Analysis of Trends and Causes of Fatal Motor Vehicle Incidents

Andrew Cockerill
Brett Keim
Catherine Rivier

October 21, 2014

Abstract

With approximately 35,000 traffic deaths in 2013, driving remains the most risky activity that Americans perform on a daily basis [**CSMonitor**]. As the number of licensed drivers continues to increase, it is paramount that citizens and emergency personnel understand the patterns of how these incidents occur. This project is an inquiry of fatal vehicle incident data (FARS) compiled by the National Highway Traffic Safety Administration for the year 2010. We perform analyses of a variety of factors related to each case in order to provide insight into the chief circumstances likely to cause a serious accident. Though this database provides dense information about every incident, we lack the same information about non-incident occurrences - perhaps making it difficult to make strong conjectures on the population as a whole. To remedy this, we plan to obtain census and other full population data related to these factors for comparison. By comparing this FARS data to the population as a whole, we can provide better insight into the problem as opposed to methods that analyze crash data alone. This will allow future researchers to explore the efficacy of methods to curb the driving risks discussed in this project.

1 Introduction

The topic of motor vehicle safety continues to be an issue of great concern in the United States today. Indeed, according to the CDC, motor vehicle incidents are among the leading causes of death for persons under age 44. With this knowledge in hand, it is paramount that drivers understand the circumstances that can cause these serious accidents. Likewise, it is vital that emergency personnel and legislators understand the current trends and causes of these vehicular events. This report seeks to illuminate these factors through an analysis of motor vehicle incident data compiled in the Fatality Analysis Reporting System (FARS) [**FARS**].

It is significant to note that although vehicle safety remains a serious issue, the matter has seen considerable improvement in recent years according to the Office of Highway Policy Information [**OHPI**]. As indicated by Figure 1, the number of fatal vehicular incidents has seen a marked decline since 2005 in spite of an increase in licensed drivers. However, 2012 saw an end to this trend. It therefore behooves the investigation into what risk factors should be addressed in order to return to this desired downward pattern.

1.1 Document reproducibility

This report is constructed is prepared using the **R** [**R-base**] package `knitr` [**R-knitr**]. This project may be imported into the `RStudio` environment and compiled by researchers wishing to reproduce this work for future datasets.

2 About the data

As described earlier, the data primarily used in this analysis is derived from the Fatality Analysis Reporting System (FARS), a census of fatal traffic incidents compiled yearly by the National Highway Traffic Safety Administration. Each year's census consists of a myriad of recorded data pertaining to each incident. This

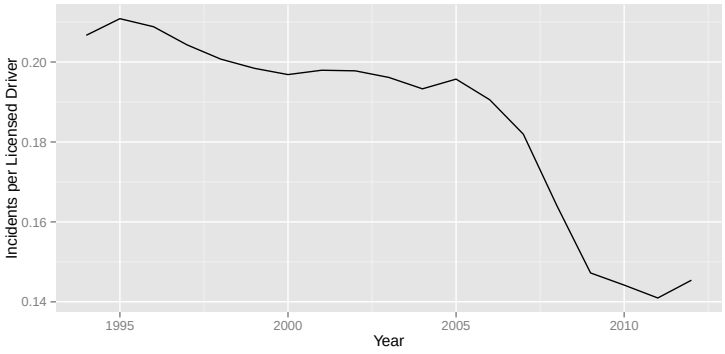

Figure 1: Yearly totals of fatal motor vehicle incidents from 1994-2012.

includes personal driver information, dates and times, locations, behavioral factors, and climate factors. In the data's raw form, these items are scattered and can also be difficult to understand. This difficulty stems from the fact that many of these factors are represented by identifying numbers, rather than as textual information. This was remedied by means of a data cleaning process as described in the following subsection.

2.1 Preparing data

Data preparation was a major piece of this project. It involved both the data itself, and the code definition file.

The data itself was in an easy format to work with, easy to download in its various formats. There are three different data forms: by individual accident, by each vehicle involved, and by each person involved. All of these can be tied together by a unique accident ID, and each vehicle and person received its own unique addition onto this ID. Together this will allow us to join on the additional files of data per accident. There are additional details of violations issued, damage done, and vehicles involved - all of which get into more specific detail than offered in the overall accident, vehicle, or person files. In addition, there are event files, which sequence through each event per accident. All of these supplemental files have been brought it to perhaps bring more detail as we need it going forward in this analysis. They will all bring in their own challenges, given they each may have multiple records per accident/vehicle/person. It should also be noted that there are files dating back to 1975 - though currently only 2010 is being worked with to develop our techniques and methodology before applying to more years.

The biggest challenge came with the code definition file - e.g. the file that would allow us to decode each numeric value in the fields in the data file. Though the coding of each variable is documented thoroughly by the NHTSA, that documentation exists in PDF only. The principle task that existed was to extract those code tables from that PDF and put them into lookup tables.

Because this process involved so much text mining and a large amount of prepatory work, we will document it here.

The data dictionary for this census has been published in the NHTSA's FARS user manual [**USERS**]. The typical variable may have a page as seen on page 59. A proper format for the lookup would entail a table with the fields: Variable Number, Variable Name, Variable Definition, Year (the code was applicable), Attribute Code, and Attribute Definition. To turn the PDF text (400 pages) into this involved text mining to extract those elements. This is the summary of the process:

1. Each line of the text (variable Text) was given a unique, ascending numerical index for reference (new variable RowID).

2. Variable number RowIDs were flagged as the [RowID when the Text began with "Definition:"]-1.

3. The RowID range of each variable number was then calculated by taking its RowID, to one less than the RowID of the next variable number.

   By getting the RowID range for each variable, we could then attribute everything extracted after this point to the proper variable.

4. Those Variable number rows were then reduced to split at the first space, breaking them into two pieces: the Variable Number, and the Variable Description.

5. Variable name rows were flagged as when the Text began with "SAS Name:". (Note it was possible for there to be more than one name listed after that start.)

6. Those Variable name rows were then reduced to remove the words "SAS Name: ", and split by a comma, if it existed. Now those rows had at least one Variable Name.

   Now we had Variable Number, Variable Name, and Variable Text. Extracting the Attributes was more challenging:

7. Extracting the attribute headers came first:

   - These were all Text that began with a year, or the term "Later".
   - These proved complex, as they covered multiple lines and multiple columns. Additionally, there could be multiple sets of headers per variable. The solution was again using the RowID and calculating the range as with Variable number - with the end of the range being the next Variable Number's RowID -1.
   - If that range was only two lines, the lines were merged together. Then the existing years were ordered (with "Later" being last). The presence of a hyphen meant the header indicated a range. And this created the full text of each header.
   - The headers were then counted and noted per each range for use with the attributes below.

   Now we had attribute headers, their RowID range, and the count of headers (e.g. attribute columns) in that range.

8. Next came extracting the attribute codes and definitions.

   - The attribute codes and definitions were all space-separated.
   - If the range header count was one, the first space was used as the split. If two, the first and second. If three, the first, second, and third. And so forth, until all possibilities were ccovered. This left us with all codes and definitions per variable.

   Finally, cleaning up the results:

9. Each Variable Number and its corresponding Name and Description was copied to each row of the Variable Number's RowID range, as three new fields (Variable Number and Variable Name and Variable Description).

10. Each Attribute Header was copied to each row of the header RowID range, as new field Year.

11. Every row except those within the Attribute Header range (which should be just the Attribute Code and Definition rows) can then be deleted.

    At this point, we had a file with a Variable Number, Variable Name, Variable Description, Attribute Header, Attribute Header Count, Attribute Code, and Attribute Definition. The only thing missing was applying the correct Attribute Header to the right Attribute Code.

12. The rows with an Attribute Header Count of 2 were duplicated, 3 were triplicated, etc. If those rows were duplicated, the original set a new Attribute Header column to use the left-most Attribute Header (with space as the delimiter); the rest were the second-from-the-left header, and so forth. This was done in a single step to avoid any issues. And now each row had a year range. This needed to be translated so there was a row for each year in the range.

13. The start date and end date (which might be "Only" or "Later") were extracted by using a space or dash as the splitter.

14. Going year by year, if that year was between the start and end date (replacing "Later" with the year 2015 and "Only" with the start date), the row would be duplicated and a new variable Attribute Year was given that year as its value.

This final step gave us exactly what we needed. We had a lookup table with the fields Variable Number, Variable Name, Variable Definition, Attribute Year, Attribute Code, and Attribute Definition. Now the data table can be joined to this lookup table per each variable, code, and year. And, the codes can be replaced with the Attribute Definition in the table. The resulting dataset was a collection of tables formed as a Microsoft Access Database file, which can be accessed in **R** by means of the RODBC package [**RODBC**].

2.2 Shortcomings

It is worth noting that although the FARS data is indeed extensive, it lacks information on the yearly population of licensed drivers. It is intuitively evident that this is an important piece of information, as making inferences into proportional risks requires a total population to compare the raw data to. To remedy this, further data was obtained from the Office of Highway Policy Information (OHPI) on the annual total number of licensed drivers.

3 Methods

The census in question provides a breakdown of factors involved in each type of vehicular incident for a given year. As such, it is reasonable to approach this data by means of an analysis of the frequency of crashes given certain parameters. To accomplish this, we present an assortment of data product figures as described in the following subsection. This will allow the inference into how likely crashes are to occur given certain circumstances, and further allows the faceting of these circumstances among different demographics.

3.1 Data product

The types of data prodcuts and figures presented in this report are indicative of the types of factors analyzed in each case. This is done in an effort to provide the reader with a clear understanding of this project's findings and conclusions.

An important element of this investigation is that of the date and time that serious accidents occur. As such, we present these in the form of time series plots, which have been specified for inferences according to month, day, and hour. We are then able to visualize when accidents are more or less likely to occur, as well as hypothesize why these discrepancies exist.

Geographic information plays a vital role in this inquiry. Individual states can differ in many aspects in the approach to vehicle safety. For instance, the maximum speed limits on rural and urban highways varies from state to state, as well as enforced penalties for violations. It is advisable then to consider the rates of incidents at the state level. This can be accomplished in multiple ways. First, we employ a mapping of the United States for visual representation of which states are more prone to fatal incidents. Secondly, we can provide box-plots for the various divisions of maximum speed limits to see if crash frequency differs between these groups.

In conjunction with knowing where and when serious motor events occur, it is also necessary to deduce the environmental and behavioral factors that are most likely to cause these incidents. To achieve this goal, we present frequency plots in the form of histograms to compare these data. In particular, we can observe the difference in crash frequency caused by elements such as weather events, as well as the effect of different driver behaviors at the time of incident.

4 Results

An analysis of the data allows us to visualize elements of vehicle crashes in a variety of ways. Specifically, this inference has first yielded a summary of the frequency of incidents as they relate to time on an hourly, daily, and monthly basis. We also examine incidents on a state-to-state basis, obtaining the incident frequency to licensed driver ratio. Additionally, we compare this to other state factors including enforced speed limits and emergency response times for each state. Lastly, environmental and behvaioral factors are illustrated by analyzing available weather data as well as information on driver impairments such as alcohol and other distractions.

4.1 Overview

We begin with a brief summary of the fatal crash statistics for 2010 as shown in Table 1. This provides the statistical breakdown of fatal incidents for all 50 states and the District of Columbia.

	Value
Total	30296.00
Average	594.04
SD	587.41
Range1	24.00
Range2	2746.00

Table 1: Summary of fatal crash statistics for 2010. Note the relatively large spread of incident frequency.

We first observe that the year saw a total of 30296 fatal crashes. Notice the range of accidents is rather large. This illustrates that the number of incidents can vary greatly from state to state. Along with risk factors, we must acknowledge that this discrepancy can also be due in large part to differences in the driving population for each state. This provides the reasoning in why we elect to analyze crash to driver ratios in relevant subsequent sections.

4.2 Time of incident

We examine the frequency of incidents for given time frames. Figure 2 illustrates the incident frequency for each month of the year. Observe that more crashes occur in the summer and beginning of fall, particularly in the span from July to October. Conversely, it is interesting to note that crash rates drop noticeably in the winter season, especially from December to March. One conjecture concerning this relationship is that there may simply be fewer cars venturing onto the road in the colder seasons.

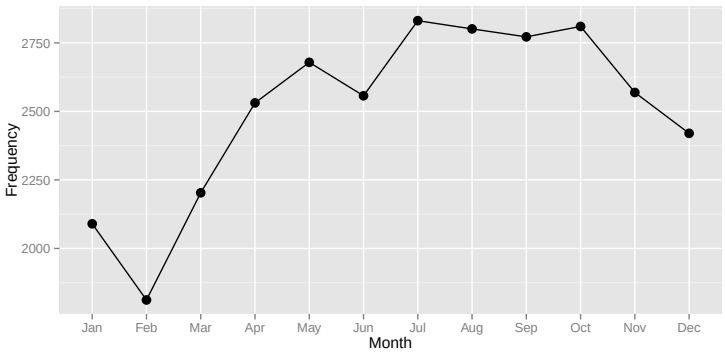

Figure 2: Fatal crash frequencies for each month of 2010. Note the rise in occurences during the early-to-late summer.

Narrowing the scope to a daily basis also yields noteworthy results. According to Figure 3, we observe a clear trend as the week progresses from Sunday to Saturday. In this case, have a marked decrease in incidents as we approach the middle of the week, then a great increase toward the weekend. Again, this may be due to fewer cars on the road in the middle of the week (during which time people are generally at work or school). By contrast, more vehicles may be on the road during the weekend, resulting in this higher number.

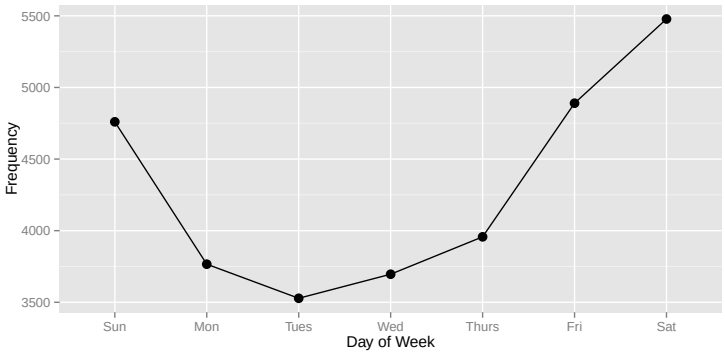

Figure 3: Fatal crash frequencies for days of the week. Observe the rise in weekend rates.

Further specifying the time frame to hours of the day, we construct Figure 4. This displays the total number of accidents compiled hourly, faceted by day of the week. Here, we see a common trend in which we see a small spike in the morning hours around 6-7 AM (likely due to rush hour traffic), which then follows with a steady increase into the evening. This gain begins to fall during the very early morning hours.

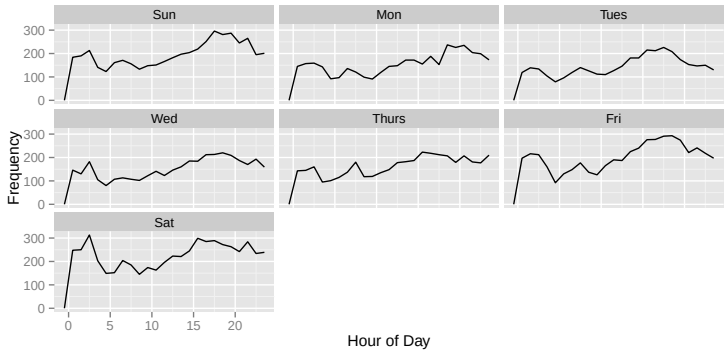

Figure 4: Fatal crash frequencies for each hour of each day of the week.

4.3 State incidents

To begin describing the FARS data at the state level, Figure 5 displays the standardized ratio of number of incidents to number of licensed drivers for each contiguous state. In the plot, blue represents a lower (better) ratio, while pink displays a higher (worse) ratio. Unexpectedly, we find that highly populated states with busy highways such as New York and California actually have a lower ratio. Conversely, more rural areas such as Wyoming and Mississippi have greater ones.

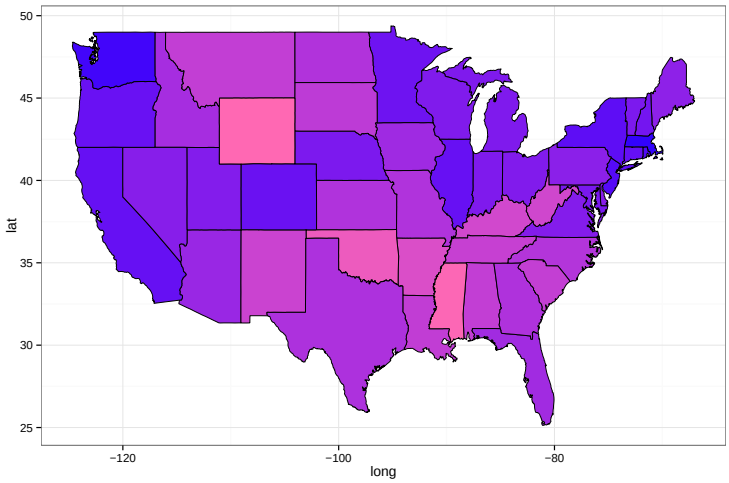

Figure 5: Ratio of incident count to licensed drivers per state. The spectrum from blue to pink indicates an increasing (Worse) ratio.

To discern possible reasons for state discrepancies, we first consider the maximum speed limits imposed on urban and rural highways in different states, based on the posted state speed limits provided by the Governors Highway Safety Association [**GHSA**]. This can be illustrated in two ways to gain valuable information. Figure 6 displays a mapping of the speed limits for the contiguous states, faceted by urban and rural highways. We see that in general, rural areas are more apt to allowing greater highway speeds.

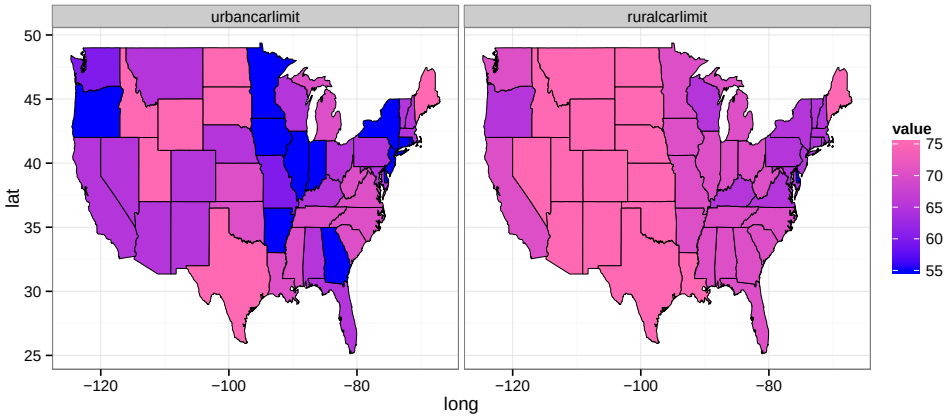

Figure 6: Speed limits for the contiguous states separated by urban and rural highways.

Given this plot, it is fitting to investigate the influence of speed limit on crash frequency further. Figure 7 displays a box plot for the ratio of incidents to licensed drivers for each state, grouped according to maximum speed limit and faceted for rural and urban highways. We find that there is little difference in ratio for lower speed limits from 55-65 miles per hour. However, the ratio sees a noticeable increase for higher limits of 70-75 miles per hour. Thus, there may be some correlation between maximum speed limit and incident rates.

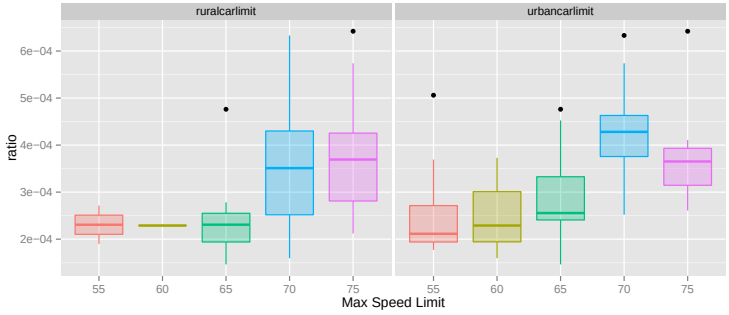

Figure 7: Box plot of state speed limits compard with the incident ratio for that state.

Another necessary consideration at the state-level is the time it takes for rescue workers to reach the scene of an accident This is due to the fact that a serious crash can become fatal if there is a significant delay in the delivery of medical care. Figure fig:map-resp illustrates this concern. In this case, the scale from blue to pink indicates an increasing (worse) response time. States colored gray do not have any response time data.

Surprisingly, we discover that emergency response time for most states does not vary that much. In fact, some of the only states with significantly higher times include Wyoming and Mississippi, two states that also happen to have higher fatal incident ratios. Therefore, although most states have similar ability to promptly dispatch emergency personnel, it appears that keeping response times lower may help reduce the number of fatal incidents.

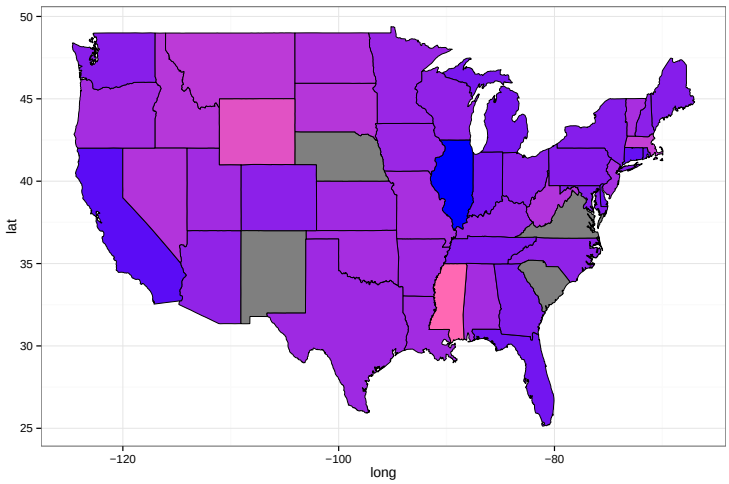

Figure 8: Emergency resopnse times for the contiguous states. The spectrum of blue to pink indicates an increasing (worse) time.

4.4 Environmental and behavioral factors

Turning our discussion to other crash related factors, we first investigate the influence of anomalous weather. We define this type of weather as neither clear nor cloudy as described in the FARS data. For 2010, this amounted to a total of 3030 incidents, or about 10.0013 percent of all cases. Figure 9 gives a count of of incidents involving these anomalous weather patterns faceted by month.

Interestingly, we do not find many occurences of incidents caused primarily by severely reduced visibility. Indeed, items like fog and blowing snow do not contribute greatly to this total. Instead, we see an enormous contribution by rain throughout the year, as well as a sizeable amount of snow incidents from December to February. While these conditions do cause reduced visibility, another factor to consider is the effect these conditions can have towards reduced vehicle control, possibly increasing the incidence of crashes.

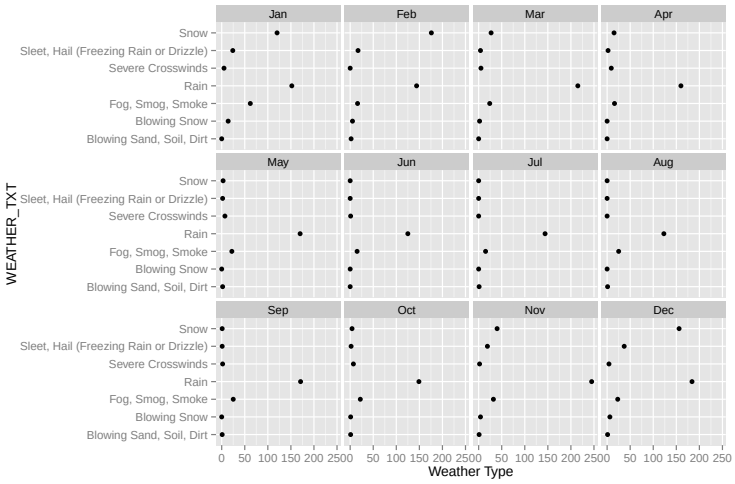

Figure 9: Dot plot of weather related incidents faceted by month of 2010.

If we consider the aspect of driver behavior, we are able to analyze the influence of impairments and distractions for all drivers involved in a fatal accident. Figure 10 illustrates the number of drivers known

to be inflicted by some sort of impairment in 2010 as tracked by FARS. Here, we find that causes of driver impairment are heavily dominated by the influence of not only alcohol, but drugs/medicine as well. This greatly reinforces common knowledge that mind altering drugs can have serious repurcussions in a driver's ability to safely operate a vehicle.

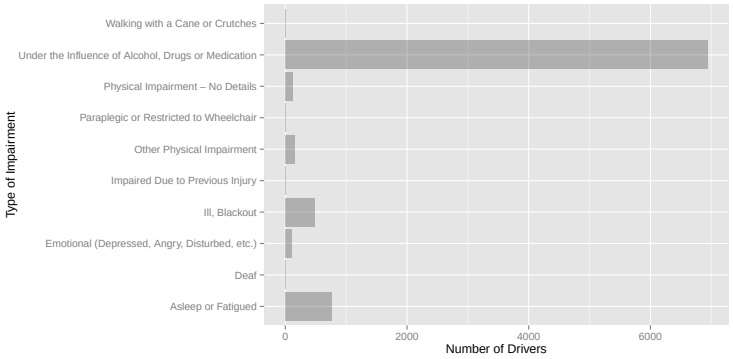

Figure 10: Histogram of the number of drivers involved in fatal incidents inflicted by types of impairments.

Aside from physical impairments on a driver's ability to safely navigate a roadway, we must also consider the role that distractions can play. Figure 11 displays this concern. While common distractoins like cell phone usage do provide a fair contribution to these incidents, it is also surprising to find that other disractions play a strong role.

For instance, the top distraction for drivers in this case that in which he or she is lost in or preoccupied with other thoughts. This illustrates an interesting notion that emotional state can have a dramatic impact on driving safety. Other noteworthy distractions that may be common practice include conversing with other occupants in the car, as well as by items outside the vehicle such as people, billboards, etc.

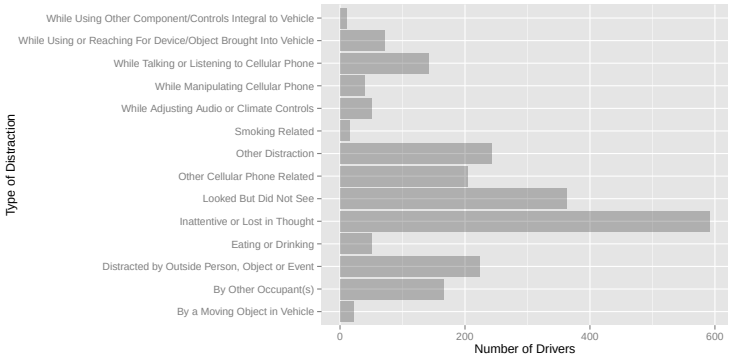

Figure 11: Histogram of the number of drivers involved in fatal incidents while distracted by given elements.

5 Conclusion

In summary, we have performed an analysis of the FARS census performed by the National Highway Traffic Safety Administration in 2010. This has been done in order to obtain an understanding in recent trends in fatal automobile accidents, as well as some of the risk factors associated with them. Here, we have focused our attention to the elements of time of incidents, state-related factors, as well as environmental and behavioral circumstances surrounding each incident.

Assorted time series analyses show that crash frequency does indeed vary with time period. For instance, we have found that incidents seem to increase in the summer months, then fall off during the winter season. During a typical week, it is evident that crash rates fall during the middle of the week before rising sharply during the weekend. On a daily basis, these accidents tend to occur in the evening hours, with some increases occuring during the morning rush hour.

On the state-level, it has been shown that the risk of an incident can vary from state to state. Interestingly enough, we find that more rural states with higher speed limits tend to have a greater number of accidents relative to the number of registered drivers in that state. Another element of concern lies with the fact that states that find themselves with larger delays in emergency response arrival also tend to have these higher fatal incident rates, showing the need for fast medical care.

Lastly, we have found that particular factors stand out in the arena of environmental and behavioral risks. On the subject of environment, rainy weather serves as the most predominant anomalous road condition related to fatal accidents. This may be due to the fact that rain is a more common weather pattern, and thus should be one of the first concerns for drivers and car manufacturers. Driver behavior also shows significant risk factors, including alcohol and drug use, as well as emotional states that cause drivers to be lost in thought. Again, these are important elements that drivers and law enforcement should be aware of.

A possible conduit for further research exists in the exploration of how different methods to reduce driving risk affect the number of fatal motor vehicle incidents. These methods could include items such as tougher laws on alcohol use while driving, public awareness programs, new speed limits, and so forth. In this way, we hope that this research will prove useful in continuing the downward trend in fatal accidents from 2005-2011, providing for a safe and secure driving environment.

References

[1] Richard Read, *Estimated 35,200 US traffic deaths reported in 2013*, The Christian Science Monitor, `http://www.csmonitor.com/Business/In-Gear/2014/0215/Estimated-35-200-US-traffic-deaths-reported-in-2013`, 2014

[2] National Highway Traffic Safety Administration, *Fatality Analysis Reporting System (FARS)*, `http://www.nhtsa.gov/FARS`, 2014

[3] Office of Highway Policy Information, *Publications*, `https://www.fhwa.dot.gov/policyinformation/`, 2014

[4] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, `http://www.R-project.org/` , 2014

[5] Yihui Xie *knitr: A general-purpose package for dynamic report generation in R*, `http://yihui.name/knitr/`, 2014

[6] National Highway Traffic Safety Administration, *Fatality Analysis Reporting System (FARS) Analytical User's Manual 1975-2012*, `http://www-nrd.nhtsa.dot.gov/Pubs/811855.pdf`, 2014

[7] Brian Ripley and Michael Lapsley, *RODBC: ODBC Database Access*, `http://cran.r-project.org/web/packages/RODBC/index.html`, 2014

[8] Governors Highway Safety Administration, *State Speed Limits*, `http://www.ghsa.org/html/stateinfo/laws/speedlimit_laws.html`, 2014