Connected Vehicle Data Safety Applications

September 2023 | Final Report

SAFE-D
SAFETY THROUGH DISRUPTION

Texas A&M Transportation Institute

San Diego State University
Leadership Starts Here

VIRGINIA TECH TRANSPORTATION INSTITUTE
VIRGINIA TECH

Disclaimer

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. TTI-05-01	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Connected Vehicle Data Safety Applications		5. Report Date September 2023 6. Performing Organization Code:
7. Author(s) Michael Martin Lingtao Wu Mahin Ramezani Xiao Li Shawn Turner Sophia Stutes Faiza Hasan Michael Potter		8. Performing Organization Report No.
9. Performing Organization Name and Address: Safe-D National UTC Texas A&M Transportation Institute		10. Work Unit No. 11. Contract or Grant No. 69A3551747115/TTI-05-01
12. Sponsoring Agency Name and Address Office of the Secretary of Transportation (OST) U.S. Department of Transportation (US DOT) State of Texas		13. Type of Report and Period Final Research Report 02/2020 – 09/2023 14. Sponsoring Agency Code

15. Supplementary Notes

This project was funded by the Safety through Disruption (Safe-D) National University Transportation Center, a grant from the U.S. Department of Transportation – Office of the Assistant Secretary for Research and Technology, University Transportation Centers Program, and, in part, with general revenue funds from the State of Texas.

16. Abstract

The large-scale assessment of how driving behavior affects traffic safety and ongoing surveillance is hindered by data collection difficulties, small sample sizes, and high costs. Connected vehicles (CV) now offer massive volumes of observed driving behavior data from newer vehicles with myriad electronics and sensors that monitor the state of the vehicle, environmental conditions, and the driver's actions. This project evaluated the viability of CV data in roadway safety applications with the objective of improving existing predictive crash methods, measuring traffic speed and its relationship to crashes, and determining whether CV data could be used to evaluate pavement marking products. The research team developed safety performance functions (SPFs) for rural two-lane segments and urban intersections in Texas. The results showed that the SPFs improved with the addition of hard braking and hard acceleration counts in a majority of areas. Further, a variety of CV speed measures were generated from the CV data and were shown to have conflicting correlations with crash risk and counts. Lastly, the research team developed the data processing methods for evaluating pavement marking products but was unable to perform an evaluation due to the lack of detailed construction project records.

17. Key Words Connected Vehicle, Big Data, Crash Model, Cloud Computing, SPF		18. Distribution Statement No restrictions. This document is available to the public through the Safe-D National UTC website, as well as the following repositories: VTechWorks, The National Transportation Library, The Transportation Library, Volpe National Transportation Systems Center, Federal Highway Administration Research Library, and the National Technical Reports Library.	
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 21	22. Price $0

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Abstract

The large-scale assessment of how driving behavior affects traffic safety and ongoing surveillance is hindered by data collection difficulties, small sample sizes, and high costs. Connected vehicles (CV) now offer massive volumes of observed driving behavior data from newer vehicles with myriad electronics and sensors that monitor the state of the vehicle, environmental conditions, and the driver's actions. This project evaluated the viability of CV data in roadway safety applications with the objective of improving existing predictive crash methods, measuring traffic speed and its relationship to crashes, and determining whether CV data could be used to evaluate pavement marking products. The research team developed safety performance functions (SPFs) for rural two-lane segments and urban intersections in Texas. The results showed that the SPFs improved with the addition of hard braking and hard acceleration counts in a majority of areas. Further, a variety of CV speed measures were generated from the CV data and were shown to have conflicting correlations with crash risk and counts. Lastly, the research team developed the data processing methods for evaluating pavement marking products but was unable to perform an evaluation due to the lack of detailed construction project records.

Acknowledgements

This project was funded by the Safety through Disruption (Safe-D) National University Transportation Center, a grant from the U.S. Department of Transportation – Office of the Assistant Secretary for Research and Technology, University Transportation Centers Program. Special thanks to Darren Torbic for his expert review and feedback.

Table of Contents

List of Figures

List of Tables

Introduction

Traffic safety is a major public health issue in the United States, with motor vehicle crashes being a leading cause of death for people ages 1 to 75 and a leading cause of death for children, youth, and young adults ages 5 to 24 [1]. In addition to certain physical roadway characteristics contributing to crashes, prior studies have shown that specific driving behaviors, such as impaired and distracted driving, are related to increased crash potential. However, the large-scale assessment of the impacts of these aberrant driving behaviors on traffic safety via traditional naturalistic driving studies and ongoing surveillance is hindered by data collection difficulties, small sample sizes, and high costs.

Connected vehicles (CVs) now offer massive volumes of observed driving behavior data from newer vehicles that come equipped with numerous electronics and sensors to monitor the state of the vehicle, environmental conditions, and the driver's actions. Similar to phones, homes, and other personal electronics, vehicles are now a part of the internet of things generating and transmitting huge amounts of data. These CVs generate hyper-local measurements, including speed, trajectory, operational status, and driver events like hard braking and seat belt latching, using onboard sensors and then transmit selected measurements via cellular modems back to the vehicles' original equipment manufacturer (OEM). The data are then productized and resold by a third-party data aggregator as anonymized disaggregate data products. Current CV data consumer markets include auto insurance, marketing, finance, and traffic management, with substantial growth in the number of CVs projected for this decade (Figure 1) [2].

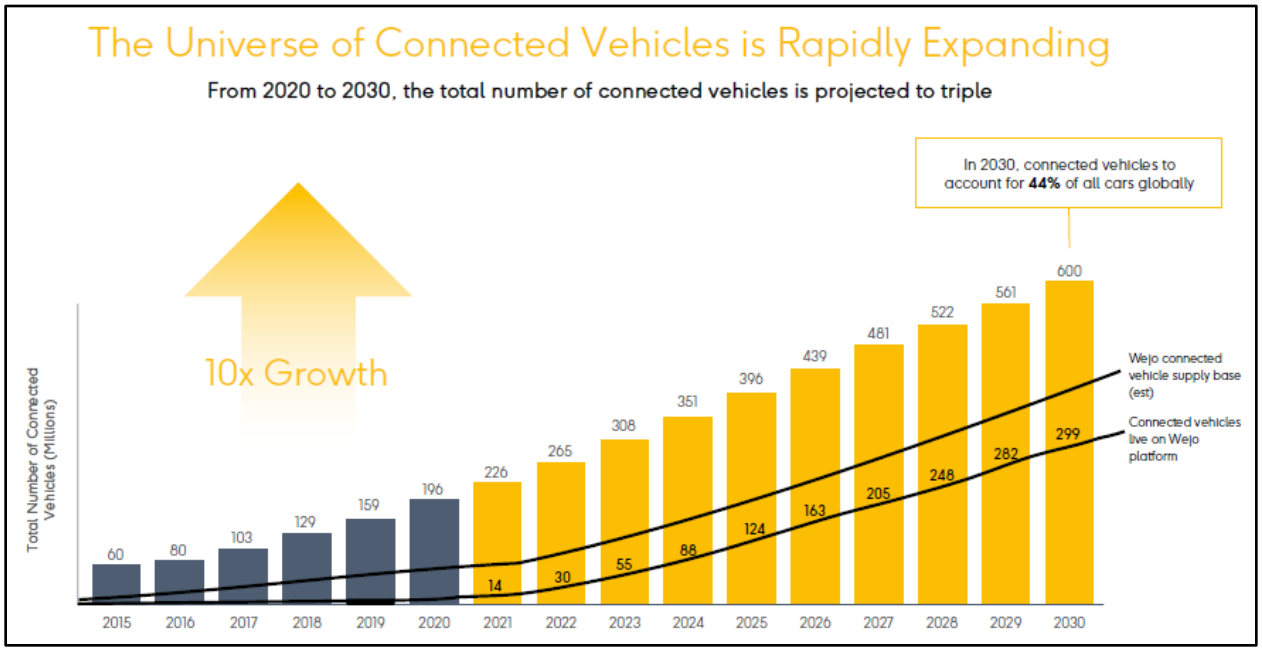

Figure 1. Graph. Wejo's investor update (November 2021).

To address the potential for this data to help save lives, this project evaluated the viability of commercially available CV data in roadway safety applications with the objective of improving the state-of-the-practice predictive crash methods. The guiding idea is that if additional crash contributing factors can be identified through CV data that cannot be identified through conventional data sets used for crash analysis, then crash predictive methodologies can be enhanced to detect crashes before they occur, thereby saving lives, time, and resources.

This project is comprised of three separate roadway safety applications:

1. **Predictive Crash Modeling**
   Determine if existing predictive crash methods can be improved by adding hard-braking, hard-acceleration, and vehicle speeds from the CV data at specific locations.

   <u>New contextual data:</u>

   - Statewide roadway intersection point locations
   - Posted speed limit segmentation
   - Intersection traffic control type (signalized/unsignalized)

   <u>Statistical models:</u>

   - Segment – rural, two-lane roadways
     - Safety performance functions (SPFs; single district and statewide)
     - Machine learning (ML; single district)
   - Intersection – urban roadways
     - SPFs (single district)

2. **CV Speed and Crash Correlation Analysis**
   Use CV speed data to generate different speed-based measures and examine their statistical relationship with crash potential on roadways.

3. **Pavement Marking Product Evaluation**
   Evaluate the use of CV data to detect the influence that retroreflective pavement marking products may have on driving behaviors in wet conditions. New contextual data included rain event locations and intensity records.

Each application required exploring and developing new cloud-based analytical methods and supporting contextual data. These creations alone comprised most of the effort spent on this project due to the complexity and scale that each had to consider, both in the amount of data and their geographic extent. Thus, the value of these methods and supporting data to others may be equal to or greater than that of the statistical findings from each application, so a commensurate level of detail is offered in the methods section.

Methodology

This section describes the methodology involved in developing the cloud computing architecture, supporting data inputs, and statistical models for each CV data safety application.

Cloud Computing Architecture

Due to the amount of CV data, this research team chose cloud computing services as the best option for storage, processing, querying, and analysis in terms of cost and computational efficiency. The research team developed the storage, computational, and analytical architecture for this research in the Microsoft Azure Cloud. Cloud services and spatial algorithms were mainly selected and developed based on their ability to spatially process the very large amounts of CV data and complex contextual data layers.

The team created the following cloud architecture:

- **Cloud Environment**
  Microsoft Azure cloud services were selected to host the project's data needs and security requirements.
- **Data Lake Storage**
  Blob storage accounts were used to store data in an unstructured data lake. This allowed a variety of data formats to be stored in one location with no extraction or transformation. The following are the data sets and their respective formats:
  - CV data as parquet files
  - Roadway segments as Esri shapefiles and JavaScript Object Notationfiles
  - Crash data as comma-separated valuesfiles
  - Weather data as Network Common Data Form raster files
- **Data Processing and Analysis Platform**
  The research team chose Azure Databricks for spatial processing, querying, and analysis needs. Databricks is a software-as-a-service platform that manages Apache Spark distributed-computing environments, also known as clusters. Spark distributes the analysis tasks across a cluster of virtual machines (VM) in the cloud for parallel data processing. Clusters are scalable in terms of number of VMs but also their number of cores and memory. Apache Sedona and Databricks Mosaic spatial libraries were employed to perform spatial joins.

Database Development

CV Data

This project explored roadway safety applications of CV data from July and October 2019 for the entire state of Texas. The data were purchased from a third-party vendor that aggregates data from OEMs and then licenses their use to customers. The CV data consist of vehicle movements with a 3-second waypoint frequency and driver events for individual vehicle trips from a single OEM.

Memory storage sizes and record counts are provided in Table 1. To illustrate the data's geographic scale, Figure 2 shows a point density maps of vehicle movements and driver events from 12:00 p.m. on October 13, 2019.

Table 1. Texas CV Data Set Quantities

Storage	Data	July 2019	October 2019
Memory Size	**Vehicle Movements**	3.4 TB	3.7 TB
Memory Size	**Driver Events**	26 GB	32 GB
Record Count	**Vehicle Movements**	49 billion	59 billion
Record Count	**Vehicle Journeys**	128 million	158 million
Record Count	**Driver Events**	406 million	518 million

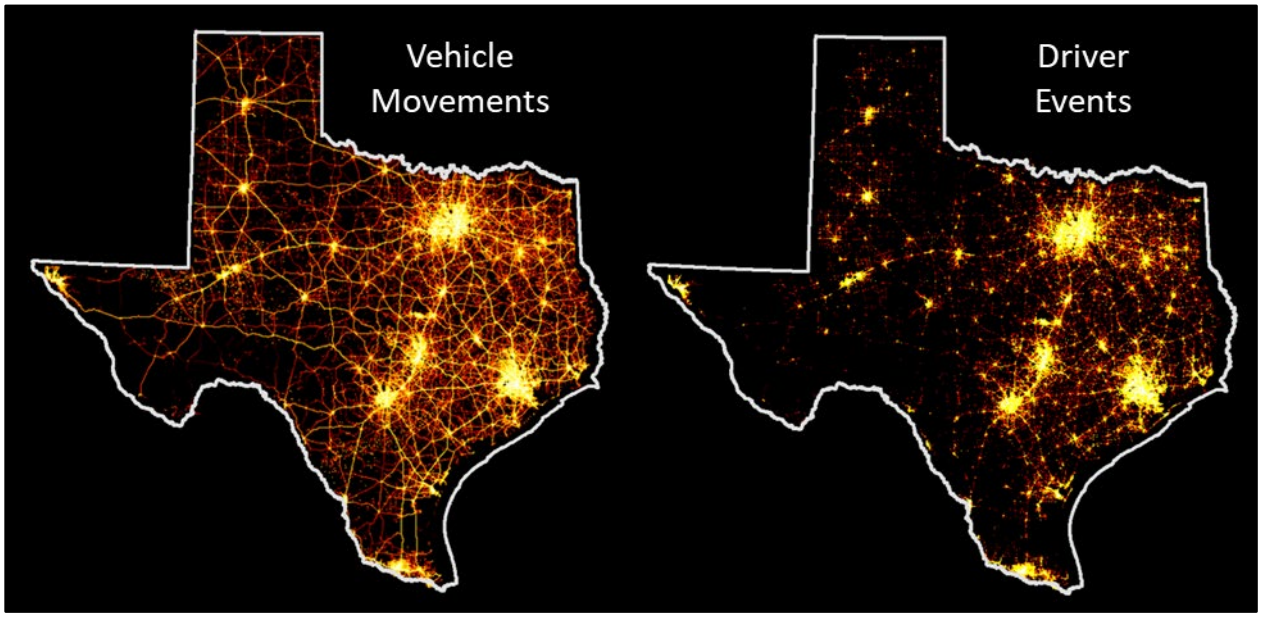

Figure 2. Maps. Vehicle movement and driver event points from 12:00 p.m. on October 13, 2019.

The data did not contain personally identifiable information (PII) and possessed no common identifier between the vehicle movements and driver events. The data use license agreement precluded users from linking the two data sets to preserve privacy. Despite having no PII, the data contained very precise times and locations of complete vehicle journeys. Journeys are defined as activity between ignition-on and -off events. New journey IDs are created for each new ignition-on event. Table 2 provides the full list of data attributes for each CV data set.

To help manage this vast amount of data, the data were repartitioned by 3-character geohash and stored in Azure blob storage accounts in parquet format. Data processing was done in Azure Databricks using the Apache Spark Sedona library [2]. The team spatially joined the movements and events to the study segments using a point-to-polygon method on a per-district basis. Point counts per segment were tabulated, exported to an Azure PostgreSQL database, and used as statistical model inputs. To accommodate the high density of CV data in metropolitan Texas

Department of Transportation (TxDOT) districts, individual geohash partitions were processed separately. This allowed faster and more cost-effective data processing.

Table 2. Vehicle Movement and Driver Event Data Attributes

Vehicle Movements	Driver Events
Data point ID	Data point ID
Journey ID	Journey ID
Timestamp	Timestamp
Latitude	Latitude
Longitude	Longitude
Geohash	Geohash
Speed	Speed
Heading	Heading
Squish VIN	Ignition status: Off, Crank, Run, ACC
Year	Acceleration type: Hard brake and Hard acceleration
Make	Seat belt latching: Latch, Unlatch
Model	Journey: Start, End, Periodic
Ignition status: Key on, Mid trip, Key off	Speed threshold (>80 mph): Above limit, Below limit

Weather Precipitation Data

One safety application of this project was to increase understanding of whether the presence of retroreflective pavement markings during wet weather conditions influenced driver behaviors. This requires knowing precise information on precipitation events such as when, where, and how much it rained. Typically, weather condition data are drawn from areawide reports or nearby weather stations. However, these methods are low resolution, producing generalizations without enough spatial or temporal precision and accuracy to capture the natural variability of weather events. For example, a strong summer afternoon thunderstorm can suddenly generate very localized flash flood conditions for a short time, whereas nearby locations remain completely dry.

Weather radar raster data for the entire United States from Iowa State University's Iowa Environmental Mesonet [3] offer precipitation intensity and accumulation data in approximately 1-square kilometer resolution, as shown in Figure 3. The raster data represent a location and a duration of time where each pixel contains the amount of accumulated precipitation in millimeters.

Finding rain events and assigning them to other events or features, like crashes, waypoints, or roadways, requires spatially processing large amounts of data. This project developed a data processing method that automatically downloads and scans the raster data to find and record rain events at point locations. The challenge is that, much of the time, it is not raining at any given point location; therefore, to identify areas of rain, the method involves first taking the daily total, then the hourly, and then the 2-minute, after performing checks to determine if there truly was rain in that area, on that day, for the road segment. Timestamp vectors from the raster layers are then populated with every 2-minute interval that contained rain, followed by an iteration of all the

timestamps to check for consecutively occurring rain events. Once a non-consecutive timestamp is found, a new rain interval is created. As each timestamp is being iterated through, the precipitation amounts corresponding to that timestamp are extracted and recorded in a running total of rain amounts per location. Once a new non-consecutive rain event is established, a new running total is started until every timestamp has been processed. Finally, after iterating through to check for consecutive 2-minute rain events, the events are then transformed into an easy-to-read table, as shown in Figure 3.

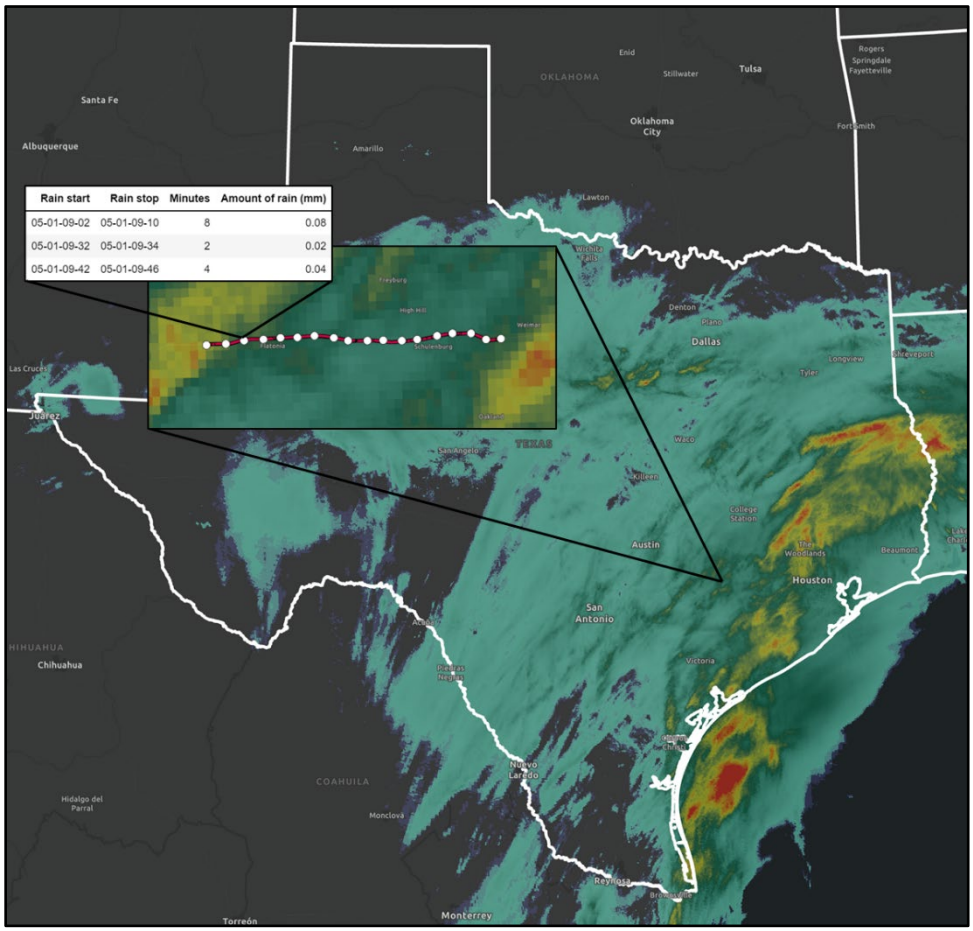

Figure 3. Map. Example rain event record for point locations along an I-10 project segment.

Roadway Intersection Data

Currently, there is no authoritative inventory of Texas roadway intersections despite intersection safety being identified as one of the emphasis areas in the Texas Strategic Highway Safety Plan; one of the strategies proposed is building a statewide intersection database. To respond to this need, the research team set out to develop its own statewide roadway intersection data set that includes location, number of legs, and traffic control type for the purpose of controlling for locations with higher likelihood of braking, acceleration, and conflicting movements.

The research team began by reviewing the TxDOT roadway inventory network but determined that it could not support automated development of intersection points due to gaps between

intersecting segments and overlapping segments. As an alternative, the research team purchased a 1-year license for the HERE roadway network data set in November 2020, including the optional feature with intersection-related data. HERE offers highly detailed, navigation-quality geospatial roadway network data from which intersection locations can be derived, along with their traffic control type and characteristics of the intersecting roadway segments, as shown in Figure 4.

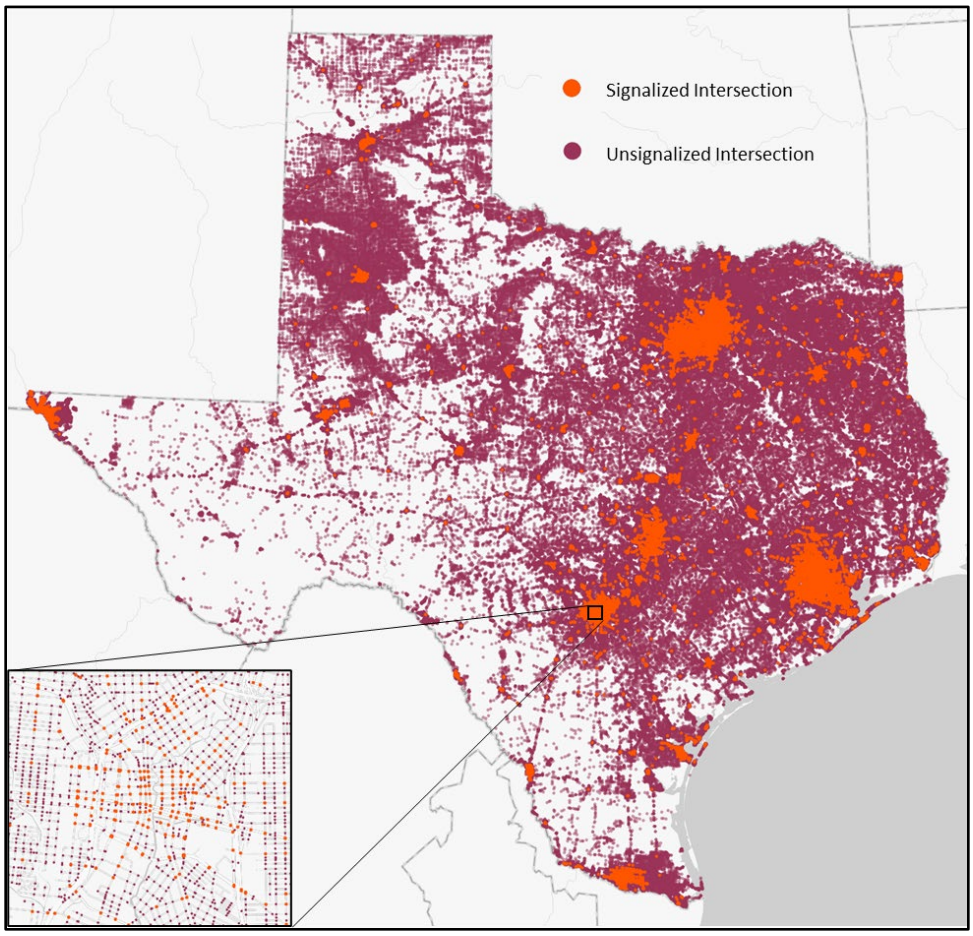

Figure 4. Map. Derived intersection locations from the HERE roadway network.

Roadway intersections were developed from the link and node tables from the HERE relational data format product stored in a local PostgreSQL database. Additional attributes were then joined to the intersections and their respective legs to eliminate link vertices and descriptive features, such as street name, functional class, ramp, number of lanes, speed limit, and traffic control type. Further processing was completed in ArcGIS Pro to reduce the number of nodes per intersection.

Roadway Segment Data

The TxDOT roadway inventory annual data [4], in shapefile format, served as the basis for spatially aggregating the CV and crash data to roadway segments. The inventory data includes physical characteristics, such as number lanes, median type, area type, shoulder type, segment length, and traffic volumes. These characteristics were used to filter the 2019 TxDOT inventory

shapefile for rural, two-lane, undivided, on-system (TxDOT maintained), and centerline features. Rural roadways were chosen because they accounted for 51% of the state's traffic fatalities in 2021 [5] and constituted 73% of TxDOT maintained roadways by centerline mileage [4]. Texas is largely a rural state by area, so using rural roadways allowed the team to maximize the CV data's expansive geographic capacity to compensate for its relatively small temporal scale of only 2 months. Roadway inventory selection criteria are provided in Table 3.

Table 3. TxDOT Roadway Inventory Selection Criteria

Characteristics	Column Name	Selection
Rural	RU	1 = Rural
Two-lane, undivided	HWY_DES1	2 = Two-way, Undivided
Number of lanes	NUM_LANES	2
Median type	MED_TYPE	0 = Without median
On-system	REC	1 = On-System Main Lanes 2 = On-System Right Frontage Road 3 = On-System Left Frontage Road
Centerline	RDBD_ID	KG = centerline

Roadway segments were further processed to be a maximum length of 2 mi, homogeneous in terms of physical characteristics, and spatially clipped with 250-ft roadway intersection point buffers to assign intersection-related labels, as shown in Figure 5. Because the CV data did not come with roadway information already provided, assigning CV GPS points to the appropriate roadway segment at intersections is challenging without a complex map matching or linear referencing algorithm. The team chose to treat segments within 250 ft of an intersection as one feature per intersection. This allowed crashes and CV data to be joined to intersections in a cost-effective manner and at an appropriate spatial scale for network screening. Final segments were buffered by 30 ft to allow for point-to-polygon spatial joining of crashes and CV data.

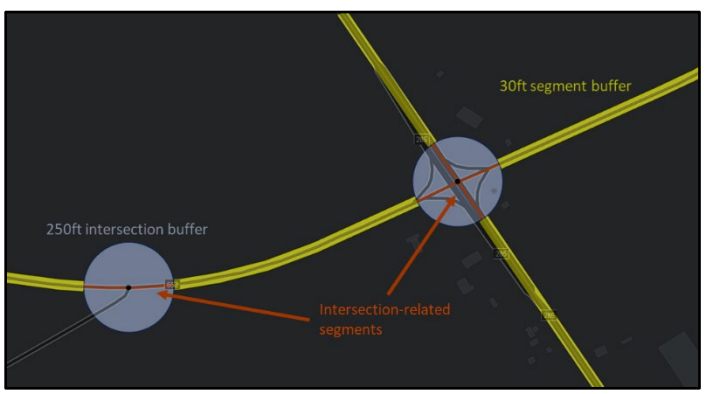

Figure 5. Diagram. Example of 30-ft segment buffers and 250-ft intersection buffers.

Crashes

Crash data from the TxDOT Crash Record Information System (CRIS) [6] were spatially joined to roadway segments in the same manner as the CV data to explore possible statistical relationships between driving behaviors and crashes. All severities were included for the years 2015-2019 for

the segment-based models and 2017-2021 for the intersection-based models and speed correlation. Apache Sedona Scala code was used in the Databricks environment to join crash points with segment polygons.

CV Speed Measures

The following segment-based speed statistics were calculated via point-to-polygon (roadbed) spatial join between the Wejo CV data and TxDOT roadway inventory study segments:

- Space-Mean Speed (SMS) is the journey distance traveled divided by the journey travel time. Segment SMS is the average journey SMS per segment. Journey distance traveled was calculated based on the waypoint speed and a 3-second frequency and then summed per journey. Only journey distances within +/-10% of the measured segment length were included.
- Time-Mean Speed (TMS) is the average of all waypoint speeds per journey. Segment TMS is the average journey TMS per segment.
- Variance for SMS and TMS per segment.
- 15th, 50th (Median), 85th, and 95th percentiles for SMS and TMS per segment.
- Speed differential between the 15th and 85th percentiles
- Speed differential between the posted speed limit and the 85th percentile
- 10-mph Pace is the 10-mph SMS and TMS range with the most journey observations per segment.

Predictive Crash Model Development

Segment – Rural, Two-lane

SPFs (Single District and Statewide)

SPFs have been commonly used by safety researchers and practitioners in roadway safety management, including network screening (also known as hot spot identification), and safety effectiveness evaluation. Conventionally, the SPFs are developed using roadway safety data, including traffic volume, segment length, and roadway characteristics. As emerging data become available to safety researchers, it is possible to develop more robust SPFs, which can potentially provide more accurate results. The research team aimed to explore the feasibility of including new attributes derived from CV data and to examine if the inclusion of CV attributes improved the predictive performance of SPFs. It is worth noting that although safety analysts have proposed various types of statistical models for developing SPFs, the negative binomial (NB) model is still the most commonly used and is recommend by the *Highway Safety Manual* (HSM) [7]. Hence, this project developed NB regression models.

The research team first prepared segment-level safety data on rural two-lane highways for each district in Texas and obtained the harsh brake event count and harsh acceleration event count for the segments. Table 4 lists the summary statistics of rural two-lane roadway segment safety data in Bryan. Rows “Hard Acceleration (Acc.) Count” and “Hard Brake Count” are CV attributes

derived from the Wejo data. They represent the event count of hard acceleration and hard brake, respectively, on each segment.

Table 4. Summary Statistics of Rural Two-Lane Roadway Segment Data in Bryan District, Texas

Variable	Mean	Std.	Min	Max
ADT	1,783.6	1,662.6	216.2	6,606.6
Length	0.2	0.3	0.10	1.95
PSL (mph)	64.3	7.7	20	75
Shoulder Width (ft)	3.0	3.1	0	16
K Factor	10.4	1.9	5.1	30.0
D Factor	55.3	5.6	50.0	84.0
Truck ADT Percentage (%)	11.8	7.3	1.8	48.2
Lane Width (ft)	11.5	0.9	10.0	13.0
Hard Acc. Count (0.1K)	0.1	0.3	0	5
Hard Brake Count (0.1K)	0.2	0.4	0	3
Crash Count (5-yr)	0.7	1.1	0	5

Note. K Factor = peak-hour factor; D Factor = directional distribution factor.

To examine whether including the CV data will improve the SPFs, the research team developed three types of SPFs using the safety data:

- Basic model, with segment length and traffic volume as independent variables (referred to as basic model).
- "Best" model considering conventional safety data (referred to as "best" conventional model).
- "Best" model considering both conventional data and CV attributes (referred to as "best" CV model).

Because the rural two-lane segment safety data in this project include many roadway characteristics, the research team first conducted manual review and validation (e.g., removing outliers and segments with missing values) of the data and then developed the three types of models. In terms of "best" model, the team utilized a stepwise model selection algorithm to determine the model that minimizes the model Akaike information criterion (AIC) for the given safety data. In evaluating the performance of the SPFs, the team primarily examined:

- Whether the CV attributes are selected in the "best" CV model and their parameters if selected.
- The dispersion parameter (also known as theta) of the models.

The dispersion parameter represents the shape of distribution of the estimated long-term crash mean. This parameter plays an important role in the Empirical Bayes (EB) analysis and crash prediction confidence intervals calculation. A higher dispersion parameter means a more stable NB model. (Note that certain publications, e.g., *HSM*, use "overdispersion parameter" to describe

how dispersed the model is. A lower overdispersion parameter indicates higher reliability. The overdispersion parameter is the reciprocal of the dispersion parameter used in this report.)

ML (Single District)

Crash prediction is a critical step in proactively reducing crash frequency and severity. Utilizing the capabilities of ML and automated machine learning (AutoML) in Microsoft Azure Databricks, the team developed a crash prediction model to identify potential contributing factors to crashes and initiate preventative measures against future crashes.

Azure Databricks provides a collaborative and scalable environment for ML and data engineering tasks. Leveraging this platform, large volumes of historical crash data were processed and analyzed, including variables such as road characteristics and driver behavior. By applying ML algorithms, the team identified patterns and correlations within the data indicative of potential crashes. AutoML in Azure Databricks further simplifies the process by automating various stages of model development, such as feature engineering, algorithm selection, and hyperparameter tuning. This not only saves time and effort but also ensures that the crash prediction model achieves optimal performance. AutoML algorithms explore different combinations of features and models, iterating through various configurations to find the best solution.

The team used three data sets to design an ML model that can predict the crash frequency:

- CV data that contains 3-second speed, location, and vehicle events such as hard braking or accelerations.
- Crash data from TxDOT CRIS.
- Geospatial roadway inventory database data that has roadway design elements and traffic volumes.

The team leveraged data from these three data sets to identify key attributes to train the model for the Bryan district. Attributes utilized in the analysis included average daily traffic (ADT), truck average annual daily traffic (AADT), lane width, shoulder width, posted speed limit, SMS average speed, SMS standard deviation of speed, ratio of SMS speed to the posted speed limit, harsh braking events, harsh acceleration events, intersection density, segment classification as an intersection or non-intersection, and total number of crashes within the segment and intersection.

Utilizing a variety of ML algorithms and techniques revealed that the XGBoost model outperformed others regarding prediction accuracy. With an R^2 value of 0.6, the XGBoost model demonstrated a reasonably strong correlation between predicted crash probabilities and actual crash occurrences. An R^2 value of 0.6 indicates that approximately 60% of the variance in the crash data could be explained by the model, which suggests that the selected attributes play a substantial role in predicting crash probabilities. By considering factors such as traffic volume, roadway design, speed-related metrics, driver behavior indicators, and intersection characteristics, the model provides valuable insights into the relationship between these attributes and crash occurrences. This information can assist transportation planners and road safety professionals in

identifying areas with high crash potential, implementing targeted countermeasures, and ultimately improving road safety for all road users.

Intersection – Urban (Single District)

The research team prepared safety data at intersections in the urban area of Tyler, Texas. The safety data include locations, control type (signalized, unsignalized), number of legs, traffic volumes, and crash frequency (2017-2021) at the identified intersections. Crashes that occurred within 250 ft of an intersection are counted. In addition, the team spatially joined the intersections and CV data and obtained the counts of harsh braking and harsh acceleration events, separately, for each intersection. The objective is similar to that of rural two-lane undivided segments, that is, examining whether including additional features from CV data will improve the performance of intersection SPFs.

The intersections are categorized into three groups: three-leg unsignalized, four-leg unsignalized, and signalized. The three groups are the same as SPF development projects conducted in Texas, such as TxDOT RTI 0-7083 [8]. The summary statistics of the safety data by intersection group, after data validation and cleaning, are shown in Tables 5, 6, 7.

Following the same procedures documented in the rural two-lane undivided roadway SPF development, the team developed SPFs for each type of urban intersection. Specifically, two forms of SPFs were developed for each type of intersection:

- Basic SPF with major- and minor-road ADT only.
- "Best" SPF considering all variables.

Similar to the segment SPFs, the "best" model includes major-road ADT, minor-road ADT, intersection characteristics, and attributes obtained from CV data (hard acceleration count and hard brake count in this project). Stepwise model selection with objective minimizing AIC is used to optimize the model.

Table 5. Summary Statistics of Urban Intersection STOP-controlled 3-leg (*N* = 1,130) Data in Tyler District, Texas

Variable	Mean	Std.	Min	Max
Major ADT	4,911.6	6,543.1	70	38,851
Minor ADT	424.7	1,425.2	1	27,323
PSL (mph)	40.5	12.6	30	70
Major Rd Lane Number	2.5	1.0	2	6
Minor Rd Lane Number	2.0	0.2	2	4
Hard Brake Count (1K)	1.1	2.0	0	19
Hard Acc. Count (1K)	0.8	2.0	0	39
Crash Count (5-yr)	2.7	4	0	20

Table 6. Summary Statistics of Urban Intersection STOP-controlled 4-leg (N = 206) Data in Tyler District, Texas

Variable	Mean	Std.	Min	Max
Major ADT	6,481.5	6,774.3	118	40,519
Minor ADT	640.0	1,332.0	26	13,167
PSL (mph)	44.6	12.6	30	70
Major Rd Lane Number	2.7	1.0	2	6
Minor Rd Lane Number	2.0	0.2	1	4
Hard Brake Count (1K)	2.1	3.0	0	17
Hard Acc. Count (1K)	1.7	3.0	0	20
Crash Count (5-yr)	4.8	5.9	0	25

Table 7. Summary Statistics of Urban Intersection Signalized (N = 196) Data in Tyler District, Texas

Signalized (*N* = 196) Variable	Mean	Std.	Min	Max
Major ADT	15,975.6	8,180.5	1,198	46,059
Minor ADT	3,334.7	3,583.9	65	19,369
PSL (mph)	44.9	9.5	30	70
Major Rd Lane Number	4.0	0.9	2	6
Minor Rd Lane Number	2.4	0.8	2	4
Hard Brake Count (1K)	9.8	6.6	0.164	38
Hard Acc. Count (1K)	12.0	9.2	0.101	48
Crash Count (5-yr)	30.1	23.4	0	97

Results

Predictive Crash Models

Segment – Rural, Two-lane

SPF (Single District and Statewide)

Because each district may have its own geographic and roadway characteristics, this project developed district-level SPFs rather than developing a single state-wide SPF. This section takes the Bryan district as an example for documenting the analysis results. Table 8 documents the modeling results of the three SPFs.

Table 8. Modeling Results of Rural Two-Lane Roadway Segment Data in Bryan District, Texas

Variable	Basic Model			"Best" Conventional Model			"Best" CV Model		
	Estimate	Std. Error	*p*-value	Estimate	Std. Error	*p*-value	Estimate	Std. Error	*p*-value
Intercept	-2.757	0.123	<0.001	-2.131	0.172	<0.001	-2.599	0.163	<0.001
ADT	0.422	0.017	<0.001	0.431	0.017	<0.001	0.411	0.016	<0.001
Length	0.283	0.011	<0.001	0.295	0.012	<0.001	0.373	0.011	<0.001
PSL	-	-	-	-0.010	0.002	<0.001	-	-	-
K Factor	-	-	-	-	-	-	-0.017	0.008	0.030
Hard Acc.	-	-	-	-	-	-	0.299	0.044	<0.001
Hard Brake	-	-	-	-	-	-	0.851	0.039	<0.001
Theta	1.453	0.081	<0.001	1.471	0.082	<0.001	2.704	0.194	<0.001
AIC	21,503.3	-	-	21,478.4	-	-	20,107.9	-	-

The modeling results indicate that both hard acceleration count and hard brake count are selected in the "best" CV model, and the parameter estimates are 0.299 and 0.851, respectively. The two values are statistically significant at the 99.9% level. Both parameters are positive, meaning greater hard acceleration count and hard brake count on segments are associated with more crashes. Specifically, as the hard acceleration count increases by 100, crash frequency increases by 35% (i.e., $e^{0.299} - 1 = 35\%$); similarly, as hard brake count increases by 100, crash frequency increases by 134% (i.e., $e^{0.851} - 1 = 134\%$). In addition, the dispersion parameter (i.e., theta) is also improved after including the two CV attributes (i.e., 2.704 vs. about 1.5 in both the basic and "best" conventional models). In short, including hard acceleration and hard brake counts in the SPFs improves the performance of the models.

Following the same procedure, the research team developed models for each of the 25 districts in Texas. Overall, the results are similar to that of Bryan. All 25 districts include hard brake count in the "best" CV model, and all the parameters are positive and statistically significant at the 99.9% level. For hard acceleration count, 15 out of 25 districts select it in the "best" CV model, and the parameter estimate is positive and statistically significant at the 99.9% level. Six out of 25 districts select it in the "best" CV model with a negative estimate (i.e., higher hard acceleration count is associated with fewer crashes); however, most parameters are not statistically significant at the 95.0% level. Hard acceleration count is not selected in the remaining six districts. Figure 6 shows the distribution of hard brake and hard acceleration parameter estimates in the 25 districts.

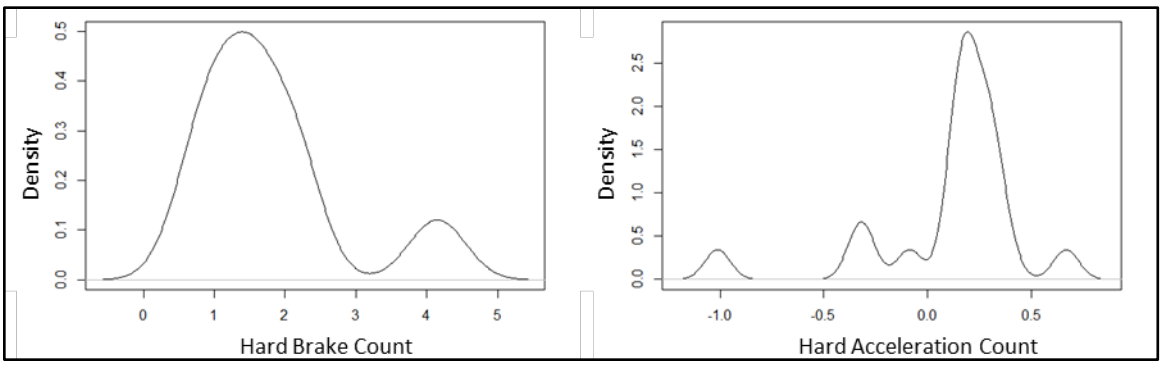

Figure 6. Graphs. Parameter distribution among 25 districts in Texas.

ML (Single District)

Figure 7 shows a scatter plot illustrating the correlation between the observed number of crashes and the predicted number of crashes generated by an ML model. Each data point in the scatter plot corresponds to a specific segment, with the color of the point indicating the associated level of crash potential categorized as low, moderate, high, or very high. These crash potential categories, known as crash risk assessment categories, were developed for TxDOT by the Texas A&M Transportation Institute and are described by Wunderlich [9]. The *x*-axis represents the number of observed crashes, providing a baseline for comparison, while the *y*-axis represents the predicted number of crashes derived from the ML model.

Figure 7 also illustrates the results of an analysis investigating the relationship between observed crashes and posted speed limit. "Observed" corresponds to the actual crash counts recorded over a 5-year period. "Predicted (SPF)" provides the forecasted crash counts obtained from an SPF model. "Expected (EB)" refers to the anticipated number of crashes, calculated using EB estimation. Finally, "Predicted (ML)" presents the projected crash counts generated by the ML model. It is observed that as the speed limit increases, the median number of observed crashes initially rises and then levels off, indicating a potential saturation point. However, the expected, predicted (SPF), and predicted (ML) crash frequencies show a consistent increase with higher speed limits. This suggests that the estimated and predicted crash counts tend to rise in proportion to the speed limit, possibly due to the model data set's potential bias for lower speed limits or external factors like better road design or stricter law enforcement at higher speed limits.

The lower right graph in Figure 7 visually represents the relationship between crash potential (i.e., categorized as low, moderate, high, and very high) and the number of crashes observed, predicted and expected by the SPF model, and predicted by the ML model. Notably, as the crash potential increases from low to moderate, there is a general increase in the number of crashes across all categories. Moving from moderate to high crash potential, the observed crashes and ML predicted crashes remain relatively stable, while the expected crashes and SPF predicted crashes continue to rise. However, as the crash potential progresses from high to very high, there is a decrease in the number of crashes across all categories.

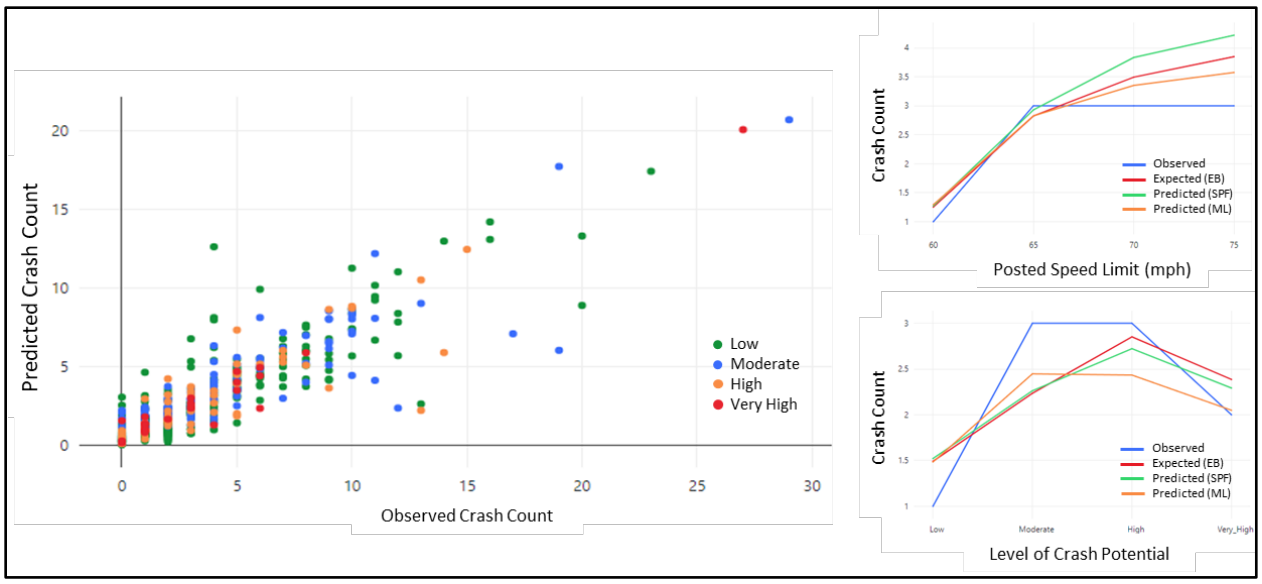

Figure 7. Graphs. ML crash prediction model results.

Intersection – Urban (Single District)

The intersection modeling results are shown in Table 9. Comparison between the basic model and "best" model indicates that for all three intersection types, the "best" model outperformed the basic model. First, the AIC of "best" models are always smaller than those of basic models. This is expected because the stepwise model selection is based on optimizing AIC values. Second, attributes obtained from the CV data are selected in the "best" models. Both the STOP-controlled three-leg intersection model and the signalized intersection model include "Hard Brake Counts," whereas the STOP-controlled four-leg intersection model includes "Hard Acceleration Counts." The estimated parameter of CV attributes in all three models are positive, indicating that more hard brake/acceleration events are related to more crashes at intersections. The results are statistically significant at a 99.9% confidence level. Finally, the dispersion parameter (i.e., Theta in Table 7) of the NB model is improved in the "best" models. For example, in the signalized intersection model, the "best" model has one more variable (i.e., Hard Brake Counts) than the basic model. By including the CV attribute in the model, the dispersion parameter, which plays an important role in the commonly used EB method, increased by 16% (i.e., 3.335 vs. 2.880).

In summary, including CV attributes in intersection SPF development improves the modeling results, and all results suggest that hard brake and hard acceleration events are positively associated with intersection crashes.

Table 9. Intersection Modeling Results

Variable	Basic Model			"Best" Model		
	Estimate	Std. Error	*p*-value	Estimate	Std. Error	*p*-value
STOP-controlled 3-leg						
Intercept	-4.713	0.231	<0.001	-3.445	0.254	<0.001
Major ADT	0.549	0.024	<0.001	0.485	0.026	<0.001
Minor ADT	0.220	0.033	<0.001	0.102	0.034	0.003
PSL				-0.011	0.003	<0.001
Hard Brake Counts				0.190	0.016	<0.001
Theta	1.405	0.117	<0.001	1.662	0.143	<0.001
AIC	4,170.4	-	-	4,066.6	-	-
STOP-controlled 4-leg						
Intercept	-6.249	0.606	<0.001	-5.233	0.659	<0.001
Major ADT	0.780	0.069	<0.001	0.775	0.089	<0.001
Minor ADT	0.181	0.055	0.001	0.103	0.055	0.062
Hard Acc. Count				0.1011	0.0199	<0.001
PSL				-0.010	0.005	0.072
Major Rd Lane Number				-0.119	0.080	0.135
Theta	1.811	0.288	<0.001	2.353	0.414	<0.001
AIC	963.5	-	-	942.8	-	-
Signalized						
Intercept	-6.845	0.761	<0.001	-4.657	0.826	<0.001
Major ADT	0.931	0.081	<0.001	0.705	0.087	<0.001
Minor ADT	0.165	0.030	<0.001	0.106	0.031	<0.001
Hard Brake Counts				0.032	0.006	<0.001
Theta	2.880	0.329	<0.001	3.335	0.393	<0.001
AIC	1,600.4	-	-	1,577.3	-	-

CV Speed and Crash Association Analysis

The relationship between speed and crash potential does not appear to be straightforward or intuitive. Figure 8 shows the correlation coefficients between a selection of the TMS speed measures that were calculated using the CV data on rural two-lane high-speed (60-75 mph) roadways in a single TxDOT district (n = 1,523). It is unclear why there is a negative correlation between the 15th to 85th percentile speed differentials and observed crash counts from all crash potential categories and each individual category crash count. This would mean that as the differential becomes wider, the number of crashes would decrease. Alternatively, the average TMS speed has an understandably positive correlation with observed crash counts and risk categorized crash counts.

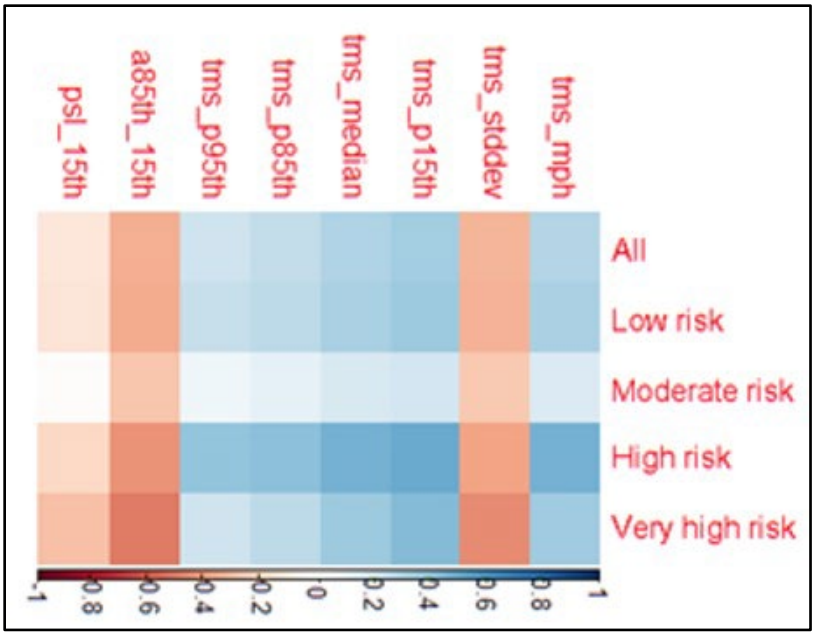

Figure 8. Graph. Correlation of speed with crash counts by crash potential category.

Pavement Marking Product Evaluation

The team collaborated with an industry partner to find locations where the industry partner's all-weather pavement marker product had recently been installed. Two sites were found in the TxDOT Yoakum district: US 183 between I-10 and Gonzales, TX, and I-10 in Fayette County, TX. However, the TxDOT construction project records in Project Tracker and in the project plans were not detailed enough for the research team to know exactly when and where the pavement products were installed. Other attempts were made to find locations and more detailed records, without success.

Conclusions

Predictive Crash Methods

- Leveraging driver behaviors in the form of hard acceleration and hard brake counts from CV data can improve state-of-the-practice HSM predictive crash methods for rural roadway segments and urban intersections by further explaining variance in crash performance. More investigation is needed to discover other CV data variables that CVs have to offer, such as emergency braking and traction control engagements.
- ML models show promise in providing similar results to HSM predictive crash methods; however, more research is needed to compare results and underlying model techniques.
- Contextual data plays an important role in understanding the relationship between drivers, their environment, and crash events. Roadway intersections, traffic control type, and speed limits should be considered fundament pieces of the data puzzle and maintained as an authoritative source by departments of transportation (DOTs).

- Considerable effort is involved in working with very large data sets like CV data. Cloud computing costs and learning new computer architecture and analysis platforms should be considered before attempting to include CV data in projects.

Speed Metrics

- Various speed measures at segment level can be derived from CV data on a large-scale roadway network, including SMS, TMS, percentile speeds, operating speed standard deviation, and more.
- Mixed relationships between operating speed measures and observed crash counts are observed in this study. Operating speed measures (average TMS, median TMS, percentile speeds) are overall positively associated with crash counts; speed differentials (e.g., standard deviation, 85th-15th) are negatively associated with crash counts, which is counterintuitive.
- The relationship between speed measures and crash occurrence is complex. Future research is needed to explore how operating speed affects crashes. This study demonstrates that detailed operating speeds can be obtained from the CV data, which covers almost all of the roadway network and all days. Having this type of data available across the network will help to investigate the relationship between speed and crashes.

Pavement Marking Product Evaluation

This project was unable to evaluate pavement marking products using CV data due to the lack of detail in the construction project records and sites that aligned with the available CV data timeframes. However, this project made great progress in developing methods that would facilitate a future product evaluation test using CV data. The research team recommends partnering closely with a DOT district office to coordinate construction and product installation details. Control sites should be selected to test the potential influence that the pavement marking product may have on driving behaviors drawn from the CV data. Changes in speed and the number of hard braking and acceleration events may serve as dependent variables.

Additional Products

The Education and Workforce Development and Technology Transfer products created as part of this project are described below and are listed on the Safe-D website https://safed.vtti.vt.edu/projects/connected-vehicle-data-safety-applications/. The final project data sets are located on the https://doi.org/10.15787/VTT1/GO97E4.

Education and Workforce Development Products

Regarding student funding, three undergraduate computer science/statistics students, one computer science master's level graduate student, and two urban and regional science Ph.D. level graduate students were hired and played key roles in data engineering, analysis, and education tasks. They helped develop model data sets in cloud computing services and GIS. This research

project offered hands-on cloud computing and big data analysis experience, which is often out of reach for students due to the high costs.

The research team is currently developing a new transportation big data lecture for the Landscape Architecture and Urban Planning department at the Texas A&M University College of Architecture. Dr. Xinyue Ye is helping to develop lecture material that introduces higher-education students to transportation big data sources and analysis techniques developed by this project.

The following Education and Workforce Development activities have or will be completed:

- Present the Transportation Data Science Seminar April 15, 2021
- Create higher education course lecture Summer 2023
- Conduct higher education course lecture Fall 2023

Technology Transfer Products

The research team completed the following:

- Researchers worked closely with an identified industry partner (3M) so that their needs would be understood and incorporated into the project work plan.
- Researchers built relationships with automotive OEM (General Motors) and CV (Wejo) data vendors by offering the opportunity to review and advise on the project.
- Research methods and results were adopted and reviewed by TxDOT to ensure the research products have practical value across Texas.

Data Products

The team uploaded to the https://doi.org/10.15787/VTT1/GO97E4 the following two model data sets: Segment – rural, two-lane (statewide) and Intersection – urban (single district)

The data sets are comprised of geometric roadway characteristics, traffic volumes, crash counts for 5 years, and CV data variables, including hard brake, hard acceleration, and CV vehicle counts. Support metadata is included to describe the source, description, and coding of categorical variables.

References

[1] Centers for Disease Control and Prevention. Transportation Safety: CDC's Injury Center Uses Data and Research to Save Lives, May 31, 2023. https://www.cdc.gov/transportationsafety/pdf/CDC-DIP_At-a-Glance_Transportation_508.pdf.

[2] Wejo. Investor Update. https://investors.wejo.com/. Accessed May 31, 2023.

[3] Apache. Apache Sedona. https://sedona.apache.org/latest-snapshot/. Accessed 2023.

[4] Herzmann, D. IEM RASTER Information. https://mesonet.agron.iastate.edu/GIS/rasters.php?rid=8. Accessed 2022.

[5] TxDOT. Roadway inventory annual data. https://www.txdot.gov/data-maps/roadway-inventory.html. Accessed 2021.

[6] TxDOT. Texas Motor Vehicle Traffic Crash Facts Calendar Year 2021, 2023. https://ftp.txdot.gov/pub/txdot-info/trf/crash_statistics/2021/01.pdf. Accessed May 31, 2023.

[7] TxDOT. Crash Record Information System. https://cris.dot.state.tx.us/.

[8] AASHTO. *Highway Safety Manual*. American Association of State Highway and Transportation Officials, 2011.

[9] Geedipally, S. R. *Calibrating the Highway Safety Manual Predictive Methods for Texas Highways* (Technical Report 0-7083-R1). Texas A&M Transportation Institute, 2022.

[10] Wunderlich, R. *Making Every Day Count: Applying Data-Driven Safety Analyses in a TxDOT District* (Technical Report 5-9052-01-R1). Texas A&M Transportation Institute, 2020. https://static.tti.tamu.edu/tti.tamu.edu/documents/5-9052-01-R1.pdf.