DigitalCommons @WayneState

Journal of Transportation Management

Volume 26 | Issue 1 Article 5

7-1-2015

Using visual data mining in highway traffic safety analysis and decision making

Yao-Te Tsai
Auburn University, yzt0007@auburn.edu

Huw D. Smith
Auburn University, hds0003@auburn.edu

Stephen M. Swartz
Auburn University, steve.swartz@auburn.edu

Fadel M. Megahed
Auburn University, fmm0002@auburn.edu

Recommended Citation

Tsai, Yao-Te, Smith, Huw D., Swartz, Stephen M., & Megahed, Fadel M. (2015). Using visual data mining in highway traffic safety analysis and decision making. Journal of Transportation Management, 26(1), 43-60. doi: 10.22237/jotm/1435709040

USING VISUAL DATA MINING IN HIGHWAY TRAFFIC SAFETY ANALYSIS AND DECISION MAKING

Yao-Te Tsai
Huw D. Smith
Stephen M. Swartz
Fadel M. Megahed
Auburn University

ABSTRACT

An ongoing, two-fold challenge involves extracting useful information from the massive amounts of highway crash data and explaining complicated statistical models to inform the public about highway safety. Highway safety is critical to the trucking industry and highway funding policy. One method to analyze complex data is through the application of visual data mining tools. In this paper, we address the following three questions: a) what existing data visualization tools can assist with highway safety theory development and in policy-making?; b) can visual data mining uncover unknown relationships to inform the development of theory or practice? and c) can a data visualization toolkit be developed to assist the stakeholders in understanding the impact of public-policy on transportation safety? To address these questions, we developed a visual data mining tool-kit that allows for understanding safety datasets and evaluating the effectiveness of safety policies.

INTRODUCTION AND LITERATURE REVIEW

Transportation accidents levy a significant cost on societies in terms of personal death or injury in addition to the economic costs. Road traffic injuries are the eighth leading cause of death, and the leading cause of death for individuals aged 15-29 (Lozano et al., 2012; World Health Organization, 2008). In 2010, transportation injuries have resulted in 1.24 million fatalities worldwide according to the World Health Organization (WHO), World Health Organization (2013, p. v). In addition to the lost lives, the costs associated with road traffic crashes runs to billions of dollars (Jacobs, Aeron-Thomas, & Astrop, 2000). These numbers are unacceptably high, especially since many of these fatalities can be avoided with evidence-driven road safety interventions.

Road safety interventions can be effective in reducing the number of accidents and/or mitigating their effects. The WHO states that "adopting and enforcing legislation relating to important risk factors – speed, drunk–driving, motorcycle helmets, seat-belts and child restraints – has been shown to lead to reductions in road traffic injuries" (World Health Organization, 2013, p. v). These five risk factors are a sample of a larger pool of behavioral factors that lead to accidents. There are increasing regulations worldwide that have been passed to cover these behavioral factors. However, "in many countries these laws are either not comprehensive in scope or lacking altogether. Governments must do more to ensure that their national road safety laws meet best practice, and do more to enforce these laws" (World Health Organization, 2013, p. v) The problem is complex in the U.S., since highway safety policies can be different in neighboring states and the identification of best practice is often unclear (Governors Highway Safety Association, 2013).

One approach to identifying best practices is to investigate the causes of vehicle crashes, assess the factors that are correlated with high severity/ frequency accidents, and propose interventions that can prevent/mitigate these accidents. Examples are provided in the works of Shibata and Fukuda (1994), Massie, Campbell, and Williams (1995), Shankar, Mannering, and Barfield (1995), Al-Ghamdi (2002), K. K. W. Yau (2004), Aarts and van Schagen (2006), and Kent, Coulter, and Coulter (2011). These papers followed a common framework that started with identifying (or using previously identified) causal factors and then validating how these factors contribute to traffic crashes. While these approaches are built on a solid statistical foundation, they are often difficult to understand by the stakeholders due, in a large part, to the number/complexity of variables and relationships in the data. Additionally, it is difficult to evaluate whether differences among locations affect the generalizability of their conclusions across geographical regions with different environmental and behavioral conditions.

Another approach to identify the best practices is to retrospectively evaluate whether safety regulations have been effective in reducing accident, injury and/or fatality rates. It should be noted that such studies not only capture the differences pre and post regulation changes, but they can also assess the impact of varying enforcement levels (especially if they compare across states and/or counties). Thus, these studies can be seen to measure whether the policies are comprehensive (or effective), an important consideration highlighted above in the WHO report. These studies investigated several behavioral-related regulations, including: a) the impact of hand-held cell phone bans on reducing fatalities (Jacobson et al., 2012; Nikolaev, Robbins, & Jacobson, 2010; Sampaio, 2012); b) the impact of medical marijuana legislation on reducing fatal crashes involving alcohol through substitution effects (Anderson & Rees, 2011); and c) the effectiveness of seatbelt laws in reducing the number of teenage traffic fatalities (Carpenter & Stehr, 2008). The results of this research are usually explained by statistical summaries and p-value tables, which are singularly unconvincing to non-scientist public policy decision makers. In addition, the general public must frequently be convinced politically to support the rationale behind changes to existing laws; and complex statistical analyses can be ineffective in making a convincing argument.

This research proposes that there is a need for new and innovative data-driven traffic safety models that can be both useful for researchers in uncovering promising areas for safety research, and tools for improving how safety research findings can be presented to and understood by the different stakeholders (general public, policy-makers, researchers, etc.). In this paper, a new approach to showing how visualization tools can address this gap is presented, with a focus on their use in detecting trends in highway safety and affecting safety policy making.

In the following sections, the field of Visual Data Mining (VDM) is discussed, and the concept of using this method to generate insights from spatiotemporal datasets is introduced. Following that we present a brief description of the methodology and the datasets used in providing some examples of how VDM tools could be used for this purpose. We try to provide some simple, obvious examples of how data visualization tools can uncover relationships that may not be captured by traditional modeling methods. Finally, examples are used to demonstrate how the developed visualization toolkit can assist in evaluating the impact of safety policy changes.

VISUAL DATA MINING

VDM is a tool which can aid in exploring hidden information and uncovering trends and patterns from other non-visual methods, It is a data mining approach that is based on the integration of multiple concepts. Visualization techniques which provide powerful and useful visual capabilities in a hypothesis testing mode for users are designed to support data mining tasks before the analysis actually begins (De Oliveira & Levkowitz, 2003). Visual data mining can overcome the gap between interacting with massive datasets and acquiring more intelligent information from the the analysis(Simoff, Böhlen, & Mazeika, 2008, p. i). Presenting data visually has been found to be one of the simplest and most effective way to discover trends in data so that humans can make better decisions (Greitzer, Noonan, & Franklin, 2011; Han & Kamber, 2011; Keim, Müller, & Schumann, 2002; Simoff et al., 2008).

The use of "visualizations" (visualization applications) to uncover patterns is not a new phenomenon in public safety. In disease control, in 1854, John Snow plotted a massive cholera outbreak overlaid on a map of the city of London in order to discover the cause of cholera. This is said to be the first geographical analysis of disease data (Rajaraman, Leskovec, & Ullman, 2012, p. 3-4). Wickham (2013) briefly reviews the history of how statisticians used visualization techniques to assist in the interpretation of complex analytical results. Good design of graphical displays can help to understand complicated procedures and algorithms and solve complex problems without making assumptions. When automated data mining tools fail, visualization for exploring data can be used to support model explanations and lead to better results.(Keim et al., 2002).

It is important to mention that not all graphical representations of data are useful, and some can be misleading. Wickham (2013, p. 39-40) includes a list of some of the formal instructions on the effective use of visualizations that was written in 1901 by the International Institute of Statistics. Below, we repeat these oft-forgotten recommendations:

1) "We must keep symbols to a minimum, so as not to overload the reader's memory. Some ancient authors, by covering their cartograms with hieroglyphics, made them indecipherable."
2) "One of us recommends adopting scales for ordinate and abscissa so the average slope of the phenomenon corresponds to the tangent of the curve at an angle of 45 degrees."
3) "Areas are often used in graphical representations. However, they have the disadvantage of often misleading the reader even though they were designed according to indisputable geometric principles. Indeed, the eye has a hard time appreciating areas."
4) "We should not, as it is sometimes done, cut the bottom of the diagram under the pretext that it is useless. This arbitrary suppression distorts the chart by making us think that the variations of the function are more important than they really are."
5) "To increase the means of expression without straining the reader's memory, we often build cartograms with two colors. And, indeed, the reader can easily remember this simple formula: 'The more the shade is red, the more the phenomenon studied surpasses the average; the more the shade is blue, the more phenomenon studied is below average."

While there are no universal visuals that will work for every application domain and problem, there are several factors/guidelines that can help in selecting/developing informative statistical graphics. For example, Tufte (1983, p. 13-15) introduced the term graphical excellence to reflect on graphics that communicate complex ideas with clarity, precision and efficiency. Keim et al. (2002) provided some general rules for expressive and effective information visualization. These guidelines were used in developing the graphics for traffic safety presented in this paper.

METHODS AND DESCRIPTION OF DATASETS

The purpose of this investigation is to develop and apply a small subset of data visualization tools to a large, complex dataset of transportation safety data for the purposes of addressing three research questions:

RQ1. What existing data visualization tools might be appropriate to inform researchers in theory development and decision makers in setting transportation safety policy?

RQ2. Can a data visualization tool be developed to assist in uncovering previously unknown constructs/relationships to inform the development of theory or practice?

RQ3. Can a data visualization tool be developed to assist decision makers in applying and evaluating public policy choices in order to improve transportation safety in practice?

RQ1 will be answered through a focused literature review on data visualization tools used in the context of transportation and safety, as well as a consideration of tools applied successfully in other contexts. A small number of tools will be developed based on this review. RQ2 and RQ3 will be answered by applying the tools developed against "real world" transportation data in an effort to demonstrate efficacy and at least minimal utility of the general approach.

Transforming accident related data into graphical information in order tofacilitate further analysis is our basic principle. Three common types of data in the transportation safety area of interest are temporal, spatiotemporal, and the effectiveness of policymaking (before-after comparison). The temporal data was analyzed by a calendar-based clustering application, and the graphical results show the characteristic of the clusters; thereby aiding researcher insight and theory development. Next, the mapping tool combines the geographic data with accident related information and statistical reports displayed on a map as an example of the treatment of spatiotemporal data. It consists of the Visual Basic Application-based (VBA) dataset and Microsoft MapPoint. The constraint for the mapping tool is the limitation of VBA functions. The speed of executing the tool relies on the quality and quantity of the VBA codes and dataset. While the current application is scalable, additional refinement is needed when the volume of the dataset increases. The current application can handle up to 1,048,576 rows by 16,384 columns of data and its execution speed is a function of computing resources.

In this paper, we have used two datasets to depict the effectiveness of the proposed/ developed visual data mining tools in enhancing our understanding of emerging patterns and trends that are related to traffic safety. Both datasets are collected for U.S. traffic by state and/or governmental agencies. The first dataset consists of traffic flow counts per hour collected by the Alabama Department of Transportation (ADoT) using a traffic camera between January 2005 and December 2010 (Alabama Department of Transportation, 2011). We did not use the 2011 and 2012 data in our dataset since they involved a massive amount of missing data. The data captures directional traffic flows caught hourly by the sensors on a busy interstate highway (I-85, sensors located 6.0 miles south of Macon Co. Line). The sensor captures whenever a motor vehicle passes the location and counts for a record. It is important to note that traffic volume is one of the measures traditionally used to normalize accident data (Gregoriades et al., 2011; Ivan, 2004; Laessig & Waterworth, 1970; Stamatiadis, Agent, & Bizakis, 1997). However, the estimation of the target variable may contain misleading information because the traffic volumes vary by location and time. Thus, we investigate how visual data mining tools can be used to disaggregate traffic flows to account for seasonal variations and emerging patterns.

The second dataset was gathered using the Fatality Analysis Reporting System (FARS) from the National Highway Traffic Safety Administration (NHTSA) (National Highway Safety Adminstration, 2012). For the purpose of our analysis, the captured data frame is for commercial vehicles (trucks and buses) involved in the fatal accidents from 2002 to 2011 within 12 U.S. southeastern states. Commercial vehicles are trucks with gross vehicle weight greater than 10,000 lb. or buses that can hold more than 10 passengers. The 12 southeastern states under consideration are: Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia, and West Virginia. These states have similar demographic factors and weather conditions, which reduces the exogenous effects of weather, demographics and urban/rural differences. We have used this dataset to demonstrate the reality of visualized spatiotemporal safety data and the impact of policy-making on commercial vehicle safety.

REVIEW OF RELEVANT DATA VISUALIZATION TOOLS SUBSET

Visualization tools can be used in research to observe the occurrence of traffic events and explore the information from the events. One of the concerns for practitioners is to provide more efficient and understandable results for the stakeholders by improving the current approaches or developing other useful techniques. The key is to know how to connect the data and the result presentation. We introduce the highlights of recent transportation safety research and emphasize some of the appropriate data visualization tools that can be used.

For a large dataset, it is difficult to move forward without understanding the basis of the contents. The histogram, one of the basic visualization tools, can indicate the distribution of information. The daily pattern and seasonal trends of the fatal crash data from 2001 to 2011 can be explored in histograms, according to N. Yau (2013). From the results, he stated that most of the crashes occur in the evening in terms of time of day, and trends can also be found regarding the day of the week and month. Instead of looking at the meaning of each data point, the aggregate values of the data sometimes extract more hidden information.

Another example can be found in traffic volumes, which researchers started to study in the mid 1960's (Roddick & Spiliopoulou, 2002). Except for various plots, cluster analysis was initially adopted to analyze traffic volumes in the 1990's (Black, 1991; Flaherty, 1993). Weijermars and van Berkum (2005) classified highway flow patterns by the daily flow profile chart and defined the characteristic of clusters by a summary table. Based on the data selected, only 118 days were included in their analysis. In addition, the summary table and the presentation of the daily pattern charts from the dataset do not interpret the results clearly. Van Wijk and Van Selow (1999) developed a calendar-based clustering application to present daily patterns and seasonal changes in employee and power demand, which displays the results on a calendar. This was the approach adopted for this research used to analyze traffic flows. The details are interpreted in the following section.

Some researchers have used Geographic Information Systems (GIS) to transform data into a map to analyze traffic accidents (Erdogan et al., 2008; Liang, Mo'soem, & Hua, 2005; Yi et al., 2001). Besides using a GIS, a large number of practitioners constructed internet-based mapping tools to monitor different safety related incidents such as the Global Incident Map and the National Incident Map. They used graphics to display where the incidents occurred. A text box with detailed information is also shown with the selected incident. The text box may contain useful information, but the detailed information sometimes causes difficulty in interpretation and communication. In addition, the information could not be extracted or synthesized among multiple points. Some advanced mapping tools are designed and displayed on the websites such as Baton Rouge Traffic Incident Map, CrashMap, and English Road Safety Comparison to show the incidents with selected variables.

The behavioral factor determination includes finding important factors in the accident analyses and evaluating the effectiveness of safety policymaking. A histogram and bar chart showing the data exploration are usually embraced in the analyses. Nevertheless, none of the transportation safety research has used visualized statistical results such as heat map. The heat map can show the significance of testing results transforming from the statistical reports.

DEVELOPING A VDM TOOLKIT

In the last section, we presented some of the visualization tools that are relevant to transportation safety.. Here, two tools are used to demonstrate the power of visualization in uncovering unknown contrasts, which can be used in informing the relevant stakeholders. First, we present how the approach of Van Wijk and Van Selow (1999) can be used to detect seasonal trends in the Alabama traffic flows. Animation can then be used to highlight how spatiotemporal traffic accidents can be depicted on a map.

Uncovering Seasonal Patterns in Traffic Flows

Using the ADoT dataset and the approach of Van Wijk and Van Selow (1999), we examine the traffic flows for 2005. Since the approach is based on k--means clustering, we also explore the effect of the number of clusters (k) on the observed patterns. One can think of the choice of the value for k as the degree of granularity required for the data analysis. Typical values of k can range from 2-12 depending on the application and a priori theoretical framework. The result for k=2 is portrayed in Table and Figure 1.

The table shows a “traditional” calendar-based approach to showing the layout between two clusters on a day-day basis. For each month in the calendar, every day of the week in that month is coded for each respective means cluster, while the grayed out cells represent days falling into an adjacent month. By contrast, the Figure breaks down the daily data into hourly segments- and reveals the reason why the two clusters are statistically separable. The y-axis represents the average traffic volume per cluster, the x-axis represents the time of the day, and the two lines correspond to the different clusters. Prior to the analysis, one could hypothesize that using k=2 should result in distinguishing between weekdays and holidays. However, the two clusters depicted in Figure 1 show that in general Fridays have a different pattern than weekdays with larger counts of vehicles on the road starting from around 9 am until midnight. This pattern is also observed on spring/summer Sundays as well as days around holidays (Martin Luther King, Jr. Day, Fourth of July, Thanksgiving, Christmas, etc.) as depicted by the calendar based table.

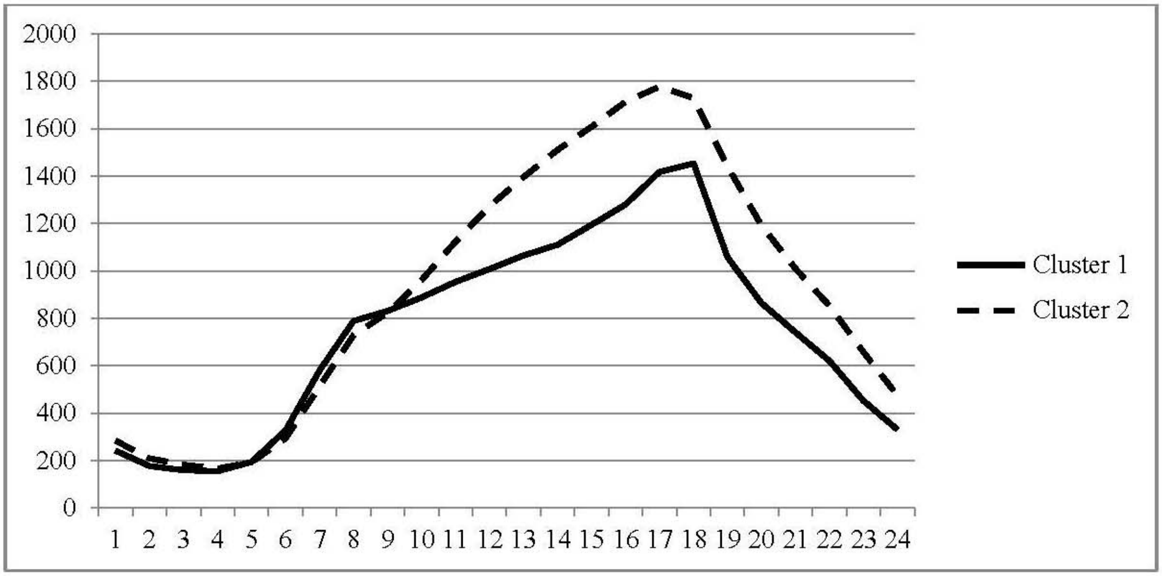
If more detail is needed, the VBA tool allows the user to pick the value of k and see the corresponding effect. Table and Figure 2 provides the results for k=5 clusters. Here, one can see that the patterns become different (i.e. not only magnified, but having different shapes) for the different clusters. For example, the cluster corresponding to Saturdays (cluster 3) shows a uniform traffic flow peak between 9 am and 5 pm. Such a pattern is quite distinct from the regular weekday patterns (clusters 4 and 5) where the peak is around 5 pm when most day shift employees travel home. Cluster 4 indicates that employees leave work earlier (and in greater numbers) on Fridays, which is an expected outcome. Thus, the use of clustering can provide evidence for daily and seasonal effects, which can inform hypotheses and research regarding traffic volumes. The calendar view allows the observer to capture all the information in one screen (which presents the contribution of the work of Weijermars and van Berkum (2005)). We present the case of k=8 in Table and Figure 3. While the increase in number of clusters have resulted in clusters that are almost identical (i.e.

clusters 5 and 6), we were able to capture a significant departure from the patterns described above in cluster 3. Cluster 3 is significant even though it captures only five Saturdays in the Fall. This result suggests an underlying phenomenon not previously anticipated by the two principle investigators due to their unfamiliarity with local custom. Upon further investigation, the researchers uncovered a local cultural phenomenon driving increased traffic volumes. These days correspond to five out of eight Saturdays when the Auburn University college football team played home games. With a stadium capacity over 87,000, these football games result in heavy commuter traffic on I-85 which passes through the city of Auburn. It is interesting to mention that all these five games were morning/early-afternoon games with a latest start of 2:30 PM local time. On the other hand, the remaining three games (Sept. 3rd, Oct 1st and Nov. 12th) were all evening games with an earliest start of 6 PM local time. Therefore, cluster 3 captured a coherent set of events that have a tremendous impact on the local community and that has a unique traffic pattern consistent with alumni driving to Auburn from Mobile and Montgomery (two of the largest cities in the state) to watch the game. Note that Station 44 is approximately 40 miles away from Auburn and I-85 is the only interstate which can be used to drive to Auburn University. This simple (somewhat obvious to North American followers of NCAA football) example serves simply to illustrate the the ability of this form of analysis to highlight unusual, event driven phenomenon.

TABLE 1
TRAFFIC VOLUME, 2 CLUSTER MEANS, TABULAR

S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S
January							**February**							**March**						
						1			1	1	1	2	1			1	1	1	2	1
2	1	1	1	1	1	1	2	1	1	1	1	2	1	2	1	1	1	1	2	1
1	1	1	1	1	2	1	1	1	1	1	1	2	1	2	1	1	1	1	2	1
1	2	1	1	1	1	1	1	1	1	1	1	2	1	2	1	1	1	1	2	1
1	1	1	1	1	1	1	1	1						1	1	1	1	1		
1	1																			
April							**May**							**June**						
					2	2	2	1	1	1	1	2	1				1	1	2	1
2	1	1	1	1	2	2	2	1	1	1	1	2	1	2	1	1	1	1	2	1
2	1	1	1	1	2	1	2	1	1	1	1	2	1	1	1	1	1	1	2	1
2	1	1	1	1	2	1	2	1	1	1	2	2	1	2	1	1	1	1	2	2
2	1	1	1	1	2	1	1	2	1					2	1	1	1	2		
July							**August**							**September**						
					2	1		1	1	1	1	2	2					1	2	2
1	2	2	1	2	2	2	2	1	1	1	1	2	1	1	1	1	1	1	2	1
1	1	1	1	1	2	1	2	1	1	1	1	2	1	1	1	1	1	1	2	1
2	1	1	1	1	2	1	2	1	1	1	1	2	1	1	1	1	1	1	2	1
2	1	1	1	1	2	2	2	1	1	1				1	1	1	1	1	2	
2																				
October							**November**							**December**						
						2			1	1	1	2	1					1	2	1
1	1	1	1	1	2	1	2	1	1	1	1	2	1	1	1	1	1	1	2	1
2	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1
2	1	1	1	1	2	1	1	1	2	2	1	2	2	1	1	1	1	2	2	1
2	1	1	1	1	2	1	2	1	1	1				1	2	2	2	2	2	1
1	1																			

FIGURE 1
TRAFFIC VOLUME, 2 CLUSTER MEANS, GRAPHICAL



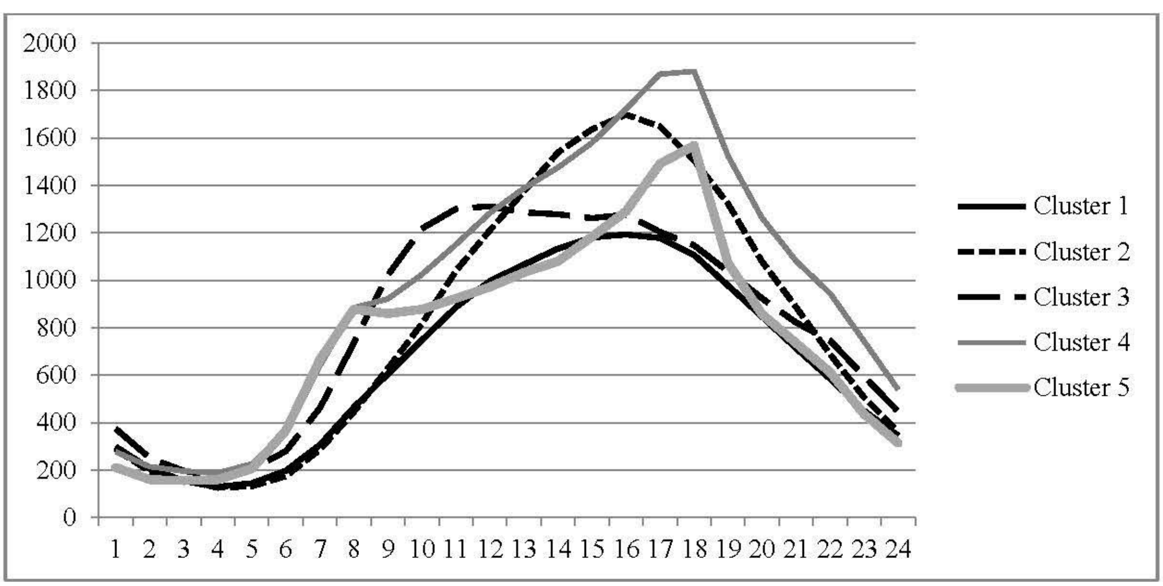
While this may seem like an “obvious” factor to consider for those familiar with the popularity of American college football, two points need to be made. First, not all transportation safety researchers are familiar with all characteristics of human behavior that may drive traffic patterns. The tool can discover “local” or “unique” factors the researchers may be unaware of *a priori*. Second, this visualization can also be used to uncover a previously unaccounted for phenomenon, and explain or highlight its impact on traffic patterns. This was the case for two of the researchers on this project. For example, the tool can then be used to drive policy decisions for government (increased policing on game days, re-routing of commercial traffic, postponement of lane closures, etc.) and industry (alternative routing during congestion, timing of travel through congested lanes, etc.) alike. The tool can be used for both discovering and alleviating the impact of “predictable” event-driven safety and efficiency factors.

This tool is also useful for exploratory disconfirmation. In recent years, there has been a strong geographical shift in the location of manufacturing plants in North America. Specifically, major assembly plants (and their suppliers) have been relocating to the Southeast due to the availability of a trained, efficient and lower cost (non-union) workforce, proximity to major logistics hubs, as well as state support and financial incentives in the form of lower taxes. From a research standpoint, one would anticipate greatly increased volumes of commercial traffic along the major corridors linking vendors and manufacturers. One would expect that this might lead to increased accident rates and reduced transportation efficiencies. Because the ADoT provided traffic volume reports since 2005, the comparison of traffic volumes before and after automotive plants arriving was possible. However, this has not been shown to be the case with traditional analytical methods. The use of the tool by the investigators revealed a potential explanation for the lack of evidence. The economic crisis also hit the automotive industry hard in 2008, and this moderating factor could have suppressed any expectation of increased activity. While it is quite possible that other tools could have been used to uncover and explore the influence of counter-balanced factors (economic activities), we found it particularly useful in guiding our investigation into this hypothesis.

TABLE 2
TRAFFIC VOLUME, 5 CLUSTER MEANS, TABULAR

S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S
January							**February**							**March**						
						1			5	5	5	4	1			5	5	5	4	3
2	5	5	5	5	5	1	2	5	5	5	5	4	1	2	5	5	5	5	4	1
2	5	5	5	5	4	1	2	5	5	5	5	4	3	2	5	5	5	5	4	3
1	2	5	5	5	5	1	2	5	5	5	5	4	1	2	5	5	5	5	4	1
1	5	5	5	5	5	1	2	5						1	5	5	5	5		
2	5																			
April							**May**							**June**						
					4	2	2	5	5	5	5	4	3				5	5	4	3
2	5	5	5	5	4	2	4	5	5	5	5	4	3	2	5	5	5	5	4	1
2	5	5	5	5	4	1	2	5	5	5	5	4	3	1	5	5	5	5	4	3
2	5	5	5	5	4	1	2	5	5	5	5	4	3	2	5	5	5	5	4	3
2	5	5	5	5	4	1	1	2	5					2	5	5	5	4		
July							**August**							**September**						
					4	3		5	5	5	5	4	3					5	4	2
1	4	4	5	4	4	4	2	5	5	5	5	4	3	1	1	5	5	5	4	3
1	1	5	5	5	4	3	2	5	5	5	5	4	3	1	5	5	5	5	4	3
2	5	5	5	5	4	3	2	5	5	5	5	4	3	1	5	5	5	5	4	3
2	5	5	5	5	4	3	4	1	5	5				1	5	5	5	5	4	
2																				
October							**November**							**December**						
						2			5	5	5	4	3					5	4	1
2	5	5	5	5	4	3	2	5	5	5	5	4	1	1	5	5	5	5	4	1
2	5	5	5	5	4	3	2	5	5	5	5	4	3	1	5	5	5	5	4	3
2	5	5	5	5	4	1	1	5	4	4	3	2	2	1	5	5	5	4	4	1
2	5	5	5	5	4	3	4	5	5	5				1	2	2	2	4	4	3
1	5																			

FIGURE 2
TRAFFIC VOLUME, 5 CLUSTER MEANS, GRAPHICAL



Visualizing Spatiotemporal Transportation Safety Data

Compared to the relatively expensive commercially available GIS analytical software systems, the researchers developed a low cost VBA-based interactive visualization tool that can transform accident data into an animated map. In addition to displaying location and time on the map, the tool is able to show three other variables dynamically denoted by symbols, shapes, and colors selected by the user. Also, the built-in information (such as population per state) can be displayed on the map. This multidimensional visualization tool is proposed as a potentially informative and flexible way to provide an effective overview for users to analyze the data. While this is impossible to replicate in grayscale, it is hoped that Figure 4 can provide a reduced fidelity version of what is possible.

In Figure 4, the map layer is shaded by population size for each state. The darkest hue of the states represents a population size greater than 30 million. The shade of the symbols show the accidents that have occurred on the different days of the week. It should be noted that the user can select the number of days that will be shown on the map, which allows for visualizing different hypotheses and research questions. The symbols (on the right side of Figure 4) indicate the number of fatalities per traffic accident. Additional relevant information can be depicted with the histogram, line chart, bar chart and/or pie chart, which we provide as a part of the toolkit. In Figure 4, we summarize the counts of types of routes where fatal accidents occurred in 2011. State highways show the highest occurrences of accidents. We believe that this snapshot view of the data can be very informative, especially with the ability of the user to query and select specific ranges (or values) for the variables he/she would want to depict.

This visualization tool provides a complete accident monitoring system, seeing different accident-related variables associated with the location and time in the same graph. The users can now know the frequency of accidents based off of location, time, weather conditions, type of vehicle or any other variable of interest they select. Researchers or public policy decision makers can get precise information on how to interpret the data in the data preparation stage and then move forward to analyze the data.

DEVELOPING A DATA VISUALIZATION TOOLKIT TO EVALUATE SAFETY POLICY CHANGES

In this last section, we use our data visualization toolkit to demonstrate the hidden patterns from the traffic flows and examine the animated transformation from spatiotemporal traffic accidents onto a map. In this section, a new data visualization form from the toolkit is introduced. A statistical analysis is performed initially to catch the potential impact of policy changes. The conclusions are displayed on a heat map where different colors shade the significance of policy adoption and symbols simply represent the results of interest.

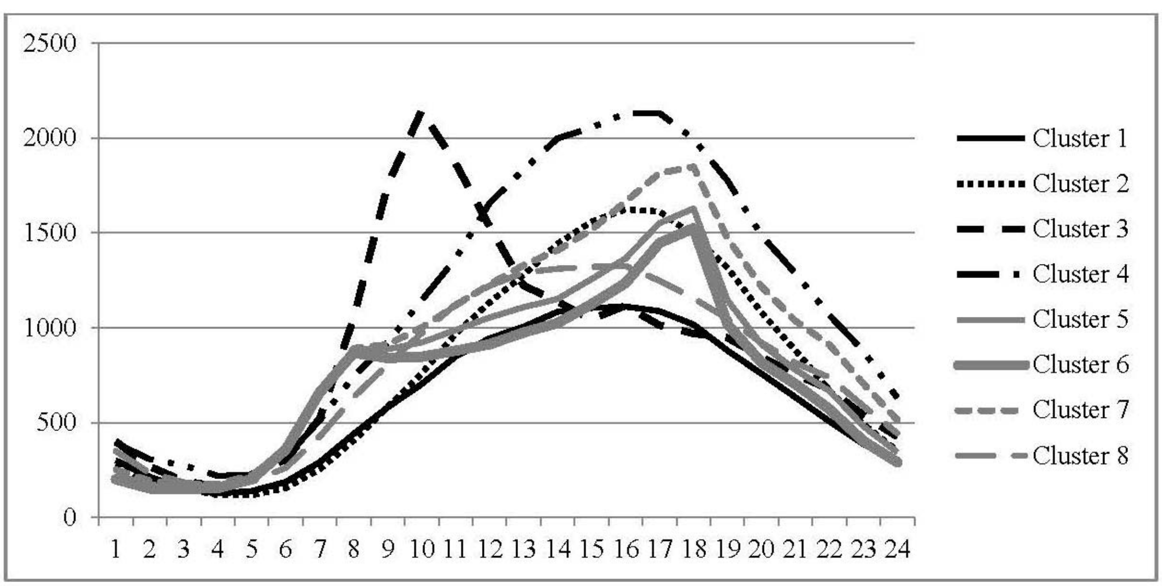
Examining the Potential Impact of Safety Policy on Accidents

The example used for this demonstration is the effectiveness of “Distracted Driving Laws” which have been enacted in recent years. Within the 12 southeastern states, seven have banned texting while driving including Arkansas, Georgia, Kentucky, Louisiana, North Carolina, Tennessee, and Virginia before January 2012. The laws apply to all vehicle drivers including CMV drivers. There are two accident measures within southeastern states for the time periods before and after text messaging ban laws were enacted. These two measures are fatal injury rate and non-fatal injury rate standardized by 100 million vehicle miles traveled by all vehicles as a measure of overall highway safety. Based on the state size, there are four different indicators:

TABLE 3
TRAFFIC VOLUME, 8 CLUSTER MEANS, TABULAR

S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S
January							**February**							**March**						
						1			6	6	6	7	8			6	6	5	7	8
2	5	5	6	6	5	1	2	5	6	5	5	7	8	2	6	6	6	5	7	8
2	6	6	6	6	7	8	2	6	6	6	5	7	8	2	6	6	6	5	7	8
1	2	6	6	6	5	1	2	5	6	6	5	7	8	2	6	6	5	5	7	1
1	6	6	6	6	5	1	2	6						2	5	6	5	6		
2	6																			
April							**May**							**June**						
					7	8	2	5	6	5	5	7	8				5	5	7	8
2	5	5	5	5	7	2	7	5	6	5	5	7	8	2	6	5	5	5	7	1
2	6	6	6	5	7	8	2	6	5	5	5	7	8	2	6	6	5	5	7	8
2	6	6	6	5	5	8	2	6	6	5	5	7	8	2	5	5	5	5	7	8
2	6	6	6	5	7	1	1	4	5					2	5	5	5	7		
July							**August**							**September**						
					7	8		5	5	5	5	7	8					6	7	8
8	4	7	5	5	4	4	2	5	6	5	5	7	8	1	2	6	6	6	7	3
1	1	6	5	5	7	8	2	5	6	6	5	7	8	2	6	6	6	6	7	3
2	5	6	5	5	7	8	2	6	6	6	6	7	8	2	6	6	6	5	7	3
2	5	5	5	6	7	8	4	1	6	6				1	6	6	6	5	7	
2																				
October							**November**							**December**						
						8			6	6	5	7	8					5	7	8
2	6	6	6	5	7	8	2	6	6	6	5	7	1	1	6	6	6	5	7	8
2	5	6	6	5	7	8	2	6	6	6	5	4	3	1	6	6	6	5	7	8
2	6	6	6	5	7	8	1	6	7	4	5	2	4	8	5	5	5	7	7	1
2	6	6	6	5	7	3	4	6	6	6				1	4	7	2	7	4	8
2	6																			

FIGURE 3
TRAFFIC VOLUME, 8 CLUSTER MEANS, GRAPHICAL



100 million vehicle miles traveled, 100,000 population, 100,000 registered vehicles, and 100,000 licensed drivers, for calculating the fatality rate as well as injury rate. Nevertheless, 100 million vehicle miles traveled is more accurate for measurement. The main target is to test the efficiency of those states that have adopted the distracted driving law, and to transform the results into a graphical pattern. In other words, better efficiency means there is significant evidence to show that the fatality rate after embracing the distracted driving law is lower than before.

The first step of the analysis is conducted at a 5% significance level by testing the hypothesis that the text messaging ban law had a positive influence on reducing fatal and non-fatal injury rates (Table 4). The F-test is applied first to determine if equal variances are used to compare two populations in the given time frame from 2002 to 2011, which are pre-law periods and post-law periods. Four out of seven states have a p-value lower than 0.05 in the t-test. Considering compound effects such as other policies deployed, the result can only provide relatively sufficient evidence that the distracted driving law may reduce the fatal injury rate in these four states. The hypothesis to test the impact of policy change for the non-fatal injury rates led to different results. Georgia did not have significant results for reducing non-fatal injury rates while Arkansas joined the group with a low p-value. Meanwhile, North Carolina and Tennessee are the only two states to have high p-values demonstrating the influence of the distracted driving law.

Visualizing the Statistical Results of Policy Changes

In order to improve the understanding of the impact of policy changes, we visualized the outcomes by plotting the results on the U.S. map according to the p-values (Figures 5 and 6). Three categories are classified for the potential impact of policy making by symbols: the smiley face, mad face, and stop sign. A "smiley face" indicates where the policy has had significant impact in reducing fatality and injury rates, the "mad face" shows the opposite result, and the "no sign" means the state has not adopted the selected policy. In addition, the gradient of the grayscale specifies the significance of each state's results. A stronger result or higher degree of statistical significance is represented by clear or no shading; less strong results are indicated by a darker shading as indicated in the legend on the left lower panel. In this case, North Carolina, which is painted dark gray, is affected most positively by the policy. Virginia, which shows "mad face" and darkest color, indicates a lack of evidence that there was any reduction in either fatal or non-fatal accidents.

For policy makers who do not have a statistical background, graphical information would attract their attention and provide clearer results. Therefore, they are able to have quick responses for the effectiveness of traffic policies and are able to ultimately make better decisions. For those audiences who know how to read statistical reports, they may not be willing to focus on understanding the complicated tables due to busy work or other issues; the graphical results can provide clear directions in less time. The example is only to provide one single policy evaluation with 12 southeastern states; however, once the results correspond to the multiple policy evaluation in the entire U.S., the complexity of reading and interpreting the reports increase. It is suggested that transforming p-values into a colored map could improve clarity and understanding.

This easy-to-use and straightforward visualization tool can also show the impact of any policy making after gaining the results from statistical analysis. Although discussing a single traffic policy may ignore the compound effects with other policies, this tool shows a potential way to present how policies impact the drivers' behavior in easy to understand visual form.

CONCLUDING REMARKS

To effectively learn from the ever-increasing volume and complexity of traffic data collected,

FIGURE 4
SPATIOTEMPORAL DATA MAPPING TOOL INTERFACE

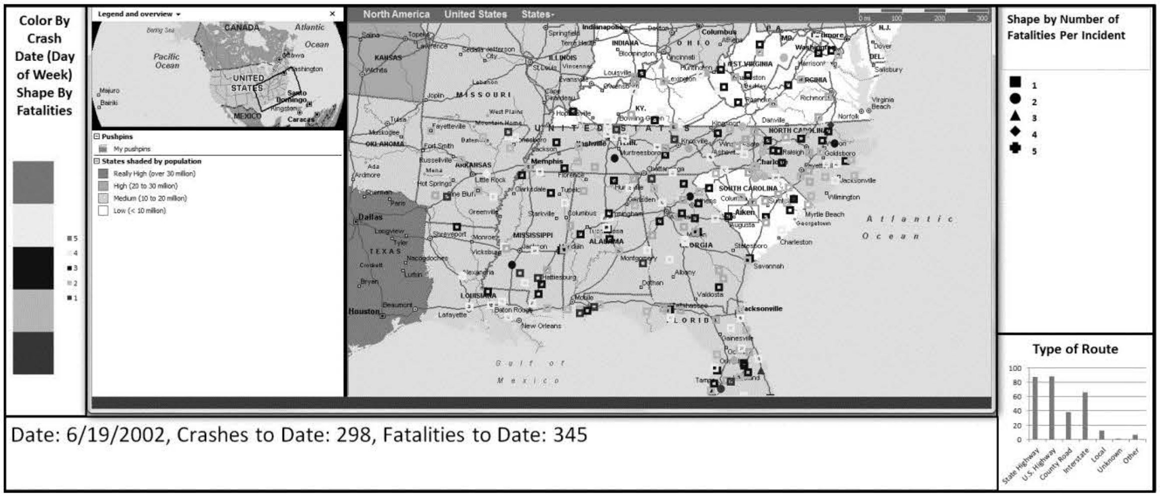


TABLE 4
PRE AND POST LAW COMPARISON FOR FATAL AND NON FATAL INJURY RATES

	Fatal			Non-Fatal		
	F test	Test type	T test	F test	Test type	T test
Alabama		NA			NA	
Arkansas	0.0643	Pooled	0.1413	<0.0001	Not Pooled	<0.0001
Florida		NA			NA	
Georgia	0.0850	Pooled	0.0306	0.0730	Pooled	0.0748
Kentucky	0.3056	Pooled	0.0729	0.1706	Pooled	0.3795
Louisiana	0.8000	Pooled	0.0786	0.1950	Pooled	0.1784
Mississippi		NA			NA	
North Carolina	0.1058	Pooled	<0.0001	0.02512	Not Pooled	<0.0001
South Carolina		NA			NA	
Tennessee	0.0612	Pooled	0.0029	0.0001	Not Pooled	<0.0001
Virginia	0.7328	Pooled	0.7542	<0.0001	Not Pooled	0.318
West Virginia		NA			NA	

this research effort explored how visualization techniques can be used to extract trends and to gain further understanding about factors of interest affecting transportation safety. Traffic safety is a critical issue to the trucking industry and is also very relevant to highway funding decisions. First, we examined how existing visual data mining tools can assist in theory development and policy making. The focus was on presenting some of the tools that have not been heavily used in the exploratory data analysis of transportation datasets. Some of these tools have been used for exploring similar traffic datasets (e.g. the work of N. Yau (2013)), while others have been applied outside the transportation field.

Using the insights from these methods and some fundamentals of visual data mining, we developed a new and potentially useful "visualization toolkit" that can be used to uncover unknown relationships in traffic volume; primarily in the clustering of traffic flows and visualizing the patterns associated with each cluster. We also provided a spatiotemporal multi-characteristic plot that allows practitioners and researchers to simultaneously visualize up to five variables of interest on a map; the model used included time and space variables. Such a tool can be useful when studying a dataset for the first time and in evaluating the validity of modeling assumptions.

Our third proposed contribution is based on the development and use of a “p--value heat map” that assists policy-makers, researchers and the general public to succinctly see the potential impact of public policy on the reduction of nonfatal and fatal crashes. This tool helps policy-makers who may not have a strong statistical background to understand statistical outputs of policy-analysis models.

FIGURE 5
P-VALUE MAP: FATAL INJURY

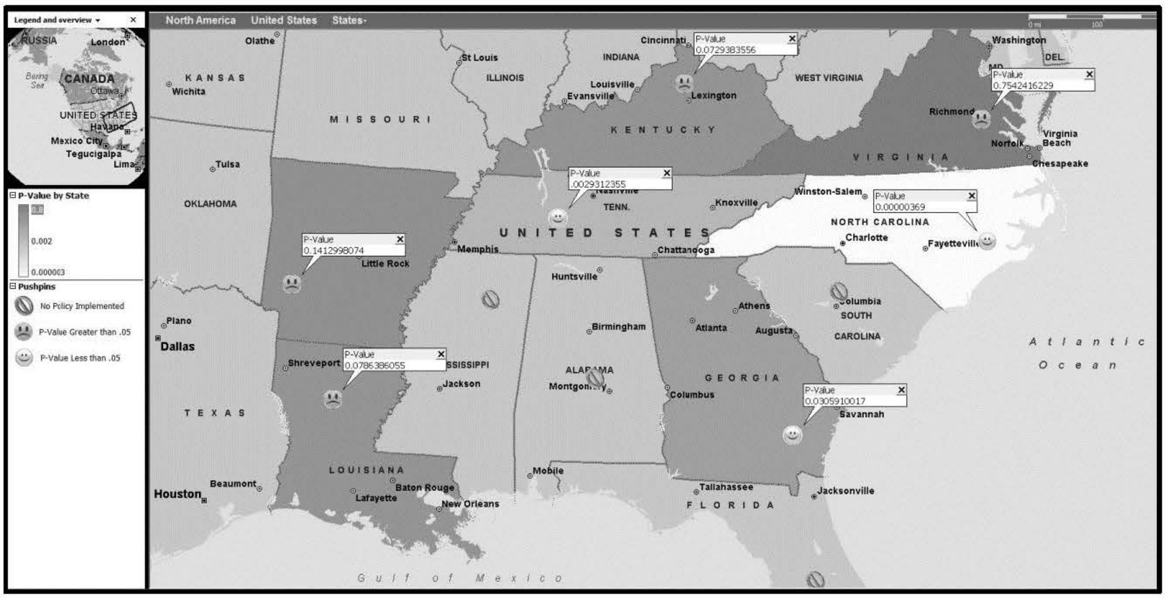
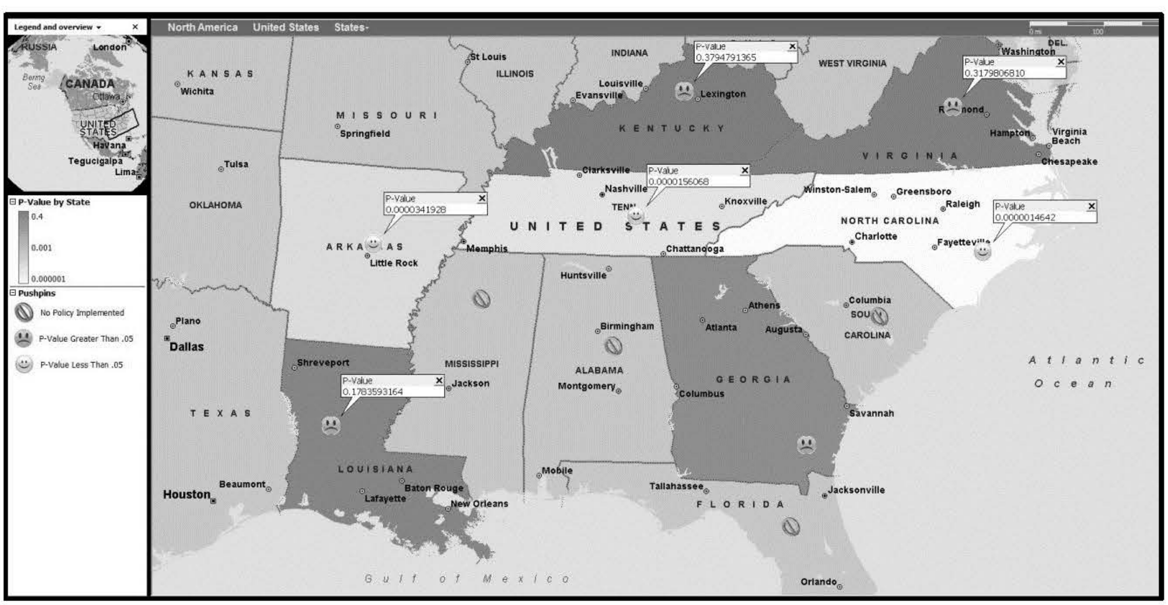


FIGURE 6
P-VALUE MAP: NON-FATAL INJURY



The results from this paper attempted to demonstrate the power of visualization and how they can assist with both theory development and explaining the results of statistical models. There remains significant work to be done in this area, including integrating visualization methods with traffic databases for real-time visualization of emerging trends, better understanding of the limitations of these approaches, and ensuring that these tools can be generalized to multiple application domains.

Supplemental Material - The clustering tool is located at: https://www.dropbox.com/s/lyq1tqhaquipjq3/CA2005.xlsm. The other tools are in the process of being revised for public use. Upon publication, we will provide the most up-to-date version at the corresponding author's website.

Acknowledgement - This work was supported by the NIOSH Deep South Center for Occupational Safety and Ergonomics, under grant number: T42OH008436. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of CDC – NIOSH.

REFERENCES

Aarts, Letty, & van Schagen, Ingrid (2006), "Driving Speed and the Risk of Road Crashes: A Review," *Accident Analysis & Prevention, 38*(2): 215-224. doi: http://dx.doi.org/10.1016/j.aap.2005.07.004

Al-Ghamdi, Ali S. (2002), "Using Logistic Regression to Estimate the Influence of Accident Factors on Accident Severity," *Accident Analysis & Prevention, 34*(6): 729-741. doi: http://dx.doi.org/10.1016/S0001-4575(01)00073-2

Alabama Department of Transportation (2011), *Directional Monthly Volume Report*, Retrieved 8/15/2013, from http://aldotgis.dot.state.al.us/atd/default.aspx

Anderson, D Mark, & Rees, Daniel I. (2011), "Medical Marijuana Laws, Traffic Fatalities, and Alcohol Consumption," Discussion Paper series, Forschungsinstitut zur Zukunft der Arbeit.

Black, WR. (1991), "Highway Accidents: A Spatial and Temporal Analysis," *Transportation Research Record,* (1318).

Carpenter, Christopher S, & Stehr, Mark (2008), The Effects of Mandatory Seatbelt Laws on Seatbelt Use, Motor Vehicle Fatalities, and Crash-related Injuries Among Youths," *Journal of Health Economics, 27*(3): 642-662.

De Oliveira, Maria Cristina Ferreira, & Levkowitz, Haim (2003), From Visual Data Exploration to Visual Data Mining: A Survey," *Visualization and Computer Graphics, IEEE Transactions,* 378-394.

Erdogan, Saffet, Yilmaz, Ibrahim, Baybura, Tamer, & Gullu, Mevlut (2008), "Geographical Information Systems Aided Traffic Accident Analysis System Case Study: City of Afyonkarahisar," *Accident Analysis & Prevention, 40*(1): 174-181.

Flaherty, Joe (1993), "Cluster Analysis of Arizona Automatic Traffic Recorder Data," *Transportation Research Record,* (1410).

Governors Highway Safety Association (2013), *Highway Safety law Charts*, Retrieved 09/15/13, from http://www.ghsa.org/html/stateinfo/laws/

Gregoriades, Andreas, Mouskos, Kyriacos, Parker, Neville, Hadjilambrou, Ismini, et al. (2011), *An Intelligent System to Enhance Traffic Safety Analysis,* Paper presented at the PESARO 2011, The First International Conference on Performance, Safety and Robustness in Complex Systems and Applications.

Greitzer, Frank L, Noonan, Christine F, & Franklin, Lyndsey R. (2011), *Cognitive Foundations for Visual Analytics*. (PNNL-20207). Pacific Northwest National Laboratory Retrieved from http://www.pnl.gov/main/publications/external/technical_reports/PNNL-20207.pdf.

Han, Jiawei, & Kamber, Micheline (2011), *Data Mining : Concepts and Techniques* (3rd ed.). Burlington, MA: Elsevier.

Ivan, John N. (2004), New Approach for Including Traffic Volumes in Crash Rate Analysis and Forecasting," *Transportation Research Record: Journal of the Transportation Research Board, 1897*(1): 134-141.

Jacobs, G., Aeron-Thomas, A., & Astrop, A. (2000), *Estimating Global Road Fatalities*, Crowthorne, Berkshire: Transportation Research Laboratory, Department for International Development.

Jacobson, Sheldon H., King, Douglas M., Ryan, Kevin C., & Robbins, Matthew J. (2012), Assessing the Long Term Benefit of Banning the Use of Hand-held Wireless Devices While Sriving," *Transportation Research Part A: Policy and Practice, 46*(10): 1586-1593. doi: http://dx.doi.org/10.1016/j.tra.2012.08.007

Keim, Daniel A, Müller, Wolfgang, & Schumann, Heidrun (2002, Sep. 2-6, 2002),*Visual Data Mining. State of the Art Report*. Paper presented at the Eurographics' 2002, Saarbruecken, Germany.

Kent, John L, Coulter, Ronald L, & Coulter, Mary. (2011), "Driver Safety and Motor Carrier Profitability: Identifying and Understanding Drivers in the Fleet," *Journal of Transportation Management, 22*(1):7.

Laessig, Ronald H, & Waterworth, Kathy J. (1970), "Involvement of Alcohol in Fatalities of Wisconsin Drivers," *Public health reports, 85*(6): 535.

Liang, Lim Yu, Mo'soem, DM, & Hua, Law Teik. (2005), "Traffic Accident Application Using Geographic Information System," *Journal of the Eastern Asia Society for Transportation Studies, 6*: 3574-3589.

Lozano, Rafael, Naghavi, Mohsen, Foreman, Kyle, Lim, Stephen, et al. (2012), "Global and Regional Mortality from 235 Causes of Death for 20 Age Groups in 1990 and 2010: A Systematic Analysis for the Global Burden of Disease Study 2010, *The Lancet, 380*(9859): 2095-2128. doi: http://dx.doi.org/10.1016/S0140-6736(12)61728-0

Massie, Dawn L., Campbell, Kenneth L., & Williams, Allan F. (1995), "Traffic Accident Involvement Rates by Driver Age and Gender, *Accident Analysis & Prevention, 27*(1): 73-87. doi: http://dx.doi.org/10.1016/0001-4575(94)00050-V

National Highway Safety Adminstration, (2012), *Fatality Analysis Reporting System (FARS) Encyclopedia*, *NCSA Data Resource Website.* Retrieved 2013/8/15, from http://www-fars.nhtsa.dot.gov/Main/index.aspx

Nikolaev, Alexander G., Robbins, Matthew J., & Jacobson, Sheldon H. (2010), "Evaluating the Impact of Legislation Prohibiting Hand-held Cell Phone Use While Driving," *Transportation Research Part A: Policy and Practice, 44*(3): 182-193. doi: http://dx.doi.org/10.1016/j.tra.2010.01.006

Rajaraman, Anand, Leskovec, Jure, & Ullman, Jeffrey D. (2012), *Mining of Massive Datasets* Retrieved from http://i.stanford.edu/~ullman/mmds/book.pdf

Roddick, John F, & Spiliopoulou, Myra (2002), "A Survey of Temporal Knowledge Discovery Paradigms and Methods," *Knowledge and Data Engineering, IEEE Transactions on, 14*(4): 750-767.

Sampaio, Breno (2012), "Identifying Peer States for Transportation Policy Analysis with an Application to New York's Handheld Cell Phone Ban," *Transportmetrica,* 1-14. doi: 10.1080/18128602.2012.688073

Shankar, Venkataraman, Mannering, Fred, & Barfield, Woodrow (1995), "Effect of Roadway Geometrics and Environmental Factors on Rural Freeway Accident Frequencies," *Accident Analysis & Prevention, 27*(3): 371-389. doi: http://dx.doi.org/10.1016/0001-4575(94)00078-Z

Shibata, Akira, & Fukuda, Katsuhiro (1994), "Risk Factors of Fatality in Motor Vehicle Traffic Accidents," *Accident Analysis & Prevention, 26*(3): 391-397. doi: http://dx.doi.org/10.1016/0001-4575(94)90013-2

Simoff, Simeon, Böhlen, Michael H, & Mazeika, Arturas (2008), *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics* (Vol. 4404): Springer.

Stamatiadis, Nikiforos, Agent, Kenneth R, & Bizakis, Apostolos (1997), "Guidelines for Left-turn Phasing Treatment," *Transportation Research Record: Journal of the Transportation Research Board, 1605*(1): 1-7.

Tufte, Edward R. (1983), *The Visual Display of Quantitative Information*, Cheshire, Conn. (Box 430, Cheshire 06410): Graphics Press.

Van Wijk, Jarke J, & Van Selow, Edward R. (1999), *Cluster and Calendar Based Visualization of Time Series Sata,* Paper presented at the Information Visualization, 1999 (Info Vis' 99) Proceedings, 1999 IEEE Symposium on.

Weijermars, Wendy, & van Berkum, Eric (2005). *Analyzing Highway Flow Patterns Using Cluster Analysis,* Paper presented at the Intelligent Transportation Systems, 2005, Proceedings. 2005 IEEE.

Wickham, H. (2013), "Graphical Criticism: Some Historical Notes," *Journal of Computational and Graphical Statistics, 22*(1): 38-44. doi: Doi 10.1080/10618600.2012.761140

World Health Organization (2008), *Global Burden of Disease*, Retrieved 11/12/2013, from http://www.who.int/healthinfo/global_burden_disease/estimates_regional/en/index.html

World Health Organization (2013), Global Status Report on Road Safety 2013: Supporting a Decade of Action. from http://www.who.int/violence_injury_ prevention/road_ safety _status/2013/en/index.html

Yau, Kelvin K. W. (2004), "Risk Factors Affecting the Severity of Single Vehicle Traffic Accidents in Hong Kong," *Accident Analysis & Prevention, 36*(3): 333-340. doi: http://dx.doi.org/10.1016/S0001-4575(03)00012-5

Yau, Nathan (2013), *Data Points : Visualization That Means Something*, Indianapolis, IN: John Wiley & Sons.

Yi, Ping, Xiao, Yingcai, Ciccolini, Anthony, Frommer, Greg, et al. (2001), "Rule-based Model for Traffic Accident Visualization and Analysis," *Journal of Computing in Civil Engineering, 15*(2): 129-136.

BIOGRAPHIES

Yao-Te Tsai is a Ph.D. student in the Department of Industrial and Systems Engineering at Auburn University. His research interests include transportation safety, visual analytics, and predictive modeling. Email: yzt0007@auburn.edu

Huw D. Smith is a graduate student Auburn University pursuing a Master of Industrial and Systems Engineering degree and a Master of Business Administration (dual program). He was the main author of a paper accepted in QREI: Quality and Engineering International focusing on SPC, Big Data, and Data Visualization (DOI 10.1002/qre.1706). His research interests include SPC, Big Data, Data Visualization, Simulation, Process Improvement, Lean, and Manufacturing. He currently works as a consultant implementing custom software systems that structure data-driven continuous improvement programs while completing his Masters program. Email: hds0003@auburn.edu

Stephen M. Swartz is an Associate Professor of Supply Chain Management at Auburn University. He earned his Ph.D. in Business Administration (major in Operations and Sourcing Management; minor in Logistics) from Michigan State University. His current research focus is in transportation safety with an emphasis in behavioral factors and decision making. His work has been published in the International Journal of Logistics Management, Transportation Journal, the Journal of Transportation Management, the Journal of the Operational Research Society, Mathematical and Computer Modelling, and others. Email: steve.swartz@auburn.edu

Fadel M. Megahed is an Assistant Professor of Industrial and Systems Engineering at Auburn University. He received his Ph.D. and M.S. in Industrial and Systems Engineering from Virginia Tech, and a B.S. in Mechanical Engineering from the American University in Cairo. Current research focuses on creating new tools to store, organize, analyze, model, and visualize large data sets associated with manufacturing, healthcare and service environments. His work has been funded by the Proctor and Gamble Fund of the Greater Cincinnati Foundation, Amazon Web Services, Windows Azure (Microsoft), the NIOSH Deep South Center for Occupational Safety and Ergonomics, and the National Science Foundation. Email: fmm0002@auburn.edu