Review

Bridging the Gap between Optimization and Statistical Modeling of Large Commercial Vehicles Safety: A Review — Part 1: Data Collection and Exploration

Firstname Lastname [1,†,‡], Firstname Lastname [1,‡] and Firstname Lastname [2,*]

[1] Affiliation 1; e-mail@e-mail.com
[2] Affiliation 2; e-mail@e-mail.com
* Correspondence: e-mail@e-mail.com; Tel.: (optional; include country code; if there are multiple corresponding authors, add author initials) +xx-xxxx-xxx-xxxx (F.L.)
† Current address: Affiliation 3
‡ These authors contributed equally to this work.

Version November 27, 2019 submitted to Sensors

Abstract: A single paragraph of about 200 words maximum. For research articles, abstracts should give a pertinent overview of the work. We strongly encourage authors to use the following style of structured abstracts, but without headings: (1) Background: Place the question addressed in a broad context and highlight the purpose of the study; (2) Methods: Describe briefly the main methods or treatments applied; (3) Results: Summarize the article's main findings; and (4) Conclusion: Indicate the main conclusions or interpretations. The abstract should be an objective representation of the article, it must not contain results which are not presented and substantiated in the main text and should not exaggerate the main conclusions.

Keywords: keyword 1; keyword 2; keyword 3 (list three to ten pertinent keywords specific to the article, yet reasonably common within the subject discipline.)

1. Introduction and motivation

Transportation crashes are a pressing global public health issue. The World Health Organization estimated that *road injuries* are the 8$^{\text{th}}$ leading cause of death worldwide, resulting in 1.4 million deaths annually[1]. Perhaps more importantly, the incidence of such crashes and their severity are on the rise. By 2030, traffic-related deaths are predicted to become the 7$^{\text{th}}$ leading cause of death worldwide [1]. The increase in annual deaths is seen in low- and high-income countries alike. For example, in the U.S., an estimated 37,133 people died in motor vehicle crashes in 2017 [2], which constituted a 7.5% increase from the average annual deaths recorded in 2012-2016 [3]. In addition to the massive loss of life, motor vehicle crashes cause significant economic losses. According to the [1], "road traffic crashes cost most countries 3% of their gross domestic product." In the U.S., it is estimated that the total value of societal harm from vehicle crashes exceeds $830 billion annually [4], which is equivalent to $\approx 4.4\%$ of the country's gross domestic product [5].

Large commercial vehicles (e.g., large trucks) are often involved in the most severe crashes. In the U.S., "large trucks and buses account for 12% of the traffic fatalities" [6], while accounting for only 9% of the total miles driven in the U.S. [7]. If one considers the working environment for truck drivers, in particular, there are four main reasons for the larger involvement of trucks in fatal (and non-fatal) crashes. First, truck drivers can encounter different routes/paths, weather, traffic conditions, and locations each time they take a trip. Second, truck drivers are on the road for long hours with little supervision or contact with fellow employees [8]. Third, the driving times of truck drivers can vary significantly over time since they are affected by the scheduling requirements of the motor carrier, shipper, and receiver [9]. Fourth, drivers sleep quality and duration is often negatively affected by the working environment [9]. For example, over-the-road (OTR) drivers can be away from their homes for several consecutive weeks. These characteristics of truck drivers' environment can increase their cognitive demands and/or fatigue rates when compared to other drivers [10].

Add introduction to statistical modeling and optimization research streams.

The remainder of this paper is organized as follows. To do done after the draft is ready.

2. Literature Review: A Bibliometric Analysis

There is a large body of literature dedicated to improving transportation/trucking safety. Based on our initial literature review, we have identified 856 documents (i.e., published articles, proceeding papers, and book chapters). To categorize these documents, a bibliometric analysis of the documents was performed using the "bibliometrix" **R** package [11], with the goals of: (a) examining the co-occurrences of keywords within documents since this shows a link between the topics captured by these keywords; and (b) constructing a conceptual structure map of the literature based on a more streamlined keywords list ("Keyword Plus", refer to [12] for a detailed introduction on how they are constructed). The results from these two analyses are shown in Figures 1a and 1b, respectively.

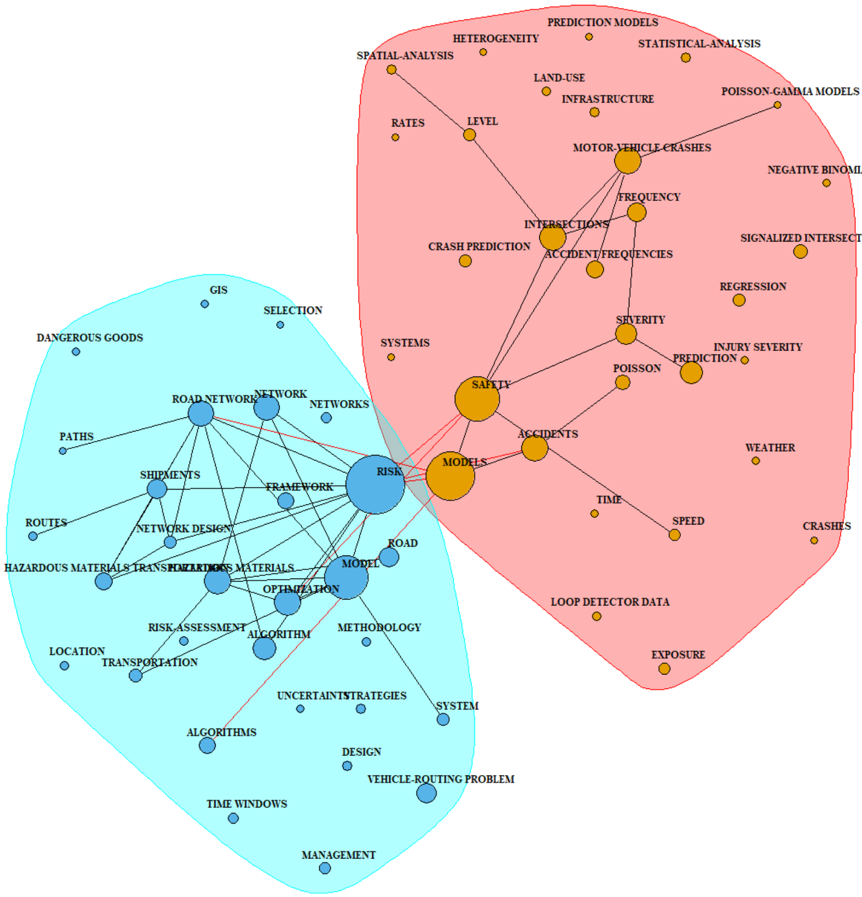
Based on Figure 1, two important conclusions can be drawn. First, the literature can be grouped into two main groups: (a) an explanatory/predictive modeling stream, where the keywords emphasize the collected data (loop detector data), predictors (traffic, weather, time and/or infrastructure), models used (regression, spatial-analysis, Poisson-gamma and negative binomial), and model outcomes (rates, crash frequencies, and crash prediction); and (b) a prescriptive modeling stream, where the focus is on developing algorithms to manage risk, particularly for hazardous materials transportation, through the selection of paths and routes. Second, the cluster agreement between the *keyword co-occurrence network* and the *concept map generated using the Keywords Plus* implies that there is a clear division between both research streams. This is somewhat surprising since the *outputs* from the first stream should be *inputs* for the optimization models used for prescriptive decision-making. Based on the second insight, a thorough examination of the relevant operations research (OR) literature was performed. From the analysis, we learned that the OR literature largely ignores the recent results on factors influencing crash risk. Particularly, most hazardous materials (hazmat) optimization models assume that the crash probability is time-invariant [13,14], and is in the range of 10^{-8} to 10^{-6} per mile [15]. This contradicts the findings from the first stream (e.g., see the reviews of [16] and [17]).

Against this backdrop, the primary purpose of our paper is to help bridge the gap between the different research streams that relate to the modeling and minimization of crash risk through a detailed review and taxonomy of the literature. Our goal is to bring the research into better focus and to encourage future work that crosses the siloed divisions within the literature. To construct our taxonomy, we will frame crash risk modeling applications using a data analytics framework. Thus, one can categorize the literature into the following applications: (a) *descriptive*, where the goal is to understand crash-related factors through visualizations and other exploratory data analysis approaches; (b) *predictive*, where the goal is to construct models that can predict crash outcomes/probabilities based on time-dependent factors such as traffic flows, weather, and road surface conditions and/or covariates pertaining to the road geometrical descriptors (number of lanes, distance between exits, etc); and (c) *prescriptive*, where outputs from the predictive models are used as input parameters for a decision-making model. Note that the two clusters in Figure 1, do not capture the keywords associated with the descriptive applications.

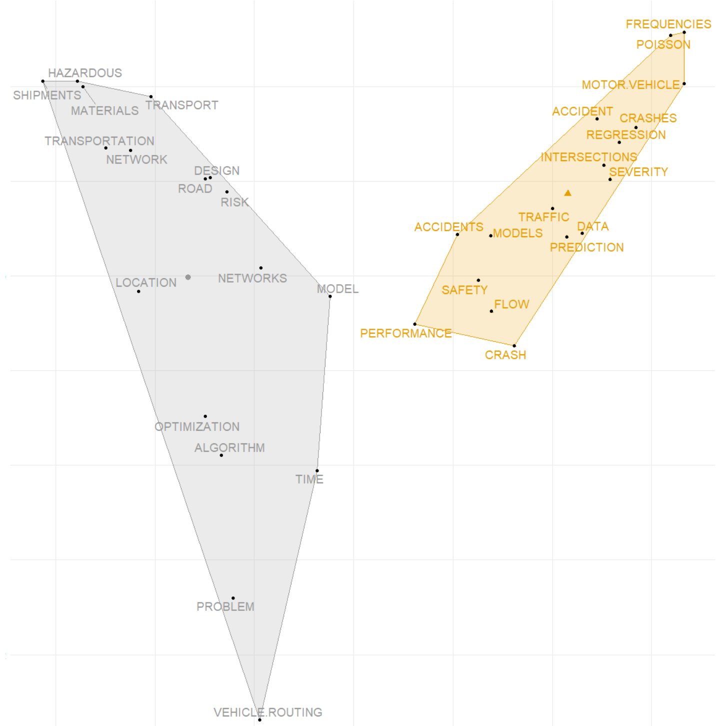
3. Data Collection

3.1 Outcome variables

In this section, we highlight some of the different online sources that can be utilized to scrape/collect the data needed for analyzing most crash risk modeling research projects. We hypothesize that one potential reason for the gap between the predictive and prescriptive analytic communities is the "large start-up burden" associated with the lack of sufficient/targeted documentation for how data on different predictor variables and covariates can



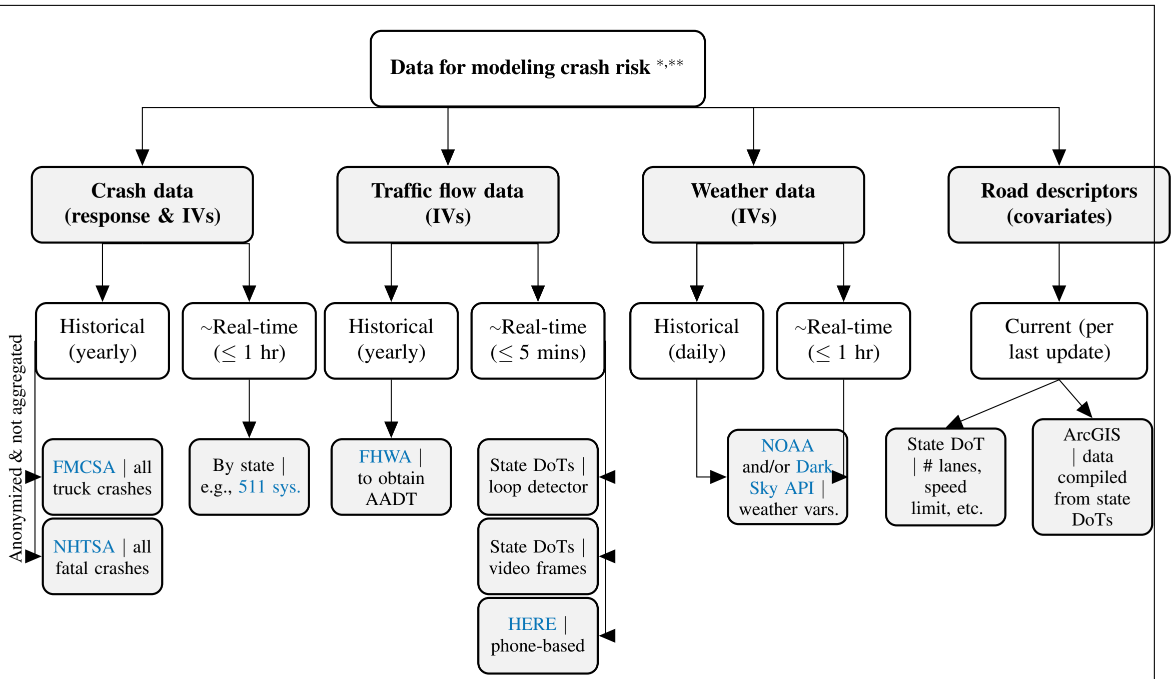
(a) A keyword co-occurrence network of the literature, depicting the 60 most used keywords. The nodes correspond to the keywords, with node size reflecting relative frequency. The links are limited to keywords that co-occurred at least five times (black and red lines correspond to between and within clusters, respectively). The network plot divides the literature into two clusters: prescriptive modeling (left), and explanatory/predictive modeling (right).



(b) A data-driven conceptual structure map based on “Keywords Plus” (keywords tagged by the ISI or SCOPUS database scientific experts) and the application of multiple correspondance analysis and k-means clustering. The nodes are limited to keywords that have occurred ≥ 5 times, and the gray circle and orange triangle depict the corresponding cluster center. Similar to Fig 1a, the concept map also divides the literature into the same two clusters.

Figure 1. A bibliographic analysis of the literature using the *bibliometrix* package in **R**.

be obtained. Thus, in this section, our goal is not to present a comprehensive review for all the potential venues for data collection, but to provide an introduction that reduces the burden on the OR researchers so that they are more likely to consider the outputs from state-of-the-art crash prediction models as inputs to their analyses. It is assumed that OR researchers posses sufficient knowledge and resources for storing the collected data in suitable databases. For this reason, we do not discuss how the data should be stored and focus instead on how the data should be collected. An overview of the four sets of variables used in the modeling of crash risk and their respective data providers is shown in Figure 2. These sets of variables include independent variables (IVs), response variables, and covariates.



* **Acronyms:** IVs = independent variables, FMCSA = Federal Motor Carrier Safety Administration, NHTSA = National Highway Traffic Safety Administration, AADT = Annual Average Daily Traffic, FHWA = Federal Highway Administration, DoT = Department of Transportation, and NOAA = National Oceanic & Atmospheric Administration.

** **Code:** To simplify the data collection process, we present the **R** code needed for scraping these different data sources at: https://caimiao0714.github.io/TrafficSafetyReviewRmarkdown/.

Figure 2. A hierarchical view of data pivotal to crash risk modeling. The first level describes the data type (in context of predictive models), the second level describes the frequency of update, and the third level denotes the source from where the data can be scraped/collected.

3.2 Predictor variables: traffic, weather, and road geometry

The first set of variables includes crash-related data, where the location, time, weather, road surface conditions, vehicles/pedestrians involved in crash, crash outcomes and/or root-causes. Information extracted from crash reports can form the dependent and/or the independent variables in the predictive modeling stage. Depending on the nature of the application, one may want to collect historical data (e.g., model training) and/or near real-time data (e.g., for routing and scheduling purposes). Historical data are released by different DoT divisions depending on the types of vehicles involved and whether the crash resulted in a fatality. On the other hand, the near real-time data can primarily be obtained from the different reporting systems used by each state. The 511 reporting system, highlighted in Figure 2, is the most common since it is used by more than 45 states [18].

The second set focuses on information pertaining to traffic flows. The analysis of historical data often starts with obtaining the AADT for different road segments (e.g., [19]). The AADT data can be downloaded from the FHWA's website [20]. The downloaded "shapefiles" can be converted to different data formats using the provided R code. For short duration traffic volume estimation, monthly and weekly factors are often used for adjusting the AADT [21]. Prescriptive applications requiring near real-time data can capitalize on: (a) data provided by the different state DoTs, which include speed, volume, and occupancy data extracted from loop detectors (e.g., [22]), and (b) estimates of the data provided by the different state DoTs based on data extracted from users' smartphones or sampled floating cars [23]. Recent research shows that the estimates based on phone usage are accurate and reliable when compared to the official data collected by government agencies [23].

Set three contains weather variables that can either affect the likelihood of crashes and/or their severity. From this set, variables affecting crash likelihood include, but are not limited to, the following: (a) visibility, (b) rain and snow accumulation, and (c) the potential for icy conditions. From the perspective of hazmat risk minimization, wind speed and direction are of interest since they can increase the severity of the release of toxic materials as a consequence of a crash. The aforementioned variables in this list can be extracted using the NOAA application programming interface (API) [24] and/or the Dark Sky API [25].

The fourth set is comprised of the geometric road segment descriptors. Since the geometric descriptors are essentially roadway design parameters, we will refer to them as covariates throughout this paper. These include: (a) number of lanes, (b) speed limits, (c) longitudinal grade, (d) whether the road segment of interest contains a straight, merge and/or diverge sections, and (e) information pertaining to shoulder width and the presence of construction. This information can be obtained from DoT of each state. For example, Florida has published a handbook for roadway features and characteristics which is accessible to the public [26].

To facilitate and encourage the examination of these important factors (per the reviews of [16] and [17]) in future *prescriptive* studies, we developed **R** code that can be used in scraping data from several different data sources. The code is freely available online through the link provided in our *Supplementary Materials* Section. Prior to examining the **R** files, we recommend the reader to refer to our R Markdown file (see supplementary materials), where we provide example queries and their resulting outputs.

4. Descriptive analytic tools used for understanding crash data

Once the data are collected and stored, descriptive analytic (DA), also known as exploratory data analysis (EDA), techniques are used to examine the data. These methodologies become important pre-processing steps when dealing with large datasets, where any analysis can be computationally intensive. Therefore, in the context of modeling crash risk which typically deals with huge amount of data (e.g., traffic and weather data), EDA plays an important role in any project. In our estimation, there are two main EDA goals: (a) provide statistical summaries of the variables and visualize the data; and (b) reduce the data's dimensionality (in terms of both the number of observations and variables to be examined). Figure 3 presents a hierarchy of the two main EDA goals and their corresponding analytic methods. Note that these techniques may not be mutually exclusive and could be done iteratively. For example, data visualization techniques may be used to gain insight into how the data dimension could be reduced, which would then lead to better data visualization. We discuss each of these goals/techniques in the following subsections.

4.1 Data summarization and visualization

As a first step to understanding the data, several *descriptive statistical* techniques are often used to summarize the data. These techniques include both univariate (e.g., describing the distribution of traffic flows, estimating its central tendency parameters and its dispersion) and multivariate methods (e.g. computing the correlation between road surface conditions and precipitation). Since these approaches are somewhat standard and are deployed in almost every paper, we do not discuss them further. The reader is referred to [27] for a detailed introduction on the application of descriptive statistical tools for transportation data analysis.

To complement the descriptive statistical techniques, data visualization offers a succinct and simple approach to understanding trends, patterns and/or anomalies within the dataset. Most of the published transportation-related

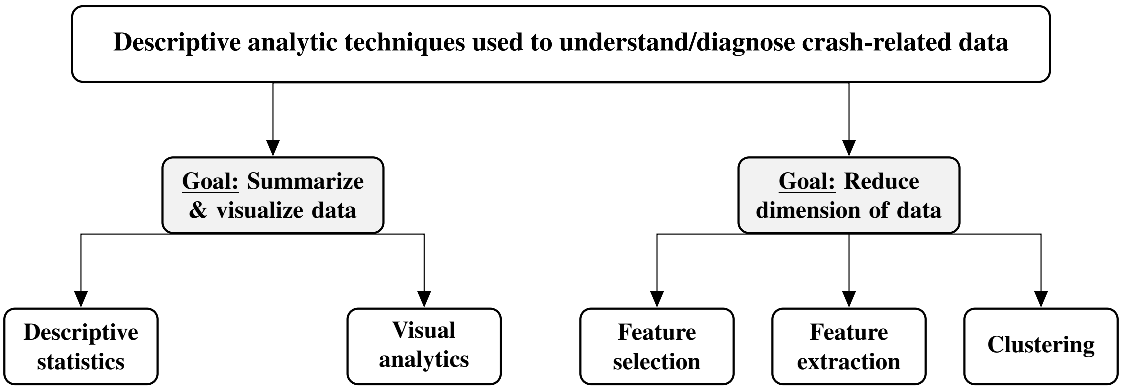


Figure 3. A hierarchical view of the EDA methods used for understanding crash-related data. The hierarchy consists of two levels. The first level summarizes the two EDA goals, and the second captures the methodological framework used to achieve each goal.

data visualization papers have focused on visualizing traffic data. [28] presented an excellent survey of those papers, and categorized the data visualization approaches into four groups. Those are methods for visualizing: (a) temporal data; (b) spatial data; (c) spatiotemporal data; and (d) multivariate data. In our estimation, this framework is suitable for most (if not all) transportation datasets. For example, it can be used to visualize weather-related factors/features. Table 1 presents a summary of relevant applications of data visualization to transportation datasets. Those papers are discussed in further detail in the subsections below.

Table 1. A framework for categorizing visualization techniques for transportation data. The framework is adopted from [28].

Variable type (Main group)	Subgroup	Visualization techniques	Examples
Time-series data	*Linear time*	Line and stacked graphs	[29], [30] and [31]
	Periodic time	Radial layout and cluster-and-calendar based visualization	[32] and [31]
	Branching time	Storylines	[33]
Spatial	*Point-based*	Symbol maps	[34]
	Line-based	Line maps, edge bundling, and kernel density estimation charts (KDE)	[35] and [36]
	Region-based	Radial metaphor charts, choropleth, proportional symbol maps, and heat maps	[37] and [38]
Spatiotemporal	-	Space-Time-Cube (STC), animated maps, GeoTime, and stacking-based STC	[39], [40] and [41]
Multiple properties	-	Parallel coordinates plot, trellis plot, and multidimensional scaling	[42], [43], [44] and [45]

4.1.1 Visualization of time-oriented data

In his seminal paper, [46] has stated that time can be abstracted/conceptualized through a number of different models (i.e. measurement methods). Based on this idea, [47] used the dimensions of orthogonal design to categorize the different “types of times". From a visualization perspective, Aigner *et al.* [48, p. 48] identified three criteria that are most important to constructing appropriate visualizations:

- *Linear time versus cyclic time.* Linear time assumes a starting point and defines a linear time domain with data elements from past to future. On the other hand, many natural processes are cyclic, for example, the cycle of the seasons ...

- *Time points versus time intervals.* Discrete time points describe time as abstractions comparable to discrete Euclidean points in space. Time points have no duration. In contrast to that, interval time uses an interval-scaled time domain like days, months, or years. In this case, data elements are defined for a duration, delimited by two-time points. ...
- *Ordered time versus branching time versus time with multiple perspectives.* Ordered time domains consider things that happen one after the other. For branching time, multiple strands of time branch out, which facilitates description and comparison of alternative scenarios (for example, for project planning). This type of time supports decision-making processes, where only one alternative will actually happen. Time with multiple perspectives allows more than one point of view at observed facts (for example, eye-witness reports).

In addition, [48] noted that there is no single visualization technique that can consider all the aspects of time. Thus, the visualizations are specialized and depend on the aforementioned criteria. Using their insights/recommendations, we have sub-grouped the different visualization techniques used for time-series type data in the transportation literature into three subcategories. These are highlighted in the second column in Table 1, with the corresponding examples on the right.

Time-series visualizations play an important role in transportation analytics. Line graphs are the most commonly used chart for that purpose, where the x-axis is used to capture time and a transportation-related variable is depicted on the y-axis. In our view, most (if not all) traffic modeling studies utilize line charts as an integral component of their data exploration/analyses. For an example application, we refer the reader to [30] who used a line chart to visualize: (a) the number of trips performed by taxi drivers in New York City for 2011 and 2012, and (b) the dollar amount of tips per trip and fare per miles-driven for trips originating in different neighborhoods of the city. Other examples include: (i) visualizing carbon monoxide pollution over the course of the day in London [49], (ii) visualizing traffic volumes in cities such as: Beijing, China [50] and Porto, Portugal [51], and (iii) effect of road surface conditions and time of day on traffic volumes [52]. [28] duly noted that the ease-of-use of line charts deteriorates as the number of depicted variables increases. In this case, other time-series based charts are more suitable. [28] suggested using the *Theme River stacked chart*. The chart developed by [53], and uses a flowing river metaphor to capture changes in several variables of interest over time. The reader is referred to [29] for a transportation application of the chart. Contrary to [28], we believe that the interpretation of this chart is somewhat difficult/confusing since it does not have a traditional y-axis. Specifically, variables depicted at the bottom of the *theme river* visual are not negative since the height is measured in absolute terms from the closest wave that is in the same direction. In our estimation, this makes it difficult to discern patterns. Instead, we recommend the use of panels of line charts to capture changes in multiple time-series.

If one would like to depict the periodic/cyclic nature of the data, there are three main visualization approaches. First, one can utilize the *radial layout* chart to visualize data exhibiting a cyclic behavior [28]. [32] has used this technique to show traffic information in different days and times. In their approach, each ring was used to represent a day, time was shown on the circular axis, and the color was to used to capture low to high traffic volume. Second, the *cluster and calendar based visualization* approach of [54] can be used to depict seasonality in daily patterns. In this approach, the days are clustered based on the hourly data. Then, the average patterns for the clusters are visualized through a line graph with multiple time-series (each corresponding to a cluster and color-coded accordingly). The line chart is supplemented with a calendar heat map, where each day is colored according to the cluster it belongs to. By implementing this methodology, [31] divided traffic patterns into eight clusters. These clusters not only identified workday and weekend effects, but it was also able to identify game-day traffic for college football (which are considered major sporting events in the U.S.) and unusual travel on or near major holidays. Third, statistically derived plots (based on time-series analysis techniques) can be used to quantify the periodic/seasonal nature of the data. From a time-series analysis perspective, the data can be decomposed into: (a) seasonal, (b) trend, and/or (c) cyclical components within a season. These components can be visualized, along with the autocorrelation function (ACF) and the partial autocorrelation function (PACF) for the differenced series to provide an understanding of what type of time-series models to use. The reader is referred to [27] for detailed coverage of time-series modeling applied to transportation data analyses.

In some cases, researchers and practitioners may want to visualize the data based on order and sequence of events instead of using time as the main exploratory variable. We agree with [28] that story-line visualizations (e.g., see [33]) can be effective in showing the sequential nature of events and their effect on a certain response. The reader should note that this approach has not been used in the context of the literature so far. However, we have highlighted it here since it can be: (a) effective in depicting the effectiveness of accident response teams in transporting injured commuters to hospitals and/or clearing the roads; and (b) a useful visualization tool in vehicle routing applications.

4.1.2 Visualization of spatial and spatiotemporal data

Location of vehicles, origin, destination, construction sites, road closures, and/or crashes provide a spatial dimension to transportation datasets. [28] classified the visualization of spatial data (with a fixed time period) into three groups based on the aggregation level of the location-based information. These groups are: (a) point-based visualizations, where no aggregation is performed; (b) line-based visualization, where a first order aggregation occurs; and (c) region-based visualizations, where a second-ordered aggregation is performed to provide a macro perspective of location. In point-based visualizations, each symbol on a map represents the position of an object at a given point in time. A popular implementation of a symbol map in crash modeling is within the [34] dashboard for visualizing traffic fatalities. Their dashboard contains a symbol map, and six filters that allow for removing unwanted data from the visualization. The filters can be applied to: person type (occupant vs. non-occupant), month, day, hour, state name, and roadway type. In Figure 4, we capture an example output from the dashboard with data limited to occupant fatalities occurring on Saturdays in December 2016. With a first-order aggregation, line maps are widely-used in visualizing travel routes, and traffic flow/volume. This type of map has been popularized by the ubiquity of modern navigation applications, where route overviews and a color scheme indicating the corresponding traffic speeds in different segments are often provided. Due to their widespread use, we will not discuss them further. For region-based transportation analytics, several visualization techniques have been deployed. These include: (a) proportional symbols map [55], which the size of a point/symbol in a map is proportional to the number of observations in that location; (b) choropleth maps [see e.g., 37,56,57], where areas/regions in maps are shaded, colored, or patterned relative to the value of the metric of interest; and (c) radial metaphors, which were used by [38] to visualize interchanging traffic patterns among different regions of a city.

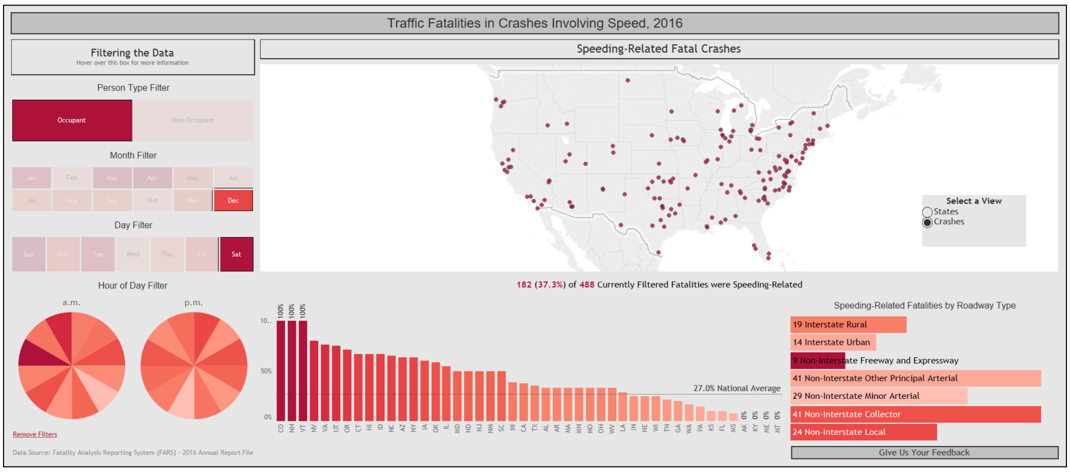


Figure 4. The location of vehicle occupants killed in speed-related crashes occurring on Saturdays in December, 2016. The data was filtered and visualized using the interactive dashboard developed and hosted by [34].

When one wants to depict changes in both space and time, there are two main approaches. The first approach is to add a time effect to the aforementioned spatial charts. This can be achieved through animation effects, which would allow us to see changes from one period to the next. Examples of this approach are often depicted in popular blogs such as: [58] and [59]; however, this approach is somewhat limited since it is difficult to discern changes over long periods of time. The second approach is through the use of dedicated visualization methodologies. An example of a popular dedicated technique is the space-time-cube (STC) visualization method of [39]. In this method, the x and y axis represents the spatial information and the temporal information is shown on the z axis. In our estimation, this method can be implemented in: (a) planning public transportation where the space time-paths of individuals or public transportation vehicles are depicted by the standard STC, (b) traffic analysis where the changes in a traffic-related variable of multiple vehicles across time and space is shown by stacking-based STC [41], and (c) crash analysis where crashes/incidents are displayed and tracked based on their spatial-temporal information by an enhanced version of standard STC [40].

4.1.3 Visualization of multidimensional and high-dimensional datasets

For multidimensional and high-dimensional data, there are several visualizations that are possible. In our estimation, the choice of the visualization is often dependent on the amount of preprocessing/analysis that is performed prior to the visualization. On the lower end of the spectrum, *parallel coordinates plots* (PCP) and *trellis* (small multiples of bar charts or scatter plots) are commonly used visualization techniques that require limited preprocessing. [42], [57] and [60] used a PCP to visualize the correlation/interaction among several crash descriptors including: cars involved, day/month effects, incident type, and road condition. Additionally, the trellis plot was used by [43] to visualize variations in the number of crashes by different census tracts. On the upper end of the analytical spectrum, visualizations are preceded with the application of projection methods to reduce the problem's dimensionality. Examples include: (a) [45] where cluster analysis and multidimensional scaling were used to produce a 2-dimensional (2D) plot of the relationship between the different constructs and types of drivers examined the study; (b) [61] who utilized multiple correspondence analysis (MCA) to present a proximity map of key factors contributing to wrong-way driving in a 2D space; and (c) [62] where the multivariate time-series data capturing the driver behavior were reduced to a 3D feature space using deep learning techniques and then, visualized using a *driving color map*.

4.2 Dimension reduction

In the previous subsection, we highlighted how projection methods can be used to reduce the data dimensionality and assist in its visualization. Here, we discuss how dimension reduction techniques can be used to prepare the data for the predictive modeling stage. In general, there are three main goals for dimension reduction: (a) *feature selection*, where important variables are identified and selected; (b) *feature extraction/generation*, where the variable set is projected into lower subspace without losing significant information and/or predictive ability; and (c) *clustering*, where similar observations are grouped together. Note that researchers can combine these approaches in their analysis; hence, we classified dimension reduction methods according to their *goals*.

4.2.1 Feature selection

One of the recommended steps before the use of statistical and machine learning models is to identify and use only the variables/features deemed important for the analysis since this [63]: (a) avoids over-fitting, (b) reduces the computational complexity in the analysis, and (c) leads to better prediction performance. This step is often referred to as variable or feature selection. In the context of crash prediction models, variable selection play an important role since there are many potential predictors (e.g., traffic, weather, road geometry related variables) which may have effect on the probability of a crash. In addition, in order to capture the spatial and temporal effects of these variables, new variables need to be introduced in the model. For instance, [64] developed a crash prediction model where each traffic-related variable is collected prior to the crash from two upstream and two downstream sensors. This means that the information for each traffic variable is divided across four variables, and that these variables contain some redundant information within them. In such cases, feature/variable selection will improve model performance as shown in [65–69]. For the sake of conciseness, hereafter we use the term *feature selection* to denote feature and variable selection methods.

Feature selection methods can be classified into three groups: filter, wrapper and embedded methods [70]. In the filter methods, the process of selecting a subset of features is independent from the statistical and machine learning model used, i.e., a subset of features will be selected according to an algorithm (e.g., Pearson Correlation or Mutual information Criterion), and then the selected features will be inputs to the explanatory/predictive model. Advantages of filter methods include: (a) simplicity, (b) computational efficiency, (c) speed, and (d) reducing the risk of over-fitting. However, they can ignore the dependency between features and do not guarantee the selection of an optimal set of features [70,71]. In contrast to filter methods, wrapper methods considers the prediction performance of the classifier (while accounting for the dependencies/interactions between features) and subsets the feature space using heuristic searching algorithms such as: genetic algorithms [72] and particle swarm optimization [73]. While they can improve performance when compared to filter methodologies, they are computationally inefficient. In addition, they also do not guarantee optimality and over-fitting remains a possibility [70,71]. To avoid such problems, feature selection is a part of the model training process in embedded approaches, which makes them the preferred approach in many crash risk modeling scenarios [see e.g., the use of *random forests (RF)* for feature selection and determining variable importance in 22,68,69]. For more information about the feature selection methods and their applications, we refer the reader to [71,74,75].

4.2.2 Feature extraction

Feature extraction methods offer an alternative approach to dimension reduction through the projection of the input space to a more efficient dimension space. The projection/transformation allows for combining input variables, reducing the problem's complexity, and presenting a useful abstraction of the data [76]. Thus, feature extraction differs from feature selection as the focus is not on dropping unimportant variables, but rather to combine the information across the variables through a mathematical transformation. Principal Component Analysis (PCA) is the most commonly used feature extraction method in the crash prediction literature [e.g., see 77–82]. Through an orthogonal transformation, PCA transforms the original variables into a set of linearly uncorrelated variables (i.e., principal components, PCs). Typically, the variation in the data can be explained with a few PCs, which allows for reducing the dimensionality of the problem without the loss of information. The determination of the number of PCs to retain is often determined through a scree plot or through a threshold for the eigenvalues [83]. Since PCA was originally designed for numeric variables that can be linearly combined, there are several extensions to PCA which do not require such assumptions. These include: (a) probabilistic PCA [84], (b) non-linear PCA [76], and (c) kernel-based PCA [85]. These methods have also been implemented extensively in the literature [see 76, for a detailed review].

4.2.3 Clustering

Contrary to feature selection and extraction, clustering approaches attempt to group observations together with the goals of maximizing the similarity within a cluster (i.e., minimizing distance between observations) and minimizing the similarity between clusters (i.e., maximizing the distance between cluster centers/centroids)[86,87]. Since one does not have a label in advance for each observation, clustering is an unsupervised machine learning method. Generally speaking, clustering approaches can be divided into: partitioning-based, hierarchical-based, density based, grid-based and model-based methodologies [86,88].

Crash risk modeling datasets have a number of characteristics that make clustering a viable and useful approach for dimension reduction. For example, if you consider traffic datasets, the goal is typically to understand the impact of traffic conditions on crash likelihood, which is typically achieved through: (a) classifying traffic into different states, and then (b) evaluating the impact of each traffic state (e.g., congested or not congested) on the crash likelihood [16]. Historically, step (a) was achieved through an analysis of traffic flow characteristics [e.g., see 89–91]. A limitation of such an approach is that the modeling can be influenced by researchers' biases and perceptions. Alternatively, one can use an assumption-free, data-driven approach to identify how observations can be clustered. [31] showed how clustering can be used to identify logical, but hard to model, groupings of the data. Applications of clustering include, but are not limited, to: (a) traffic categorization [31,92,93], (b) identifying accident clusters [94–96], and (c) grouping of weather conditions [97]. To demonstrate how an optimal number of clusters (k^*) can be obtained, we provide a detailed example in the supplementary materials where we use $k-$means clustering and the elbow method to determine the k^* clusters for traffic data.

5. Conclusion

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing–original draft preparation, X.X.; writing–review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y.", please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

Funding: This work was supported in part by: the National Science Foundation (CMMI-1635927 and CMMI-1634992); the 817 Ohio Supercomputer Center (PMIU0138 and PMIU0162); the American Society of Safety Professionals (ASSP) 818 Foundation; the University of Cincinnati Education and Research Center Pilot Research Project Training Program; 819 the Transportation Informatics Tier I University Transportation Center (TransInfo); a Google Cloud Platform re820 search grant for data management; and a Dark Sky grant for extended API access (i.e., they increased the number of 821 possible queries per day).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
DOAJ	Directory of open access journals
TLA	Three letter acronym
LD	linear dichroism

References

1. World Health Organization. WHO | The Top 10 Causes of Death. http://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death, 2018. [Online; accessed 23-February-2019].
2. National Highway Traffic Safety Administration, NHTSA. U.S. DOT Announces 2017 Roadway Fatalities Down. https://www.nhtsa.gov/press-releases/us-dot-announces-2017-roadway-fatalities-down, 2018. [Online; accessed 23-February-2019].
3. Insurance Institute for Highway Safety. Fatality Facts - IIHS. The Insurance Institute for Highway Safety and the Highway Loss Data Institute, http://www.iihs.org/iihs/topics/t/general-statistics/fatalityfacts/overview-of-fatality-facts, 2018. [Online; 23-February-2019].
4. Blincoe, L.; Miller, T.R.; Zaloshnja, E.; Lawrence, B.A. The Economic and Societal Impact of Motor Cehicle Crashes, 2010 (Revised). U.S. Department of Transportation, National Highway Safety Administration, Report No.: DOT HS 812 013, https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812013, 2015. [Online; accessed 28-April-2018].
5. GDP (current US $) | Data: United States. World Bank national accounts data, and OECD National Accounts data files. https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=US, 2018. [Online; accessed 28-April-2018].
6. Commercial Motor Vehicle: Traffic Safety Facts. U.S. Department of Transportation, https://www.fmcsa.dot.gov/sites/fmcsa.dot.gov/files/docs/safety/data-and-statistics/84856/cmvtrafficsafetyfactsheet2016-2017.pdf, 2017. [Online; accessed 28-April-2018].
7. Table VM-1 - Highway Statistics 2016 - Policy | Federal Highway Administration. U.S. Department of Transportation, Office of Highway Policy Information, https://www.fhwa.dot.gov/policyinformation/statistics/2016/vm1.cfm, 2017. [Online; accessed 28-April-2018].
8. Zohar, D.; Huang, Y.h.; Lee, J.; Robertson, M.M. Testing extrinsic and intrinsic motivation as explanatory variables for the safety climate–safety performance relationship among long-haul truck drivers. *Transportation Research Part F: Traffic Psychology and Behaviour* **2015**, *30*, 84–96.
9. Crum, M.R.; Morrow, P.C. The influence of carrier scheduling practices on truck driver fatigue. *Transportation Journal* **2002**, pp. 20–41.
10. Crizzle, A.M.; Bigelow, P.; Adams, D.; Gooderham, S.; Myers, A.M.; Thiffault, P. Health and wellness of long-haul truck and bus drivers: A systematic literature review and directions for future research. *Journal of Transport & Health* **2017**, *7*, 90–109.

11. Aria, M.; Cuccurullo, C. bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics* **2017**, *11*, 959–975.
12. Garfield, E.; Sher, I.H. KeyWords Plus^{TM}—algorithmic derivative indexing. *Journal of the American Society for Information Science* **1993**, *44*, 298–299.
13. Erkut, E.; Tjandra, S.A.; Verter, V. Hazardous materials transportation. *Handbooks in Operations Research and Management Science* **2007**, *14*, 539–621.
14. Androutsopoulos, K.N.; Zografos, K.G. A bi-objective time-dependent vehicle routing and scheduling problem for hazardous materials distribution. *EURO Journal on Transportation and Logistics* **2012**, *1*, 157–183.
15. Abkowitz, M.; Cheng, P.D.M. Developing a risk/cost framework for routing truck movements of hazardous materials. *Accident Analysis & Prevention* **1988**, *20*, 39–51.
16. Theofilatos, A.; Yannis, G. A review of the effect of traffic and weather characteristics on road safety. *Accident Analysis & Prevention* **2014**, *72*, 244–256.
17. Roshandel, S.; Zheng, Z.; Washington, S. Impact of real-time traffic characteristics on freeway crash occurrence: Systematic review and meta-analysis. *Accident Analysis & Prevention* **2015**, *79*, 198–211.
18. Federal Highway Administration. Real-Time System Management. Department of Transportation, https://ops.fhwa.dot.gov/511/index.htm, 2016. [Online; accessed 03-August-2018].
19. Wang, L.; Abdel-Aty, M.; Lee, J. Safety analytics for integrating crash frequency and real-time risk modeling for expressways. *Accident Analysis & Prevention* **2017**, *104*, 58–64.
20. Highway Performance Monitoring System. AADT Traffic. Department of Transportation, https://www.fhwa.dot.gov/policyinformation/hpms/shapefiles.cfm, 2018. [Online; accessed 04-August-2018].
21. Minnesota Department of Transportation. AADT Adjusment Factors. http://www.dot.state.mn.us/traffic/data/docs/tvp/AADT_Adjustment_Factors_for_Short_Duration_Traffic_Volume_Counts.pdf, 2017. [Online; accessed 03-May-2017].
22. Xu, C.; Wang, W.; Liu, P. A Genetic Programming Model for Real-Time Crash Prediction on Freeways. *IEEE Transactions on Intelligent Transportation Systems* **2013**, *14*, 574–586.
23. Sun, J.; Sun, J. A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data. *Transportation Research Part C: Emerging Technologies* **2015**, *54*, 176–186.
24. National Centers for Environmental Information. NOAA API. National Oceanic and Atmospheric Administration, https://www.ncdc.noaa.gov/cdo-web/webservices/v2, 2018. [Online; accessed 04-August-2018].
25. Dark Sky. Dark Sky API. The DarkSky Company, LLC, https://darksky.net/dev, 2019. [Online; accessed 24-February-2019].
26. Florida Department of Transportation. Roadway Characteristics Inventory Features & Characteristics eBook. http://www.fdot.gov/statistics/rci/, 2018. [Online; accessed 04-August-2018].
27. Washington, S.P.; Karlaftis, M.G.; Mannering, F. *Statistical and econometric methods for transportation data analysis*; Chapman and Hall/CRC, 2010.
28. Chen, W.; Guo, F.; Wang, F.Y. A survey of traffic data visualization. *IEEE Transactions on Intelligent Transportation Systems* **2015**, *16*, 2970–2984.
29. Guo, H.; Wang, Z.; Yu, B.; Zhao, H.; Yuan, X. Tripvista: Triple perspective visual trajectory analytics and its application on microscopic traffic data at a road intersection. 2011 IEEE Pacific Visualization Symposium (PacificVis). IEEE, 2011, pp. 163–170.
30. Ferreira, N.; Poco, J.; Vo, H.T.; Freire, J.; Silva, C.T. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *IEEE Transactions on Visualization and Computer Graphics* **2013**, *19*, 2149–2158.
31. Tsai, Y.T.; Alhwiti, T.; Swartz, S.M.; Megahed, F.M. Using visual data mining in highway traffic safety analysis and decision making. *Journal of Transportation Management* **2015**, pp. 43–60.
32. Pu, J.; Liu, S.; Ding, Y.; Qu, H.; Ni, L. T-Watcher: A new visual analytic system for effective traffic surveillance. Mobile Data Management (MDM), 2013 IEEE 14th International Conference on. IEEE, 2013, Vol. 1, pp. 127–136.
33. Tanahashi, Y.; Ma, K.L. Design considerations for optimizing storyline visualizations. *IEEE Transactions on Visualization and Computer Graphics* **2012**, *18*, 2679–2688.
34. NHTSA. FARS Speeding Data Visualization. United States Department of Transportation, https://www.nhtsa.gov/press-releases/usdot-releases-2016-fatal-traffic-crash-data, 2018. [Online; last accessed 7-Sept-2018].
35. Cui, W.; Zhou, H.; Qu, H.; Wong, P.C.; Li, X. Geometry-based edge clustering for graph visualization. *IEEE Transactions on Visualization and Computer Graphics* **2008**, *14*, 1277–1284.

36. Xie, Z.; Yan, J. Kernel density estimation of traffic accidents in a network space. *Computers, Environment and Urban Systems* **2008**, *32*, 396–406.
37. Liu, S.; Pu, J.; Luo, Q.; Qu, H.; Ni, L.M.; Krishnan, R. VAIT: A visual analytics system for metropolitan transportation. *IEEE Transactions on Intelligent Transportation Systems* **2013**, *14*, 1586–1596.
38. Zeng, W.; Fu, C.W.; Arisona, S.M.; Qu, H. Visualizing interchange patterns in massive movement data. Computer Graphics Forum. Wiley Online Library, 2013, Vol. 32, pp. 271–280.
39. Kraak, M.J. The space-time cube revisited from a geovisualization perspective. Proceedings of the 21st International Cartographic Conference. Citeseer, 2003, pp. 1988–1996.
40. Kapler, T.; Wright, W. GeoTime information visualization. *Information Visualization* **2005**, *4*, 136–146.
41. Tominski, C.; Schumann, H.; Andrienko, G.; Andrienko, N. Stacking-based visualization of trajectory attribute data. *IEEE Transactions on visualization and Computer Graphics* **2012**, *18*, 2565–2574.
42. Pack, M.L.; Wongsuphasawat, K.; VanDaniker, M.; Filippova, D. ICE–visual analytics for transportation incident datasets. Information Reuse & Integration, 2009. IRI'09. IEEE International Conference on. IEEE, 2009, pp. 200–205.
43. Cottrill, C.D.; Thakuriah, P.V. Evaluating pedestrian crashes in areas with high low-income or minority populations. *Accident Analysis & Prevention* **2010**, *42*, 1718–1728.
44. Chu, D.; Sheets, D.A.; Zhao, Y.; Wu, Y.; Yang, J.; Zheng, M.; Chen, G. Visualizing hidden themes of taxi movement with semantic transformation. Visualization Symposium (PacificVis), 2014 IEEE Pacific. IEEE, 2014, pp. 137–144.
45. van Huysduynen, H.H.; Terken, J.; Martens, J.B.; Eggen, B. Measuring driving styles: a validation of the multidimensional driving style inventory. Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. ACM, 2015, pp. 257–264.
46. Stevens, S. On the Theory of Scales of Measurement. *Science* **1946**, *103*, 677–680.
47. Frank, A.U., Different types of "times" in GIS. In *Spatial and temporal reasoning in geographic information systems*; Oxford University Press, New York, 1998; chapter 3, pp. 40–62.
48. Aigner, W.; Miksch, S.; Müller, W.; Schumann, H.; Tominski, C. Visual methods for analyzing time-oriented data. *IEEE Transactions on Visualization and Computer Graphics* **2008**, *14*, 47–60.
49. Croxford, B.; Penn, A.; Hillier, B. Spatial distribution of urban pollution: civilizing urban traffic. *Science of the Total Environment* **1996**, *189*, 3–9.
50. Han, W.; Wang, J.; Shaw, S.L. Visual Exploratory Data Analysis of Traffic Volume. MICAI 2006: Advances in Artificial Intelligence; Gelbukh, A.; Reyes-Garcia, C.A., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2006; pp. 695–703.
51. Alam, I.; Ahmed, M.F.; Alam, M.; Ulisses, J.; Farid, D.M.; Shatabda, S.; Rossetti, R.J. Pattern mining from historical traffic big data. IEEE Region 10 Symposium (TENSYMP), 2017. IEEE, 2017, pp. 1–5.
52. Nookala, L.S. Weather impact on traffic conditions and travel time prediction. M.S. Thesis. Department of Computer Science, University of Minnesota Duluth., 2006. Online; last accessed 04-Sept-2018.
53. Havre, S.; Hetzler, B.; Nowell, L. ThemeRiver: Visualizing theme changes over time. Information visualization, 2000. InfoVis 2000. IEEE symposium on. IEEE, 2000, pp. 115–123.
54. Van Wijk, J.J.; Van Selow, E.R. Cluster and calendar based visualization of time series data. Information Visualization, 1999.(Info Vis' 99) Proceedings. 1999 IEEE Symposium on. IEEE, 1999, pp. 4–9.
55. Kraak, M.J. Visualising spatial distributions. *Chapter* **1999**, *11*, 157–173.
56. Erdogan, S. Explorative spatial analysis of traffic accident statistics and road mortality among the provinces of Turkey. *Journal of Safety Research* **2009**, *40*, 341–351.
57. Wongsuphasawat, K.; Pack, M.; Filippova, D.; VanDaniker, M.; Olea, A. Visual analytics for transportation incident data sets. *Transportation Research Record: Journal of the Transportation Research Board* **2009**, pp. 135–145.
58. Romero, B. Traffic Accidents. http://brettromero.com/traffic-accidents-cyclists/, 2015. [Online; accessed 09-September-2018].
59. Galka, M. Traffic Accidents. http://metrocosm.com/map-us-traffic/, 2016. [Online; accessed 09-September-2018].
60. Pack, M.L. Visualization in transportation: challenges and opportunities for everyone. *IEEE Computer Graphics and Applications* **2010**, *30*, 90–96.
61. Das, S.; Avelar, R.; Dixon, K.; Sun, X. Investigation on the wrong way driving crash patterns using multiple correspondence analysis. *Accident Analysis & Prevention* **2018**, *111*, 43–55.
62. Liu, H.; Taniguchi, T.; Tanaka, Y.; Takenaka, K.; Bando, T. Visualization of driving behavior based on hidden feature extraction by using deep learning. *IEEE Transactions on Intelligent Transportation Systems* **2017**, *18*, 2477–2489.

63. Sawalha, Z.; Sayed, T. Traffic accident modeling: some statistical issues. *Canadian Journal of Civil Engineering* **2006**, *33*, 1115–1124.
64. Shi, Q.; Abdel-Aty, M. Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. *Transportation Research Part C: Emerging Technologies* **2015**, *58*, 380–394.
65. Hassan, H.M.; Abdel-Aty, M.A. Predicting reduced visibility related crashes on freeways using real-time traffic flow data. *Journal of Safety Research* **2013**, *45*, 29–36.
66. Hossain, M.; Muromachi, Y. A real-time crash prediction model for the ramp vicinities of urban expressways. *IATSS Research* **2013**, *37*, 68–79.
67. Yu, R.; Abdel-Aty, M. Utilizing support vector machine in real-time crash risk evaluation. *Accident Analysis & Prevention* **2013**, *51*, 252–259.
68. You, J.; Wang, J.; Guo, J. Real-time crash prediction on freeways using data mining and emerging techniques. *Journal of Modern Transportation* **2017**, *25*, 116–123.
69. Basso, F.; Basso, L.J.; Bravo, F.; Pezoa, R. Real-time crash prediction in an urban expressway using disaggregated data. *Transportation Research Part C: Emerging Technologies* **2018**, *86*, 202–219.
70. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Computers & Electrical Engineering* **2014**, *40*, 16–28.
71. Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517.
72. Goldberg, D.E.; Holland, J.H. Genetic algorithms and machine learning. *Machine Learning* **1988**, *3*, 95–99.
73. Kennedy, R. J. and Eberhart, Particle swarm optimization. Proceedings of IEEE International Conference on Neural Networks IV, pages, 1995, Vol. 1000.
74. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *Journal of Machine Learning Research* **2003**, *3*, 1157–1182.
75. Jović, A.; Brkić, K.; Bogunović, N. A review of feature selection methods with applications. Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on. IEEE, 2015, pp. 1200–1205.
76. Khalid, S.; Khalil, T.; Nasreen, S. A survey of feature selection and feature extraction techniques in machine learning. Science and Information Conference (SAI), 2014. IEEE, 2014, pp. 372–378.
77. Nagendra, S.S.; Khare, M. Principal component analysis of urban traffic characteristics and meteorological data. *Transportation Research Part D: Transport and Environment* **2003**, *8*, 285–297.
78. Lee, H.C.; Cameron, D.; Lee, A.H. Assessing the driving performance of older adult drivers: on-road versus simulated driving. *Accident Analysis & Prevention* **2003**, *35*, 797–803.
79. Li, Q.; Jianming, H.; Yi, Z. A flow volumes data compression approach for traffic network based on principal component analysis. 2007 IEEE Intelligent Transportation Systems Conference. IEEE, 2007, pp. 125–130.
80. Caliendo, C.; Guida, M.; Parisi, A. A crash-prediction model for multilane roads. *Accident Analysis & Prevention* **2007**, *39*, 657–670.
81. Guo, F.; Fang, Y. Individual driver risk assessment using naturalistic driving data. *Accident Analysis & Prevention* **2013**, *61*, 3–9.
82. Lee, J.; Abdel-Aty, M.; Shah, I. Evaluation of surrogate measures for pedestrian trips at intersections and crash modeling. *Accident Analysis & Prevention* **2018**.
83. Cook, R.D. Principal components, sufficient dimension reduction, and envelopes **2018**.
84. Tipping, M.E.; Bishop, C.M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **1999**, *61*, 611–622.
85. Schölkopf, B.; Smola, A.; Müller, K.R. Kernel principal component analysis. International conference on artificial neural networks. Springer, 1997, pp. 583–588.
86. Berkhin, P. A survey of clustering data mining techniques. In *Grouping multidimensional data*; Springer, 2006; pp. 25–71.
87. Rai, P.; Singh, S. A survey of clustering techniques. *International Journal of Computer Applications* **2010**, *7*, 1–5.
88. Fahad, A.; Alshatri, N.; Tari, Z.; Alamri, A.; Khalil, I.; Zomaya, A.Y.; Foufou, S.; Bouras, A. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing* **2014**, *2*, 267–279.
89. Hall, F.L.; Hurdle, V.; Banks, J.H. Synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relationships on freeways **1993**.

90. Kerner, B.S.; Rehborn, H. Experimental properties of complexity in traffic flow. *Physical Review E* **1996**, *53*, R4275.
91. Wu, N. A new approach for modeling of Fundamental Diagrams. *Transportation Research Part A: Policy and Practice* **2002**, *36*, 867–884.
92. Golob, T.F.; Recker, W.W. A method for relating type of crash to traffic flow characteristics on urban freeways. *Transportation Research Part A: Policy and Practice* **2004**, *38*, 53–80.
93. Xu, C.; Liu, P.; Wang, W.; Li, Z. Evaluation of the impacts of traffic states on crash risks on freeways. *Accident Analysis & Prevention* **2012**, *47*, 162–171.
94. Steenberghen, T.; Dufays, T.; Thomas, I.; Flahaut, B. Intra-urban location and clustering of road accidents using GIS: a Belgian example. *International Journal of Geographical Information Science* **2004**, *18*, 169–181.
95. Xie, Z.; Yan, J. Detecting traffic accident clusters with network kernel density estimation and local spatial statistics: an integrated approach. *Journal of Transport Geography* **2013**, *31*, 64–71.
96. Shen, L.; Lu, J.; Long, M.; Chen, T. Identification of Accident Blackspots on Rural Roads Using Grid Clustering and Principal Component Clustering. *Mathematical Problems in Engineering* **2019**, *2019*.
97. Kwon, O.H.; Park, S.H. Identification of Influential Weather Factors on Traffic Safety Using K-means Clustering and Random Forest. In *Advanced Multimedia and Ubiquitous Engineering*; Springer, 2016; pp. 593–599.