What (and Whom) to Avoid On the Roads

Desmond Cole -
Teerth Patel -
Yunbin Peng

April 15, 2018

Introduction

Executive Summary

In the United States, approximately 35-40,000 people die in vehicle-related incidents every year. The fatality rate declined significantly in the second half of the 20th century, and continued to decline in the 21st, however the decline has plateaued in recent years, to a rate of just over 1 death per 100 million vehicle miles traveled (VMT). Understanding and tackling the problem of crash fatalities has motivated considerable improvements and investments in vehicle design, road design, and our knowledge of human motor skills and reaction times.

With this analysis, we set out to survey the distribution of vehicle-related fatalities across geography and time, and explore a number of driver-specific characteristics to assess dominant contributors to the incidence and extent of vehicle-related deaths. We use this analysis to develop a limited profile of the circumstances associated with traffic fatalities, and of things to avoid if you're considering taking a drive.

Data

The data used for this analysis are gathered by the National Highway Traffic Safety Administration (NHTSA), a branch of the U.S. Department of Transportation. Traffic fatality data is contained within NHTSA's Fatality Analysis Reporting System (FARS), a nationwide census of various information regarding fatalities occurring in motor vehicle crashes. FARS contains a rich set of different features, with information on crash circumstances, areas of damage, police reports, emergency response times, and vehicle characteristics. We focus specifically on data from 2015 and 2016, and subset a choice set of predictors of interest, paying particular attention to indicators of driver behavior.

Exploratory Analysis and Visualization

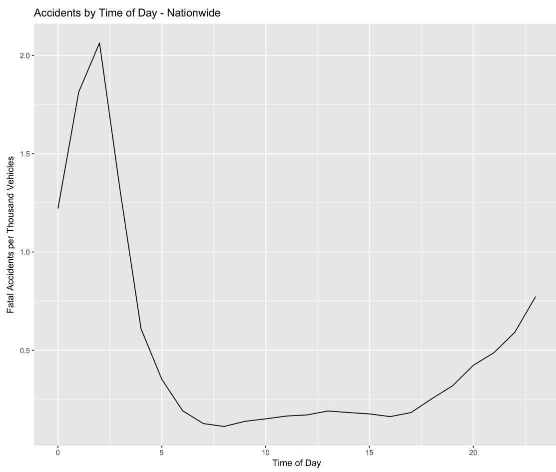
Geographic Patterns

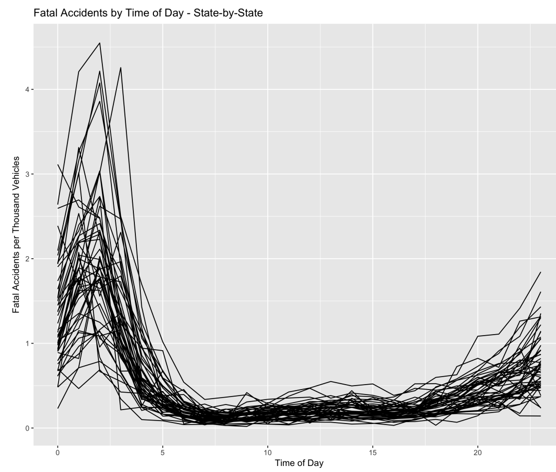
Time Trends

Daily Cycle

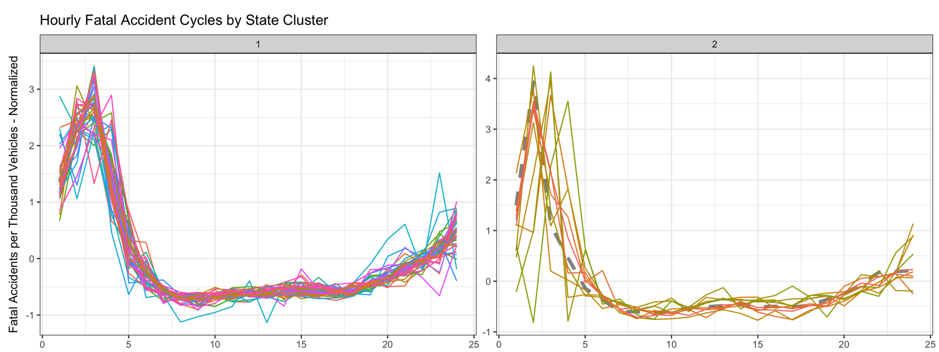
At the national and state level, the cycle of fatal accidents throughout the day is fairly consistent. The below plot shows aggregate fatal accidents per thousand vehicles per hour for the United States, between 2015 and 2016. To account for traffic variability, we weighted the rate of fatal incidents 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 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 some typical time trend. Further detail on relevant methods is available in work by Gravano and Paparrizos (2015). Testing different numbers of clusters yielded little overall diversity in terms of performance. A plot of results obtained using 2 clusters is below.[1] There is little indication of significant variation across clusters, and the clustering results (and performance statistics, shown in the below table) are highly sensitive to initial randomization.



[1]Across multiple iterations, a 2-cluster separation tends to yield the best performance. For reference in the Validity Index table, optimal clustering will minimize the Silhouette ("Sil"), Davies-Bouldin ("DB" and "DBstar") indices, and maximize the Dunn index ("D"), Calinski-Harabasz index ("CH"), and Score Function ("SF"). 2 clusters were chosen to be optimal on this basis.

	NumClusters	Sil	SF	CH	DB	DBstar	D	COP
1	2	0.508	0.602	23.664	0.899	0.899	0.072	0.181
2	3	0.267	0.595	17.453	1.461	2.080	0.013	0.135
3	4	0.167	0.589	12.171	1.929	2.510	0.019	0.121
4	5	0.133	0.572	9.163	1.345	2.451	0.017	0.153
5	6	0.391	0.576	15.172	0.683	1.219	0.056	0.084

Table 1: Time Series Cluster Validity Indices (CVIs)

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

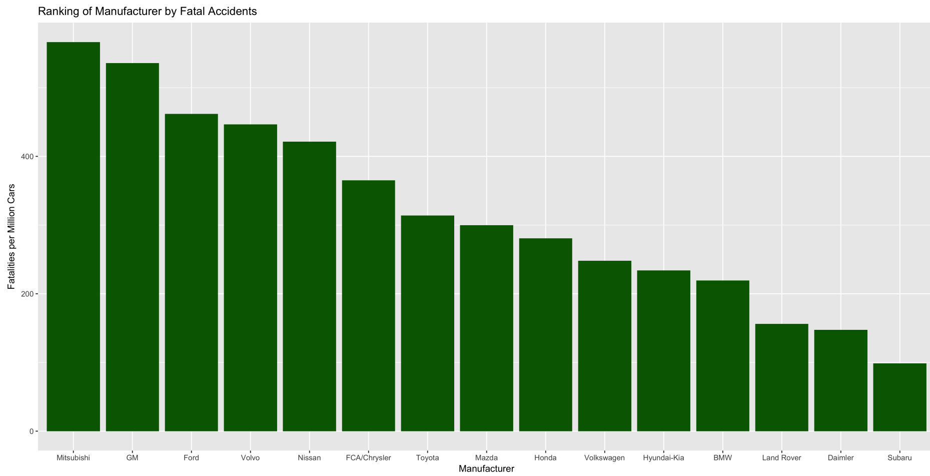
Drugs/Alcohol Distraction

Automakers

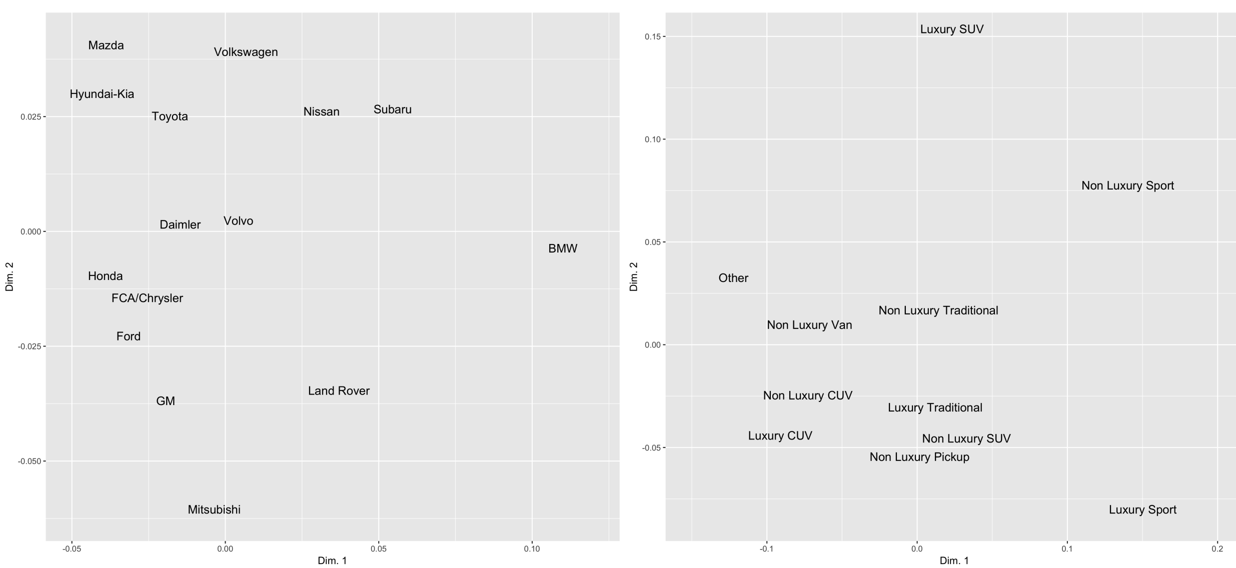
This section explores in some detail different fatality rates for different car types (by automaker, vehicle body, etc.).[2] For this assessment, we focused on the 15 largest automakers that together produce more than 95% of the cars and light trucks on American roads. In addition, we subsetted the data 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 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.

[2] 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.



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 2: 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 drivers to be somewhat distinct, as noted by the outlying values for the BMW/Land Rover[3] in figure (a), and for Luxury SUV and Sport vehicles in figure (b).

[3]Note that this analysis considers all Jaguar models to be part of Land Rover

To numerically explore this issue more fully, we draw on results of a logistic regression model (discussed more fully in the next section) to consider how a vehicle's manufacturer affects the likelihood of a multi-fatality collision.[4]

	Odds.Ratios	Automaker
1	0.85	Volkswagen
2	0.70	Honda
3	0.62	Mazda
4	0.50	Ford
5	0.41	Toyota
6	0.39	Volvo
7	0.39	FCA/Chrysler
8	0.33	GM
9	0.27	Hyundai-Kia
10	0.21	Mitsubishi
11	0.10	Daimler
12	0.06	Nissan
13	-0.07	Subaru
14	-0.34	Land Rover

Table 2: Automaker Odds Ratios for Predicting Multi-Fatality Accident (Relative to BMW)

Multi-fatality Incidents

Here we discuss multi-fatality incidents more thoroughly. 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 logistic regression, support vector machine, and adaptive boosting methods.

	Logistic Regression	Supp. Vec. Machine	AdaBoost
Accuracy	0.75	0.80	0.79

Table 3: Test Set Prediction Accuracy

We also examine the importance of certain features in predicting an accident with multiple fatalities. The below output illustrates variable importance rankings for the SVM and AdaBoost classifiers. The models don't completely agree on relative variable importance, but there are some

[4]Note again that the data here are drawn solely from accidents involving at least one fatality. Therefore, classification methods such as this are predicting the severity of an accident, *given that one has already occurred.*

key patterns. For example, key for both models is the fact that the hour of the crash is a dominant predictor of the likelihood a fatal accident resulted in multiple deaths. In addition, a driver's record, potential drug use, and driving speed are especially significant for predicting multi-fatality accidents. Interestingly, whether or not a driver was distracted (using a cellphone, for example), and whether the weather was clear or not, appear to be relatively less significant determinants of the likelihood that a fatal accident involved multiple deaths. Further assessment of this question would consider the affect of these less important predictors on non-fatal accidents, as well as fatal ones.

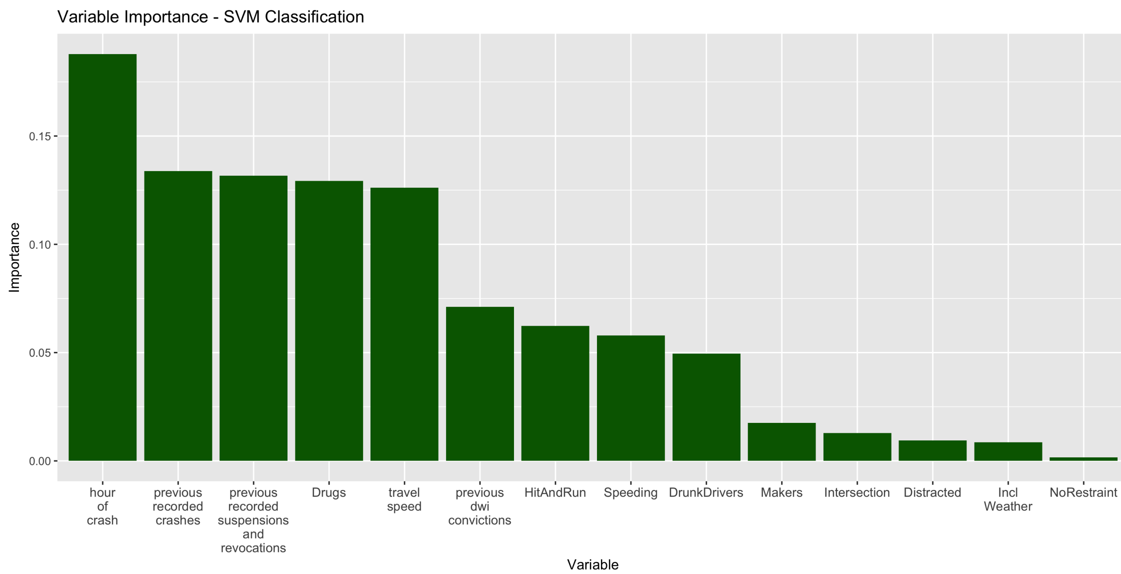


Figure 3: Variable Importance - Support Vector Machine

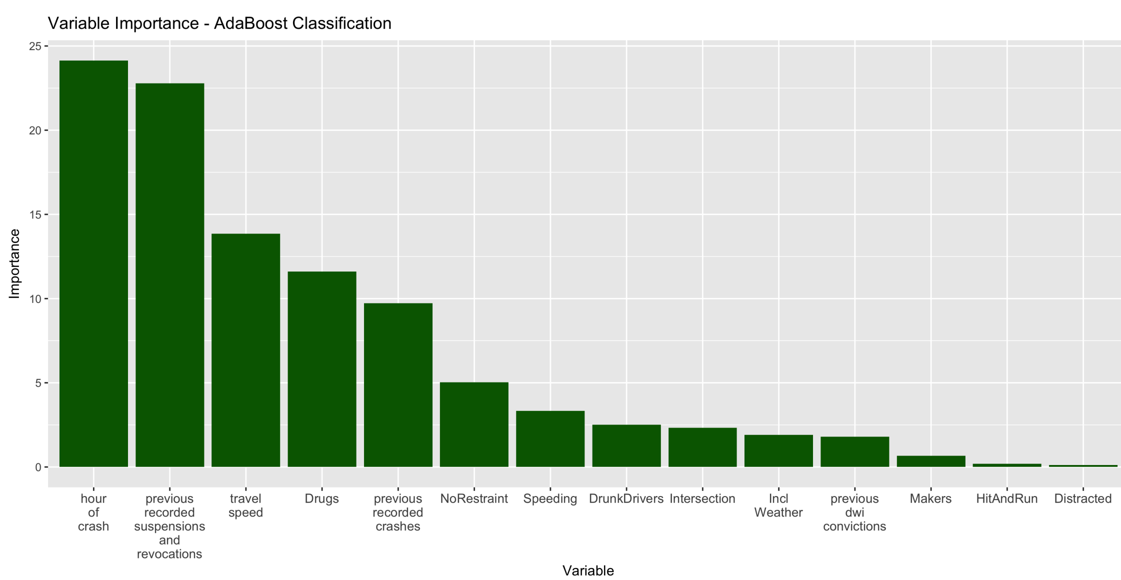


Figure 4: Variable Importance - AdaBoost Algorithm

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.

"National Motor Vehicle Crash Causation Survey: Report to Congress." *U.S. Department of Transportation - National Highway Traffic Safety Administration.* July 2008.