (e.g., snow days in the winter), then there may be only one distribution, and the question being answered might be: how did the system perform, followed by: how do we make it perform better? If the focus is on a time period, like the AM or the PM peak, then the operating condition is a "mixed bag". It involves lots of different conditions under which the system was operating. Here, we simply classify those conditions into "normal" and "abnormal" as described earlier. Hence, two distributions are developed, one for the observations during "normal" conditions, and a second for the "abnormal" conditions. The implicit question being posed is: how can the performance under both normal and abnormal conditions be improved? Likely, that means there are two answers, one for each of those conditions. (And for the abnormal condition, there might be multiple answers.)
7) *Draw conclusions* / *prepare recommendations*. This step, which is not within the bounds of this case study, asks the question, what can be done to improve performance. What

the analysis presented here does is to defensibly describe the changes in performance that arise between and among the operating conditions.

8) *Gather additional operating condition data.* This is a step that might prove necessary if the findings look suspect. If something about the distributions looks awry. Experience tells us that this happens when the observations have been mis-classified. That is, data for one or more probe trips, in this case, have been put in the "normal" group while they should have been put in the "abnormal" group. Since the incident data are perishable, it is hard to find the evidence after the event has occurred, probe observations that occurred during the incident are put in the "normal" group because no record of the incident could be found. We have seen two causes for this wrong binning. In the first, not enough operating condition information was collected to correctly label the observation (e.g., there was an incident, but it went unrecorded; or there was an incident on a nearby facility (e.g., a cross-street) and it was not identified. In the second, the incident occurred (e.g., travel rates became abnormally high), but no label was added to indicate that this was the case. The problem with these labeling errors is that they then "taints" the analysis. The picture of "normal" operation that is created includes probe travel rates that occurred during incidents; and vice versa, those during abnormal operating conditions are missing the misclassified observations.

In the case of this I-5 analysis, many of these steps were accomplished during SHRP-2 L02 when the Sacramento I-5 datasets were created, and this analysis benefits from that prior work. The effort that is new involves applying steps 6) through 8), and primarily 6) and 7) to the I-5 data.

3.5.3. Analysis

The analysis begins with the selection of a spatial and temporal context. We choose the spatial limits of the locations of the Bluetooth sensors, which were 5.6 miles apart (Figure 3-1). We will pretend, in the context of MAP21, that this is one TMC. We must do this because the datasets for the pairwise matches involving the in-between monitoring stations lack the ancillary information about weather, incidents, and other external influences. (In Phase 2 we will add this information to those datasets and prepare a study that pretends that the end-to-end section is comprised of three TMCs in each direction.) The posted speed limits in the segment are twofold: 55 mph for trucks and 65 mph for all other vehicles.

Insofar as the time context is concerned, we elected to focus on the weekdays. Moreover, we divided the days into four time periods: early morning (midnight to 5 am), AM peak (5 am to 10 am), midday (10 am to 3 pm), and evening (3 pm to midnight). While this breakdown is logical, other traffic engineers might argue for other choices. Without loss of generality, the methodology still applies.

FHWA's expectation is that the agency responsible for managing the facility will instrument it in such a way that its "delivery of speed" performance can be assessed. Breakpoints are established, based on agency policy, which allow measurement of performance for varying

degrees of acceptability. And then the percent time the facility spends in each of these operating conditions is assessed. We elected to use five categories: 0-15 mph, 15-30 mph, 30-45 mph, 45-60 mph, and 60 mph+. The implicit assumption is that speeds of 45 mph or greater are acceptable during the times of congestion.

We did not follow the FHWA guidance completely. We did not count the number of 5-minute intervals during which the segment had average speeds based on the above breakpoints. Instead, we focused on counts of the number of probes (users) that experienced those ranges of space-based speeds (in actuality, the associated travel rates) to see what performance they experienced. Our sense is that this is the true focus of the FHWA mandate. But most, if not all, TMCs nationwide, lack "user data" to do such an assessment. Rather, they must rely on observations of spot speeds (from system detectors) or average probe travel rates, say from vendors like INRIX or HERE.COM, without complaint, as the data source.

Table 3-2 presents our findings from studying the 144,179 observations of southbound travel rates. As can be seen, the speed range with the most observations are 60 mph and faster. Next most common is 45-60 mph. This is particularly true in the early morning, the AM peak, and the evening. Understandably, during the PM peak, when demand induced congestion occurs, the percentage of travel rates in these two categories is the lowest, although the shift to lower speed categories is not dramatic.

TABLE 3-2: DISTRIBUTIONS OF SOUTHBOUND I-5 INDIVIDUAL PROBE TRAVEL RATES BY TIME-OF-DAY CATEGORY AND OPERATING CONDITION (NORMAL OR ABNORMAL)

	Normal						Abnormal				
SpdRng\TP	Early	AM	Midday	PM	Evening	SpdRng\TP	Early	AM	Midday	PM	Evening
60+	6501	15674	31111	26095	17107	60+	55	608	1221	2579	231
45-60	3774	4892	8245	15398	4623	45-60	46	247	552	2104	76
30-45		56	98	1485		30-45			7	349	1
15-30		9	30	356		15-30				383	
0-15				192		0-15				74	
Total	10275	20631	39484	43526	21730		101	855	1780	5489	308
	Normal						Abnormal				
SpdRng\TP	Early	AM	Midday	PM	Evening	SpdRng\TP	Early	AM	Midday	PM	Evening
60+	63.3%	76.0%	78.8%	60.0%	78.7%	60+	54.5%	71.1%	68.6%	47.0%	75.0%
45-60	36.7%	23.7%	20.9%	35.4%	21.3%	45-60	45.5%	28.9%	31.0%	38.3%	24.7%
30-45		0.3%	0.2%	3.4%		30-45			0.4%	6.4%	0.3%
15-30		0.0%	0.1%	0.8%		15-30				7.0%	
0-15				0.4%		0-15				1.3%	

These data are presented graphically in Figure 3-11. The contrast between the time periods in the break between 60 mph or faster and 45-60 mph is easier to see; and the contrast between normal and abnormal operating conditions is more apparent.

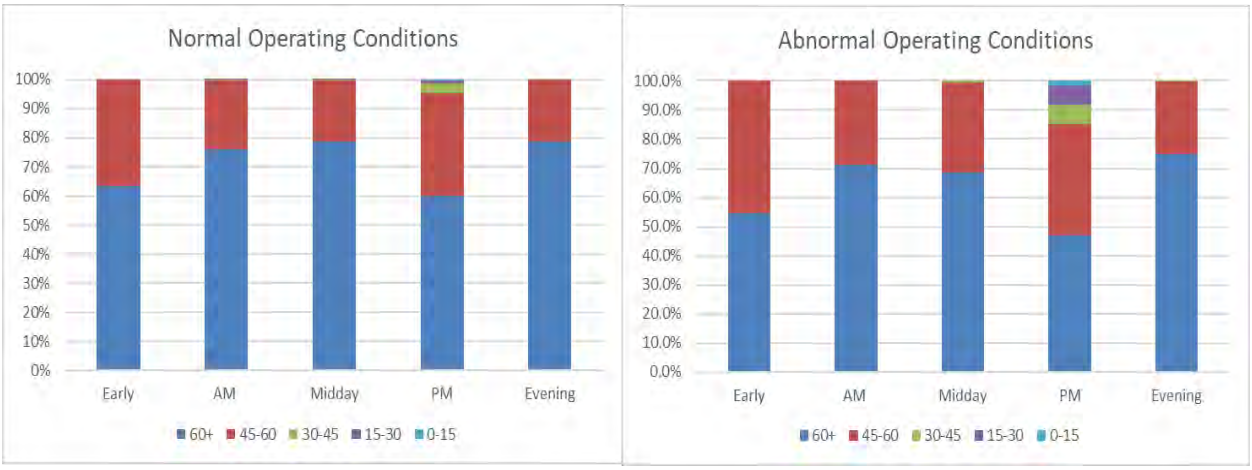


FIGURE 3.11: TRENDS IN THE DISTRIBUTION OF SPACE-BASED SPEEDS DEPENDING UPON THE OPERATING CONDITIONS, NORMAL OR ABNORMAL, AND THE TIME OF DAY ON WEEKDAYS FOR SOUTHBOUND I-5

Table 3-3 presents our findings from studying the 138,333 observations of northbound travel rates. As can be seen, the speed range with the most observations are again 60 mph and faster as was the case southbound. The next most common is 45-60 mph. This is again particularly true in the early morning, the PM peak, and the evening. Understandably, during the AM peak, when the northbound demand induced congestion occurs, the percentage of travel rates in those two categories is the lowest.

TABLE 3-3: DISTRIBUTIONS OF NORTHBOUND I-5 INDIVIDUAL PROBE TRAVEL RATES BY TIME-OF-DAY CATEGORY AND OPERATING CONDITION (NORMAL OR ABNORMAL)

	Normal						Abnormal				
SpdRng\TP	Early	AM	Midday	PM	Evening	SpdRng\TP	Early	AM	Midday	PM	Evening
60+	7603	21666	29643	22882	12860	60+	334	799	817	2214	296
45-60	3742	10274	7294	4799	4044	45-60	96	887	262	520	115
30-45		4587	56	1	1	30-45		803	25	11	
15-30		1634				15-30		45	2	21	
0-15						0-15					
Total	11345	38161	36993	27682	16905		430	2534	1106	2766	411
	Normal						Abnormal				
SpdRng\TP	Early	AM	Midday	PM	Evening	SpdRng\TP	Early	AM	Midday	PM	Evening
60+	67.0%	56.8%	80.1%	82.7%	76.1%	60+	77.7%	31.5%	73.9%	80.0%	72.0%
45-60	33.0%	26.9%	19.7%	17.3%	23.9%	45-60	22.3%	35.0%	23.7%	18.8%	28.0%
30-45		12.0%	0.2%	0.0%	0.0%	30-45		31.7%	2.3%	0.4%	
15-30		4.3%				15-30		1.8%	0.2%	0.8%	
0-15						0-15					

As we have mentioned elsewhere, the shift to lower speed categories is more dramatic in this northbound direction than is observed southbound.

These data are presented graphically in Figure 3-12. The contrast between the time periods in the break between 60 mph or faster and 45-60 mph is easier to see; and the contrast between normal and abnormal operating conditions is more apparent.

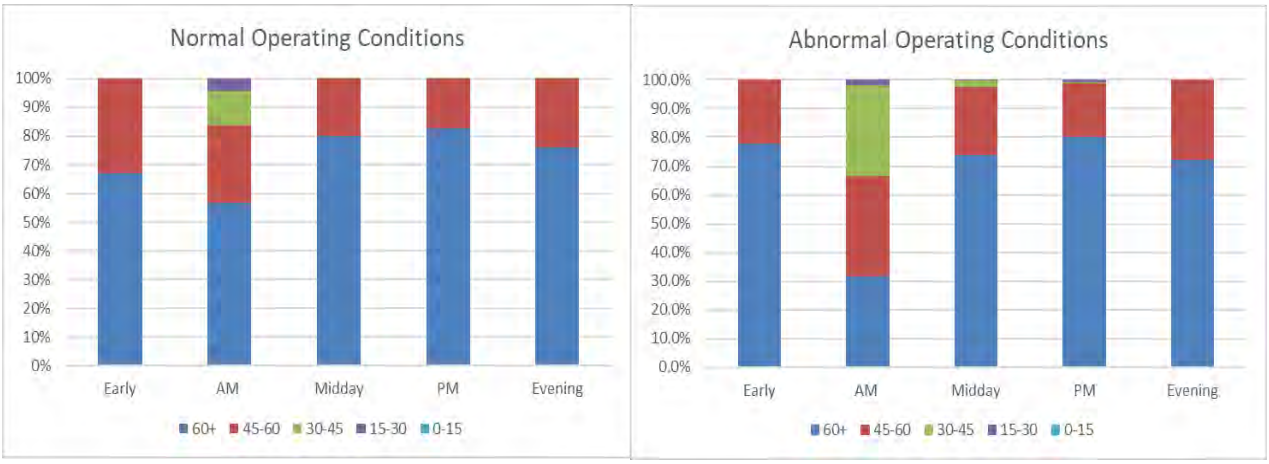


FIGURE 3.12: TRENDS IN THE DISTRIBUTION OF SPACE-BASED SPEEDS DEPENDING UPON THE OPERATING CONDITIONS, NORMAL OR ABNORMAL, AND THE TIME OF DAY ON WEEKDAYS FOR NORTHBOUND I-5

Information like that presented in Tables 3-4 and 3-5 is the expectation of FHWA in the context of the MAP 21 monitoring requirements. A deficiency of the assessment presented in these tables is that they do not examine trip-based travel times or rates that would be experienced by vehicles that traverse this segment. Understandably, no trip uses only these two segments (northbound or southbound).

3.5.4. Conclusions

This section has presented an illustration of performance assessment based on the probe data for I-5 northbound and southbound. The analysis is different from List et al. (2018) and the FHWA mandate in that it focuses on the "service" experienced by individual vehicles (users) as opposed to the service provided by the segment without regard to the number of vehicles (users) involved. It also treats the 5.6-mile segments, both northbound and southbound, as being single "TMC" segments, because of limitations in the finer grained data between all of the four Bluetooth monitoring stations.

3.6. SUMMARY

In this case study, we developed a tool—using travel rate data of probes—that highway system managers can use to detect demand-induced congestion (DIC) and incident-induced congestion (IIC) on freeways. It revealed that for groups of consecutive probes, the faster vehicles tend to slow down sooner than the rest as traffic density gradually increases before DIC arises. Moreover, at the beginning of IIC, the spread in the travel rates increases abruptly. We chose two thresholds through trial-and-error: the 5th percentile travel rate for such a group exceeding 0.9 min/mi, indicating an onset of DIC; and the spread between the 95th and 5th percentiles exceeding 0.4 min/mi, indicating an occurrence of IIC. Meeting both criteria indicates that a disruptive incident happened during a DIC.

A by-hand assessment of the tool revealed that there were no false negatives. Of the DIC alarms in daytime, 86%–88% agreed with the by-hand assessment. This percentage was low for

the IIC alarms in daytime (47%–61%) due to considerable misclassifications. The same is true in the case of DIC and IIC occurring simultaneously. Overall, algorithm did better in the daytime than at night, possibly because of sparsity of observations in the latter case.

The tool, in its current form, is evidently useful for real-time detection of DIC if probe data are available in adequate concentration. Given this pre-requisite on data-concentration is met, its ability to detect congestion in the case of a disruptive incident is not bad either. However, the substantial number of misclassifications implies that an individual needs to verify the flag. In future research, we will incorporate the headway of probes into the algorithm and investigate the sensitivity of the outcomes to the thresholds to improve the performance of the tool, particularly in the nighttime. We should note that the performance of the tool here is its ability to detect the occurrence of DIC, not predicting a trend toward its occurrence. in terms of how early it can detect a DIC is not evaluated here. We will incorporate that measure in our future research where the tool will be evaluated based on a more robust approach.

4. TAMPA-HILLSBOROUGH CASE STUDY

Connected vehicles (CV) will provide an important source of data to support real-time management of traffic operation and off-line analysis of traffic operations. CV data allows the derivation of metrics not currently computed for use in freeway management such as the standard deviation of speed, acceleration/deceleration, and jerk. This study explores the utilization of CV data to derive such metrics for use in traffic management. The study then investigates the use of the derived metrics in combination with data analytic techniques to assess and predict the onset of congestion on freeways in real-time operations.

CV data will be an important source of information to support real-time management of traffic operations. They will also be useful for off-line analysis of traffic operations. Data from CVs are transmitted using messages communicated utilizing cellular communication (C-V2X) or dedicated short-range communication (DSRC) to roadside receivers. The CV message formats are specified in the Society of Automotive Engineers (SAE) J2735 standards (SAE International 2016) and various SAE J2945 standards. The basic safety message (BSM), specified in J2735, which contains vehicle safety-related information, are broadcasted to surrounding vehicles, but can be also captured by the infrastructure. The BSM, as defined in the J2735 standards, consists of two parts. Part 1 is sent in every BSM message broadcasted 10 times per second. It contains core data elements, including vehicle position, heading, speed, acceleration, steering wheel angle, and vehicle size. BSM Part 2 consists of a large set of optional elements such as precipitation, air temperature, wiper status, light status, road coefficient of friction, Antilock Brake System (ABS) activation, Traction Control System (TCS) activation, and vehicle type. However, a large proportion of these parameters are currently unavailable from every vehicle, and they are not expected to be available. Connected vehicle data can be captured by a roadside unit or can be sent to the cloud for processing and use.

Some of the measures that can be estimated using CV data includes travel time, origin-destination, vehicle classification, queue length/back of queue, stops, accelerations and decelerations, standards deviations of speeds, intersection movement-level delays and queues, near-misses, emission, and route choice.

Zou et al. (2010) used simulation to examine the accuracy of travel time estimation based on CV and found an average error percentage of 27.6%, 12.5%, and 8.2% for 1%, 5%, and 10% market penetrations, respectively. Hadi et al. (2018) assessed the quality of travel time estimation based on CV data and found that a low market penetration (1%-2%) is generally sufficient to produce an error that is lower than 10% for high volume urban freeway segments. For urban street segments, however, this data quality cannot be achieved until the market penetration of CV exceeds 10%-15%. Hadi et al. (2020) examined the use of detailed metrics in combination with the usually used macroscopic metrics for the estimation and prediction of traffic safety and mobility. The utilized disturbance metrics are the standard deviation of speed between vehicles, standard deviation of speed of individual vehicles, acceleration, jerk rate, number of

oscillations and a measure of disturbance durations in terms of the time exposed time–to–collision (TET). The authors used data clustering for better off-line categorization of the traffic states. These measures can be used in the calibration and validation of simulation models.

There has been limited use of data from large-scale CV deployments in estimating and predicting performance measures. Many of the existing studies use simulation outputs to examine the use of CV to examine this estimation and prediction. Recently, CV data became available from the Connected Vehicle Pilot Deployment Program sponsored by the USDOT, which is a national effort to deploy, test, and operationalize CV applications (USDOT 2018). The USDOT selected three sites for the pilot CV deployment. These sites are the New York City, New York; Wyoming; and Tampa-Hillsborough Expressway Authority (THEA), Florida sites.

This study demonstrates the use of CV data to estimate performance measures using a freeway segment that is part of the Tampa CV Pilot Deployment Site. The Tampa Connected Vehicle Pilot has equipped buses, streetcars, and privately-owned vehicles with CV technology to enable them to communicate information with each other, as well as with infrastructure and pedestrians who use a smartphone app (Vadakpat, 2018; USDOT, 2017). The pilot aimed at deploying onboard CV units on 1,600 privately owned vehicles, ten buses, and 10 streetcars. Forty roadside units were installed at the busiest intersections of the pilot area. The remaining of this chapter will discuss the Tampa Case Study used in this project and the initial results obtained from the investigation.

4.1. DESCRIPTION OF THE CASE STUDY SITE

A 1.6-mile portion of the northbound direction of the Selmon Expressway in Tampa, Florida that is contiguous to the downtown area was selected as case study for the analysis since this is the only freeway segment covered by the CV deployment. THEA owns and manages the Selmon Expressway including the Reversible Express Lanes. Selmon Expressway constitutes an important commute which endpoint intersect various major arterials into and out of the downtown commercial business district in Tampa, Florida. It has been observed that drivers may experience significant delay during the morning and PM peak in this area. Figure 4-1 shows the study area as it related to the THEA CV pilot location.

FIGURE 4.1: CONNECTED VEHICLE PILOT DEPLOYMENT IN DOWNTOWN TAMPA

The THEA pilot includes the deployment several vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) applications. employing Dedicated Short-Range Communication (DSRC) to enable data transmissions (See Table 4-1).

TABLE 4-1: THEA PILOT SITE CV DEVICES

Tampa (THEA) – Devices	Estimated Number
Roadside Unit (RSU)	47
Vehicle Equipped with On-Board Unit (OBU)	~1,000
HART Transit Bus Equipped with OBU	10
TECO Line Street Car Equipped with OBU	8
Total Equipped Vehicles	~1,018

Source: USDOT

4.2. DATA SOURCES

The data used in this project has been collected from two sources. First, CV data was acquired from the U.S. DOT ITS Connected Vehicle Pilot Sandbox for the THEA pilot. The available data from this source consists of records with a resolution of 1/10th of a second including BSM (Basic Safety Messages), SPAT (Signal Phasing and Timing Massages) and TIM (Traveler Information Massages) files which are available via either the web interphase to download individual batched data files or, by using a Sandbox export script in python. The study also utilized data from ClearGuide which is an analytic platform developed by Iteris that provides average speeds aggregated into a 5-minute resolution. In both cases, data belonging to a pre-pandemic period

(January-December, 2019) were analyzed utilizing the records from the PM peak period (15:00 – 19:00).

4.3. PERLIMINARY ANALYSIS

4.3.1. Identification of Congested Days during the Study Period

The selected segment is not congested in every day of the year. This study first conducted a preliminary analysis to identify the congested days during the year 2019 utilizing data from ClearGuide. The data in the Iteris ClearGuide system for the Tampa area is collected from a third party vendor that provided travel time/speed estimates based on tracking GPS-enabled devices. The provided information in the system includes different performance measures and information including speed, travel time, reliability, data quality, volume-based data, and regional dashboard data of the roadways. From the CV data and ClearGuide data, the study found that the northbound direction has congested days in the PM peak period. Thus, further analysis was performed using the data for the northbound direction of the freeway segment. For this part of the project, the average flow and the average travel time measurement in ClearGuide were used to analyze the congested hours of different days for the study roadway segment. These data have been collected for 365 days a year and twenty-four hours a day in 2019. To identify the congested hours of different days, first the cumulative distribution function (CDF) of average travel time and average traffic flow were calculated. The identification of these hours was important to explore potential parameters to identify and predict congestion based on CV data also. Further analysis was conducted to identify the most congested month and week in 2019 based on the average speed data obtained from ClearGuide for the segment under study.

4.3.2. Data Preprocessing

Once the most congested days were identified, the corresponding records from the CV data were obtained from the USDOT ITS Connected Vehicle Pilot Sandbox. The BSM files were formatted into CSV files. In its original form, the retrieved data contains records that belongs not only to Selmon Expressway but to all the Tampa Downtown area (See Figure 4-2).

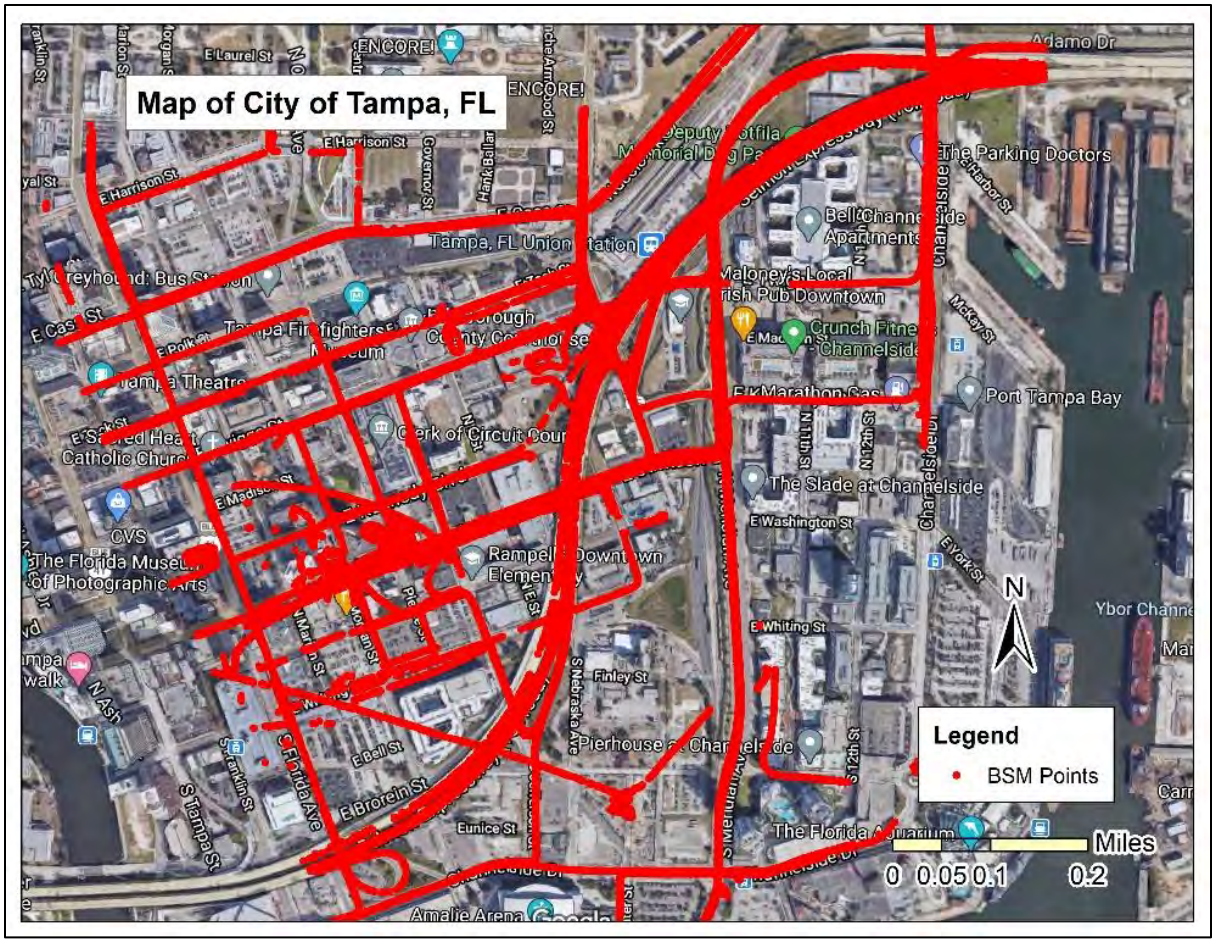


FIGURE 4.2: DOWNTOWN AREA OF THE CITY OF TAMPA WITH VISUALIZATION OF THE BSM DATA POINTS

A python script was then utilized to isolate the data points for the segment of interest from those of the rest of the network. The utilized script allows the definition of a geometry that defines the segment study area. The corresponding data points within the outlined geometry are then isolated for further analysis. Figure 4-3 shows the isolated data points belonging to a subsegment of the study segment selected as the focus of the case study analysis. The selected segment for the analysis corresponds to the last 300 ft segment of the 1.6-mile segment case study of the Selmon Expressway in the northbound direction. This segment was selected due to the more frequent congestion occurring on this subsegment.

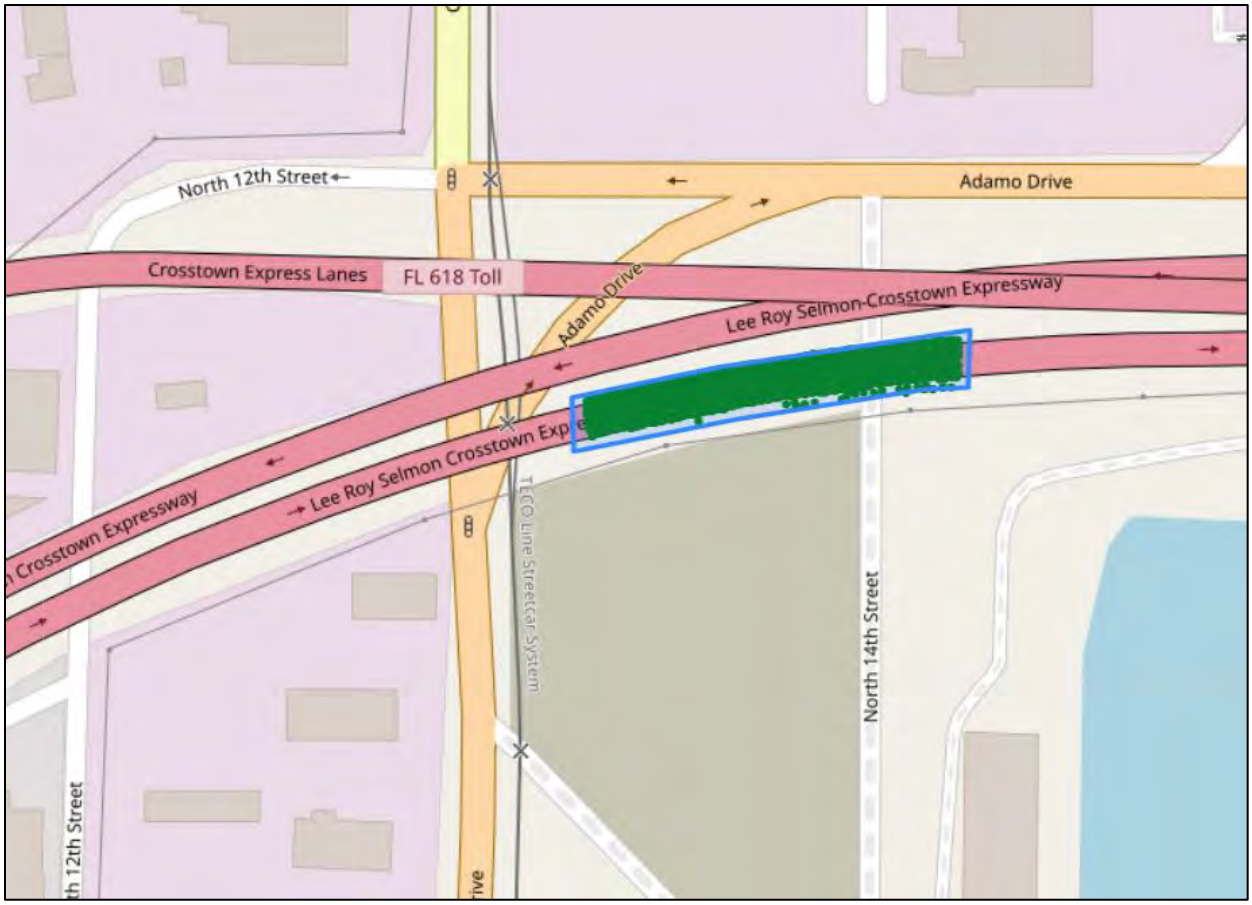


FIGURE 4.3. DATA POINTS SPECIFICALLY EXTRACTED FOR A SUBSEGMENT BASED ON THE OUTLINED GEOMETRY

In order to get a better understanding of the CV data, one of the first tasks involved analyzing the number of records available in the BSM database in terms of the number of readings (records) available per second, as well as the number of records available per individual vehicle from all the vehicles traversing the study segment. Figure 4-4 shows the CDF plot that depicts the distribution of the reading counts in vehicles per second. The median is shown to be close to five. On the other hand, Figure 4-5 shows that the average count of readings by ID is around 40 readings throughout the 300 ft segment whereas the 50th percentile is around 70 readings by ID.

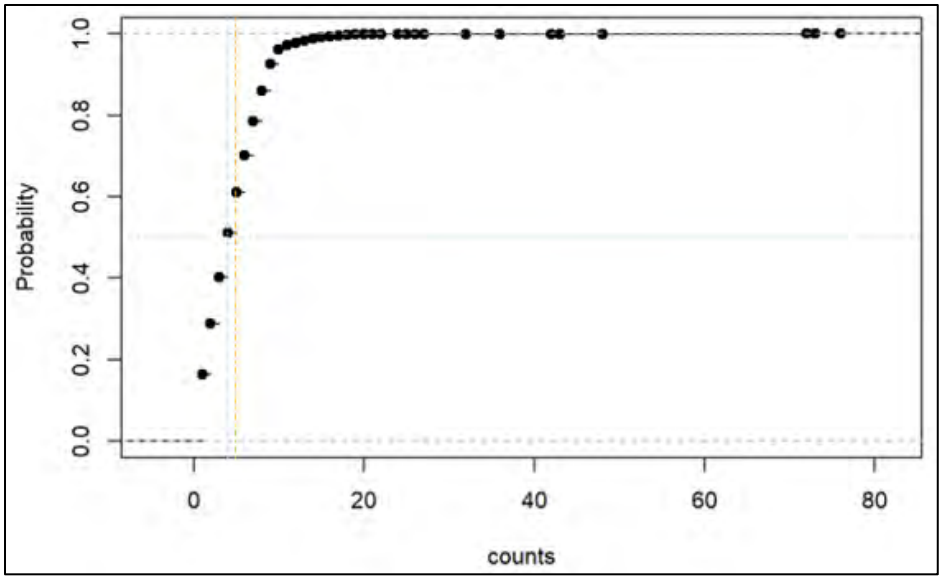


FIGURE 4.4: CDF OF READING COUNT IN VEHICLES PER SECOND

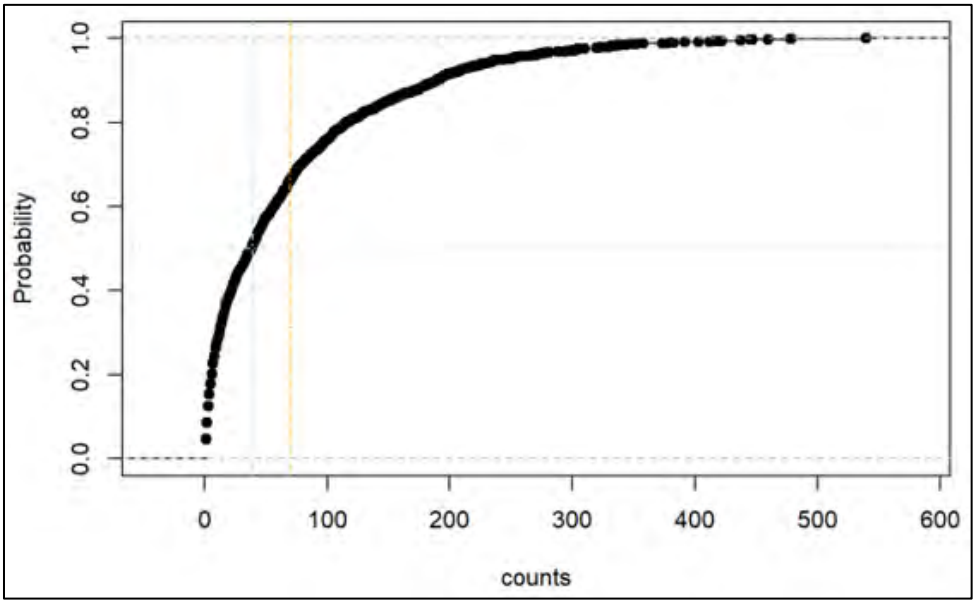


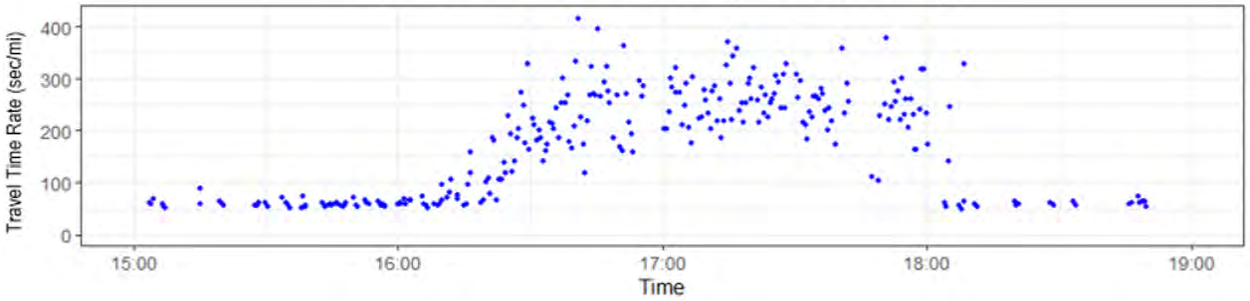
FIGURE 4.5: CDF OF READING COUNTS BY ID

4.3.3. Utilization of Standard Deviation to Predict the Onset of Congestion

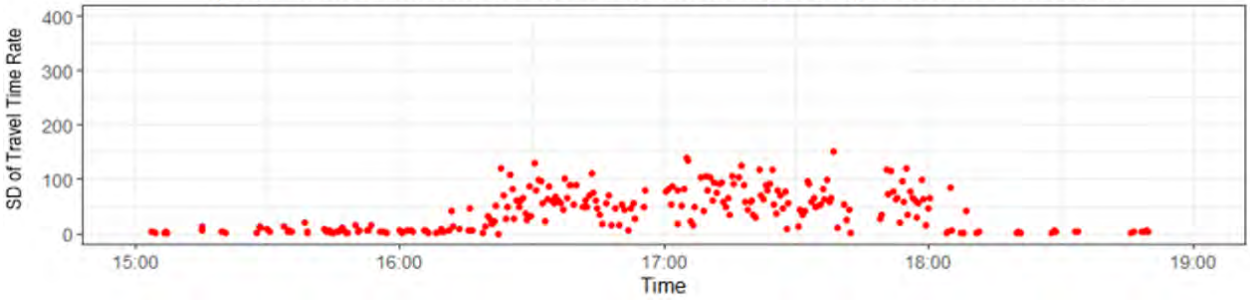
One of the objectives of the study is to explore alternative variables that could be useful in real-time operations to identify the occurrence of congestion ahead of time (i.e., variables that could

present significant variations before the actual congestion occurs) or else, variables with significant predictive power such that they could be incorporated in a machine learning model capable of predicting congestion. In that sense, one of the first explored approaches was to use the standard deviation of speed and travel time rates. Multiple approaches to compute the standard deviation were utilized. For example, the exploration included the use of standard deviation of data points using bins of a fixed time length (e.g., 30 seconds, 1 minute, 5 minutes), and the standard deviation of travel time rates using a moving window of specific sizes in terms of the number of data points in the window (i.e., a rolling standard deviation for consecutive data points groups composed of "n" data points). Figure 4-6 shows a group of time series charts that compare the mean of the travel time rates aggregated every 30 seconds (Figure 4-6a, the standard deviation between the data points also estimated in 30 seconds size bins Figure 4-6b, and the rolling standard deviation for every 50 data points (moving window group size, n=50) with an overlap of twenty five data points Figure 4-6c. This means that consecutive groups of 50 data points are used to compute the standard deviation with 25 data points in common (overlap).

a) Time series of travel time rates (30 seconds average)



b) Time series of standard deviation: standard deviation of dp within 30 seconds bins



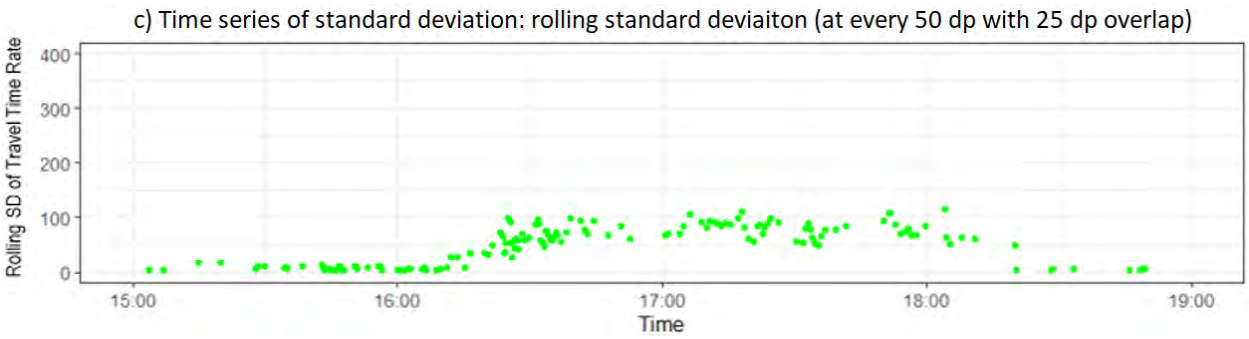


FIGURE 4.6: TIME SERIES OF MEAN AND STANDARD DEVIATIONS OF TRAVEL TIME RATES

Significant variations in the standard deviation are present in both Figure 4-6b (fixed bin standard deviation) and Figure 4-6c (the rolling standard deviation). And it seems that that the standard deviations start increasing just about the same time as the travel time rates start increasing, that is, such variations in the standard deviations start showing practically at the same time, which is also the point when the travel time rates increase due to the onset of congestion as shown by Figure 4-6a.

4.3.4. Connected Vehicles vs Vendor's (HERE) Data

Another concern addressed by this study is the comparison of the speeds or travel rates between CV data and vendor's data on the selected segment. The comparison of these two data sources was done to validate the potential of using CV data to estimate travel time compared to the data source currently used in the region. It is known that ClearGuide utilizes HERE data. The evaluation of the two data sources was made by comparing their distributions by plotting them together in a CDF chart. Given the difference in the resolution of the two datasets (5 minutes from ClearGuide vs. 1/10th of a second from the CV data), it was necessary to compute the average travel time rate per vehicle for a bin size of 5 minutes using the CV data to make it comparable with the vendor's data (ClearGuide). Figure 4-7 shows the CDFs of the two data sources and indicates that the travel times measured based on the CV data are lower than those estimated based on the vendor data for the measurements that are higher than the 60th percentile. The two CDFs seem to be close to each other though they show a difference of about two seconds per mile that remain constant up to the 60th percentile where the two distributions start splitting apart. In this case, the red line (travel rate based on CV data) appears to depict significantly higher travel time rates relative to the vendor's data (blue line). As Figure 4-7 shows, the horizontal dotted line on the top represents the 90th percentile, where the two data sources show a difference on their travel rates of approximately 45 seconds per mile. Figure 4-8 shows a time series of the travel time rates that also confirms this pattern described above by showing a substantial difference in the travel time rates between the vendor's data vs the travel rates from CV data during the most congested hour during the PM peak. During this hour, the red dots (Travel rates from CV data) show significantly higher values relative to the blue triangles (travel rates based on vendor's data). The difference between the travel time rate measurements from the two sources could be due to the use of a smoothing algorithm in the vendor's data. Such algorithms have been used by vendors to reduce the noise in travel time measurements due to small sample sizes.

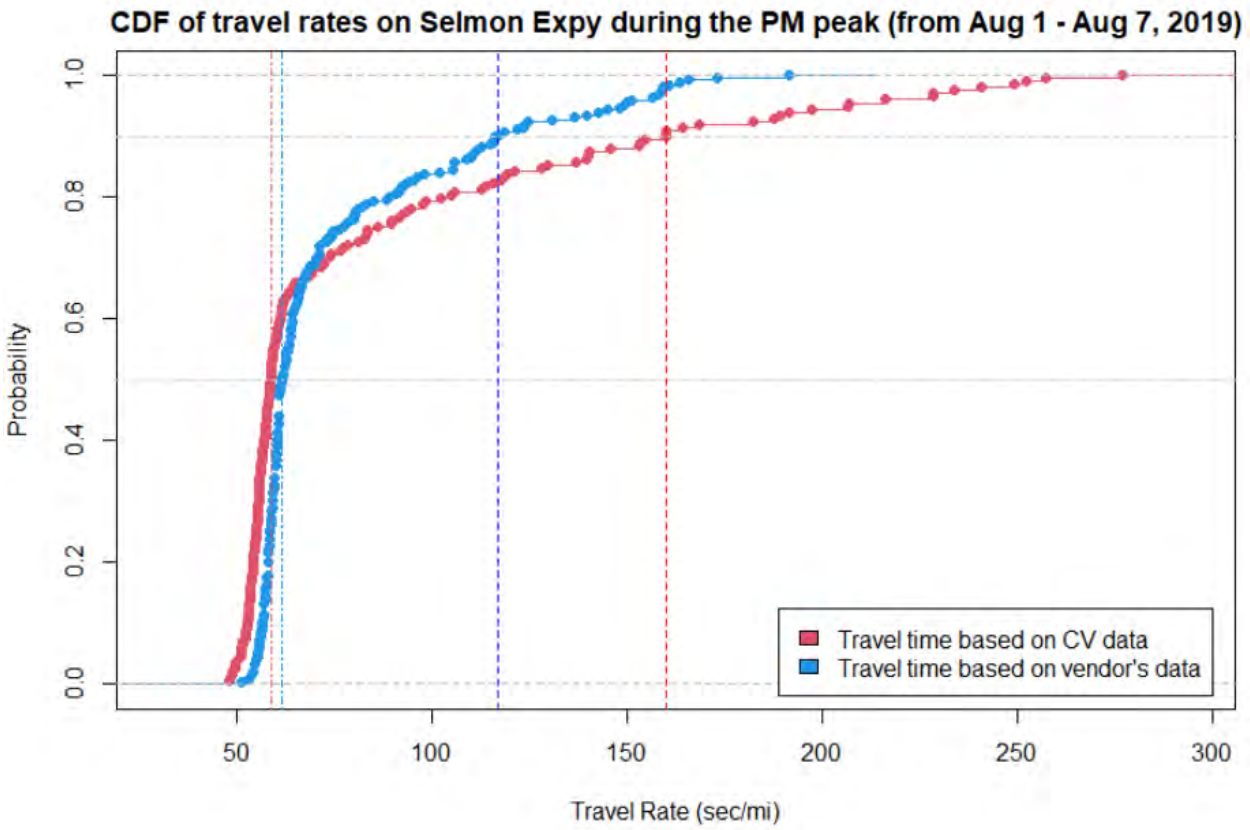


FIGURE 4.7: CDF OF TRAVEL RATES FROM CV DATA COMPARED TO THE CDF OF TRAVEL TIME RATES BASED ON VENDOR'S DATA

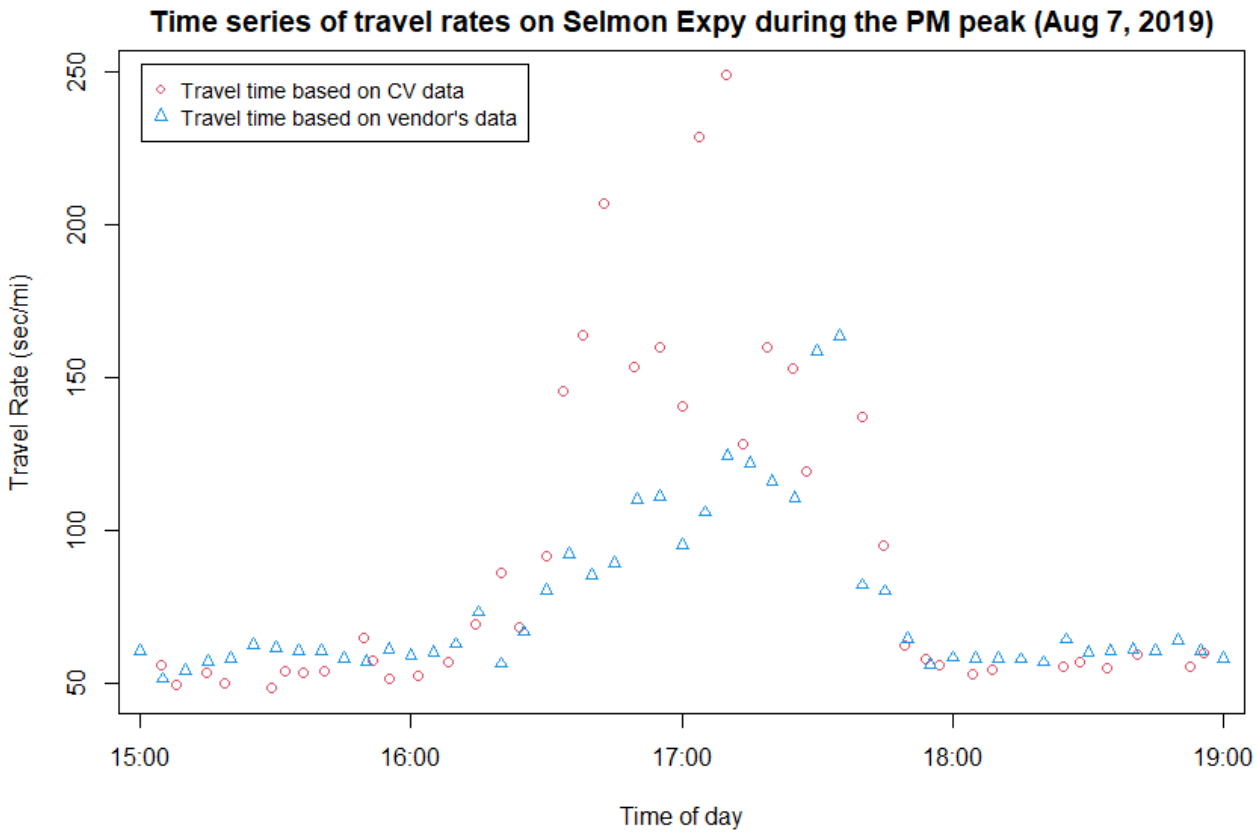


FIGURE 4.8: TRAVEL TIME RATES BASED ON CV DATA COMPARED TO THOSE BASED ON VENDOR'S DATA

Lane-by-Lane Analysis

The next step of the analysis consisted of the implementation of a Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm for the identification of the CV data points for each lane of the subsegment under study. The purpose of identifying the data points from each lane was to perform a comparison of the travel time rates as well as the acceleration distributions for each lane to determine if there are significant trends that are unique for a specific lane during the peak hour. The DBSCAN algorithm utilized the geodetic coordinates of each data point to identify the areas with high-density (e.g., data points in separate lanes) that are separated from each other by low-density areas (e.g., space between lanes). The density-based approach allows the differentiation of data points as being in the interior of a dense region which are known as core points, on the edge of a dense region (or border points), and in a sparsely occupied region (noise).

The DBSCAN algorithm can be summarized as follows:

1. Label all points as core, border, or noise points
2. Eliminate noise points
3. Put an edge between all core points within a distance *Eps* of each other.
4. Make each group of connected core points into a separate cluster.
5. Assign each border point to one of the clusters of its associated core points.

The DBSCAN algorithm worked pretty well to identify the data points belonging to each one of the lanes in the 300 ft segments. An example of the visualization of the output of the algorithm is shown in Figure 4-9.

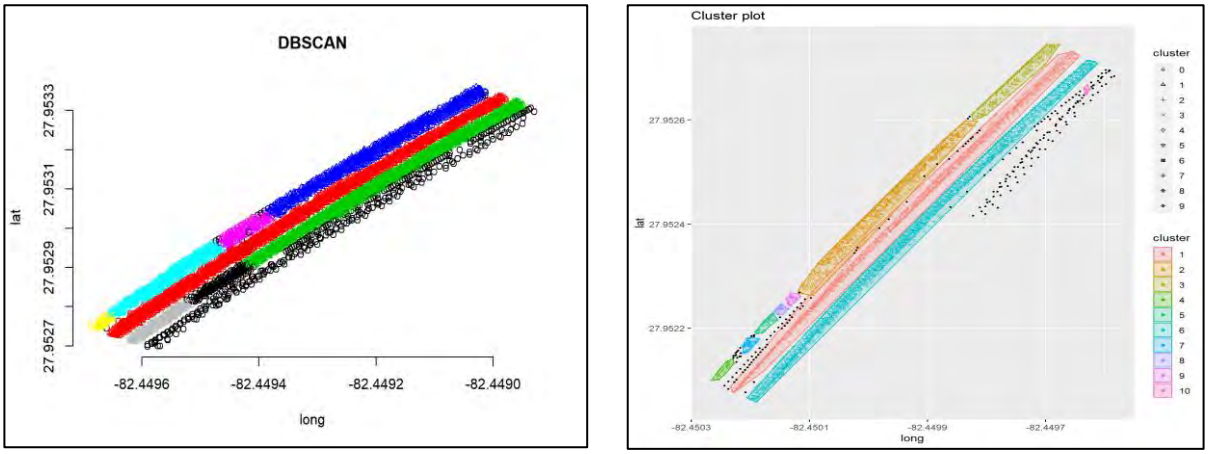


FIGURE 4.9: EXAMPLE OF THE OUTPUT OF THE DBSCAN ALGORITHM TO IDENTIFY THE DATA POINTS BY LANE

Once the data points were labeled according to their respective lane, a comparison of the distribution of the travel rates from each lane was performed by visualizing their respective CDFs as shown in Figure 4-10. The lanes of the three-lane freeway segment of the case study were identified as Lanes 1, 2 and 3, with Lane 1 is the lane on the median side, and Lane 3 is the outer lane (or the rightmost lane). Figure 4-11 shows a time series of the travel time rate during the PM and Figure 4-12 shows a time series of the acceleration/deceleration by lane for the same period. Usually, the lane on the median side ~~is~~ the fast lane. However, as Figures 9 and 10 depict, this is not the case with the studied segment where Lane 3 (the outer lane) displays a distribution with much lower travel time rates (higher speeds) relative to Lane 1 (median lane). This phenomenon may be due to the disturbance and queuing that occur on Lane 1 from the exit to the express lane that occurs one mile downstream of the study segment that is on the left side of the highway (accessed by the vehicles positioned on the median lane, which is Lane 1.

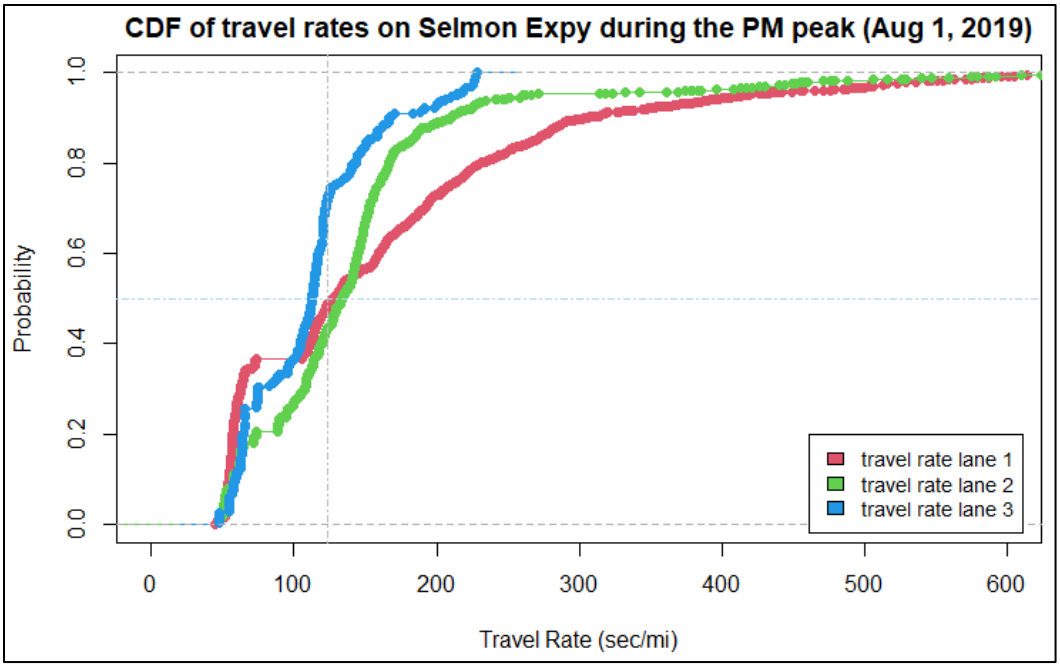


FIGURE 4.10: CDFS OF TRAVEL TIME RATE BY LANE DURING THE PM PEAK

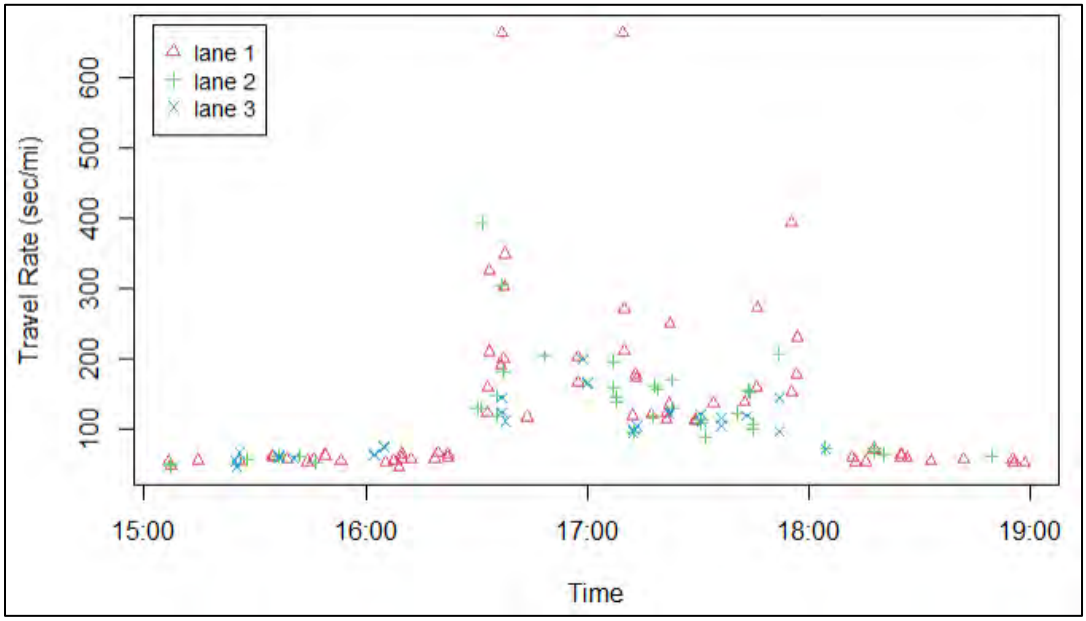


FIGURE 4.11: TIME SERIES OF TRAVEL TIME RATE ON THE STUDY SEGMENT DURING THE PM PEAK

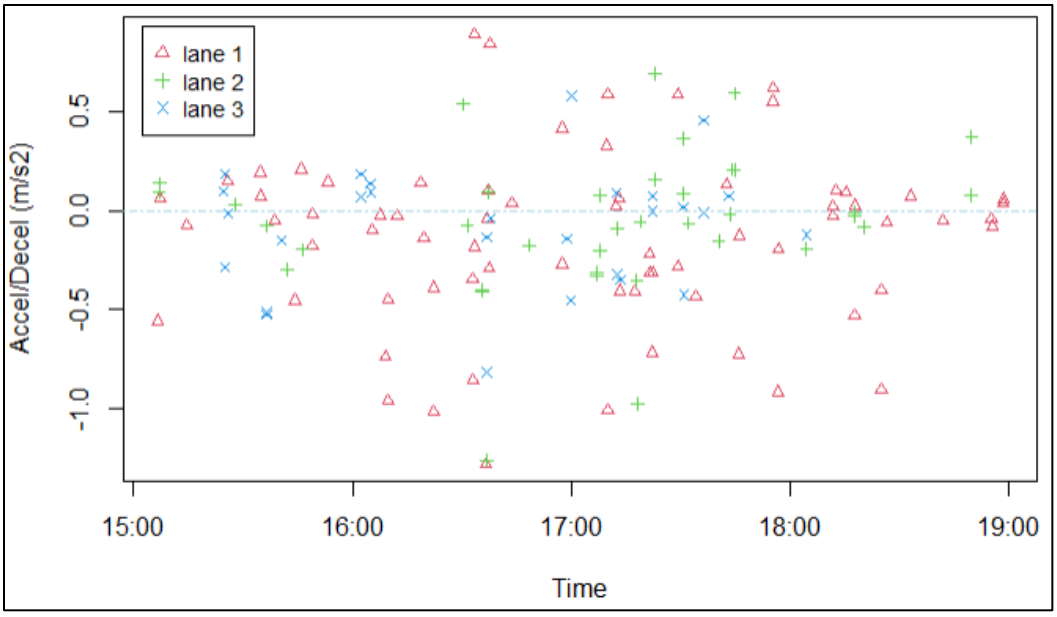


FIGURE 4.12: TIME SERIES OF THE ACCELERATION /DECELERATION DURING THE PM PEAK ON THE STUDY SEGMENT

Figure 4-11 confirms the pattern initially depicted by Figure 4-10 regarding the travel rates on Lane 1. That is, the travel time rates on Lane 1 (median lane) are slightly higher than the travel rates on the other two lanes during the most congested period during the PM peak (around 16:45 to 18:00). On the other hand, Figure 4-12 does not show any clear patterns that can differentiate between the three lanes in terms of their acceleration/deceleration values throughout the PM peak.

4.3.5. Additional Variables Derived from Connected Vehicles Data

The BSM files allowed the estimation of speed/travel time rate and acceleration/deceleration as described in the previous section. However, additional metrics were derived in this project to study the traffic dynamics and the relationship of this metrics with mobility and safety. The derived metrics include the following:

- **Standard deviation of speed:** The standard deviation of speeds is related to the shockwave and platoon formation during the onset of congestion. Thus, the standard deviation of speed can be an important measure in assessing and predicting safety and mobility. For the purposes of the analysis, the data were first aggregated into bins of five minutes each. Three different variants for the computation of the standard deviation of speed were calculated for each five-minute bin as listed below.
  - *The standard deviation between data points (SDdp)* is computed as the standard deviation of all the data points contained in a five-minute bin. SDdp represents a disturbance metric that captures the variability of the speeds of all the data points contained in a bin (of five minutes) regardless of if these measurements are coming from one vehicle or multiple vehicles. The calculation of SDdp is as in equation 4-1 below.

- $SDdp_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(S_i - \bar{S})^2}$, (Eq. 4-1)

  where, $SDdp_x$ is the standard deviation between data points contained in a bin of size x; n is the number of data points within the bin of size x; S_i is the speed for the i_{th} record, and $\bar{S}$ is the average speed from all the speed records in the bin of size x.

- *The standard deviation of individual vehicles (SDv)* is the standard deviation at of the speeds records from an individual vehicle traversing the segment. This metric captures the disturbance of the traffic flow by reflecting the variability of the speeds for each individual vehicle. The computation of *SDv* is based on equation 4-2 below.
  - $SDv_j = \sqrt{\frac{1}{n_j-1}\sum_{i=1}^{n_j}\left(S_{ij} - \bar{S}_j\right)^2}$, (Eq. 4-2)

    where, SDv_j is the standard deviation of the speeds of vehicle j; n_j are the number of speed records available for vehicle j; S_{ij} is the i_{th} speed record from the j_{th} vehicle; and $\bar{S}_j$ is the average speed of vehicle j.
- *The standard deviation between individual vehicles (SDbv)*. This is the standard deviation between the average speeds of individual vehicles that traverse the study segment in a five-minute period (bin size). Equation 4-3 provides the computation of this metric.
  - $SDbv = \sqrt{\frac{1}{n_k-1}\sum_{i=1}^{n_k}\left(\bar{S}_{jk} - \overline{S_k}\right)^2}$, (Eq. 4-3)

    where, n_k is the number of individual vehicles within the five minutes bin; $\bar{S}_{jk}$ is the j_{th} average speed record within the group or bin from k vehicles, and $\overline{S_k}$ is the average of the mean speed from all k vehicles in each group or bin.
- **Jerk: Jerk** is the second derivative of velocity. It indicates the rate of change of acceleration as expressed in Equation 4-4.

$$j(t) = \frac{da(t)}{d(t)}, \quad \text{(Eq. 4-4)}$$

where, j represents Jerk; a represents acceleration/deceleration, and t represents the time.

Multiple plots were produced in an effort to explore the relationship between the proposed metrics. As depicted in Figure 4-13, the SDdp and SDv increases significantly with the decrease in speed.

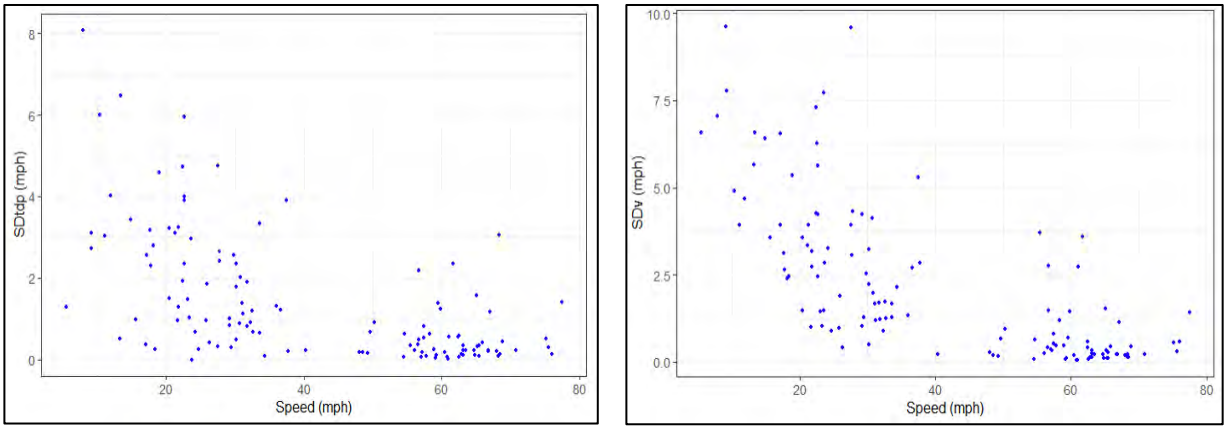


FIGURE 4.13: RELATIONSHIP BETWEEN SPEED AND STANDARD DEVIATION BETWEEN DATA POINTS (SDDP), AND STANDARD DEVIATION BETWEEN INDIVIDUAL VEHICLES (SDV)

Figure 4-14 on the other hand, depicts the relationship between the Acceleration and Jerk with respect to speed. Again, Figure 14 shows that both the acceleration/deceleration and jerk tend to display a wider range of values during lower speeds (below 35 mph).

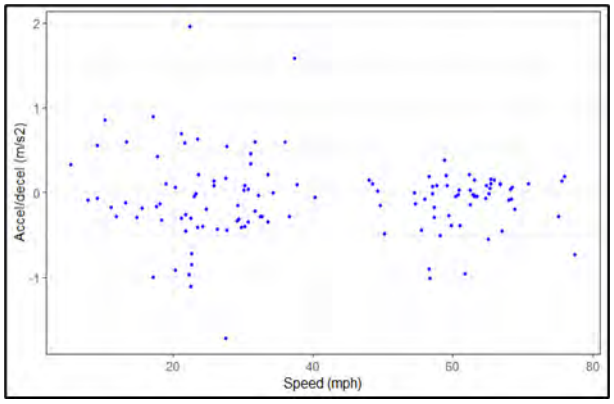


FIGURE 4.14: RELATIONSHIP BETWEEN ACCELERATION/DECELERATION AND JERK WITH SPEED

To continue exploring the relationships between the proposed variables, a scale color for the data points was added as an additional dimension to the plots. The data points were classified into four speed regimes including: 0-20 mph, 21-40 mph, 41-60 mph, and > 60 mph. Figure 4-15 shows the relationship between the SDv and SDdp, this time including an additional dimension to show the speed information as well. Figure 4-14 shows that for the higher speed groups (i.e., 41-60 mph, and > 60 mph), the SDv and SDdp seem to be highly correlated. However, as the mean speed drops below 40 mph the data points spread out. Moreover, there are some cases where relatively high values of SDv exist for a corresponding low value of SDdp perhaps indicating that the vehicles are generally moving at similar speeds, but each vehicle has high speed variations due to stop and go operations.

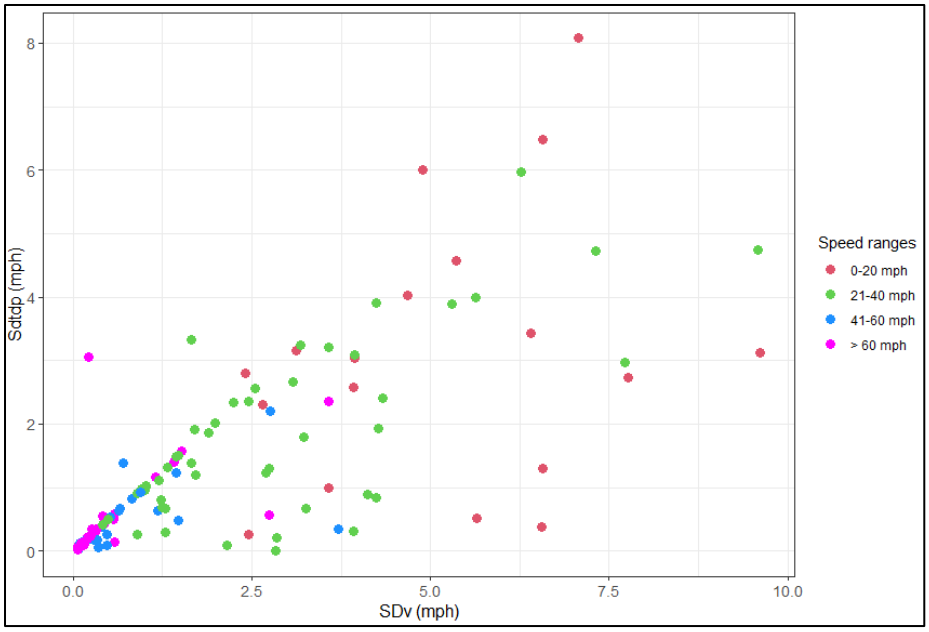


FIGURE 4.15: RELATIONSHIP BETWEEN STANDARD DEVIATION OF INDIVIDUAL VEHICLES (SDV) AND STANDARD DEVIATION BETWEEN DATA POINTS (SDDP)

Figure 4-16 shows a similar relationship between the Standard deviation of individual vehicles (SDv), and the standard deviation between vehicles (SDbv). One evident pattern that can be highlighted by looking at Figure 4-16 is that the standard deviation between vehicles (SDbv) tend to occasionally show higher values as the standard deviation of individual vehicles (SDv) increases. Again, this seems to be characteristic of the lower speed regimes (21-40 mph, and 0-20 mph), which may imply traffic instability characterized by the propagation of traffic oscillations during traffic congestion.

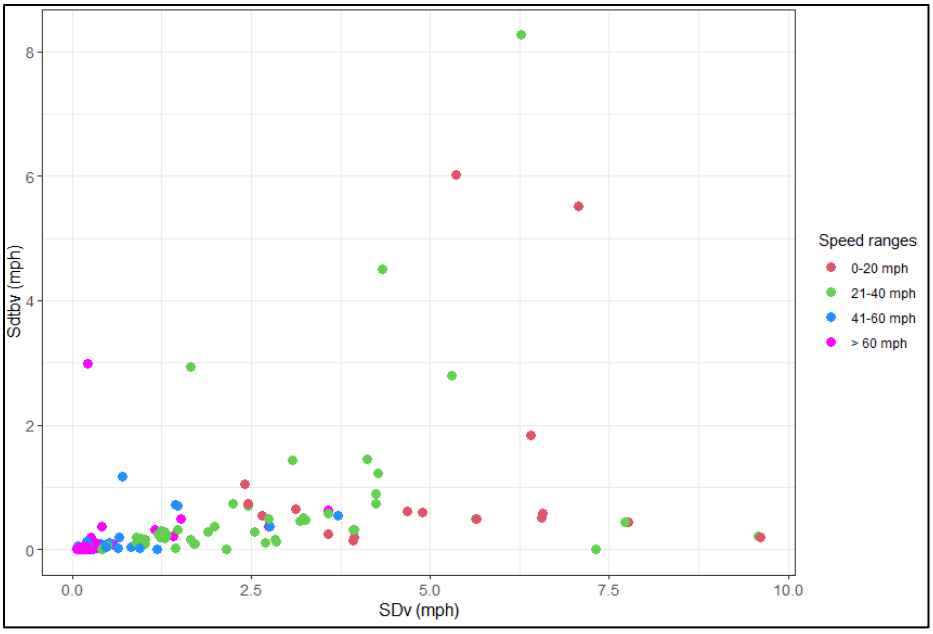


FIGURE 4.16: RELATIONSHIP BETWEEN SDV AND SDBV

Figure 4-17 shows the relationship between the standard deviation between vehicles (SDv) and the acceleration/deceleration of vehicles. Whereas Figure 4-18, shows the relationship between the SDv and Jerk. Figures 17 and 18 confirm that a higher magnitude in the values of acceleration/deceleration and jerk is observed, as the standard deviation for individual vehicles (SDv) increases which is a characteristic of the lower speed regimes (0-20 mph, and 21-40 mph).

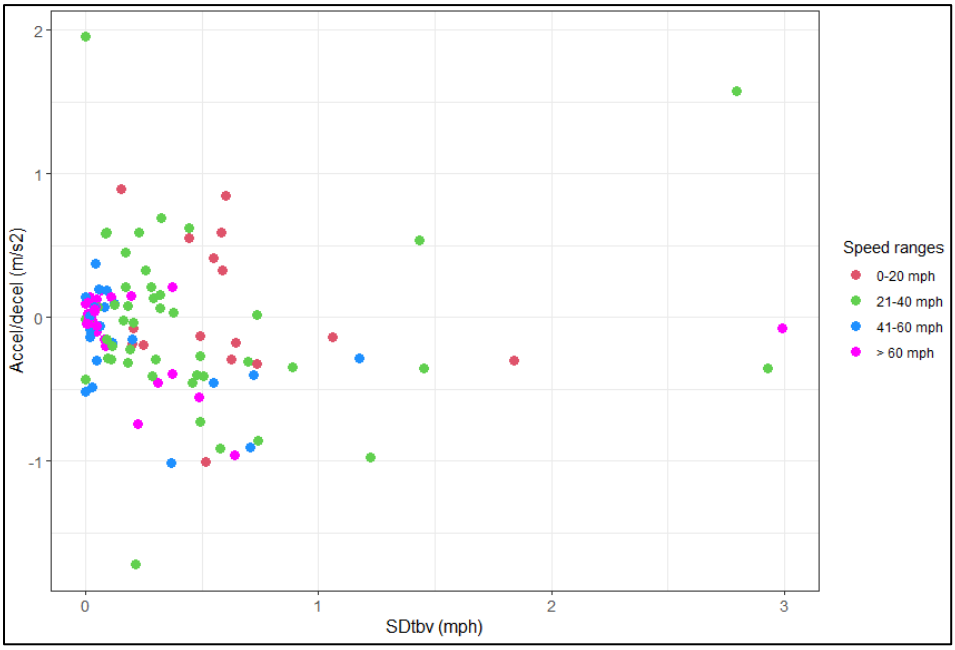


FIGURE 4.17: RELATIONSHIP BETWEEN SDV AND ACCELERATION DECELERATION

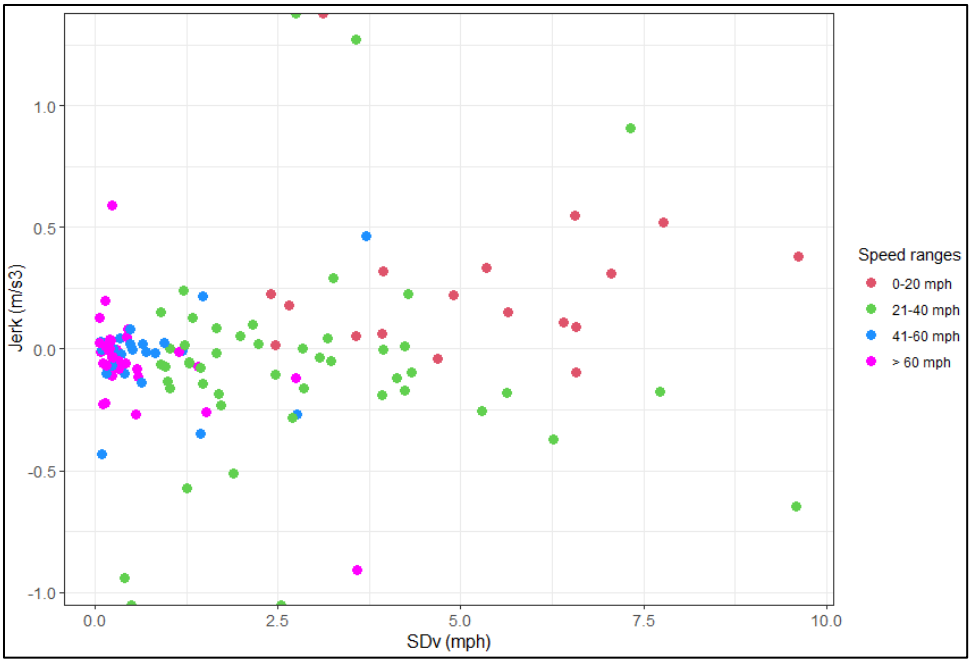


FIGURE 4.18: RELATIONSHIP BETWEEN SDV AND JERK

4.4. CONGESTION IDENTIFICATION

The preliminary analysis presented in the previous section including the exploration of disturbance metrics derived from the CV data revealed that these metrics, such as the standard deviation between data points, the standard deviation between individual vehicles, acceleration/deceleration, and jerk could be useful to detect the onset of congestion. This section builds on these findings by investigating the use of a clustering algorithm to group the data points into different clusters and examining the results to identify signals for the onset of congestion. Then, a Recursive Partitioning and Regression Tree was implemented to extract logic-based rules that define each one of the produced clusters. Given the high dimensionality of the data, Principal Components (PCs) was implemented before the implementation of the clustering algorithm.

4.4.1. Implementation of the Clustering Algorithm

This investigation utilized the k-mean clustering algorithm. The k-means algorithm cluster data by separating the samples into groups of equal variances by minimizing the sum of squares within each cluster (Tan et al.). Having a training set $x^{(1)}, \ldots, x^{(m)}$ from which the data is needed to be clustered into disjoint clusters. As usual feature vectors $x^{(i)} \in \mathbb{R}^n$ are given, however, no labels $y^{(i)}$ are provided which makes it an unsupervised learning problem. The objective of the algorithm is to predict *k* centroids and a label $c^{(i)}$ for each one of the data points. The k-Means clustering algorithm can be then described as follows:

1. The first step is to randomly initialize the cluster centroids $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$
2. Then repeat until convergence: {
   For every *i,* set

$$c^{(i)} := \underset{j}{arg\,min} \left\| x_i - \mu_j \right\|^2 \quad (5)$$

For each *j,* set

$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}} \quad (6)$$

}

The implemented model considered a total of 11 features including speed, acceleration, jerk, standard deviation between datapoints, standard deviation between vehicles, standard deviation of individual vehicles, variance of speed, and log of speed.

The principal components were utilized to reduce the dimensionality of the data. Principal components are a data analytic technique that finds new attributes that are orthogonal to each other and are linear combinations of the original attributes. The new attributes capture the maximum amount of variation in the data. Figure 4-19 shows the amount of variance that is explained by different number of components. From Figure 4-19, we can observe that five components already explain up to 90% percent of the variance in the model, therefore five components were selected for the input of the clustering algorithm.

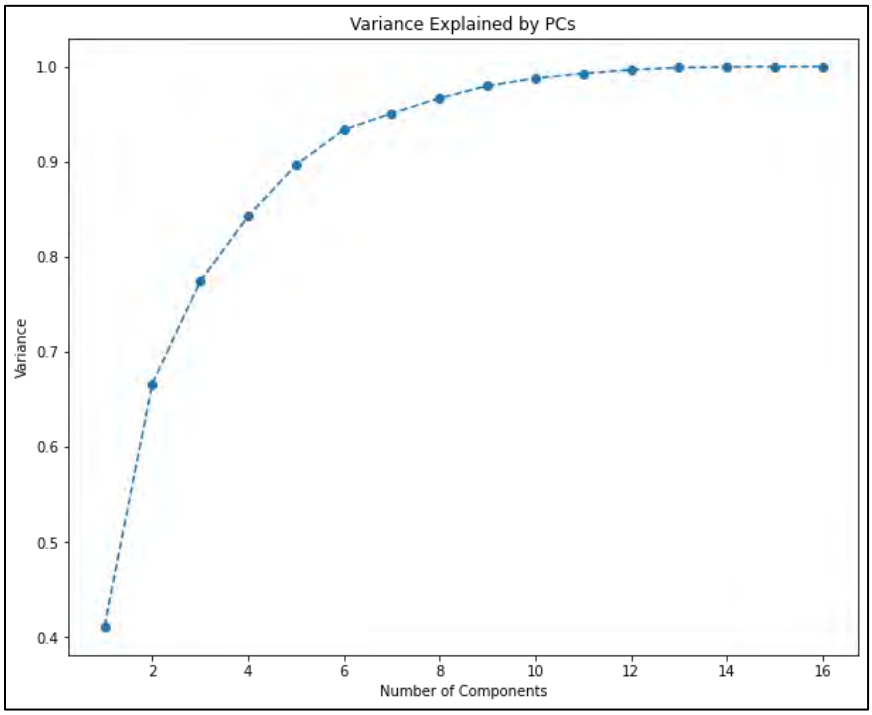


FIGURE 4.19: VARIANCE EXPLAINED BY NUMBER OF COMPONENTS

The K-means algorithm requires the specification of the number of clusters as an input to the algorithm. This study obtained the number of cluster (k) based on the elbow technique. The elbow technique is a plot that depict the total within clusters sum of squares (WCSS) for each value of k. The k value is selected at the point in the graph where the decrease in WCSS stop being significant as the value of k increases (See Figure 4-20). In this example, Figure 4-20 shows that the WCSS decreases at a significant rate up to k=6, and after that, the decrease in WCSS for each additional cluster is negligible, hence, it was determined that six clusters were the optimal value for k in this example.

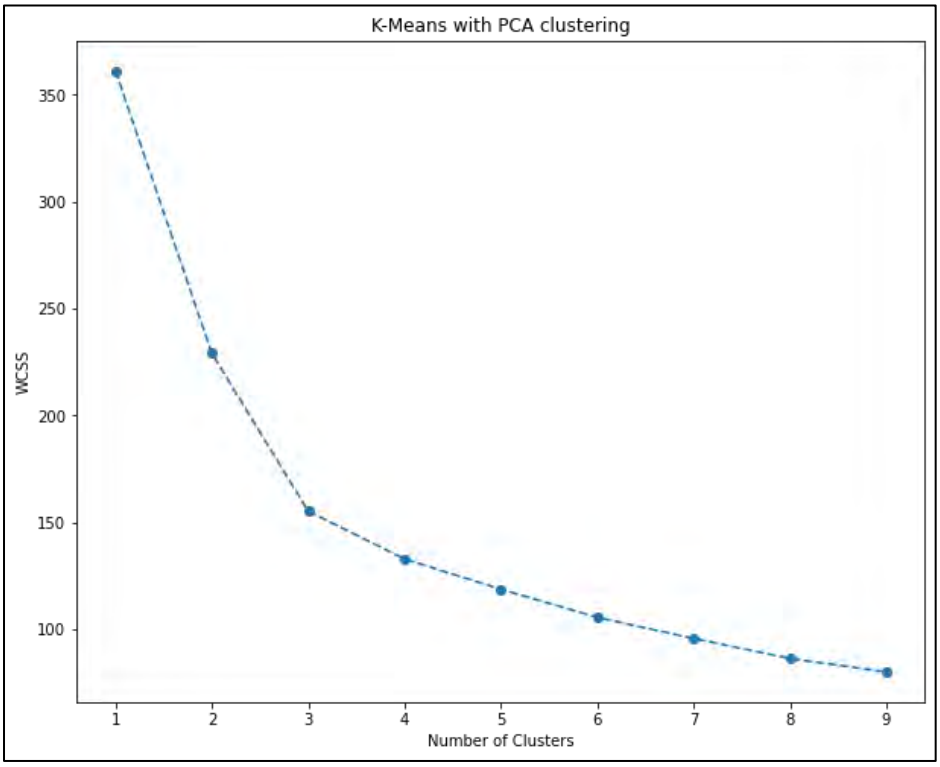


FIGURE 4.20: TOTAL WITHIN CLUSTERS SUM OF SQUARES WCSS FOR DIFFERENT VALUES OF K

The output of the k-means clustering is shown in Figures 21 and 22. Figure 4-21 shows the relationship between speed and SDdp for each cluster whereas Figure 4-22 shows the relationship of SDdp and acceleration/deceleration for each cluster. The clustering algorithm classified the data points into six clusters, where clusters 1, 2, and 3 corresponds to the data points with higher speed regime (speed>=45 mph), but clusters 1 and 3 show lower values for SDdp while cluster 2 depicts much higher values of SDdp. Cluster four contains data points that belong to both high and low speed regimes that are also characterized by having the highest deceleration values (see Figure 4-22). Figure 4-21 also shows clusters 5 and 6 belonging to the lower speed regime (speed < 45 mph). The main differentiation between cluster 5 and 6 is that cluster five represents data points with lower values of SDdp whereas cluster 6 contains data points with much higher values of SDdp indicating the occurrence of breakdown (stop-and-go operation).

Figure 4-22 shows the clusters categorized into different ranges of SDdp and acceleration/deceleration values. For example, Cluster 1 represents all the data points close to free flow speed with the lowest standard deviation and acceleration/deceleration. Data points in Cluster 1 represents the non-congested regime with no indication of breakdown. Cluster 2 represents the vehicles when they are still traveling near the posted speed limit, but with much higher values of SDdp ($6<SDdp_2<12$) indicating some degree of disturbance. However, according to Figure 4-22, cluster two does not present extremely high values for acceleration or deceleration. Cluster 3 on the other hand, shows data points that tend to have deceleration ($-0.75<decel_3<-0.17$), possibly indicating that the queue from a bottleneck happening downstream has reached the study location so the vehicles have to start decelerating at a considerable rate. As mentioned before, cluster 4 is characterized by data points distributed along both the high speed and the low-speed regime but having much higher deceleration values ($-1.6< decel_4<-0.43$) as well as considerably high SDdp values ($6<SDdp_4<20$) indicating the onset of congestion. Clusters 5 and 6 belong to the low-speed regime where cluster 5 shows lower values of SDdp ($0<SDdp_5<6$) and does not present intense acceleration or deceleration events. On the other hand, cluster 6 represents breakdown with low speed combined with high standard deviation values ($6<SDdp_6<25$) and a significantly wide range of acceleration/deceleration values (($-0.8< accel/decel_6<1.0$).

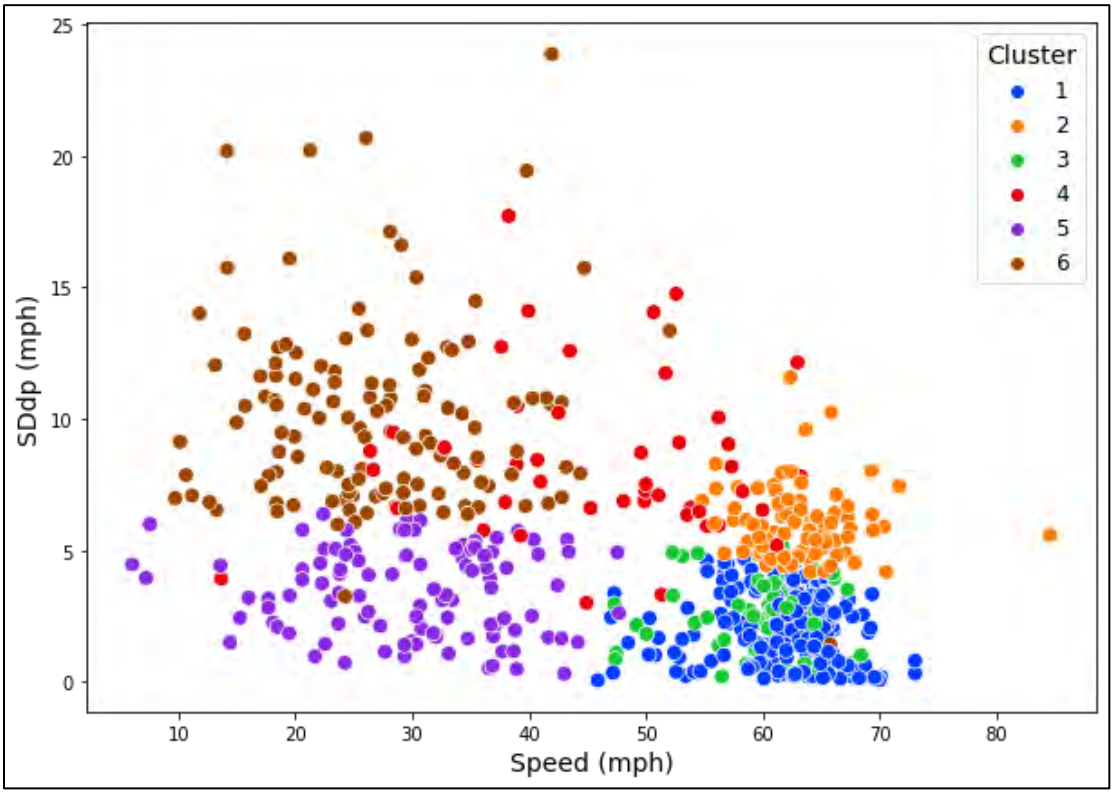


FIGURE 4.21: K-MEANS OUTPUT: OUTPUT: SCATTERPLOT OF SPEED AND SDDP FOR THE IDENTIFIED CLUSTERS

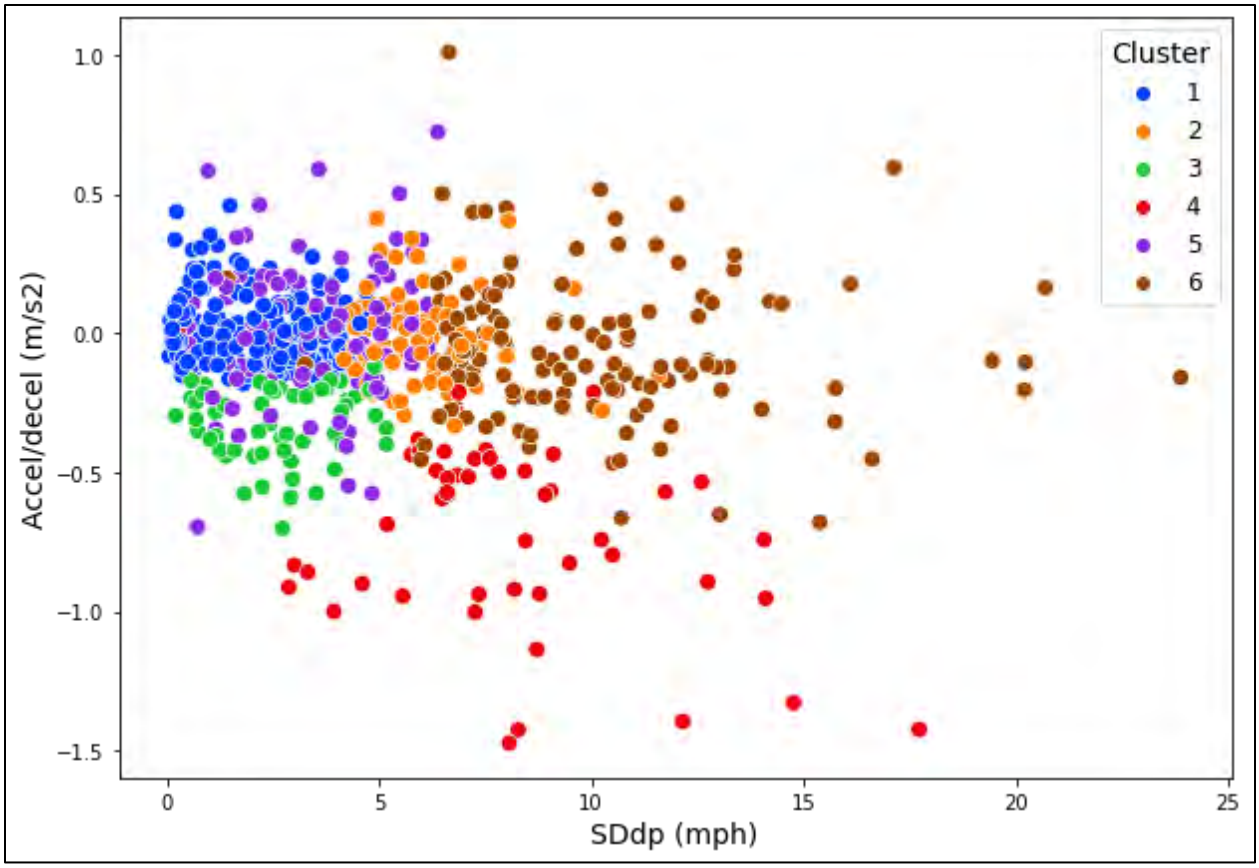


FIGURE 4.22: K-MEANS OUTPUT: SCATTERPLOT OF SDDP ACCELERATION/DECELERATION FOR THE IDENTIFIED CLUSTERS

The above discussion indicates that there may be a relationship between the occurrence of Cluster 2, Cluster 3, and Cluster 4 conditions in the uncongested regime and the onset of congestion in the near future. The researchers in this study examined a large number of time series of traffic conditions and were able to confirm that such a relationship may exist. For example, Figure 4-23a shows a time series of speed and Figure 4-23b shows a time series of the SDdp, as well as the occurrence of different clusters by time of day. Figure 4-23 shows the occurrence of Clusters 3 and 4 about 30 minutes before the actual breakdown, and two appearances of cluster 4 (the transition cluster) occurring just 10 minutes before the breakdown. The results in Figure 4-23b show how the standard deviations increase gradually as the time approaches the breakdown point.

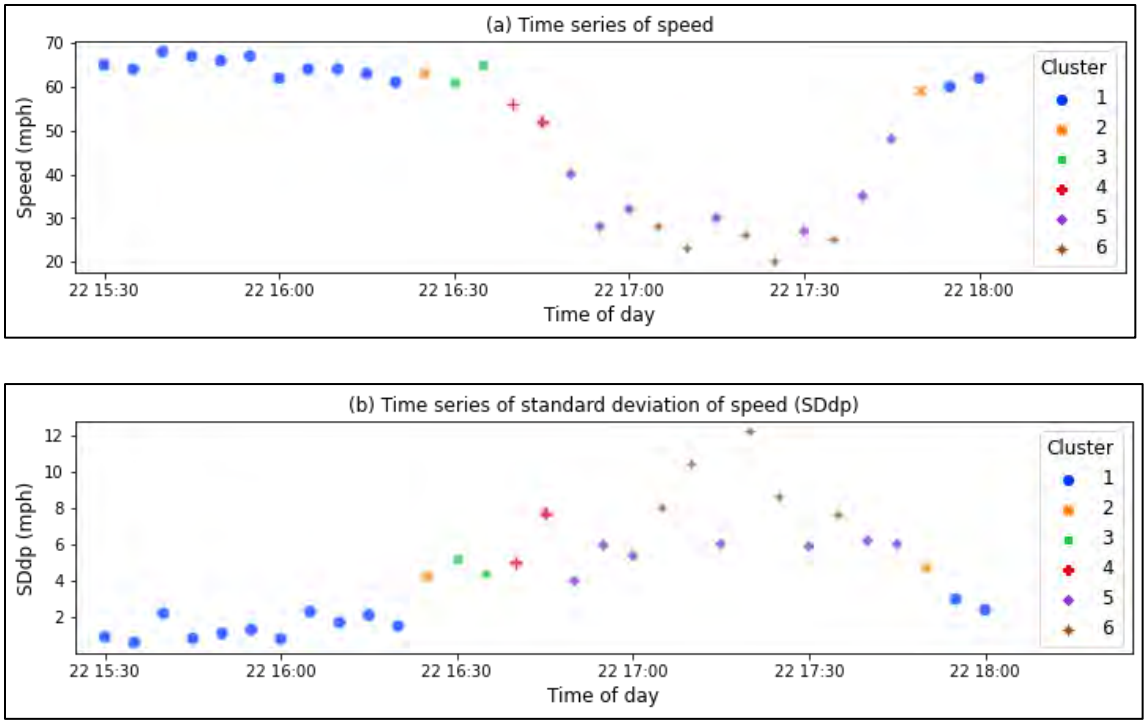


FIGURE 4.23: TIME SERIES OF SPEED AND STANDARD DEVIATIONS OF SPEED AND THE CORRESPONDING CLUSTERS

4.4.2. Variable Importance and Rules Extraction Using a Decision Tree

Decision trees are a non-parametric approach for creating classification models that does not require any pre-existing assumption regarding the probability distribution of the label and data features; hence, decision trees are applicable to a vast range of data sets and classification problems. It can be developed in an efficient manner and produces results that are easy to present and understand. (Han and Kamber, 2006). The tree-based methods can be used for both regression and classification, and they imply the stratification (or segmentation) of the predictor space into a number of simple regions. The implemented decision tree used for this study uses Classification and Regression Trees (CART) algorithm. CART constructs binary trees using both the threshold and the feature that generates the largest information gain at each node. The decision tree classifier algorithm can be summarized as follows:

1. Given the vectors $x_i \in \mathbb{R}^n, i = 1, \dots, i = n$, and a label vector $y \in \mathbb{R}^n$, a decision tree iteratively partitions the space so that the samples with the same labels are grouped together.
2. For the data *Q* at node *m,* each candidate split $\theta = (j, t_m)$, that is the feature *j* and the threshold t_m partition the data into $Q_{left}(\theta)$ and $Q_{right}(\theta)$ as follows:

$$Q_{left}(\theta) = (x, y)\big|x_j \leq t_m \quad (7)$$

$$Q_{right}(\theta) = Q \backslash Q_{left}(\theta) \quad (8)$$

3. The impurity at *m* is calculated by using the impurity function *H(),* which is determined by the type of problem to be solved, be it regression or classification (in this case is classification).

$$G(Q,\theta) = \frac{n_{left}}{N_m} H\left(Q_{left}(\theta)\right) + \frac{n_{right}}{N_m} H\left(Q_{right}(\theta)\right) \quad (9)$$

3.1. Since the objective is classification with an outcome taking values 0, 1, 2, …, *k – 1,* for node *m,* representing a region R_m with N_m observations, then:

$$p_{km} = 1/N_m \sum_{x_i \epsilon R_m} I(y_i = k) \quad (10)$$

Represents the portions of the class *k* observations in the node *m*

3.2. In this case, the Gini function throw the best results so, it is used as a measure for impurity:

$$H(X_m) = \sum_k p_{mk}(1 - p_{mk}) \quad (11)$$

4. The following function represents the selection of the parameters that minimize the impurity:

$$\theta^* = argmin_\theta G(Q,\theta) \quad (12)$$

5. Repeat for subsets $Q_{left}(\theta^*)$ and $Q_{right}(\theta^*)$ until it gets to the maximum allowable depth, that is $N_m < min_{samples}$ *or* $N_m = 1$

A classification decision tree was used to determine the variable importance in categorizing traffic conditions in the cluster analysis. This classification allows identifying rules for data partitioning and classification based on the identified features by the decision tree algorithm. This technique provides a better understanding of the selected clusters in terms of which variables are more relevant to the categorization of each cluster. The resulting decision tree structure can be converted into crisp (if-then) rules as shown in table 4-2 below.

TABLE 4-2: CRISP (IF-THEN) RULES EXTRACTED FROM THE DECISION TREE

cluster	Rules															
1	when	speed	>=	45	&	accel/decel	>=	-0.17	&	SDbv	<	4.1				
2	when	speed	>=	45	&	accel/decel	>=	-0.41	&	SDbv	>=	4.1				
3	when	speed	>=	45	&	accel/decel	<	-0.17	&	SDbv	<	4.1				
4	when	speed	<	45	&	accel/decel	<	-0.43					&	SDdp	>=	6
4	when	speed	>=	45	&	accel/decel	<	-0.41	&	SDbv	>=	4.1				
5	when	speed	<	45									&	SDdp	<	6
6	when	speed	<	45	&	accel/decel	>=	-0.43					&	SDdp	>=	6

Table 4-2 shows the rules produced based on the decision tree that can be used to classify traffic conditions. According to the rules in Table 4-1, the most significant features in the model are the standard deviation between data points (SDdp), the standard deviation between

vehicles (SDbv), the average speed, and the acceleration/deceleration. The output also confirms that a threshold of 45 mph separates the high-speed clusters (Clusters 1, 2, and 3) from the low-speed clusters (Clusters 5 and 6). In the high-speed cluster group, a threshold of -0.43 in the deceleration defines Cluster 4.

4.4.3. Implementation of Decision Tree for Prediction

This study was then derived a decision tree, to predict the occurrence of breakdown within the next 30 minutes based on the frequency of data points belonging to each cluster during the current 30 minutes. For this purpose, the data was first aggregated into bins with a size of 30 minutes with each bin containing the number of five minutes that belong to each cluster during that time period. Additionally, the breakdown incidence for the next 30 minutes was labeled as a binary variable and used as the dependent variable in the decision tree with the independent variables being the frequencies of five-minute data points in each cluster during the current 30-minute period.

The graphical output of the prediction model using the decision tree is shown in Figure 4-24. Starting on the root node, the CART algorithm first divides the dataset based on the frequency of five-minute data points belonging to Cluster 5 (the cluster with low speed and low SDdp). In the case that there are at least two instances of Cluster 5 during the current 30 minutes, then the model determines that breakdown will occur during the next 30 minutes, contrarily, if cluster 5 has less than two instances during the current 30 minutes, then the tree will evaluate the frequency of Cluster 4 instances next (first child node to the left) which is the cluster with the highest standard deviation and higher deceleration values group among the high speed clusters. As the tree depicted in Figure 4-24 shows, the occurrence of Cluster 4 instances is critical to predict the onset of congestion. In case that there is at least one instance of Cluster 4, then the model determines that breakdown is likely to happen during the next 30 minutes. On the other hand, if no instances of Cluster 4 occur during the current 30 minutes, the decision tree proceed to evaluate Clusters 1, 2, and 3 at their respective thresholds.

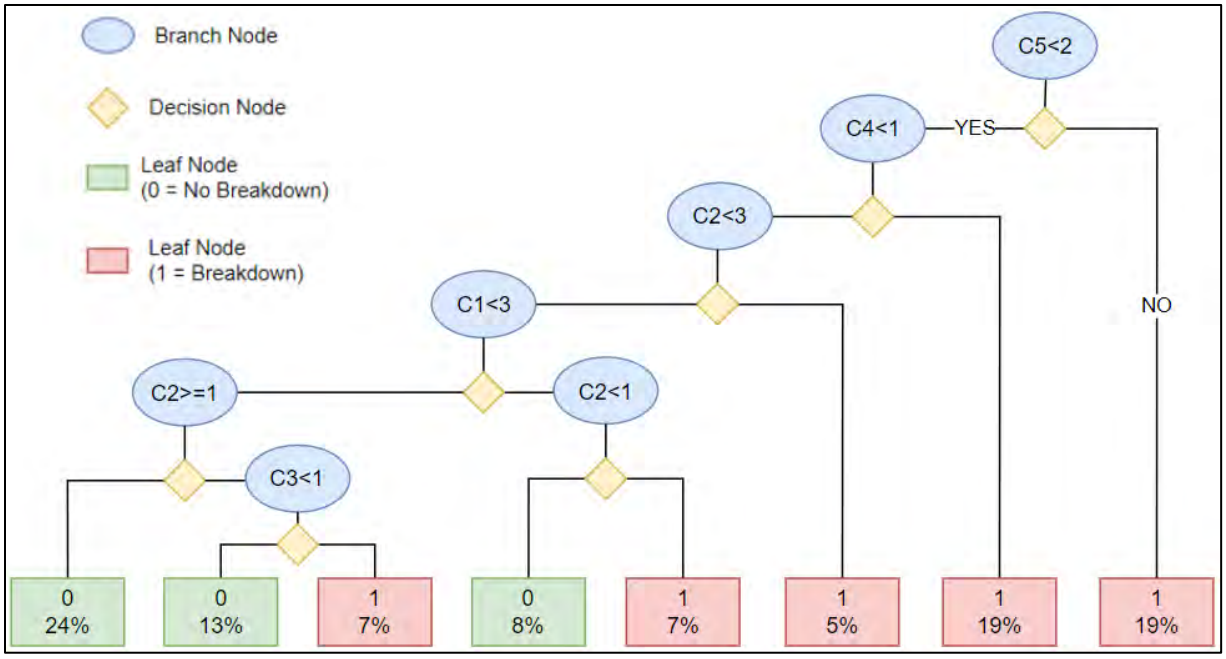


FIGURE 4.24: GRAPHICAL OUTPUT OF THE DECISION TREE IMPLEMENTED FOR PREDICTION OF BREAKDOWN

For the purposes of the evaluation of the performance of the prediction model, the dataset was partitioned in a proportion of 70% observations and 30% for testing. For evaluation purposes, the outputs of the model were evaluated utilizing the testing set that represents 30 percent of the data. In assessing the performance, the study used measures commonly used in evaluating machine learning models such as accuracy score, sensitivity, specificity, as well as precision, recall, and F1-score.

Table 4-3 presents a summary of the evaluation metrics that allows the assessment of the performance of the decision tree, the FRBS and the ANFIS. From Table 4-3, it can be concluded that decision tree performs well in prediction resulting in 81.82% accuracy. Ian terms of sensitivity, which is a ratio that express how many observations were predicted as positive for the breakdown versus how many observations were actually positive, Table 4-3 shows that the model achieved a good rate in predicting the breakdown occurrence (82.61%). Another important measure to evaluate the model is the Precision Rate, especially the precision of the positive class (Class 1) since this indicates the ability of the model to avoid predicting false positives. Based on the Precision Rate assessment results in Table 3, the precision value is 82.61%. A last measure used in the evaluation is the F1 Score, which is a weighted average of the precision and recall ratios. The F1 Score allows an integral evaluation of the models by taking into consideration both the false positives and false negatives into account The decision tree achieved a F1 Scores for both classes that are higher than 80%.

.

TABLE 4-3: CRISP (IF-THEN) RULES EXTRACTED FROM THE DECISION TREE

	Overall Model			*Breakdown (1)*			*no breakdown (0)*		
	Acc.	Sens.	Spec.	Prec.	Recall	F1	Prec.	Recall	F1
Decision Tree	0.8182	0.8261	0.8095	0.8261	0.826	0.826	0.8095	0.8095	0.8095

Note: Acc. = Accuracy; Sens. = Sensitivity; Spec. = Specificity; Class 1 = breakdown; Class 0 = no breakdown.

4.5. SUMMARY

The case study presented in this chapter using data from the connected vehicle deployment in Tampa explored the use of performance measures derived based on real world connected vehicle data for the assessment and prediction of congestion on a freeway segment. The study examined the use of multiple measures including speed, acceleration/deceleration, jerk, standard deviation between datapoints, standard deviation between vehicles, standard deviation of individual vehicles. The analysis revealed that the traffic states could be classified into groups with six different traffic conditions based on speed, standard deviation of speed between vehicles, standard deviation between points, as well as the deceleration values. The study also developed a methodology for the prediction the breakdown based on connected vehicle data. A prediction model utilizing a decision tree achieved good results with an accuracy rate of 82%. The Precision Rate for the positive class indicated that the model had a relatively low chance of predicting false positives.

5. ATLANTA CASE STUDY

System detector data, from loops, video cameras, or other sources, is historically the oldest source of data about system operation, and in many places nationwide, it is still the principle one. This study aimed to identify early indicators of the onset of congestion and to develop a prediction model for congestion events. The Atlanta Metro area was used as the study area. The high levels of variability in congestion levels and the availability of high-quality high-fidelity vehicle volume-speed data over hundreds of miles of roadway made this an ideal study area. However, the analysis ensured that the methodology development was sufficiently generic to ensure transferability of the methods, and potentially the models, to other regions that suffer from similar congestion issues.

Problems explored through the lens of ML techniques can be broadly classified into supervised and unsupervised learning problems. Further, supervised learning problems can be categorized into regression and classification problems. Since the objective of this study is concerned with traffic state prediction, the problem is analyzed as a supervised learning problem. Under this framework, the traffic state, characterized by available speed and flow data from the detector system, is labelled as belonging to one of the two classes, pre-congestion and the rest. This type of formulation is common under classification problems handled under the supervised learning framework. The choice of ML algorithms in the first set of experiments in this study was driven by the positive results reported in the study by (Filipovska and Mahmassani 2020).

For the analysis, the study uses point detector data (speed and flow) from detector stations equipped with Video Detection System (VDS) owned and operated by the Georgia Department of Transportation's Transportation Management Center. Section 5.1 describes the site and the collected data and the quality related issues. Section 5.2 presents a discussion on the problem formulation and other related challenges. Section 5.3 discusses the implementation of the discussed ideas. Section 5.4 presents a discussion on results and lessons learnt.

5.1. ATLANTA DATASET

5.1.1. Site and Data Description

The advanced traffic management system in Atlanta has approximately 1,645 VDS stations installed along most major interstates around the metro Atlanta area (Wells 2016). The system mostly uses Video Detection System based detectors and Microwave based detectors to provide vehicle count aggregates and average speeds in 20-second intervals (Cho 2017). For this study, a one-mile section of Southbound I-285 near SR-6 Camp Creek Pkway, Atlanta, Georgia, was used as the study site, as shown in the Figure 5-1 below. The selected section provided a section of freeway with three adjacent VDS stations without any corridor access or egress points in between. This setup helped reduce confounding variables emerging from flow ingress/egress in form of ramp traffic. For this site, speed and count data aggregated at a 20-second frequency is available through stations equipped with Video Detection System (VDS).

Each station has a camera mounted on a pole and is located at approximately every 1/third mile. The available dataset covers from October 2007 to July 15th, 2021.

FIGURE 5.1: ONE-MILE SECTION OF THE I-285-SOUTHBOUND FREEWAY CORRIDOR (NORTH OF MILE MARKER 1), ATLANTA, GEORGIA. SOURCE: GOOGLE EARTH

5.1.2. Data Pre-processing

The detection system generates vehicle counts and average speeds aggregated over 20-second intervals for each lane. In this study, this data is then further aggregated into 1-minute bins over all lanes to improve the signal to noise ratio (Guin 2004). Further investigation into the extracted data revealed two types of missing data patterns in the received datasets for the detectors under study. The first set of missing data points occurred at a regular time around 1:10 AM to 2:45 AM, as can be observed in Figure 5-2. These outages are a result of a scheduled maintenance window of the detector system database. Unlike this type of missing data pattern, the next set of missing data occurred at irregular intervals, an example for this is shown in Figure 5-2 from 10:44 AM to 13:10 PM. For this study both types of missing points were replaced by data imputed using linear interpolation. For interpolation of speeds, the means of speeds (over all lanes) over the five-minute periods immediately before and after the missing interval were used as the start and the end points of the linear interpolation. Imputed speed data for the two types of examples can be observed through Figure 5-3. In the given

figure, the vertical axis shows average speed (in mph) over all lanes for the selected site for the middle detector GDOT-STN-2851097, and the horizontal axis represents time-of-the-day.

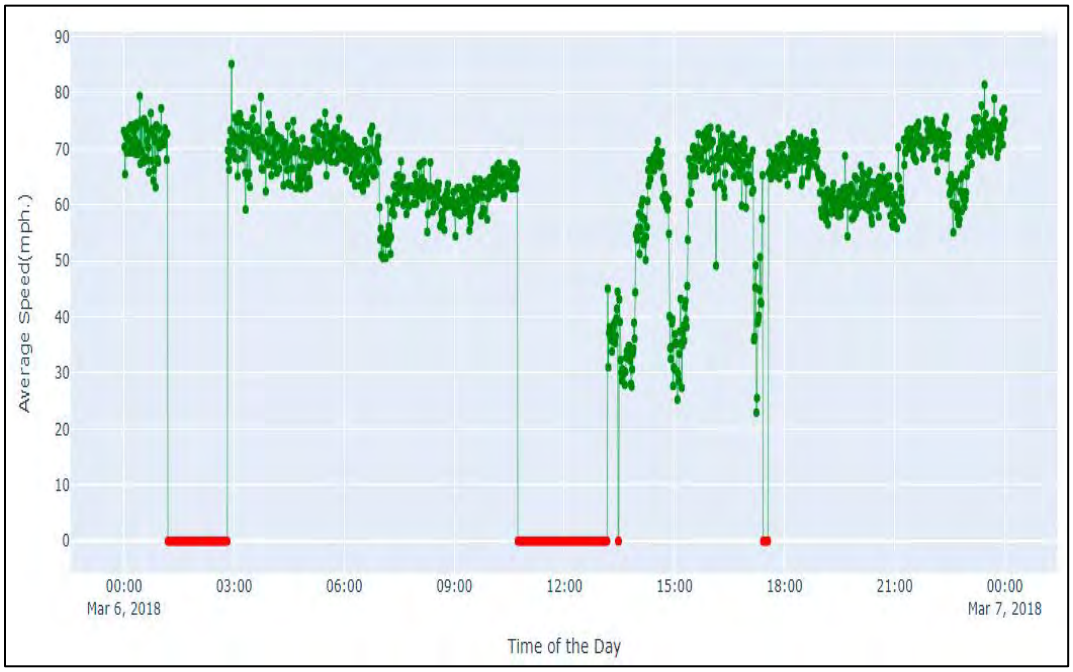


FIGURE 5.2: MISSING DATA OCCURRING AT REGULAR TIME INTERVALS FROM 01:11 AM TO 02:44 AM FOR SCHEDULED DETECTOR SYSTEM MAINTENANCE AND AT IRREGULAR TIME INTERVALS FROM 10:44 AM TO 13:10 PM ON 03/06/2018

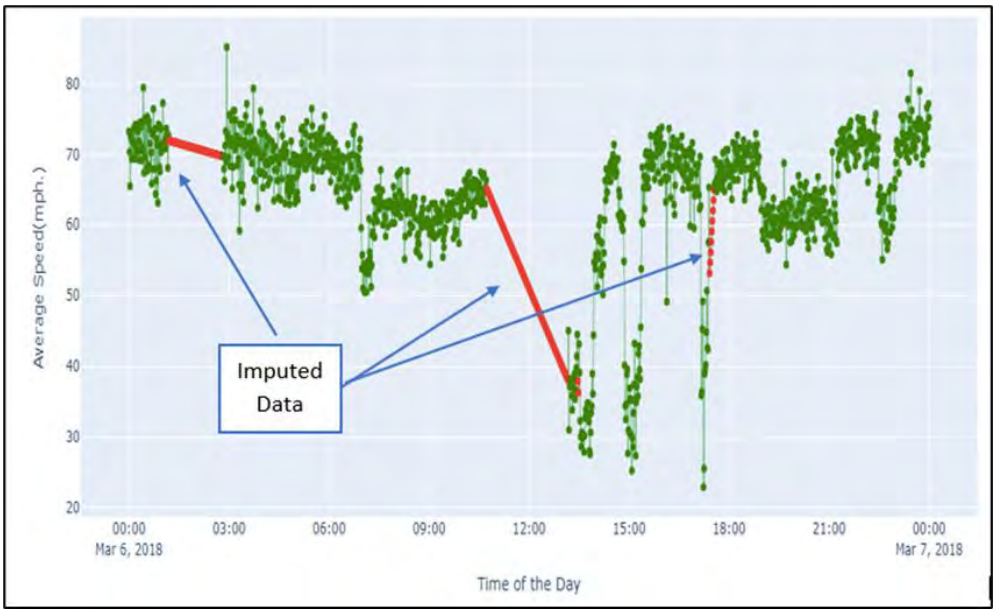


FIGURE 5.3: IMPUTED DATA FROM 01:11 AM TO 02:44 AM AND FROM 10:44 AM TO 13:10 PM ON 03/06/2018

5.1.3. Data Quality

Previous research suggests that the data quality of VDS can depend on a number of different factors, such as camera's vertical and horizontal angle (Grant, Gillis, and Guensler 2000), prevailing weather conditions, time of the day, site, and camera angle, height, and offset, etc., and can also degrade over time (Suh et al. 2015 (Suh et al. 2015), (James Bonneson 2002), (Dr. Peter T. Martin 2004), and (Avery Rhodes 2006).

The data inconsistency, as noted by (Guensler et al. 2013) and (Suh et al. 2015) is verified and demonstrated using traffic count data for the site under study. The site represents three adjacent VDS stations within this roadway section, without any intermediate corridor access or egress points. The stations are shown in Figure 5-1, namely: GDOT-STN-2851096, GDOT-STN-2851097, and GDOT-STN-2851098, listed in order of direction of travel. For this site, a general pattern has been observed where GDOT-STN-2851097 and GDOT-STN-2851096, respectively, give an over and under count estimate compared to GDOT-STN-2851098. The pattern can also be observed for a typical day (dated 04/09/2018) in Figure 5-4. This data inconsistency is seen to be further exacerbated during incident conditions. This issue is illustrated in the cumulative count curves for the three stations on one such incident-day (dated 04/19/2018) in Figure 5-5. Multiple incidents were detected during the evening peak hour for this day. The incidents were detected at locations immediately downstream of station GDOT-STN-2851098. From , it can be observed that the order of count differences between the three stations over the duration of 24 hours may be conservatively estimated as approximately 2000 vehicles. Furthermore, the cumulative curves are not overlapping according to expectations, GDOT-STN-2851096's curve is lying below both GDOT-STN-2851097 and GDOT-STN-2851098. The underlying reason for this inconsistency is likely a limitation of the field deployment of the VDS detectors whereby the camera angles lead to occlusion of vehicles in some cases, leading to undercounts, or splash-over on adjacent lanes, leading to overcounts. The splash over issue disproportionately affects larger vehicles that have a higher likelihood of activating more than one virtual detector loop because of their larger footprint in the video. Moreover, for the detectors on the lanes that are farther away from the camera, the flatter angles cause vehicles to activate detectors from multiple lanes, depending on the size of the vehicles and the position of the vehicle within the lane. For example, a vehicle traveling close to the edge of the lane nearer to the camera might not have a splash over, but even a sports utility vehicle on the farther edge could trigger the detector on the adjacent lane on the far side. This issue primarily affects the vehicle count data, since the effects of the aggregate count errors are cumulative. While this also affects the speeds, the effects are much more muted since the speeds are reported as an average. For the purpose of identification of the traffic state, the errors in the speed are not expected to have a significant impact.

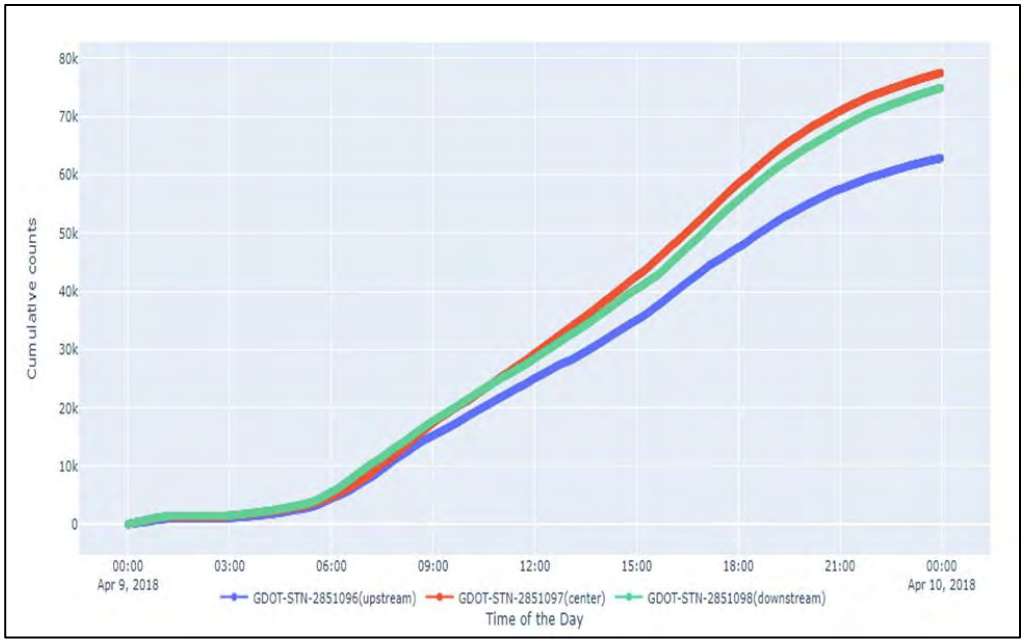


FIGURE 5.4: CUMULATIVE COUNT CURVE FOR A TYPICAL DAY (DATED 04/09/2018) AT THE SITE

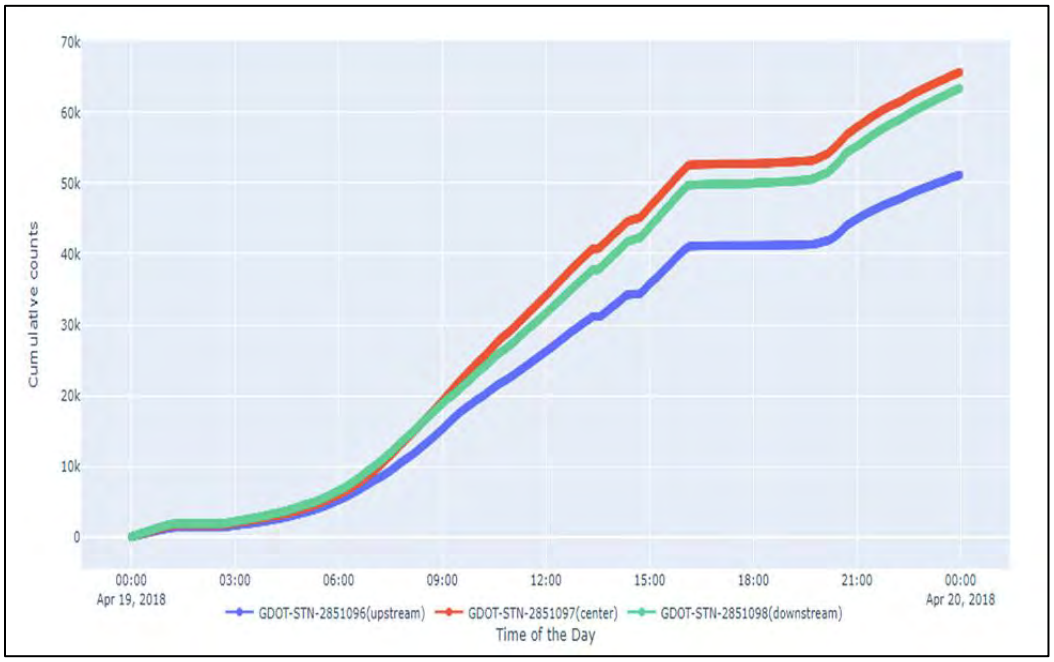


FIGURE 5.5: CUMULATIVE COUNT CURVE FOR AN INCIDENT DAY (DATED 04/19/2018) AT THE SITE

5.2. Problem Formulation

The overarching goal of this study is to identify early indicators of congestion conditions. Intuitively, early indicators of congestion can be expected to present in the past near-term of a congestion instance. So, the exploration was staged into two steps, first, identification of pre-congestion instances using the aggregated point detector data (speed and flow). Then, next

step involved training of the selected ML algorithms using the relevant input features to identify consistent pre-congestion patterns.

A brief introduction of the algorithms and their working principle is in Section 2.7. The rest of this subsection provides a detailed description of the procedure for labeling of the training dataset, a discussion on the resulting class imbalanced dataset, the training dataset, and the input features. The section concludes with a review and discussion of the selected performance metrics.

5.2.1. Labeling Dataset for Supervised learning

Typically, short-term traffic flow prediction falls within the scope of regression problems under supervised learning. However, since the objective of this study concerns a binary traffic state prediction, it is formulated as a classification problem. The first step in supervised learning is the preparation of the labels of the dataset.

Speed Threshold

To label the dataset, the overall traffic state, characterized by average speed and flow rate variables, was assigned two categories, congested and uncongested. The categorization was based on the current speed, averaged over across all lanes and aggregated at 1-minute interval, and compared to a pre-defined threshold speed. The study by (Filipovska and Mahmassani 2020) defined the threshold speed value as 20% lesser than the prevailing free-flow speed for their class labeling step. The same threshold was used in the current study.

Sustained State Change

Additionally, to avoid random data perturbations getting incorrectly coded as a change in traffic state, an additional constraint was introduced whereby the change in speed was required to be sustained for a fixed time duration for the logic to identify it as a change in state. This constraint helped prevent the triggering of multiple pre-congestion alarms during a single congestion period. Multiple trials were performed with durations ranging from 5 to 30 minutes to determine a suitable state-persistence check duration. Figure 5-6 and Figure 5-7Error! R eference source not found. demonstrate the types of pre-congestion alarms generated as a result of using a 5-minute and a 10-minute window, respectively. In the figures, the vertical axis represents average speed (in mph) over all lanes for the middle detector GDOT-STN-2851097 and the horizontal axis represents time-of-the-day; the detected pre-congestion points in both cases are highlighted with red circles. A state-persistence duration of 10-minutes was found to be sufficient in ensuring robustness of the labeling process.

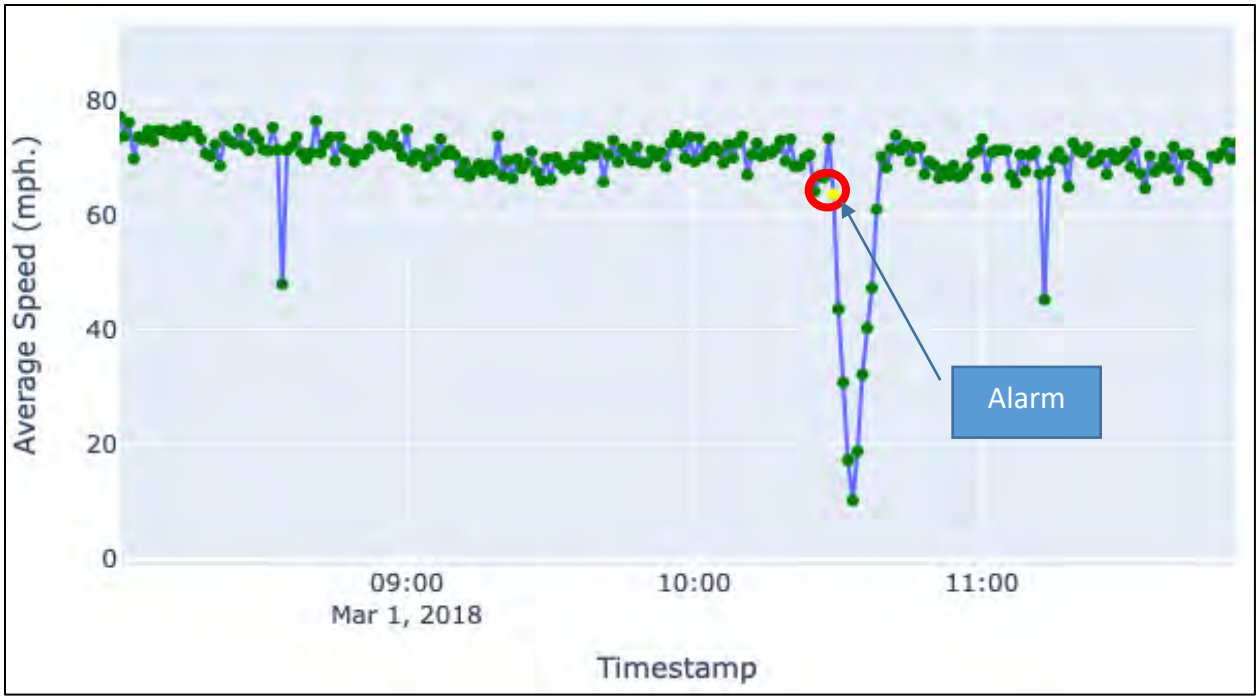


FIGURE 5.6 PRE-CONGESTION ALARM USING 5-MINUTE STATE PERSISTENCE ON 03/01/2018

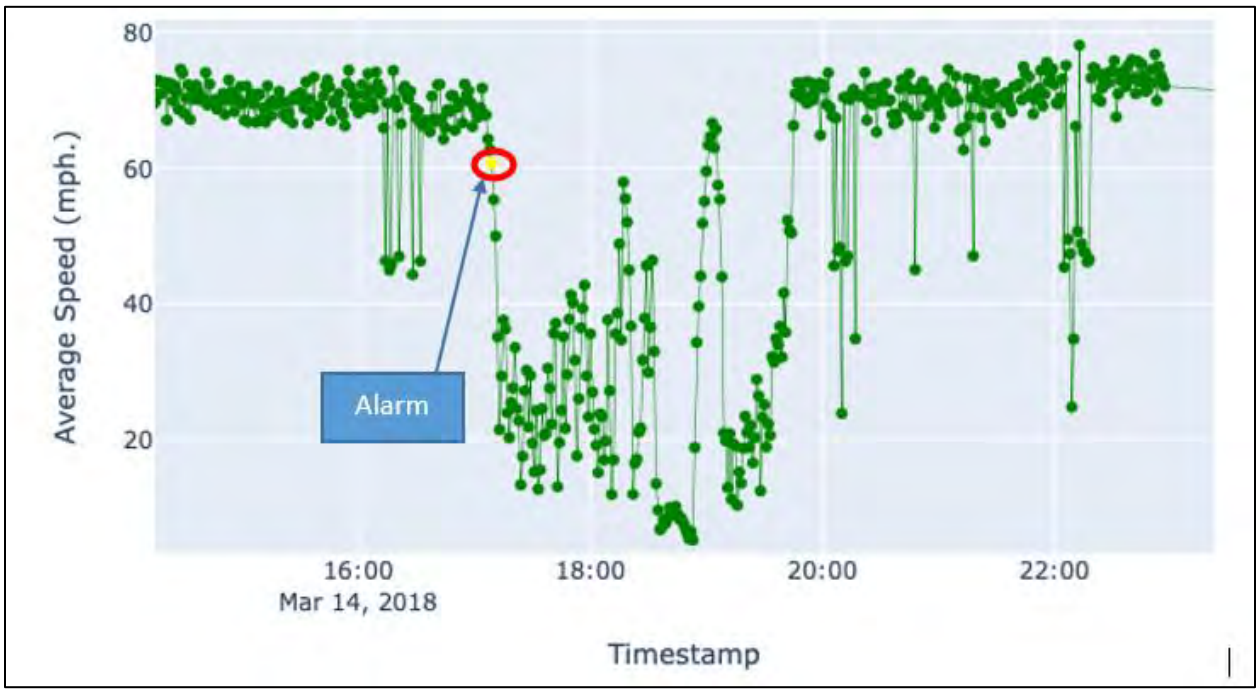


FIGURE 5.7 PRE-CONGESTION RAISED USING 10-MINUTE STATE PERSISTENCE ON 03/14/2018

Breakdown Predictor

The traffic state assignments were used to generate and assign the binary class labels that were then used for training the classification algorithms. The breakdown predictor variable was assigned a unit label for the time intervals where the current traffic state changed from

uncongested to a congested in next time interval. The null class label encompassed all other traffic states. These binary class labels field is treated as a nominal variable in the dataset. The full details of the logic for the assignment of traffic states are shown in the flowchart in Figure 5-8 and Figure 5-9.

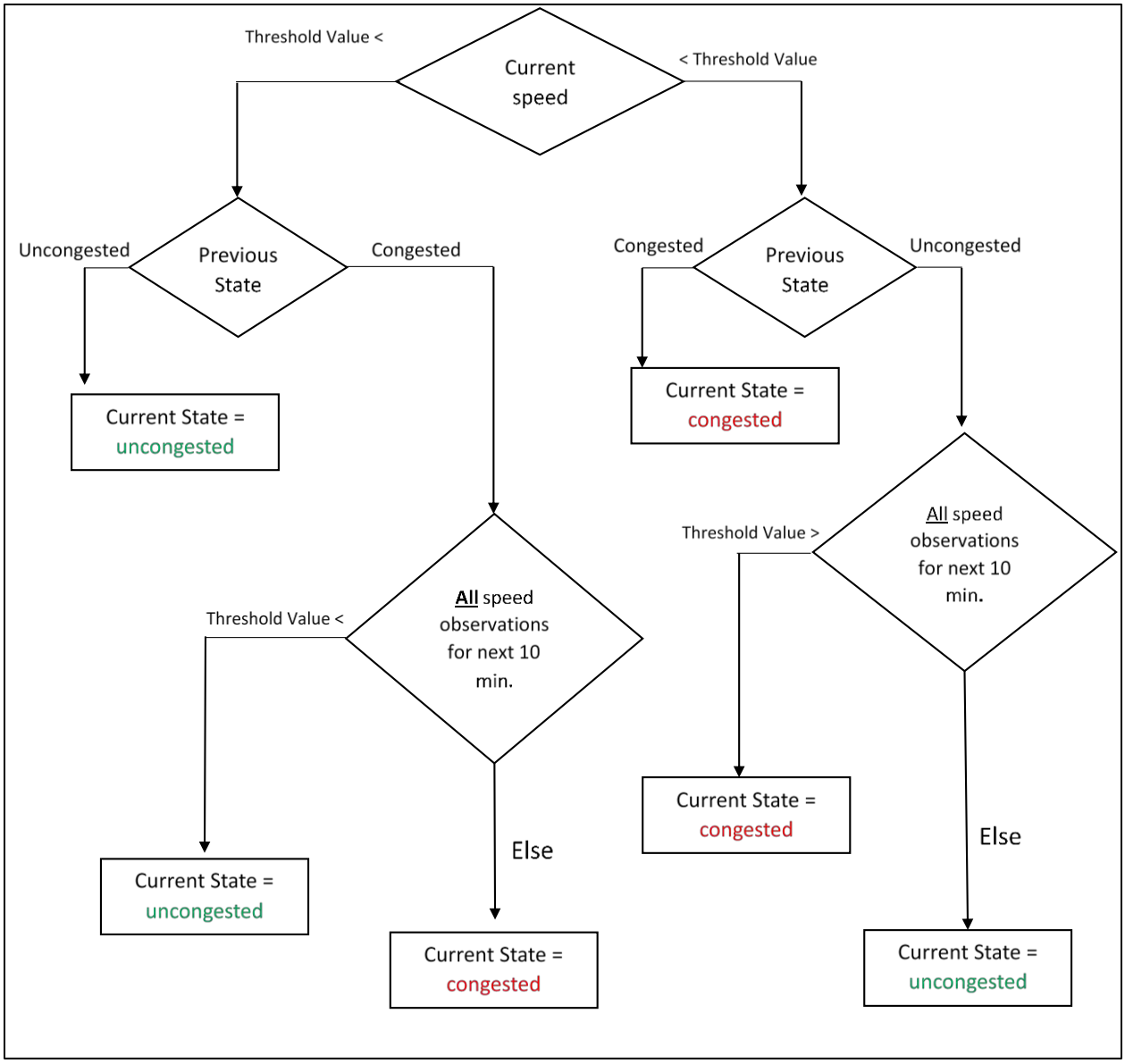


FIGURE 5.8: PROCESS FOR IDENTIFYING CONGESTED AND UNCONGESTED STATES FOR TRAINING AND TEST DATASETS

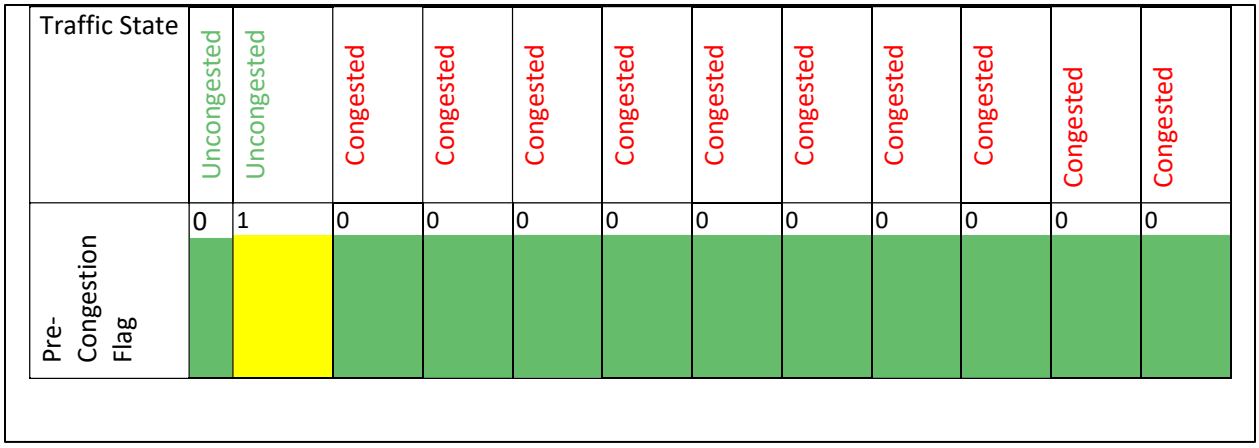


FIGURE 5.9: PROCESS FOR LABELLING PRE-CONGESTION ALARMS

5.2.2. Class-imbalanced Dataset

The binary coding of the pre-congestion variable naturally resulted in a highly skewed class distribution given that the state changes occur only 1-2 times in a day (1-2 data-points out of 280 5-minute-interval data-points in a day). For the current dataset, the class of interest represented only about 0.07% of the total data. According to (Krawczyk 2016), the class distribution where the minority class represents less than 1% of the data can be called as 'extreme class imbalance'. This type of class imbalance incentivizes the classification algorithm to train for the majority class, which would be counterproductive for the congestion-precursor detection objective.

(Krawczyk 2016) divides strategies to handle the 'extreme class imbalance' issue into data-based, algorithm-based, and hybrid approaches. Data-based approaches modify the training dataset to counter the effects of class imbalance. Algorithm-based approaches modify the cost function of the involved algorithms, while hybrid approaches represent a combination of both. Common data-based approaches include either under-sampling from the majority class and/or oversampling from the minority class, (Krawczyk 2016), (Johnson and Khoshgoftaar 2019). These approaches have shown success in offsetting class imbalance in the field of image classification (Johnson and Khoshgoftaar 2019). To implement these ideas here, data points belonging to typical non-peak-traffic periods were filtered out. The underlying assumption of this implementation is that the majority class (null class) forms the bulk of non-peak period data-points. In this study peak period refers to the period between 06:30 to 08:30 during the morning hours and, 16:00 to 18:00 during the evening hours. This under-sampling from the majority class changed the class distribution in favor of the unit class from 0.07% to approximately 0.14%. Moreover, this approach was also expected to help in reducing instances of disruptive incidents which are less predictable by nature and might not have consistent precursors to a traffic state transition and are likely to confound the classification algorithm.

5.2.3. Training dataset

Typically, ML implementations involve random division of the dataset into training and testing datasets using a pre-determined ratio. However, with the relatively sparse "true" cases in this dataset, it was necessary to split the dataset temporally, rather than randomly, to ensure that sufficient “true” cases were present in both training and testing sets. The training dataset included January to August, October, and December of 2018 and January, February, and August to part of November of 2019 (representing 295 weekdays and 115 weekend days). For testing, the last 15 days’ worth of data from November 2019, including 11 weekdays and 4 weekend days, representing about 4% of the total data was used.

5.2.4. Input Features

Important input features, identified in previous studies (Treiber and Kesting 2013), include current speed, change in speed over the past 1-2 intervals, and current flow rate from the study detector and the adjacent detectors. In addition to these variables, the experiments in this study have also used the past 15-minute traffic information (speed and flow rate) from all the three detectors involved. The final input feature vector contains a total of 99 points. These input feature vectors are presumed to contain the required spatiotemporal information about the precursors to the congested state that are necessary to predict an upcoming transition of the traffic state from uncongested to congested.

5.2.5. Performance Metric

The evaluation of the success of the methodology in addressing the problem requires identification of the appropriate performance criteria for the evaluation of the model. Generally, prediction accuracy is used to evaluate an ML classifier. However, given the inherent rarity of the class of interest, the number of actual positives is very low. Using prediction accuracy as the target metric to train the classifier can incentivize the classifier to train only for the majority class. Hence, the prediction accuracy criterion is likely to give a biased estimate of the performance and would be an unsuitable criterion. Thus, for the initial stages of this study, recall and precision scores are chosen as the target criteria. Recall and precision scores are defined in equations 5-1 to 5-2. Essentially, the precision score is a measure of the proportion of correctly classified positive labels among the total positive predictions. On the other hand, recall score is a measure of the proportion of correct positive prediction from the true positive labels. These performance criteria aim to reduce the number of false alarms received while trying to predict the majority of congestion events. Another common metric indirectly used in the context of the class-imbalanced dataset is selectivity which helps to determine the number of correctly labelled negative points among total negative labels.

$$Recall\ Score = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad \text{(Eq. 5-1)}$$

$$Precision\ Score\ = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad \text{(Eq. 5-2)}$$

$$Selectivity\ = \frac{True\ Negative}{True\ Negative + False\ Positive} \quad \text{(Eq. 5-3)}$$

Here,

True positive is the number of correct predictions for the unit class (pre-congestion label)

True negative is the number of correct predictions for the null class

False positive is the number of incorrect predictions for the unit class (pre-congestion label)

False Negative is the number of incorrect predictions for the null class

Other performance criteria popularly used for research for problems with class rarity are F score or F 1 score, Area under the Receiver operating characteristics (ROC) curve (AUC) (Provost and Fawcett 1997), and balanced accuracy (Johnson and Khoshgoftaar 2019). Among these, F 1 score represents the harmonic average of the precision and recall scores. It is mainly used as a single metric striking the balance between both precision and recall. Similar to F 1 score, to counter the trade-off between selectivity (defined in Equation. 5-3) and recall, the average of the two is used and is referred as Balanced accuracy. Compared to Balanced accuracy scores, the ROC curve gives a pictorial representation of the trade-off between recall and fail out (which is 1-selectivity) (Johnson and Khoshgoftaar 2019). The area under the ROC curve serves as the numerical score, in this case. Since False alarms generally are one of the main deterrents when it comes to the usage of predictive models by TMCs (Guin, Williams, and Ni 2004), these additional metrics need to be explored (slated for future research).

5.3. Implementation

5.3.1. Hyperparameter Tuning

Several of the ML algorithms discussed in the previous section (listed in Table 5-1), were implemented and tested for performance in this study. The algorithms used in this study are typically the simpler first-generation classification algorithms and are chosen based on their simplicity of training and use, as demonstrated in previous studies (Filipovska and Mahmassani 2020). The performance of an ML algorithm is dependent on its Hyperparameters. Hyperparameters are user-specified parameters specific to a model architecture. These special variables are not optimized during the learning process so as to prevent overfitting and reduce the generalization error. Instead, these variables are commonly optimized using methods including, manual tuning, grid search, random search, and Bayesian optimization (Bergstra et al. 2011), (Karlaftis and Vlahogianni 2011). The algorithms and respective hyperparameter

optimization discussed below are based on the Scikit-learn package's implementation (Pedregosa et al. 2011) in Python®.

Hyperparameters for SVM depend on the choice of the kernel function, which are used to transform the input data into a higher dimensional space. Based on popular choice, two types of kernel functions were used here, polynomial and Radial Basis Function (RBF). The hyperparameters are degree of the polynomial and gamma value for polynomial and RBF, respectively. Here, the gamma value for RBF determines the weight to be given to a single training point (Pedregosa et al. 2011). Other than kernel specific parameters, C which represents the regularization term that penalizes the cost function used for training for misclassifications. A higher value for C can result in overfitting over the training dataset (Pedregosa et al. 2011). For NNs the hyperparameters are number of hidden layers, number of nodes for each layer, and the activation function. For Decision Trees, the hyperparameters are, the criterion to measure the quality of sample split, maximum allowable depth of the tree, and minimum impurity decrease. Among these, minimum impurity decrease is the threshold value by which the next sample split should improve for the selected quality measurement criterion. For Random Forest classifier, other than hyperparameters from the Decision Trees, are number of decision trees to be used. Other than these algorithm specific hyperparameters Scikit-learn package also allows an additional hyperparameter to counter the class-imbalance, called 'class-weight'. However, this additional parameter is currently not available for NNs. Since the dataset in use exhibits high class-imbalance, value to 'class-weight' was set to 'balance' during optimization. The discussed hyperparameters were optimized using the grid search over the specified values of the parameters. For grid-search, different combinations of specified values for these hyperparameters were generated forming a grid-like structure. The performance, using every point of this 'grid', was evaluated using the precision score. The performance can either be tested on an out-of-sample dataset or using the training dataset divided into 'k' parts where 'k-1' parts are treated as in-sample data and the kth sample is considered as out-of-sample. For this study, the latter technique, called as K-fold cross validation, was used where value for k was set as ten. These optimized hyperparameters might require periodic re-tuning during future use which can result in significant computational cost.

5.4. Results

A closer look at the time series plots of the prediction instances reveals that in some instances classifiers have labeled observations in the neighborhood of the true pre-congestion observation as pre-congestion points. For example, Figure 5-10 shows the congestion state-change label in yellow. The previous speed drop was not labeled as a congestion state-change point because the speeds did not stay below the threshold for the required 10 minutes state-persistence criteria. Figure 5-11 shows the multiple points in red that were all identified as congestion state-change points. Mathematically, these instances negatively affect the classifiers' performance. However, from an application point of view, multiple pre-congestion

warnings are not necessarily an indication of poor performance. This further suggests the requirement of modification in the definition of currently selected performance metrics.

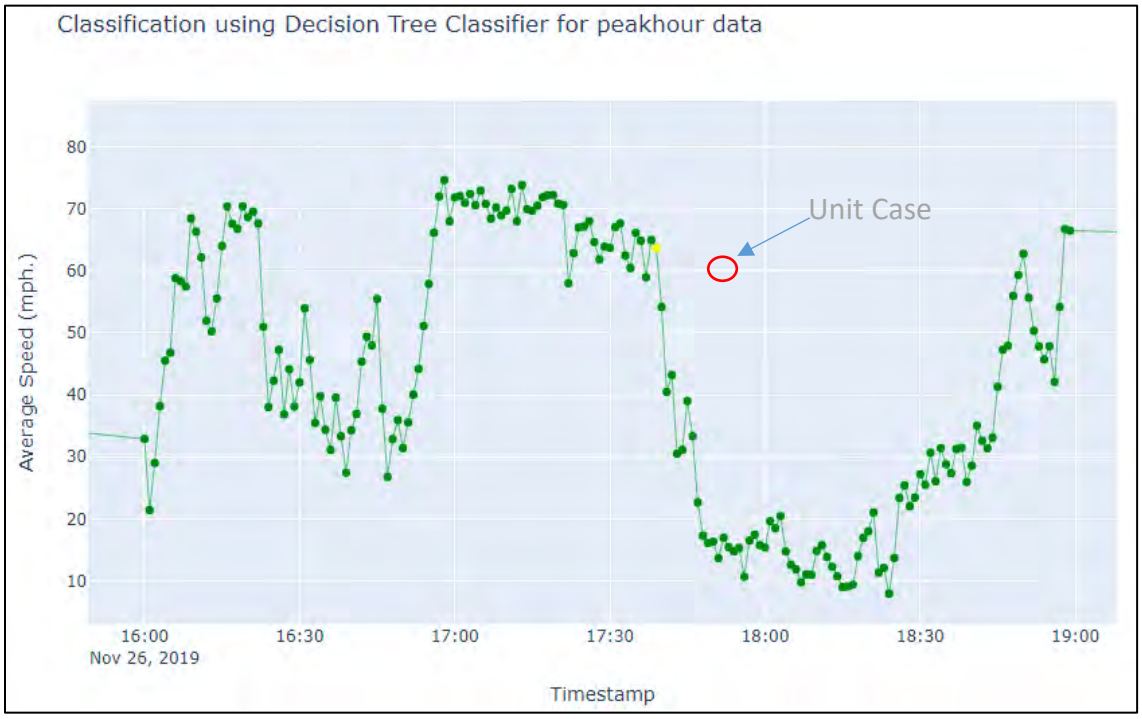


FIGURE 5.10: INPUT DATA LABELS FOR PRE-CONGESTION CASE FOR 11/26/2019

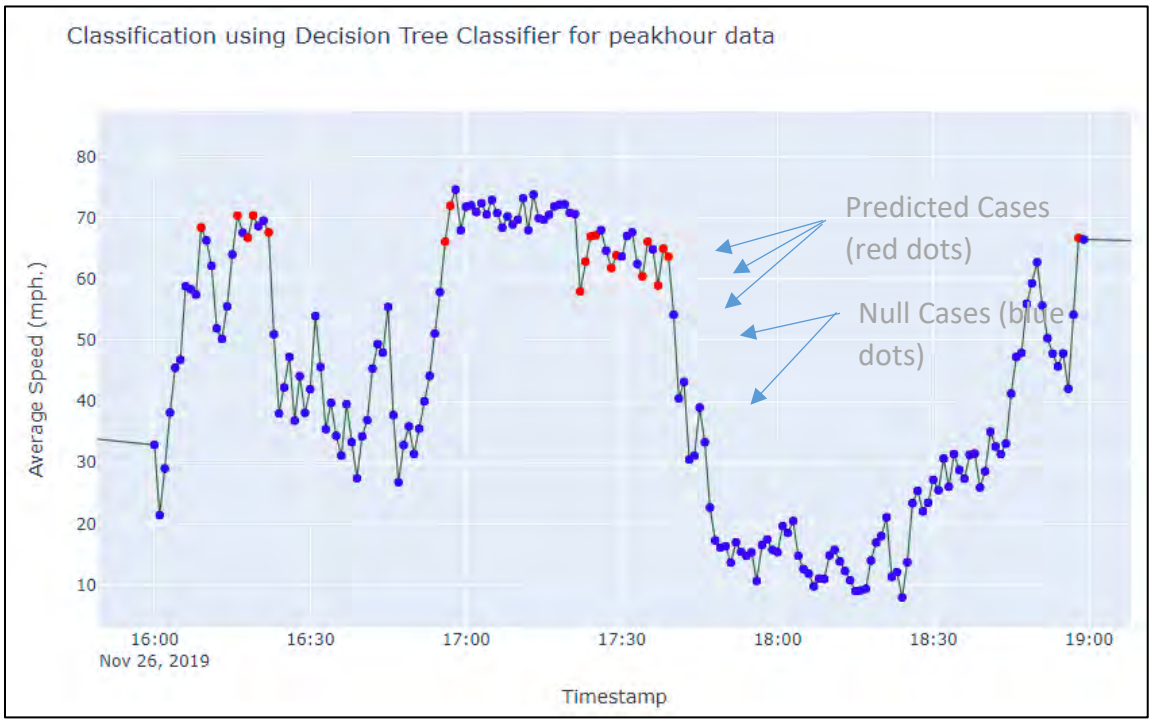


FIGURE 5.11: PREDICTED CLASSES USING DECISION TREE CLASSIFIER FOR 11/26/2019

The results also demonstrate that the current level of under-sampling, in the form of removal of off-peak-period data, is insufficient to counter the data imbalance. In the analysis of a class-imbalanced dataset, (Napierala and Stefanowski 2016), suggested examination of the 'neighborhood of minority class' datapoints. As a consequence, multiple types of minority class examples were discovered. Figure 5-12, 5-13, 5-14, 5-15 given below show a sample for such detected examples with varied congested conditions. In the figures, the vertical axis shows average speed (in mph) over all lanes for the selected site for the middle detector GDOT-STN-2851097, and the horizontal axis represents time-of-the-day. As in Figure 5-11, the yellow points represent the timestamps just before the congestion started and the green points represent all the other datapoints.

Some of the detected examples exhibited a major drop in the average speed, with values dropping to 30 mph or lower for an extended period of time. These examples were plausibly a result of a congestion around the selected site and were labeled as 'Major speed drop' cases. Other examples showed a minor drop in the speed, with the average speed still being above 50mph. In some cases, this drop in the average speed lasted for the entire peak hour duration while in other cases, only lasted for about less than 30 minutes. These examples were labeled as 'Minor speed drop' and 'Maybe' cases, respectively. These types of examples probably resulted due to recurrent congestion and were mainly observed during peak hours. (Krawczyk 2016) suggests developing different classifiers for each type of minority example. This categorization would ensure consistent prior distribution of predictor variables for each category of congestion and hence ensemble classifiers will be explored in future version.

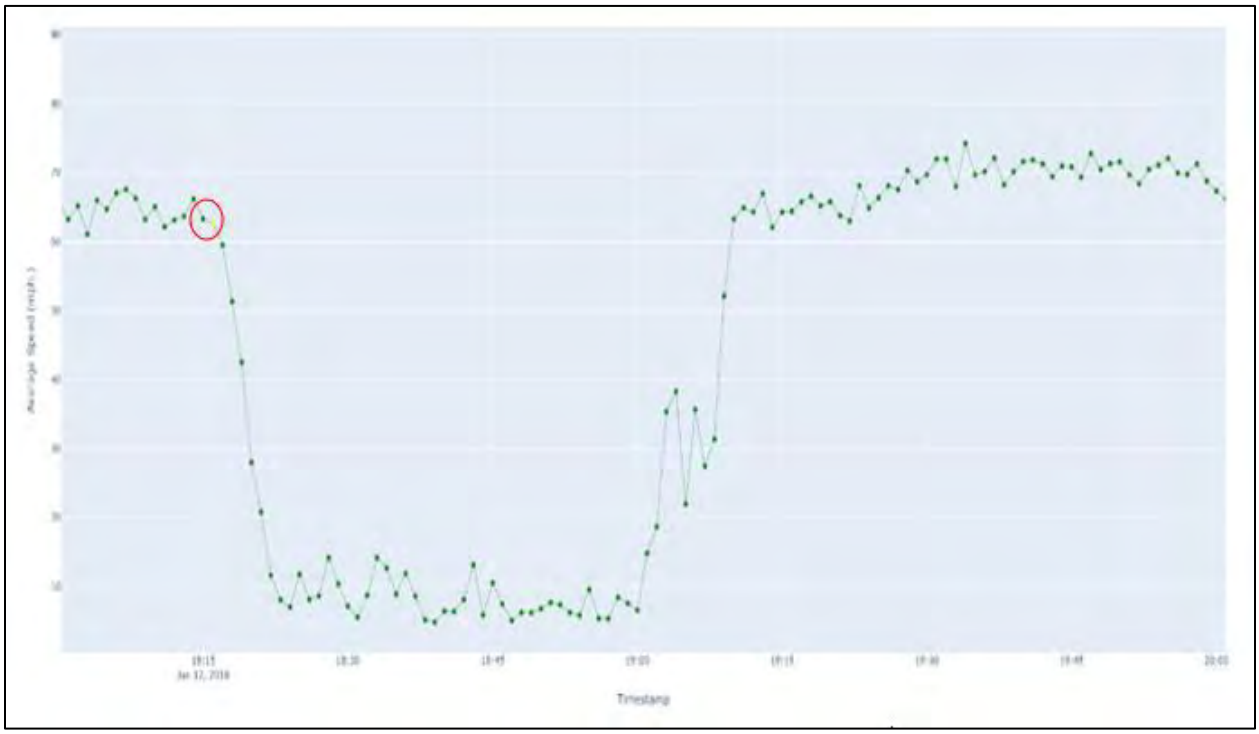


FIGURE 5.12: EXAMPLE OF MAJOR SPEED DROP CASE AT 01/25/2018 22:36

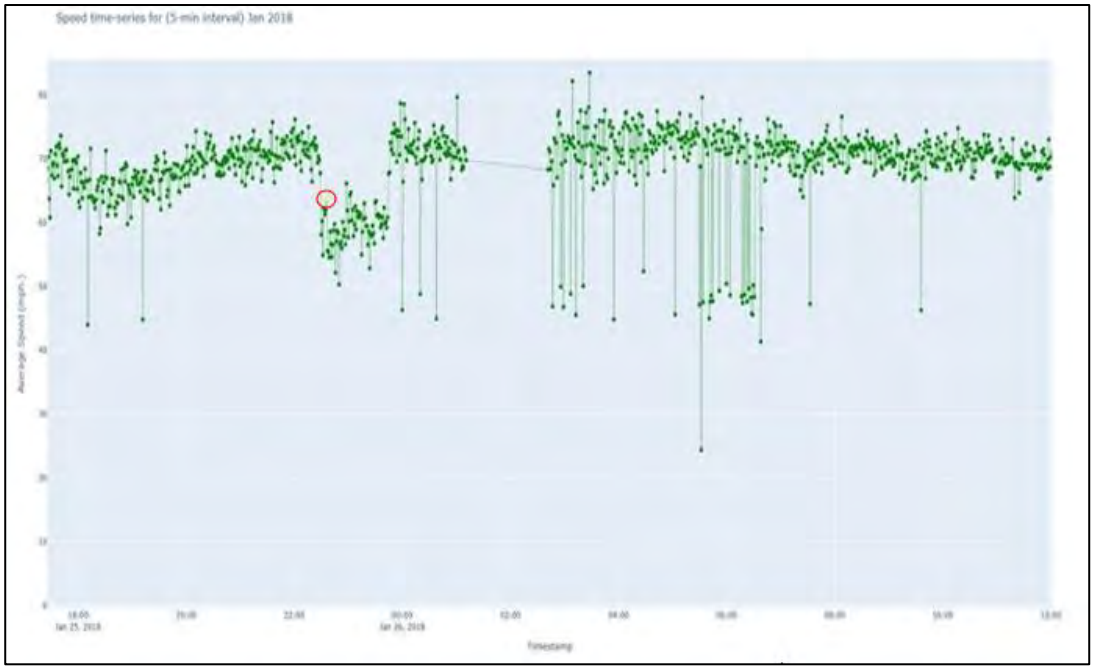


FIGURE 5.13: EXAMPLE OF MINOR SPEED DROP CASE AT 01/25/2018 22:36

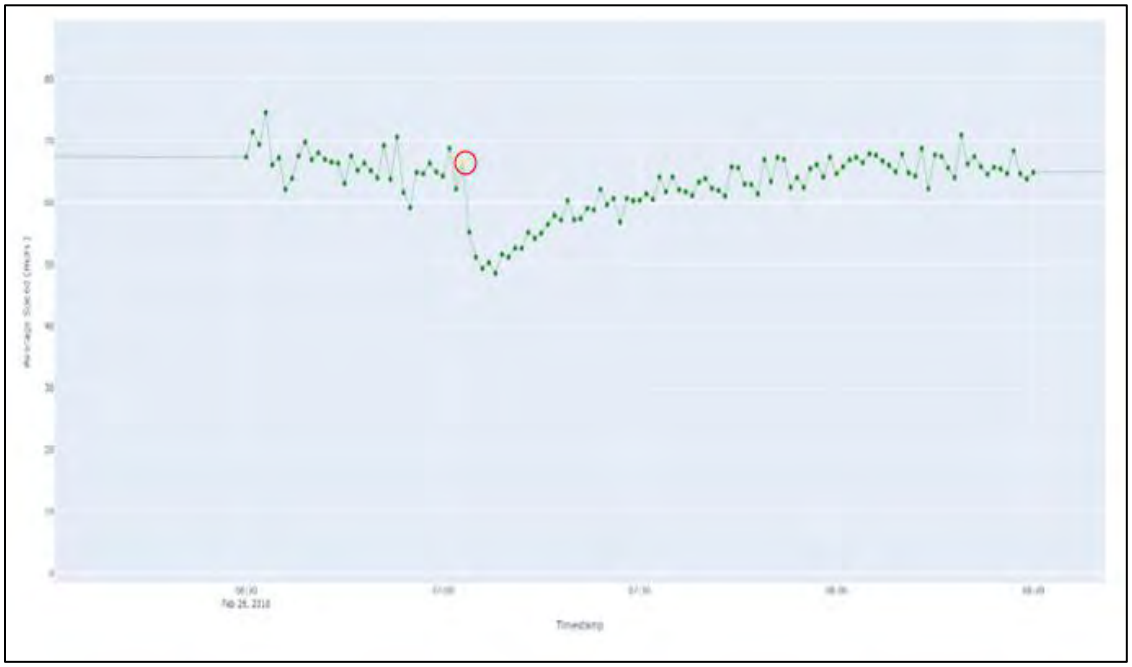


FIGURE 5.14 EXAMPLE OF CASE TAGGED AS 'MAYBE' AT 02/26/2018 07:03

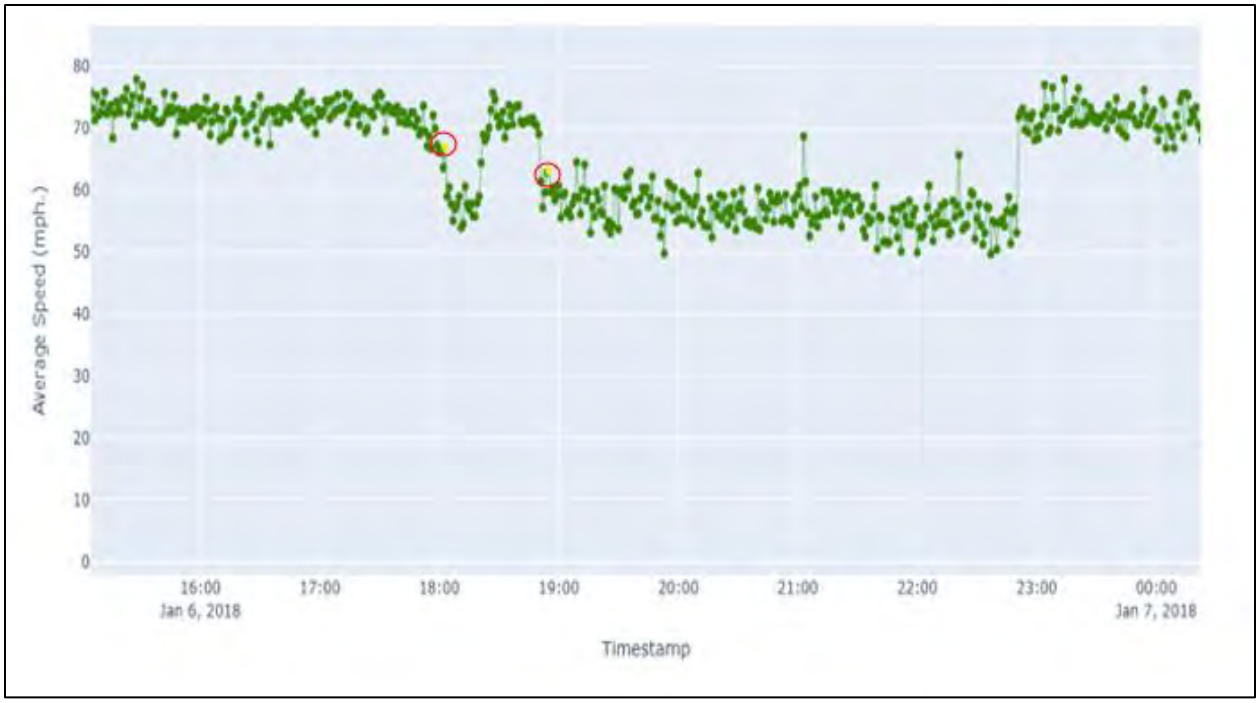


FIGURE 5.15 MULTIPLE CONGESTION ALARMS (TAGGED MINOR SPEED CASE) DETECTED AT 01/06/2018 18:03 AND 18:55 DURING THE EVENING PEAK HOUR

5.4.1. Results Summary

After hyperparameter optimization and the scaling of the input features, performance of the ML algorithms was evaluated on the test set, results of which along with the optimized hyperparameter values is given in Table 5-1 below. In the table, performance on both training and testing dataset is shown to help draw insights about the trade-off between training data requirement and overfitting. Higher performance on the training dataset, revealed by higher balanced accuracy, recall, and precision scores, compared to the same scores on the test dataset indicate that parameters learned by the model have overfitted over the training dataset. On the contrary, a lower performance on the training dataset suggests need for more training data as well as further refinement of the model. Table 5-1 shows the current classifiers are performing well on training dataset when it comes to Balanced accuracy and Recall scores, compared to the test dataset. This suggests that the classifiers need fine-tuning to reduce the generalization error and improve overfitting to the training dataset.

TABLE 5-1: PERFORMANCE ON TEST AND TRAINING DATASET

Algorithm	Balanced-Accuracy	Recall	Precision	Training Balanced-Accuracy	Training Recall	Training Precision	Specification
SVM-polynomial Kernel	0.50	0.00	0.00	1.00	0.99	0.35	C=1, kernel='poly', degree=3, tol=0.001, class_weight='balanced', random_state=seed
SVM-Radial Kernel	0.63	0.29	0.03	0.97	1.00	0.02	C=1, kernel='rbf', gamma=0.1,tol=0.001, class_weight='balanced', random_state=seed
Neural Network	0.50	0.00	0.00	0.50	0.00	0.00	hidden_layer_sizes= (20,10), max_iter=10000, random_state=seed, solver='lbfgs'
Decision Tree Classifier	0.76	0.57	0.02	0.91	0.88	0.02	class_weight='balanced', max_depth=4, min_impurity_decrease= 0.01, criterion='entropy', random_state=seed
Random Forest Classifier	0.69	0.43	0.01	0.94	0.99	0.01	class_weight='balanced', max_depth=10, min_impurity_decrease= 0.01, criterion='entropy', bootstrap=False, random_state=seed
Naïve Bayes classifier	0.69	0.43	0.01	0.71	0.77	0.00	

5.5. Discussion

From the results shown in Table 5-1 it can be observed that the Decision Tree and the Random Forest algorithms are the optimal classifiers in terms of the precision and recall scores**.** Analysis of the optimal Decision Tree revealed that the optimized classifiers only ended up using current average speed of the detector under study (GDOT-STN-2851097) and the downstream detector (GDOT-STN-2851098). This is plausible as the original deterministic method used to label the dataset has a tree-based structure. However, this severely restricts the ability to draw meaningful insights about spatiotemporal features other than ones used for labeling. Further, the current labeling of the pre-congestion data-points are dependent on the observation of and actual state change rather than the precursors. This might not provide an alarm early enough for practical use, and the objective is to identify the precursors that are likely embedded in the data in earlier data-points. Hence, a less-deterministic approach in creating the pre-congestion

class and input features will be considered in future iterations. The optimized tree structure also reveals the need for more specialized input features to help extract advance congestion warning. Additional exploration of the dataset suggests some likely candidates such as standard deviation of the average speed within a temporal window at the detector under study and downstream detector, speed-differential between lanes, etc.

5.6. SUMMARY

The case study presented in this chapter was aimed at identifying the precursors for imminent traffic congestion and developing a prediction model for congestion events. The model would allow traffic management agencies to take preemptive actions to minimize or mitigate the impacts of imminent traffic congestion. For the model development and testing, the study used vehicle detection data (speed and flow) from roadside detector stations, equipped with Video Detection System devices, in the Atlanta Metro area. The majority of previous studies have explored traffic congestion prediction problem using ML techniques in the context of short-term traffic flow or speed prediction. This study on the other hand, addressed a more complex problem and developed a methodology for identifying the precursors of congestion using freeway data. This problem was formulated as a binary classification task and resulted in a highly imbalanced dataset due to the limited datapoints corresponding to the precongestion class. The dataset was used to train a set of generative and discriminative Machine Learning classifiers. Performance of trained classifiers was tested using balanced accuracy, recall, and precision scores. Initials results have demonstrated superior accuracy performance from tree-based classifiers. Future efforts will tackle the additional complexities added by signalization on arterials.

6. CONCLUSIONS

At the outset of Phase 1 our aspiration was that we would be able to identify tools that could be used to detect DIC and IIC and assess system performance. The work in SHRP-2 L02 presented promise that this would be feasible. But until the ideas are tried, it is not clear whether they will work.

Our aspirations have been met. We have been able to create tools, fed by "big data", that highway system managers can use to identify the onset of congestion and the occurrence of incidents, for both freeways and arterials. We have also been able to create an off-line tool that they can use to assess past performance.

The tools all use the same basic idea – they look at the distribution of travel rates and flows, in real-time – and watch for trends of changing performance. The tools can be fed either vehicular data or roadside detector data, with differences in the mechanics. The tool for performance assessment processes data for travel rates and flows (by TMCs-and-5-minute intervals, each one is an *instance*) and identifies the number of instances that exceed a threshold. When it is fed vehicular data (say from Bluetooth sensors), it looks to see what percentage of the instances have at least *X*% (e.g., 85%) of the vehicles with travel rates greater than "*Y*", like 1.33 min/mi (45 mph). The reported metric is the percentage of instances that meet or exceed this criterion and where and when they occur (e.g., on I-40 for the TMCs upstream of the junction with I-540 during the AM and PM peak hours). The tool for identifying incidents works similarly. It looks for major, abrupt changes in the distribution of travel rates on these TMC-level segments. For example, If the entire distribution of rates rises abruptly, or the interval between vehicular observations increases abruptly, this suggests a blockage for all vehicles, on the TMC or upstream. If just the lower percentiles rise, this suggests a blockage in the left-most lane, because the fastest moving vehicles have slowed down. If the rates rise for the higher percentiles (the slowest moving vehicles), it is likely that there is a partial blockage of the right-most lane. The tool for congestion detection is more sophisticated. It watches for a rise in the 5th percentile travel rate (the faster moving vehicles) in conjunction with an upward shift in the 85th percentile travel rate (the slower moving vehicles) and predicts how long it will be (minutes) until a certain percentage (e.g., 15%) of the travel rates exceed the acceptable performance criterion (e.g., they will have a travel rate greater than 1.33 min/mi). Its finding ranges from "green" – it is not likely in the next 30 minutes – to "red" – it is likely to happen in the next 5 minutes. This phase I report describes what we have done to develop these tools; the ideas we tried; those that worked and did not work; the data we used; and the current status of their refinement.

Our anticipation is that these tools will reduce the severity of the impacts from congestion and incidents because their occurrence will be detected sooner, especially for congestion, and more reliably. Also, the performance assessment tool will help network managers identify where system improvements are needed and defend quantitatively, their benefits.

In the next phase of this project, our intent is to further refine and test the tools to apply in real-time for monitoring and detecting congestion onset. To this end, we will engage stakeholders from different transportation agencies, which would also help developing a software and a user manual. We also plan to extend the effort demonstrated here to arterial streets, with a view to testing the potential of big data to enhance the current congestion management strategies.

7. FUTURE WORK

Our expectation for Phase 2 is that we will continue to enhance and refine the tools we have developed in Phase 1. We will test the quality and reliability of their detection and assessment capabilities using datasets from Sacramento, Atlanta, and Tampa. We will strive to obtain additional datasets from other areas so we can do further and additional testing.

We will also engage in substantial outreach, to our identified stakeholders and to other interested organizations and individuals so we can make them aware of the work we have done and the tools we have developed. With the easing of pandemic restrictions, this will be more engaging and easier to do. We also plan to extend the effort demonstrated here to arterial streets, with a view to testing the potential of big data to enhance the current congestion management strategies.

The culmination of Phase 2 will be a final report that describes what we have done and final deliverables including the tools we have created and electronic copies of the tests we have conducted and our findings.

8. LIST OF ACRONYMS

ABS	Antilock Brake System
AUC	Area Under Roc
AV	Autonomous Vehicles
AVI	Automatic Vehicle Identification
AVL	Automatic Vehicle Location
BSM	Basic Safety Message
BSM	Basic Safety Message
BT	Bluetooth
CART	Classification and Regression Trees
CDF	Cumulative Distribution Function
CV	Connected Vehicles
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DIC	Demand-induced Congestion
DSRC	Dedicated Short-range Communication
GARCH	Generalized Autoregressive Conditional Heteroscedasticity
IIC	Incident-induced Congestion
MAC	Media Access Control
MARE	Mean Absolute Relative Error
ML	Machine Learning
OBU	On-board Unit
OD	Origin-destination
PC	Principal Components
PDF	Probability Density Function
RBF	Radial Basis Function
ROC	Receiver Operating Characteristics
RSSI	Received Signal Strength Indicator
RSU	Roadside Unit
SAE	Society of Automotive Engineers
SDbv	Standard Deviation Between Individual Vehicles
SDdp	Standard Deviation Between Data Points
SDv	Standard Deviation of Individual Vehicles
SPaT	Signal Phasing and Timing
SVM	Support Vector Machines
TCS	Traction Control System
TET	Time-to-collision
THEA	Tampa-Hillsborough Expressway Authority
TIM	Traveler Information Massages)
TTRMS	Travel Time Reliability Monitoring System
V2I	Vehicle-to-infrastructure
V2V	Vehicle-to-vehicle

VDS	Video Detection System
WCSS	Within Clusters Sum of Squares

9. REFERENCE LIST

TRB Publication

Ahmed, I., Rouphail, N. M., & Tanvir, S. (2018). Characteristics and temporal stability of recurring bottlenecks. Transportation Research Record, 2672(42), 235–246.

Ahmed, I., Williams, B. M., & Samandar, M. S. (2018). Application of a Discontinuous Form of Macroscopic Gazis–Herman–Rothery Model to Steady-State Freeway Traffic Stream Observations. Transportation Research Record, 2672(20), 51–62.

Ahmed, M. S., and A. R. Cook. 1979. 'ANALYSIS OF FREEWAY TRAFFIC TIME-SERIES DATA BY USING BOX-JENKINS TECHNIQUES', Transportation Research Record.

Azizi, L., and Hadi, M. (2020). Utilizing Traffic Disturbance Metrics to Estimate and Predict Freeway Traffic Breakdown and Safety Events. Transportation Research Record: Journal of the Transportation Research Board. Washington, D.C., August 2020.

Chen, C., Skabardonis, A., & Varaiya, P. (2003). Travel-time reliability as a measure of service. Transportation Research Record, 1855(1), 74–79.

Clark, S. D., Grant-Muller, S., & Chen, H. (2002). Cleaning of matched license plate data. Transportation Research Record, 1804(1), 1–7.

Haghani, A., Hamedi, M., Sadabadi, K. F., Young, S., & Tarnoff, P. (2010). Data collection of freeway travel time ground truth with bluetooth sensors. Transportation Research Record, 2160(1), 60–68.

Hellinga, B., & Knapp, G. (2000). Automatic vehicle identification technology-based freeway incident detection. Transportation Research Record, 1727(1), 142–153.

Hua, J., and A. Faghri. 1994. 'APPLICATIONS OF ARTIFICIAL NEURAL NETWORKS TO INTELLIGENT VEHICLE-HIGHWAY SYSTEMS', Transportation Research Record.

INRIX, 2022. https://inrix.com/press-releases/safety-alerts/

List, G. F., Rouphail, N., Smith, R., & Williams, B. (2018). Reliability Assessment Tool: Development and Prototype Testing. Transportation Research Record, 2672(14), 29-38

Mahmassani, H. S., Hou, T., & Dong, J. (2012). Characterizing travel time variability in vehicular traffic networks: deriving a robust relation for reliability analysis. Transportation Research Record, 2315(1), 141–152.

Malinovskiy, Y., Wu, Y.-J., Wang, Y., & Lee, U. K. (2010). Field experiments on bluetooth-based travel time data collection. (No. 10-3134)

Moghaddam, S. S., & Hellinga, B. (2014a). Algorithm for detecting outliers in Bluetooth data in real time. Transportation Research Record, 2442(1), 129–139.

Moghaddam, S. S., & Hellinga, B. (2014b). Real-time prediction of arterial roadway travel times using data collected by bluetooth detectors. Transportation Research Record, 2442(1), 117–128.

Saeedi, A., Park, S., Kim, D. S., & Porter, J. D. (2013). Improving accuracy and precision of travel time samples collected at signalized arterial roads with bluetooth sensors. Transportation Research Record, 2380(1), 90–98.

Samandar, M Shoaib, Williams, B. M., & Ahmed, I. (2018). Weigh Station Impact on Truck Travel Time Reliability: Results and Findings from a Field Study and a Simulation Experiment. Transportation Research Record, 2672(9), 120–129.

Van Boxel, D., Schneider IV, W. H., & Bakula, C. (2011). Innovative real-time methodology for detecting travel time outliers on interstate highways and urban arterials. Transportation Research Record, 2256(1), 60–67.

Van Lint, J. W. C., & van Zuylen, H. J. (2005). Monitoring and predicting freeway travel time reliability: Using width and skew of day-to-day travel time distribution. Transportation Research Record,

1917(1), 54–62.
Xu, Y., Williams, B. M., Rouphail, N. M., & Chase, R. T. (2013). Development of an Oversaturated Speed–Flow Model Based on the Highway Capacity Manual. Transportation Research Record, 2395(1), 41–48.

Book

Bishop, Christopher M. 2006. Pattern recognition and machine learning (New York : Springer, [2006] ©2006).
Breiman, L., J. Friedman, C.J. Stone, and R.A. Olshen. 1984. Classification and Regression Trees (Taylor & Francis).
Han, J., and Kamber, M. (2012). Data Mining Concepts and Techniques. 3rd. Ed. Elsevier Direct.
Newland, D. E. Random Vibrations: Spectral and Wavelet Analysis. John Wiley & Sons, Inc., New York, 1998

Book Chapter

Shunk, G. A. Urban Transportation Systems. In Transportation Planning Handbook (J. D. Edwards, Jr., ed.), Prentice Hall, Englewood Cliffs, N.J., 1992, pp. 88–122

Government Report

Hadi, M., Xiao, Y., Iqbal, M. S., Khazraeian, S., Sturgeon, I. I., & Purser, K. 2018. Utilization of Connected Vehicle Data to Support Traffic Management Decisions (No. BDV29-977-21). FDOT. Office of Research and Development.
List, G. F., Williams, B. M., Rouphail, N. M., Hranac, R., Barkley, T., Mai, E., Ciccarelli, A., Rodegerdts, L., Pincus, K., & Nevers, B. (2014). Establishing Monitoring Programs for Travel Time Reliability [supporting datasets]. United States. National Transportation Library [distributor].
Margiotta, R., Lomax, T. J., Hallenbeck, M. E., Turner, S. M., Skabardonis, A., Ferrell, C., & Eisele, W. L. (2006). Guide to effective freeway performance measurement: Final report and guidebook. (No. NCHRP Project 3-68)
Tan, P., Steinbach, M., Karpatne, A., and Kumar, V. (2019). Introduction to Data Mining. 2nd Ed. Pearson, NY, USA. ISBN-10: 0-13-312890-3.
United States Department of Transportation (USDOT). (2018). Connected Vehicle Pilot Deployment Program. https://www.its.dot.gov/pilots/
United States Department of Transportation (USDOT). (2017). CV Pilot Deployment Program: Tampa Factsheet. https://www.its.dot.gov/factsheets/pdf/TampaCVPIlot_Factsheet.pdf
Vadakpat, G. (2018). Tampa (THEA) CV Pilot Site. CV Deployment Pilot Program Presentation, https://www.its.dot.gov/pilots/pdf/CVP-TampaTHEA_v4.pdf, Accessed May 20, 2018.

United States Department of Transportation (USDOT). (2018). Connected Vehicle Pilot Deployment Program. https://www.its.dot.gov/pilots/
United States Department of Transportation (USDOT). (2017). CV Pilot Deployment Program: Tampa Factsheet. https://www.its.dot.gov/factsheets/pdf/TampaCVPIlot_Factsheet.pdf
Vadakpat, G. (2018). Tampa (THEA) CV Pilot Site. CV Deployment Pilot Program Presentation, https://www.its.dot.gov/pilots/pdf/CVP-TampaTHEA_v4.pdf, Accessed May 20, 2018.
Von Quintus, H. L., and A. L. Simpson. Documentation of the Back calculation of Layer Parameters for LTPP Test Sections. Publication FHWA-RD-01-113. FHWA, U.S. Department of Transportation, 2002.
Zou, Z., Li, M., & Bu, F. (2010). Link travel time estimation based on vehicle infrastructure integration probe data. In ICCTP 2010: Integrated Transportation Systems: Green, Intelligent, Reliable (pp. 2266-2276).

Website

Federal Highway Administration (2017c), *Two Performance Management Final Rules Take Effect*, https://www.fhwa.dot.gov/tpm/rule.cfm, 2017.

National Archives and Records Administration (2017), *National Performance Management Measures; Assessing Performance of the National Highway System, Freight Movement on the Interstate System, and Congestion Mitigation and Air Quality Improvement Program*, https://www.federalregister.gov/documents/2017/01/18/2017-00681/national-performance-management-measures-assessing-performance-of-the-national-highway-system, 2017.

Federal Highway Administration (2017a), *How is Transportation Performance Monitoring being Implemented*?, https://www.fhwa.dot.gov/tpm/about/how.cfm, accessed July 2017.

Retrieved from http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

State and Local Policy Program. Value Pricing. Hubert H. Humphrey Institute of Public Affairs, University of Minnesota, Minneapolis. www.hhh.umn.edu/centers/slp/vp/vp_org. Accessed Feb. 5, 2008.

Transportation Research Board (2017b), *The Second Strategic Highway Research Program (2006-2015),* http://www.trb.org/StrategicHighwayResearchProgram2SHRP2/Blank2.aspx, accessed 2017.

Other Publications/Journals

Abdulhai, Baher, Himanshu Porwal, and Will Recker. 2002. 'Short-Term Traffic Flow Prediction Using Neuro-Genetic Algorithms', Journal of Intelligent Transportation Systems, 7: 3-41.

Avery Rhodes, Edward J. Smaglik, Darcy M. Bullock. 2006. "Vendor Comparison of Video Detection Systems " In, 41. Joint Transportation Research Program, Indiana Department of Transportation and Purdue University, West Lafayette, Indiana.

Bates, J., Polak, J., Jones, P., & Cook, A. (2001). The valuation of reliability for personal travel. Transportation Research Part E: Logistics and Transportation Review, 37(2–3), 191–229.

Bergstra, James, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. "Algorithms for hyper-parameter optimization." In Proceedings of the 24th International Conference on Neural Information Processing Systems, 2546–54. Granada, Spain: Curran Associates Inc.

Bhaskar, A., & Chung, E. (2013). Fundamental understanding on the use of Bluetooth scanner as a complementary transport data. Transportation Research Part C: Emerging Technologies, 37, 42–72.

Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. "A training algorithm for optimal margin classifiers." In Proceedings of the fifth annual workshop on Computational learning theory, 144–52. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery.

Breiman, Leo. 2001. 'Random Forests', Machine Learning, 45: 5-32.

Castro-Neto, Manoel, Young-Seon Jeong, Myong-Kee Jeong, and Lee D. Han. 2009. 'Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions', Expert Systems with Applications, 36: 6164-73.

Cho, Hyun Woong. 2017. 'Modeling and simulation of congestion control strategies on freeways: Pricing, ramp metering, and variable speed limit', Doctoral, Georgia Institute of Technology.

Chun-Hsin, Wu, Ho Jan-Ming, and D. T. Lee. 2004. 'Travel-time prediction with support vector regression', IEEE Transactions on Intelligent Transportation Systems, 5: 276-81.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. Journal of the

American Statistical Association, 74(368), 829–836.
Delhome, R., Billot, R., & El Faouzi, N.-E. (2017). Travel time statistical modeling with the Halphen distribution family. Journal of Intelligent Transportation Systems, 21(6), 452–464.
Dion, F., & Rakha, H. (2006). Estimating dynamic roadway travel times using automatic vehicle identification data for low sampling rates. Transportation Research Part B: Methodological, 40(9), 745–766.
Dougherty, Mark. 1995. 'A review of neural networks applied to transport', Transportation Research Part C: Emerging Technologies, 3: 247-60.
Dr. Peter T. Martin, Gayathri Dharmavaram and Aleksandar Stevanovic. 2004. "Evaluation of UDOT'S Video Detection Systems " In.: Department of Civil and Environmental Engineering, University of Utah Traffic Lab.
Dunne, Stephen, and Bidisha Ghosh. 2012. 'Regime-Based Short-Term Multivariate Traffic Condition Forecasting Algorithm', Journal of Transportation Engineering, 138: 455-66.
Filipovska, Monika, and Hani S. Mahmassani. 2020. 'Traffic Flow Breakdown Prediction using Machine Learning Approaches', Transportation Research Record, 2674: 560-70.
Friedman, Nir, Dan Geiger, and Moises Goldszmidt. 1997. 'Bayesian Network Classifiers', Mach. Learn., 29: 131–63.
Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. Deep Learning (The MIT Press).
Grant, Christopher, Bret Gillis, and Randall Guensler. 2000. 'Collection of Vehicle Activity Data by Video Detection for Use in Transportation Planning', ITS Journal - Intelligent Transportation Systems Journal, 5: 343-61.
Guensler, Randall, Vetri Elango, Angshuman Guin, Michael Hunter, Jorge Laval, and et al. 2013. 'Atlanta I-85 HOV-to-HOT conversion : analysis of vehicle and person throughput'.
Guin, Angshuman. 2004. 'An Incident Detection Algorithm Based On a Discrete State Propagation Model of Traffic Flow', Dissertation, Georgia Institute of Technology.
Guin, Angshuman, Billy M. Williams, and D. Ni. 2004. "Assessment of the Current Status of Incident Detection Algorithms: Results of a Nationwide Survey." In.
Hand, David J., and Keming Yu. 2001. 'Idiot's Bayes: Not So Stupid after All?', International Statistical Review / Revue Internationale de Statistique, 69: 385-98.
Impedovo, Donato, Fabrizio Balducci, Vincenzo Dentamaro, and Giuseppe Pirlo. 2019. 'Vehicular Traffic Congestion Classification by Visual Features and Deep Learning Approaches: A Comparison', Sensors, 19: 5213.
James Bonneson, Montasir Abbas. 2002. "Video detection for intersection and interchange control." In, 96. Texas Department of Transportation: Texas Transportation Institute.
Johnson, Justin M., and Taghi M. Khoshgoftaar. 2019. 'Survey on deep learning with class imbalance', Journal of Big Data, 6: 27.
Kamarianakis, Yiannis, Angelos Kanas, and Poulicos Prastacos. 2005. 'Modeling Traffic Volatility Dynamics in an Urban Network', Transportation Research Record, 1923: 18-27.
Karlaftis, M. G., and E. I. Vlahogianni. 2011. 'Statistical methods versus neural networks in transportation research: Differences, similarities and some insights', Transportation Research Part C: Emerging Technologies, 19: 387-99.
Krawczyk, Bartosz. 2016. 'Learning from imbalanced data: open challenges and future directions', Progress in Artificial Intelligence, 5: 221-32.

Kim, J., & Mahmassani, H. S. (2015). Compound Gamma representation for modeling travel time variability in a traffic network. Transportation Research Part B: Methodological, 80, 40–63.
Lan, Guanghui (George). 2020. "Support Vector Machines and Stochastic Subgradient Descent." In ISyE/CSE 6740 Lectures- Spring 2020, edited by Guanghui (George) Lan. Georgia Institute of

Technology: Lan, Guanghui (George).
Lee, Sangsoo, and Daniel B. Fambro. 1999. 'Application of Subset Autoregressive Integrated Moving Average Model for Short-Term Freeway Traffic Volume Forecasting', Transportation Research Record, 1678: 179-88.
Lippi, M., M. Bertini, and P. Frasconi. 2013. 'Short-Term Traffic Flow Forecasting: An Experimental Comparison of Time-Series Analysis and Supervised Learning', IEEE Transactions on Intelligent Transportation Systems, 14: 871-82.
Liu, H. (2008). Travel time prediction for urban networks.
Liu, Y., Xia, J. C., & Phatak, A. (2020). Evaluating the Accuracy of Bluetooth-Based Travel Time on Arterial Roads: A Case Study of Perth, Western Australia. Journal of Advanced Transportation, 2020.
Liu, Y., and H. Wu. 2017. "Prediction of Road Traffic Congestion Based on Random Forest." In 2017 10th International Symposium on Computational Intelligence and Design (ISCID), 361-64.
Lo, H. K. (2001). A cell-based traffic control formulation: strategies and benefits of dynamic timing plans. Transportation Science, 35(2), 148–164.
Margreiter, M. (2016). Automatic incident detection based on bluetooth detection in northern Bavaria. Transportation Research Procedia, 15, 525–536.
Mercader, P., & Haddad, J. (2020). Automatic incident detection on freeways based on Bluetooth traffic monitoring. Accident Analysis & Prevention, 146, 105703.
Mitsakis, E., Salanova Grau, J. M., Chrysohoou, E., & Aifadopoulou, G. (2015). A robust method for real time estimation of travel times for dense urban road networks using point-to-point detectors. Transport, 30(3), 264–272.
Moonam, H. M. (2016). Developing sampling strategies and predicting freeway travel time using Bluetooth data.
Moorthy, C. K., and B. G. Ratcliffe. 1988. 'Short term traffic forecasting using time series methods', Transportation Planning and Technology, 12: 45-56.
Napierala, Krystyna, and Jerzy Stefanowski. 2016. 'Types of minority class examples and their influence on learning classifiers from imbalanced data', Journal of Intelligent Information Systems, 46: 563-97.
Ng, Andrew Y., and Michael I. Jordan. 2001. "On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes." In Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, 841–48. Vancouver, British Columbia, Canada: MIT Press.
Park, Dongjoo, and Laurence R. Rilett. 1999. 'Forecasting Freeway Link Travel Times with a Multilayer Feedforward Neural Network', Computer-Aided Civil and Infrastructure Engineering, 14: 357-67.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M. et al (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.
Provost, Foster, and Tom Fawcett. 1997. "Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions." In Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, 43–48. Newport Beach, CA: AAAI Press.
Robinson, S., & Polak, J. (2006). Overtaking rule method for the cleaning of matched license-plate data. Journal of Transportation Engineering, 132(8), 609–617.
Samandar, Mohamad Shoaib. (2019). Freeway Travel Time Reliability: Traditional, Connected and Autonomous Vehicles Perspectives.
Sansalone, M., J. M. Lin, and W. B. Street. Determining the Depths of Surface-Opening Cracks Using Impact-Generated Stress Waves and Time-of-Flight Techniques. ACI Materials Journal, 2018. 95: 168–177.
Stathopoulos, Anthony, and Matthew G. Karlaftis. 2003. 'A multivariate state space approach for urban traffic flow modeling and prediction', Transportation Research Part C: Emerging Technologies, 11:

121-35.
Suh, Won Ho, James Anderson, Angshuman Guin, and Michael Hunter. 2015. 'Evaluation of Traffic Data Collection Method', Applied Mechanics and Materials, 764-765: 905-09.
Taylor, M. A. P. (1999). Dense network traffic models, travel time reliability and traffic management. II: Application to network reliability. Journal of Advanced Transportation, 33(2), 235–251.
Taylor, M. A. P. (2017). Fosgerau's travel time reliability ratio and the Burr distribution. Transportation Research Part B: Methodological, 97, 50–63.
Treiber, Martin, and Arne Kesting. 2013. Traffic Flow Dynamics: Data, Models and Simulation (Springer-Verlag Berlin Heidelberg).
Vapnik, Vladimir N. 1995. The nature of statistical learning theory (Springer-Verlag).
Vasudevan, Meenakshy, Chris Curtis, Alexa Lowman, and James O'Hara. 2016. 'Big data analytics : predicting traffic flow regimes from simulated connected vehicle messages using data analytics and machine learning'.
Vlahogianni, Eleni I. 2007. 'Prediction of non-recurrent short-term traffic patterns using genetically optimized probabilistic neural networks', Operational Research, 7: 171-84.
Vlahogianni, Eleni I., Matthew G. Karlaftis, and John C. Golias. 2014. 'Short-term traffic forecasting: Where we are and where we're going', Transportation Research Part C: Emerging Technologies, 43: 3-19.
Wang, J., X. Li, S. S. Liao, and Z. Hua. 2013. 'A Hybrid Approach for Automatic Incident Detection', IEEE Transactions on Intelligent Transportation Systems, 14: 1176-85.
Wells, Bill. 2016. 'NaviGAtor', Intelligent Transportation Society Georgia http://www.itsga.org/navigator/.
Williams, Billy, and Lester Hoel. 2003. 'Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical Results', Journal of Transportation Engineering, 129: 664-72.
Woodard, D., Nogin, G., Koch, P., Racz, D., Goldszmidt, M., & Horvitz, E. (2017). Predicting travel time reliability using mobile phone GPS data. Transportation Research Part C: Emerging Technologies, 75, 30–44.
Xia, Jingxin, Wei Huang, and Jianhua Guo. 2012. 'A clustering approach to online freeway traffic state identification using ITS data', KSCE Journal of Civil Engineering, 16: 426-32.
Yu, W., Park, S., Kim, D. S., & Ko, S.-S. (2015). Arterial road incident detection based on time-moving average method in bluetooth-based wireless vehicle reidentification system. Journal of Transportation Engineering, 141(3), 4014084.
Zou, Z., Li, M., & Bu, F. (2010). Link travel time estimation based on vehicle infrastructure integration probe data. In ICCTP 2010: Integrated Transportation Systems: Green, Intelligent, Reliable (pp. 2266-2276).

Unpublished Papers

Corbett, J. J. Toward Environmental Stewardship: Charting the Course for Marine Transportation. Presented at 83rd Annual Meeting of the Transportation Research Board, Washington, D.C., 2018.
Tu, H., van Lint, H., & van Zuylen, H. (2008). The effects of traffic accidents on travel time reliability. 2008 11th International IEEE Conference on Intelligent Transportation Systems, 79–84.

10. APPENDICES

APPENDIX A. Time Until Congestion Using Regression

Factors that contribute to the variation of speed of different vehicles that are traversing a road at the same time include but are not limited to the vehicle type, lane position, and driving behavior. During free-flow conditions, this variation of speed is usually caused by differences in the desired travel speeds among the drivers. As traffic density increases, the flow becomes restrictive, and the speed variation drops. This drop is primarily attributed by the reduction in speed of the faster vehicles. The speed of the slower moving vehicles in the traffic stream, on the other hand, remains almost unchanged until this transition state progresses to congested state. These phenomena are illustrated with real-world data in the previous section. In this section, we discuss the application of these potential signals, along with other predictors, for characterizing time-to-congestion onset. Time-to-congestion onset is estimated as the time difference between the next congestion onset time and the current time. Further, this drop in speed variation typically occurs gradually when traffic state changes due to recurring bottlenecks, whereas it tends to occur abruptly due to other non-recurring events. Hence, it is also expected that the nature of this change in speed variation would be able to tell the signal for traffic state transition.

As explained Section 3.2, we elected to aggregate the travel rate observations in moving groups of *N* observations (SHRP-2 L02 used 50 with an overlap of 25. We are using groups of 30 with an overlap of 24.) We used these groups to estimate percentiles of the travel rate distributions. Vendors of probe vehicle speed and system detector data usually aggregate these data by fixed time intervals (e.g., INRIX, PeMS). However, to compare distributions of raw probe data for different groups, it is important that we keep the number of samples for each distribution fixed. With a fixed-time interval data, the number of observations will vary. For instance, late at night, a 15-minute interval may contain very few observations, whereas in the daytime, that same interval duration may contain hundreds. Hence, we chose to aggregate data based on a fixed number of observations instead of a fixed time interval. This makes the night-time distributions have long spans (the difference in time between the first and last probe observation), but one may reasonably assume that the traffic state remains almost the same for long time spans during the night-time, unless a disruptive event takes place. Note that this aggregation process must be adjusted for the incident detection tool so that the night-time disruptive incidents can be detected.

Here, we are using groups of 30 consecutive observations with an 80% overlap between successive groups. Figure A.1 shows an example of the grouping window for two successive groups. The solid and the dashed lines represent, respectively, the start and end of each group. The shaded rectangle represents the overlap between these groups. Note that for clearly explaining the time span, the start and the end lines for each group in this figure are separated by more than 30 observations.

Intuitively, higher the number of observations in a group, higher is the confidence and degree of freedom for a modeling or quantile estimation for each group. However, the tradeoff is that the time span of the groups also increases as the number of observations increases within each group. The number of vehicles detected and matched between two Bluetooth sensors represent only 6-10% of the total flow rate on a road. Hence, even to form a group of 30 consecutive probe vehicle observations, one may have to wait for several hours during night-times. During or right before the peak hours, a group of 30 probe vehicle observations usually spans over a period ranging from 4 to 10 minutes. This period is reasonable given that our objective is to predict congestion onset within a certain time (e.g., 10-30 minutes) earlier so that transportation agencies can deploy necessary operational measures.

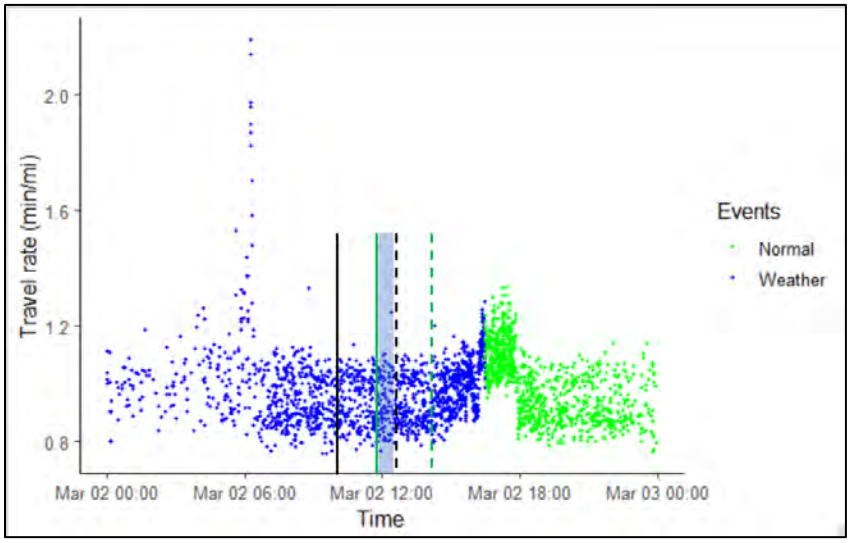


FIGURE A.1: TRAVEL RATE GROUPING TECHNIQUE FOR THE RAW DATA

The raw travel rate observations and the estimates of their percentiles in each group appear to have significant randomness. The scattered points in Figure A.2, representing the 5% and 95% travel rates for each group, clearly show this randomness. In addition to the variations of these percentiles before and during congestion, they exhibit significant random variations during other periods. Because such random variations can mask the signal of congestion onset, we tested several time-series smoothing algorithms on the raw data. Among those, the locally weighted regression function (Cleveland, 1979), also known as LOESS, seemed to perform well. Its performance was visually evaluated by observing how much randomness it can dampen without masking the actual signals and patterns.

The application of the LOESS algorithm to the grouped travel rate statistics (say, 5% travel rate) is briefly explained here. For each group *i*, it fits a weighted regression line through the 5% travel rate for *n* neighboring groups. The weight for the neighboring groups is inversely proportional to the distance from the *i*th group. Two inputs for this algorithm are the value for *n* and the sliding window (the number of data to shift to fit the line for the next group). Visually,

it appeared that an $n = 15$ and a $sliding\ window\ = 1$ generate satisfactory results. The same process can be applied to the 95% or any other percentiles of travel rate.

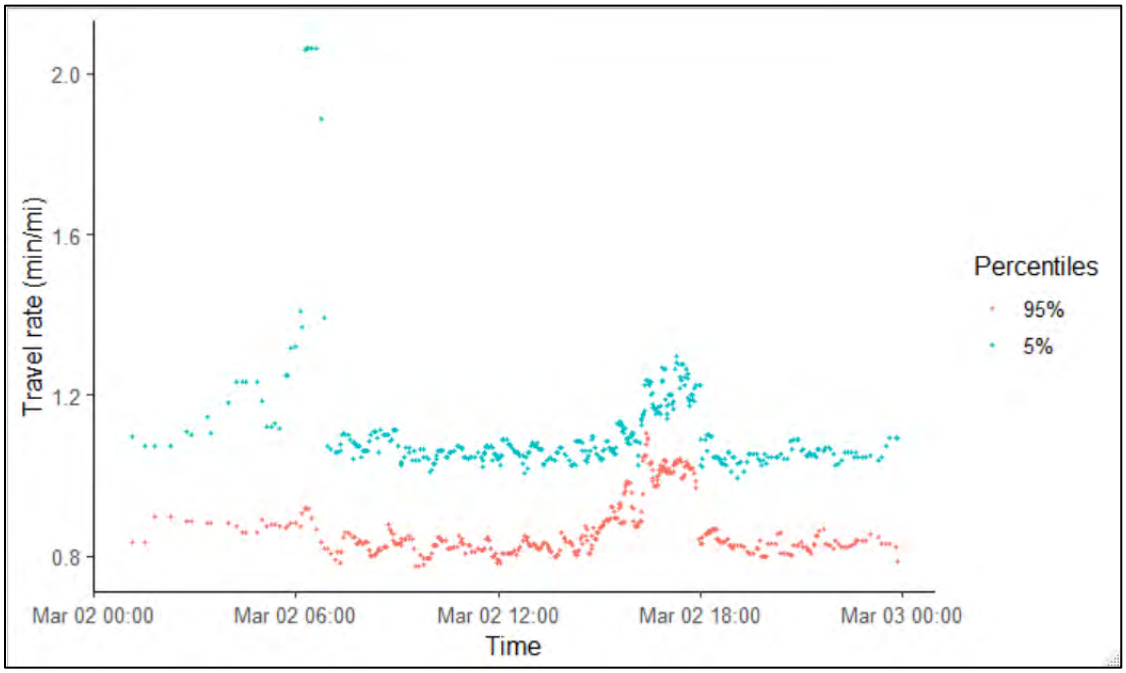


FIGURE A.2: 5% AND 95% TRAVEL RATES FOR EACH GROUP

Method

The first step for estimating time-to-congestion onset is to identify the congestion onset time. Previous studies that used probe data to identify congestion onset used a threshold on the travel rate or speed (Ahmed, Rouphail, et al., 2018). However, this selection of threshold is a policy question, i.e., what speed (or travel rate) the policy makers consider as unsatisfactory to declare a congestion onset? The answer to this question is driven by many factors, such as the surrounding locality, driving behavior, and posted speed limit. A past research (Ahmed et al., 2018) related to bottleneck activation identification on interstate freeways in Raleigh, NC used the ratio of observed speed to the free flow speed as the main criterion and declared congestion onset once it drops below 0.7. In our case, we have groups of 30 travel rate data as our data points. The criteria for congestion onset must be set up in such a way so that it does not generate a significant number of false signals. To this end, we classified a single travel rate observation as congested or uncongested based on a threshold. However, congestion onset is declared once the number of such congested data points in a group exceeds a certain value, and that happens for a certain number of consecutive groups. Thus, it was ensured that a random high travel rate observation does not trigger the alarm.

A vehicle's travel rate is considered as congested if it exceeds 1.2 minutes/mile. This travel rate threshold of 1.2 minutes/mile is equivalent to an average speed of 50 mph. We estimated the average speed for this corridor during off-peak periods as 70 mph. Considering this speed as the

free flow speed, this 1.2 minutes/mile threshold is equivalent to approximately 70% of the free flow speed.

The term congestion index (CI), which is defined in Eq. A-1 determines whether a travel rate is congested ($CI = 1$) or uncongested ($CI = 0$).

$$CI_{i,j} = \begin{cases} 1 \; if \; TR_{i,j} \geq TR^c \\ 0 \; otherwise \end{cases}, \qquad \text{Eq. A-1}$$

where, $CI_{i,j}$= congestion index for travel rate observation *i* of a group *j*. For each group, $i = 1,2, \dots, I$. And, here, *I = 30,* $TR_{i,j}$= travel rate for observation *i,* and TR^c= critical travel rate. As mentioned earlier, we chose $TR^c = 1.2 \; minutes/mi$.

A parameter called the group congestion value (GCV) is defined as the number of congested observations in a group. Mathematically,

$$GCV_j = \sum_{i}^{I} CI_{i,j}. \qquad \text{Eq. A-2}$$

We assert that a congestion onset once at least 5 observations in a group (i.e., about 16% data in a group of 30) is congested and at least three consecutive groups meet this criterion. Two parameters, namely group congestion index (GCI) and congested condition (CC) are defined to express these criteria mathematically.

$$GCI_j = \begin{cases} 1 \; if \; GCV_j \geq N^c \\ 0 \; otherwise \end{cases}. \qquad \text{Eq. A-3}$$

$$CC_j = \begin{cases} 1 \; if \; GCI_j \wedge GCI_{j+1} \wedge GCI_{j+2} \wedge \dots \wedge GCI_{j+J^c-1} \\ 0 \; otherwise \end{cases}. \qquad \text{Eq. A-4}$$

Here, N^c is the threshold for the number of congested observations in a group. J^c is the number of consecutive groups that must meet the N^c threshold to declare the traffic condition as congested. Here, we used N^c = 5 and $J^c = 3$. Congestion onset time is the latest detection time in a group *j* which is congested but its previous group is uncongested (i.e., $CC_j = 1 \wedge CC_{(j-1)} = 0$). Note that all these arbitrary values, such as N^c, J^c, *I*, and TR^c will be subjected to sensitivity tests in Phase II of the project.

The time-to-congestion onset for each group is estimated as the time difference between the next congestion onset time and that group's latest detection time. Each time a new congestion occurs, time-to-congestion onset is reset to zero.

In dealing with adverse weather and incidents, the objectives of this analysis are to detect congestion onset and to classify those by their possible causes. Because the congestion onset

signals and the detection process might be different for demand induced congestion and disruptive incidents, it is important that we divide the tasks based on the possible cause of congestion. As a part of the task of a previous project, the probe vehicle and system detector data from I-5 near Sacramento, CA were assigned tags of incidents and inclement weather condition. Besides, demand induced congestion is easy to distinguish from the disruptive ones occurring during the off-peak periods just by observing the clock time. These tags (i.e., incident, weather, and demand induced congestion) make it easier to divide the tasks by the possible cause of congestion.

However, there are instances when an incident or weather-induced congestion occurs at the time range of demand-induced congestion. These complex cases will be focused on in Phase II of the project. Here, we focused on the simplest case, which is characterizing time-to-congestion onset for demand induced congestion only.

We proposed to detect congestion onset by investigating the characteristics of probe-based travel rate for different values of time-to-congestion onset. These parameters were selected after exploring their patterns on different days. Below is a list of these characteristics that we estimated for each group of travel rates.

As illustrated in section 3.2, the lower and the upper portion of the observed travel rate band, when plotted against their detection time, exhibit different patterns before congestion starts. We attempted to leverage this signal with the help of selected percentile values close to the tails (i.e., 5% and 95% travel rate) and the middle of the travel rate distribution for each group. Figure 3-4 showed that the travel rate band tends to shrink before a demand induced congestion onset. Since similar pattern was observed on multiple days, we also estimated the standard deviation and the difference between 95% and 5% travel rate. All the percentile values were smoothed using the LOESS algorithm as explained earlier.

As shown in Eq. A-2, this parameter is directly related to number of high travel rate observations, and hence, should be considered as an important indicator of congestion onset. It is estimated as the difference in timestamp between the latest and earliest Bluetooth detection within a group. Since it is inversely related to the traffic flow rate, it can be an alternate signal to that.

Individual vehicle travel rates are expected to increase with time within a group right before congestion occurs. This parameter is included to represent the “intra-group” travel rate increase with time. It is estimated as the slope of a regression line fitted to the travel rate data against the relative clock time within a group. The 5% and 95% travel rate of groups are expected to exhibit distinguished patterns; however, these changes take place over a range of period. Hence, these percentiles for a single group may not represent the overall pattern. To better capture these patterns over time, we included the change of the 5% and 95% travel rates over multiple consecutive groups. It is estimated as the slope of the LOESS smoothed line fitted

for the corresponding percentile travel rate of each group. Details about the LOESS fitted algorithm are described earlier.

Testing

We applied the proposed analyses to the probe vehicle travel rate data obtained from I-5 southbound near Sacramento, CA. We started with exploring the pattern of the predictors by clock time and time-to-congestion. First, we chose a typical day with no inclement weather or incident induced congestion. The morning rush hour traffic does not cause any significant travel rate impact on this corridor. During the afternoon peak, starting from around 4:45 PM, this road gets congested recurrently. The smoothed average, 5%, and 95% travel rates for such a day (Jan 27, 2011) are shown against clock time in Figure A.3 The dashed vertical line represents time-to-congestion onset (TTC)= 30 minutes. Our focus is around this value of TTC because congestion onset signals should not set off more than 45 minutes before it gets congested. On the other hand, TTC<10 minutes is too close to the congestion onset for the congestion detection tool to have any practical value. It is apparent that while the 95% and average travel rate remain stable at TTC = 30 minutes, 5% travel rate exhibited an increasing trend even prior to that period. This pattern is magnified in Figure A-4, which shows the difference in 95% and 5% travel rate against clock time. However, it is difficult to use this signal for predicting congestion onset because i) such a decreasing trend of this difference was also found at other periods of the day and ii) the magnitude of the drop was found to vary across different days for the same TTC values.

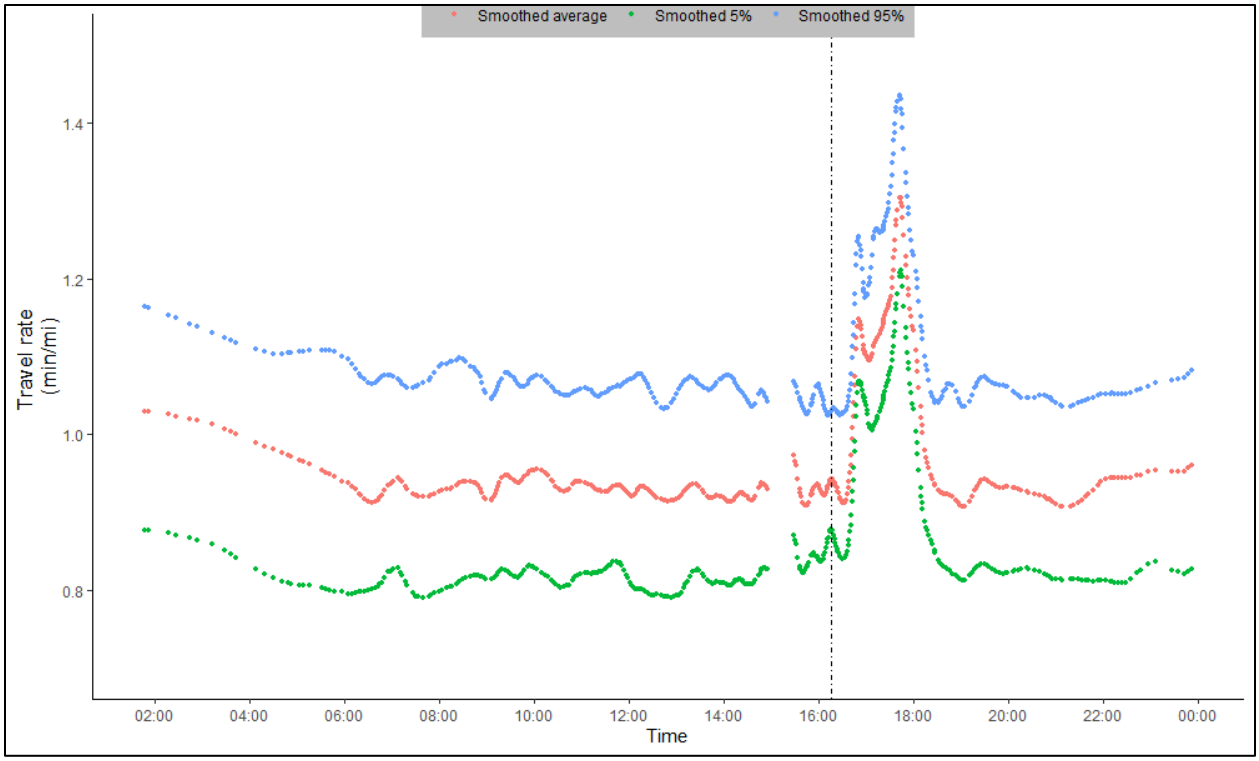


FIGURE A.3: VARIATION OF SMOOTHED 5%, 95%, AND AVERAGE TRAVEL RATE OF EACH GROUP FOR A TYPICAL DAY ON I-5 SB

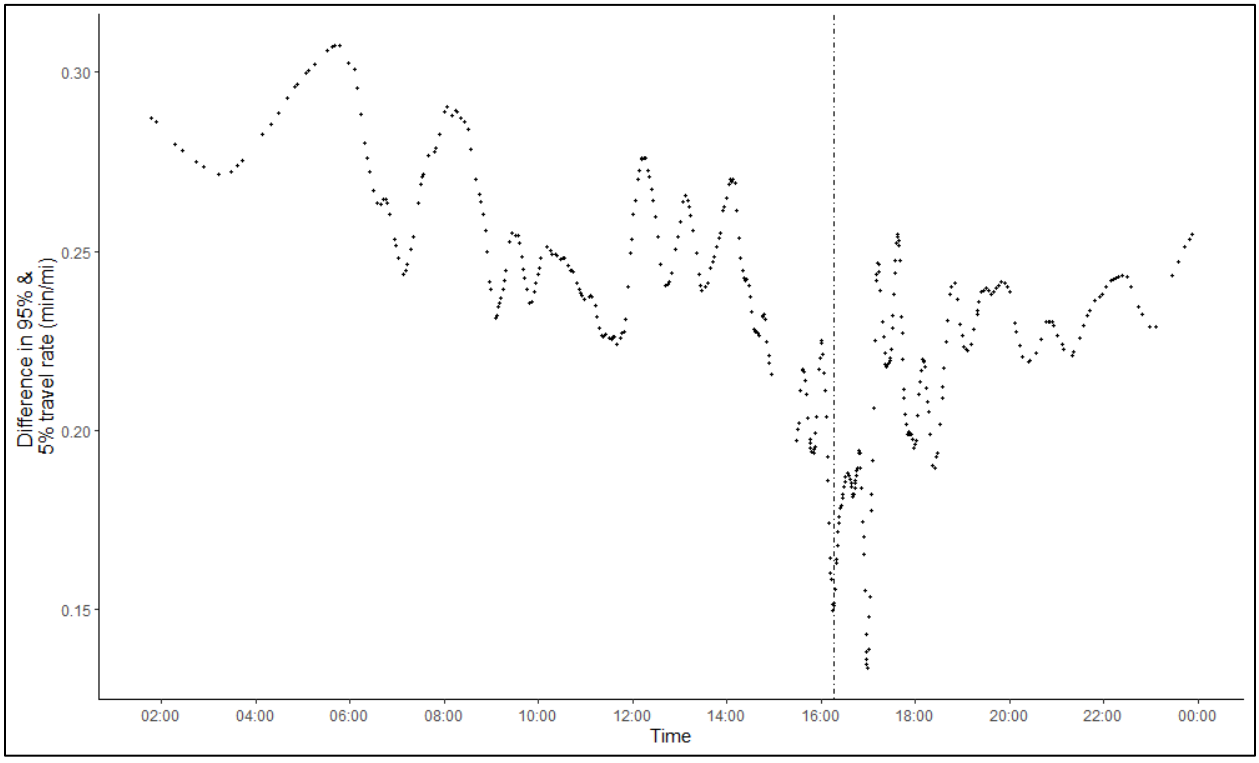


FIGURE A.4: VARIATION OF THE DIFFERENCE IN SMOOTHED 95% AND 5% TRAVEL RATE SHOWN IN FIGURE A.3

Figure A.5 presents another possible indicator of congestion onset—the standard deviation of travel rate for each group. Unlike the percentile curves in Figure A.3, the standard deviation data were not smoothed, and hence the randomness. However, both the raw and the smoothed standard deviation showed exhibited a pattern that is like the difference between the 95% and 5% travel rate.

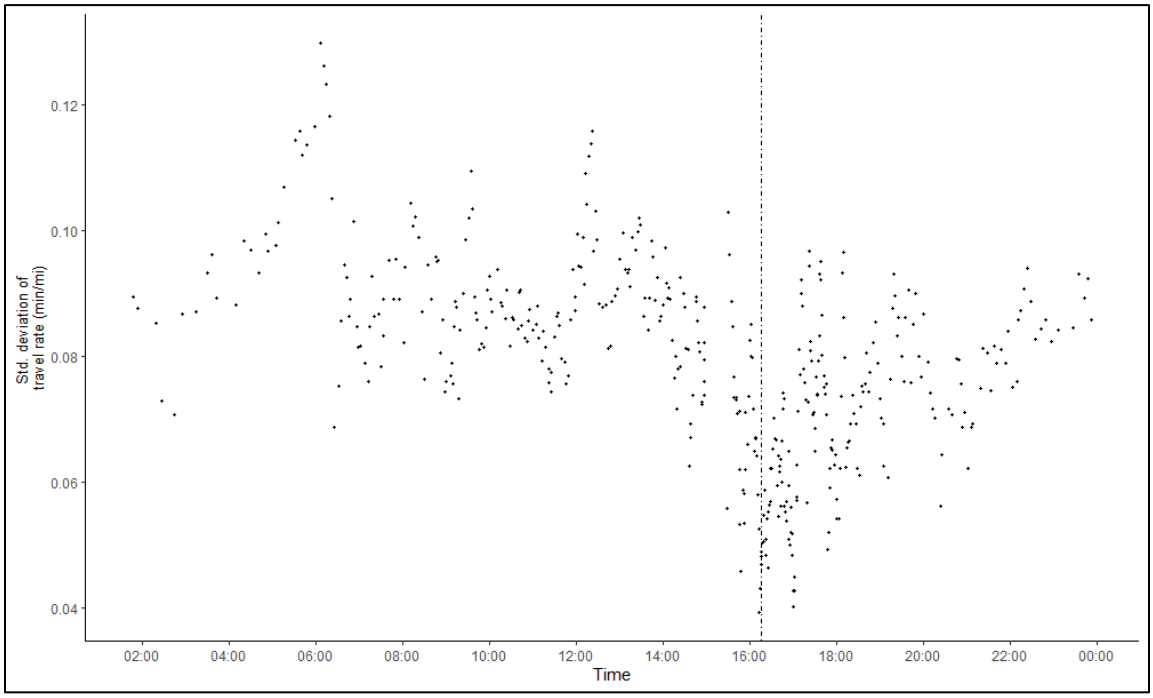


FIGURE A.5: VARIATION OF STANDARD DEVIATION OF TRAVEL RATE FOR THE SAME DATA AS FIGURE A.4

Figure A.6 below shows the GCV against clock time for the same day. Here, we have added two items: a second vertical line (dotted) that represents TTC = 0 and a zoomed view of the period right before the congestion onset time in the inset. These help to visualize the rise of GCV before the congestion onset time. It shows that GCV did not rise abruptly from 0 to 5 before the congestion onset, rather it rose gradually. Although for this day, it started increasing only a few minutes before the congestion onset, it is worth including as a potential predictor to predict congestion onset for the entire dataset. However, the non-zero GCV values during the off-peak periods might cause a false alarm of congestion onset if someone rely entirely on GCV for predicting congestion onset. This observation highlights the importance of testing all potential predictors together along with their interactions for predicting congestion onset.

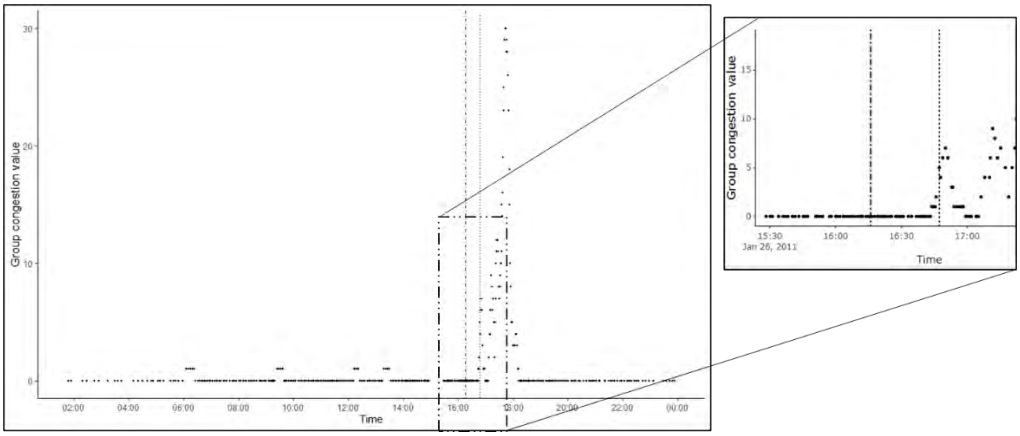


FIGURE A.6: GCV AGAINST TIME FOR THE SAME DATA USED TO GENERATE FIGURE A.4

Figures A.1 through A.6 showed that the 5% travel rate, difference between the 95% and 5% travel rate, standard deviation, and GCV have the potential to generate a combined signal for short-term prediction of congestion onset. So far, we have shown the analyses of these signals for a congested day only. There are weekdays when the study site does not get congested in the PM peak period, although the flow rate was found to reach close to the capacity (observed from the system detector data but not shown here). It is important that we generate similar plots for those uncongested days to demonstrate how these predictors individually may generate false negative signals, but their combined effects may resolve it.

Figure A.7 shows the travel rate vs. clock time plot for such a day, when the increase in the probe vehicle data is apparent in the PM peak period, but the criteria for congestion onset were not met. Figure A.8 shows the smoothed 5%, 95%, and average travel rate for each group for the same day. Figure A.9 shows the difference in 95% and 5% travel rate against clock time for that day.

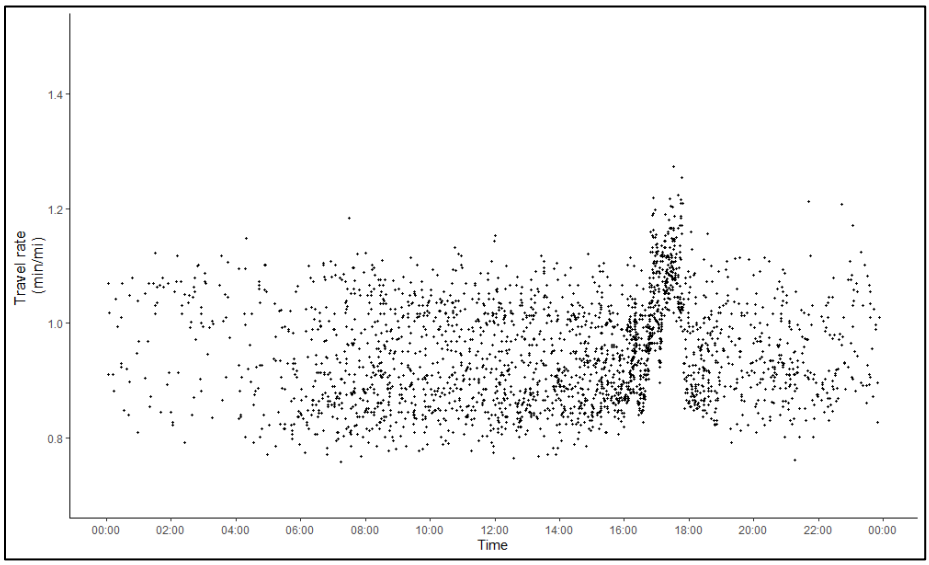


FIGURE A.7: TRAVEL RATE VS. CLOCK TIME FOR A DAY WHEN CONGESTION DID NOT OCCUR

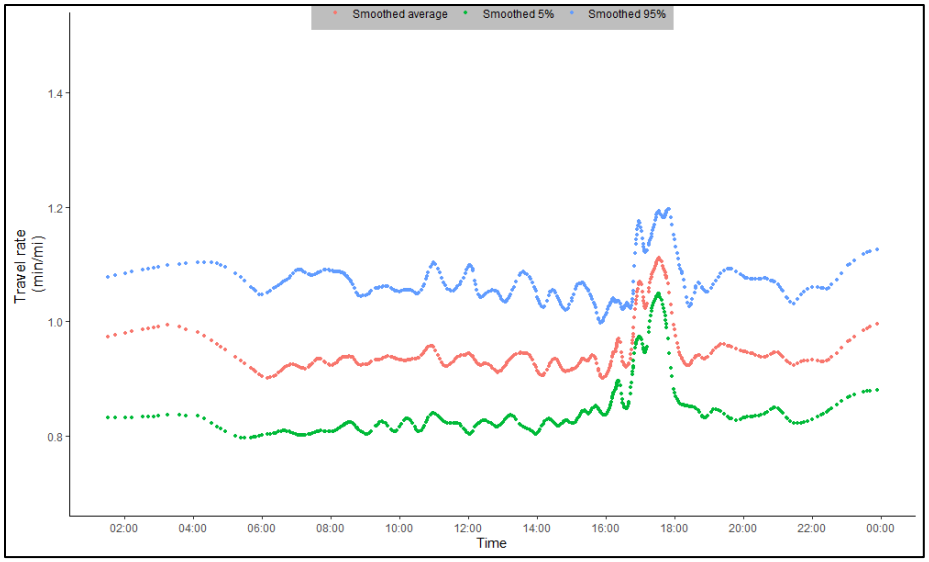


FIGURE A.8: 5%, 95%, AND AVERAGE TRAVEL RATE FOR EACH GROUP FOR THE SAME DAY AS IN FIGURE A.7

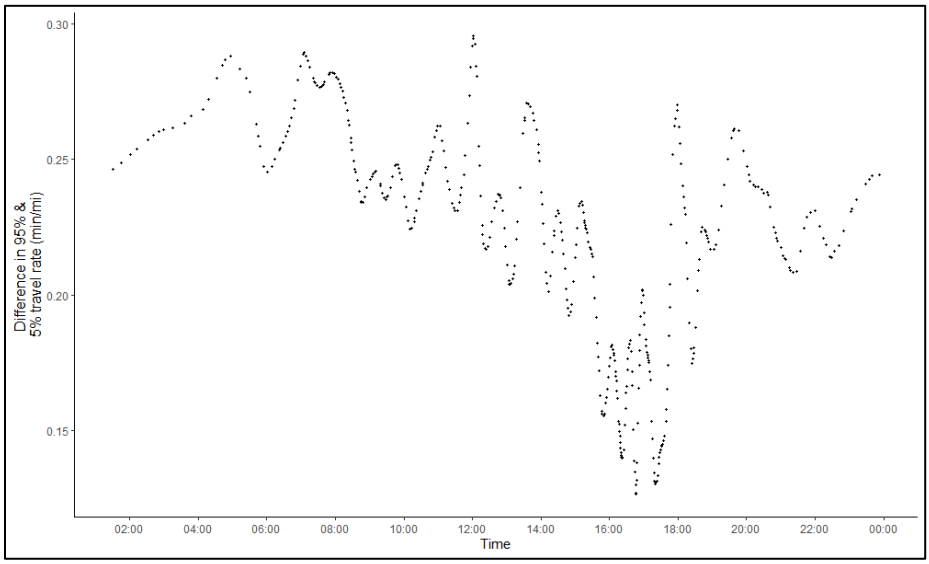


FIGURE A.9: DIFFERENCE BETWEEN 95% AND 5% TRAVEL RATE FOR EACH GROUP FOR THE SAME DAY AS IN FIGURE A.7

The patterns we see in Figure A.8 and Figure A.9 are very similar to the ones in Figure A.3 and Figure A.4, respectively. The 5% travel rate showed an increasing trend before the 95% and average travel rate did. However, the system did not breakdown although the traffic density was high enough to slow down the speed of the faster vehicles. These observations suggest that potential predictors like the 5% travel rate and the difference in 95% and 5% travel rate would generate a false congestion onset alarm for such a day, but the inclusion of predictors like GCV may resolve this issue. Note that we explored similar illustrations of other predictors we mentioned earlier, e.g., the group time span, travel rate change across different groups, and travel rate change per unit time. They did not exhibit any visually recognizable signals prior to

congestion onset. Nonetheless, any hidden signals those predictors might have and their interactions with other predictors should be investigated using machine learning tools, such as decision trees.

Conclusions

Analysis of the variation of these predictors with time and time-to-congestion onset (TTC) showed that the 5% travel rate, difference between the 95% and 5% travel rate, standard deviation of travel rate, and group congestion value exhibited distinguished patterns when TTC is between 10 to 40 minutes. The key strengths and weaknesses of this approach are noted below.

Strengths

- The choice of the predictors was driven by logical reasons established based observations of travel time variation. These reasons are likely to contribute to the generalizability of this approach when applied to different sites.
- The method uses traffic data from one source (probe vehicles). Since no other traffic data sources were used, the need for fusing multiple data sources—which often raises the question of lack of correlation between data from various sources—was avoided. Furthermore, the availability of probe vehicle-based data is increasing everyday with the advancements of technologies related to in-vehicle sensors and vehicle-to-infrastructure (V2I) connectivity.

Weaknesses

- Because the proposed approach is based mostly on the distribution of probe-based travel time data, it is important that there is no sampling bias in the data collection process. It is critical to ensure that the probes were not oversampled from slower moving vehicles.
- The proposed method is based on several thresholds and selected percentiles which, for now, are empirically chosen for demonstrating the analysis results. These arbitrary thresholds and percentiles must be subjected to sensitivity tests so that the selections are robust.

Recommended future tasks for progressing this research include:

- Exploring the joint probability distribution of TTC and other potential predictors. Given that our interest lies within a narrow range of TTC values, TTC may need to be transformed from a continuous to an ordinal or categorical variable.
- Applying a machine learning algorithm to model time-to-congestion onset using the predictors described above. Expectedly, such a model would help reducing the false positive and false negative outcomes.
- Investigating the sensitivity of the model performance on the arbitrary thresholds.

References

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association, 74(368), 829–836.

Ahmed, I., Rouphail, N. M., & Tanvir, S. (2018). Characteristics and temporal stability of recurring bottlenecks. Transportation Research Record, 2672(42), 235–246.

Ahmed, I., Williams, B. M., & Samandar, M. S. (2018). Application of a Discontinuous Form of Macroscopic Gazis–Herman–Rothery Model to Steady-State Freeway Traffic Stream Observations. Transportation Research Record, 2672(20), 51–62.

APPENDIX B. Fitting Parametric Distributions to Travel Rate Data

As explained in section 3.2, when traffic state shifts from uncongested to congested state, the distribution of travel rate for a group of vehicles at that time should change. The previous section attempted to predict time-to-congestion-onset by several predictors that are mostly the descriptive statistics of that distribution. Two critical limitations of this method are: a) the selection of the descriptive statistics measures, i.e., the 5^{th} and 95^{th} percentile travel rate for each group of probe vehicle observations is somewhat arbitrary; b) these empirical percentiles are estimated from a group of a very few observations. Consequently, these extreme percentile estimates become very sensitive to any minor random change in the distribution. One viable way to avoid this downside is to fit a parametric distribution to each group of probe vehicle travel rate, and then investigate how the fitted parameter values change. Thus, we avoid having to select an arbitrary percentile value.

Several past studies attempted to fit parametric distributions to travel time data obtained from freeway corridors (Taylor, 2017; Samandar et al., 2017). Among these, Samandar et al. (2017) showed that a bi-modal Burr distribution fits the travel time data with a satisfactory goodness of fit. This bimodality in the travel time distribution is intuitive for a typical scenario where the traffic stream consists of faster moving passenger cars and slower moving heavy vehicles. We leveraged this very logic to predict congestion onset using probe vehicle travel rate data.

This idea of fitting a parametric distribution to travel rate data for predicting congestion onset was implemented in two ways. The first approach was to investigate the change in the fitted parameter values for successive groups of probe vehicle observation. However, it is difficult to fit a mixture of distributions to a small sample of observations. For instance, a bi-modal Burr distribution has nine parameters, which would be difficult to fit to 30 data points. Hence, we did not present further details of this approach in this section. However, this approach may work very well in a scenario where all probe vehicles travel time data are available to us (such as, from a simulation model).

The second approach deals with clustering the probe vehicle observations in such a way that each cluster represents a traffic state, and then fit a Burr distribution to each cluster. This clusters can be created from all days of probe vehicle data available to us. Thus, it would avoid the sample size issue to a significant extent. However, the major challenge lies in this process is the clustering process. To this end, we incorporated a secondary data source—the system detector data collected within the boundary of the study area. The underlying assumption was that the system detector(s) data are representative of the probe vehicle travel time data. Details of this process are described below.

Description of the System Detector Data

Infrastructure-based detectors embedded in the study site included dual loop and radar detectors that are operated and managed by the California Department of Transport (Caltrans). These point detectors embedded in each lane report five-minute aggregated lane flow counts and the harmonic mean speed of all vehicles. We collected these data for all the periods for which we have probe vehicle travel time records. There were ten functional system detectors on the northbound direction of the study site five on the southbound

Methodology

The first step of this approach is to select a system detector that is representative of the traffic state of the study corridor. Later, we will show that for this site, data from one system detector cannot properly reflect it. To this end, data from multiple locations must be fused. However, for the purpose of demonstrating the proposed method, we discuss the framework that we developed and demonstrate the experimentation results using one system detector data.

Upon exploring traffic flow, speed, and density data from system detectors located at different locations on I-5 southbound, it was found that the most common cause of congestion is associated with the on-ramp from I-80 at mile-marker 519. Hence, we choose the detector just upstream of this location for this analysis. The most upstream probe sensor is also located close to this point. Traffic flow, speed, and density data aggregated in 5-minute intervals for each lane of this location were extracted.

It is important to choose the appropriate variables so that various traffic regimes and their transition from one another can be properly identified. For distinguishing various traffic regimes, average traffic speed and or its combination with flow and density are commonly used. In addition, their lane-by-lane variation is another interesting aspect to consider. Typically, traffic on the leftmost lane tends to be faster than that on the rightmost lane in uncongested conditions. Furthermore, to distinguish the transition of traffic regimes, it is also important that we investigate the change of these characteristics with time. For instance, Figure B.1 shows the average speed vs. flow rate data for a typical weekday. The data were obtained from a single lane of the selected system detector. The second from the rightmost lane of the system detector location was selected. The colors represent pre-congested (TTC between 10 to 45 minutes), congested, and normal operating conditions. Note that to incorporate the TTC values to the system detector data, we first fused it with the probe vehicle data based on clock time. The data points are connected by arrows which shows the chronological progression between two successive observations.

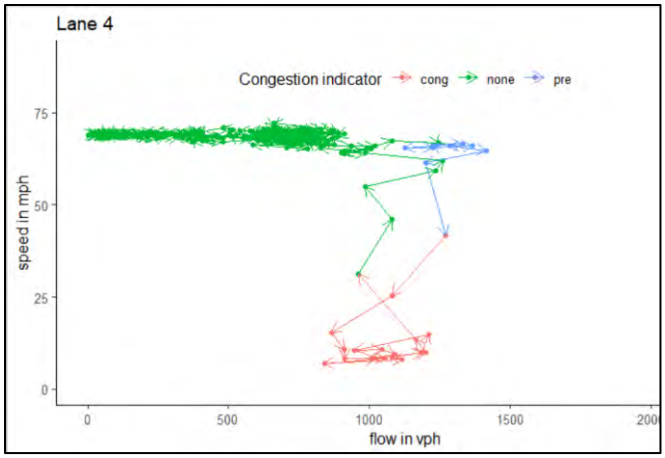


FIGURE B.1: AVERAGE SPEED VS. FLOW DIAGRAM FOR A LANE AT THE SELECTED SYSTEM DETECTOR LOCATION

It is apparent from the arrows that the transition from pre-congested to congested state and back to the normal state follows a pattern. This pattern can be quantified using the time-series variation of speed and flow data. Presumably, the temporal variation of speed and flow was highest at the transition points.

Based on these abovementioned reasonings, we came up with the following set of variables to incorporate in our cluster analysis.

- Average speed across all lanes
- Total flow rate across all lanes
- Lane-based variation of speed: It is estimated as the standard deviation of average speed of different lanes at a time-step.
- Lane-based variation of flow rate: It is estimated as the standard deviation of flow rate of different lanes at a time-step.
- Temporal variation of speed: It is estimated as the standard deviation of three observations—average speed of all lanes of the current, previous, and the future time-step.
- Temporal variation of flow rate: It is estimated as the standard deviation of three observations—total flow rate of all lanes of the current, previous, and the future time-step.

Among various clustering algorithms, we tested two popular ones: K-means clustering and model-based clustering. K-means is the most popular partitioning clustering algorithm. In it, each cluster is represented by the center or mean of the data points belonging to the cluster. The clusters are created by minimizing the total intra-cluster variation. Despite being popular, a few drawbacks of the K-means clustering method are that the number of clusters must be specified, the initial starting point is random, and the process is heuristic, i.e., it is not based on a formal model. Hence, another algorithm that avoids these drawbacks called model-based

clustering was also tested. This method assumes that the data has a certain parametric distribution (e.g., Gaussian) and create clusters by fitting multi-modal parametric distribution to the data. Of course, the main limitation of this method is that traffic data (e.g., flow, speed, and density) seldom follow a particular distribution. Therefore, only the K-means cluster's application is described here.

Upon generating the clusters from the system detector data, the next task is to assign these cluster numbers to the probe vehicle travel rate based on their timestamps. Similar to the previous analyses, the downstream detection time of each probe vehicle observation was used to fuse the cluster numbers.

The next task is to fit a parametric distribution to each regime's travel rate data. The idea is that whenever the system will receive new travel rate data, the tool will test the hypothesis whether or not these data belong to the distribution of the current traffic regime. Our assumption is that if the new data feed does not pass this hypothesis test, this may indicate that the traffic state is transitioning.

To this end, we chose the Burr distribution not only because our data exploration suggested that it may provide the best fit but also because previous studies successfully fitted this distribution to probe-based travel rate data. The equation for the cumulative distribution function (CDF) of a single-mode Burr distribution is:

$$\mathrm{F}(p) = 1 - \left(1 + \left(\frac{x - \gamma}{\beta}\right)^{\alpha}\right)^{-\kappa}, \qquad \text{Eq. B-1}$$

where,
$\mathrm{F}(p) =$ Cumulative probability function $(0 \leq F(p) \leq 1)$;
$x =$ Observed travel rate;
k is a continuous shape parameter $(k > 0)$;
α is a continuous shape parameter $(\alpha > 0)$;
β is a continuous scale parameter $(\beta > 0)$;
δ is a continuous location parameter $(\delta \geq 0)$;
The equation for a two-mode Burr distribution is:

$$\mathrm{F}(p) = \mathrm{F}_1(\mathrm{p}) * \mathrm{w} + \mathrm{F}_2(\mathrm{p}) * (1 - \mathrm{w}), \qquad \text{Eq. B-2}$$

where, w is the mixing proportion $(0 \leq w \leq 1)$.

Three or higher mode-Burr distributions' equation follow the same progression. Note that the number of parameters to be fitted increases by five for each increment of the number of modes.

The choice of number of modes to be fitted is dictated by the goodness of fit and sample size. We propose to test both single, two, and three-mode Burr distribution to each cluster of travel rate. Between the two common methods of fitting distributions, namely minimizing the sum of squared error of cumulative distribution function (CDF) and maximum likelihood method, we picked the first one since deriving the likelihood function for multi-modal Burr distribution associates many complexities. Bayesian information criteria (BIC) will be used to evaluate each fit, which takes both the error and the degree of freedom into account.

Testing

Before applying the clustering algorithm, system detector data associated with incidents were removed. Furthermore, to remove unsteady observations, a filter based on the temporal variation of speed, called the speed first difference threshold was applied. Details of this filter is described elsewhere (Xu et al., 2013;Ahmed et al., 2018)

The K-means clustering algorithm was applied to the system detector data associated with the variables mentioned above for all 35 days of observations from the southbound corridor of our I-5 study site. The variation of the total within sum of squared distance was used to make an educated guess about the number of clusters. Figure B.2 shows the variation of this distance for different numbers of clusters. It is apparent that this distance dropped rapidly from one to five clusters, but then the slope of the curve became mild. As this plot suggests, we chose six clusters.

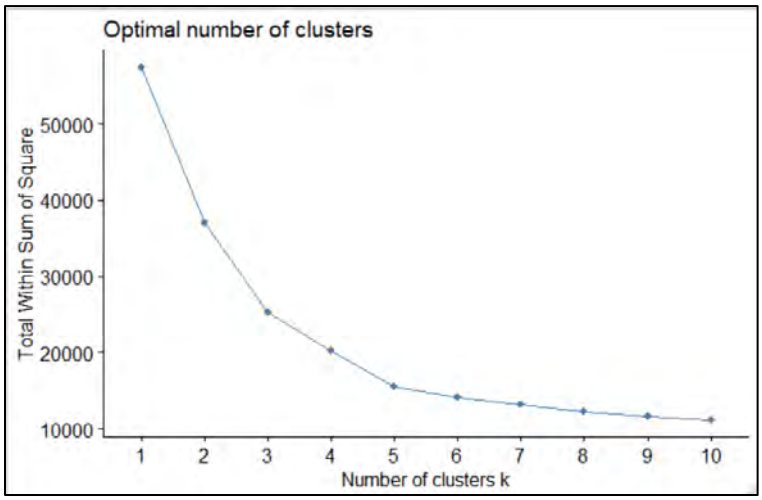


FIGURE B.2: SELECTION OF OPTIMAL NUMBER OF CLUSTERS BASED ON TOTAL WITHIN SUM OF SQUARED DISTANCE

Figure B.3 shows the outputs of the K-means clusters through a speed-flow plot.

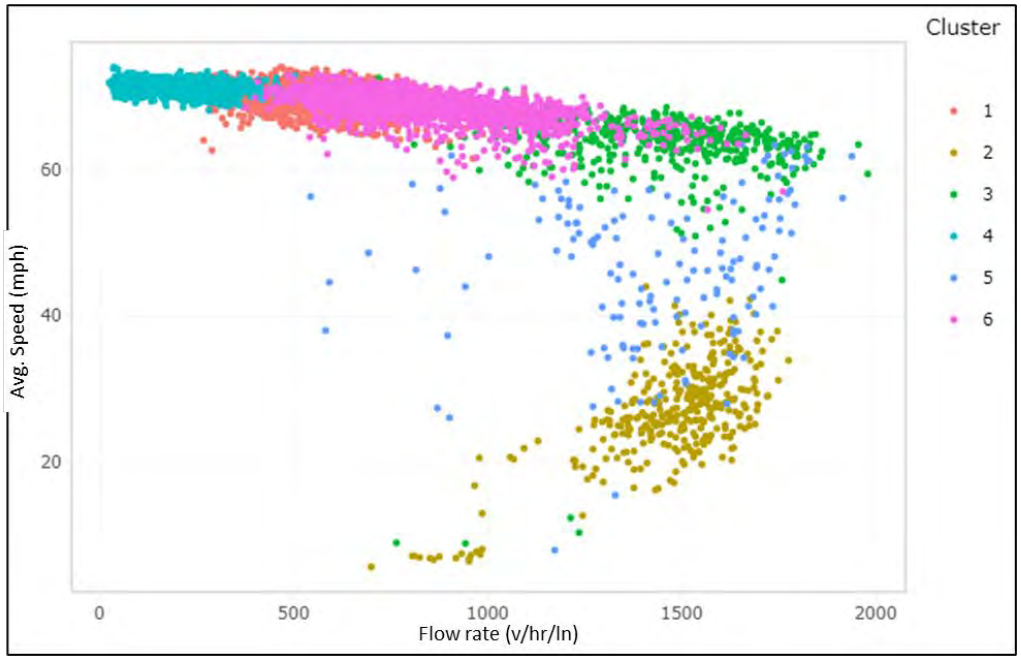


FIGURE B.3: CLUSTERING OUTPUT IN THE SPEED-FLOW PLOT

Key observations from this figure are listed below:

- Free flow regime is divided into three clusters: 1,4,6. These clusters were possibly generated due to the random variation of speed and flow and their temporal and lane-based variation in free flow conditions. Because traffic state does not shift directly from the free flow regime to the congested regime in the case of demand induced congestion, these clusters are not focused in detail in this study.
- Referring to Highway Capacity Manual's (Transportation Research Board, 2016) macroscopic model for basic freeway segment, the regime between breakpoint flow rate and maximum observed flow rate is mostly comprised of cluster 3.
- Cluster 5 constitutes the transition regime between uncongested and congested condition. This regime is not clearly defined because many of its observations appear to represent an unsteady condition.
- The congested regime is mainly comprised of cluster 2.
- The focus of this study is on clusters 2, 3, and 5. There are overlaps in terms of average speed between clusters 2 and 5 and clusters 3 and 5.
- Overall, the clustering outputs seem reasonable since it appears to successfully distinguish the traffic regimes, with some redundant clusters (from this study's perspective) in the free flow regime.

The cluster numbers associated with the 5-minute aggregated system detector data were fused with the probe vehicle travel rate data based on the detection time at the downstream sensor for each detected vehicle. Upon fusing, the first question that needs to be answered is: *does the traffic operating condition (e.g., average speed) estimated from the system detector data correspond to the probe vehicle travel rate data for each cluster?* Ideally, if a system detector could represent the operating condition of a 5.58 miles long corridor, one would expect that the average speed estimated from the system detector is negatively correlated with the probe vehicle travel rate. However, that expectation is not realistic because a) the random variation in individual vehicles' speed is substantially dampened in the aggregated system detector data b) system detectors measure speed at a point, and hence, the space mean speed is an estimate from the reported time mean speed c) traffic operating conditions may vary spatially across the corridor. To check how representative the system detector data are for the Probe vehicle travel rate, the correlation coefficient between average speed from the system detector and travel rate from the Bluetooth sensors was estimated for each cluster, as shown in Table B.1.

TABLE B.1: CORRELATION BETWEEN AVG. SPEED FROM THE SYSTEM DETECTOR AND TRAVEL TIME DATA FROM PROBE VEHICLE SENSORS FOR EACH CLUSTER

Cluster number	Correlation between avg. speed (from system detectors) and probe vehicle travel rate
1	-0.12
2	-0.76
3	-0.49
4	-0.05
5	-0.32
6	-0.61

It is apparent that the correlation is negligible for clusters belonging to the free flow regime (clusters 1 and 4). Further investigation of the free flow regime data showed that the random variation of individual vehicles' travel rate cannot be captured by the aggregated system detector data at a point. Among the three clusters in which our interest lies, i.e., clusters 2, 3, and 5, cluster 2 exhibited a satisfactory correlation coefficient. It implies that during the congested conditions, probe vehicle travel rate and average speed estimated from the system detector are highly correlated. The correlation coefficients for clusters 3 and 5 are weak. This is probably because traffic state did not remain the same throughout the entire study corridor at near capacity and transition conditions.

Despite these weak correlation coefficients, the relative speed among clusters 2, 3, and 5 observed in the system detector data is reflected in the probe vehicle travel rate data. Figure B.4 shows the cumulative distribution curves of the probe vehicle data associated with these three clusters.

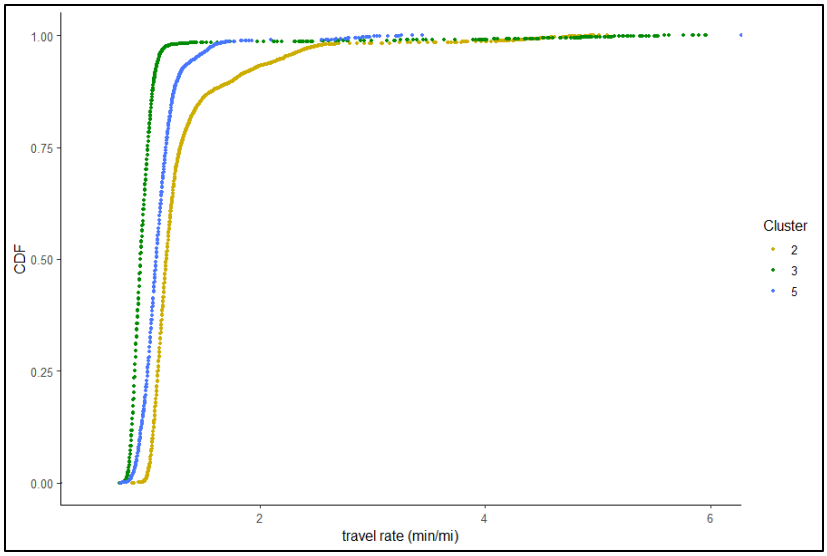


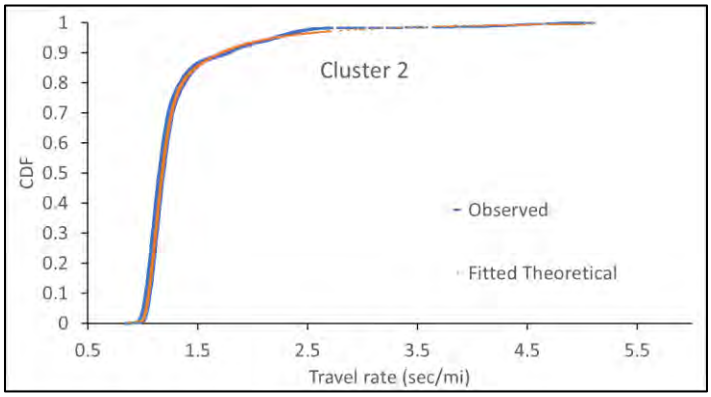
FIGURE B.4: CUMULATIVE DISTRIBUTIONS OF TRAVEL RATE FOR THREE CLUSTERS

Here, only these three clusters are shown to avoid cluttering the plot. Also, traffic regimes associated with these three clusters are of most important for the purpose of this research.

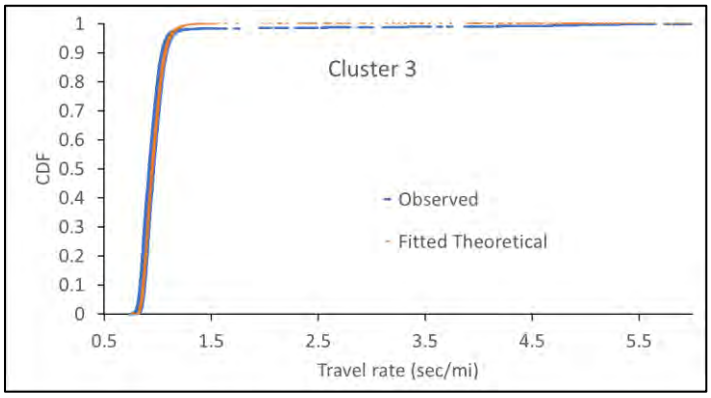
Expectedly, travel rate distribution for cluster 2, which in the system detector data constituted the congested condition, is on the rightmost side. Only at the rightmost end of the distribution, cluster 3 exceeds the other two in terms of travel rate. Such a long tail of cluster 3 poses a critical problem because the clock times associated with this cluster usually do not face congested conditions. Cluster 5, as in Figure B.3, is in between the other two. Expectedly, both Kolmogorov-Smirnov and Anderson-Darling tests showed that at a significance level of 0.01, these three distributions belong to different populations. For this research, this finding was important given that the assumption was that the clusters can separate the probe vehicle data into different traffic regimes.

Despite the long tail of travel rate distribution for cluster 3, which potentially affects the proposed plan of signaling traffic transition based on these distributions, we moved forward with the proposed analyses and fitted Burr distributions to each cluster of travel rate data.

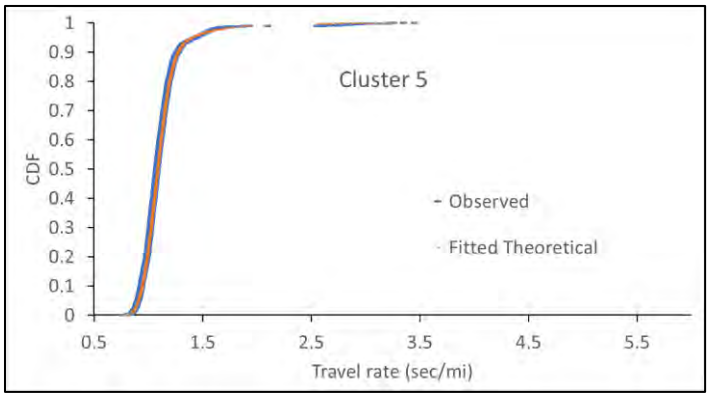
Although it is not clear from the CDF for cluster 3, the dips in the CDFs for clusters 2 and 5 in Figure B.3 indicate multi-modality of these distributions. Here, we fitted both a two-mode and a three-mode Burr to each cluster's travel rate data. For all six clusters, the two-mode Burr generated the lowest BIC. Figure B.5 shows the fitted two-mode Burr for clusters 2, 3, and 5. The relative root mean squared error of these fits (i.e., root mean squared error of CDF divided by the mean observed CDF) was found to be between 0.65% to 0.92%.



(a)



(b)



(c)

FIGURE B.5: CUMULATIVE DISTRIBUTIONS FOR OBSERVED DATA AND FITTED TWO-MODE BURR DISTRIBUTION FOR CLUSTERS (A) 2 (B) 3 (C) 5

It is apparent from Figure B.5 that for cluster 5 and most part of cluster 2, the fitted and observed CDF curves are visually indistinguishable. Near the tail of the CDF for cluster 2, the fitted values deviate from the observed ones. On the contrary, the fitted tail for cluster 3 is off from the observed curve. Clearly, the fitted Burr distribution could not properly capture the long tail of the distribution. Note that we conducted the Kolmogorov-Smirnov test, which

showed that the fitted and observed CDFs are from the same population for these three clusters.

Conclusions

In this section, we proposed to use a parametric distribution that was fitted to classified probe vehicle travel rate data to signal congestion onset. probe vehicle travel rate data were classified into different traffic regimes by clustering system detector data. The clustering outcomes for the system detector data were intuitive. However, some extremely high travel rate values in the probe vehicle data, which should belong to the transition regime, caused discrepancies.

Below are the strengths and weaknesses of this research.

Strengths:

- The approach is based on a parametric (Burr) distribution, which past studies successfully used to fit travel time data. This distribution also demonstrated a satisfactory fit when applied to the data used in this research.
- The process is independent of forming groups of travel rate with a limited sample of observations. Thus, it reduces the chance of overfitting the Burr distribution.
- Unlike the method described in section 3.2, this method does not need to define any arbitrary number, e.g., congestion onset threshold, group size, overlap percentage, and certain percentiles of travel rate distributions. The only decision parameter here is the number of clusters of the K-mean clustering algorithm, for which an educated guess can easily be made.

Weaknesses:

- The accuracy of the method depends on how well the clustering algorithm can distinguish different regimes, and how well the Burr distribution is fitted to the travel rate data.
- The system detector data from one (or multiple) sensor(s) may not represent the corridor-wide traffic state depending on the spatial variation of traffic condition.

Recommended future tasks for progressing this research include:

- Ensuring that the unexpected high travel rate in the transition regime data was not caused by any unreported incident.
- Data from a single system detector may not represent the operating condition of a long corridor. Therefore, multiple detectors' data may need to be fused to make the speed data be representative of the traffic state of the entire study corridor.

References

Ahmed, I., Williams, B. M., & Samandar, M. S. (2018). Application of a Discontinuous Form of Macroscopic Gazis–Herman–Rothery Model to Steady-State Freeway Traffic Stream Observations. Transportation Research Record, 2672(20), 51–62.

Samandar, M Shoaib, Williams, B. M., & Ahmed, I. (2018). Weigh Station Impact on Truck Travel Time Reliability: Results and Findings from a Field Study and a Simulation Experiment. Transportation Research Record, 2672(9), 120–129.

Taylor, M. A. P. (2017). Fosgerau's travel time reliability ratio and the Burr distribution. Transportation Research Part B: Methodological, 97, 50–63.

Xu, Y., Williams, B. M., Rouphail, N. M., & Chase, R. T. (2013). Development of an Oversaturated Speed–Flow Model Based on the Highway Capacity Manual. Transportation Research Record, 2395(1), 41–48.

APPENDIX C. Fourier Series Analysis of the Travel Rates

In sections 3.1, 3.2, and APPENDIX A, we discussed the ideas relating to predicting congestion onset time over a short interval and characterizing the cause of it. In addition to this idea, we also attempted to characterize the whole pattern of individual vehicles' travel rate over time. To this end, we tried a classical mathematical transformation method, called the Fourier series, to model travel rate against time. Fourier series has been widely used for decomposing various waveforms into a linear combination of sine and cosine functions. Despite being widely used in numerous scientific fields, to the authors' knowledge, it has not been tested previously for modeling travel rate variation with time.

Methodology

The Fourier series is a way of expressing a periodic function as a combination of sine and cosine functions. Eq. C-1 shows a generic form of the function for $f(z)$.

$$f(z) = \alpha_0 + \alpha_1 z + \alpha_2 z^2 + \sum_{q=1}^{Q} \left(\lambda_q \sin(qz) + \delta_q \cos(qz)\right), \qquad \text{Eq. C-1}$$

where $z =$ clock time (seconds from a reference time); $\alpha_0, \alpha_1, \alpha_2$= linear coefficients; $q = 1, 2, 3, \ldots, Q$; represents the number of pairs of sine and cosine function to integrate. Coefficients $\alpha_0, \alpha_1, \alpha_2, \lambda_q, \delta_q$, and Q are the parameters to be fitted by minimizing the Akaike Information Criteria (AIC). Therefore, the number of parameters to be fitted is $4 + 2Q$. $f(z)$= travel rate (seconds/mi).

Of the diurnal variation of travel rate data, this section is focused on the part during the recurrently congested period, i.e., building up of the congestion before either the morning or the afternoon or both rush hours, peaking, and then the recession of travel rate. Travel rate variations during the uncongested periods is random and not of much interest. Sections 3.2 and APPENDIX 1 dealt with the short-term prediction methods for congestion onset. Hence, the analyses conducted in those sections covered all periods in a day. Unlike those analyses, the Fourier series application that is proposed to apply here is for expressing the travel rate variation, not for short-term prediction. Hence, in this section, we focused on the most critical part (i.e., periods surrounding a congested condition) of travel rate variation.

Experimentation

Figure C.1 shows an example of fitting the Fourier equation to travel rate data from 6:00 AM to 9:00 AM on Feb 23, 2011 from I-5 northbound. On the portion of I-5 northbound corridor that the Bluetooth sensors covered, the morning rush-hour traffic typically causes a severe congestion starting at around 7:30 AM, and hence the selection of the time period.

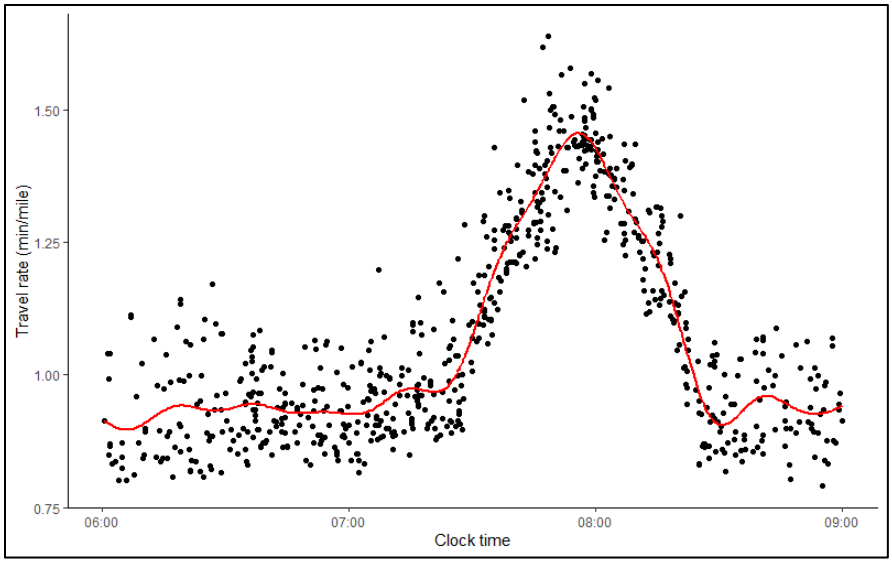


FIGURE C.1: FOURIER EQUATION FITTED TO TRAVEL RATE OBSERVATIONS AGAINST TIME

The performance measures of the fit were satisfactory—the relative RMSE was 6.9% and adjusted R-squared was 0.86. However, the AIC value did not reach a minimum till $Q = 9$. Such a high Q value means that a large integral was fitted. In fact, the total number of fitted parameters was $4 + (2 * 9) = 22$. Further investigation showed that such a high number of fitted parameters was attributed to the random variation of travel rate during the uncongested periods, i.e., from 6:00 AM to around 7:00 AM and from 8:45 AM to 9:00 AM. One way to bypass such an overfitting is by selecting a narrower region of interest and only account for the goodness of fit for that region. For instance, one may choose the bend of the upward travel rate curve (i.e., from 7:15 AM and 7:30 AM in Figure C.1) and estimate the residual sum of squared error for different values of Q ranging from 2 to 9. Figure C.2 shows the resulting output.

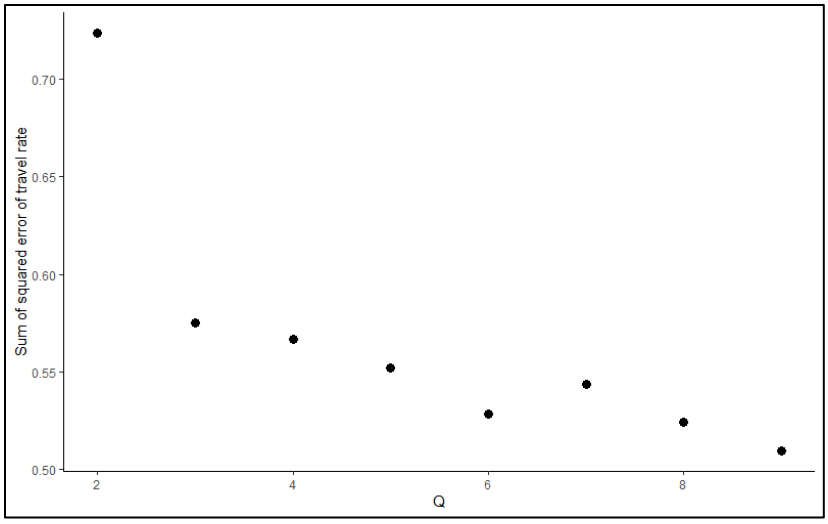
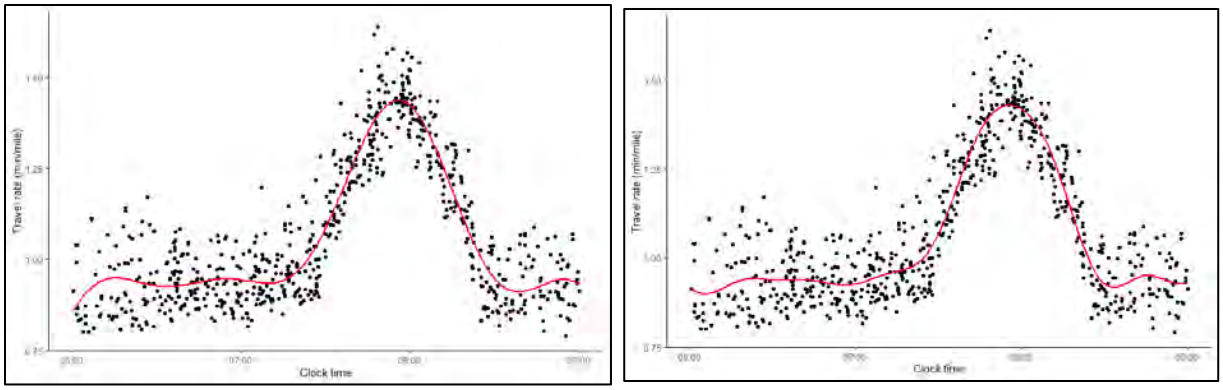


FIGURE C.2: VARIATION OF SUM OF SQUARED ERROR OF THE FITTED FOURIER SERIES WITH THE NUMBER OF INTEGRALS

In Figure C.2, the sum of squared error for this particular region of travel rate drops abruptly between Q=2 and Q=3, but then it becomes very gradual for the larger values of Q. Besides, the resulting error is very close for Q=6 and Q=8, implying that for having a reasonable fit for this chosen region of the travel rate curve, Q=6 is a reasonable selection. Figure C.3 shows the fitted curve for Q=3 and Q=6.



(a) (b)

FIGURE C.3: FOURIER EQUATION FITTED TO TRAVEL RATE DATA WITH (A) Q =3 (B) Q = 6

Figures C.3 (a) and (b) show that the fitted curve remains almost the same from right before the congestion onset till it ends for these two values of Q. Only the pattern of the curve for the uncongested portions changes. The relative RMSE and R-squared values for Q=3 were 7.2% and 0.85, which are very similar to those for Q=9. It implies that even with a few numbers of fitted parameters, a reasonable fit of the Fourier equation to the critical part of the travel rate curve, i.e., from prior to congestion onset till the congestion ends, is achievable.

Conclusions

Here, we demonstrated a successful application of Fourier series to model probe vehicle-based travel rate variation with time, with a focus on the congestion onset time for a day. It exhibited a satisfactory goodness of fit even with a very few parameters. Although we did not apply this series for the purpose of short-term prediction of congestion onset, it can be used to characterize the day-to-day variation of the congestion onset, uphill, peak, and/or the downhill part of the travel rate curve.

Below are the strengths and limitations of this approach.

Strengths

- Fourier series has a strong theoretical background, with numerous successful applications to time-series data from different fields in the past.
- The fitted Fourier series can be succinct if the focus is on a narrow part of the travel rate curve against time.

Weaknesses

- The proposed method does not have the capability to predict congestion onset

Because the demand induced congestion onset time varies day to day, its value across multiple days needs to be adjusted for comparing the change in the fitted series on different days. This can be done by using a relative scale of time for all days, where the congestion onset time will be set to zero. Further investigation along this path of idea is recommended for future research.

APPENDIX D. Standard Deviation for Detecting Disruptive Incidents

Incidents on the road network tend to occur suddenly and without any significant indication. While past studies investigated the potential of several traffic characteristics as the precursor of roadway accidents, that topic is out of the scope of this research. In this study, the aim is to propose tools that detect incidents very early once they occur. This would help agencies to take necessary actions very quickly, e.g., deploying emergency medical services and/or diverting traffic through variable message signs or other techniques.

As indicated in Section 3.1, our data base contains incident and weather-related data that were collected from secondary sources and fused with the probe-based data by matching their timestamps at the downstream detector. This resulted in giving each probe vehicle travel rate data point a color based on the ambient weather condition and the presence of incident on the travel time recording section when that probe vehicle travelled the section, as shown in Figures 3-4 and 3-5. However, it is known that incident reports are susceptible to delays and inaccuracies, and that poses a challenge to the task of real time detection of incidents based on incident reports and police records. Consequently, rather than relying on incident records, this chapter explores analytical methods through which we can utilize probe vehicle travel rates to identify the occurrence of incidents that cause disruption to the traffic flow in real time. Two methods are explored: 1) The sudden spike in standard deviation of travel rate groups 2) The difference between the 95th percentile and the 5th percentile travel rates within travel rate groups.

This chapter will explain the concept and provide application examples of the two methods mentioned above for disruptive incident detection. The limitations of these two methods will also be addressed towards the end of the chapter.

Method and Application

One of the tools that we have explored so far which may have the potential to enable agencies to confirm the occurrence of incidents that alter the state of traffic flow is the sudden spike in the standard deviation of the probe vehicle groups (the grouping technique was defined in Section 3.1). This sudden spike can be used to confirm that what we are observing represents a sudden slow-down in traffic that is a result of a disruptive incident, which is dissimilar to the gradual change in the traffic state that characterizes the onset of recurring peak conditions. However, this sudden traffic slow-down occurs when the traffic conditions prior to the disruptive incident occurrence are non-congested, not when the condition of traffic is already congested due to performance degradation associated with a recurrent peak or other disruptive event.

Figure D.1 below shows a plot of individual vehicles' travel rates against time for one day along I-5 NB, on which the peak flow occurs in the AM period. Each point represents an individual

vehicle's travel time, plotted based on the time it crossed the downstream travel time detector. The points are color-coded based on the ambient weather conditions and the occurrence of any incident along the NB travel time recording section. Furthermore, some observations may also have been recorded during a period which featured both incident occurrence and inclement weather event along the travel time recording section. These are given a different color, as is the case with some observations in Figure D.2 which are given the code "IncidentGust",Looking at the figure, the effect of the demand induced congestion can be observed by the spike in the travel rates at around 07:30 AM, whereas the sudden and severe spike in travel rates at around 03:00 PM in the afternoon is attributed to a disruptive incident.

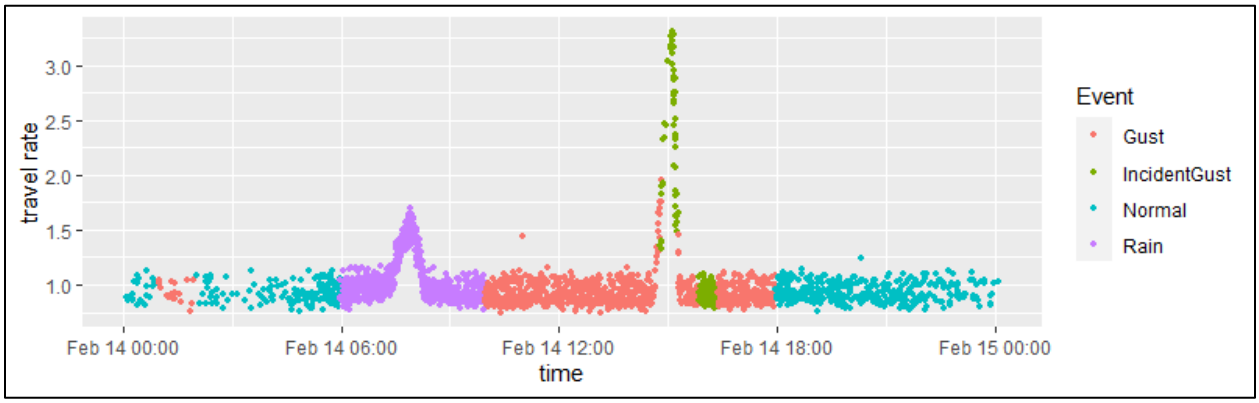


FIGURE D.1: TRAVEL RATE VS CLOCK TIME FOR A DAY WITH RECURRING CONGESTION AND SEVER INCIDENT ON I-5 NB

The standard deviations of the groups of 30 travel rate observations (with 80% overlapping) were calculated. Note that the size (30) of travel rate groups was selected because the research team perceived that this number could represent the current traffic state without missing out on the transient sub-states in the transition from free flow to congested traffic condition (or vice-versa), as indicated in Section 3.1. The group standard deviation is plotted against time of day based on the timestamp of the latest BT observation in the travel rate group, as shown in Figure D.3.

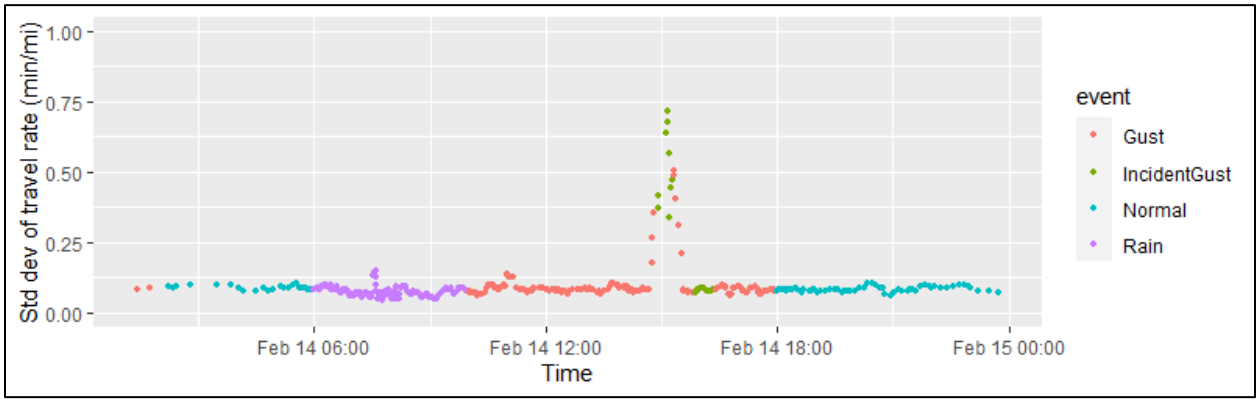


FIGURE D.2: STANDARD DEVIATION OF GROUPS OF 30 TRAVEL RATE OBSERVATIONS (WITH 80% OVERLAPPING) VS CLOCK TIME

The essence of this method for disruptive incident detection is that when a disruptive incident happens during free flow conditions, there will be a sudden drop in speeds and increase in travel rates upstream of the incident. As a result, in a group of 30 consecutive travel rate observations recorded during a transitional period from free flow to incident, or during the recovery period after clearing the incident, some observation will be normal and will belong to the pre- or post-incident free flow traffic states, whereas the remainder will be high due to the impact of the incident that has just happened. This sudden shift in traffic states within one group of travel rates recorded at the time of an incident results in a jump (a sudden spike) in the standard deviation of that group, as observed in Figure 3-30. In addition to that, Figure 3-30 also shows how the standard deviation of travel rate groups during the recurrent peak (at 07:30 am) is considerably less sharp than that of travel rates during the time of the incident.

The sudden spike in the standard deviation of a travel rate group when incidents occur may also be due to incidents that cause the blockage of one lane while others stay open. This results in variation in travel rates of vehicles travelling across the different lanes of the freeway, and this variation is then reflected by high standard deviation among the travel rate observations in a single group.

The importance of overlapping in highlighting and expanding the effect of transitions from one traffic state to another is observed by comparing the number of spiked standard deviation points during the period of the disruptive incident occurrence between Figures D.2 and D.3. In Figure D.3, it is evident that removing the 80% group overlap significantly obscures the advantage of using the standard deviation of travel rate group as an incident detection tool, because it minimizes the number of groups corresponding to transition from free flow to incident states, or vice versa.

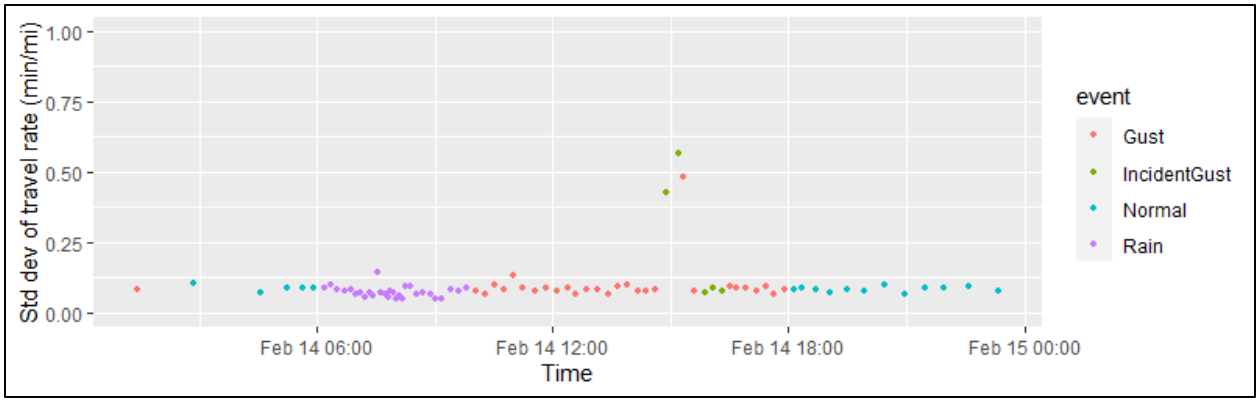


FIGURE D.3: STANDARD DEVIATION OF GROUPS OF 30 TRAVEL RATE OBSERVATIONS (NO OVERLAPPING) VS CLOCK TIME

Moreover, an important observation that can be made by observing at Figure D.1 is that not all the points exhibiting an increase in travel rates and standard deviation as a result of the incident are marked with the green (incident) stamp. This highlights the previously mentioned problem regarding the possibility of incident data inaccuracy and lag in incident reporting and underscores the need to specify an accurate operational measure by which incidents can be detected. To that end, using the standard deviation clearly succeeds in delivering an accurate and timely alarm of the degradation of the system performance as a result of disruptive incidents when the traffic conditions before the incident occurrence are uncongested.

On the other hand, looking back at Figure D.1, and knowing that recurrent peak usually occurs in the AM period on I-5 NB, one may argue that it is possible detect the occurrence of an incident at 15:00 simply by looking at clock time associated with the spike in travel rates (it happens during off-peak, so it resembles something abnormal). However, it can be difficult to differentiate the effect of incident from that of the recurrent peak when incidents occur shortly before the expected time of the recurrent peak. Figure D.4 depicts one of these instances on that happened on 02/18/2011 on I-5 SB. The recurrent peak on I-5 SB typically occurs during the PM period.

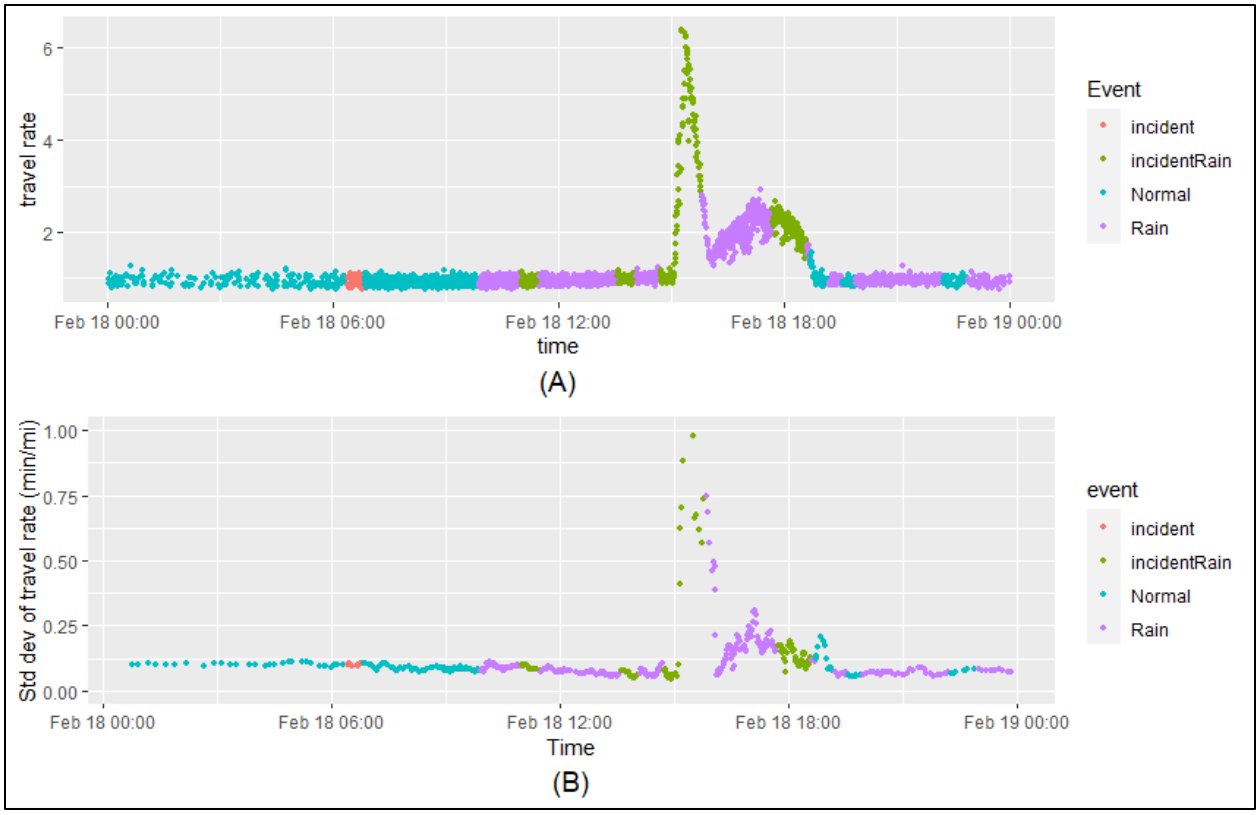


FIGURE D.4: VARIATION OF TRAVEL RATES (A) FOR ONE DAY WITH INCIDENT AND RECURRENT PEAK ON I-5 SB VS AND (B) VARIATION OF STANDARD DEVIATION OF TRAVEL RATE GROUPS

Looking at the increase in travel rates versus clock time in Figure D.4A at 15:00 in the afternoon, agencies may -at first glance- conceive that what is being observed is just an early onset of the recurrent peak. However, observing the sudden spike in the standard deviation plot during that same period (Figure D.4B) will suggest that this high standard deviation of group travel rates indicates an abrupt jump in travel rates, which has a high potential of being attributed to incident rather than a recurring peak.

Nonetheless, since this method is based on the sudden transition of the traffic condition from free flow to highly congested as a result of incident occurrence, it may fail to detect incident occurrence during already congested conditions. This is attributed to that fact that when incidents occur during congested conditions, the transition from one traffic state to another will not be as abrupt, and therefore, standard deviation of travel rates in one group will not have a high value. Furthermore, and as indicated in section 3.1, aggregating probe data in groups of 30 consecutive observations and calculating the standard deviation of travel rates in these groups once all 30 observations are obtained may results in delay in incident detection during periods of low traffic flow (such as nighttime periods). During the night period, the time required to collect the observation that form a group may be significantly increased than that during the daytime.

As discussed in the standard deviation method for incident detection, disruptive incident occurrence causes the spread of travel rate observations within one data grouping unit to propagate. This was attributed to two reasons: The first is that as soon as an incident occurs, some travel rate observations within a group will come from the pre-incident traffic state whereas the remaining travel rates will be affected by the incident occurrence. The second reason is that incidents may affect and cause blockage of some lanes while others stay open. This increase in the "spread" within one group can also be expressed using another method, which is the difference between the 95^{th} and 5^{th} percentile travel rates within the groups of 30 travel rates observations.

In Figure D.5A, the 5^{th} and 95^{th} percentile travel rates for each data group are plotted against time for the same day exhibited in Figure D.4. Figure D.5B shows the difference between these two-percentile plotted against time. From this figure, it is seen how the difference between the percentiles due to the disruptive incident (at 03:00 PM) is more pronounced than the difference during off-peak and even during the recurrent peak.

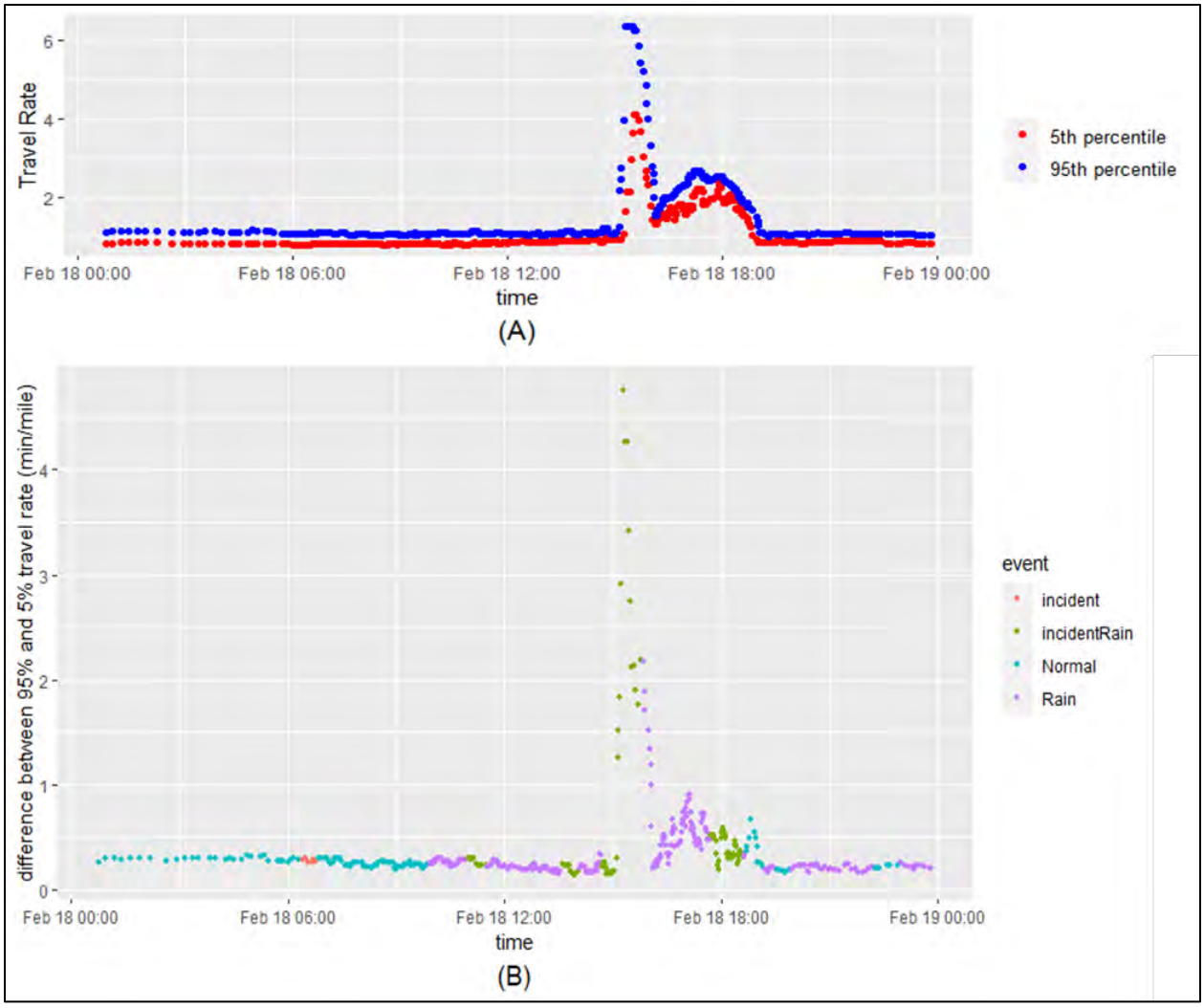


FIGURE D.5: VARIATION OF GROUP 5TH AND 95TH PERCENTILE TRAVEL RATES AGAINST TIME (A). VARIATION IN THE DIFFERENCE BETWEEN THESE PERCENTILES FOR THE SAME DAY (B)

Figure D.6A shows individual vehicles' travel rates for another day (02/17/2011) on I-5 SB. According to weather and incident reports (see color coding legend), the site had a disruptive incident at noon, in addition to rain and windy weather conditions (A). The standard deviation of travel rate groups is depicted in Figure D.6B, and the difference between the 95th and 5th percentiles is shown in Figure D.6C.

Looking at the plots, it can be observed how the disruptive incident at noon caused a sudden spike in the standard deviation and in the difference between the 95th and 5th percentiles. On the other hand, the recurring peak (starting around 16:30) does not cause any increases in the standard deviation or in the difference between the percentiles. As mentioned earlier, the role of the two explored tools is to help agencies differentiate whether the observed increase in the travel rates is associated with the recurrent peak or with an abnormal event such as an incident. However, since the difference between the percentiles notion is also based on the

same spread in the travel time observations within one group as the standard deviation tool, it may also fail to identify disruptive incidents occurrence within the peak, when the traffic is already congested and there is no study transition from free flow to congested conditions.

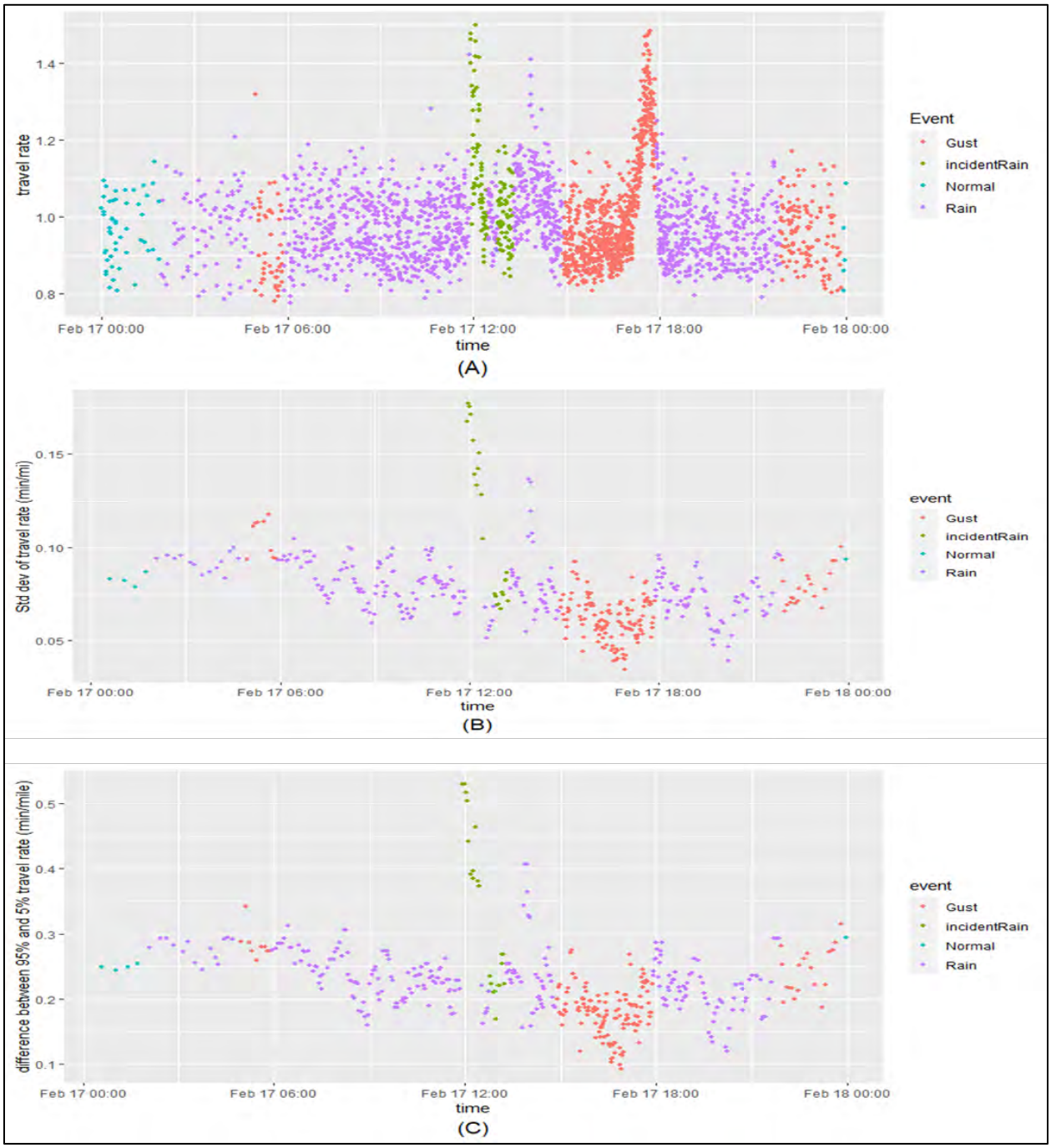


FIGURE D.6: (A): TRAVEL RATE VARIATION WITH TIME ON FEB 17,2011. (B): VARIATION OF GROUP STANDARD DEVIATION (B) AND THE DIFFERENCE BETWEEN 95TH AND 5TH PERCENTILES (C) WITH TIME DURING THE SAME DAY.

Conclusions

Early detection of disruptive incidents is important for agencies to undertake emergency actions. It is also important to investigate analytical measures that detect incident occurrence rather than just relying on incident reporting, because the latter can be missing or inaccurate on some occasions. This section presented two disruptive incident detection tools that have their basis in the increase in the spread and variation of travel rate observations within a probe data group as soon as disruptive incidents occur. The first method is based on observing a sudden and severe spike in the standard deviation of travel rate groups once a disruptive incident occurs, whereas the second method is based on observing an increase in the difference between the 5^{th} and 95^{th} percentile travel rates as due to the disruptive incident. The strengths and weaknesses of these two methods are as follows:

Strengths:

- These methods presented can help agencies overcome the lag and errors in disruptive incident reporting, which may delay traffic agencies' response to these incidents.
- These methods can help agencies in detecting disruptive incident and other events that occur during the off-peak period and can provide a tool to differentiate between incidents and recurring congestion onset when incidents happen shortly before recurring congestion onset.

Weaknesses

- The methods presented above may not be able to highlight the effect of disruptive incidents that occur when the conditions on the roadway are already congested. This case will be further investigated during phase II report.
- Grouping probe vehicles in groups of 30 may result in delaying the attainment of the std deviation and the percentile estimate during nighttime period. The overlapping technique does mitigate this problem by reducing the number of new observations needed to form a new group to 6 once the first 30 observation are collected, yet that may still take time considering the low probe vehicle matching rate. Therefore, the team suggests monitoring individual vehicle traffic rates for incidents during that night time period, and activating the grouping technique once traffic starts "picking up" in the morning.