New York City Case Study: Inclement Weather and Hazardous Driving Conditions

Carl Colena, David Zeng, Jonathan Rozario, and Samuel Cohen

***Abstract*— In this report, the effects of inclement weather are analyzed in a comprehensive case study to determine its effect in relation to vehicular collisions in New York City. It was hypothesized that inclement weather would have a statistically significant impact on collisions in New York City, and that inclement weather would lead to more injurious and fatal collisions. Collision and weather data were processed on Apache Spark and went through a number of phases in preprocessing and analysis. Through analysis, 5 key metrics were chosen to determine 'inclement' weather as opposed to 'non-inclement' weather. After post-processing the gathered statistical data, it was found that there is an increased probability of collisions during inclement weather, indicating a causative link between bad weather and dangerous driving conditions.**

I. Introduction

Whether a pedestrian, a biker, a bus commuter, or a driver, being safe in New York City is a top priority. It is well known that vehicular accidents are one of the leading causes of travel-related death in the world, and in the densely populated sprawl of New York City, this becomes even more poignant. With an average of over 600 vehicular accidents reported per day, it is worth asking why these accidents occur, and what can be done to prevent more accidents. It can also inform lawmakers as to how to make the best policy in order to keep people safe. For that reason, our project investigated the link between specific inclement weather and traffic collisions, in order to identify factors that affect safety on the roads.

For our project, we specifically investigated the effects of weather on New York City traffic. Our reasoning for focusing on New York were twofold. One, as New Yorkers we are motivated in doing all we can to make our city safer. But more importantly, the recent explosion of open data in recent years has led to New York making a centralized, easy-to-access website with all government data open to everyone, giving us the ability to comb through the entire dataset of collisions to discover a trend.

We hypothesized that there would be an increase in collisions during inclement weather such as rain, snow, or wind and that there could be certain locations that are more accident prone in specific conditions.

II. Datasets used

Our project mainly used two datasets: the New York City Police Department's database of collisions from 2012 to present from NYC Open Data, and historical hourly weather data scraped from Weather Underground for every hour of each day from the year 2011 to present.

The New York Police Department (NYPD) Motor Vehicle Collision data was made available on the NYC Open Data website. It consists of over 1.03 million rows, with each of the rows being a specific traffic incident that occurred between 2012 and the present day (the dataset is, as of May 2017, still updating every week). These rows have 29 columns: those include date, time, location, and cross street name, as well as type of collision and number of people involved, whether there were any injuries or fatalities due to the crash, contributing factors to the accident, and the types of vehicles involved, among others. For a full description of the data one can go to NYC Open Data's website. The total file size is around 220 MB.

For weather data, we initialized downloaded the daily summary data from the National Oceanic and Atmospheric Administration (NOAA) dataset for the years 2011 to 2016. This dataset was split into years and contained information about the max and min temperatures precipitation amounts for each day for all the weather stations in the nation and the total file size was over 1GB of data. However, after filtering for the central park weather station out for one year of data on a local machines, we decided that the dataset did have detailed enough information for us to work with. While the data set was large, it took time to filter for a specific station and only provided a few basic details for each day. We then switched over to manually scraping the Weather Underground website for our weather data.

The Weather Underground data was obtained via web scraping using a python script that we wrote and parsing the entire historical record of a NYC weather station from the year 2011 to present. The weather data we obtained from the site included the datetime in UTC along with the time in EDT, temperature measurements, dew point, humidity, sea pressure, visibility, wind speed and direction, gust speed, precipitation amount, events, and conditions. This data was available for every day and every 51st minute of each hour, which some special events recorded between the hourly 51st minutes as needed. This data provided us with a rich amount of historical weather data for NYC that is required for our project.

III. Big Data Challenges

Our primary challenge was mostly the volume of our data. While the datasets that we settled on for this project was not

exactly in the gigabyte range, the size of our datasets could be considered too big to work with using conventional methods on a local machine. Additionally, the datasets contained quite a large amount of features for each row of data we had. This meant that processing the data would take a lot of time and memory.

Although the collisions dataset is small enough to work on a high-memory laptop for some of the tasks, many necessary data manipulation operations were simply not possible in this kind of environment due to the scale and volume of data needed for processing. This was worsened when weather data was joined with collision data, as significant memory and CPU is required to perform the joins with respect to date and time. The resulting final dataset was required in order for us to do our analysis proved to be a pain to do on our local machines. Using Spark on a cluster allows for a more practical approach for a large amount of data with complex processing due to the speed and functionality of parallel processing.

Furthermore, the cluster allowed us to have all of our data in one central location for processing, which saved time transferring large amounts of data from personal laptop to personal laptop, just so we could work on it. A single powerful remote cluster is much more useful and power than a local computer in that respect. Moreover, by utilizing the cluster for processing, we would not need to waste our own computer's resources and laptop battery life.

IV. Methodologies

Our project used Python programming language with the Spark Python API (PySpark) for all of our computation and processing. The main parts of our code were run on NYU's CUSP cluster with direct access to the full datasets. Additionally, some pre- and post-processing was done on our local machines for convenience. We used smaller versions of the datasets on our local machines for testing our algorithms and approaches when working with PySpark.

A. Preprocessing and Gathering

One of the first things we worked on was obtaining our weather data. After checking out and abandoning the NOAA datasets due to insufficient data, we decided to obtain our data from Weather Underground. We did this because Weather Underground gave better granularity in weather resolution with hourly data. This was achieved using a python script, "wunderground_parse.py"[A.1], that was able to scrape the hourly weather updates for each day in a range of specified years. The scraped data was then written to and saved to CSV. Each year of weather data took roughly 5 minutes to be completely scrapped since the script has to access each via http requests. For the years 2011 to 2017, it took roughly 35 minutes to complete.

Next, we decided on weather classifications to apply to our data. We settled on using boolean values 0 for false and 1 for true in relation to our categories of rain, snow, freezing, windy, and low visibility. This was then appended to the existing weather data.

The thresholds for Rain and Snow are to look for the words 'rain' or 'snow' in the conditions column (as well as any similar phrases such as drizzle or sleet) The threshold for freezing was set to be less than or equal to 32 degrees Fahrenheit (0 degrees Centigrade) to denote icy conditions. The threshold for wind was set for gust speeds meeting and/or exceeding 30 mph.

The next step is to prepare our data for joining. The date and time in EDT were provided in our collision data in two separate columns while the datetime information in the weather data was in UTC. To work around this, we added a new column DATETIME in the collisions CSV file using "add_datetime_to_collisions.py"[A.1] and added a DatetimeEDT column in the weather CSV in our "wunder_preprocessing.ipynb" file. This allowed for a common key column to provide a base to join the two datasets. The joining was achieved in the "join_weather_collisions.py"[A.1].

Our approach to joining was to use RDDs to map the two datasets to a common Key ID, "datehour", which consists of the date and hour of each row of the two datasets. Our assumption was that weather conditions probabilistically will not drastically change within an hour time span. Working from this assumption, we mapped each dataset to a tuple "(datehour, p)" where p is the entire row of the dataset. With the two mapped RDD partitions, we then join the two datasets using the datehour key and take the first weather entry from each weather entry of the date hour. This leaves us with a joined RDD that we convert into a dataframe with a schema in order to write to a CSV file, that we end up naming "coalesce_joined_weather_collisions_sorted_fixed.csv". This process took around 11 minutes to complete and yielded a 300MB sorted CSV file.

B. Processing

Once we have the primary CSV file with weather and collision data joined, we were able to work with the resulting dataset for our analysis. We decided that we wanted to look at the data as a whole as well as on a yearly basis. To deal with this, we used "split_csv_to_years.py"[A.1] to load the CSV into a dataframe, automatically get the distinct number of years in the CSV, using a udf, and then apply an SQL select query for each distinct year and saved them to individual CSV files under our " partitioned_joins_by_year" folder. This took 3 minutes to complete.

Next, we moved to the main bulk of our analysis, the statistics. We loaded the full CSV file and the yearly CSV files into SQL dataframes and processed them through statistical filtering and grouping. Using SQL queries on the SQL dataframes, along with general computations we were able to obtain the key metrics necessary to analyze the statistical significance of each type of inclement weather

condition on collisions, injuries, and deaths. This included the average collisions per day for the whole year and for individual months in a year and data on the death and injury rates and counts. The resulting data was saved in a JSON format for post-processing use. This process took roughly 30 minutes to complete for the entire dataset on the CUSP Cluster.

C. Post Processing and Analysis

With the JSON statistics, the data was entered into python scripts that analyzed the occurrence of inclement weather normalized with collisions that occurred during inclement weather. In addition, graphs and trends were extracted from the JSON statistic files and used to show how collision data changed on many different dimensions, including time of year, weather condition, fatality rate, and death rate. This gave a broad scope for leading us to the results we have.

Finally, we processed the data for CartoDB use. This was done in "carto_coords.py"[A.1] as a proof of concept and then further normalized in "get_carto_weather_normalized.py"[A.1]. These files go through the joined CSV and filter for specific weather classifications in order to get the aggregate counts for each distinct geo-location. This took about a minute on the cluster. Using the aggregate counts data, we could upload our dataset to CartoDB and show the locations with the greatest number of collisions that occurred with a specific weather pattern. However, this data was not normalized with the number of collisions that occurred at that location. We then needed to normalize our data using "get_carto_weather_normalized.py"[A.1]. This applied the same concept but aggregated the weather classification counts along with the percentages. This left us with the ability to show collision locations that have only occurred during specific weather patterns.

V. Results

In our analysis, we looked at the data through a number of dimensions in order to gain a more thorough and comprehensive view at the meaning of the data, and what trends exist in the data. We looked at the following metrics: Inclement weather both in aggregate and individually by type (Rain, Snow, Wind, Freezing, and Low Visibility); Time of year, month of year, and year itself; Location of collision; Injury rate, Death rate, and Contributing Factors. In Figure 1, we analyze the effect of the time of year with aggregate collisions. We see that in general there is a statistically significant uptick during the middle months of May, June, and July.

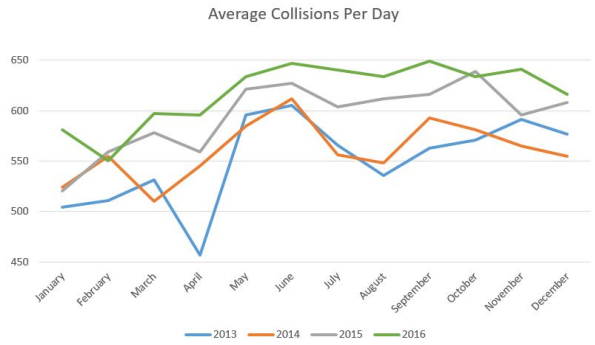


Figure 1. Average Collisions per Day by Year and Month

In the second figure, we have the % of collisions which resulted in at least one injury per month. For a trend we see that during the beginning of the year and sometimes towards the end of the year, the injury rate drops.

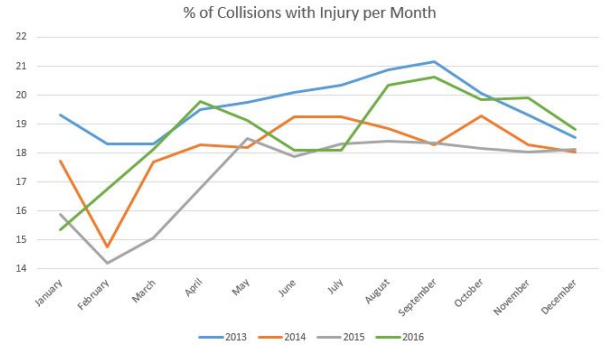


Figure 2. Percent of Collisions with Injury per Month

In the third figure, we show the rates of fatal collisions per month. It's not clear that there is any significant trend, besides, similar to that of the injuries in Figure 2, that there is a slight yet marked decline in fatalities at the beginning of each year. These trends may be due to the fact that vehicles tend to drive slower during snowy days.

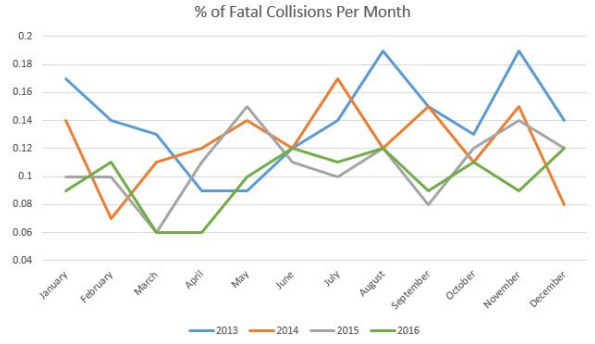


Figure 3. Percent of Collisions with Death per Month

In Figure 4, we show the yearly injury rates for each weather condition. It is quite clear that compared to the baseline of the double black line, rain is significantly higher on the list than any of the other weather conditions, which warrants investigation, as well as points to a potential causal relationship between collision rate and weather condition.

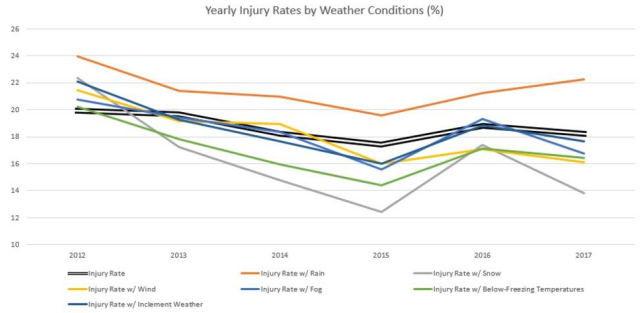


Figure 4. Yearly Injury Rates by Weather Conditions

In Figure 5, we show the yearly death rates by weather conditions. We do not see any particular trend besides that death rates in general are declining as the years go on. One likely reason of this is the enactment of the New York City Vision Zero Initiative in mid 2012 in an effort to enforce speed restrictions to prevent fatal crashes from occurring.

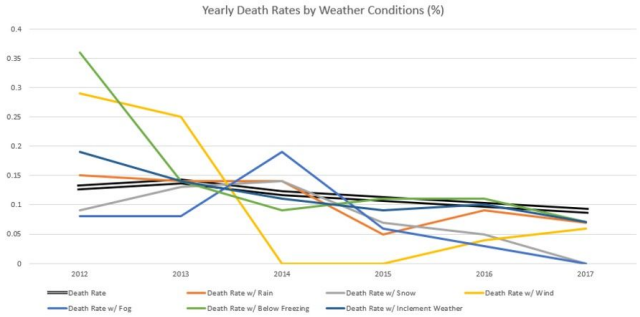


Figure 5. Yearly Death Rates by Weather Conditions

VI. Analysis

With all of our data compiled at the end of the project, we were able to make a number of conclusions about road safety and weather. More work needs to be performed to normalize these data with traffic density and other statistics. Nonetheless, they represent a good initial investigation into the relative danger of driving under varying weather conditions.

A. Collision Frequency

The main thing we analyzed was the relative frequency of collisions in inclement weather, specifically as compared to the frequency of weather events in general. In particular, if weather has no effect on the frequency of collisions, then one would expect the likelihood of collisions under weather to match the likelihood of weather effects in general. So, for example, if in a given month 30% of the weather samples were rainy, we would expect 30% or so of the crashes to be rainy. We found that for rain, snow, and high winds, this was not true. There were more traffic incidents in those forms of inclement weather than there were records of that weather, indicating a causative link between the weather and increased collisions. Months were chosen because weather is similar within each month, while years were not chosen because weather obviously varies drastically.

```
Rainy collisions occurred more than rain 86.6666666667% of months
Foggy collisions occurred more than fog 45.0% of months
Snowy collisions occurred more than snow 86.6666666667% of months
Windy collisions occurred more than wind 83.3333333333% of months
Freezing collisions occurred more than freezing 46.6666666667% of months
```

Figure 6. Post-Processing Analysis of Normalized Collision Rates by Inclement Weather Condition.

We see in Figure 6 that crashes during rain occurred more often than rain did in 86% of months, a similar number with collisions under snow, and 83% of the months, windy collisions happened more often than wind did. However, such a relationship was not found with fog and freezing temperatures. Windy weather surprised us, but it is possible that it is highly correlated with storm conditions.

B. Collision Locations

Another aspect we investigated were locations in New York City where collisions occur more frequently than others. A snapshot of this can be seen in Figure 7. We tried to discover if there were locations which were normally safe to drive in, that became less safe in bad weather. However, we found that the data was overwhelmed by the densest traffic locations, such as Time Square or the various highways. Those were the most dangerous either in bad weather or not, indicating that rather than making new locations more dangerous, weather intensifies the accidents happening in places where they usually do. Our CartoDB map demonstrates these effects.

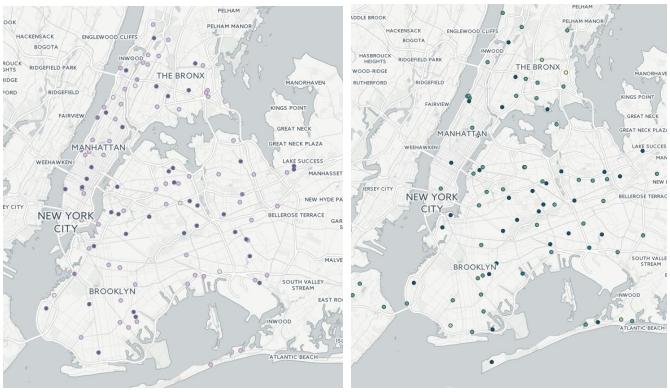
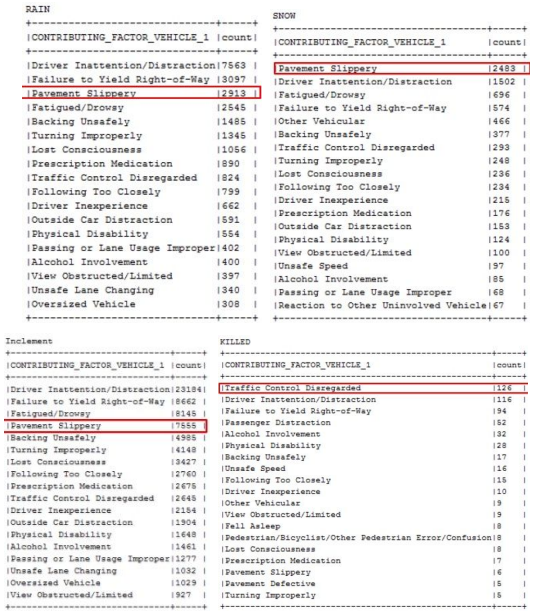


Figure 7. Map of Collision Locations on Inclement Weather Condition. Left: Rain. Right: Low Visibility. [A.2]

C. Causes of Collisions

The traffic collision data includes a column on "contributing factor to accident", a human-entered data field that attempts to record the reason for the accident. This can be seen in Figure 8 for four metrics. We found a change in this field under specific weather conditions. In particular, slippery pavement was the 10th most common cause of accidents in all of our samples, but second in our rainy samples and first out of of snowy incidents. This shift in reasons for traffic accidents, combined with a higher density of incidents overall under inclement weather, indicates that the danger of driving shifts under bad weather.



RAIN

CONTRIBUTING_FACTOR_VEHICLE_1	count
Driver Inattention/Distraction	7563
Failure to Yield Right-of-Way	3097
Pavement Slippery	2913
Fatigued/Drowsy	2545
Backing Unsafely	1485
Turning Improperly	1345
Lost Consciousness	1056
Prescription Medication	890
Traffic Control Disregarded	824
Following Too Closely	799
Driver Inexperience	662
Outside Car Distraction	591
Physical Disability	554
Passing or Lane Usage Improper	402
Alcohol Involvement	400
View Obstructed/Limited	397
Unsafe Lane Changing	340
Oversized Vehicle	308

SNOW

CONTRIBUTING_FACTOR_VEHICLE_1	count
Pavement Slippery	2483
Driver Inattention/Distraction	1502
Fatigued/Drowsy	696
Failure to Yield Right-of-Way	574
Other Vehicular	466
Backing Unsafely	377
Traffic Control Disregarded	293
Turning Improperly	248
Lost Consciousness	236
Following Too Closely	234
Driver Inexperience	215
Prescription Medication	176
Outside Car Distraction	153
Physical Disability	124
View Obstructed/Limited	100
Unsafe Speed	97
Alcohol Involvement	85
Passing or Lane Usage Improper	68
Reaction to Other Uninvolved Vehicle	67

Inclement

CONTRIBUTING_FACTOR_VEHICLE_1	count
Driver Inattention/Distraction	23184
Failure to Yield Right-of-Way	8662
Fatigued/Drowsy	8145
Pavement Slippery	7555
Backing Unsafely	4985
Turning Improperly	4148
Lost Consciousness	3427
Following Too Closely	2760
Prescription Medication	2675
Traffic Control Disregarded	2648
Driver Inexperience	2184
Outside Car Distraction	1904
Physical Disability	1648
Alcohol Involvement	1461
Passing or Lane Usage Improper	1277
Unsafe Lane Changing	1032
Oversized Vehicle	1029
View Obstructed/Limited	927

KILLED

CONTRIBUTING_FACTOR_VEHICLE_1	count
Traffic Control Disregarded	126
Driver Inattention/Distraction	116
Failure to Yield Right-of-Way	94
Passenger Distraction	52
Alcohol Involvement	32
Physical Disability	28
Backing Unsafely	17
Unsafe Speed	16
Following Too Closely	15
Driver Inexperience	10
Other Vehicular	9
View Obstructed/Limited	9
Fell Asleep	8
Pedestrian/Bicyclist/Other Pedestrian Error/Confusion	8
Lost Consciousness	8
Prescription Medication	7
Pavement Slippery	6
Pavement Defective	5
Turning Improperly	5

Figure 8. Top 20 Contributing Factors for Given Weather Conditions. Red Highlights Key Points

D. Shifts in Collision Frequency

Despite our findings of increased collision rates in bad weather, we identified certain trends in the fatality and injury rates over our dataset. In particular, we discovered that accidents in general were higher in the summer months and lowered in the winter, perhaps correlating with how much people drive in general. Fatalities showed a similar trend. Additionally, we found long-term trends in our data set: over the 4 years we recorded traffic incidents, fatality numbers went down, but injury numbers rose. This could correlate with the Vision Zero initiative by NYC, which has a focus on lowering traffic fatalities but does not focus much on injuries. (Vision Zero is named after a goal of zero traffic fatalities in a year).

VII. Conclusion

This project represents an important first step in analyzing the effects of weather on traffic accidents. Although its scope was somewhat limited, being that we stuck to collisions in New York City, and within a short timespan of four years, we have identified trends in traffic accidents that should still be useful for keeping our roads safe. Lawmakers could use our conclusions in order to design better urban road systems. Drivers and pedestrians can use our conclusions to ensure they remain safe on the road. Our work only scratches the surface, but the results and conclusions made are still suggestive of underlying trends.

Future work that builds on our project should utilize available traffic density data to make better conclusions about the danger of driving in certain weather as well as certain road conditions. This data is available online but due to limited time, we did not get a chance to incorporate it into our project's analysis of our data due to the fact that they are not geo-tagged. Manually tagging would be time consuming but would help make the traffic datasets more manageable.

Some challenges we identified during this project were dealing with inconsistent data, such as missing rows or columns in our data, which made processing it difficult at times. For example, there were a few extremely rare days where Weather Underground was missing weather day. Additionally, many rows in the collision data did not contain detailed location information. We also found it helpful to use Spark's built in libraries, such as data frame functions, as we progressed through the project even though there is a learning curve involved. Spark provides a powerful way to perform computations on big data projects that is very valuable.

Another learning experience came from utilizing the CUSP cluster. It was sometimes frustrating to convert a ipynb file to a py file and fix the errors due to the 40 second start up time for each job. Errors were also difficult to find and access. It also made computations that were simply not possible on our local machines take only minutes. However, we found that it required significant tweaking of the memory allocated to the executors as well as the number of nodes used. In the future we hope to get a better understanding of how to optimally tweak the execution, rather than just relying on trial and error as we did for this project.

IX Appendix

[A.1] Github Repository:
https://github.com/slamwell17/CCNYBigData

[A.2] Carto Interactive Map, Collision Locations:
https://asdfblarg.carto.com/builder/5676480a-413e-11e7-815a-0e3ff518bd15

X References

[1] "NYPD Motor Vehicle Collisions | NYC Open Data", *Data.cityofnewyork.us*, 2017. [Online]. Available: https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95. [Accessed: 26- May- 2017].

[2] "Weather Forecast & Reports - Long Range & Local | Wunderground | Weather Underground", *Wunderground.com*, 2017. [Online]. Available: https://www.wunderground.com/. [Accessed: 26- May- 2017].

[3] "How Do Weather Events Impact Roads? - FHWA Road Weather Management", *Ops.fhwa.dot.gov*, 2017. [Online]. Available: https://ops.fhwa.dot.gov/weather/q1_roadimpact.htm. [Accessed: 26- May- 2017].

[4] B. Leard and K. Roth, "How Climate Change Affects Traffic Accidents | Resources for the Future", *Rff.org*, 2016. [Online]. Available: http://www.rff.org/research/publications/how-climate-change-affects-traffic-accidents. [Accessed: 26- May- 2017].

[5] "Warning/Advisory Threshold Reference", *Weather.gov*, 2017. [Online]. Available: http://www.weather.gov/btv/wwa_reference. [Accessed: 26- May- 2017].

[6] "Data Catalog", 2017. [Online]. Available: https://data.noaa.gov/dataset. [Accessed: 26- May- 2017]