Breakdown Prediction Utilizing Data Mining Algorithms in Combination With Performance Measures Derived From Connected Vehicle Data

Hector Donaldo Mata (hmata010@fiu.edu)
Florida International University

Atika Jabin
Florida International University

Mohammed Hadi
Florida International University

Research Article

Keywords: Connected Vehicle Data, Basic Safety Messages, Traffic State Prediction, Data Analytics, Machine Learning

Posted Date: January 2nd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2406755/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.
BREAKDOWN PREDICTION UTILIZING DATA MINING ALGORITHMS IN COMBINATION WITH PERFORMANCE MEASURES DERIVED FROM CONNECTED VEHICLE DATA

Authors and Affiliations

Hector Mata (corresponding author)
Research Assistant
Department of Civil and Environmental Engineering
Florida International University
10555 West Flagler Street, EC 3730, Miami, FL 33174
Email: hmata010@fiu.edu

Atika Jabin
Research Assistant
Department of Civil and Environmental Engineering
Florida International University
10555 West Flagler Street, EC 3730, Miami, FL 33174
Email: ajabi002@fiu.edu

Mohammed Hadi, Ph.D., PE
Professor
Department of Civil and Environmental Engineering
Florida International University
10555 West Flagler Street, EC 3605, Miami, FL 33174
Email: hadim@fiu.edu

Submission Date: December 22, 2022
ABSTRACT

Connected vehicles (CV) will provide an important source of data to support real-time management of traffic operations 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.

The analysis revealed using cluster analysis and classification decision tree that the traffic states of the freeway facility used as a case study could be classified into groups representing 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 compared the performance of three data mining/machine learning techniques for the prediction of congestion using a decision tree, a fuzzy rules-based system that utilizes the decision tree results, and a neuro-fuzzy inference system model utilizing the backpropagation optimization method for optimization. The comparison of the performance of the three prediction models demonstrated that the neuro-fuzzy inference system model achieved the best performance in terms of various performance measures that are commonly used to assess machine learning performance.

Key Words: Connected Vehicle Data, Basic Safety Messages, Traffic State Prediction, Data Analytics, Machine Learning.
STATEMENTS AND DECLARATIONS

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

Data Availability
The datasets generated during and/or analyzed during the current study are available in the Github repository, https://github.com/hctrd242/CV-Data-Selmon-Expressway

Authors Contributions
The authors confirm contribution to the paper as follows: study conception and design: Mohammed Hadi, Hector Mata, Atika Jabin; Data collection: Hector Mata, Atika Jabin; Data Preparation: Hector Mata, Atika Jabin; Analysis and interpretation of results: Hector Mata, Mohammed Hadi; Draft manuscript preparation: Hector Mata, Atika Jabin. Manuscript Review and Editing: Mohammed Hadi, Hector Mata. All authors reviewed the results and approved the final version of the manuscript.
INTRODUCTION

Freeway management is an important component of Transportation System Management and Operations (TSMO) that aims to improve the performance of the system. Freeway management systems utilize advanced strategies and decision support systems supported by data from multiple sources. Connected vehicles (CV) and connected automated vehicles (CAV) will provide an important source of data to support real-time management of traffic operations and offline analysis of traffic operations. CV data generated from vehicles are transmitted utilizing cellular communication (C-V2X) or dedicated short-range communication (DSRC). 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 the SAE J2735 standards, contains vehicle safety-related information that is broadcasted to surrounding vehicles but can be also captured by the infrastructure for use in traffic management. 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. However, a large proportion of these parameters are currently unavailable from every vehicle. Nevertheless, the information provided in the BSM Part 1 can provide an important source of data for traffic messages. Some of the measures that can be estimated using CV and AV data include 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.

Agencies have started implementing and testing CV-based applications that are expected to provide a large amount of data for potential use in traffic management. An important effort in this regard is the Connected Vehicle Pilot Deployment Program (USDOT, 2018a). The program aims to accelerate deployment, promote interoperability, and support the use of CV data. In September 2015, the USDOT awarded three sites to pilot CV deployment. CV data is becoming available from these pilot sites, which will allow further research and assessment of the use of the data.

This study explores the utilization of metrics derived based on the CV data. This 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. The study first investigates the use of a clustering algorithm to group the CV data points into clusters to identify metric values that can be used as signals for the onset of congestion. Then, the study uses the clustering results as input to a Recursive Partitioning and Regression Tree to extract logic-based rules to identify the onset of congestion. Finally, three prediction models are implemented and compared to predict the occurrence of traffic breakdown in the next 30-minute horizon based on the frequency of data points that belong to each of the clusters within the current 30-minute period. For this purpose, the study combines a decision tree and a Fuzzy Rules-Based System to produce rules for use in real-time operations to predict the breakdown in real-time operations. Finally, a third model is implemented, consisting of an Adaptive Neuro-Fuzzy Inference System (ANFIS) that utilizes backpropagation optimization to tune the membership function parameters in the fuzzy system.
LITERATURE REVIEW

There are a number of studies that have investigated using CV data for performance measurement. However, these studies have mainly focused on measures that can be estimated by existing technology such as travel time. There have been limited efforts to investigate the potential of taking advantage of the more detailed data obtained from CV in deriving additional measures for use in assessing system operations.

With the purpose of examining the accuracy of travel time estimation based on CV data (Zou et al. 2010) used simulation and found an average error percentage of 27.6%, 12.5%, and 8.2% for 1%, 5%, and 10% market penetrations, respectively. Another study (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%.

(Vasudevan et al. 2015) summarized a method to estimate the travel time and back of queue location using CV data. This research provided an alternative approach for predicting congestion by combining big-data analytics for analyzing nearly 4 billion BSM generated by the safety pilot model. This approach was able to estimate the severity of congestion with 77% accuracy. (Argote et al. 2012) estimated measures of effectiveness based on real-world vehicle trajectories including queue length, speed, number of stops, acceleration noise, and average delay per unit distance. They developed a strategy to avoid queue spillovers based on the location of the last queued vehicle and found that this concept can provide over 90% correct estimates of queue lengths with a CV penetration rate of 25% at oversaturated conditions. (Mousa et al. 2018) investigated the effectiveness of tree-based ensemble (XGB) algorithms in detecting imminent lane change maneuvers using vehicle trajectory data obtained from data collected as part of the Next Generation Simulation (NGSIM) program funded by the Federal Highway Administration (FHWA). They found that the tree-based XGB algorithm is superior to other algorithms like gradient boost and random forest, providing a 99.7% accuracy value. (Xie et al. 2019) explored the potential of using the connected vehicle data to identify high-risk locations in a more proactive manner without depending on the historical crash data. They used deceleration rate, initial speed, time of disturbance, length of vehicle, and distance between leading and following vehicle as variables. From the results, they found that time to collision with disturbance (TTCD) has a higher correlation with rear-end crash rate than other traditional surrogate safety measures (SSMs). (Emami et al. 2019) used connected vehicle data to predict the traffic flow approaching an intersection based on a Kalman filter technique. They found that there is a positive correlation between the model’s accuracy and CV penetration rate. The study concluded that a penetration rate of at least 20% is necessary to produce a good model performance. Another study by Florida International University (Azizi and Hadi, 2020) proposed using detailed metrics in combination with macroscopic parameters to achieve a better estimation of traffic performance. However, the study mainly used emulated CV data based on the trajectories produced using simulation and trajectory data collected as part of the NGSIM program.

(Park, 2002) explored the application of a neuro-fuzzy system for the short-term prediction of traffic volumes on freeway. The study implemented a hybrid system consisting of a fuzzy C-
means module to classify the traffic patterns into clusters combined with a radial-basis function neural network to provide the prediction. The study concluded that the proposed model was able to overperform other methods such as the Kalman filtering-based dynamic linear model and the Kohonen neural network. (Pribyl et al. 2003) implemented a neuro-fuzzy inference system (ANFIS) to model travel behavior and concluded that the predictive capabilities of ANFIS were superior to other traditional methods like ordinary least squares (OLS) and negative binomial model (NBM) regression.

**UTILIZED DATA**

The data utilized in this project is collected as part of the Tampa-Hillsborough Expressway Authority (THEA) CV pilot sites, which is one of the three sites funded by the USDOT CV pilot study, mentioned earlier. The CV deployment in this pilot site includes several vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) applications. The data archived in the effort are collected from CV BSM, Signal Phasing and Timing Messages (SPaT), and Traveler Information Messages (TIM) messages that are compliant with the SAE J2735 and J2945 standards. 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 pedestrian (Vadakpat, G. 2018) (USDOT, 2017). The Pilot aimed at deploying onboard CV units on privately owned vehicles, buses, and streetcars in addition to the installation of forty roadside units at the busiest locations of the pilot area.

An important consideration when working with connected vehicle data is the market penetration of connected vehicles, which is the proportion of CV in the traffic stream. The market penetration was calculated in this study by counting the number of CV based on the unique vehicle ID of each vehicle at a location as obtained from the BSM messages. The market penetration was then estimated as the CV counts obtained as described above divided by the total traffic volume estimated based on data downloaded from the regional traffic data archive, which utilizes the Iteris Clearguide platform. The estimated market penetration was approximately 2% of the total traffic flow of the study segment.

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 data was obtained from the ITS Data Sandbox (USDOT, 2018b) in the form of a JSON file containing multiple types of data grouped in one file. The data used in the analysis is based on the BSM messages received by the roadside equipment (RSE) and archived in the system. The data items utilized in the analysis based on the BSM messages are the timestamp, vehicle position, heading, speed, and acceleration. The used data includes records of weekdays only from January to December 2019 for the selected segment during the PM peak period (from 4:00 pm to 5:30 pm). Only days with no incidents were considered for the analysis.
As mentioned earlier, data from the BSM files obtained from the Tampa CV Pilot was utilized for the analysis. The obtained BSM files are specified in the SAE J2735, and they contain core data elements such as vehicle position, heading, acceleration, speed, and the related timestamp for each record. Besides the utilization of acceleration and speed records obtained directly from the BSM file, the study calculated additional measures to be used in the analysis. A brief description of each of the utilized measures is given below:

- **Acceleration/deceleration:** The BSM files provide both longitudinal and lateral acceleration. This study utilized only the longitudinal acceleration for the analysis. The longitudinal acceleration provides the acceleration value along the vehicle’s longitudinal axis of the vehicle, where a positive acceleration value is in the direction of the travel of the vehicle. The acceleration is expressed in the BSM file in 0.01m/s$^2$. Thus, the acceleration rate used in this study is calculated as follows

\[
\text{Acceleration (ft/s}^2) = X \times 0.01 \times 3.28084
\]

where $X$ is the acceleration value compliant with SAE J2735 utilized units provided in the BSM file.

- **Speed:** In compliance with the SAE J2735 standards, the speed records provided in the BSM file are expressed in 0.02 meters per second, thus, to convert the speed values from their native format into miles per hour, the following equation was applied:

\[
\text{Speed(mph)} = X \times 0.02 \times 2.23694
\]

where $X$ is the speed value compliant with SAE J2735 units provided in the BSM file.

- **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):** 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 3 below.

\[
\text{SDdp}_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (S_i - \bar{S})^2}
\]
where, $SD_{dp}$ 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 ($SD_v$)** 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 $SD_v$ is based on equation 4 below.

$$SD_v_j = \sqrt{\frac{1}{n_j-1} \sum_{i=1}^{n_j} (S_{ij} - \bar{S}_j)^2} \quad (4)$$

where, $SD_v_j$ 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 ($SD_{bv}$)**. This is the standard deviation between the average speeds of individual vehicles that traverse the study segment in a five-minute period (bin size). At least 5 vehicles where observed in each bin. Equation 5 provides the computation of this metric.

$$SD_{bv} = \sqrt{\frac{1}{n_k-1} \sum_{i=1}^{n_k} (\bar{S}_{jk} - \bar{S}_k)^2} \quad (5)$$

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 $\bar{S}_k$ 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 (6).

$$j(t) = \frac{da(t)}{dt} \quad (6)$$

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

**METHODOLOGY**

Figure 1. depicts a flowchart that provides a graphical description of the different steps involved in the proposed analytical method where the objective is to use metrics derived from CV data to identify and anticipate congestion. As reviewed in the previous section; the utilized metrics include the speed, the acceleration/deceleration, jerk, and multiple variants of the standard deviation of speed.
The analysis started with the data ingestion where the information was retrieved from Amazon S3 services using AWS Command Line interface. Once the data was retrieved it could be accessed and then prepared for the analysis. Data preprocessing consisted of converting the files from its native JSON format to data frames. Once formatted, the data was filtered, and geo processed to make possible to isolate records from the study area using a geofence. Finally, the new metrics were computed in order to proceed to the next stages of the analysis.

![Workflow of the implemented analytical process](image)

*Figure 1. Workflow of the implemented analytical process that shows the different stages involved in the proposed methodology including data ingestion, data preparation, clustering, and implementation and evaluation of three different prediction models.*

After preprocessing the data, cluster analysis was implemented to categorize the measurements into different traffic states. Cluster analysis and more specifically, the k-means algorithm is a widely used technique when it comes to traffic pattern identification. The k-means algorithm cluster the data by separating it into sample groups of equal variances by minimizing the sum of squares within each cluster (Tan et al. 2019). The implemented model considered a total of 11 features including speed, acceleration/deceleration, jerk, standard deviation between datapoints, standard deviation between vehicles, standard deviation of individual vehicles, as well as the log of speed. Principal components analysis (PCA) was utilized to reduce the dimensionality of the data. Principal components is a data analytic technique that finds new attributes that are orthogonal to each other and are linear combinations of the original attributes (James et al. 2017). The new attributes capture the maximum amount of variation in the data. In
this case, four components were able to explain more than 80% of the variance in the model thus four components were computed and used as input to the clustering algorithm.

The k-means algorithm requires the specification of the number of clusters as an input to the algorithm. This study identified the optimal number of clusters (k) based on the elbow technique, which is a commonly used technique for this purpose. The elbow technique is a plot that depicts the total within clusters sum of squares (WCSS) for different consecutive values of k. The k value is selected at the point in the graph where the decrease in the WCSS stops being significant as the value of k increases (Umargono et al. 2019).

This study used decision trees for two purposes. First, the study used a decision tree to classify the traffic conditions based on the currently collected data in real-time operations considering the results from the cluster analysis. A classification decision tree was used to determine the variable importance in categorizing traffic conditions in the cluster analysis and to identify rules for the data partitioning and classification based on the features identified in the classification based on the decision tree. Decision trees are a non-parametric approach for creating classification models that do not require any pre-existing assumption regarding the probability distribution of the label and data features (Song & Lu, 2015) Thus, 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. The implemented decision tree used for this study uses the 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 study utilized a decision tree in a second implementation for predicting the occurrence of breakdown within the next 30 minutes based on the frequency of data points related to each cluster during the current timestep (current 30 minutes). However, one of the issues with the decision tree output is the sharp decision boundaries associated with the crisp rules that can be extracted from the tree structure, which can reduce the accuracy of the results. Even though the decision tree model usually performs well, it is believed that the use of a decision tree in combination with a fuzzy rules-based system (FRBS) to convert the results from the decision tree to fuzzy rules can improve the results of the model. The use of the FRBS in this study is based on the Mamdani model that allows expressing the output of the decision tree prediction in terms of degree of membership in fuzzy sets (Didier & Prade, 1980). The main elements of a FRBS are the rules, a fuzzifier, fuzzy inference engine, and the output processor used for defuzzification (Kaufmann & Gupta, 1985). The Mamdani systems have the advantage of being easy to understand by utilizing a set of linguistic control rules (Mamdani & Assilian, 1975). In a FRBS model, the objects are represented by fuzzy if-then rules that relate the input membership function to the corresponding output membership functions (Lee, 1990). A fuzzy set is defined without a crisp defined boundary, as the elements contained in it only represent a partial degree of membership (Sugeno, 1985). The fuzzy set is the output of the related membership function and the implication method of the fuzzy inference system. The fuzzy set outputs from all rules are then combined into a single fuzzy set by applying an aggregation method and finally, the combined fuzzy set is defuzzified to obtain the final crisp output value by using a defuzzification method (Olaru & Wehenkel, 2003).
Finally, an ANFIS is implemented for prediction and compared with the decision tree and FRBS predictive models. The ANFIS classifier is a fuzzy inference system developed in a form of an adaptive fuzzy neural network in order to combine the best characteristics of the fuzzy systems and the artificial neural network (ANN) (Hosseini & Zekri, 2012). The ANFIS is based on a fuzzy Sugeno model that can be implemented when there exists a collection of input and output data available for modeling (supervised model). Among its main advantages, the ANFIS provides great help to produce if-then fuzzy rules based on the neural network results in order to properly describe the interactions of the variables in a complex model. Also, the human expertise that is usually required in fuzzy systems (and sometimes not available) is not needed for the ANFIS implementation since neural networks are used to help extracting these rules (Basri, 2008). The implemented ANFIS model utilizes backpropagation as optimization method which employs gradient descent to tune the parameters of the membership functions (Talpur et al. 2017).

CASE STUDY

A portion of the Selmon Expressway located in Tampa, Florida is used as a case study in this paper. The Selmon Expressway is managed and operated by the Tampa-Hillsborough Authority (THEA). A 1.6 mile segment of the Selmon Expressway is part of the THEA connected vehicle pilot program funded by the USDOT mentioned earlier. Figure 1 shows the THEA pilot location and the study segment within the location. As shown in Figure 1, CV data is available for the portion of the Selmon Expressway that is adjacent to the downtown area. This study utilized data from the 320 ft located in the most downstream section for the Selmon Expressway segment that is equipped with CV devices in the northbound direction. This segment was selected for use as a case study because of the large variation in the day-of-day traffic conditions compared to other Selmon Expressway segments that are covered by the CV pilot.

Figure 1. Connected vehicle pilot deployment in downtown Tampa and case study segment location. The study segment consists of 320 ft of freeway segment on the northbound direction.
ANALYSIS AND DISCUSSION

The results in Figure 2 were used to determine the number of clusters for use in the analysis. Figure 2 shows that the WCSS decreases at a significant rate up to six utilized clusters. The rate of decrease in WCSS is significantly lower for each additional cluster after six clusters. Hence, six clusters were used in the analysis.

Figure 2. Total within clusters Sum of Squares (WCSS) for different numbers of clusters. The graph shows that a significant decrease in WCSS is achieved for six clusters. For more than six clusters the decrease in WCSS is not as significant, therefore it is inferred that six clusters is the optimal number of clusters to categorize the traffic states.

The output of the k-means clustering is shown in Figures 3 and 4. Figure 3 shows the relationship between speed and SDdp for each cluster whereas Figure 4 shows the relationship between SDbv and acceleration/deceleration for each cluster. The clustering algorithm classified the data points into six clusters. Clusters 1, 2, and 3 include data with relatively high-speed regime (speed>=45 mph), but Clusters 1 and 3 data have lower SDdp values compared to Cluster 2, possibly indicating that the data points in Cluster 2 reflect traffic conditions that are about to transit to a congestion regime. Cluster 4 contains data points that are characterized for having high standard deviation and scattered over both low and high-speed areas. Figure 3 shows that Clusters 5 and 6 belong to the lower speed regime (speed < 45 mph). The main differentiation between Cluster 5 and Cluster 6 is that Cluster 5 represents data points with lower values of SDdp whereas Cluster 6 contains data points with much higher values of SDdp, possibly indicating stop-and-go operation.
Figure 3. $k$-means output: scatterplot of speed and $SD_{dp}$ for the identified clusters. Clusters 1, 2, and 3 represent the high-speed regime. Cluster 4 represents the transition regime, and cluster 5 and 6 represent the low-speed regime, being cluster 6 the break down state.

Figure 4. $k$-means output: scatterplot of $SD_{bv}$ and acceleration/deceleration for the identified clusters. Note that the high-speed regime clusters (1 to 3) present lower values of both $SD_{bv}$ and accel/decel. Whereas cluster 4 (transition) present the highest deceleration values indicating the onset of congestion.

Figure 4 shows $SD_{bv}$ and acceleration/deceleration values for different clusters. Cluster 1 represents uncongested conditions with the lowest standard deviation and acceleration/deceleration. Data points in Cluster 1 may represent a regime with no indication of breakdown in the near future. Cluster 2 represents traffic traveling near the posted speed limit, but with much higher values of $SD_{bv}$ ($6<SD_{bv}<12$) indicating some degree of disturbance. However, according to Figure 4, Cluster 2 does not present extremely high values of acceleration or deceleration. Cluster 3 shows data points that tend to have significant deceleration ($-0.75<\text{deceleration}<-0.17$), possibly indicating that the queue from a downstream bottleneck is
reaching the study location so the vehicles start decelerating at a considerable rate. Cluster 4 is characterized by data points scattered over both the high speed and the low-speed regimes while exhibiting higher deceleration values \((-1.6 < \text{decel}4 < -0.43)\) and SDbv values \((6 < \text{SDbv}4 < 20)\) indicating the onset of congestion. Cluster 5 and 6 involve data in the low-speed regime. Cluster 5 shows lower values of SDbv \((0 < \text{SDbv}5 < 6)\) with no intense acceleration or deceleration. On the other hand, Cluster 6 represents low speed combined with high-speed standard deviation \((6 < \text{SDbv}6 < 25)\) and a wide range of acceleration/deceleration values \((-0.8 < \text{accel/decel}6 < 1.0)\).

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 investigation in this study examined a large number of time series of traffic conditions that indicated that such a relationship may exist. For example, Figure 5 shows a time series of speed and the SDbv, as well as the occurrence of different clusters by time of day. Each point in Figure 5 is based on data collected for a five-minute interval. Figure 5 shows the occurrence of Clusters 3 and 4 about 30 minutes before the occurrence of the actual breakdown, and two consecutive occurrences of Cluster 4 just 10 minutes before the breakdown. Figure 5(b) shows that SDbv increases gradually as the time approaches the breakdown point.

![Figure 5](image)

**Figure 5.** Time series of speed, standard deviation of speed (SDbv) with cluster identification. The graphs confirm that transition from high-speed regime to low-speed regime is preceded by a significant increase in SDbv characterized by cluster 4 (transition). Figure 5(b) also shows that the highest values of SDbv are observed during breakdown due to a stop-and-go operation.

**Variable Importance and Rules Extraction Using a Decision Tree**

As stated earlier, 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 (Olaru & Wehenkel, 2003). Table 1 shows the rules produced based on the decision tree that can be used to classify traffic conditions. According to the rules in Table 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.

Table 1. Crisp (If-Then) Rules Extracted from the Decision Tree

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>if speed $\geq$ 45 &amp; accel/decel $\geq$ -0.17 &amp; SDbv $&lt; 4.1$</td>
</tr>
<tr>
<td>2</td>
<td>if speed $\geq$ 45 &amp; accel/decel $\geq$ -0.41 &amp; SDbv $\geq 4.1$</td>
</tr>
<tr>
<td>3</td>
<td>if speed $\geq$ 45 &amp; accel/decel $&lt; -0.17$ &amp; SDbv $&lt; 4.1$</td>
</tr>
<tr>
<td>4</td>
<td>if speed $&lt; 45$ &amp; accel/decel $&lt; -0.43$ &amp; SDdp $\geq 6$</td>
</tr>
<tr>
<td>5</td>
<td>if speed $\geq 45$ &amp; accel/decel $&lt; -0.41$ &amp; SDbv $\geq 4.1$ &amp; SDdp $&lt; 6$</td>
</tr>
<tr>
<td>6</td>
<td>if speed $&lt; 45$ &amp; accel/decel $\geq -0.43$ &amp; SDdp $\geq 6$</td>
</tr>
</tbody>
</table>

Implementation of Decision Tree for Prediction

The implementation of the decision tree to predict the occurrence of breakdown within the next 30 minutes was 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 6. 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 6 shows, the occurrence of Cluster 4 instances is critical in predicting 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 proceeds to evaluate Clusters 1, 2, and 3 at their respective thresholds.
Figure 6. Graphical output of the decision tree implemented for prediction of breakdown. To determine the probability of breakdown, the decision tree evaluates first the number of instances of low-speed and transition clusters (clusters 5 and 4 respectively), where transition cluster is characterized by higher values of deceleration.

Implementation of Fuzzy Rule-Based System

The rules necessary for implementing the FRBS are shown in Table 2. The membership functions for the FRBS are defined based on the output of the decision to convert the crisp logic input into a membership function representing the linguistic control rules from each fuzzy set. Each membership function has parameters and shapes for both the input and output variables (Zadeh, 1989).

<table>
<thead>
<tr>
<th>cluster 5</th>
<th>cluster 4</th>
<th>cluster 2</th>
<th>cluster 1</th>
<th>cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>low</td>
<td>medium</td>
<td>low</td>
<td>not relevant</td>
</tr>
<tr>
<td>low</td>
<td>low</td>
<td>medium</td>
<td>low</td>
<td>not relevant</td>
</tr>
<tr>
<td>low</td>
<td>low</td>
<td>low</td>
<td>high</td>
<td>not relevant</td>
</tr>
<tr>
<td>low</td>
<td>low</td>
<td>medium</td>
<td>high</td>
<td>not relevant</td>
</tr>
<tr>
<td>low</td>
<td>low</td>
<td>low</td>
<td>low</td>
<td>high</td>
</tr>
<tr>
<td>low</td>
<td>low</td>
<td>high</td>
<td>not relevant</td>
<td>not relevant</td>
</tr>
<tr>
<td>low</td>
<td>low</td>
<td>not relevant</td>
<td>not relevant</td>
<td>not relevant</td>
</tr>
<tr>
<td>high</td>
<td>not relevant</td>
<td>not relevant</td>
<td>not relevant</td>
<td>not relevant</td>
</tr>
</tbody>
</table>

Trapezoid and triangular shapes were used to define the shapes of the membership functions. As previously mentioned, the linguistic terms and membership function parameters were defined based on the structure and data partitioning provided by the decision tree in the preceding step.
Implementing Adaptive Neuro-Fuzzy Inference System (ANFIS).

For the purposes of the implementation of the ANFIS, the dataset was partitioned in a proportion of 70% observations for training and 30% of the observation for testing. In the ANFIS, the membership functions were optimized using the backpropagation method. The ANIFS architecture resembles the architecture of a typical neural network. There is an input layer, as well as an output layer, and some hidden layers containing the membership functions and the fuzzy rules (Talpur et al. 2017). The nodes in the input layer that take the values corresponding to the number of instances in the current 30 minutes of each of the clusters obtained from the output of the clustering algorithm. The hidden layer contains the input membership functions that are mapped to the corresponding input values (Talpur et al. 2017). The hidden layer also contains the fuzzy rules receiving information from the fuzzification neurons. The output of each node in the hidden layer is the product of all the incoming signals from the nodes in the input layer which is then used as the input to the output layer (Basri, 2008). The output layer performs defuzzification by combining the weights from the output membership functions into a unique fuzzy set (Hosseini & Zekri, 2012). The output layer is connected to a node providing the rules to be activated allowing the node to estimate the occurrence of the traffic breakdown.

Comparison of the Three Prediction Models

For the purposes of the evaluation of the performance of the three prediction models, the dataset was partitioned in a proportion of 70% observations and 30% for testing. For evaluation purposes, the output of the three models were evaluated utilizing the testing set that represents 30 percent of the data. In assessing the performance of the three models, 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 3 presents a summary of the evaluation metrics that allows comparing the performance of the decision tree, the FRBS, and the ANFIS. From Table 3, it can be concluded that the ANFIS model over-performed the other two models not only in terms of accuracy (88.64% for the ANFIS vs 81.82% for both the decision tree and the FRBS) but also in terms of sensitivity, which is a ratio that expresses how many observations were predicted as positive for the breakdown versus how many observations were actually positive (Guruprasad, 2019). The summary shows that the ANFIS model achieved a better rate in predicting the breakdown occurrence whereas the other two models underperformed in both the sensitivity and specificity metrics. The specificity evaluation results indicate that ANFIS was more accurate in predicting the negative class (no breakdown). Another important measure to evaluate the models 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 (Renuka, 2016). Based on the Precision Rate assessment results in Table 3, the ANFIS showed a significant improvement compared to the other two models by achieving a better precision ratio of 0.95 versus the 0.85 achieved by the FRBS and 0.82 achieved by the decision tree. The 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 (Scikit-learn, 2020). Again, the ANFIS model exhibits a clear edge over the decision tree and the FRBS by achieving significantly higher F1 Scores for both classes.

Table 3. Performance of the Decision Tree, FRBS and ANFIS Prediction Models

<table>
<thead>
<tr>
<th></th>
<th>Overall Model</th>
<th>Breakdown (1)</th>
<th>no breakdown (0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree</td>
<td>0.8182</td>
<td>0.8261</td>
<td>0.8095</td>
</tr>
<tr>
<td>FRBS</td>
<td>0.8182</td>
<td>0.7826</td>
<td>0.8571</td>
</tr>
<tr>
<td>ANFIS</td>
<td>0.8864</td>
<td>0.8438</td>
<td>0.9524</td>
</tr>
</tbody>
</table>

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

CONCLUSIONS

This study 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 acceleration/deceleration values. The study also developed a methodology for the prediction of the breakdown based on connected vehicle data. An initial 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. However, two additional predictive models based on fuzzy inference systems were implemented. First, a fuzzy rule-based system was implemented utilizing the crisp rules extracted from the decision tree mentioned above. The results showed that the FRBS achieved the same accuracy as the decision tree model, and it only showed a slight improvement in terms of specificity and precision for the positive class compared to those produced by the decision tree. Then, a second fuzzy inference system was derived using the adaptive neuro-fuzzy inference system. The comparison of the performance of the three prediction models demonstrated that the adaptive neuro-fuzzy inference system model utilizing the backpropagation optimization method achieved the best performance in terms of accuracy, sensitivity, precision, and F1 score when compared to the results obtained with the decision tree and the FRBS models.

The produced study demonstrates the importance of CVs as a data source to identify the real-time traffic state for better use in traffic operations and management. The traffic management agencies can apply the derived metrics to predict the onset of congestion, enabling the implementation of proactive strategies such as proactive ramp metering, variable speed limit/speed harmonization, dynamic lane assignment, and so on to mitigate or delay the effects of congestion. Future studies can investigate the effectiveness of such proactive strategies under different conditions.
Different variations and applications of the presented algorithms can be developed. It would be desirable to investigate the effectiveness and transferability of the presented model for its implementation in other locations where connected vehicle data is available. Taking into consideration that the ANFIS model produced the best results for this case study, it is recommended that future studies investigate different architectures of neural networks to explore the implementation of predictive models based on connected vehicle data.
REFERENCES


