Big Data Visualization and Spatiotemporal Modeling of Risky Driving

July 2020 | Final Report

M 14
Ann Arbor Charter Township
M 153
Barton Hills
Superior Charter Township
Scio Township
167
US 23
Ann Arbor
M 17
Ypsilanti
+
−

Source: United States Department of Transportation, Preparing for the Future of Transportation: Automated Vehicle 3.0, October 4, 2018, Washington,DC. https://www.transportation.gov/av/3. Accessed February 3, 2020.

SAFE-D
SAFETY THROUGH DISRUPTION

VT | TRANSPORTATION INSTITUTE VIRGINIA TECH.

Texas A&M Transportation Institute

San Diego State University
Leadership Starts Here

Disclaimer

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

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. 03-087	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Big Data Visualization and Spatiotemporal Modeling of Risky Driving		5. Report Date July 2020 6. Performing Organization Code:
7. Author(s) Arash Jahangiri Charles Marks Sahar Ghanipoor Machiani Atsushi Nara Mahdie Hasani Eduardo Cordova Ming-Hsiang Tsou Joshua Starner		8. Performing Organization Report No. Report 03-087
9. Performing Organization Name and Address: Safe-D National UTC San Diego State University Virginia Tech Transportation Institute		10. Work Unit No. 11. Contract or Grant No. 69A3551747115/Project 03-087
12. Sponsoring Agency Name and Address Office of the Secretary of Transportation (OST) U.S. Department of Transportation (US DOT)		13. Type of Report and Period Final Research Report 14. Sponsoring Agency Code

15. Supplementary Notes

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

16. Abstract

Statistical evidence shows the role of risky driving as a contributing factor in roadway collisions, highlighting the importance of identifying such driving behavior. With the advent of new technologies, vehicle kinematic data can be collected at high frequency to enable driver behavior monitoring. The current project aims at mining a huge amount of driving data to identify risky driving behavior. Relational and non-relational database management systems (DBMSs) were adopted to process this big data and compare query performances. Two relational DBMSs, PostgreSQL and PostGIS, performed better than a non-relational DBMS, MongoDB, on both nonspatial and spatial queries. Supervised and unsupervised learning methods were utilized to classify risky driving. Cluster analysis as an unsupervised learning approach was used to label risky driving during short monitoring periods. Labeled driving data, including kinematic information, were employed to develop random forest models for predicting risky driving. These models showed high prediction performance. Open source and enterprise visualization tools were also developed to illustrate risky driving moments in space and time. These tools can be used by researchers and practitioners to explore where and when risky driving events occur and prioritize countermeasures for locations in highest need of improvement.

17. Key Words Publication, guidelines, report, brochure, communication, marketing Driver behavior monitoring, risky driving identification, cluster analysis, database management, big data	18. Distribution Statement No restrictions. This document is available to the public through the Safe-D National UTC website, as well as the following repositories: VTechWorks, The National Transportation Library, The Transportation Library, Volpe National Transportation Systems Center, Federal Highway Administration Research Library, and the National Technical Reports Library.

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 24	22. Price $0

Abstract

Statistical evidence shows the role of risky driving as a contributing factor in roadway collisions, highlighting the importance of identifying such driving behavior. With the advent of new technologies, vehicle kinematic data can be collected at high frequency to enable driver behavior monitoring. The current project aims at mining a huge amount of driving data to identify risky driving behavior. Relational and non-relational database management systems (DBMSs) were adopted to process this big data and compare query performances. Two relational DBMSs, PostgreSQL and PostGIS, performed better than a non-relational DBMS, MongoDB, on both nonspatial and spatial queries. Supervised and unsupervised learning methods were utilized to classify risky driving. Cluster analysis as an unsupervised learning approach was used to label risky driving during short monitoring periods. Labeled driving data, including kinematic information, were employed to develop random forest models for predicting risky driving. These models showed high prediction performance. Open source and enterprise visualization tools were also developed to illustrate risky driving moments in space and time. These tools can be used by researchers and practitioners to explore where and when risky driving events occur and prioritize countermeasures for locations in highest need of improvement.

Acknowledgements

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

Table of Contents

List of Figures

List of Tables

Introduction

The formal concept of risky or aggressive driving may date back to 1968 when Meyer Parry's monograph, *Aggression on the Road*, was published. Parry declared that "the increasing stress involved in motoring nowadays makes the psychological efficiency of the driver a more important factor than the mechanical efficiency of the vehicle he drives" [1]. Examples of aggressive behaviors include tailgating, driving faster than other drivers, running stop lights and stop signs, and improper lane changes [2]. The term "risky driving" has been also used instead of "aggressive driving." Risky driving mainly involves drinking and driving or driving without wearing a seat belt and excludes some behaviors associated with aggressive driving, such as horn honking [2].

Understanding driving style helps with the evaluation of traffic safety, and the impact of aggressive driving on traffic safety has drawn researchers' and practitioners' attention. The National Highway Traffic Safety Administration found that aggressive driving is one of the most important factors affecting traffic safety, with aggressive driving behavior observed in two-thirds of fatal crashes [3]. In support of this, many studies have revealed the effects of aggressive driving behavior on crash rates [4-7]. Research by the AAA Foundation revealed that in 55.7% of the fatal crashes that occurred from 2003 to 2007, at least one driver had already committed one or more aggressive behavior [8]. Paleti et al. [9] also revealed a positive association between aggressive driving and injury severity.

It is therefore important to identify when and where risky or aggressive driving moments occur so that appropriate actions can be taken. However, in most cases, there is no evidence that shows risky behaviors in time and space. The present work aims at identifying and visualizing risky driving moments in a large, real-world driving dataset, the Safety Pilot Model Deployment (SPMD), where there is no hard evidence to confirm such moments.

Literature Review

Driving styles can be explored and evaluated by monitoring instantaneous driving decisions as reflected in vehicle kinematic data [10-15]. Speed has been identified as the main factor in determining a driver's performance when assessing driving style [16,17]. Acceleration has also been used as an intuitive measure to identify aggressive driving [16,18,19]. Certain values of motion-related variables have been determined to be representative of aggressive driving behavior as well. As the main focus of this project is on kinematic data corresponding to driving style, aggressive driving was mainly studied and discussed with respect to kinematic data.

Definitions

Driving style is the way a driver chooses to drive or the way the driver has become habituated to driving over time [20,21]. While aggressive driving can be considered a driving style, there is no consensus among researchers and experts as to a concrete definition of "aggressive driving."

Mizell et al. defined aggressive driving incidents as those in which an angry or impatient driver kills or injures or attempts to kill or injure another driver or passenger or pedestrian in an unfavorable traffic condition [22]. According to the National Highway Traffic Safety Administration, Mizell's definition is a definition of a "road rage" criminal offense, while aggressive driving behavior is associated with lesser traffic offenses [23]. Shinar defined aggressive behavior as one's intention to inflict physical or psychological harm on a person. He also noted the difference between aggressive driving and aggressive drivers—the former is a kind of behavior displayed by many drivers less frequently, while the latter are individuals who drive aggressively most of the time [24]. Aggressive drivers can also be simply defined as careless drivers [25], and aggressive driving has been referred to as a driving behavior where a driver intentionally tends to increase the risk of accident with contempt toward other drivers [26]. Some factors increase the likelihood of aggressive driving behavior, such as being in an angry mood or in congested traffic [2]. One study suggested the following definition, which captures several definitions in one: "A driving behavior is aggressive if it is deliberate, likely to increase the risk of collision, and is motivated by impatience, annoyance, hostility and/or an attempt to save time" [2]. Since the focus of this study is on identifying aggressive driving from kinematic data only, things like deliberate actions or driver impatience are unknown, as there is no way to determine intentions or driving conditions. For that reason, we opted to use the term "risky driving" instead of aggressive driving, as additional variables would be required to determine that behavior. Also, since risky driving has been defined differently in different studies, in the present work risky driving is defined as any driving behavior that is not considered the norm and that is more likely to increase the probability of collisions. It is important to note that a driving behavior may not be aggressive but may still be risky. For example, a swerve to avoid a collision with a child running into the street is not an aggressive behavior, but since the driver makes an abnormal maneuver, it is still considered risky.

Driving Style Categorization

Studies have categorized driving style using different variables and methods. Appendix A summarizes the approaches found in the literature. The table includes columns for method name, type (supervised or unsupervised), and accuracy. The "variables" column indicates the variables applied to classify driving style. In the "boundary" column, a threshold was specified for variables to identify aggressive driving. Some researchers used binary categorization, such as aggressive versus non-aggressive, while others used multi-class categorization. The "driving style categories" column lists previous studies' driving style categorizations.

Previous studies have utilized different approaches to perform binary categorization, including supervised machine learning, unsupervised machine learning, and traditional methods. In a supervised method, a set of labeled driving behavior events were used to classify new unlabeled events. In one study, 120 labeled behaviors were used to perform a k-nearest neighbor analysis using dynamic time warping to categorize driving behavior [27]. Another study applied a naïve Bayes classifier to evaluate drivers' aggressiveness according to questionnaire responses and collected driving features such as maximum and average speed, acceleration, and throttle position [28]. Yu et al. used a smartphone sensor and applied support vector machine and neural networks as classifiers to identify abnormal—weaving, swerving, sideslipping, fast U-turn, turning with a wide radius, and sudden braking—and normal driving behavior [18]. The random forest model is another supervised approach to classify aggressive and normal driving at a horizontal curve [29]. In addition to supervised machine learning methods, some studies applied unsupervised machine learning techniques to categorize driving style into a binary categorization. For instance, Lee et al. applied a three-step procedure: abrupt change detection, feature extraction, and a two-level clustering algorithm, including a self-organizing map and k-means, to classify driving style. A framework was proposed to classify drivers' behavior into aggressive and normal driving based on speed, yaw rate, and acceleration. [19]. Jahangiri et al. employed k-means as an unsupervised learning method to identify aggressive driving events using variables such as speed, acceleration, and yaw rate measured over some monitoring period [15].

In addition to machine learning techniques, several studies applied more traditional statistical methods for classification, such as linear regression models, nonlinear regression models, and *t*-test analyses. Wang et al. characterized drivers styles' as calm versus volatile by categorizing vehicular jerk. A driving style can be identified as a volatile behavior when the acceleration exceeds the mean plus or minus 1 standard deviation for a certain speed range. A similar approach was used on vehicular jerk to detect volatility [16].

Kinematic Data and Aggressive Driving

As explained earlier, kinematic data have a significant role in determining aggressive driving behavior. The investigation of the relationship between unsafe driving behavior and kinematic data is not limited to driving style studies. Several researchers have defined specific thresholds to stratify kinematic data ranges into various categories, such as safe, unsafe, and comfortable, investigating traffic-safety topics such as the impact of specific driver behavior on driving style, and comparing young and adult driver style [30-35]. Appendix B summarizes variables, thresholds, and recommended categories for some of these studies. For instance, the American Association of State Highway and Transportation Officials recommended a deceleration of 3.4 m/s^2 (considered comfortable for most drivers) to determine stopping sight distance [36]. Another study investigated driver behavior based on crash involvement data. Drivers were categorized into crash and non-crash groups based on self-reported survey data of past crash involvement. Speed and acceleration data were also collected for both groups based on GPS data. The impact of hard deceleration on crash involvement was then evaluated. A threshold of 6 mph/s was employed to define hard deceleration events. The frequency of hard deceleration events was statistically different between the two groups, which showed that more hard deceleration events are associated with crash involvement, implying that deceleration rates of more than 6 mph/s can be attributed to aggressive driving behavior that has a potential of leading to crashes [37]. In another study, Fazeen et al. proposed a driving assistance system that analyzes road and driving conditions and advises users about unsafe situations using a smartphone with GPS, headphones, and accelerometer. Safe and unsafe accelerations and decelerations were separated based on accelerometer data recorded using a threshold of 3 m/s^2. The differentiation between safe and unsafe events was based on whether the acceleration and deceleration were gradual. However, it is not clear exactly how the threshold of 3 m/s^2 was determined [38]. Additionally, some studies identified maximum and minimum values for acceleration and deceleration data that can help identify outliers (i.e., risky driving behaviors). These thresholds are presented in Appendix C.

Methods and Approach

The present work describes the development of database management systems and Web-based analytics tools to identify and visualize risky driving behavior across space and time. Risky driving behavior was investigated by monitoring the kinetic information of vehicles. A suite of methods was explored to efficiently store, process, and analyze the dataset. Four main steps are shown in Figure 1 and described below: data exploration, database development, risky driving classification, and data visualization and tool development.

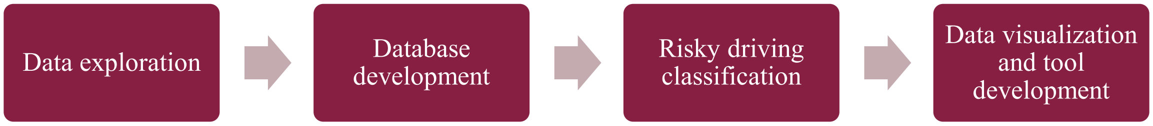

Data exploration
Database development
Risky driving classification
Data visualization and tool development

Figure 1. Flowchart. Overview of processes.

Data Exploration

As part of the SPMD program, large transportation datasets were collected in Ann Arbor, Michigan, and were made publically available via the Federal Highway Administration’s Research Data Exchange. The SPMD data collection made use of approximately 3,000 onboard vehicle units and 30 roadside equipment units that provided vehicle-to-vehicle and vehicle-to-infrastructure communications data. Basic Safety Messages containing vehicle operation information were communicated via dedicated short-range communications. The available SPMD data include text-based files along with a handbook and metadata document.

This study aims to detect and analyze risky driving events within the SPMD data. The BSMP1 dataset, containing latitude, longitude, and kinematic data, such as speed, acceleration, and yaw rate (see Table 1), was used. Continuous data were collected from vehicles at a rate of 10 Hz, resulting in large amounts of data. The public-access BSMP1 dataset corresponds to 2 months of data (April and October 2013), which are 295.5 GB in size, uncompressed. The SPMD dataset comprises 38 data tables in a comma-separated value (csv) file format. For database development, we used the largest data table in the April 2013 dataset, which contained more than 1.5 billion GPS points (205 GB). In addition to the SPMD dataset, we used geographic information system (GIS) layers obtained from the City of Ann Arbor’s Data Catalog website (https://www.a2gov.org/services/data/Pages/default.aspx) to provide geospatial contextual information. Since analyzing large amounts of data was time-consuming, we opted to use one week (the first week of April 2013) of data for the risky driving identification. In addition to the Research Data Exchange, there are other sources of data that frequently provide datasets in different domains. These sources were also explored (see Appendix D) to see if other kinematic data were available for analyzing risky driving.

Table 1. Variables Used in BSMP1 Dataset

Variable Name	Data Type	Units	Description
RxDevice	Integer	None	Unique ID of vehicle
Gentime	Integer	None	A more secure form of Epoch time, influence by 1609.2 of the IEEE 1609 family of standards-related network management and security
Latitude	Float	Degrees	Current latitude of vehicle
Longitude	Float	Degrees	Current longitude of vehicle
Heading	Real	Degrees	Vehicle direction
Speed	Real	m/s	Vehicle speed
A_x	Real	m/s^2	Vehicle longitudinal acceleration
A_y	Real	m/s^2	Vehicle lateral acceleration (due to measurement error seen in many vehicles, this variable was excluded from analyses)
Yaw rate	Real	Degrees/s	Vehicle yaw rate

Database Development

Two types of open source database management systems (DBMSs) were utilized to store, query, and analyze the SPMD data with GIS layers: (1) an object-relational database (PostgreSQL and PostGIS), and (2) a NoSQL document-oriented database (MongoDB).

Relational DBMS

We implemented a relational DBMS using PostgreSQL and PostGIS. PostgreSQL is a relational database that stores data in a set of strictly defined tables, making it ideal for structured data. Structure Query Language (SQL) was used to build, manage, and query the stored data. PostGIS is a spatial database extender for PostgreSQL. It provides rich spatial operators, spatial functions, spatial data types (including vector, raster, and network types), and spatial indexing enhancements to PostgreSQL, allowing sophisticated GIS analyses. The project database consists of 57 tables, 38 of which were created from SPMD data and 19 of which were created from GIS data. To speed up spatial queries, we created a GiST (Generalized Search Tree) spatial index on geometry columns. To further improve query performance on the large data table, we implemented vertical database partitioning on BSMP1 data tables using the timestamp field.

Non-relational DBMS

MongoDB, a NoSQL document-oriented database, makes the integration of very large datasets easier and faster by storing records in a JavaScript Object Notation (JSON) format. A NoSQL database does not have a strict table structure and does not support relationships between tables, allowing unstructured data to be stored [39]. MongoDB uses JavaScript for its query language. We implemented the MongoDB database by importing the SPMD csv files and generating collections, which are analogous to tables in relational databases. In each collection, SPMD data were stored as unstructured documents, which were composed of field-value pairs. The value of a field can be any of the BSON (JSON documents in binary-encoded format) data types, including other documents, arrays, and arrays of documents. For example, a document of a GPS point record can be stored in a MongoDB collection as follows:

```
{ _id: 1, type: “Feature”,
  properties: {gid: 8, speed: 30},
  geometry: {type: “Point”, coordinates: [-83.62, 42.24]} }
```

We implemented a 2dsphere index on geometry fields to execute a spatial query efficiently in MongoDB.

Database Comparison

Database query performance tests were conducted between PostgreSQL/PostGIS and MongoDB, which stored identical data derived from the SPMD dataset. Our test was focused on nonspatial and spatial queries. A spatial query, supported by spatial databases, considers the spatial relationships among geometries of location data. PostgreSQL/PostGIS supports several spatial data types and over 300 functions for working with these spatial types, while MongoDB supports four spatial queries (geoIntersects, geoWithin, near, nearSphere). For an accurate comparison of the two databases, a simple intersection query was used, as both databases support this feature.

Database Settings

PostgreSQL/PostGIS and MongoDB were installed on the same machine. To conduct performance tests in similar database settings, each database was non-partitioned and had identical data tables generated from the same SPMD csv files. The computer used for testing was a Windows 10 operating system with 16 GB of RAM, 2 TB of storage, and an Intel(R) Xeon(R) W-2123 (8.35M Cache, 3.60 GHz). We installed PostgreSQL 11 with PostGIS 2.5.1 and MongoDB 4.0.5.

In MongoDB, we used the default RAM configuration, which takes advantage of approximately 50% of the available RAM minus 1 GB when there is more than 1 GB available [40]. The default setting of PostgreSQL was limited for a small database server environment; therefore, the configuration for PostgreSQL was changed so that memory usage was on par with MongoDB (shared_buffers = 8 GB, effective_cache_size = 8 GB, work_mem = 1 GB). Nevertheless, configuring the same settings for memory usage in the two databases was not straightforward, as both databases use memory differently.

For the performance test, we used the April BSMP1 data, which has 21 attribute columns with approximately 1.5 billion GPS points. To add the timestamp field on both databases, we converted the *gentime* field into the *epoch timestamp*. In addition, we created a point geometry column on the April BSMP1 table in each database using latitude and longitude. To test database performance in terms of scalability by different database sizes, we created nine subset tables from the April BSMP1 data for each database (see Appendix E).

Query Designs

We designed two nonspatial and two spatial queries to test the two databases' performances:

> Query 1: a nonspatial query to retrieve the number of records in a given range of time;
>
> Query 2: a nonspatial query to retrieve the number of records above a specific speed value;
>
> Query 3: a spatial query to retrieve the number of GPS points that intersect within a specified road buffer; and
>
> Query 4: a spatial query to retrieve the number of points that intersect within randomly distributed circles with two different radii.

Queries 1 and 2 utilize all nine subset tables (Appendix E). Queries 3 and 4 only use the Subset 9 table with 500,000,000 GPS points. In Query 3, we examined the performance of the spatial intersection query between GPS points and a major road buffer and the effect of the intersected road buffer size. Road buffers were created along North/South Main Street in Ann Arbor with lengths of 29, 58, 116, 232, 464, 927, 1,855, 3,710, and 7,420 ft and a width of 50 ft (Appendix F). The largest buffer length of 7,420 ft was selected for its proximity to Ann Arbor's town center. The smaller buffers were then created by reducing the buffer lengths by half. In Query 4, we used the spatial intersection query between GPS points and circle buffers at random locations to examine the effect of locations on query performance. To create random circle buffers, we randomly selected 100 GPS points from the Subset 9 table. For each of those randomly selected 100 points, we created a 50-ft and a 1,400-ft circular diameter buffer. Query 4 used the exact same intersection query as Query 3. These two buffer sizes were selected based on the width of a road (50 ft) and the length of a neighborhood block in Ann Arbor (1,400 ft). We executed each query 10 times and obtained the average execution time. The query language examples used for the performance tests in each database are detailed in Appendix G.

Results

Figure 2, Figure 3, Figure 4, Figure 5, Table 2, and Table 3 present the results of the performance tests. Our experiments show that PostgreSQL and PostGIS perform better than MongoDB on both nonspatial queries (Query 1 and 2) and spatial queries (Query 3 and 4). In particular, PostgreSQL and PostGIS outperformed MongoDB when the data size was larger.

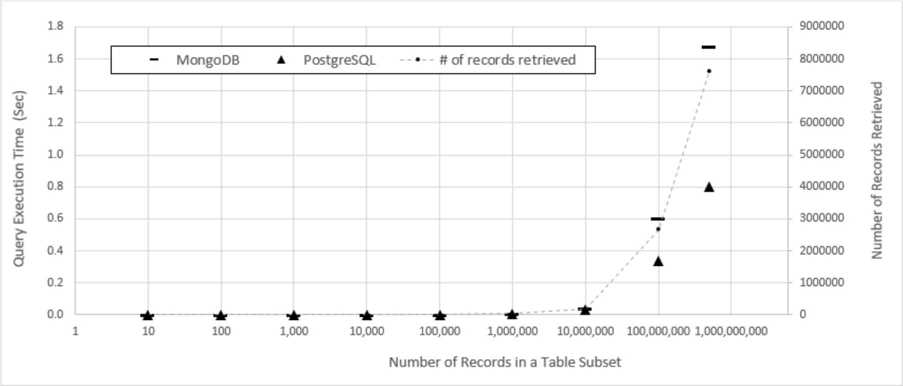


Figure 2. Graph. Query 1: Nonspatial query performance using a timestamp filter.

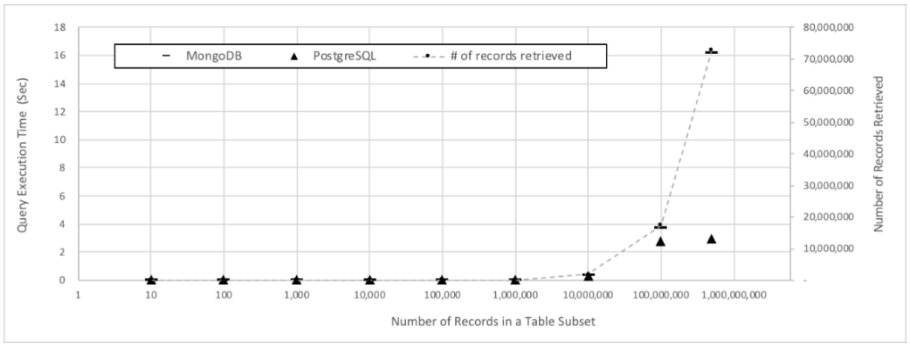


Figure 3. Graph. Query 2: Nonspatial query performance using a speed filter.

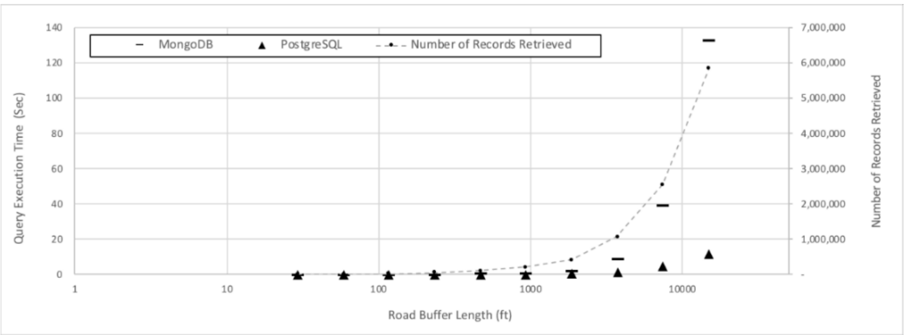


Figure 4. Graph. Query 3: Spatial query performance using road buffers.

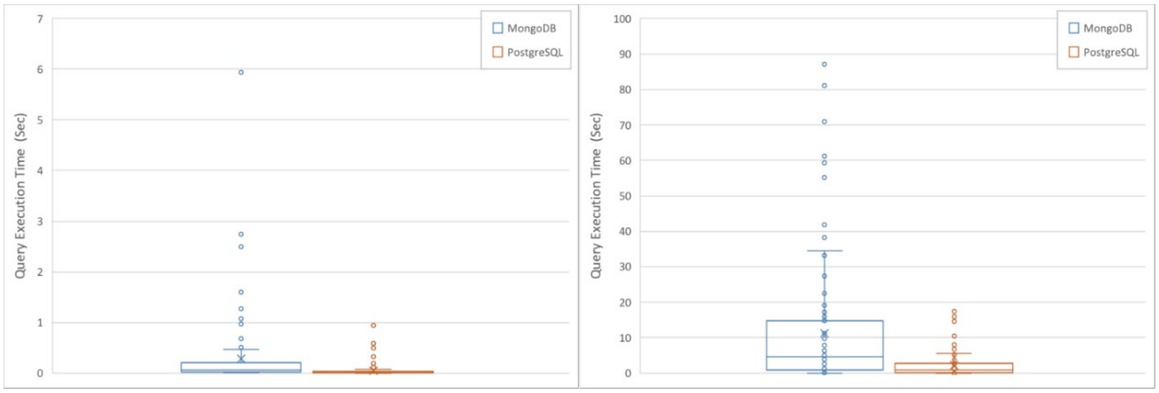


Figure 5. Charts. Query 4: Spatial query performance using 100 random circles (left: 50-ft circles, right: 1,400-ft circles). The box-whisker plot shows the minimum, first quartile, median, third quartile, and maximum of the query execution time.

Table 2. Execution Time of Spatial Query Using 100 Random Circles (50-ft Circle)

Database	Avg. (s)	Med. (s)	Min. (s)	Max (s)	Std. (s)
MongoDB	0.278	0.055	0.002	5.940	0.727
PostgreSQL	0.054	0.012	0.001	0.943	0.127

Number of records retrieved: Avg. = 43967.9; Med. = 7114.5; Min. = 36; Max = 832346; Std. = 114534.3

Table 3. Execution Time of Spatial Query Using 100 Random Circles (1,400-ft Circles)

Database	Avg. (s)	Med. (s)	Min. (s)	Max (s)	Std. (s)
MongoDB	11.142	4.645	0.012	87.029	17.521
PostgreSQL	2.149	0.790	0.005	17.431	3.409

Number of records retrieved: Avg. = 1744790.6; Med. = 663038.0; Min. = 2517; Max = 12655130; Std. = 2657178.5

Risky Driving Classification

Two approaches were adopted to detect risky driving behavior: (1) a simple approach that defines risky driving as driving events during which speeding is observed, and (2) a more advanced approach that identifies risky driving using unsupervised and supervised learning methods.

Risky Driving Detection – Speeding

The BSMP1 data include speed, location, direction, yaw rate, and heading collected at approximately 10 Hz. The major road line dataset contains major road conditions and speed limits from the Southeast Michigan Council of Government's Annual Average Daily Traffic program.

To find cases of speeding, we used sampled BSMP1 data (0.1%) with Esri ArcGIS Pro and buffered the roads by a width of 5 meters, since major roads should be at least 5 meters wide. We then used a spatial join to aggregate vehicle point data into the road buffers. We compared the speed recorded at each vehicle point with the speed limit and extracted four speeding clusters:

Class 1: speeding 1–5 mph over speed limit;
Class 2: speeding 5–10 mph over speed limit;
Class 3: speeding 10–20 mph over speed limit; and
Class 4: speeding more than 20 mph over speed limit (Figure 6).

We visualized these data points on GIS maps to identify their spatial patterns and point density. Figure 6 illustrates the severity of speeding and associated cluster patterns. Speeds more than 10 mph over the speed limit mostly occurred at the intersections of major highway segments and ramps. Speeding was also seen where the speed limit changes between two local road segments. Most speeding cases of less than 10 mph over the speed limit were in downtown Ann Arbor. The spatial pattern of speeding activity can be further evaluated in the future using advanced machine learning methods.

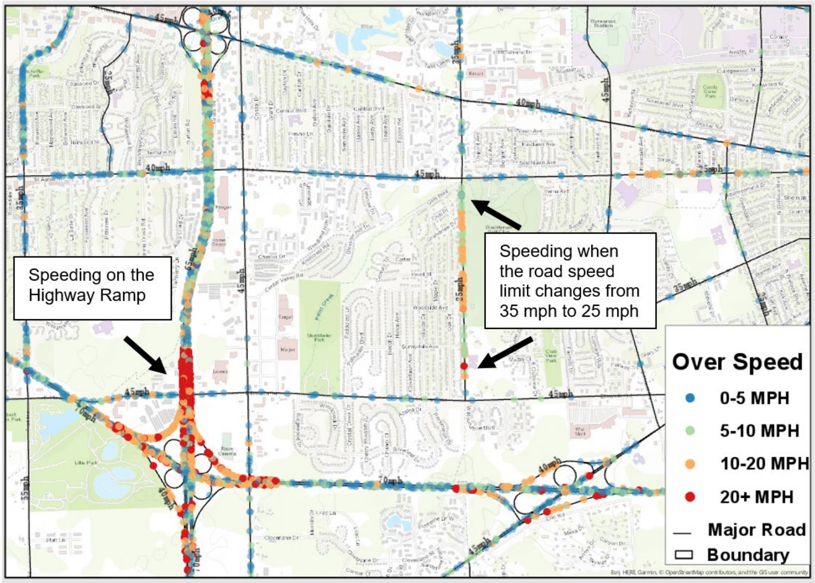

Speeding on the Highway Ramp
Speeding when the road speed limit changes from 35 mph to 25 mph
Over Speed
● 0-5 MPH
● 5-10 MPH
● 10-20 MPH
● 20+ MPH
Major Road
Boundary

Figure 6. Map. Identifying over-speeding locations in major roads using BSM_p1 GPS trajectory datasets.

Risky Driving Detection – Unsupervised and Supervised Learning

The process of risky driving classification was divided into five primary stages: subsetting the BSMP1 data, restructuring the data for classification, labeling the data as either risky or not (i.e., unsupervised learning), training predictive models based on the labeled data (i.e., supervised learning), and utilizing the predictive models to then label the BSMP1 data. Analyses were conducted in R. The goal was to develop a framework in which unsupervised and supervised learning methods can be applied to identify risky driving in a large unlabeled dataset. It should be noted that within this framework, one can utilize other unsupervised and supervised learning methods instead of the ones used in this study. This could be a future direction of the work to explore new deep-learning algorithms within this framework.

Subsetting BSMP1 Data

First, the BSMP1 data for the month of April 2013, stored in a PostGreSQL database, were subsetted by day. For classification, we considered the first seven days (April 1–7, 2013). For each of the seven days, a subset of data corresponding to 100 unique vehicle IDs was extracted.

Restructuring the Data

The BSMP1 data represent single time points, measured every decisecond of subject vehicles as they traveled. These data points, individually, lack the temporal context to identify and classify instances of risky driving. As such, prior to classification, the data were reformatted from a time point format into a *monitoring period* format. Monitoring period data were generated by taking time point datapoints representing *x* seconds at *y*-second intervals and generating data such as the average, maximum, minimum, and standard deviation of speed, acceleration, and yaw rate over those *x* seconds. Converting from time point to monitoring period data made it possible to better identify distinct driving behaviors.

Restructuring each of the datasets for April 1–7 involved a series of preparatory steps. First, the time point data had to be ordered by vehicle and by time. Then, the data points corresponding to a single, continuous trip needed to be identified and grouped together, as it would be inappropriate to include data from the end of one trip with the data from the beginning of another into a single monitoring period. Next, the time point data were converted into monitoring period data. We opted to create monitoring periods corresponding to 3 seconds at 1-second intervals. Since time point data were measured every decisecond, each monitoring period was computed using 30 time point data points. Further, computing monitoring periods every second reduced the size of the datasets by approximately one order of magnitude.

Labeling the Monitoring Period Data – Unsupervised Learning

Once the monitoring period data for each day were created, risky driving data points could be labeled and classified. This was done by utilizing a k-means clustering algorithm, determining heading thresholds to subset the data, and then utilizing the density-based spatial clustering of applications with a noise algorithm in an iterative fashion to identify risky driving periods. Although many unsupervised methods could be adopted, we opted to use k-means since this method is easy to implement and computationally efficient. We also applied density-based spatial clustering of applications with noise (DBSCAN) due to previous experience using this method in identifying risky driving [41]. The general principle of the labeling approach was that there are a set of elementary driving behaviors (EDM) that occur (such as accelerating, making a U-turn, merging onto the highway, etc.) and that these EDMs will likely have similar statistical profiles. Potentially risky driving behaviors, then, are identified as the data points that are further outliers from their prescribed cluster. This is meant to capture abnormal EDM instances. The limitation of using unsupervised learning methods in the labeling step is the fact that there is no guarantee that a behavior is correctly classified as risky. That is basically the nature of unsupervised learning problems, where validation is challenging. In the case of risky driving classification, video data could help with class validation. Although video data were collected as part of the SPMD project, we did not have access to the video data as they have not been made available to the public. Video validation can be further explored if the data become available in the future.

K-means and change-in-heading thresholds were first utilized to identify the EDMs. This was done by first running k-means on the full dataset, clustering on only the average speed variable, resulting in three distinct speed classes (low speed, medium speed, and high speed). Then, these subsets were further subsetted into five groups based on change in heading (left turns, left curves, straight driving, right curves, and right turns). After experimenting with different thresholds for change in heading, the following values were used: change in heading greater than 45 degrees for left and right turns; change in heading between 10 and 45 degrees for left and right curves; and change in heading under 10 degrees for straight. Subsequently, k-means was run on each of these 15 subsets, using the sum of squared distances "elbow" method to identify the optimal number of clusters (clustering variables were average, maximum, and standard deviation of speed; average, maximum, minimum, standard deviation, and jerk of acceleration; and average, maximum, minimum, standard deviation, and jerk of yaw rate).

For each of the k-means clusters identified in each of the 15 subsets, DBSCAN was performed iteratively (using the Iterative-DBSCAN [I-DBSCAN] method) [41]. Since the data have been clustered into EDMs, the dataset will be dense, and each iteration of DBSCAN will cluster most of the data together. DBSCAN returns *n* clusters and one set of noise (i.e., unclustered data). One iteration of I-DBSCAN is as follows. First, DBSCAN is run on the dataset, utilizing the "elbow" method to determine the optimal *epsilon* parameter. Second, the "normal" cluster is identified as the cluster consisting of at least 90% of the dataset; if no such "normal" cluster exists, I-DBSCAN is terminated and run again from the beginning. Third, all data identified as noise are extracted and labeled as risky. Fourth, if any additional clusters are identified, they are extracted and labeled as risky. If no such additional cluster is identified, then researchers verify that this was the third time no additional cluster was found. If so, I-DBSCAN is terminated and the results are returned. If not terminated, another I-DBSCAN iteration is undertaken utilizing the "normal" cluster as the dataset. In a sense, this process is similar to peeling the layers off an onion, where the furthest outlying data points are "peeled away" and labeled as risky and the dense set of data in the middle is labeled as not risky. While utilizing I-DBSCAN, principle component analysis (PCA) was adopted to reduce the data dimensions (i.e., number of variables) and thus reduce the data size. This is a great technique for handling big data because PCA has the potential to extract important information from many variables and generate a few new variables to use instead of the original variables without significant loss of information [42]. After I-DBSCAN was run on all the generated subsets, the labeled datasets were merged back together.

Training Predictive Models – Supervised Learning

Each of the labeled datasets was then used to train a random forest classification model for that day. Random forest was chosen after comparing the performance of logistic regression, random forest, and neural network classification models utilizing 5-fold cross-validation. This process involves dividing the data into five groups, training models on four-fifths of the dataset, and then testing the model on the remaining one-fifth to gauge performance at predicting whether the events were risky or not. A random forest model was then fitted to April 1, 2, and 4–7. Due to data file corruption, April 3 data were not used in developing the random forest classification model.

Using the Predictive Models to Label the Full Dataset

The random forest models for April 1, 2, and 4–7 were trained on the subsets of BSMP1 data from each day. These models were then utilized to label all the data in each of these datasets. To do this, data were extracted from each dataset by vehicle ID, converted into monitoring period data format, and then labeled utilizing the random forest model. The labeled datasets were then saved in the database by day, such that a database table with risky labels was created for each day. In addition, a table consisting of all of the risky-labeled data points in all six of these data tables was created.

Results

Subsetting BSMP1 Data

BSMP1 data were subsetted in the PostGreSQL database by calendar day. For analysis, datasets corresponding to April 1–7, 2013 were utilized (see Table 4 for the number of data points in each table and the corresponding number of vehicles). Data corresponding to 100 randomly selected vehicles were extracted (see Table 4). As noted, the April 3 sample set was corrupted during the full analysis; the remaining results correspond to the other six days.

Table 4. Subsetting BSMP1 Data

Date	Day	Approximate Database Size (Number of Data Points)	Number of Vehicles	100-vehicle Sample Size (Number of Data Points)
April 1	Monday	44.5 Million	1,395	3.61 Million
April 2	Tuesday	51.4 Million	1,418	3.03 Million
April 3	Wednesday	51.7 Million	1,440	NA
April 4	Thursday	50.0 Million	1,430	3.27 Million
April 5	Friday	50.0 Million	1,405	2.97 Million
April 6	Saturday	39.7 Million	1,133	3.37 Million
April 7	Sunday	32.6 Million	1,072	3.14 Million

Restructuring the Data

At this stage, the six remaining datasets were converted from their initial time point data format into the monitoring period data format. Data conversion resulted in the datasets' size being reduced by an order of magnitude. Table 5 shows the number of data points before and after data conversion, as well as the number of distinct continuous trips identified in each sample.

Table 5. Restructuring the Data

Date	Day	Dataset Size Prior to Conversion (Number of Data Points)	Dataset Size Post Conversion (Number of Data Points)	Distinct Vehicle Trips
April 1	Monday	3.61 Million	291,155	1,383
April 2	Tuesday	3.03 Million	257,752	1,350
April 4	Thursday	3.27 Million	277,634	3,085
April 5	Friday	2.97 Million	250,467	1,225
April 6	Saturday	3.37 Million	203,073	1,773
April 7	Sunday	3.14 Million	212,488	811

Labeling the Monitoring Period Data – Unsupervised Learning

The clustering protocol was utilized on each of the six datasets separately to label monitoring data as either risky or not risky. Table 6 displays the number of risky monitoring periods labeled in each set and the proportion of the total table labeled as risky. The proportion of each dataset labeled as risky ranged from 8.25% to 10.0%, indicating that the clustering algorithm behaved in a consistent manner.

Table 6. Labeling the Monitoring Period Data

Date	Day	Risky Monitoring Periods	Proportion of Dataset
April 1	Monday	24,021	8.25%
April 2	Tuesday	23,063	8.95%
April 4	Thursday	26,296	9.5%
April 5	Friday	25,227	10.0%

April 6	Saturday	19,672	9.69%
April 7	Sunday	19,666	9.26%

Training Predictive Models – Supervised Learning

With the labeled data now procured, random forest models were tested and evaluated on each of the six datasets utilizing 5-fold cross validation. The area under the curve (AUC) scores for each of the six models are presented in Table 7. AUC scores ranged from 0.974–0.979, indicating that random forests models were highly accurate at identifying data labeled as risky. Random forest models were then trained on each of the six datasets.

Table 7. Training Predictive Models

Date	Day	AUC
April 1	Monday	0.977
April 2	Tuesday	0.976
April 4	Thursday	0.975
April 5	Friday	0.979
April 6	Saturday	0.977
April 7	Sunday	0.974

Using the Predictive Models to Train the Full Data

The six random forests models fitted in the prior step were then applied to label all the data in the PostGreSQL database corresponding to the same date. Data were extracted from the PostGreSQL by day and by vehicle, reformatted into the monitoring period structure, labeled utilizing the corresponding random forest model, and then inserted into a new table in the PostGreSQL database. Table 8 displays the size of the original database table, the size of the new labeled database table, and the percentage of the entries labeled as risky for each day. Final datasets ranged in size from 2.43 million to 4.47 million data points and, in each table, between 6.89% and 8.9% of all data points were labeled as risky.

Table 8. Using the Predictive Models to Train the Full Data

Date	Day	Approximate Database Size (Number of Data Points)	Size of Labeled Database Table (Number of Data Points)	Proportion Labeled Risky
April 1	Monday	44.5 Million	3.92 Million	7.10%
April 2	Tuesday	51.4 Million	4.32 Million	7.54%
April 4	Thursday	50.0 Million	4.60 Million	7.93%
April 5	Friday	50.0 Million	4.47 Million	8.90%
April 6	Saturday	39.7 Million	2.92 Million	7.62%
April 7	Sunday	32.6 Million	2.43 Million	6.89%

Data Visualization and Tool Development

Spatiotemporal analysis was used to visualize monitoring periods of risky driving behavior and investigate how these moments were distributed both spatially and temporally. This effort also demonstrates the feasibility of big data visualization and spatiotemporal modeling and analytics through Web-based GIS tools utilizing DBMSs.

Two types of data visualization approaches were used. The first approach utilized open source software tools. R Shiny with the Leaflet package, OpenLayers, and D3 were adopted, along with the open source databases developed earlier in this work, to map and visualize multilayered geographic information and statistics. The second approach utilized currently available data visualization software, including Tableau, Insights for ArcGIS, GeoAnalytics Server, and GeoEvent Server. Details of these approaches are presented below.

Open Source Software

In order to visualize the risky driving behaviors identified within the BSMP1 dataset, the project team implemented two Web-based applications, one using R Shiny, an open source R package, and the other using Node.js, a JavaScript runtime environment.

R Shiny Application

The R Shiny interactive app was developed in RStudio, using the Leaflet package to create the map visualizations. Two datasets were loaded into the application: the first, a subset (*n* = 10,000) of the risky-labeled BSMP1 data; the second, a dataset in which each observation represents a different intersection and the number of risky monitoring period moments identified at that intersection (this represents all of the risky-labeled BSMP1 data).

Three specific visualizations were generated for the first dataset, the sample of risky-labeled BSMP1 data. The first is a cluster display, in which risky driving observations are clustered on the map in dynamic points which can be selected in order to dive deeper into that cluster. For each level zoomed in, the clusters separate and represent smaller regions. This provides a way to visualize where risky observations are concentrated. The second is a heat map (see Figure 7) of the risky monitoring periods across the map, where an intensification of color (blue being least intense, red being most intense) indicates a higher concentration of risky observations. The third is a map with each observation labeled as an individual point. For all three options, data can be subsetted by time of day (as a range), day of week (as a range), road type, and turning behavior (right turn, left turn). In addition, a table of all data is presented below the maps in the app environment; clicking on an observation in this table highlights that observation on the map.

For the second dataset, the intersection data with risky observation counts, a visualization was generated in which each intersection is labeled with a point. The color of the point is darker based on the increasing number of risky driving monitoring periods observed there. Similar to the prior visualizations, data can be subsetted by time of day, day of week, and turning behavior. The data table below the map can be used to explore the dataset and to identify specific intersections on the map. More details about the R Shiny tool are provided in Appendix H.

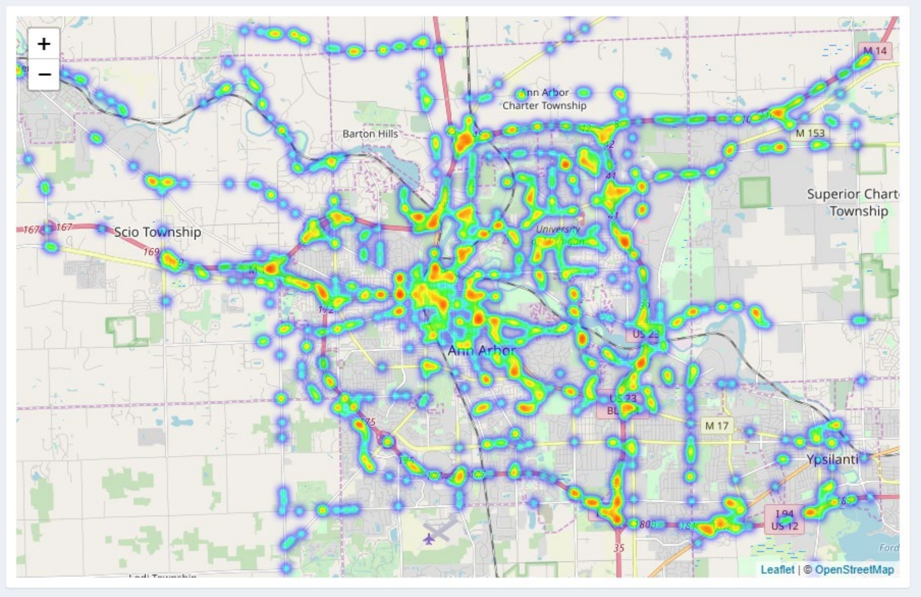
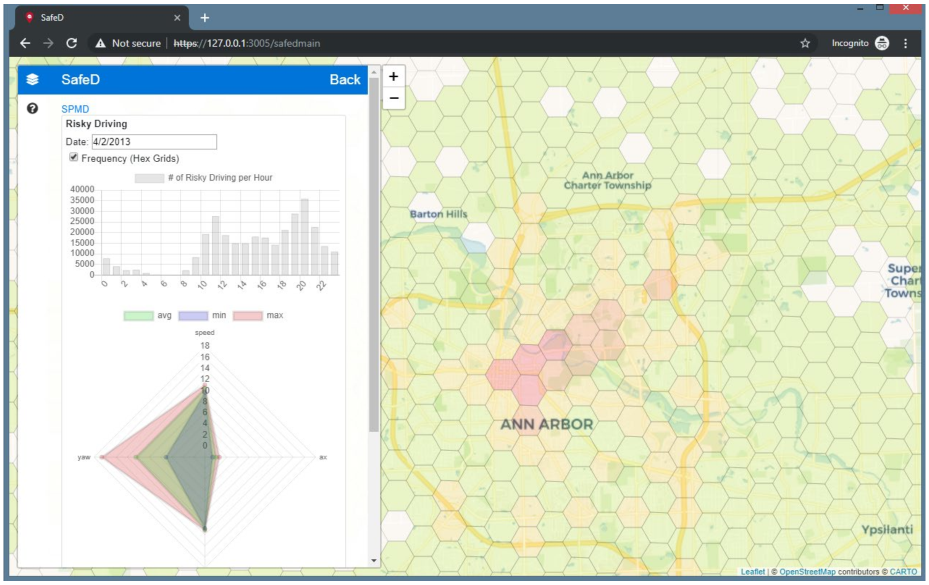


Figure 7. Map. R Shiny heat map of risky driving.



Risky Driving

Date: 4/2/2013

☑ Frequency (Hex Grids)

Figure 8. Screen capture. A Web-GIS application using Node.js.

Node.js Web Application

A Web-based GIS application using Node.js was developed to visualize the spatiotemporal distribution of risky driving behavior (Figure 8). Node.js is a JavaScript runtime built on Chrome's V8 JavaScript engine and uses an event-driven, non-blocking I/O model for developing a lightweight and efficient Web application. Node.js uses JavaScript on the server scripts. Key open source JavaScript libraries for visualization include jquery.js, bootstrap.js, leaflet.js, and chart.js.

In the developed Web application, risky driving incidents were summarized in hexagonal grids at five spatial resolutions, where the heights of a hexagon correspond to 250 m, 500 m, 1 km, 2 km, and 4 km. In the app environment, as a user zooms in on and out of the map, hexagonal grids at an appropriate spatial resolution will be displayed. The color on the hexagonal grids represents the level of risky driving frequency, with red indicating a higher frequency. A bar chart on the top shows the hourly frequency of risky driving on a selected day. The radar chart summarizes risky driving in terms of speed, acceleration, and yaw rate. Both charts are responsive to user interaction and update as a user moves or zooms in/out on the map.

Risky driving data were stored in PostgreSQL and data are dynamically retrieved in response to users' interaction with the application. Secure data transfer between clients and the application server was achieved using a socket protocol. This Web-based application is also capable of displaying additional GIS layers, including a base map, city/county GIS layers, and real-time Waze traffic alerts and jams using Waze APIs.

Enterprise Software

We compared four different commercial GIS data visualization software applications (ArcGIS Pro, ArcGIS Desktop, ArcGIS Insights, and Tableau) and evaluated their performance in handling different dataset sizes. Appendix I illustrates the comparison results. Data visualization examples using each of these tools are provided in Appendix J. According to the testing results for ArcMap, ArcGIS Pro, Tableau, and ArcGIS Insights, in general ArcGIS Pro performed faster than ArcMap in displaying a map on the fly and performing spatial analysis, but neither displays a map completely when the database size is larger than 0.02 GB. In Tableau, a database size smaller than 1 GB allows the map to render completely, but as the dataset gets larger, the operation time increases. ArcGIS Insights does not allow users to upload files over 0.1 GB, and the operation times outs when attempting to handle data sizes close to 0.1 GB. However, the map does display completely with an appropriate file size (0.05 GB and below). Overall, data sizes around 0.05 GB to 0.1 GB were a good fit to perform spatial analysis, as the operation time was less than 1 minute. If 30 minutes is used as a threshold, regardless of displaying problems on the fly, data sizes of 1 GB data or less would be suitable for GIS analysis using traditional GIS software.

Conclusions and Recommendations

This work presents an array of methods and tools that can be applied to store, process, analyze, and visualize large amounts of data for the purpose of identifying risky driving events. Two open source DBMSs, PostgreSQL and PostGIS, were used to develop a relational DBMS. MongoDB was utilized to build a non-relational DBMS. The efficiency of MongoDB and PostgreSQL in handling the data, query attributes, and spatial data analysis was investigated. PostgreSQL and PostGIS performed better than MongoDB on both nonspatial and spatial queries. PostgreSQL and PostGIS outperformed MongoDB as the data size increased. Vertical database partitioning was implemented to improve query performance; however, performance was still limited by the capacity of the database server. Supervised and unsupervised learning algorithms were employed to develop and implement the techniques for creating classification models to identify risky driving events. This involved the creation of the I-DBSCAN algorithm to identify and label risky driving events, the implementation of this algorithm on a subset of the BSMP1 data, the subsequent training of random forest models, and the application of these random forest models to label larger sets of the BSMP1 data. Two open source data visualization tools (one using R Shiny and the other using Node.js) were developed to identity risky driving events in space and time. The following recommendations are provided for dealing with large datasets.

- As part of I-DBSCAN, PCA was used to reduce the data dimensions. PCA can thus contribute to data size reduction, an effective technique when dealing with big data.
- The BSMP1 data points were converted into short monitoring period data, and risky driving was assessed in these monitoring periods. This not only provided additional information in terms of better identifying distinct driving behaviors (e.g., turning), it also significantly reduced data size. We created monitoring periods corresponding to 3 seconds at 1-second intervals. Since time point data were measured every decisecond, each monitoring period was computed using 30 time point data points. Further, this reduced the data size by approximately one order of magnitude.
- Several software packages for data analysis and visualization were evaluated. Current desktop and Web GIS software were not able to handle very large data effectively (>1 GB). Most GIS software was found to have a capacity of 20–50 MB (150–270 K records), which is considerably small. Tableau was determined to be one of the best data visualization tools for BSMP1 data.

It should be noted that the original project included a small study to be performed by a student at the Virginia Tech Transportation Institute designed to determine if there were correlations between the locations where risky driving moments occurred and environmental GIS data. Unfortunately, this portion of the study ended after the student’s unexpected departure from the project and no suitable replacement could be found to continue the work. However, the research team would like to acknowledge the student’s efforts in completing an in-depth literature review to identify priorities and appropriate spatial resolutions and perspectives for the data (see Appendix L). It is recommended that future work be developed around the student’s investigation and findings.

Additional Products

The Education and Workforce Development and Technology Transfer products created as part of this project are located on the project page of the Safe-D website. The final project dataset is located in the Safe-D Collection on the VTTI Dataverse.

Education and Workforce Development Products

The following Education and Workforce Development items resulted from project activities:

1. Two Ph.D. (Charles Marks and Yulu Chen) and two Master's (Eduardo Cordova and Haihong Huang) students were involved in this project. The students learned several methods and skills, such as literature review, cluster analysis, supervised/unsupervised learning, database development, and data visualization tool development. The students also contributed to the publications resulting from this project. Eduardo Cordova's thesis is based on this project.
2. The project contributed to Big Data Science and Analytics Platforms (GEOG-594), taught by Dr. Dr. Ming-Hsiang Tsou during Fall 2018 and again in Fall 2019. The Safe-D project was introduced in the week 2 lecture (What is Big Data?). A dataset from this project is listed as one potential group project topic. One student used the GPS dataset and ArcGIS Insights for a technical demo.
3. The project contributed to Data Management for GIS (GEOG-580), taught by Dr. Atsushi Nara during Fall 2018 and again in Spring 2020. The Safe-D project and its database design using the E-R (Entity-Relationship) diagram were introduced.
4. The project team had a booth display and technology demonstration of Safe-D projects, including the current one, on March 17, 2018, for SDSU Explore Day event at San Diego State University (https://admissions.sdsu.edu/tours_events/explore).

Technology Transfer Products

The following T2 products resulted from project activities:

1. The following journal and conference papers were produced or are underway:
   - A journal paper is currently underway.
   - Chen, Y., M.-H. Tsou, and A. Nara. Analyzing Transportation Big Data with GIS: Detecting Over-speeding Vehicles from Traffic GPS Data. *CSU Geospatial Review,* Vol. 16, 2019, pp. 10-11, https://csugis.sfsu.edu/sites/default/files/19CSU_GeospatialReview_Web.pdf
   - Marks, C., A. Jahangiri, and S. G. Machiani. Iterative DBSCAN (I-DBSCAN) to Identify Aggressive Driving Behaviors within Unlabeled Real-World Driving Data. 22nd Intelligent Transportation Systems Conference, Auckland, New Zealand, 27-30 October 2019.
   - Jahangiri, A., S. G. Machiani, M.-H. Tsou, and A. Nara. Big Data Visualization and Spatiotemporal Modeling of Aggressive Driving using Connected Vehicle Data. 2018 Summer Specialist Meeting (Workshop), Analyzing Social Perception and Amplification using Social Media and Big Data in Human Dynamics, San Diego, CA, August 7-8, 2018.
   - Tsou, M.-H., A. Nara, A. Jahangiri, and S. G. Machiani. Developing Web-based Spatiotemporal Analytics Software Tools for Analyzing Connected Vehicle Data and Aggressive Driving Behaviors. The First Workshop: Geospatial Software:

Connecting Big Data with Geospatial Discovery and Innovation, Los Angeles, CA, January 28-30, 2018.

2. Web-based tools were developed to visualize where and when risky driving occurred as follows:
    - https://charles-marks.shinyapps.io/AggressiveMapper/
    - https://130.191.118.107:3005 (not publicly available due to security issues with large database management systems)
3. A webinar will be held to present the project outcomes.

Data Products

- Link to Dataset – https://doi.org/10.15787/VTT1/KOT55T

- Project Description – The study goal was to identify risky driving behavior using data mining methods within a large dataset. The data used in this study were obtained from the SPMD study conducted in Ann Arbor, Michigan.

- Data Scope – One week of SPMD data was processed to create a data table (each row representing a monitoring period) in csv format, resulting in a table of 320 million observations with 32 variables (i.e., columns).

- Data Specification – A detailed description of each variable in the dataset can be found in Appendix K.

- Citation Metadata:
    - Title of datasets: “SafeD-03-087-Data_x.csv”, x={Apr1,Apr2,Apr4, Apr5, Apr6, Apr7}
    - Author list with researcher ORCIDs
        - Charles Markes, 0000-0002-3893-1914
        - Arash Jahangiri, 0000-0002-8825-961X
        - Ming-Hsiang Tsou, 0000-0003-3421-486X
        - Atsushi Nara, 0000-0003-4173-7773
        - Sahar Ghanipoor Machiani, 0000-0002-7314-2689
    - Contact information (email) for corresponding author: AJahangiri@sdsu.edu

Keywords: risky driving behavior, big data analytics, cluster analysis, data visualization, data mining

References

1. Parry, M. H. *Aggression on the Road : A Pilot Study of Behaviour in the Driving Situation*. London ; New York : Tavistock Publications, 1968.
2. Tasca, L. *A Review of the Literature on Aggressive Driving Research*. Ontario Advisory Group on Safe Driving Secretariat, Road User Safety Branch, Ontario Ministry of Transportation Ontario, Canada, 2000.
3. *National Highway Traffic Safety Administration (1998): Aggressive Drivers View Traffic Differently. (US Department of Transportation Traffic Tech Number 175). Washington DC: US Department of Transportation*. 1998.
4. Boyce, T. E., and E. S. Geller. An Instrumented Vehicle Assessment of Problem Behavior and Driving Style:: Do Younger Males Really Take More Risks? - ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S0001457500001020. Accessed Feb. 25, 2018.
5. Simons-Morton, B. G., Z. Zhang, J. C. Jackson, and P. S. Albert. Do Elevated Gravitational-Force Events While Driving Predict Crashes and Near Crashes? *American Journal of Epidemiology*, Vol. 175, No. 10, 2012, pp. 1075–1079. https://doi.org/10.1093/aje/kwr440.
6. Klauer, S. G., T. A. Dingus, V. L. Neale, J. D. Sudweeks, and D. J. Ramsey. Comparing Real-World Behaviors of Drivers with High Versus Low Rates of Crashes and Near Crashes. 2009.
7. Evans, L. *TRAFFIC SAFETY*. 2004.
8. Aggressive Driving Research Update 2009. 2009, p. 12.
9. Paleti, R., N. Eluru, and C. R. Bhat. Examining the Influence of Aggressive Driving Behavior on Driver Injury Severity in Traffic Crashes. *Accident Analysis & Prevention*, Vol. 42, No. 6, 2010, pp. 1839–1854. https://doi.org/10.1016/j.aap.2010.05.005.
10. Feng, F., S. Bao, J. R. Sayer, C. Flannagan, M. Manser, and R. Wunderlich. Can Vehicle Longitudinal Jerk Be Used to Identify Aggressive Drivers? An Examination Using Naturalistic Driving Data. *Accident Analysis & Prevention*, Vol. 104, 2017, pp. 125–136. https://doi.org/10.1016/j.aap.2017.04.012.
11. Murphey, Y. L., R. Milton, and L. Kiliaris. Driver's Style Classification Using Jerk Analysis. Presented at the 2009 IEEE Workshop on Computational Intelligence in Vehicles and Vehicular Systems, 2009.
12. Li, Y., C. Miyajima, N. Kitaoka, and K. Takeda. Measuring Aggressive Driving Behavior Using Signals from Drive Recorders. Presented at the 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), 2014.
13. González, A. B. R., M. R. Wilby, J. J. V. Díaz, and C. S. Ávila. Modeling and Detecting Aggressiveness From Driving Signals. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 15, No. 4, 2014, pp. 1419–1428. https://doi.org/10.1109/TITS.2013.2297057.
14. Danaf, M., M. Abou-Zeid, and I. Kaysi. Modeling Anger and Aggressive Driving Behavior in a Dynamic Choice–Latent Variable Model. *Accident Analysis & Prevention*, Vol. 75, 2015, pp. 105–118. https://doi.org/10.1016/j.aap.2014.11.012.

15. Jahangiri, A., S. G. Machiani, V. Balali, S. G. Machiani, and V. Balali. Big Data Exploration to Examine Aggressive Driving Behavior in the Era of Smart Cities. *Data Analytics for Smart Cities*. https://www.taylorfrancis.com/. Accessed Jul. 23, 2019.
16. Wang, X., A. J. Khattak, J. Liu, G. Masghati-Amoli, and S. Son. What Is the Level of Volatility in Instantaneous Driving Decisions? *Transportation Research Part C: Emerging Technologies*, Vol. 58, 2015, pp. 413–427. https://doi.org/10.1016/j.trc.2014.12.014.
17. Haglund, M., and L. Åberg. Speed Choice in Relation to Speed Limit and Influences from Other Drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 3, No. 1, 2000, pp. 39–51. https://doi.org/10.1016/S1369-8478(00)00014-0.
18. Yu, J., Z. Chen, Y. Zhu, Y. (Jennifer) Chen, L. Kong, and M. Li. Fine-Grained Abnormal Driving Behaviors Detection and Identification with Smartphones. *IEEE Transactions on Mobile Computing*, Vol. 16, No. 8, 2017, pp. 2198–2212. https://doi.org/10.1109/TMC.2016.2618873.
19. Lee, J., and K. Jang. A Framework for Evaluating Aggressive Driving Behaviors Based on In-Vehicle Driving Records. *Transportation Research Part F: Traffic Psychology and Behaviour*, 2017. https://doi.org/10.1016/j.trf.2017.11.021.
20. Elander, J., R. West, and D. French. Behavioral Correlates of Individual Differences in Road-Traffic Crash Risk: An Examination Method and Findings. *Psychological bulletin*, Vol. 113, No. 2, 1993, pp. 279–294.
21. Lajunen, T., and T. Özkan. Self-Report Instruments and Methods. In *Handbook of Traffic Psychology* (B. E. Porter, ed.), Academic Press, San Diego, pp. 43–59.
22. Mizell, L., M. Joint, and D. Connell. AGGRESSIVE DRIVING: THREE STUDIES. 1997.
23. *National Highway Traffic Safety Administration (2009). Countermeasures That Work: A Highway Safety Countermeasure Guide for State Highway Safety Offices. Report No. DOT HS 811081. National Highway Traffic Safety Administration, Washington, DC.* 2009.
24. Shinar, D. Aggressive Driving: The Contribution of the Drivers and the Situation1Keynote Address Presented at the International Congress of Applied Psychology, August 13, 1998, San Francisco, CA.1. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 1, No. 2, 1998, pp. 137–160. https://doi.org/10.1016/S1369-8478(99)00002-9.
25. Beck, K. H., M. Q. Wang, and M. M. Mitchell. Concerns, Dispositions and Behaviors of Aggressive Drivers: What Do Self-Identified Aggressive Drivers Believe about Traffic Safety? *Journal of Safety Research*, Vol. 37, No. 2, 2006, pp. 159–165. https://doi.org/10.1016/j.jsr.2006.01.002.
26. Balogun, S. K., N. A. Shenge, and S. E. Oladipo. Psychosocial Factors Influencing Aggressive Driving among Commercial and Private Automobile Drivers in Lagos Metropolis. *The Social Science Journal*, Vol. 49, No. 1, 2012, pp. 83–89. https://doi.org/10.1016/j.soscij.2011.07.004.
27. Johnson, D. A., and M. M. Trivedi. Driving Style Recognition Using a Smartphone as a Sensor Platform. Presented at the 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC), 2011.
28. Hong, J.-H., B. Margines, and A. K. Dey. A Smartphone-Based Sensing Platform to Model Aggressive Driving Behaviors. New York, NY, USA, 2014.
29. Jahangiri, A., V. J. Berardi, and S. G. Machiani. Application of Real Field Connected Vehicle Data for Aggressive Driving Identification on Horizontal Curves. *IEEE*

Transactions on Intelligent Transportation Systems, Vol. PP, No. 99, 2017, pp. 1–9. https://doi.org/10.1109/TITS.2017.2768527.

30. Merrikhpour, M., and B. Donmez. Towards Mitigating Teenagers' Distracted Driving Behaviors: Comparison of Real-Time and Post-Drive Feedback in a Simulator Study. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 60, No. 1, 2016, pp. 1879–1883. https://doi.org/10.1177/1541931213601428.
31. Metz, B., A. Landau, and V. Hargutt. Frequency and Impact of Hands-Free Telephoning While Driving – Results from Naturalistic Driving Data. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 29, 2015, pp. 1–13. https://doi.org/10.1016/j.trf.2014.12.002.
32. Romoser, M., M. Deschamps, H. Wilson, and D. Fisher. Investigating Differences Between Experienced Adult Drivers and Teen Drivers with Low-Cost Vehicle Data Recorder. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2321, 2012, pp. 79–87. https://doi.org/10.3141/2321-11.
33. van der horst, R., and J. Hogema. TIME-TO-COLLISION AND COLLISION AVOIDANCE SYSTEMS. 1994.
34. Hankey, J. M., M. A. Perez, and J. A. McClafferty. Description of the SHRP 2 Naturalistic Database and the Crash, Near-Crash, and Baseline Data Sets. 2016.
35. Martins, J., M.-A. Fénart, G. Feltrin, A.-G. Dumont, and K. Beyer. Defining a Braking Probability to Estimate Extreme Braking Forces on Road Bridges. 2015.
36. American Association of State Highway and Transportation Officials. *AASHTO Green: A Policy on Geometric Design of Highways and Streets*. 2001.
37. Jun, J., J. Ogle, and R. Guensler. Relationships Between Crash Involvement and Temporal-Spatial Driving Behavior Activity Patterns: Use of Data for Vehicles with Global Positioning Systems. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2019, 2007, pp. 246–255. https://doi.org/10.3141/2019-29.
38. Fazeen, M., B. Gozick, R. Dantu, M. Bhukhiya, and M. C. González. Safe Driving Using Mobile Phones. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 13, No. 3, 2012, pp. 1462–1468. https://doi.org/10.1109/TITS.2012.2187640.
39. Leavitt, N. Will NoSQL Databases Live Up to Their Promise? *Computer*, Vol. 43, No. 2, 2010, pp. 12–14. https://doi.org/10.1109/MC.2010.58.
40. MongoDB (n.d.). FAQ: MongoDB Storage — MongoDB Manual. *https://github.com/mongodb/docs/blob/v4.0/source/faq/storage.txt*. https://docs.mongodb.com/manual/faq/storage. Accessed Jul. 31, 2019.
41. Charles Marks, Arash Jahangiri, and Sahar Ghanipoor Machiani. Iterative DBSCAN (I-DBSCAN) to Identify Aggressive Driving Behaviors within Unlabeled Real-World Driving Data. Presented at the 22nd Intelligent Transportation Systems Conference, ITSC 2019, Auckland, New Zealand, 2019.
42. Abdi, H., and L. J. Williams. Principal Component Analysis. *Wiley interdisciplinary reviews: computational statistics*, Vol. 2, No. 4, 2010, pp. 433–459.
43. Constantinescu, Z., C. Marinoiu, and M. Vladoiu. Driving Style Analysis Using Data Mining Techniques | Constantinescu | International Journal of Computers Communications & Control. http://univagora.ro/jour/index.php/ijccc/article/view/2221. Accessed Feb. 11, 2018.

44. MacADAM, C., Z. BAREKET, P. FANCHER, and R. ERVIN. Using Neural Networks to Identify Driving Style and Headway Control Behavior of Drivers. *Vehicle System Dynamics*, Vol. 29, No. sup1, 1998, pp. 143–160. https://doi.org/10.1080/00423119808969557.
45. Simons-Morton, B. G., K. Cheon, F. Guo, and P. Albert. Trajectories of Kinematic Risky Driving among Novice Teenagers. *Accident Analysis & Prevention*, Vol. 51, 2013, pp. 27–32. https://doi.org/10.1016/j.aap.2012.10.011.
46. Abou-Zeid, M., I. Kaysi, and H. Al-Naghi. Measuring Aggressive Driving Behavior Using a Driving Simulator: An Exploratory Study. Presented at the 3rd International Conference on Road Safety and SimulationPurdue UniversityTransportation Research Board, 2011.
47. *Hard Brake and Hard Acceleration*. Publication Verizon Telematics Technical Information Bulletin.
48. St John, A. D., and D. R. Kobett. Grade Effects on Traffic Flow Stability and Capacity. *NCHRP report*, No. 185, 1978.
49. Mehar, A., S. Chandra, and S. Velmurugan. Speed and Acceleration Characteristics of Different Types of Vehicles on Multi-Lane Highways. *Trasporti europei (Online)*, No. 55, 2013.
50. Reymond, G., A. Kemeny, J. Droulez, and A. Berthoz. Role of Lateral Acceleration in Curve Driving: Driver Model and Experiments on a Real Vehicle and a Driving Simulator. *Human Factors*, Vol. 43, No. 3, 2001, pp. 483–495. https://doi.org/10.1518/001872001775898188.
51. Kenda, J., and J. Kopač. Measurements and Analyses of Lateral Acceleration in Traffic of Vehicles. *Tehnički vjesnik*, Vol. 18, No. 2, 2011, pp. 281–286.
52. Hamersma, H. A., and P. S. Els. Longitudinal Vehicle Dynamics Control for Improved Vehicle Safety. *Journal of Terramechanics*, Vol. 54, 2014, pp. 19–36.
53. Lee, T., J. Kang, K. Yi, K. Noh, and K. Lee. *Integration of Longitudinal and Lateral Human Driver Models for Evaluation of the Vehicle Active Safety Systems*. Publication 2010-01–0084. SAE International, Warrendale, PA, 2010.
54. Yamakado, M., and M. Abe. An Experimentally Confirmed Driver Longitudinal Acceleration Control Model Combined with Vehicle Lateral Motion. *Vehicle System Dynamics*, Vol. 46, No. sup1, 2008, pp. 129–149. https://doi.org/10.1080/00423110701882363.
55. Wang, J., K. K. Dixon, H. Li, and J. Ogle. Normal Acceleration Behavior of Passenger Vehicles Starting from Rest at All-Way Stop-Controlled Intersections. *Transportation Research Record*, Vol. 1883, No. 1, 2004, pp. 158–166.
56. Kraft, W. H. *Traffic Engineering Handbook*. Institute of Transportation Engineers, Washington, 2010.
57. Levin Vehicle Telematics. https://www.kaggle.com/yunlevin/levin-vehicle-telematics. Accessed May 22, 2018.
58. Data.Gov. *Data.gov*. https://www.data.gov/. Accessed May 22, 2018.
59. DataSF | San Francisco Open Data. https://datasf.org/opendata/. Accessed May 22, 2018.
60. Multi-Modal Intelligent Traffic Signal Systems (MMITSS) Basic Safety Message | Department of Transportation - Data Portal. https://data.transportation.gov/Automobiles/Multi-Modal-Intelligent-Traffic-Signal-Systems-MMI/5tsh-j288. Accessed May 21, 2018.

61. Advanced Messaging Concept Development Basic Safety Message | Department of Transportation - Data Portal. https://data.transportation.gov/Automobiles/Advanced-Messaging-Concept-Development-Basic-Safet/eezi-v4pm. Accessed May 21, 2018.
62. Wyoming CV Pilot Basic Safety Message One Day Sample | Department of Transportation - Data Portal. https://data.transportation.gov/Automobiles/Wyoming-CV-Pilot-Basic-Safety-Message-One-Day-Samp/9k4m-a3jc. Accessed May 21, 2018.
63. Mizell, L. *Aggressive Driving*. AAA Foundation for Traffic Safety, 1997, pp. 4–18.
64. Asbridge, M., R. G. Smart, and R. E. Mann. Can We Prevent Road Rage? *Trauma, Violence, & Abuse*, Vol. 7, No. 2, 2006, pp. 109–121. https://doi.org/10.1177/1524838006286689.
65. Stuster, J. *Aggressive Driving Enforcement: Evaluations of Two Demonstration Programs*. US Department of Transportation, National Highway Traffic Safety Administration, 2004.
66. Burns, R. G., and M. A. Katovich. Examining Road Rage / Aggressive Driving: Media Depiction and Prevention Suggestions. *Environment and Behavior*, Vol. 35, No. 5, 2003, pp. 621–636. https://doi.org/10.1177/0013916503254758.
67. Wang, X., A. J. Khattak, J. Liu, G. Masghati-Amoli, and S. Son. What Is the Level of Volatility in Instantaneous Driving Decisions? *Transportation Research Part C: Emerging Technologies*, Vol. 58, 2015, pp. 413–427. https://doi.org/10.1016/j.trc.2014.12.014.
68. Balogun, S. K., N. A. Shenge, and S. E. Oladipo. Psychosocial Factors Influencing Aggressive Driving among Commercial and Private Automobile Drivers in Lagos Metropolis. *Social Science Journal*, Vol. 49, No. 1, 2012, pp. 83–89. https://doi.org/10.1016/j.soscij.2011.07.004.
69. Casey, S. M., and A. K. Lund. Changes in Speed and Speed Adaptation Following Increase in National Maximum Speed Limit. *Journal of safety research*, Vol. 23, No. 3, 1992, pp. 135–146.
70. Holland, C. A., and M. T. Conner. Exceeding the Speed Limit: An Evaluation of the Effectiveness of a Police Intervention. *Accident Analysis & Prevention*, Vol. 28, No. 5, 1996, pp. 587–597.
71. Dula, C. S., and M. E. Ballard. Development and Evaluation of a Measure of Dangerous, Aggressive, Negative Emotional, and Risky Driving. *Journal of Applied Social Psychology*, Vol. 33, No. 2, 2003, pp. 263–282.
72. Rajalin, S., S.-O. Hassel, and H. Summala. Close-Following Drivers on Two-Lane Highways. *Accident analysis & prevention*, Vol. 29, No. 6, 1997, pp. 723–729.
73. Hennessy, D. A., and D. L. Wiesenthal. The Relationship between Traffic Congestion, Driver Stress and Direct versus Indirect Coping Behaviours. *Ergonomics*, Vol. 40, No. 3, 1997, pp. 348–361. https://doi.org/10.1080/001401397188198.
74. Hauber, A. R. The Social Psychology of Driving Behaviour and the Traffic Environment: Research on Aggressive Behaviour in Traffic. *Applied psychology*, Vol. 29, No. 4, 1980, pp. 461–474.
75. Cinnamon, J., N. Schuurman, and S. M. Hameed. Pedestrian Injury and Human Behaviour: Observing Road-Rule Violations at High-Incident Intersections. *PLOS ONE*, Vol. 6, No. 6, 2011, p. e21063. https://doi.org/10.1371/journal.pone.0021063.
76. Deery, H. A. Hazard and Risk Perception among Young Novice Drivers. *Journal of safety research*, Vol. 30, No. 4, 1999, pp. 225–236.

77. Cackowski, J. M., and J. L. Nasar. The Restorative Effects of Roadside Vegetation: Implications for Automobile Driver Anger and Frustration. *Environment and Behavior*, Vol. 35, No. 6, 2003, pp. 736–751. https://doi.org/10.1177/0013916503256267.
78. Aggressive Driving and The Law: A Symposium. https://one.nhtsa.gov/people/injury/drowsy_driving1/text.htm. Accessed Mar. 3, 2018.
79. NHTSA. Speeding. *NHTSA*. https://www.nhtsa.gov/risky-driving/speeding. Accessed Mar. 3, 2018.
80. Turner, C. W., J. F. Layton, and L. S. Simons. Naturalistic Studies of Aggressive Behavior: Aggressive Stimuli, Victim Visibility, and Horn Honking. *Journal of personality and social psychology*, Vol. 31, No. 6, 1975, pp. 1098–1107.
81. Ellison, P. A., J. M. Govern, H. L. Petri, and M. H. Figler. Anonymity and Aggressive Driving Behavior: A Field Study. *Journal of Social Behavior and Personality; Corte Madera, CA*, Vol. 10, No. 1, 1995, pp. 265–272.
82. Szlemko, W. J., J. A. Benfield, P. A. Bell, J. L. Deffenbacher, and L. Troup. Territorial Markings as a Predictor of Driver Aggression and Road Rage. *Journal of Applied Social Psychology*, Vol. 38, No. 6, 2008, pp. 1664–1688.
83. Kaysi, I. A., and A. S. Abbany. Modeling Aggressive Driver Behavior at Unsignalized Intersections. *Accident Analysis & Prevention*, Vol. 39, No. 4, 2007, pp. 671–678. https://doi.org/10.1016/j.aap.2006.10.013.
84. Fang, F. C., and H. Castaneda. Computer Simulation Modeling of Driver Behavior at Roundabouts. *International Journal of Intelligent Transportation Systems Research*, Vol. 16, No. 1, 2018, pp. 66–77. https://doi.org/10.1007/s13177-017-0138-2.
85. Madanat, S. M., M. J. Cassidy, and M.-H. Wang. Probabilistic Delay Model at Stop-Controlled Intersection. *Journal of transportation engineering*, Vol. 120, No. 1, 1994, pp. 21–36.
86. Raff, M. S. A Volume Warrant for Urban Stop Signs. 1950.
87. Rakha, H., I. El-Shawarby, and J. R. Setti. Characterizing Driver Behavior on Signalized Intersection Approaches at the Onset of a Yellow-Phase Trigger. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 8, No. 4, 2007, pp. 630–640. https://doi.org/10.1109/TITS.2007.908146.
88. Shinar, D., M. Bourla, and L. Kaufman. Synchronization of Traffic Signals as a Means of Reducing Red-Light Running. *Human factors*, Vol. 46, No. 2, 2004, pp. 367–372.
89. Shinar, D. Aggressive Driving: The Contribution of the Drivers and the Situation1Keynote Address Presented at the International Congress of Applied Psychology, August 13, 1998, San Francisco, CA.1. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 1, No. 2, 1998, pp. 137–160. https://doi.org/10.1016/S1369-8478(99)00002-9.
90. Dollard, J., N. E. Miller, L. W. Doob, O. H. Mowrer, and R. R. Sears. Frustration and Aggression - I Definitions. In *Frustration and Aggression*, pp. 1–26.
91. Deffenbacher, J. L., E. R. Oetting, and R. S. Lynch. Development of a Driving Anger Scale. *Psychological reports*, Vol. 74, No. 1, 1994, pp. 83–91.
92. Harwood, L. C., and Z. R. Doerzaph. Visualization of the Effect of Topography on Connected Vehicle Communications Using LiDAR-Derived Models and Interactive Mapping Techniques. *INTERNATIONAL JOURNAL OF TRANSPORTATION*, Vol. 6, No. 1, 2018, pp. 15–28.

93. Staubach, M. Factors Correlated with Traffic Accidents as a Basis for Evaluating Advanced Driver Assistance Systems. *Accident Analysis & Prevention*, Vol. 41, No. 5, 2009, pp. 1025–1033. https://doi.org/10.1016/j.aap.2009.06.014.

94. Tandy, C. The Isovist Method of Landscape Survey. *Methods of Landscape Analysis*, 1967.

95. Lynch, K. Managing the Sense of Region. 1976.

96. Benedikt, M. L. To Take Hold of Space: Isovists and Isovist Fields. *Environment and Planning B: Planning and design*, Vol. 6, No. 1, 1979, pp. 47–65.

97. US Department of Commerce, N. O. and A. A. What Is LIDAR. https://oceanservice.noaa.gov/facts/lidar.html. Accessed Mar. 5, 2018.

98. Harwood, L. C., and Z. R. Doerzaph. Lidar: Another Potential Data Source. Blacksburg, Virginia, Aug 25, 2014.

99. Wehr, A., and U. Lohr. Airborne Laser Scanning—an Introduction and Overview. *ISPRS Journal of photogrammetry and remote sensing*, Vol. 54, No. 2–3, 1999, pp. 68–82.

100. de Santos-Berbel, C., M. Essa, T. Sayed, and M. Castro. Reliability-Based Analysis of Sight Distance Modelling for Traffic Safety. *Journal of Advanced Transportation*, Vol. 2017, 2017, pp. 1–12. https://doi.org/10.1155/2017/5612849.

101. Castro, M., and C. De Santos-Berbel. Spatial Analysis of Geometric Design Consistency and Road Sight Distance. *International Journal of Geographical Information Science*, Vol. 29, No. 12, 2015, pp. 2061–2074. https://doi.org/10.1080/13658816.2015.1037304.

102. Castro, M., J. A. Anta, L. Iglesias, and J. A. Sánchez. GIS-Based System for Sight Distance Analysis of Highways. *Journal of computing in civil engineering*, Vol. 28, No. 3, 2013, p. 04014005.

103. Bartie, P., and M. P. Kumler. Route Ahead Visibility Mapping: A Method to Model How Far Ahead a Motorist May View a Designated Route. *Journal of Maps*, Vol. 6, No. 1, 2010, pp. 84–95. https://doi.org/10.4113/jom.2010.1107.

104. Llobera, M. Extending GIS-Based Visual Analysis: The Concept of Visualscapes. *International journal of geographical information science*, Vol. 17, No. 1, 2003, pp. 25–48.

105. Castro, M., C. De Santos-Berbel, and L. Iglesias. A Comprehensive Methodology for the Analysis of Highway Sight Distance. 2017. https://doi.org/10.1201/9781315281896.

106. Macdonald, E., A. Harper, J. Williams, and J. A. Hayter. *Street Trees and Intersection Safety*. Publication 2006,11. Institution of Urban and Regional Development, Berkeley, California, 2006, p. 104.

107. Kato, A., L. M. Moskal, P. Schiess, M. E. Swanson, D. Calhoun, and W. Stuetzle. Capturing Tree Crown Formation through Implicit Surface Reconstruction Using Airborne Lidar Data. *Remote Sensing of Environment*, Vol. 113, No. 6, 2009, pp. 1148–1162. https://doi.org/10.1016/j.rse.2009.02.010.

108. T. J. Boyle Associates. *DPS Verizon Waterbury Communications Tower Project*. Public Services Department, Waterbury, Vermont, 2015.

Appendix A. Summary of Studies on Driving Style Categorization

Reference	Driving Style Categories	Variables	Boundary	Method	Supervised or Unsupervised	Accuracy of model	Number of observation	Data Source
Johnson and Trivedi (*27*)	Typical (Non-aggressive), Aggressive	NA	NA	Dynamic Time Warping (DTW) system and smartphone-based sensor-fusion, KNN	Supervised	97% correctly identified	200 driver events (about 50 aggressive events)	Real Field Data
Hong, Margines, and Dey (*28*)	Violator, Non-violator, Aggressive, and Calm	NA	NA	Naïve Bayes classifier	Supervised	90% with violation-class and 81% with questionnaire-class	22 drivers	Real Field Data
González et al. (*13*)	Smooth and Aggressive	Lateral and Longitudinal Accelerations, Speed	NA	Gaussian mixture model (GMM) and maximum likelihood classifier	Unsupervised	92.3%	10 drivers	Real Field Data
Jahangiri, Berardi, and Machiani (*29*)	Aggressive and Normal	TLC	Less than 0.5 seconds as aggressive	Random Forest	Supervised	NA	About 2700 vehicle in SPMD	Open source data-SPMD
Constantinescu, Marinoiu, and Vladoiu (*43*)	Aggressivity Level: moderately low, very low, moderately high, high, and neutral	Speed, Acceleration, Braking, Mechanical work	Based on ranges for Principle Components	Hierarchical cluster analysis with Ward's method and Euclidian distance	Unsupervised	NA	23 drivers and 2 additional controlled test drivers	Real Field Data
Wang et al. (*16*)	Volatile and Typical Driving	Jerk, Acceleration	mean plus/minus one standard deviation	Mixed-effect regression model	Supervised	NA	1653 drivers, 51370 trips	Open source data-Atlanta Regional Commission in 2011

Murphey, Milton, and Kiliaris (*11*)	Calm, Normal, Aggressive driving, No speed	Jerk	Jerk ratio < 1.0 (a division of jerk standard deviation by jerk mean)	Self-developed classification method by using Jerk	Supervised	NA	11 drivers	Open source data-PSTA (Powertrain System Analysis Toolkit)
MacADAM et al. (*44*)	Aggressivity level based on a proposed index	Distance from the lead vehicle to the host vehicle	Distance between leading vehicle and host vehicle	Neural Network	Supervised	NA	36 drivers (drivers graph on page 155)	Real Field Data
Simons-Morton et al. (*45*)	Low-risk, High-risk	Acceleration, deceleration, yaw rate	NA	latent class model	Unsupervised	NA	42 drivers	Naturalistic Teenage Driving Study (NTDS)
Lee and Jang (*19*)	Aggressive and normal driving	Speed, yaw rate, acceleration	Based on clusters	self-organizing map (SOM) and k-means	Unsupervised	NA	43 taxi drivers	Real Field Data
Feng et al. (*10*)	Aggressive and normal driving	Two jerk-based metrics	NA	Receiver Operating Characteristic (ROC)	Supervised	NA	108 drivers	Real Field Data
Li et al. (*12*)	Aggressiveness level based on a score	Velocity, acceleration, deceleration	Aggressiveness Score 1 to 5	multiple linear regression Principle Component Regression (PCR)	Supervised and Unsupervised	NA	78 drivers	Real Field Data
Yu et al. (*18*)	Normal driving and abnormal driving such as Weaving, Swerving, Sideslipping, Fast U-turn, Turning with a wide radius, and Sudden Braking	Acceleration and Orientation	NA	SVM and NN	Supervised	95%	20 drivers	Real Field Data

Abou-Zeid, Kaysi, and Al-Naghi (*46*)	aggressive and timid driver	Based on Questionnaire	Score> 3: aggressive	t-test analysis for calculating drivers' variable (speed, acceleration) after classification	NA	NA	27 timid and 17 aggressive drivers	Real Field Data

Appendix B. Kinematic Data Thresholds Suggested by Different Studies

Reference	Variable	Threshold	Categories
Merrikhpour and Donmez (*30*)	acceleration	0.6 g	Unsafe braking and safe braking
Metz, Landau, and Hargutt (*31*)	acceleration	-4 m/s^2 to -6 m/s^2 (-0.4 g to -0.6 g)	Sharp braking
Romoser et al. (*32*)	acceleration	3.5 m/s^2 and -6 m/s^2 (0.35 g and -0.6 g)	Hard start, hard stop
Fazeen et al. (*38*)	acceleration	3 m/s^2 (0.3 g)	Safe and unsafe acceleration and deceleration
Jun, Ogle, and Guensler (*37*)	acceleration	6 mph/s (0.3 g)	Hard acceleration
van der horst and Hogema (*33*)	acceleration	-5 m/s2 (-0.5 g)	Hard braking
Martins et al. (*35*)	acceleration	4 m/s2 (0.4 g)	Hard braking
American Association of State Highway and Transportation Officials (*36*)	acceleration	3.4 m/s2 (0.34 g)	Comfortable deceleration
Hankey, Perez, and McClafferty (*34*)	Longitudinal acceleration, lateral acceleration, swerve, yaw rate, longitudinal jerk	Longitudinal deceleration: -0.65g Longitudinal acceleration: 0.5g Freeway deceleration: -0.3g Lateral acceleration: -0.75g Swerve: ±15 deg/s/s Yaw rate: ±8deg/s Jerk: -0.1g/s	NA
Feng et al. (*10*)	Longitudinal acceleration	0.6 g	High-crash risk, low-crash risk
Verizon (*47*)	acceleration	Hard braking: 3.92 m/s2 (0.4 g) Hard acceleration: 3.53 m/s2 (0.36 g)	Hard braking and hard acceleration

Appendix C. Lateral and Longitudinal Acceleration Extremes

To detect outliers of lateral and longitudinal acceleration variables, finding the common range of the variables for normal drivers is crucial. In this regard, several studies were reviewed.

John and Kobette found that maximum acceleration (presumable longitudinal acceleration) for passenger car and heavy vehicles are 3.36 m/s^2 and 5.19 m/s^2 respectively (*48*). Another study stated that 2.5 m/s^2 is the maximum longitudinal acceleration for passenger cars and heavy vehicles (*49*).

Reymond et al. investigated the lateral acceleration of drivers at curves by modeling the lateral acceleration based on experimental data. The model can categorize drivers into normal or fast based on the variation of model parameters. For normal driving, the range of lateral acceleration experiments reached by one of the experiments was about 7 m/s^2 and 8.5 m/s^2 for normal and fast drivers, as shown in the figure below (*50*).

Kenda et al found that maximum lateral acceleration is 9.7 m/s^2, longitudinal acceleration equals 4.8 m/s^2 and braking deceleration is -2 m/s^2 (*51*). While they prescribe that the maximum lateral and longitudinal accelerations for a comfortable driver are 2.65 m/s^2 and 2.5 m/s^2.

Hamersma et al. stated that the maximum lateral and longitudinal accelerations measured during a double lane change were 4 m/s^2 and 3.2 m/s^2 (*52*). Another study observed a maximum lateral acceleration of 7 m/s^2 in a simulation study (*53*).

Yamakado declared that a vehicle acceleration never reaches 10 m/s^2 . The maximum range of lateral acceleration they observed was -6 m/s^2 to +8 m/s^2 and the maximum range of longitudinal acceleration was -3 m/s^2 to +8 m/s^2 (*54*). Alternatively, maximum normal lateral and longitudinal accelerations of passenger vehicles measured about 5 m/s^2, as shown below (*55*).

Maximum acceleration rate documented in the ITE Traffic Engineering Handbook is 9.3 ft/s^2 (~3 m/s^2) (*56*).

According to the aforementioned studies, maximum threshold for lateral acceleration can be assumed as 9.7 m/s^2 (-9.7 m/s^2 to 9.7 m/s^2), and longitudinal acceleration range as -3 m/s^2 to +8 m/s^2. The shorter range can also be assumed for both variables. However, in this study we took a more conservative approach to avoid removing aggressive drivers as outliers.

Appendix D. Other Data Sources

Kaggle is one of the platforms in which companies and users share datasets. Data scientists and researchers compete to propose solutions to solve different problems using Kaggle datasets. Different keywords related to the study were searched on Kaggle, including car, vehicle, vehicle kinematic data, vehicle speed, connected vehicle, etc. Among all suggested datasets, "Levin vehicle telematics" included some kinematic data, such as speed and acceleration for around 30 vehicles in 4 months; however, none of this data contained longitude and latitude information (*57*). Data.gov and data.sfgov.org are other data sources that were checked. Several keywords were used, but no dataset was found related to our project (*58*, *59*).

The Atlanta Regional Commission (ARC) is a regional coordination agency conducting management and planning projects (e.g., planning new transportation options, wisely managing such things as water resources, etc.). In 2011, ARC conducted a Household Travel Survey that was used in (*16*). It appears that the data collected contained the vehicle's kinematic data. However, the data was not available on their website.

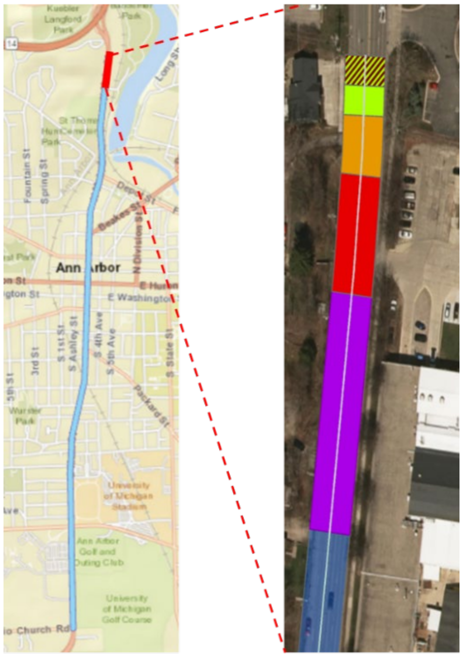
Transportation.gov is another website that has many datasets in different categories. The Automobile category contains many transportation-related studies and their datasets. Among those studies, several datasets with kinematic vehicle data that could potentially be used for risky driving identification are as follows:

- Multi-Modal Intelligent Traffic Signal Systems (MMITSS): in the metadata, there are 11 datasets, 8 of which are available online and three of which are missing. Among 8 datasets, 2 had vehicle location and vehicle kinematic data (*60*).
- Advanced Messaging Concept Development (AMCD): in the metadata, there are five datasets, all of which are available online. Three out of five datasets included vehicle location and kinematic data (*61*).
- Wyoming Connected Vehicle (CV) Pilot: the metadata was not found online. However, two datasets are available, and one has vehicle locations and kinematic data (*62*).

Appendix E. Database Subsets

Subset Number	Number of GPS Points	Size in a csv format
1	10	2 KB
2	100	19 KB
3	1,000	182 KB
4	10,000	1,765 KB
5	100,000	17 MB
6	1,000,000	176 MB
7	10,000,000	1.756 GB
8	100,000,000	17.696 GB
9	500,000,000	88.871 GB

Appendix F. Road Buffers for an Intersection Query



Appendix G. Query Language Examples

	PostgreSQL (SQL)	MongoDB (Javascript)

Query 1	SELECT COUNT(*)
	FROM subset1
	WHERE
	EXTRACT(HOUR FROM timestamp) = 1;
		db.subset1.aggregate(
		  [
		    {
		      $project:{
		        hour: {$hour:"$timestamp"}
		      }
		    },
		    {$match: {hour:1}},
		    {$count: "timestamp"}
		  ]
		)

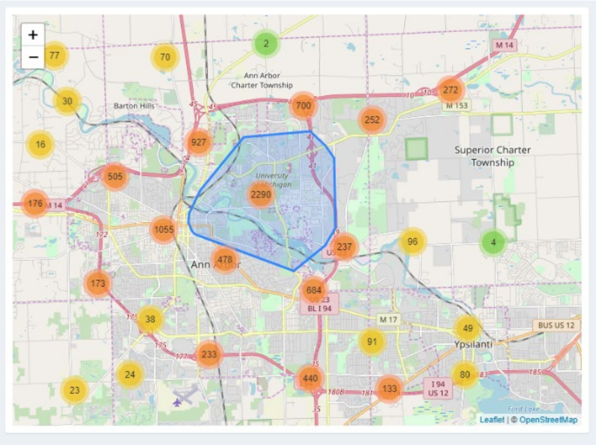
Query 2	SELECT COUNT(*)
	FROM subset1
	WHERE speed > 30;
		db.subset1.find(
		  {"speed": {$gt: 30}}
		).count()

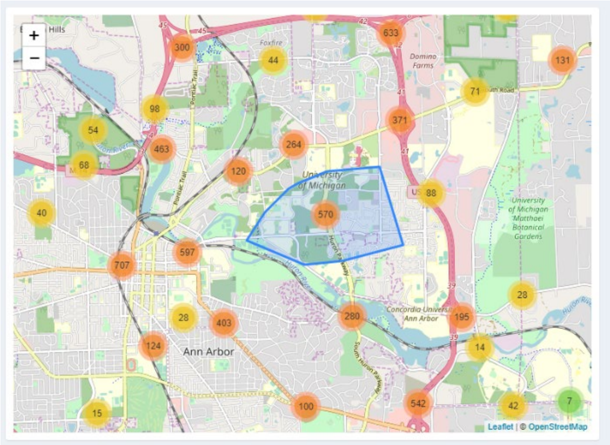
Query 3	SELECT COUNT(*)
	FROM road_buff t1, subset t2
	WHERE
	t1.road_id=1
	AND
	ST_Intersects(t1.geom, t2.geom);
		var road = db.road_buff.findOne({“_id” : 1})
		
		db.subset1.find{
		  {
		    geometry: {
		      $geoWithin: {
		        $geometry: road.geometry
		      }
		    }
		  }
		}.count()

Query 4	SELECT COUNT(*)
	FROM circle t1, subset t2
	WHERE
	t1.gid=1
	AND
	ST_Intersects(t1.geom, t2.geom);
		var circle =
		db.circles.findOne({"properties.gid" : 1})
		
		db.subset1.find{
		  {
		    geometry: {
		      $geoWithin: {
		        $geometry: circle.geometry
		      }
		    }
		  }
		}.count()

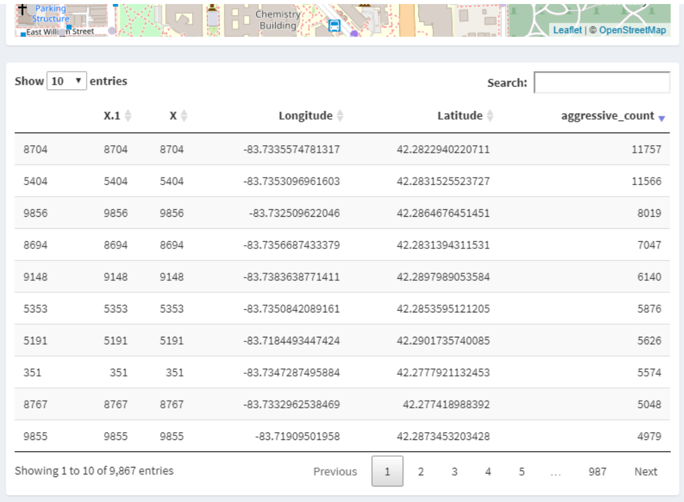
Appendix H. R Shiny Visualization App

The R-Shiny application has four modules, as described: three for the first dataset of all risky driving observations and one for the second dataset of intersection-specific counts of risky driving. The first is the cluster marker setting, in which clusters in a similar region on the map are clustered into clickable group. When clicked on, the map zooms in on the cluster and splits the cluster into sub-clusters, as depicted in the following images**.** This clustering method allows for the identification of regions (whether larger regions of the county or specific localities in the area) in which concentrations of risky driving have been observed and identified.





For the second dataset, the intersection specific data, each intersection was labeled as a point and the darker (closer to purple) and more opaque color corresponded to a greater number of risky driving points. Hovering over individual intersection points displays a label with the number of observed risky driving observations at specific intersection. In order to aid in identifying intersections with the highest concentrations of risky driving, an interactive data table was included as depicted in the images below. The table can be sorted by number of risky incidents and then observations can be clicked, resulting in the highlighting of the corresponding intersection on the map. This allows users to identify the specific intersections at which the highest concentrations of risky driving have been identified.



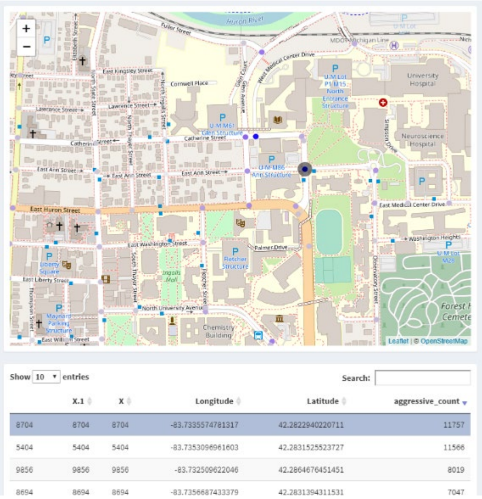
Show 10 ▾ entries

Search:

	X.1	X	Longitude	Latitude	aggressive_count
8704	8704	8704	-83.7335574781317	42.2822940220711	11757
5404	5404	5404	-83.7353096961603	42.2831525523727	11566
9856	9856	9856	-83.732509622046	42.2864676451451	8019
8694	8694	8694	-83.7356687433379	42.2831394311531	7047
9148	9148	9148	-83.7383638771411	42.2897989053584	6140
5353	5353	5353	-83.7350842089161	42.2853595121205	5876
5191	5191	5191	-83.7184493447424	42.2901735740085	5626
351	351	351	-83.7347287495884	42.2777921132453	5574
8767	8767	8767	-83.7332962538469	42.277418988392	5048
9855	9855	9855	-83.71909501958	42.2873453203428	4979

Showing 1 to 10 of 9,867 entries

Previous 1 2 3 4 5 … 987 Next



Show 10 ▾ entries

Search:

	X.1	X	Longitude	Latitude	aggressive_count
8704	8704	8704	-83.7335574781317	42.2822940220711	11757
5404	5404	5404	-83.7353096961603	42.2831525523727	11566
9856	9856	9856	-83.732509622046	42.2864676451451	8019
8694	8694	8694	-83.7356687433379	42.2831394311531	7047

Appendix I. Commercial GIS Software

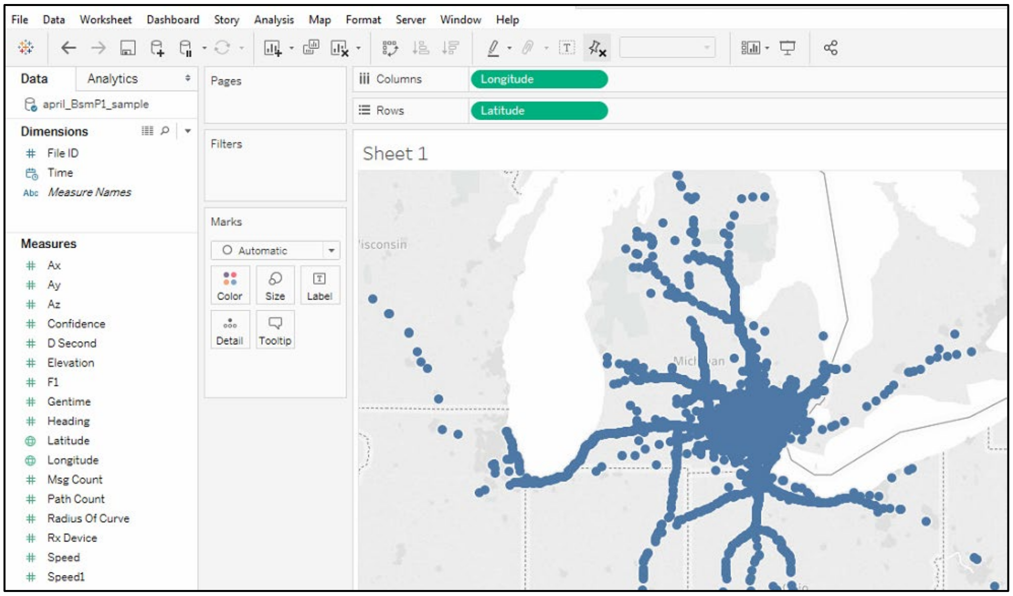
Software name	Size of BSM_p1 (csv_ files					
	11GB 57,764,257 records	1GB around 5,000,000 records	0.5GB around 2,500,000 records	0.1GB around 500,000 records	0.05GB 270,000 records	0.02GB 156,411 records
ArcGIS Pro	• 1 hrs + • layer didn’t show up • program slowed down greatly and other function cannot execute properly	• processing time : 7 mins 34 seconds • layer didn’t show up completely • program slowed down	• processing time : 3 mins 57 seconds • layer didn’t show up completely • software slow down	• processing time : 43.55 seconds • layer didn’t show up completely • other functions like spatial join still can be execute properly	• processing time : 21.81 seconds • layer didn’t show up completely • other functions execute properly	executed very fast and data display fine
ArcGIS Desktop	More than 30 mins	• processing time : around 15 mins • layer didn’t show up completely • other functions like spatial join still can be execute properly	• processing time : around 8 mins • layer didn’t show up completely • other functions like spatial join still can be execute properly	• processing time : around 1 mins • layer didn’t show up completely • other functions like spatial join still can be execute properly	• processing time : around 38 seconds • layer didn’t show up completely • other functions like spatial join still can be execute properly	executed very fast and data display fine
ArcGIS Insight	do not accept csv files over 100MB			• takes around 1 min to load csv file • time out for location enabling function	• takes around 35 second to load csv file • take around 1.5 minutes to enable location function • could not display all features	works fine, every execution takes around 6 seconds
Tableau	execution of query very slow, more than 30 mins	• Executions take around 47 seconds • Map layers show up • operation slow down (6 seconds each)	• Executions take around 38 seconds • Map layers show up • operation (drag and drop) slow down	• Map Layer show up in a few seconds • operation slow down a little	• Map Layer show up • operation works fine	• Map Layer show up • operation works fine

Appendix J. Data Visualization Examples with Commercial Software

Tableau:

Tableau is one of the most popular data visualization software packages. As shown in

Figure 9 The BSM data visualization using Tableau software, we used this software with database connection to create a set of small tables by initial query (test for 11 GB, 1 GB, 0.5 GB, 0.05 GB, and 0.02 GB). The loading speed of database connection is fine with 11 GB of data, but for operations such as drag and drop for visualization, the execution of the query was very slow (more than 30 minutes). When we used the 1 GB csv file, the execution took around 47 seconds; Map layers showed up but operations slowed down (6 seconds for each operation). When we used 0.05 and 0.02 GB sample csv data, the software worked smoothly.



Columns: Longitude

Rows: Latitude

Figure 9. The BSM data visualization using Tableau software.

We also used the Tableau to perform other data visualization tasks as follows:

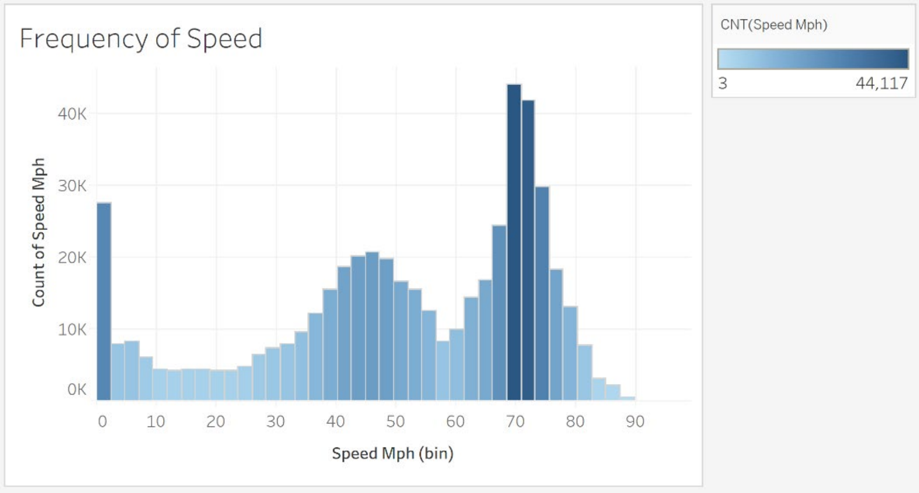


Figure 10. Frequency of speed.

Figure 10 is a histogram of speed using Tableau. The continuous numeric speed data has been converted into several bins. This chart shows the highest frequency speed is 70 mph, which indicates the highway speed.

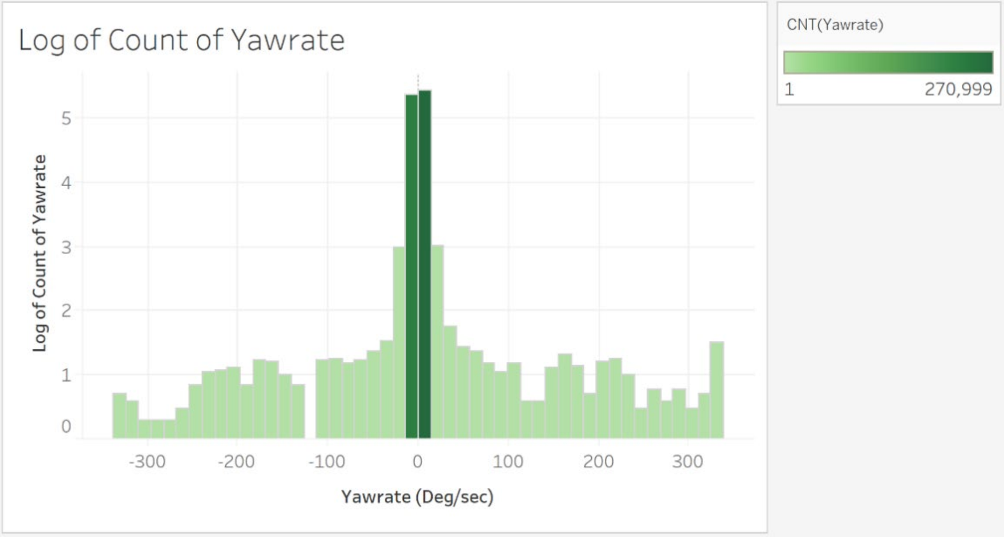


Figure 11. Log of count of yaw rate.

Figure 11 is a histogram of yawrate. The continuous numeric yawrate data has been converted into several bins. Log transformation has been used to show the frequency of yawrate. The frequency of a value of 0 is much higher than the frequency of other values, which indicates driving straight without any turning.

ArcGIS Desktop:

ArcGIS is one of the most popular Geographic Information Systems (GIS) software packages for mapping and spatial analysis. ArcGIS cannot handle csv files larger than 0.1 GB effectively. When we used a 0.05 GB csv file, it took 21.81 seconds for the map to finish displaying. Figure **12** shows that only certain points are displayed properly, while other points are hidden in the dark lines. However, operations such as spatial join still work fine. When we used a 0.02 GB file, the program worked smoothly. ArcGIS has better spatial analysis functionality compared to Tableau.

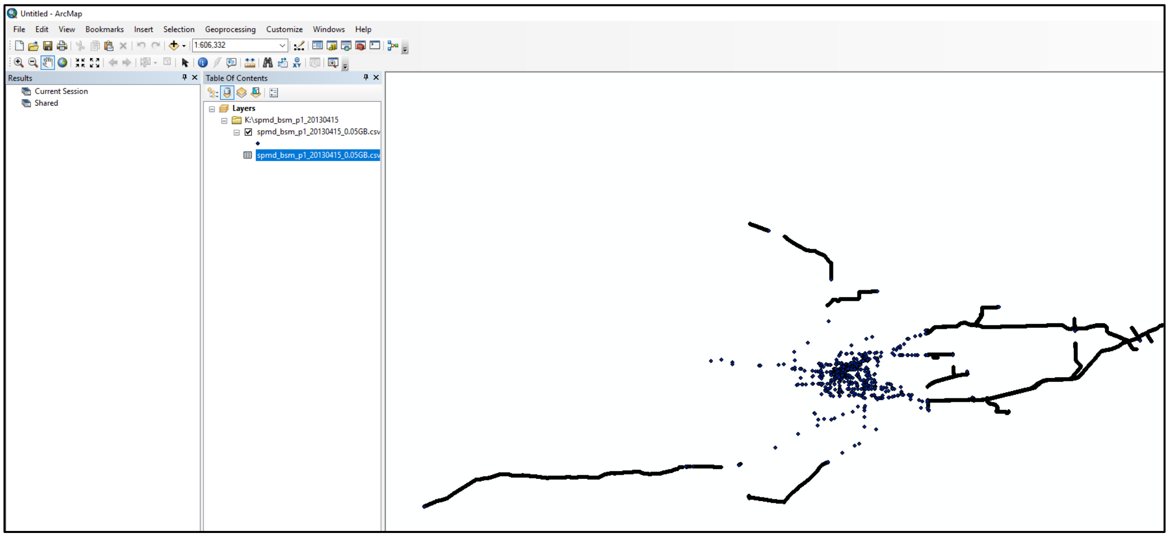
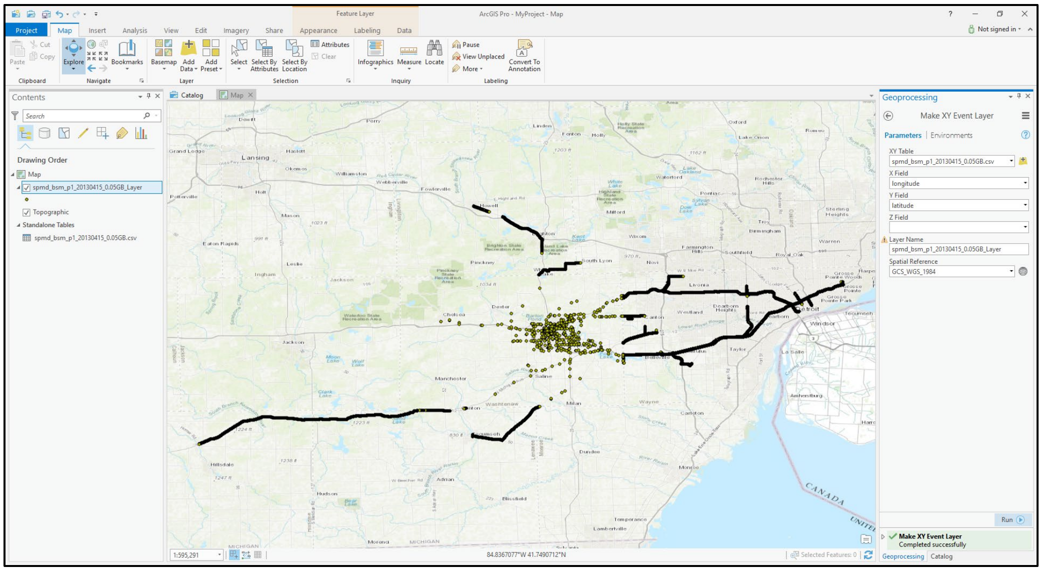


Figure 12. Partial display problem in ArcGIS when using 0.05GB file size.

ArcGIS Pro:

ArcGIS Pro is the latest professional GIS software from Esri (replacement of ArcGIS desktop). The performance of ArcGIS Pro performed slightly better than ArcGIS desktop in our testing. All major functions and data visualization tools are very similar to ArcGIS Desktop. Figure 13 is shows a screenshot of the ArcGIS Pro with the BSM data.



Geoprocessing

Make XY Event Layer

Parameters | Environments

XY Table: spmd_bsm_p1_20130415_0.05GB.csv

X Field: longitude

Y Field: latitude

Z Field:

Layer Name: spmd_bsm_p1_20130415_0.05GB_Layer

Spatial Reference: GCS_WGS_1984

Run

Make XY Event Layer
Completed successfully

Figure 13. Using ArcGIS Pro to display the 0.05GB file.

ArcGIS Insights

ArcGIS Insights is a web-based data analytics tool for exploring spatial and nonspatial data. However, it does not allow upload size over 100 MB (0.1 GB) with a regular user account. Figure 14 and Figure 15 are the visualization of 0.02 GB and 0.05 GB files.

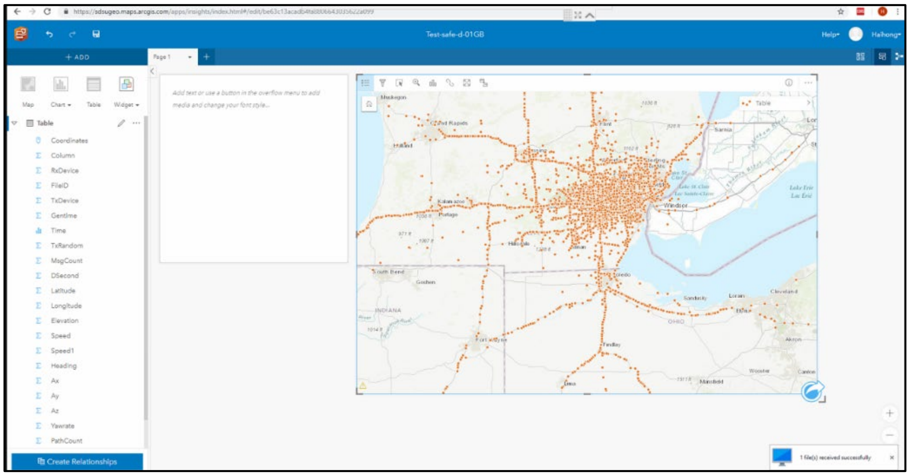
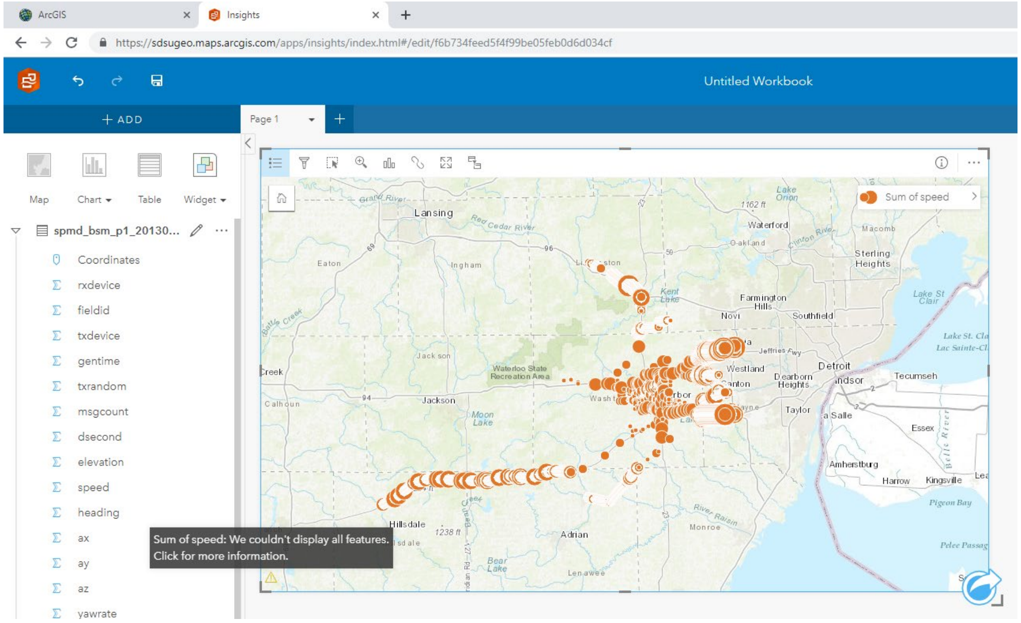

https://sdsugeo.maps.arcgis.com/apps/insights/index.html#/edit/be63c13acadb4fa8806643035622a099
Test-safe-d-01GB
Help
+ ADD
Page 1
Map
Chart
Table
Widget
Table
Coordinates
Column
RxDevice
FileID
TxDevice
Gentime
Time
TxRandom
MsgCount
DSecond
Latitude
Longitude
Elevation
Speed
Speed1
Heading
Ax
Ay
Az
Yawrate
PathCount
Create Relationships
Add text or use a button in the overflow menu to add media and change your font style...
1 file(s) received successfully

Figure 14. ArcGIS insights with 0.02GB csv file.



https://sdsugeo.maps.arcgis.com/apps/insights/index.html#/edit/f6b734feed5f4f99be05feb0d6d034cf

Sum of speed: We couldn't display all features.
Click for more information.

Figure 15. Visualization of speed info using ArcGIS insights.

Appendix K. Monitoring Period Data Specification

The monitoring period data generated from the BSMP1 data was used to identify risky driving (using a monitoring period length of 3 seconds, taken at 1-second intervals). Monitoring period data for the first week of April 2013 have been made available and the data is specified as follows:

Variable Name	Type	Unit	Description
Vehicle_id	Integer	None	SPMD vehicle identifier
Trip_id	Integer	None	Trip identifier, used to identify points representing a continuous trip
Starttime	Datetime	None	Datetime object representing the beginning of the monitoring period (MP)
Start_latitude	Float	Degrees	Latitude at beginning of MP
Start_longitude	Float	Degrees	Longitude at beginning of MP
End_latitude	Float	Degrees	Latitude at end of MP
End_longitude	Float	Degrees	Longitude at end of MP
Avg_speed	Real	m/s	Average (mean) speed across entire MP
Max_speed	Real	m/s	Max recorded speed across MP
Min_speed	Real	m/s	Min recorded speed across MP
Sd_speed	Real	None	Standard deviation (SD) of speed across MP
Avg_ax	Real	m/s^2	Average acceleration across MP
Max_ax	Real	m/s^2	Max recorded accel across MP
Min_ax	Real	m/s^2	Min record accel across MP
Sd_ax	Real	None	SD of accel across MP
Jerk_ax	Real	m/s^3	Jerk of accel (comparing end of MP against start)
Avg_ay	Real	m/s^2	Average lateral accel across MP (subject to high levels of measurement error)
Max_ay	Real	m/s^2	Max recorded lateral accel across MP
Min_ay	Real	m/s^2	Min recorded lateral accel across MP
Sd_ay	Real	None	SD of lateral accel across MP
Avg_yaw	Real	Deg/sec	Average yaw rate across MP
Max_yaw	Real	Deg/sec	Max yaw rate across MP
Min_yaw	Real	Deg/sec	Min yaw rate across MP
Sd_yaw	Real	None	SD of yaw rate across MP
Jerk_yaw	Real	Deg/sec^2	Jerk of yaw rate (comparing end of MP against start)

Road_type	Text	None	Categorical variable indicating roadclass (such as highway, local, etc)
Initial_heading	Float	Degrees	Heading of vehicle at beginning of MP (0 = 360 = North)
End_heading	Float	Degrees	Heading of vehicle at end of MP
Sd_heading	Real	None	SD of heading across monitoring period
Change_heading	Float	Degrees	Change in heading from beginning of MP to end
Pid	Text	None	A list of the identifiers that can be used to match the MP data up with the original BSMP1 data
Risky	Boolean	None	If True, identified as risky in our analysis

Appendix L. Literature Review: Identifying Aggressive Driving Locations and Environmental Attributes

This following literature review was conducted by Joshua Starner, a master's student at Virginia Tech, during the initial stages of this project. This work informed the project and is thus included here for the reader, but was not incorporated into the analytic basis of the project. This in-depth literature review will be useful in developing future work in correlating environmental factors and locations where risky or aggressive driving behaviors occur.

Review of Relevant Literature

The objectives of this project include quantifying environmental factors to correlate aggressive driving with the physical environment surrounding a roadway. A literature review of psychological, sociological, transportation engineering, and other related research was conducted to identify priorities and appropriate spatial resolutions and perspectives for the data.

Correlating the Physical Environment and Aggressive Driving

Across fields, a large portion of the research treats aggressive driving as a form of violence, where a driver has the intent to injure or kill another individual (*63*), which has allowed the near interchangeable use of the terms "road rage" and "aggressive driving." In many cases, this overlap of terms is intentional, where the term "road rage" is used to describe an extreme form of aggressive driving. Within transportation, however, most definitions do not require violent intent and allow for "road rage" and "aggressive driving" to occur entirely independently of each other (*64*). The definition most frequently used within transportation and law enforcement identifies aggressive driving as any driver behavior that has the potential to endanger persons or property (*65*). This definition of aggressive driving does not require, or exclude, the presence of anger, an intentional action, or aggression in the literal sense. It also allows aggressive driving to be identified by a combination of observable actions, often in the form of recognizable traffic violations that can be readily associated with the situational variables of the physical driving environment.

Improvements in the collection of spatially referenced data and geographic information analysis have created opportunities to explore possible correlations between aggressive driving, the physical environment, and transportation infrastructure. Transportation research has applied these technologies to explore and advance the understanding of how automobiles respond to the physical environment. This has led to the development of intelligent safety features and warning systems found in commercial production cars.

Even though the significance of environmental factors is not generally disputed, the discussion of how the human operator responds and the specific role that physical traits of the environment play in aggressive driving has received less attention. This gap in the literature was noted by Burns and Katovich (*66*) in a study which suggested that aggressive driving could be better addressed through traffic facilitation based on the principles of Crime Prevention Through Environmental Design. More specifically, Burns and Katovich (*66*) proposed that an effort to understand the interaction between the transportation environment and aggressive driving would allow planning and design decisions to become more effective at preventing aggressive driving than traditional enforcement.

Wang et al. (*67*) quantified driver behavior using onboard sensors recording at high-frequency intervals to take advantage of the potential of “Big Data.” Their study directly focused on the development of methods to identify the instantaneous volatility of decisions made while driving, which will prove useful to the current project’s task of identifying occurrences of aggressive driving behavior. Wang et al. (*67*) also noted a limitation that this project will work to satisfy: their research was unable to include geographically specific data in the analysis due to privacy concerns. They also indicated that location-specific data at a 1-second temporal resolution will be critical to identifying factors of aggressive driving and provided examples of environmental factors, including road type, road geometry, surrounding land use, traffic counts, traffic facilities, and the network attributes related to the road segments traveled. The Safety Pilot Model Deployment (SPMD) data containing the vehicle and driver observations for the behavioral portion of this study also include a geographic location for each observation. The geographic locations attached to the behavioral data will provide the ability to identify any correlations between aggressive driving and location-specific traits such as those suggested by Wang et al. (*67*), as well as any other relevant traits discovered in the review of other current literature.

Physical Environment Variables

While not as widely studied as psychological factors, the role of environmental elements in cases of aggressive driving existed prior the prevalence of the term “aggressive driving” in the U.S. The following subsections cover the research literature relating environmental factors to aggressive driving.

To allow for appropriate quantification and identification of environmental factors independent of the SPMD observations, this project will focus on behaviors identified through vehicle maneuvers and traffic violations. These behaviors have been used as both dependent and independent variables in several studies and have been associated with both instrumental and hostile aggressive driving. The most common behaviors listed in aggressive driving reports are violations of traffic laws including exceeding the speed limit, disregard for traffic control devices, tailgating, failure to yield, frequent lane changes, and weaving in traffic (Stuster 2004; Balogun, Shenge, and Oladipo 2012; and others).

Speed and Other Network Attributes

Studies have found that several of the violations associated with aggressive driving are correlated with network attributes. Casey and Lund (*69*) use the term "adaptation effect" to describe the tendency of operators who travel on roads with higher speed limits to also travel at faster speeds on roads with lower speed limits. This suggestion when applied to a local scale invites investigation into the proximity of one road segment to another with a higher posted rate of speed (*70*). Legal speeds may be present as an attribute of road edge features found in some GIS data. If present, this information can be used to accurately identify posted speed differences between roads or variations in the speed limit between segments of individual roads. In cases where these data are not available, a universal approach may be to use broader categories, such as road class, to identify where vehicles enter a highway or city street from an interstate.

Speed-related aggressive driving does not always involve a higher rate of absolute speed. Rapid acceleration may be combined with tailgating, weaving in traffic, and failure to yield when an individual driver perceives driving as a competitive sport and races other drivers from one traffic control device to the next (*71*). Prime locations for this type of behavior are roads with multiple lanes and traffic control devices, especially when they are spaced at frequent intervals. Datasets identifying these roadway features should be investigated to pinpoint locations where aggressive driving behavior may occur.

Similar to speeding, following too closely, or "tailgating," may not be specifically targeted at the driver of the forward vehicle, but may simply be a result of difference in the desired speed of two drivers when there is not an immediate opportunity to pass (*72*). Tailgating has been identified as being most prevalent in congested areas (*73*) and where vehicles frequently reduce speed to turn off onto an alternate road in locations without deceleration lanes (Rajalin, Hassel, and Summala 1997). Congestion for the study site in this project could potentially be modeled using a combination of the traffic counts, the number of intersecting road segments, and the presence of land-use categories related to the identification of pedestrian generators.

Pedestrian Generators and Crosswalks

"The Social Psychology of Driving Behaviour and the Traffic Environment" (*74*) measured the aggressive reactions of drivers at an uncontrolled pedestrian crossing. The focus of this study was not on whether interactions with pedestrians were a common trigger of aggressive driving but rather on quantifying the intensity of the aggressive response. Cinnamon et al. (*75*) suggested that land use can be used to identify pedestrian generators that may lead to the disruption of traffic. This reduces the dependency on obtaining known crosswalk locations and allows the inclusion of pedestrian impacts that may occur outside a legal crossing. The disruption potential of a location can be identified through a single point such as a bus stop or polygon identifying the area of a strip shopping mall, public parking, or other sources of pedestrian foot traffic near road ways (*75*, *76*). Land-use zoning, parcel, or building information may be used to indicate areas that lead to an increase in the level of hazard already present in the roadway, such as strip-mall retail shopping locations (*76*).

Land Use

A 2003 article by Cackowski and Nasar 2003 studied the effects of roadside vegetation on anger and frustration tolerance while driving. The study was constructed as a simulated driving experience using a prerecorded video of three different roads: one with a largely built-up environment, one with a mixed environment, and one that was a scenic parkway. Participants were given anagrams that had no possible valid solutions and were instructed that if they could not solve the anagram, they could request an alternate anagram. The anagram test was conducted both before and after each simulated drive (*77*). The amount of time that respondents spent before requesting an alternate anagram was recorded and used to estimate their frustration tolerance after the drive relative to the amount of time spent on each puzzle before. Researchers found that all participants were willing to spend more time prior to giving up after "driving" along the scenic parkway, suggesting that the view of increased vegetation increased the ability of the respondents to focus and manage frustration (*77*). While the non-tasked simulated environment and the small number of sample groups limit the Cackowski and Nasar study, the research findings suggest that varying degrees of vegetation and built-environment obstructions (e.g. buildings) along the edges of the road may impact driver behavior by reducing stress, increasing the subject's ability to manage frustration, and potentially diffusing anger. The correlation between roadside vegetation and aggressive driving can be useful in the current study and can be analyzed using publicly available datasets.

Visibility

Visibility is frequently a consideration in traffic accidents but is mentioned less often in research specific to aggressive driving. Visibility, or the area that can be seen by the vehicle operator, can represent two factors related to aggressive driving: visual communication with other drivers and the visible length of the roadway. The inability to see the driver of another vehicle increases the possibility that a driver will consider other vehicles as objects rather than people when assessing risk (*78*, *79*). This results in a perception of anonymity likely to increase the level of aggressive reactions in several situations (*80–82*).

Roadway Geometry

The length of the driver's field of view determines the amount of time a driver has to evaluate roadway information, make a decision, and execute the chosen action. At intersections that do not have active traffic signals and points where traffic from a minor road may enter a major road, drivers must use the information within the visible distance of the roadway to assess the speed and distance of oncoming traffic and make a determination to accept or reject the available gap in the flow of traffic (*83*, *84*). Acceptable gap size is defined dynamically (*85*). Fang and Castaneda (*84*) reported that the acceptable gap size at four roundabouts ranged from 3.1 to 4.7

seconds, while gap acceptance for intersections with a stop bar fell below 50% probability at similar lengths to the 5.4 seconds defined by the Raff method (*84*, *86*). The difference in the decisions individual drivers make at intersections has been measured by Rakha et al. (*87*), Kaysi and Abbany (*83*), Shinar et al. (*88*), and others. These differences in decision-making may result in frustration, aggressive behavior, and high-risk driving choices (*83*). Rakha et al. (*87*) found that when a vehicle was 51 meters from a changing traffic signal, the decision to stop or proceed was the least predictable; this presents the potential for conflict between two subsequent vehicles. Many road users may choose to avoid or disregard traffic control devices as a result of perceived time constraints or the traffic control device being too slow or inefficient (*89*).

Specific locations of interest that may trigger aggressive driving include four-way stops, locations where U-turns are permitted, sequential traffic lights, merge lanes, and traffic circles. Increased attention should be given to locations where traffic from a minor road is combined with dense traffic on a major road (*83*). In addition to these variables, it may also be possible to include additional environmental characteristics from available geospatial data incorporating psychological factors related to aggressive driving such as frustration (*90*) and the social environment (*91*).

Topographical Obstructions

The term "topography" encompasses both the natural (terrain or surface of the earth, vegetation) and unnatural (built environment) features which exist in an area. Specific locations may possess topographical features that permanently limit visibility (*92*). These features may obstruct the view of oncoming traffic, the roadway, or roadway signage. One study identified failure to yield as the cause for 37% of the accidents investigated (*93*). This study also found that reduced visibility due to buildings, fixtures, terrain, and vegetation were a factor in 33% of the accidents (*93*).

Vehicle-mounted sensors can provide a real-time scene of surrounding objects; however, these sensors are limited to specialized vehicles that have been outfitted specifically for data collection. In support of this study's goal to create a set of methods that are transferable and reproducible at any location, this section will focus on environmental analysis using data that are readily available for the most-populated areas. Remote identification of potential topographical obstructions is possible with geographical datasets, including remotely sensed aerial imagery, georeferenced representation such as building footprints, and data collected by Light Detection and Ranging (LiDAR) sensors.

Current GIS capabilities provide several tools for measuring visibility and have evolved from the concepts of environmental modeling predating current computational abilities. The term "isovist" appears in Tandy (*94*) to indicate the visibility around the solid walls and objects common in urban environments. This term seems to be less frequently referenced in current literature, and it is possible that it has been encompassed by the current understanding of a "viewshed". A viewshed is a model of the visible field from a given point based on the topography of an area (*95*). The basic viewshed model identifies the area visible above a surface from a specific location based on the relative elevation of the surface. Often referred to as 2.5D, the elevation is established by the values contained in a raster such as a Digital Elevation Model (DEM), Digital Surface Model (DSM), or Digital Terrain Model (DTM). The 2.5D model of visibility is improved by processing the input surface so that the values include a representation of solid objects in addition to the earth's terrain alone. Building footprint polygons have been used to represent an obstructed view (*96*) and may be incorporated into the GIS model by relating elevation (z) values of the earth's surface to the building footprints, thus giving "height" to a 2D building footprint polygon. GIS data containing building footprints are commonly available, and obstruction caused by buildings may play an important role in the analysis of visibility at intersections. However, buildings represent only one portion of the information needed to determine the overall visibility to a driver. In cases where buildings or additional obstructions are not included in the GIS, remotely sensed imagery and LiDAR data are especially useful.

LiDAR is described by the National Oceanic and Atmospheric Administration (NOAA) as a form of remote sensing that is used to collect 3D information about built and natural environments (*97*). Airborne LiDAR requires an aircraft to carry an active sensor utilizing wavelengths within the near infrared portion of the spectrum. LiDAR is consistently available for U.S. coastal areas, with growing coverage for many other regions across the U.S. The 3D information provided by LiDAR is processed into a series of reflected returns containing information indicating both the elevation of the reflecting object and the intensity of the reflected light. The intensity of the reflected light will vary by the type of surface encountered. The elevation of the reflecting object is typically used to create a DEM or a DSM, either of which can be modeled as a 2D raster or with a 3D Triangulated Irregular Network (TIN), or Terrain. The DEM represents the surface of the earth, while the DSM represents the tops of objects such as trees or built structures that protrude above the surface of the earth (*98*). The intensity values from LiDAR data, in addition to being able to differentiate between concrete or asphalt and vegetation, increase the ability to differentiate between types of vegetation or types of building materials (*99*).

These methods have been used to identify whether there is a clear line of sight from the eye of a driver to a point along a roadway (*92*, *100–103*). The traditional 2.5D viewshed model is limited by the assumption that all objects represented by the surface are continuous features that intersect the surface of the ground. This vertically continuous assumption does not allow for instances where a driver may still be able to see under or through an object such as trees or other overhanging features. It has been suggested that determining the visual permeability of these features will reduce the impacts of this limitation (*104*).

Several previous studies have presented methods and GIS tools that support the modeling of obstructions to driver vision (*100–102*, *105*). These methods determine the visible length of the roadway between the current vehicle location and a series of points along the projected path.

This type of analysis requires the surface inputs of a viewshed analysis as well as target points located along the line that represents the road path. The observer point is established at the current location of the vehicle with a z value representing the eye level of the driver. In previous studies of driver vision, the height of the observer points have been set at the American Association of State Highway and Transportation Officials suggested height of 3.5 feet (*101*) or a height of 3.6 to 3.75 feet based on the vehicle used (*106*). Target points have previously been given a z value of 0.2 meters based on highway safety standards (*105*). For the current work focused on a driver's ability to see another vehicle, it may be reasonable to make the target point height closer to vehicle bumper height or the driver eye level of the target location.

While most potential obstructions can be included in a DSM, the desire to more accurately model trees and other features that may overhang the line of sight without being vertically continuous or intersecting with the ground must be represented using other methods. There are several methods for developing a digital canopy height model of trees based on subtracting the DTM from the DSM; however, this does not identify the bottom of the canopy, or the height of the visible area under the lowest branch. One possible solution is the use of existing multipatch objects (*105*). ArcGIS Pro allows multipatch objects to be created using z values from an attribute field of a polygon such as a building footprint, but this will not account for the unique shapes of individual trees unless accurate field data are present. Where there has been a need to represent unique objects with varying clearances, airborne LiDAR data have been used to create a wrapped surface. The wrapped surface method proposed by Kato et al. (*107*) begins with a digital canopy height model (DCHM) that allows individual trees to be located based on density or height-dependent segmentation. The LiDAR points are then grouped by tree based on the segmentation, extracted, and points representing the surface are selected. An isosurface method can then be produced where the exterior of the tree crown is represented by a value of zero. Against field measurements, the height of the lowest branch was found to have a root mean square error of 1.54 meters for coniferous trees (R^2=0.72) and 1.73 meters for deciduous trees (R^2=0.51). In the absence of accurate field data such as terrestrial LiDAR, the wrapped surface method offers improvement to a visibility model that must consider vision under the lowest branch.

With a focus on the impacts of aesthetics as opposed to visual obstruction, vegetation along the roadway was quantified relative to other locations along the road in Cackowski and Nasar (*77*). The researchers used a 0.5-inch grid to measure the area of vegetation on the monitor that displayed the video of the driving route. Cackowski and Nasar (*77*) determined that measuring the vegetated area at 30-second intervals was appropriate for comparing the different routes. While video from a driver perspective would be ideal for assessment, our study of a larger area could utilize the National Land Cover Database (NLCD) to identify areas that are generally built up or offer a vegetated scene. NLCD data have been used in combination with viewshed analysis to establish the vegetated viewshed visible from an observer point for the purposes of aesthetic analysis reports (*108*). NLCD data available at a 30-meter resolution provide the ability to assess the aesthetics visible from a 35-mph road at a 2-second temporal resolution. In addition to the NLCD data, other data containing land-use information may help our model properly specify where the number of occurrences of aggressive driving is lower than what is predicted based on other environmental variables.

Summary

The research literature demonstrates the significance of understanding the role that environmental variables may play in aggressive driving. While detailed measurements of the correlation have not been presented previously, the body of work has drawn attention to the need to evaluate how natural and built features impact driver behavior.

The portion of the roadway that is visible to a driver at a given point in time has been shown to play a significant role in decisions impacting vehicle control and the nature of interactions with other drivers. The study of the visible road using GIS employed two separate methods. The general impacts of topography were automated using GIS, while the specific visual field around unique objects was manually built using GIS and then manually assessed using either video or 3D rendering technology. The opportunity exists to extend automated 2D GIS analysis by incorporating 3D details, while also covering a complete network of roads. Existing work has introduced the combined use of remote-sensing data and GIS for managing and measuring transportation-related variables, highlighting that the evaluation process can be automated by developing software specific to the needs of the project. Well-suited for a study intended to explore Big Data, this automated approach will allow the interactions between several factors over a large area to be assessed.