Here's What (and Whom) to Avoid On the Roads

Desmond Cole, Teerth Patel, Yunbin Peng

April 15, 2018

Introduction

We analyze traffic fatality data provided by the National Highway Traffic Safety Administration (NHTSA) to assess various predictors of traffic fatalities and develop a limited profile of the circumstances associated with traffic fatalities.

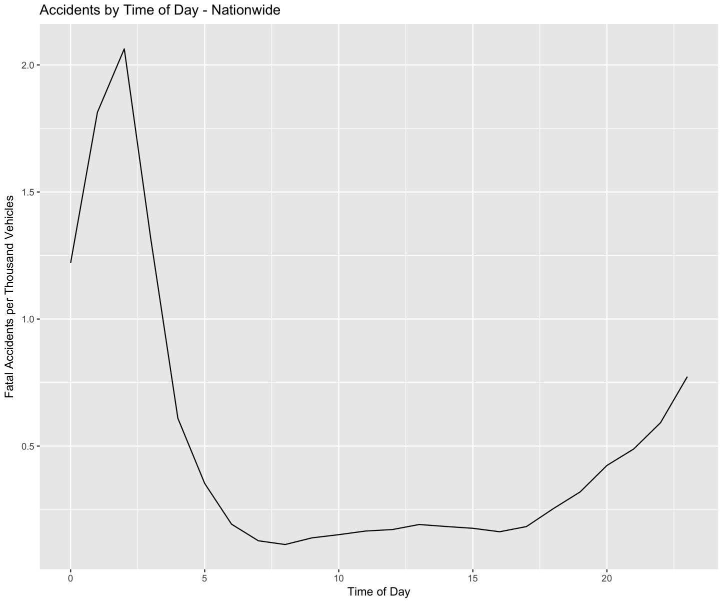
Exploratory Analysis and Visualization

Geographic Patterns

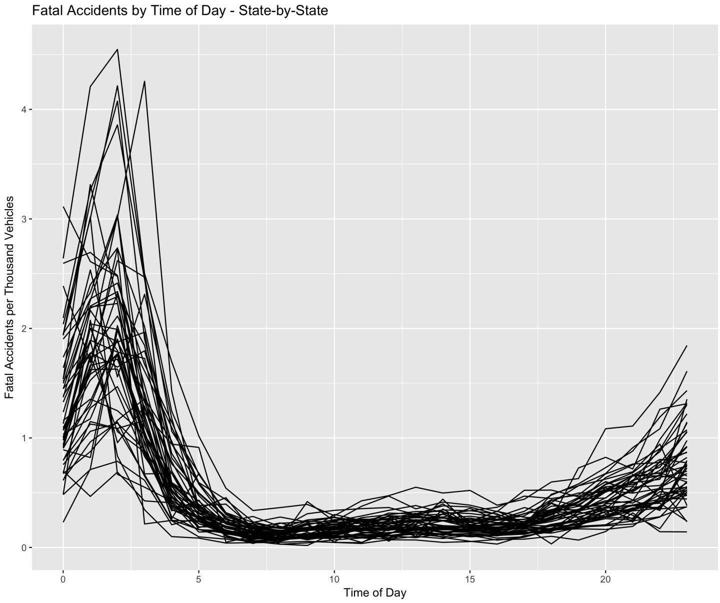
Time Trends

Daily Cycle

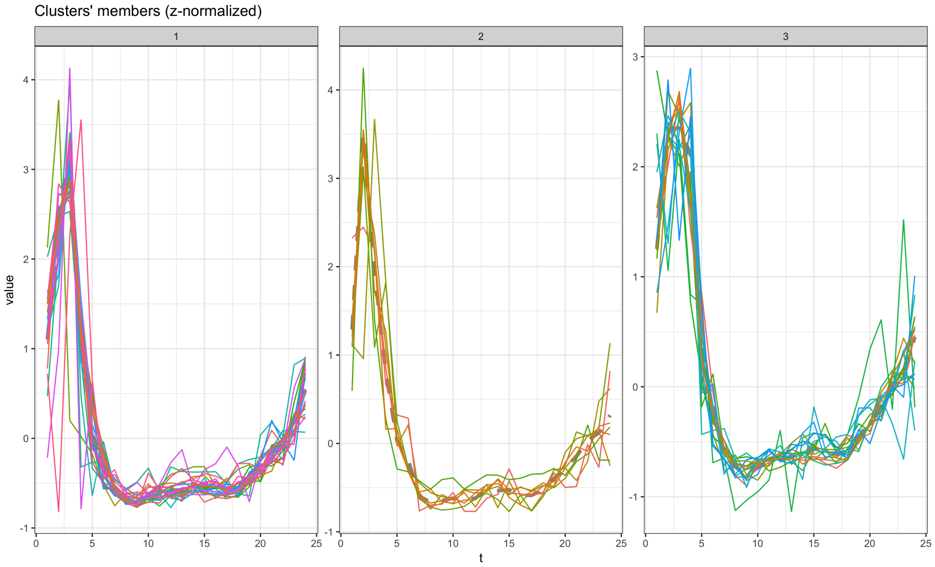
At the national and state level, the cycle of fatal accidents throughout the day is fairly consistent. The below plot shows aggregate fatal accidents per thousand vehicles per hour for the United States, between 2015 and 2016. To account for traffic variability, we weighted the rate of fatal incidents according to hour-by-hour traffic flow allocations estimated by Batterman, et al. (see references). These same flow allocations are used to weight state-by-state hourly trends.



At the state level, the daily fatal accident cycle is roughly similar. Estimating the overall traffic flow for each state using state-level vehicle registrations, combined with the flow allocations estimated by Batterman, et al. (2015), returns state-level cycles that closely mirror the overall national trend.



To tease out any potential state-level variation in daily cycles, we used shape-based time series clustering, available in R's dtwclust package. Roughly speaking, this method clusters different time series based on their similarity to different centroids, where a centroid in this case is defined as a typical time series. Further detail on relevant methods is available in work by Gravano and Paparrizos (2015). Testing several different numbers of clusters yielded little overall diversity in terms of performance. A plot of results obtained using 3 clusters is below. There is little indication of significant variation across clusters, suggesting that hourly fatal incident cycles vary little across the United States.



(Insert table of clustering performance)

Weekly Cycle

Risks of Traffic Fatalities

We used a mix of classifiers to assess factors relevant to incident fatalities, with a particular focus on driver behavior and vehicle manufacturer. In many cases, we had to grapple with significant class imbalance. For example, there are far more single-fatality incidents than multi-fatality incidents. As a result, optimal classification in the naive setting without any resampling will occasionally identify all incidents as single-fatality.

Driver Behavior

Drugs/Alcohol Distraction

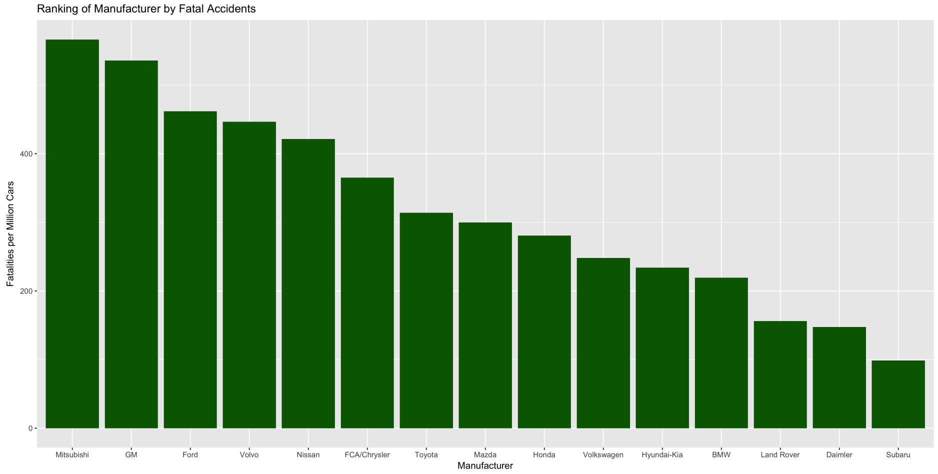
Automakers

This section explores in some detail different fatality rates for different car types (by automaker, vehicle body, etc.).[1] For this assessment, we focused on the 15 largest automakers that together produce more than 95% of the cars and light trucks on American roads. In addition, we subsetted the data to focus specifically on smaller vehicles, excluding commercial vehicles, semi-trucks, etc.

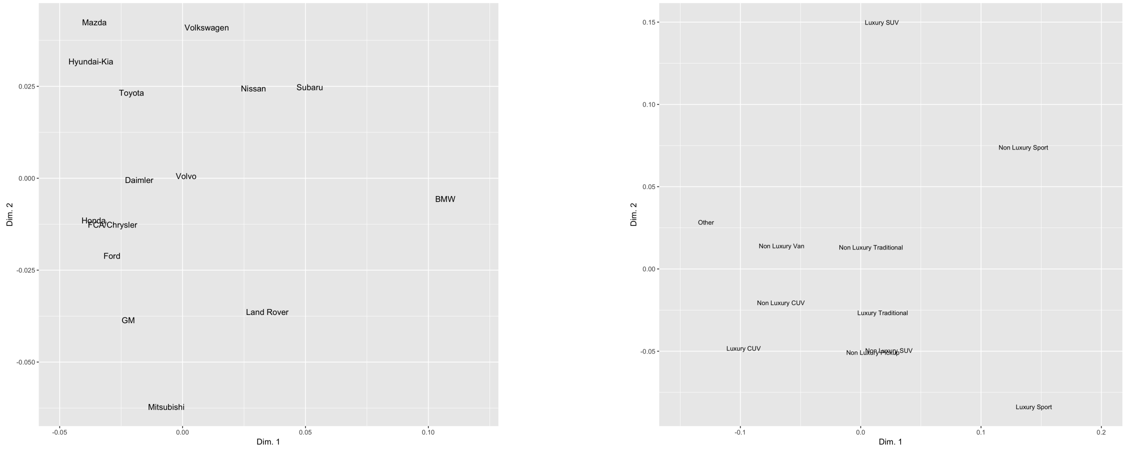
The plot below shows a basic ranking of car manufacturers according to the number of fatalities per million vehicles. This is scaled and subsetted according to an approximation of the number and

[1] Note that the numbers in this assessment focus on car brands *involved* in fatal incidents. Thus, they do not imply specifically that the vehicle types considered here actually caused death(s), or that the driver(s) of the vehicle(s) themselves died.

types of vehicles each manufacturer has on the road, but due to a lack of thorough data, it is likely that not all variation simply attributable to manufacturer size and vehicle class has been accounted for. With that said, this plot does suggest that there is meaningful variation across manufacturers in terms of their safety record, with makers such as Subaru and luxury makers such as BMW and Land Rover having significantly less involvement in fatal incidents than carmakers like GM, Ford, and Volvo.



The plot above, although suggestive of meaningful difference across manufacturers, fails to fully account for the various contextual differences across vehicle types which may be unobservable. To further explore the potential relevance of car manufacturer and car type to crash risks, we applied multidimensional scaling to assess visually the differences between vehicles in terms of various driver-specific crash-relevant pieces of information, such as use of restraints, drug/alcohol use, fires, and speeding, with the goal being to assess whether a certain vehicle type or manufacturer predicts the behavior of the vehicle's driver.



(a) Manufacturer MDS

(b) Vehicle Type MDS

Figure 1: Multidimensional Scaling for Vehicle Manufacturer and Vehicle Type

The plots don't indicate strong patterns of differentiation, but there are some noteworthy indicators.

In particular, these results suggest the behavior of luxury sport/SUV vehicles to be somewhat distinct, as noted by the outlying values for BMW/Land Rover in figure (a), and for Luxury SUV and Sport vehicles in figure (b).

we consider the relationships between car model and various fatality-related predictors, to understand if a given manufacturers' products tend to be associated with high-risk behaviors or other crash factors. The table below shows the results from various multi-class classifications of car manufacturer ran using a set of crash-relevant predictors .... The results do not suggest any substantial mechanism(s) by which a car's make determines the risk of it being involved in a fatal incident.

(Insert table of classifications and results)

Multi-fatality Incidents

Another specific category of interest is multi-fatality incidents. Every observation in the available data describes an instance of a given person dying, but what about multiple deaths? To explore this issue, we generated a binary 1-0 indicator for whether an incident involved more than 1 fatality, and tested the classification performance of a series of different models, the results of which are shown below. The resampling procedure used here is the SMOTE algorithm (for "Synthetic Minority Over-sampling TEchnique"), developed by Chawla, et al. (2002). SMOTE uses a combination of over-sampling of the minority class and under-sampling of the majority to improve class balance.

The below table shows relative test error rates for classification using the logistic regression, support vector machine, and adaptive boosting methods.

(Insert table of test error rates)

Given that a key interest in this paper is inference about relevant variables, we also examine the importance of certain features in predicting an accident with multiple fatalities.

(Insert table of variable importance, using adaboost and (if relevant) SVM/Logit results.)

Conclusion

References

Batterman, Stuart, Richard Cook, and Thomas Justin. "Temporal variation of traffic on highways and the development of accurate temporal allocation factors for air pollution analyses." *Atmos Environ.* Apr. 2015.

Chawla, Nitesh, Kevin Bowyer, Lawrence Hall, and W. Philip Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research.* 2002.

Choi, Eun-Ha. ”Crash Factors in Intersection-Related Crashes: An On-Scene Perspective.” *U.S. Department of Transportation - National Highway Traffic Safety Administration.* Sept. 2010.

Gravano, Luis and John Paparrizos. ”k-Shape: Efficient and Accurate Clustering of Time Series.” *SIGMOD.* June 2015.

Zador, P.L, S.A. Krawchuk, and R.B. Voas. ”Relative Risk of Fatal Crash Involvement by BAC, Age, and Gender.” *U.S. Department of Transportation - National Highway Traffic Safety Administration.* Apr. 2000.