Estimating Traffic Crash Counts Using Crowdsourced Data

Pilot analysis of 2017 Waze data and Police Accident Reports in Maryland

Dan F.B. Flynn, PhD
Michelle M. Gilmore
Erika A. Sudderth, PhD

2018 Project Summary

DOT-VNTSC-BTS-19-01

Prepared for:
US DOT, Bureau of Transportation Statistics
Office of Director
Washington, D.C.

Notice

This document is disseminated under the sponsorship of the Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents or use thereof.

The United States Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this report.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	November, 2018	

4. TITLE AND SUBTITLE	5a. FUNDING NUMBERS
Estimating Traffic Crash Counts Using Crowdsourced Data: Pilot analysis of 2017 Waze data and Police Accident Reports in Maryland	OR02A2
6. AUTHOR(S) Dan B. Flynn, Michelle M. Gilmore, Erika A. Sudderth	5b. CONTRACT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) U.S. Department of Transportation Erika Sudderth (Project Manager) Volpe National Transportation Systems Center 55 Broadway, Cambridge, MA 02142	8. PERFORMING ORGANIZATION REPORT NUMBER DOT-VNTSC-BTS-19-01
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) US DOT, Bureau of Transportation Statistics Office of Director Washington, D.C.	10. SPONSORING/MONITORING AGENCY REPORT NUMBER N/A
11. SUPPLEMENTARY NOTES	
12a. DISTRIBUTION/AVAILABILITY STATEMENT	12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

The U.S. Department of Transportation (DOT) is leading a Safety Data Initiative (SDI), to enhance the analysis and visualization that informs safety policy decisions. In support of the SDI, Volpe applied machine learning techniques to assess the potential for crowd-sourced roadway data such as Waze alerts to serve as a reliable indicator of police-reportable crashes. Using six months of Waze alerts and police reported traffic accident data for Maryland in 2017, Volpe developed random forest models to estimate the number of police-reported crashes. The specific spatial and temporal patterns of the estimated crashes from the models is close to the observed police reported crashes, but not exact. The model underestimates crashes during early morning hours, and overestimates at commuting times, when the volume of Waze data is highest. The Waze crash models appear to capture unreported crashes, including minor crashes which might not require a police presence, but can seriously impact congestion. Near real time estimates of police-reported crashes using crowd-sourced traffic data such as Waze could offer an early indicator of traffic crash risk. In the next phase of the project, the team will work with state and local partners to implement case studies demonstrating specific applications of the Waze crash estimation models.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Safety Data Initiative, crowdsourced roadway data, police reported crashes, traffic crash estimates, data integration, Random Forest	47
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18
298-102

Contents

List of Figures

List of Tables

Executive Summary

The U.S. Department of Transportation (DOT) is leading a Safety Data Initiative (SDI), to enhance the analysis and visualization that informs safety policy decisions. As part of the SDI, the Bureau of Transportation Statistics (BTS) supported by The Volpe Center (Volpe) led the first phase of a pilot project to integrate and analyze Waze data and other transportation data resources. The SDI Waze pilot aims to gain insight from crowdsourced roadway event data to estimate the number, pattern, and severity of crashes which rise to the level of a police response, in near real time at large spatial scale. The specific objectives of phase 1 of the SDI pilot were to: 1) develop an analytical capability to support integration of DOT data resources at large scales, and 2) assess the potential for crowd-sourced Waze data to provide a rapid indicator of traffic crashes. Near real time estimates of police-reported crashes could offer an early indicator of traffic crash risk and support the Office of the Secretary of Transportation, Policy (OST-P) efforts to promote the use of data insights to reduce traffic fatalities.

The SDI Waze pilot builds on existing crash modeling tools, using an innovative combination of crowdsourced data on roadway incidents and machine learning approaches. The Waze data are available as part of the Waze Connected Citizens Program[1]. Electronic Data Transfer (EDT) is a program to transfer police accident reports (PARs) electronically from states to a federal database. Nine states have implemented EDT as of fall 2018, with seven offering complete data. By combining Waze and EDT data with additional information on the roadway network, historical crashes, demographics, and weather, the Volpe team aimed to generate modeled estimates for where and when police-reportable crashes have occurred. The estimates could serve as an early indicator of reportable traffic crashes for states that have not yet implemented EDT, and for times when EDT reports are not yet available. Waze-based traffic crash estimates could also supplement traditional crash count methods used in establishing safety policy and facilitating traffic operations.

Volpe developed cloud-based computing methods to integrate transportation data sets that were not all intended for traffic safety assessment. To assess the potential for Waze data to serve as a reliable indicator of police-reportable crashes where users are active, Volpe applied machine learning techniques to understand which features of Waze are most associated with crashes, over space and time. Using six months of data for Maryland in 2017 for this pilot study, we found that models could identify the number of actual police-reportable crashes with high accuracy. The specific spatial and temporal patterns of the estimated crashes from the models is close to the ground-truth EDT PARs data, but not exact. During the day, on higher functional classification roadways such as interstates, and at commuting times, the volume of Waze data is highest. The approach also appears to captures a wider range of crashes, including minor crashes which might not ordinarily require a police presence, but can seriously impact congestion.

[1] https://www.waze.com/ccp

In Phase 1 of the SDI Waze pilot, Volpe established the analytical pipeline and data integration methods that provide a strong foundation for using Waze data as an indicator of traffic safety to inform operational decisions. Using the Waze models trained on ground-truth EDT data, incoming Waze data could potentially be used to quickly identify police-reportable crashes that occurred in close to real-time. In phase 2 of the SDI Waze pilot, Volpe will support OST-P to assess specific applications of the crash estimation models developed in phase 1 to traffic safety questions. The team will identify State or local DOT partner(s) and support case studies that apply Waze data insights to address operational or transportation safety problems. For example, the models could be used to track estimated crash counts in specific areas over time and flag anomalous patterns. The Waze models could also be further developed to help Traffic Management Center (TMC) operators quickly identify the highest risk events when presented with an enormous amount of incoming data. By applying new techniques and data typically used in the private sector to traffic safety analysis, DOT is working to better understand roadway risk to help inform policy and decision making and improve safety.

1. Introduction

The DOT-led Safety Data Initiative aims to integrate transportation data and new "big" data to enhance safety analysis and visualization, and inform policy decisions. As part of the SDI, BTS supported by Volpe led phase 1 of a pilot project to 1) develop analytical capability to support integration of DOT data resources at large scales, and 2) assess the potential for crowd-sourced Waze traffic data to provide a rapid indicator of crashes. The Waze pilot aims to gain insight from crowdsourced data on roadway incidents to estimate the number, pattern, and severity of crashes which rise to the level of a police response, in near real time across a large spatial area (states to nation-wide).

The SDI Waze pilot relies on crowdsourced data from the mobile phone application Waze, where users report roadway incidents including accidents, hazards, jams, or road closures. The data have been made available by Waze as part of their Connected Citizens Program[2]. By combining these data with additional data on the roadway network, historical crashes, demographics, and weather, the SDI Waze project aims to generate estimates for where and when police-reportable accidents have occurred in near-real time. Timely crash estimates could provide an early indicator of emerging safety risks. The project findings could be applied across scales (regions and states) to supplement traditional crash count methods used in establishing safety policy and facilitating traffic operations.

The Waze crash data can be compared to PARs to understand the relationship between crowd-sourced traffic accident reports and crashes that the police respond to. However, PARs are typically not available outside of each individual State. A pilot EDT program for several States was established to allow States to share PARs with the National Highway Traffic Safety Administration (NHTSA) daily. EDT records should include all crashes which required a police response, and include time of incident and geospatial information. However, fatal crashes are typically not added until the police investigation is complete, often 30 days or more. EDT records are available for nine states as of fall, 2018. In contrast, Waze data in theory could provide an indicator of roadway incidents occurring across all 50 States in near real time, where users are active.

One goal of the SDI Waze pilot is to develop an analytical process to estimate the number, pattern, and severity of EDT-level crash events from Waze data. The analysis relies first on training and testing models using observed EDT and Waze data where they are linked, then identifying the location types and time periods where the model performs well. Once the model is established, the incoming Waze data can be used to generate estimates of where and when EDT-level crash events are likely to have occurred, for places and times where actual EDT data are not available. Ultimately, the models could use near real time streams of Waze data to provide estimates of the number of EDT-level crashes in locations types and time periods where model performance is sufficient (e.g., highways and primary roads during commute hours). Crash estimates could also be used to rapidly (within weeks to months) flag locations and time periods with higher than typical crash rates for further local investigation.

[2] https://www.waze.com/ccp

This SDI Waze pilot differs from previous analysis approaches in both the type of data and the methods used. For example, the Fatality Analysis Reporting System (FARS) curated by NHTSA provides a detailed description of a large number of important variables. This data is robust, and forms the basis for a large number of interventions directed at reducing serious crashes. The FARS data is most useful for retrospective analyses over multiple years. In contrast, the Waze data lacks much of the detail about individual crashes, but is available at a high frequency (in theory, every 2 minutes), for all roads in the US where Waze users are active. The use of private sector, crowdsourced data is a novel approach for DOT.

The analysis methods Volpe used in the SDI Waze pilot also differ from traditional crash analysis approaches (e.g., Lord and Mannering 2010). The analysis in this pilot study relies on random forests (RF), a machine learning approach. Similar to traditional statistical analysis, such as linear regression, there are predictor variables including weather, road type, and number of non-accident events. The response variable is binary (yes or no), indicating whether there was a Waze accident reported close to each EDT crash in one mile area grid cells during each hour of the study period (from April to September, 2018).. Random forests are based on decision trees, identifying the "branches" for combinations of data that have the most predictive power.

1.1 Report Overview

In this report we summarize phase 1 of the SDI Waze pilot, and provide a detailed overview of the data sets and analysis platform utilized by the Volpe team. We also present an extensive summary of the methods we developed to integrate the Waze, EDT, and other data sets that were not all intended for traffic safety assessment. Phase 1 of the SDI Waze pilot included three analysis iterations. In the first iteration, we developed a process to match EDT and Waze events based on defined buffers in space and time. The Volpe team then utilized Classification and Regression Tree (CART) methods, and completed preliminary assessments using RF methods, to understand the factors driving the matches of the EDT and Waze data. The first analysis iteration focused on one month of point data, where each Waze event and each EDT crash was represented as a separate row of data.

For the second iteration, we focused on a RF analysis of Waze and EDT data for grid-count data, and expanded the analysis to include 3 months of data. We aggregated the Waze, EDT, and auxiliary data to hourly counts in each one square-mile hexagonal grid cells where we had EDT or Waze observations in Maryland. We compared the performance of three initial RF models when we adjusted the spatial grain (size of the hexagonal grid cells), and included information about the six neighboring cells as predictors in the RF models.

For the third iteration, we expanded the analysis to include the full data set from April to September, 2017, across Maryland, and incorporated additional supplemental data. We tested 48 models in the third iteration, specifically assessing the performance of the model when different combinations of supplemental data were included. The supplemental variables included counts of historical fatal crashes from FARS and average annual daily traffic (AADT) from the Highway Performance Monitoring System

(HPMS). We present a summary of the key results from each iteration, discuss potential implications of the results, and conclude with a summary of the next steps for phase 2 of the SDI Waze pilot.

2. Data Summary

2.1 Waze data

The input data for the SDI Waze Pilot are derived from Waze incident data. There are four event "Types": accident, hazard, jam, or road closure, and over 20 "Sub-types". Event sub-types include major or minor accident, specific types of weather hazards such as hail or heavy rain, and specific types of jams, major or minor. Waze events can repeat over time, especially jams or road closures. For this project, the Waze event data were processed from individual events to monthly, gridded data, and matched to EDT crash reports. These gridded data were then used for analysis with RF methods to estimate the number and pattern of EDT-level crashes.

Both Waze and EDT data enter the pipeline as point datasets, with a single spatial location and either a single (EDT) or range (Waze) of time values. These point data are grid aggregated for each hour of each target month. In short, the process involves the following steps:

- Read in the merged Waze / EDT data, organized by unique identifiers for Waze and EDT events, as well as the spatial layer for the hexagonal-shaped grid cells.
- Create an expanded data frame for day, hour, and grid ID for the target month.
- Populate this data frame with the count of Waze events, the count of EDT crashes that match Waze events, the count of EDT crashes that match Waze accidents, and a large number of other aggregated variables from the Waze data.
- Remove grid cells and hour combinations without Waze or EDT observations.

Input Waze event data have features for location, time, event type, event subtype, and road type. Each row represents a single Waze event identified by a unique identifier. From April to September 2017, there were 4,538,868 unique Waze events in Maryland. A total of 148,435 of the entries were accidents reported by Waze users. Waze accident data show distinctive spatial and temporal patterns: Crashes are more common along expressways or freeways than on local roads, during weekdays, and at commuting hours (Figure 1).

2.2 Electronic Data Transfer

EDT PARs for Maryland were provided by NHTSA. There were a total of 54,030 crash records from April to September 2017 in Maryland. The EDT data have similar distributions as the Waze data, but are less spatially clustered around expressways and freeways, with a broader spread across the state (Figure 2). Strong temporal patterns still exist, with the highest number of crashes occurring on Thursdays and Fridays, in the late afternoon commuting time.

Waze and EDT data were matched spatially and temporally. For each EDT report, Waze accidents within

a timeframe of 60 minutes before or after, and within a distance of 0.5 miles were considered a match. The integration of the EDT and Waze data form the basis for the analysis, since the goal was to design a model which can estimate the number of EDT-level crashes from the Waze data.
After matching Waze and EDT data, the gridded, aggregated data take the following form:

Table 1. Excerpt of gridded data, with matched EDT crash reports (nMatchEDT_buffer_Acc) and number of Waze crash reports (nWazeAccident) shown.

GRID_ID	day	hour	DayOfWeek	nMatchEDT_buffer_Acc	nWazeAccident
A-40	91	11	Saturday	0	0
A-40	91	12	Saturday	0	0
A-40	91	13	Saturday	0	0
A-40	91	15	Saturday	1	1
A-40	91	17	Saturday	0	0
A-40	92	14	Sunday	0	1
A-40	93	7	Monday	0	0

In this gridded dataset, additional columns also summarize the count of weather hazard, jam, or road closure events, Waze event subtypes, and other variables derived from the Waze events which occurred in the grid cell at the given hour. Individual Waze events are not identifiable in the derived data.
Each row represents a grid ID at one hour of the month of interest. Columns are counts of Waze events or continuous values representing values such as median report reliability. From April to September 2017, using 1 square mile hexagonal grids, there were 2,057,791 grid ID × hour combinations.

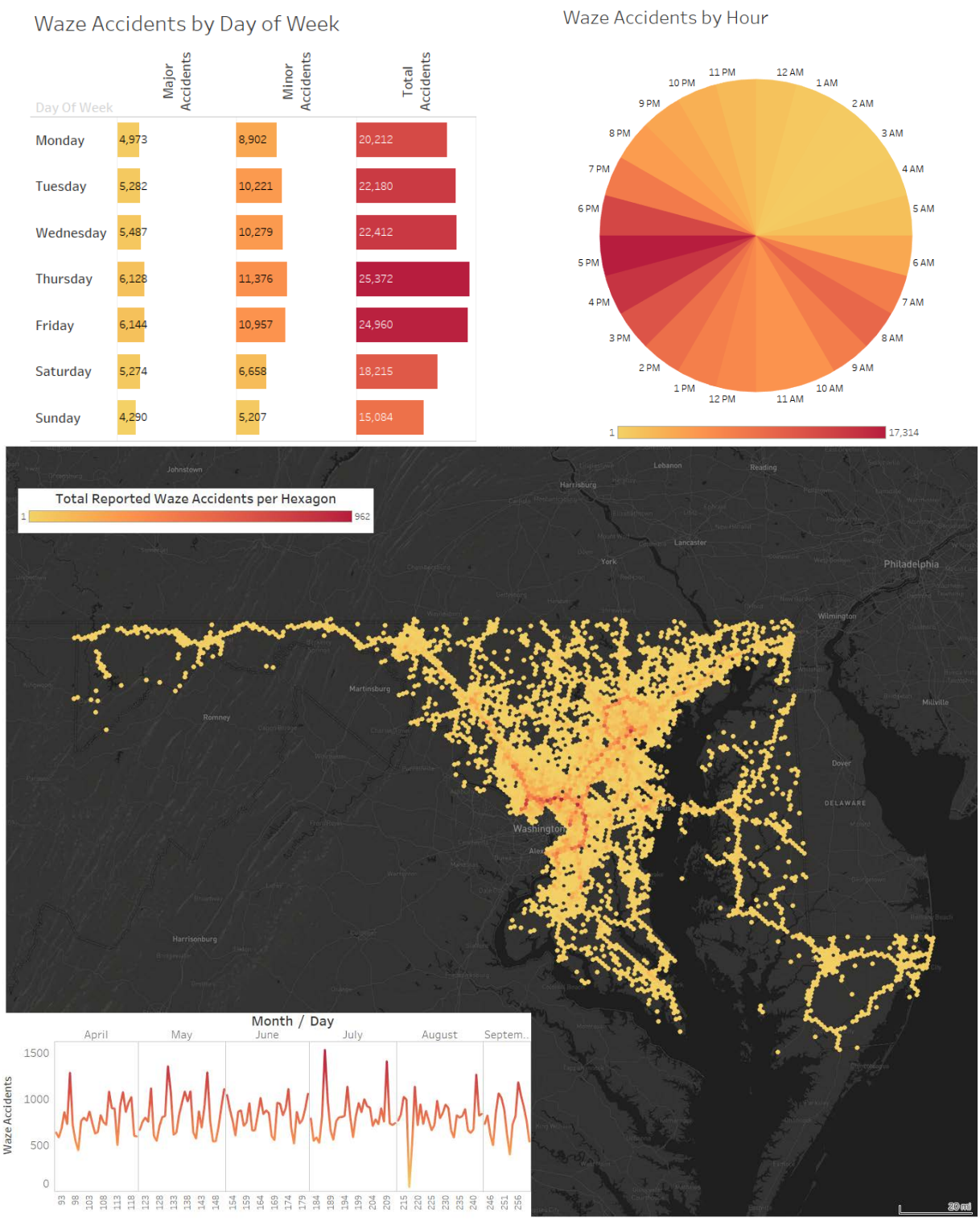

Figure 1. Spatial and temporal patterns of Waze accident reports, April - September 2017. Mapped incident data are aggregated to 1 square mile hexagonal grid cells. The temporal pattern of Waze-reported accidents is by day of week, hour of day, and day of month. Most Waze accidents occur late in the work-week (Thursday and Friday – top left panel), during commute hours (7am-9am – top right panel), and in the Capital Beltway region (dark red in the map).

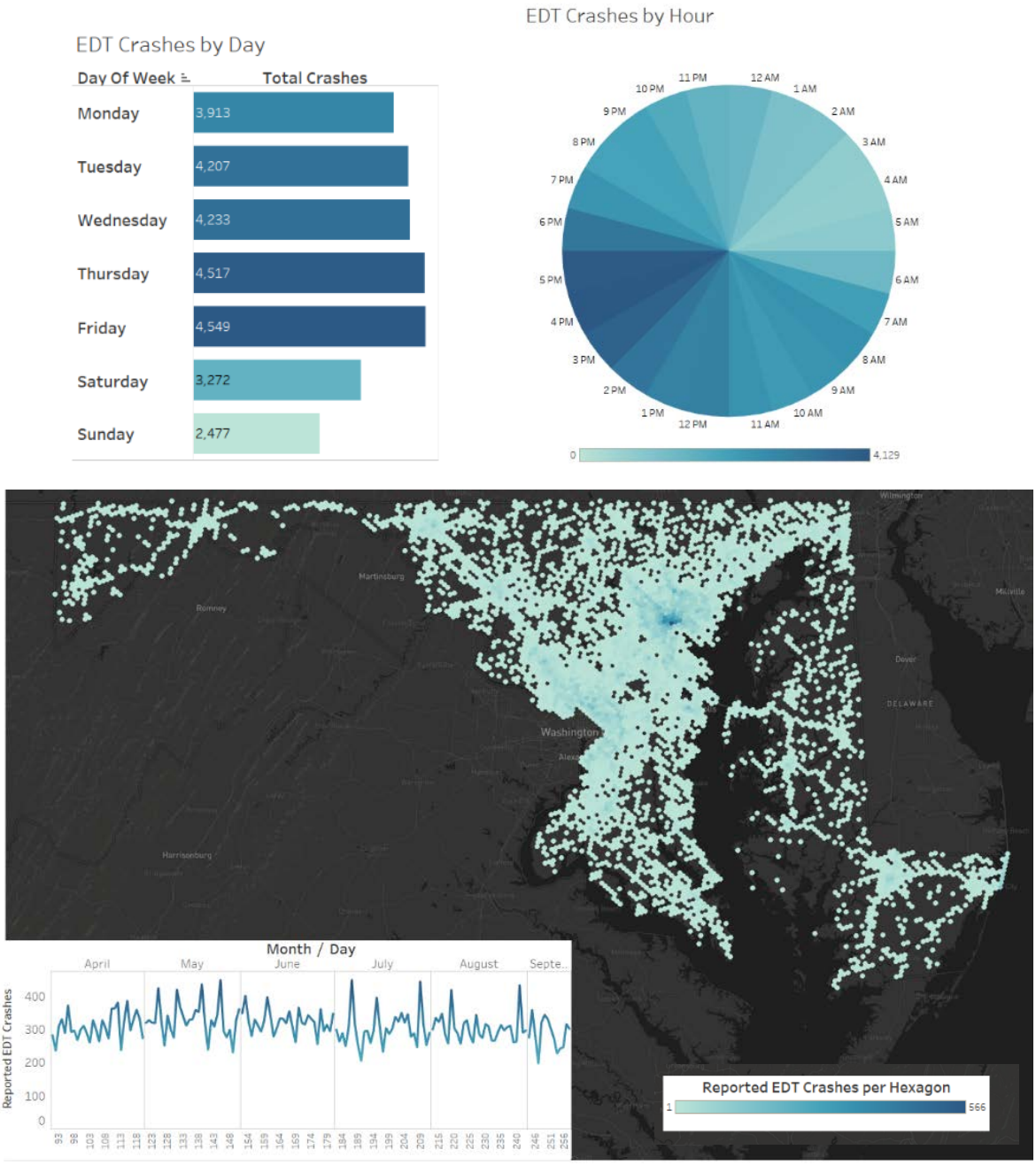

Figure 2. Spatial and temporal patterns of the EDT PARs, April-September 2017. Similar patterns over time are observed for the EDT and the Waze data shown in Figure 1. However, in the EDT data there are more police-reported crashes in Baltimore compared to the Capital Beltway.

2.4 Supplemental data

In addition to the Waze and EDT data, several supplemental datasets were integrated for this analysis. These variables were largely intended to address exposure, in terms of population, volume of traffic, miles of roadway, and historical crashes. In addition, a weather variable was brought in to address exogenous factors contributing to crashes, beyond the spatial and temporal factors already included. These variables were tested individually and in combinations to assess how inclusion would improve model performance.

Weather

- Reflectivity from the NEXRAD radar network[3], pulled hourly and merged with gridded data. This was the only supplemental variable which varied over time for each grid cell. Reflectivity can be used to represent the intensity of precipitation.

LEHD

- Economic data from Longitudinal Employer-Household Dynamics (LEHD) data set, namely the LEHD Origin-Destination Employment Statistics (LODES) dataset of the U.S. Census Bureau[4]. This includes the total number of jobs, for several earnings levels and by sex from the Residence Area Characteristic data, and total jobs, by earnings levels, by sex, and by firm size from the Workplace Area Characteristic data.

HPMS

- Road functional class: Miles of roads of each functional class, from the Highway Performance Monitoring System (HPMS)[5].
- AADT, by sum of the AADT in the roads in a grid cell, from HPMS.

FARS

- Fatal Accident Reporting System (FARS)[6] counts of fatal accidents from 2012-2016 for each grid cell.

[3] https://www.ncdc.noaa.gov/data-access/radar-data/nexrad
[4] https://lehd.ces.census.gov/
[5] https://www.fhwa.dot.gov/policyinformation/hpms.cfm
[6] https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars

3. Modeling Approach

A goal of the SDI Waze pilot is to produce estimated EDT-level crash counts which best fit the observed data in terms of overall accuracy and precision in spatial and temporal patterns. The challenge of crowd-sourced data is that there can be substantial variation in when and where users report crashes, independent of the variation in when and where crashes occur. A statistical modeling approach can be useful in this situation, to isolate the high-confidence signal of events from the noisy stream of incoming data. The primary approach we used in this pilot study was RF, a machine learning technique. We also explored the use of regularized regression, a technique based on standard statistical approaches. The two approaches can generate similar outputs (counts or presence of estimated EDT-level crashes). The distinction between these two approaches is that machine learning tools seek to generate the best match between estimates and test output, regardless of the interpretability of the model used. Regression approaches are used to test specific hypotheses about the relationships between predictors (independent variables) and a response (dependent variable).

Machine learning approaches like RF can be useful when processing a large number of input variables, and may provide estimates of the number and spatio-temporal pattern of EDT-level crashes which is sufficient for the SDI Waze project. The random forest approach also provides insight into which predictor variables have the greatest importance in estimating EDT-level crashes.

Volpe implemented the crash estimation models over three iterations, in order to assess initial results and build on the findings. In the first iteration, we matched EDT and Waze events using defined buffers in space and time, then utilized CART and RF methods to understand which factors are associated with the linkages. We began with one month of point data, where each Waze event and each EDT crash was represented as a separate row of data.

For the second iteration, we aggregated three months of the Waze, EDT, and auxiliary data to hourly counts in each one square-mile hexagonal grid cells where we had EDT or Waze observations in Maryland. We used RF models to resolve initial questions about which features to include, as well as the consequence of adding months of data and the spatial grain of the data aggregation. The first complete RF analysis used April, May, and June 2017 data from Maryland. Additional features beyond Waze data included hourly weather features, roadway characteristics, and socio-economic variables from census sources. We trained the RF models on 70% of the observations, then tested model performance on the remaining 30% of the data.

For the third iteration, we expanded the analysis to the full six months of available data and extended the random forest models to include a complete set of predictor features, including historical FARS accident counts, and AADT. We compared model performance for 48 different combinations of features (input variables). For a complete set of models tested, see Appendix 3.

3.1 Random forests overview

Random forests expand on the concept of a Classification and Regression Tree (CART), by creating a large number of possible trees composed of different sets of predictor variables to eliminate issues of overfitting. CART approaches, and decision trees in general, are a well-understood tool for classifying categorical or continuous outcomes (Brieman et al. 1984). These approaches all share the same goal, to produce estimated values with the closest match to known observed values. For classification problems, this means creating a decision tree where nodes most effectively split the outcome into a single class for each terminal node. As a simplified example, if 80% of all crashes in a given dataset occurred after 1600 hours, the top node might be time of day with two branches: time ≤ 1600 hours and time > 1600 hours. Subsequent nodes would identify additional data features that best separated the crashes into groups that are the most similar. CART approaches have been used in transportation safety research for analyses such as classification of road safety risk factors (Kwon et al. 2015) or crash severity (Chang et al. 2006). CART and other tree-based approaches are suitable for problems with a large number of potential predictors directed toward one outcome of interest.

An issue with decision trees is that they can be highly sensitive to the input data, and lack generalization when new data are provided. This problem is generally referred to as "overfitting". Random forests (Brieman 2001) minimize the problem of overfitting by creating a large number of separate decision trees, made with different subsets of the predictor variables. This provides context for how strong each predictor is in general, independent of the other predictors used. The trees are then generalized by "majority vote" of the output estimations for each observations. For the SDI Waze project, this process is implemented in the statistical programming environment R using the *randomForest* package (Liaw & Wiener 2002).

Random forests are not the only solution to analyzing this type of data. An alternative approach to a classification problem is logistic regression. Logistic regression is a well-understood regression approach where a binary response variable is converted by the logit transformation (i.e., the log of the probability of the event occurring, divided by the probability of the event not occurring), and then estimated by linear regression for any combination of predictor variables. Compared to RF, regression approaches have the advantage of generating coefficients, with associated confidence intervals and p-values, for each of the predictor variables. However, regression approaches are challenging to implement and interpret when a large number of predictor variables are provided, and where there may be a large number of interactions between these predictors. Regression is typically most appropriate when testing a specific hypothesis about a relationship between some independent (predictor) and dependent (response) variables. For the first phase of the Waze pilot project, we were most interested in accurately estimating the number of police-reportable crashes, and therefore focused on RF methods.

Regularized regression provides a technique to combine advantages of machine learning and regression approaches (Friedman et al. 2010). In regularized regression, a large number of potentially co-linear variables with unknown relationships to the response variable can be used. There are two approaches taken in regularized regression, *ridge regression* and *lasso regression* (least absolute shrinkage and

selection operator, Tibshirani 1996). Ridge regression keeps a large number of predictors, but shrinks the size of the coefficients towards each other if they are highly correlated; lasso regression keeps only the set of predictors which generate the best fit to the data, and sets the coefficients of the remaining predictors to zero. Friedman et al. (2010) proposed an implementation the *elastic-net penalty* which balances between the two approaches, and produces models which are highly interpretable (since only the most important predictors are kept) and have high performance even for a large number of sparse predictors. Elastic net generalized linear models are implemented in R using the *glmnet* package. The regularized regression models were initially tested for the SDI Waze project using the same set of input predictors and output response variables (presence and count of EDT crashes matching Waze events) as the random forest approach. In initial tests, the regularized regression models performed similarly to the random forest models when a relatively number of predictor variables were used; the SDI Waze pilot focused on random forest models because RF models were overall more accurate and performed well with large number of predictor variables.

For the SDI Waze project, the analysis Volpe used is an estimation of EDT crashes, based on Waze predictors. There were six months of data where geolocated EDT and Waze data are available, from April – September, 2017, for Maryland. The models developed in this project are training on data for which the presence of an EDT crash (binary) is modeled based on a large number of predictors from the Waze data, such as number of Waze accidents, the type of Waze events, number of Waze events in total, and other variables. A random forest model trained on known data can be employed on data for which only Waze data are provided. The method can be applied to use incoming Waze data to estimate the number and spatial/temporal pattern of EDT-level crashes in near real time or for times or states when EDT data are not available. In the model development process, the testing is done for a subset of the data where the known EDT values are held back, and then the estimates produced by the model fit to training data can be compared to the known data.

3.2 Model evaluation

There are multiple criteria for evaluating classification and regression models. For all the models, we used two different data sets to train and test the model. *Training* refers to fitting the model parameters with a large set of known EDT crashes and associated Waze events and other predictors, while *testing* refers to applying the fitted model parameters to a new set of Waze events and other predictors, generating estimated EDT crashes. The estimated EDT crashes are then compared to the known, observed EDT crashes in the test data set to evaluate model performance.

For binary classification models, it is possible to create a 2x2 table where columns are observed negative and positive, and rows are predicted negative and positive. This is known as a *confusion matrix*, and shows four quantities to represent model performance (Table 2).

Table 2. Confusion matrix definition.

		OBSERVED	
		Positive	Negative
PREDICTED	Positive	TP	FP
	Negative	FN	TN

False positives (FP) are considered Type I errors, and false negatives (FN) are considered Type II errors. The following quantities can be calculated from this matrix, and are used to evaluate model performance:

- *Accuracy* = (TN + TP) / All observations
  True positives and true negatives divided by all observations. A high value indicates that the observed occurrences and absences of EDT crashes are correctly being estimated.

- *Precision* = TP / (FP + TP)
  True positives divided by all predicted positives. A high value indicates that there are relatively few false positives (locations and times where a crash is estimated, but did not actually occur).

- *Recall* = TP / (FN + TP)
  True positives divided by all observed positives. This is also called *Sensitivity*, or the *true positive rate*. A high value indicates that there are relatively few false negatives (locations and times where a crash was not estimated, but did actually occur).

- *False Positive Rate* = FP / (TN + FP)
  False positives divided by all observed negatives. A low value indicates that there are relatively few false positives compared to all observed absences of EDT crashes.

- *Specificity* = TN / (TN + FP)
  True negatives divided by all observed negatives, also called the true negative rate. For the SDI Waze analysis, most observations are “0”, meaning no EDT crashes occurred, so much of the model performance is driven by accurately predicting these “0” (no crash) values.

Balancing between high specificity (where false positives are avoided) and high sensitivity (where false negatives are avoided) is an important decision point in evaluating a model. In discussions with a cross-modal working group within USDOT, modal representatives expressed that minimizing false negatives would be more important than minimizing false positives; false negatives indicate times where a crash is reported in the EDT data, but the model did not correctly estimate the occurrence of a crash.

Plotting the false positive rate versus the true positive rate visualizes this balance, and is known as the ‘receiver-operator characteristic (ROC) curve’ (Figure 3). The larger the area under the ROC curve, the more high specificity is maximized with low loss of sensitivity. Area under the ROC is abbreviated AUC,

and is a single value which can represent how well a model does in minimizing both false positives and false negatives. An area of 0.5 is equivalent to flipping a coin; and area of 1 is perfect estimation, with no false positives or false negatives. As a rule of thumb, areas of 0.6 or greater are generally considered to represent useful classification models; areas of 0.9 or greater are considered to represent very good classification models.

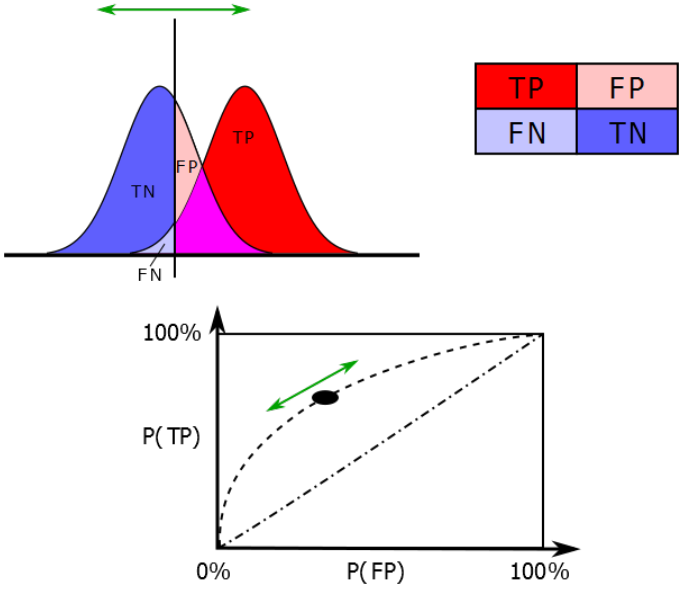

Figure 3. ROC curve interpretation, in terms of True Negative (TN), True Positive (TP), False Negative (FN), and False Positive (FP) rates. The area under the curve (AUC, bottom) shows the balance between sensitivity and specificity. When TP is high (observed crashes estimated nearly all the time correctly) and FP is low (observed non-crashes are estimated nearly all the time correctly), the curve reaches to the upper left of the plot, and the area under the curve is large. AUC for all models in this project was >0.9, indicating high accuracy.
https://commons.wikimedia.org/w/index.php?curid=44059691

4. Results

4.1 Iteration 1: Linking Waze events to EDT reports

In the first iteration of the Waze data analysis, two CART approaches were taken. First, using April 2017 data from Maryland as the test case, all EDT crashes were matched in location and time with Waze events. Of the 9,308 EDT crashes, 5,310 were co-located with a Waze event within our defined distance and time parameter (0.5 miles in radius and 60 minutes prior or following the Waze event). From the perspective of the Waze events, of the 439,562 events (including jams, accidents, road closures, and hazards), 26,268 matched EDT crashes within the spatial and temporal buffers (Table 3).

Table 3. Summary of April 2017 Maryland EDT crash and Waze event matching.

Data	Matching	Non-Matching	Total	Percent Matching
EDT	5,310	3,998	9,308	57.05
Waze	26,268	413,294	439,562	5.98

To understand the factors driving the matches of EDT to Waze and Waze to EDT, we completed two CART analyses. For EDT to Waze matching, the predictors of the decision tree were EDT features including light conditions, atmospheric conditions, damage extent, total fatal count, hour of day, day of week, and urban area classification from the U.S. Census Bureau. This first analysis showed that EDT crashes which occurred in daylight, resulted in disabling damage, and occurred in urban areas matched at least one Waze event approximately 80% of the time. Conversely, EDT crashes which occurred in the dark, before 4am, and resulted in non-disabling damage matched reported Waze events less than 20% of the time.

For Waze to EDT matching, the predictors of the decision tree were Waze features including median report reliability, number of records, type of Waze event, road classification, hour of day, day of week, and urban area classification. The highest proportion of matches with EDT crashes were Waze events which were major accidents, with more than 9 records, a median reliability of greater than 5.5, and occurring on an interstate. These Waze events matched EDT crashes 80% of the time. Together with the EDT to Waze CART analyses, these results demonstrate where and when a model of EDT-level crashes should be expected to provide useful insights.

This work was extended to preliminary assessments of the random forests method, taking multiple subsets of EDT predictors to classify Waze matches, and taking multiple subsets of Waze predictors to classify EDT matches. The first iteration focused on point data, where each Waze event and each EDT crash was represented as a separate row of data. The accuracy of the initial EDT random forest model was 65%, with precision of 66% and recall of 81%. Accuracy refers to the sum of true positives and true negatives as a proportion of all data, while precision refers to the proportion of estimated values which are correctly assigned to the true positive category, and recall refers to the proportion of observed values which are correctly estimated. Balancing high precision, recall, and accuracy is a goal of classification models.

4.2 Iteration 2: Crash count estimation using Random Forests

For the second iteration, we applied the RF method to gridded Waze and EDT data, where the unit of analysis was 1 square mile hexagonal grid cells. Within each grid cell the Waze event data are tabulated by day of the year, and hour of the day (see Gridded data section in Appendix 2 for more details). Here, 20 variables derived from Waze data were used as predictors, with the response being a binary variable of whether or not an EDT crash matching a Waze crash was detected in each grid cell and hour.

Three random forest models were constructed to test how the random forest approach performed with an increase in the size of the observed data set and an increase in the complexity of the predictions:

- Model 1 used April 2017 data from Maryland, with a random subset of 70% of the data used for training, and 30% of the data held back for testing.
- Model 2 used April and May 2017 in combination, again with 70% of the data used for testing.

- Model 3 tested a complete set of April and May 2017 data, and generated estimates on June 2017 Waze data. This model most closely represents the end goal of the SDI Waze project, which is to generate estimates of the number and pattern of EDT crashes when only Waze data are available.

The three models performed similarly. For Model 1, using 93,630 observations, the accuracy was 99%, but with precision of only 48%, and recall of 67%. This indicates that while overall the total number of correct zeros (no EDT crash expected) and ones (at least one EDT crash expected) was correct, there was not high precision in where these estimated EDT crashes were located in space and time. Given that Model 3 had very similar performance, even with 657,319 observations, it was clear that increasing the size of the training data set alone was unlikely to result in improved model performance.

4.2.1 Spatial grain

In the second iteration we tested three different scales of spatial aggregation (0.5 mile, 1 mile, and 4 mile area hexagon grid cells), and selected 1 mile grid size for subsequent work. Models at the larger spatial grain (4 square mile hexagonal grid cells) performed slightly worse than 1 mile grid cells, while the smaller spatial grain (0.5 square mile hexagonal grid cells) performed slightly better in most metrics. The small performance gain at the 0.5 mile grain size in some cases comes at the cost of substantially longer run time for data preparation. With approximately twice as many grid cells for the 0.5 mile compared to the 1-mile area grain size, the computational time to prepare weather variables was around 36 hours per month of data, compared to about 6 hours, respectively. The model fitting processes also took approximately 50% longer to complete with the smaller grid cells. . When training models over multiple states, for multiple months of data, such performance differences could be barriers to success, depending on the computational resources for a given application.

For future analyses, performance gains need to be weighed against the higher computational resources needed for data aggregation and modeling. The subsequent models all use 1 square mile grid cells, but we will revisit the potential gain from a smaller spatial resolution in future analyses. Note that of the evaluation metrics, AUC was very high for nearly all models. This is driven by the high accuracy of these models, especially in estimating the true zeros (times and locations where no EDT crash occurred).

Table 4. Summary of spatial grain testing for Waze-EDT random forest models.

Time period and test data	Model number and spatial grain	Accuracy	Precision	Recall	False Positive Rate	AUC
April 70/30	01 - 1 mile	98.39	57.23	65.86	0.97	0.9907
April 70/30	04 - 4 mile	97.65	56.83	67.93	1.49	0.9876
April 70/30	06 - 0.5 mile	98.51	56.52	66.61	0.92	0.9907
April-May, test June	03 - 1 mile	98.4	54.21	66.97	1.03	0.9900
April-May, test June	05 - 4 mile	97.72	54.9	68.28	1.50	0.9876
April-May, test June	07 - 0.5 mile	98.56	53.58	66.99	0.93	0.9909

4.2.2 Neighbors

One approach to addressing spatial dependence is to consider the counts of Waze events in neighboring grid cells as additional predictors of an EDT-level crash occurring. Addition of neighbors drove minor increases in performance at both the one-month (April) and three month (April-May, test on June) data sets. Recall (minimization of false negatives) was higher when neighboring grid cells were included as predictors. Neighboring grid cells are useful additional predictors, as Waze-reported events like traffic jams or stalled cars are likely to affect near-by areas

4.3 Iteration 3: Identifying the best performing models

For the third iteration, we expanded the data sets to six months (April to September, 2017, across Maryland), and incorporated additional supplemental data. Dividing Maryland into 1 square mile grid cells, and using each hour of this time period, a total of over 2 million observations were included in the data. These data were split into training and test sets by grid cell and hour, with models trained on 70% the data, and tested on the remaining 30%.

A total of 48 models were assessed to address the inclusion of additional supplemental data variables not used in the initial models, namely count of historical fatal crashes from FARS and AADT. Models were designed to sequentially test each supplemental variable in turn, and then to combine them in series. For a full set of models, see Appendix 3. All models exhibited excellent accuracy and overall performance.

The best performing set of models focused on Waze event “types” without the “sub-types”. The sub-types are not present for as many as half of all events in the Waze data, depending on event type. Excluding sub-types improved model fit over all. The base model (Model 24, in Table 5) included only Waze event type variables.

Excluding Waze event types and only including sub-types drove a substantial degradation in model performance (not shown; see Appendix 3 for full set of models). Recall dropped substantially, and model fit was clearly lower than Set A or B models. This finding demonstrates that models should be constructed with either a combination of Waze types and sub-types, or just Waze event types.

Including additional variables from the Waze data marginally improved recall, but also slightly increased the false positive rate. These additional variables include the confidence and reliability metrics provided by Waze, as well as the median direction of travel, and counts of Waze events in surrounding grid cells (neighbors). Sequentially adding FARS, weather, HPMS, and LEHD sets of variables all served to improve recall over the base model without these variables. Road functional class and AADT again were the predictors which drove the largest increase in recall.

Including all of these variables together (Model 30, in Table 5) provided the best balance between recall and false positive rate, with an AUC of 0.9914. This model is the focus of the remaining presented results.

Table 5. Summary of best performing models for April-September 2017, Maryland. These models included alert 'Types' (accident, hazard, jam, or road closure) but not 'Sub-types' (e.g., major or minor accident).

Model number and supplemental data included	Accuracy	Precision	Recall	False Positive Rate	AUC
24 – Type Base	98.38	54.79	52.42	0.78	0.9866
25 – Type Neighbors	98.39	55.36	60.14	0.9	0.9897
26 – Type FARS	98.36	54.03	53.35	0.82	0.9858
27 – Type Weather	98.35	53.97	54.12	0.84	0.9875
28 – Type Road + AADT	98.38	54.25	62.44	0.96	0.9889
29 – Type Jobs	98.37	54.33	61.75	0.95	0.9897
30 – Type FARS, Weather, Road, AADT, Job	98.47	57.04	61.1	0.85	0.9914
32 – 30, minus EDT-only	98.44	56.5	62.12	0.89	0.9913
33 – 31, minus road closure only	98.44	56.13	63.52	0.92	0.9912

In addition, approximately 1.5% of the data included times and locations where only EDT values, but no Waze accidents, were present. These account for 30,875 of the 2,003,391 observations for the April-September 2017 data. Omitting these observations had a negligible effect on the overall model performance, with minor increase in recall and decrease in precision, leading to an overall minor decrease in AUC.

An even smaller quantity of data included times and locations where only Waze road closure events were reported, 2,892 of the total observations (0.1%). Omitting these interestingly lead to the highest recall of this set of model, but the best balance between recall and precision was still achieved with Model 30, including those observations.

4.3.1 Model accuracy over space

Over the state of Maryland, looking across the six months of data, the random forest models of estimated EDT-level crashes closely matched the observed EDT crash reports (Figure 4). Each 1-mile area cell shows the percent of observed EDT-level crashes estimated by the model. The dark blue grid cells indicate that at least one Waze event was reported, but no EDT-level crashes were reported (true positives). Light blue grids are locations where over six months, the model underestimates the number of EDT-level crashes (false negatives) and orange grids are locations were the model overestimates the number of EDT-level crashes (false positives). The inset table summarizes the model accuracy by grid cell, showing the majority of grid cells have zero difference between observed and estimated EDT-level crashes; the model is tuned slightly towards overestimation.

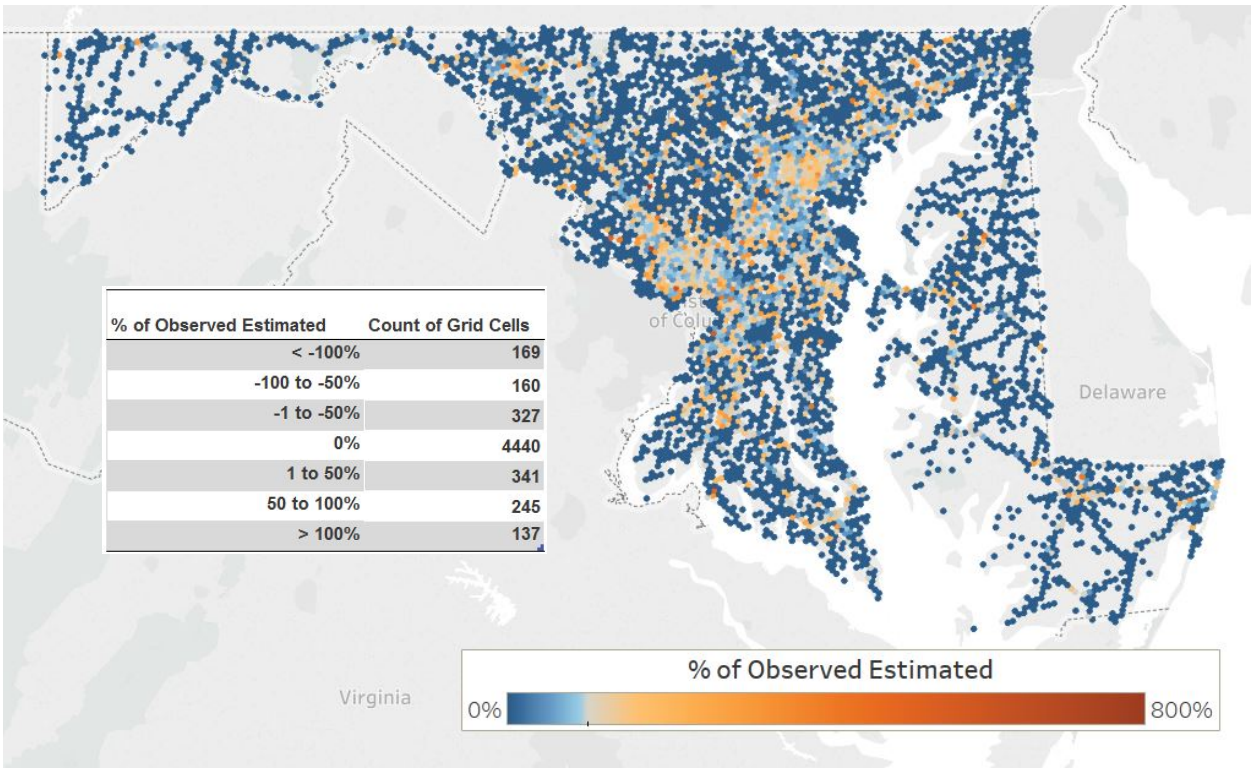

Figure 4. Model 30 outputs for estimated EDT-level crashes, April-September 2017 in Maryland. Percent of observed EDT crashes which are estimated by the model is shown graphically on the map, and in tablular format in the inset.

4.3.2 Model accuracy over time

Estimated EDT-level crashes were accurate by hour of day and day of week (Figure 5). In general, there is a large surge of Waze accidents which are reported during commuting hours, and many fewer user reports in early morning hours. The random forest modeling approach addresses such bias by training the crowdsourced data against the observed, ground-truth EDT data. However, not all the bias is

removed. By time of day, model estimates vary from 77% of the observed EDT crashes being reflected in the estimated EDT crashes in early morning. During commuting times, 111% of the number of observed EDT crashes appear in the estimated EDT crashes, representing a slight over estimate. By day of week, the biases are much smaller, with all observed and estimated EDT crashes matching with 2%. The time series plot shows on a daily basis the degree of matching between observed and estimated EDT crashes

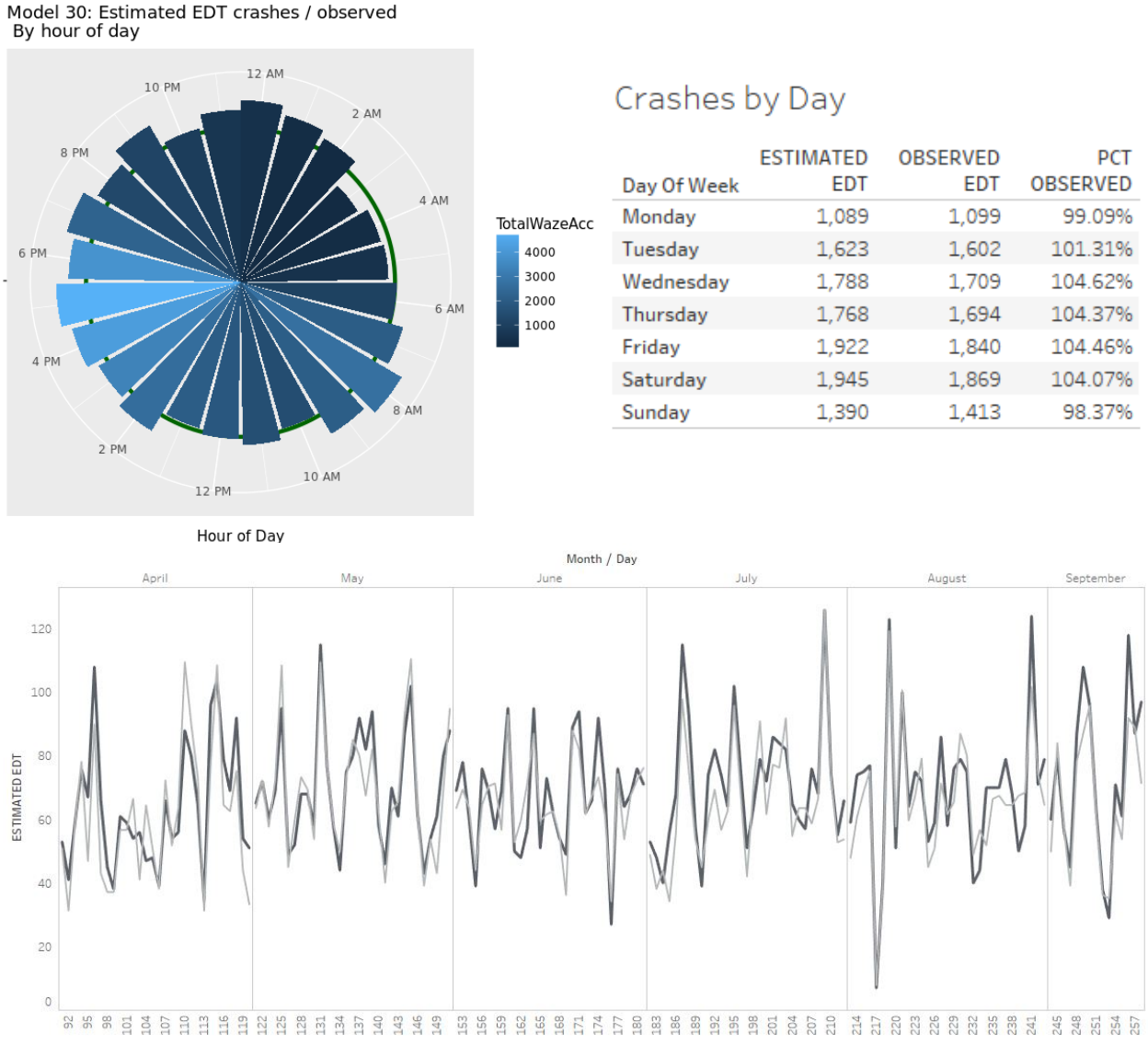

Figure 5. Model accuracy by hour of day (top left), day of week (top right), and day across six-month period (bottom).

4.3.3 False positives

The 'false positives' indicate times and places where the model, based on the crowdsourced data and all the auxiliary data, estimated that there was at least one EDT-level crash in this area, at that hour, but no crash was reported in the EDT data. This may be overestimation, or it may reveal how crowdsourced data can fill in reporting gaps in EDT.

The 2-11% overestimations shown in the model performance by time of day are one way these false positives appear. Since the ground-truth EDT data is based on police accident reports, several steps are needed before a report a crash appears in the EDT data: The crash must be called in to the police or otherwise observed, it must have been severe enough to warrant a police presence, the responding police must fill out a report, and the report must enter the state's PARs records. In Maryland, any crash resulting in an injury and fatality must be reported; property-damage only crashes may receive a police accident report[7]. In contrast, user reports that Waze assigns a sufficiently high reliability and confidence score all appear in the data stream. Therefore, it is possible that a portion of the 'false positives' represent true crash events, reported in Waze data, but not included in EDT because they did not meet the threshold for police-reportable crashes, or the crash was not reported to the police.

5. Conclusions

Using six months of data for Maryland in 2017 for this pilot study, we found that models could identify the number of actual police-reported crashes with high accuracy. The specific spatial and temporal patterns of the estimated crashes from the models is close to the ground-truth PARs data, but not exact. All models showed very high accuracy and excellent performance according to AUC. Even a single month of data presents a rich source of information for building an accurate crash estimation model. For the Maryland data, going to longer time frames and testing on novel data (training on April + May, testing on June) showed similar overall model performance as the more simple single month of data (April only). Increasing the data coverage to six months allowed a more comprehensive estimation of EDT-level crashes, but model accuracy was still relatively high in the smaller models. Additional data improved different aspects of the model. For example, weather data reduced false positives, while road functional class and jobs data reduced false negatives.

In phase 1 of the SDI Waze pilot, Volpe successfully integrated different sources of transportation data that were not intended to address safety questions in a secure cloud environment. We have also shown that Waze data can be used to generate reasonable EDT-level crash count estimates over space and time, when users of the Waze app are active. During the day, on higher functional classification roadways such as interstates, and at commuting times, the volume of Waze data is largest. In these conditions, we expect the model estimates based on crowdsourced data to accurately reflect the pattern of crashes on roadways. In early morning hours, and in locations where there are few users of the Waze app, the model estimates based on crowdsourced data will underestimate crashes. The approach described in this pilot captures a wider range of crashes than are reported in the police accident report data, including minor crashes which might not ordinarily require a police presence, but can seriously impact congestion. Extrapolating from these retrospective models, providing incoming Waze data could potentially be used to quickly identify actual, police-reportable crashes that occurred in close to real-time. Near-real time crash predictions can help emergency responders, TMCs and law

[7] Maryland Code, § 20-107

enforcement proactively allocate resources to locations with the highest crash likelihood (e.g., traffic crashes are most likely at these 8 interchanges from 3-6pm during icy conditions).

In phase 2 of the SDI Waze pilot, Volpe will support OST-P to assess specific applications of the crash estimation models developed in phase 1 to traffic safety questions. For example, the models could be used to track estimated crash counts in specific areas over time and flag anomalous patterns. The models could support a tool to detect anomalous crash trends, by comparing the estimated crash counts with observed EDT crashes and reported Waze crashes. With further model development, the Waze data may also prove useful for cross-state comparisons of specific traffic safety indicators such as incident duration, clearance times, secondary crashes, and short-term intervention assessment. In phase 2 of the SDI Waze pilot project, Volpe will work with OST-P to identify State or local DOT partner(s) and support case studies that apply Waze data insights to address operational or transportation safety problems.

The transportation sector is grappling with how best to use the ever increasing data availability that comes from the private sector in order to identify, measure, and diagnose issues, and develop solutions to improve the way it does business. This information includes not just Waze data, but also data from connected and automated vehicles and rideshare companies. By getting first hand experiences in these new techniques and data, USDOT is positioning itself to better leverage the rapidly changing technological world to positively influence safety.

6. References

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. CRC press.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

Chang, L. Y., & Wang, H. W. (2006). Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis & Prevention*, *38*(5), 1019-1027.

Friedman, J., Hastie, T., & Tibshirani, R. (2009). *glmnet*: Lasso and elastic-net regularized generalized linear models. *R Package Version*, *1*(4).

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1.

Kwon, O. H., Rhee, W. & Yoon, Y. (2015). Application of classification algorithms for analysis of road safety risk factor dependencies. *Accident Analysis & Prevention* 75**,** 1–15.

Liaw, A., & Wiener, M. (2002). Classification and regression by *randomForest*. *R News*, *2*(3), 18-22.

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice* 44 (5). Elsevier: 291–305.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Appendix 1: Data Pipeline

The project described in this report relied on two innovative cloud computing platforms recently established within DOT, the Secure Data Commons (SDC)[8] and Analytics Technology Architecture prototype (ATA). In the early stages of the project, the two system worked in conjunction with each other to curate and perform the analysis of the Waze data. In the later stages, the analytical capabilities of the ATA were incorporated into the SDC to create a single platform to curate and perform the analysis.

DOT established the SDC to serve as a data warehousing and analytics platform, for datasets such as the Waze data. The SDC provides secure, cloud-based, revocable access to complex (e.g. high volume, near-real time) and sensitive datasets in concert with analysis tools and shared computing resources. ATA served as a pilot focusing on conducting complex data analysis on a cloud computing platform; the functionality of ATA has now been incorporated within SDC.

The data processing steps described below were conducted on the SDC platform. The data analysis steps for derived Waze data, namely the random forest modeling, were initially conducted on the ATA platform. The roles of the two platforms can be summarized as follows:

Secure Data Commons (SDC)	**Analytical Technology Architecture (ATA)**
• Receives incoming Waze data • Curates data for users • Provides computing environment for integration, transformation, and aggregation of data	• Connects to SDC for derived Waze data • Houses additional data, including geospatial, census, and weather data • Provides computing environment for statistical analysis tasks

[8] https://portal.securedatacommons.com/.

The pipeline developed to translate incoming crowdsourced Waze data to insights on roadway crashes is summarized in Figure 6. The top portion of the figure describes the data and processing flow, and the lower portion identifying the key software tools.

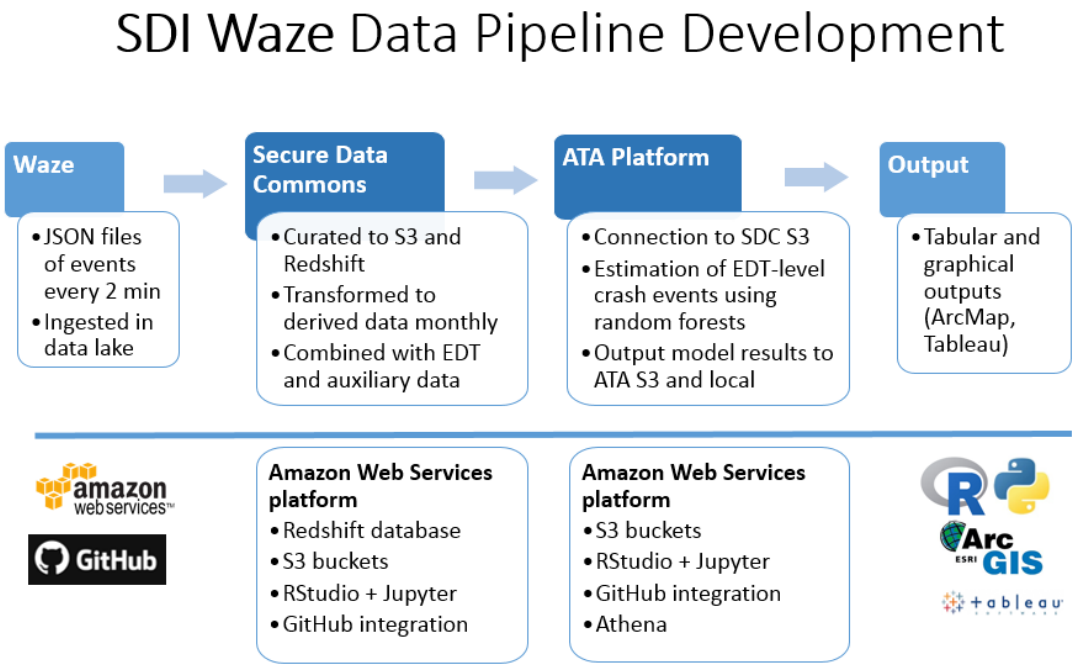

Figure 6. Data pipeline overview. S3 is the AWS Secure Simple Storage service, and provides network directories to store data within both SDC and ATA. Redshift is the AWS relational database service, and houses the curated Waze data in SDC. RStudio and Jupyter Notebooks are development environments for creating, testing, and running code in R or Python.

In this pipeline, SDC serves two roles: ingest and curate incoming Waze data, and host a computational platform to aggregate these data into a sufficiently derived form to be transferred to a secure storage on the ATA platform. In the initial stages, the ATA platform had been the platform for the statistical analysis (random forest modeling), as well as the production of model outputs for users.

With ongoing development of the Secure Data Commons, these key features of the ATA platform have now been incorporated, including:

- Rapid re-sizing of computational instances (increase the working memory, RAM, and the number of cores, CPUs) to run larger models as needed
- Import of supplemental data sets, such as FARS, weather, census, and HPMS roadway feature data
- Export derived analysis products, such as figures, tables, and dynamics reports
- Expansion of the analytical tools within the system

With these features in place, all analysis steps can now be completed within SDC.

Secure Data Commons has three features especially useful for the SDI Waze pilot:

- Securely stores the Waze data in as an incoming stream of JSON files in an S3 bucket and the subsequent curation of that data into form that facilitates subsequent analysis.
- Provides a persistent workspace for each user in an EC2 instance.

- Hosts multiple development environments, including Jupyter Notebook and RStudio, to carry out data aggregation, analysis and derivation using Python and R.

Appendix 2: Data Processing

This section describes how data are processed for the modeling. The following formatting conventions are used to distinguish data files, paths, and code files:

- `Data file / data object`
- `Path to a directory`
- *`Code`*

The input data for phase 1 of the SDI Waze pilot is from 5-minute increment Waze event data, from an area including Maryland. Waze events can repeat over time, especially jams or road closures.

The Waze data are available in the Secure Data Commons[9]. Geospatial data processing code is largely written in SQL and Python, and data analysis code is largely written in R and stored on the GitHub repository[10]. This repository is private, and is accessible to Volpe collaborators and selected other collaborators within DOT by request.

The processing workflow is summarized as follows:

1. Prepare data frames of unique Waze events, for each month of available data, from data stored in Redshift on SDC (*`ReduceWaze_SDC.R`*).
2. Clip these data to only those event within ½ mile of the border of the selected state (*`Waze_clip.R`*).
3. Match Waze and EDT events in space and time (*`Space_time_match.R`*).
4. Overlay spatial layers, including urban areas and hexagonal tessellation (*`UrbanArea_overlay.R`*).
5. Aggregate linked Waze and EDT data to 1 square mile grid cells, for each hour of the time period of interest (*`MakeSpaceTimeGrids.R`* and *`Grid_aggregation.R`*).

A2.1 Waze data

Origin

The data are accessed only via the Secure Data Commons (SDC). The raw data arrive in JavaScript Object Notation (JSON) format, and are processed to CSV format within SDC as part of the data ingestion process. The structure of these data and fields are described in the Waze Traffic Data Specification Document, Version 2.7.1 (Waze_Traffic_Data_Spec.pdf).

The data arrive in three tables for every state, every two minutes. The three tables are *alert*, *irregularity*, and *jam* data. For this pilot, we focus only on the *alert* table, which provides aggregated user reports of accidents, hazards, jams, as well as road closure reports from Waze. A single alert can persist over time (i.e., an accident can take time to be cleared, and a hazard can persist over several hours). Alerts are identified by universally unique identifiers (UUIDs), which can be used to create a data file of an individual event, all relevant characteristics of that event, and the duration. This work is done by the *`ReduceWaze_SDC.R`* script and the data are located in SDC (portal.securedatacommons.com).

[9] https://portal.securedatacommons.com
[10] https://github.com/VolpeUSDOT/SDI_Waze

Notes

Waze data were originally delivered as JSON files in 5-minute intervals, zipped into monthly directories. The pilot work shifted to SDC as that platform was developed, and then used the curated data stored in Redshift, which had already been converted from JSON files.

Monthly Files

Monthly Waze data were compiled for each state of interest, by aggregating to one row per UUID. The monthly files for Waze are also clipped to a 0.5 mile buffer around the state of interest. The working data are housed on SDC.

Origin

Redshift relational database on SDC.

Processing

ReduceWaze_SDC.R aggregates monthly .RData and .csv files.

Waze_clip.R filters the Waze data to the Maryland polygon, buffered to 0.5 miles. *Buffer_state.R* for the script to create the buffered polygon from the Census Bureau shapefiles.

Notes

For example, `MD__buffered_2017-04.RData` is one of the monthly file outputs. This has a data frame of Waze events, collapsed to a single UUID, for April 2017, for all event in Maryland, including a 0.5 mi buffer around the state.

This file also includes the relevant EDT events, see below. Waze and EDT events are linked using *Space_time_match.R* and *wazefunctions.R*. The linking was done using 0.5 mi radius and +/- 60 minute windows around EDT events, and produces `EDT_Waze_link_April_MD.csv`.

A2.2 EDT

Origin

These data were delivered from NHTSA and have been uploaded to SDC for analysis.

Processing

The EDT dataset had several issues that required adjustment prior to processing. All of the issues identified were in the main data file, `1_CrashFact.csv`. The original file was preserved and a new file with the identified solutions is saved as `1_CrashFact_edited.txt`.

Field Name	Issue	Solution	New Field Name
N/A	Tick marks (`) were inserted in character fields as part of a name.	Replaced the tick marks with apostrophes (‘).	N/A
CrashDate, UpdateDate, CreateDate, CrashFactDT, LastModifiedDT	Datetime columns in 12-hour format.	Update datetime columns to 24-hour format.	N/A

CrashDate, HourofDay, MinuteofDay	There is no datetime field with the date and the exact time of the crash. The time component is separated into multiple fields.	Combine the date from the CrashDate and join the Hour of Day and Minute of Day fields to get the whole structure.	CrashDate_Local
CheckSum	Field was non-numeric data type.	Change field to numeric.	N/A
GPSLong	Longitude values were in Eastern hemisphere.	Multiplied the values by -1 to place the locations in the correct hemisphere.	GPSLong_new
GPSLat, GPSLong, GPSLong_new	Decimal places were not persevered in the Lat/Lon fields.	Change the field structure to have 8-decimal places.	N/A

Monthly Files

A subset of the EDT data for April 2017, for Maryland, has been used for analysis with Waze events.

Origin

The made from the `1_CrashFact_edited.txt` file as the input, and `2017-04_1_CrashFact_edited.csv` and `.RData` is the output.

Processing

`Waze_clip.R` creates the monthly EDT files.

A2.3 Gridded data

Both Waze and EDT data exist as point datasets, with a single spatial location and either a single (EDT) or range (Waze) of time values. For each month, these files are stored as `wazeTime.edt.hex.XX.RData`, where XX represents the numeric month.

Origin

These are generated from the `merged.waze.edt.XX_MD.RData` files, which are produced by the *`UrbanAreas_Overlay.R`* script, in combination with the hexagonal grid tessellation from `spatial_layers/MD_hex.RData`.

Processing

These files are grid aggregated for each hour of the target month. The script *`Grid_aggregation.R`* carries out these steps. In short, the process involves the following steps:

- Reading in the merged Waze / EDT data, organized by Waze UUID and EDT CrashID, separately, as well as the spatial layer for the hexagonal-shaped grid cells.
- Creating an expanded data frame for day, hour, and grid ID for the target month.
- Populating this data frame with the count of Waze events, the count of EDT crashes matching Waze events, the count of EDT crashes matching Waze accidents, and a large number of other aggregated variables from the Waze data.

A2.4 Random Forest modeling

The outputs from the data processing steps are modeled using RF, with the *randomForest* package in R. Over 50 models were tested, and are summarized in Appendix 3.

Origin

Inputs are the `WazeTimeHexWx.Rdata` files as the input, and `Model_*_RandForest_Output.Rdata` is the output.

Location

Model outputs have been saved internally on the ATA platform, in an S3 directory.

Processing

Random forest modeling is done using a series of functions written in *wazefunctions.R*. These functions prepare the data, including adding any supplemental data such as weather or census information for that grid cell, in that hour, and separates the data into training and test datasets if the model requires that. Some models were trained on complete months of data, and then tested on subsequent months, so the scripts accommodate that structure.

Notes

Output files for each model are saved, so that re-processing with alternative classification thresholds (crash or no crash, in the outcome estimated EDT variable) is possible without needing to re-run each model.

Appendix 3: Random Forest Models Tested

Forty-eight random forest models were tested in this Waze pilot. These increased the size of data, the complexity of the Waze data and complexity of supplemental data. Models were then compared using standard goodness of fit metrics for machine learning models, as well as examination of the output. Data size varied from a single month to six months of data, with tests either based on a 70/30 split of all data, or testing on a separate month of data.

Complexity of Waze data itself had several levels. For some models, all data were used, including all counts of Waze alerts (accidents, hazards, jams, road closures), the median values for the reliability of the reports, the direction of travel, and other metrics. Additional Waze data was incorporated for the counts of Waze accidents and jams in the neighboring grid cells around a given target cell, noted by the 'Neighbors' column in the table below. Complexity of Waze data also varied by just high-level alert types (four categories), or just the lower-level alert subtypes (11 categories).

Supplemental data included NEXRAD radar-detected reflectivity (Weather), miles of roadway by functional class from HPMS (Road), AADT from HPMS, selected variables from the LHED data set (Jobs), and 3 years of historical crash data from FARS. See the

Supplemental data section for more details.

Set	Set Description	Model Number	Data	Additional Data	Neighbors
Iteration 2	1 mile	01	April, 70/30	None	No
Iteration 2	1 mile	02	April-May, 70/30	None	No
Iteration 2	1 mile	03	April-May, test on June	None	No
Iteration 2	4 mile	04	April, 70/30	None	No
Iteration 2	4 mile	05	April-May, test on June	None	No
Iteration 2	0.5 mile	06	April, 70/30	None	No
Iteration 2	0.5 mile	07	April-May, test on June	None	No
Iteration 2	1 mile + Neighbors	08	April, 70/30	None	Yes
Iteration 2	1 mile + Neighbors	09	April-May, test on June	None	Yes
Iteration 2	4 mile + Neighbors	10	April, 70/30	None	Yes
Iteration 2	4 mile + Neighbors	11	April-May, test on June	None	Yes
Iteration 2	1 mile Additional Data	12	April, 70/30	Weather	Yes
Iteration 2	1 mile Additional Data	13	April-May, test on June	Weather	Yes
Iteration 2	1 mile Additional Data	14	April, 70/30	Weather + Road	Yes
Iteration 2	1 mile Additional Data	15	April-May, test on June	Weather + Road	Yes
Iteration 2	1 mile Additional Data	16	April, 70/30	Weather + Road + Jobs	Yes
Iteration 2	1 mile Additional Data	17	April-May, test on June	Weather + Road + Jobs	Yes
Iteration 3 A	All Waze	18	April-Sept, 70/30		No
Iteration 3 A	All Waze	19	April-Sept, 70/30	FARS	No
Iteration 3 A	All Waze	20	April-Sept, 70/30	Weather	No
Iteration 3 A	All Waze	21	April-Sept, 70/30	Road + AADT	No
Iteration 3 A	All Waze	22	April-Sept, 70/30	Jobs	No
Iteration 3 A	All Waze	23	April-Sept, 70/30	FARS + Weather + Road + AADT + Job	No
Iteration 3 B	Type Counts	24	April-Sept, 70/30		No
Iteration 3 B	Type Counts	25	April-Sept, 70/30		Yes
Iteration 3 B	Type Counts	26	April-Sept, 70/30	FARS	No
Iteration 3 B	Type Counts	27	April-Sept, 70/30	Weather	No
Iteration 3 B	Type Counts	28	April-Sept, 70/30	Road + AADT	No
Iteration 3 B	Type Counts	29	April-Sept, 70/30	Jobs	No
Iteration 3 B	Type Counts	30	April-Sept, 70/30	FARS + Weather + Road + AADT + Job	No
Iteration 3 B	Type Counts	31	April-Sept, 70/30		
Iteration 3 B	Type Counts	32	April-Sept, 70/30		
Iteration 3 C	Subtype Counts	33	April-Sept, 70/30		No
Iteration 3 C	Subtype Counts	34	April-Sept, 70/30		Yes
Iteration 3 C	Subtype Counts	35	April-Sept, 70/30	FARS	No
Iteration 3 C	Subtype Counts	36	April-Sept, 70/30	Weather	No

Set	Set Description	Model Number	Data	Additional Data	Neighbors
Iteration 3 C	Subtype Counts	37	April-Sept, 70/30	Road + AADT	No
Iteration 3 C	Subtype Counts	38	April-Sept, 70/30	Jobs	No
Iteration 3 C	Subtype Counts	39	April-Sept, 70/30	FARS + Weather + Road + AADT + Job	No
Iteration 3 C	Subtype Counts	40	April-Sept, 70/30		
Iteration 3 C	Subtype Counts	41	April-Sept, 70/30		
Iteration 3 D	Test of data structure: EDT Counts vs binary	42a	April-Sept, 70/30	FARS + Weather + Road + AADT + Job	No
Iteration 3 D	Test of data structure: EDT Counts vs binary	42b	April-Sept, 70/30	FARS + Weather + Road + AADT + Job	No
Iteration 3 D	Test of data structure: EDT Counts vs binary	43a	April-Sept, 70/30	FARS + Weather + Road + AADT + Job	No
Iteration 3 D	Test of data structure: EDT Counts vs binary	43b	April-Sept, 70/30	FARS + Weather + Road + AADT + Job	No
Iteration 3 D	Test of data structure: EDT Counts vs binary	44a	April-Sept, 70/30	FARS + Weather + Road + AADT + Job	No
Iteration 3 D	Test of data structure: EDT Counts vs binary	44b	April-Sept, 70/30	FARS + Weather + Road + AADT + Job	No

Appendix 4: Annotated Bibliography of Machine Learning and Crowdsourced Data in Highway Safety Research

This annotated bibliography summarizes recent peer-reviewed publications for research into highway safety relevant to the SDI Waze pilot. This bibliography focuses on three themes: machine learning approaches in transportation safety research, the use of crowdsourced data in highway safety research, and the use of sophisticated spatial regression for crash analysis.

A4.1 Machine learning approaches in transportation safety

Abbas and Machiani (2016): Modeling the dynamics of driver's dilemma zone perception using agent based modeling techniques

- In a driving simulation study, used Agent Based Models (ABM) to investigate driver impacts of driving in a 'dilemma zone', such as too close to an intersection to safely stop. Models include the MATsim dynamic agent-based traffic simulation model.

Bahouth, Digges, and Schulman (2012): Influence of injury risk thresholds on the performance of an algorithm to predict crashes with serious injuries

- Optimization for first responders, using logistic regression on National Automotive Sampling System / Crash-worthiness Data System (NASS/CDS) to determine how variation in injury risk thresholds affects crash predictions. Standard logistic regression approach, where outcomes are binary for each type of crash (separate models, not ordinal).

Chiou, Lan, and Chen (2013): A two-stage mining framework to explore key risk conditions on one-vehicle crash severity

- This research combines data mining and a logistic regression approach to identify crash severity in one-vehicle crashes. Genetic mining rule (GMR) model developed, to identify 'rules' which correspond to variables most associated with risk of a crash. The variables were then used in a hierarchical logistic regression (mixed logit model) to identify road conditions associated with serious crashes.
- Similar to initially-proposed SDI Waze project approach, where random forests used to identify combinations of variables highly associated with EDT-level crashes, and then logistic regression used to assign probability of a crash to Waze events and test statistical significance. Use a training/validation approach for the rule-mining, 70% of data for training, 30% for validation.

Das et al. (2015): Estimating likelihood of future crashes for crash-prone drivers

- Logistic regression on 8 years of traffic crash data in Louisiana. Use road characteristics, human factors, collision type, and weather in the model; use model diagnostics to assess true positives, sensitivity, and false positive rate for model predictions. Use area under receiver-operator curve (AUC) to assess model fit. Can correctly identify responsibility of crash of 62% of crashes, with the response variable being "at-fault" true or false.

Delen et al. (2017): Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods

- Predictive analytics used for injury severity models. Refers to multinomial logistic regression (namely, ordinal logistic regression) as commonly used for injury severity analysis. Refer to previous work on FARS data to use logistic regression for estimating if a crash would be fatal (Liu and McGee 1988). Some studies have used combination of ordered probit, ordered logit, and multinomial logit in combination (Park et al. 2012). Here use machine learning methods: artificial neural networks, support vector machines, and decision trees as an ensemble to develop a ranking of risk factors for crash injury severity.
- Data from the National Automotive Sampling System General Estimates System (NASS GES), with 1% of all national automobile crashes, for 2011 and 2012. Approximately 25 predictors used. K-fold cross-validation used in model development, and models evaluated with AUC.
- Focus is on developing a ranking of risk factors, rather than estimating crash severity from new input data, differing from the goals of the SDI Waze project.

Gkritza et al. (2013): Empirical Bayes approach for estimating urban deer-vehicle crashes using police and maintenance records

- 150 highway sections in Iowa. Use the Empirical Bayes approach with zero-inflated negative binomial regression for frequency of deer-vehicle crashes. Average annual daily traffic (AADT) used as exposure for highway sections, following AASHTO 2010 Highway Safety Manual recommendations.
- Model produces rankings of which highway sections are most suitable for focused safety improvement, based on crashes per mile-year.

Gonzalez-Velez and Gonzalez-Bonilla (2017): Development of a Prediction Model for Crash Occurrence by Analyzing Traffic Crash and Citation Data

- Focus on human factors, such as traffic violation and crash history, in developing model of likelihood of crash occurrence at the driver level. Logistic regression approach, using Minitab, with model selection by AIC and assessment by AUC.

Kwon, Rhee, and Yoon (2015): Application of classification algorithms for analysis of road safety risk factor dependencies

- Severity of injury for accidents modeled from historical incident data in California, 2004-2010. Naive Bayes and decision tree (CART) used to identify risk factors of greatest importance; use logistic regression to compare the output of the two classification approaches. AUC for model assessment.
- Refer to other studies using decision tree (CART) approaches for injury severity modeling: Kashani and Mohyamany 2011, Montella et al. 2011, and others. Sohn and Shin 2001 compared ANN, logistic regression and CART for severity classification, finding each has similar classification accuracy.
- Differs from goals of SDI Waze in focusing on ranking risk factors, rather than producing estimated counts of crashes based on geospatial data. Found that the decision tree approach had best combination of true positive rate and false positive rate (AUC).

Lin (2015): Data science application in intelligent transportation systems: An integrative approach for border delay prediction and traffic accident analysis

- Thesis focusing on intelligent transportation systems (ITS) in general. Uses Seasonal Autoregressive Integrated Moving Average Model (SARIMA) and Support Vector Regression (SVR) to model traffic accident data. Use k-nearest neighbor (KNN) as well.

Lord and Persaud (2004): Estimating the safety performance of urban road transportation networks

Lord, Washington, and Ivan (2005): Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory

- In both publications, Lord et al. review common models for modelling crash count data: Poisson, Zero-inflated Poisson, and Zero-inflated negative binomial (also called Poisson-gamma) models. They point out that models which can account for zero-inflation, which can arise because of overly narrow time and space scale selection and rarity of crashes, often provide the best statistical fit, but may not characterize the underlying crash process completely.
- Provide detailed reviews of statistical theory behind these crash count models, and lay out how a zero-inflated model makes a simplifying assumption that a roadway can exist in either a 0-crash, 'perfectly safe' condition, or a non-zero crash, 'imperfectly safe' condition. They argue that having a too-small spatial or temporal scale can lead to over-estimation of the 'perfectly safe' condition.

Lord and Mannering (2010): The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives

- Excellent review of models used for crash frequency data. Discusses the commonly-used zero inflated negative binomial, as well as Poisson regression more generally. Discusses challenges with modeling crash frequency data, including overdispersion (variance exceeding the mean), correct choice of time window, and temporal and spatial correlation.
- Refers to common approaches for dealing with temporal and spatial correlation. GEE, GAM, and random effects (hierarchical) models also discussed.
  Machine learning models are briefly discussed, including neural networks and support vector machine models.

Morgan (2013): Performance Measures for Prioritizing Highway Safety Improvements Based on Predicted Crash Frequency and Severity

- Thesis on crash frequency modeling based on incident features, roadway infrastructure, demographic, and roadway network flow data. Estimate crash severity in scenarios of differing infrastructure and demographic change.
- Ordered probit model for crash frequency, which is an unusual application.

Pal et al. (2016): Factors influencing specificity and sensitivity of injury severity prediction (ISP) algorithm for AACN

- Use NASS CDS database of US vehicle accidents, 2005-2012, using a 'branching logistic regression' approach for modeling occurrence of minor or serious injury for crashes. Similar in some respects to a decision tree approach.
- Crash-level estimations of severity are the focus, rather than the number, pattern, and severity of crashes. Crash-level features include speed, impact direction, seat belt use, age, and gender.

Pande, Nuworsoo, and Shew (2012): Proactive Assessment of Accident Risk to Improve Safety on a System of Freeways

- Four freeway corridors selected, and historical crash data 2010-2011 assessed in combination with real-time traffic patterns. Logistic regression and decision trees (CART) used to assess crash or non-crash outcomes.
- Data aggregation and preparation discussed. Includes a useful literature review.
- Similar in some respects to goals of SDI Waze project, but using different data sets and with a different geographic and temporal scope.

Saha, Alluri, and Gan (2015): Prioritizing Highway Safety Manual's crash prediction variables using boosted regression trees

- Decision tree (boosted regression tree, BRT, similar to random forests, and also based on CART) approach used to evaluate the impact of individual roadway characteristics on crash predictions. The goal here was to rank roadway characteristics, to prioritize which variables should be the focus of data collection, when resources are limited for roadway monitoring. Roadway characteristics are the input for the Highway Safety Manual (HSM) empirical Bayes approach to estimating crash frequency with negative binomial regression.
- Five years of data (2008-2012): in Florida used. BRT are similar to random forests, in using an ensemble of decision trees, and can be useful when most individual decision trees produce weak statistical predictions. Implemented in *gbm* package in R.

Saleem, Asa, and Membah (2016): An Exploratory Computational Piecewise Approach to Characterizing and Analyzing Traffic Accident Data

- Six years of data (2008-2013): in North Dakota from state sources. Large number of crash-level data used. Only data analysis is fitting polynomial functions to the bivariate patterns, no statistical inference.

Shawky and Al-Ghafli (2016): Risk Factors Analysis for Drivers with Multiple Crashes

- Identifying high-risk drivers using demographic characteristics, historical violations, and specific violation types with negative binomial regression. Crash estimation model identifies the set of predictors most strongly associated with high-risk drivers. Standard regression approach, models evaluated by AIC.

Srinivasan et al. (2015): Crash Prediction Method for Freeway Facilities with High Occupancy Vehicle (HOV) and High Occupancy Toll (HOT) Lanes

- Segment-based crash frequency modeling, separate models for fatal/injury crashes and all crashes. Negative binomial regression approach, with AADT as exposure variable, segment length and number of lanes as additional important variables. Data from three states, CA, WA, and FL, from the Highway Safety Information System (HSIS). Models were run in SPSS, and spreadsheet tool developed using the fitted coefficients.

Sun, Das, and Broussard (2016): Developing Crash Models with Supporting Vector Machine for Urban Transportation Planning

- Support vector machines (SVM) unsupervised learning approach to discover patterns in crash frequency. Data from Louisiana urban roadways in 2011-2013, with crash frequency, roadway geometry, and AADT as main inputs. Little detail on model specification or application provide, largely a demonstration that SVM can be used.

K. Wang (2016): Exploration of Advances in Statistical Methodologies for Crash Count and Severity Prediction Models

- PhD thesis from University of Connecticut, focusing on the simultaneous estimation of injury severity and vehicle damage using regression models. Simultaneous estimation is done by "copula based models"", and finds high correlation between injury and vehicle damage. Spatial analysis of road intersections and segments using socio-economic variables. Thirdly, carried analysis of crash type and crash severity on rural two-lane highways, using a multivariate Poisson lognormal model.
- Crash type and severity were better predicted by the multivariate Poisson lognormal than by negative binomial or univariate Poisson lognormal models.

Wei et al. (2017): Analyzing Traffic Crash Severity in Work Zones under Different Light Conditions

- Focus on work zones in Tennessee, 2003-2015, to assess factors determining crash severity (not count). Use Classification and Regression Trees (CART) to show importance of light conditions in crash severity, as well as roadway geometry factors, driver factors, and environmental factors (e.g., Weather, clear or not clear).
- Highest proportion of injury crashes for head-on collisions, along roadways, with greater than two lanes. Create three decision trees, one for each of the light conditions, and compare results. For instance, traffic control devices were effective in reducing crash severity in daylight and dark-lighted, but not dark-not-lighted conditions.

Xie, Lord, and Zhang (2007): Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis

- Analysis of rural roads in Texas, comparing two types of neural network machine learning models, and a negative binomial regression model. Suggest that the Bayesian neural network is a useful approach for estimating crash counts in rural highways.
- Reviews the limitations of regression model approaches: need for clearly defined function relating crash frequencies and explanatory variables. Neural networks do not require *a priori* specification of a functional form relating these variables. Such models have however been criticized for over-fitting data and resulting in models without interpretable coefficients for explanatory variables. The Bayesian approach to a neural network can alleviate the former concern.
- Using a training/testing framework, neural networks outperformed negative binomial regressions for crash counts, with predictors of segment length, vehicles per day, shoulder width, and lane width.

A4.3 Crowdsourced data analysis approaches

Masino et al. (2017): Learning from the crowd: Road infrastructure monitoring system

- Data collected automatically from new vehicles used as input for a decision tree analysis of road condition. Data collection relies on GPS, Wi-Fi, and sensors of vertical acceleration and pitch rate to detect features such as potholes.

Vasudevan et al. (2016): Predicting Traffic Flow Regimes From Simulated Connected Vehicle Messages Using Data Analytics and Machine Learning

- Simulated data from a highway corridor in Seattle, to model traffic flow regimes under different conditions for connected vehicles. Three machine learning approaches were taken for traffic flow estimation: logistic regression, individual decision trees (CART), and random forests. Models were run in a Microsoft Azure cloud computing environment, using Apache Spark machine learning libraries.

- Focus is on connected vehicle configuration, operational conditions, market penetration, and estimating traffic flow rather than estimating crash counts. Useful detail on feature extraction, relying on principal component analysis (PCA) in R.

A4.2 Spatial regression for road safety

Gill et al. (2017): Comparison of adjacency and distance-based approaches for spatial analysis of multimodal traffic crash data

- Model of spatial correlation for traffic crash counts at the county level. Develop two Bayesian models to look at how much adjacency explains in crash counts. 58 counties in California for 2012. Exposure variable of daily vehicle miles traveled (DVMT) from Highway Performance Monitoring System (HPMS) of FHWA.
- Poisson model for crash counts, with errors drawn from a normal distribution. Spatial autocorrelation is built in via the hyperparameter Σ, the covariance matrix which is used as the standard deviation of the error u_{ij}, using multivariate conditional auto-regressive (MCAR) model. Do not specify the modeling tool, but Stan likely used.

K.A. Rhee et al. (2016): Spatial regression analysis of traffic crashes in Seoul

- Road segment based analysis for traffic crashes in Seoul, Korea, in 2010, using geographically weighted regression to account for spatial autocorrelation. Discuss conditional autoregressive (CAR) model, but end up using geographically weighted regression. Use Moran's I to assess strength of spatial autocorrelation. Use AIC to evaluate competing models.

Schultz (2015): Use of Roadway Attributes in Hot Spot Identification and Analysis

- Analysis of Utah "hot spots" for crashes, adding detailed roadway attribute layers such as vertical sag and grade to traditional variables such as lane width, number of lanes, shoulder width, and horizontal curvature. Use a hierarchical Bayesian Poisson mixture model.
- Use Bayesian horseshoe method for variable selection. This approach can take in a large number of possible variables, and assign a coefficient of zero to those which are unimportant. Lasso and ridge regression techniques serve a similar purpose in logistic regression models. Once variables were selected, a Bayesian Poisson regression was done on segments, using non-informative priors.
- Areas where many segments have observed crashes much greater than predicted crashes are considered hot spots. A number of specific hot spots are examined in detail.

Xu, Kockelman, and Wang (2014): Modeling crash and fatality counts along mainlines and frontage roads across Texas: The roles of design, the built environment, and weather

- Analysis of Texas highways, using spatial data on traffic, demography, land use, population and job density, rainfall, income, and education. Compare zero-inflated negative binomial, zero-inflated Poisson, and negative binomial models, finding the first preferred.
- Fully-spatial analysis (e.g., conditional autoregressive analysis) can be intractable for very large data sets, so segment-based analysis is typically used.
- Use 50-year average rainfall as the weather variable. Separate analysis for main lanes and frontage roads. Population density and job densities found to be the strongest predictors of crash counts, along with urbanization. Age and income have negative effects; average rainfall slightly positive.

Zeng and Huang (2014): Bayesian spatial joint modeling of traffic crashes on an urban road network

- Poisson, negative binomial, and conditional autoregressive (CAR) models used to model crash counts at intersections and along road segments. A combination of spatial approaches to join intersections and segment models, with the segment models having traditional crash frequency modeling. Presents one way to approach fully spatially-explicit modeling of crash frequency on a road network, but too data-intensive to be useful for SDI Waze project.

A4.4 References

Abbas, Montasir M, and Sahar Ghanipoor Machiani. 2016. "Modeling the Dynamics of Driver's Dilemma Zone Perception Using Agent Based Modeling Techniques." *International Journal of Transportation* 4 (2). Science; Engineering Research Support Society: 1–14.

Bahouth, George, Kennerly Digges, and Carl Schulman. 2012. "Influence of Injury Risk Thresholds on the Performance of an Algorithm to Predict Crashes with Serious Injuries." *Annals of Advances in Automotive Medicine* 56. Association for the Advancement of Automotive Medicine: 223.

Chiou, Yu-Chiun, Lawrence W Lan, and Wen-Pin Chen. 2013. "A Two-Stage Mining Framework to Explore Key Risk Conditions on One-Vehicle Crash Severity." *Accident Analysis & Prevention* 50. Elsevier: 405–15.

Das, Subasish, Xiaoduan Sun, Fan Wang, and Charles Leboeuf. 2015. "Estimating Likelihood of Future Crashes for Crash-Prone Drivers." *Journal of Traffic and Transportation Engineering (English Edition)* 2 (3): 145–57. doi:https://doi.org/10.1016/j.jtte.2015.03.003.

Delen, Dursun, Leman Tomak, Kazim Topuz, and Enes Eryarsoy. 2017. "Investigating Injury Severity Risk Factors in Automobile Crashes with Predictive Analytics and Sensitivity Analysis Methods." *Journal of Transport & Health* 4 (Supplement C): 118–31. doi:https://doi.org/10.1016/j.jth.2017.01.009.

Gill, G, T Sakrani, W Cheng, and J Zhou. 2017. "Comparison of Adjacency and Distance-Based Approaches for Spatial Analysis of Multimodal Traffic Crash Data." *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences* 42.

Gkritza, Konstantina, Reginald R Souleyrette, Michael J Baird, and Brent J Danielson. 2013. "Empirical Bayes Approach for Estimating Urban Deer-Vehicle Crashes Using Police and Maintenance Records." *Journal of Transportation Engineering* 140 (2). American Society of Civil Engineers: 04013002.

Gonzalez-Velez, Enrique, and Armando Gonzalez-Bonilla. 2017. "Development of a Prediction Model for Crash Occurrence by Analyzing Traffic Crash and Citation Data." Transportation Informatics Tier I University Transportation Center, University at Buffalo.

Kwon, Oh Hoon, Wonjong Rhee, and Yoonjin Yoon. 2015. "Application of Classification Algorithms for Analysis of Road Safety Risk Factor Dependencies." *Accident Analysis & Prevention* 75 (Supplement C): 1–15. doi:https://doi.org/10.1016/j.aap.2014.11.005.

Lin, Lei. 2015. "Data Science Application in Intelligent Transportation Systems: An Integrative Approach for Border Delay Prediction and Traffic Accident Analysis." PhD thesis, State University of New York at Buffalo.

Lord, Dominique, and Fred Mannering. 2010. "The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives." *Transportation Research Part A: Policy and Practice* 44 (5). Elsevier: 291–305.

Lord, Dominique, and Bhagwant N Persaud. 2004. "Estimating the Safety Performance of Urban Road Transportation Networks." *Accident Analysis & Prevention* 36 (4). Elsevier: 609–20.

Lord, Dominique, Simon P Washington, and John N Ivan. 2005. “Poisson, Poisson-Gamma and Zero-Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory.” *Accident Analysis & Prevention* 37 (1). Elsevier: 35–46.

Masino, Johannes, Jakob Thumm, Michael Frey, and Frank Gauterin. 2017. “Learning from the Crowd: Road Infrastructure Monitoring System.” *Journal of Traffic and Transportation Engineering (English Edition)* 4 (5). Elsevier: 451–63.

Morgan, Noah S. 2013. “Performance Measures for Prioritizing Highway Safety Improvements Based on Predicted Crash Frequency and Severity.” PhD thesis, Auburn.

Pal, Chinmoy, Tomosaburo Okabe, Vimalathithan Kulothungan, Narahari Sangolla, Jeyabharath Manoharan, Wang Stewart, and John Combest. 2016. “Factors Influencing Specificity and Sensitivity of Injury Severity Prediction (Isp) Algorithm for Aacn.” *International Journal of Automotive Engineering* 7 (1). Society of Automotive Engineers of Japan, INC: 15–22.

Pande, Anurag, Cornelius Nuworsoo, and Cameron Shew. 2012. “Proactive Assessment of Accident Risk to Improve Safety on a System of Freeways.” Mineta Transportation Institute; MTI Report 11-15.

Rhee, Kyoung-Ah, Joon-Ki Kim, Young-Ihn Lee, and Gudmundur F Ulfarsson. 2016. “Spatial Regression Analysis of Traffic Crashes in Seoul.” *Accident Analysis & Prevention* 91. Elsevier: 190–99.

Saha, Dibakar, Priyanka Alluri, and Albert Gan. 2015. “Prioritizing Highway Safety Manual’s Crash Prediction Variables Using Boosted Regression Trees.” *Accident Analysis & Prevention* 79. Elsevier: 133–44.

Saleem, Farhan, Eric Asa, and Joseph Membah. 2016. “An Exploratory Computational Piecewise Approach to Characterizing and Analyzing Traffic Accident Data.” *International Journal of Scientific and Technical Research in Engineering*.

Schultz, Basset, Grant G. 2015. “Use of Roadway Attributes in Hot Spot Identification and Analysis.” Utah Department of Transportation; Brigham Young University-Provo.

Shawky, Mohamed, and Abdulla Al-Ghafli. 2016. “Risk Factors Analysis for Drivers with Multiple Crashes.” *International Journal of Engineering and Applied Sciences* 3 (11): 42–48.

Srinivasan, Sivaramakrishnan, Phillip Haas, Priyanka Alluri, Albert Gan, James Bonneson, and Paul Hiers. 2015. “Crash Prediction Method for Freeway Facilities with High Occupancy Vehicle (Hov) and High Occupancy Toll (Hot) Lanes.” Florida Department of Transportation.

Sun, Xiaoduan, Subasish Das, and Nicholas Broussard. 2016. “Developing Crash Models with Supporting Vector Machine for Urban Transportation Planning.” In *17th International Conference Road Safety on Five Continents (Rs5c 2016), Rio de Janeiro, Brazil, 17-19 May 2016.* Statens väg-och transportforskningsinstitut.

Vasudevan, Meenakshy, Chris Curtis, Alexa Lowman, and James O’Hara. 2016. “Predicting Traffic Flow Regimes from Simulated Connected Vehicle Messages Using Data Analytics and Machine Learning.” ITS Joint Program Office, Department of Transportation.

Wang, Kai. 2016. “Exploration of Advances in Statistical Methodologies for Crash Count and Severity Prediction Models.” PhD thesis, University of Connecticut.

Wei, Xinxin, Xiang Shu, Baoshan Huang, Edward L Taylor, and Huaxin Chen. 2017. “Analyzing Traffic Crash Severity in Work Zones Under Different Light Conditions.” *Journal of Advanced Transportation* 2017. Hindawi.

Xie, Yuanchang, Dominique Lord, and Yunlong Zhang. 2007. “Predicting Motor Vehicle Collisions Using Bayesian Neural Network Models: An Empirical Analysis.” *Accident Analysis & Prevention* 39 (5). Elsevier: 922–33.

Xu, Jian, Kara M Kockelman, and Yiyi Wang. 2014. “Modeling Crash and Fatality Counts Along Mainlanes and Frontage Roads Across Texas: The Roles of Design, the Built Environment, and Weather.” *93rd Annual Meeting of the Transportation Research* 22 (23): 24.

Zeng, Qiang, and Helai Huang. 2014. “Bayesian Spatial Joint Modeling of Traffic Crashes on an Urban Road Network.” *Accident Analysis & Prevention* 67. Elsevier: 105–12.

U.S. Department of Transportation
John A. Volpe National Transportation Systems Center
55 Broadway
Cambridge, MA 02142-1093

617-494-2000
www.volpe.dot.gov

DOT-VNTSC--BTS-19-01