

# Roadway Factors behind Car Accidents

Authors: Chong Tan, Xiangjie Chen  
Group Project Team#30  
18755-Networks in the Real World  
Mountain View, U.S.

**Abstract**—The number of car accidents and fatalities on roadways continues to climb, so to reduce traffic accidents, analyzing and investigating the factors behind them becomes important.

**Key Words:** Car Accidents; Natural Conditions; Networks

## I. INTRODUCTION

The problems we plan to address in this project are investigating the key factors behind the high number of traffic accidents within a certain area.

Our goal is to analyze traffic accidents and visualize accident hotspot locations in order to understand the causes behind them and predict future accidents with risk, frequency, and environmental stimuli.

## II. PRIOR WORK

George Yannis and Matthew G. Karlaftis used an autoregressive model to estimate the effects of weather conditions on traffic accidents, including vehicle accidents, vehicle fatalities, pedestrian accidents, and pedestrian fatalities. The temperature rise was found to lead to increased accidents.[1]

This article investigates several natural conditions, temperature included, to explore the most significant factor causing traffic accidents.

## III. APPROACH

The data set contains 2.8 million accident records with forty-seven unique attributes for each entry, including environmental conditions, geolocation, and weather information.

### A. Clean Dataset

Before building the graph, we cleaned the data by removing empty or broken entries, so the total valid data is about 1 million. We analyzed the data and narrowed down the features to ten critical factors of an accident.

Since the dataset is too large, it was super time-consuming to build a network using all the data. We split data equally and extracted a subnetwork with ten thousand nodes for Milestone1.

### B. Generate Networks

Our graphs are generated to show the likelihood of car accidents. The nodes are car crash records, and the existence of an edge between two vertices indicates that these two records are alike in terms of the natural conditions under which accidents happened.

For instance, below Table 1 is the table of three car accident records. Records A and B are almost the same in terms of severity, temperature, humidity, pressure, visibility, and wind speed, so we connected these two nodes while leaving node C alone since the similar factors it shares with A or B do not meet our threshold.

Table 1 Sample Car Accident Records

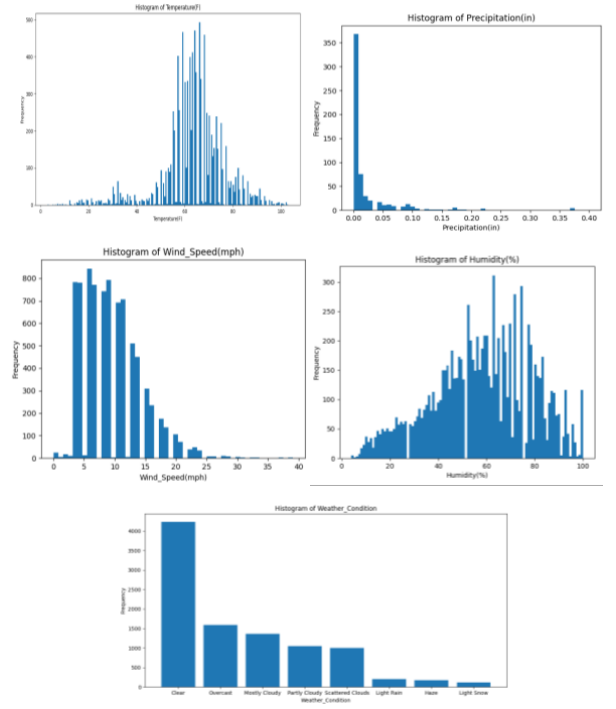
Record	Severity	Temperature	Humidity	Pressure	Visibility	Wind_speed
A	3	33.8	93%	29.76	10.0	10.4
B	3	33.1	91%	29.18	10.0	10.4
C	4	46.7	75%	29.37	9.5	10.4

## IV. EXPERIMENTAL SETUP AND RESULTS

We picked and analyzed the dataset's five weather-related and four environmental features.

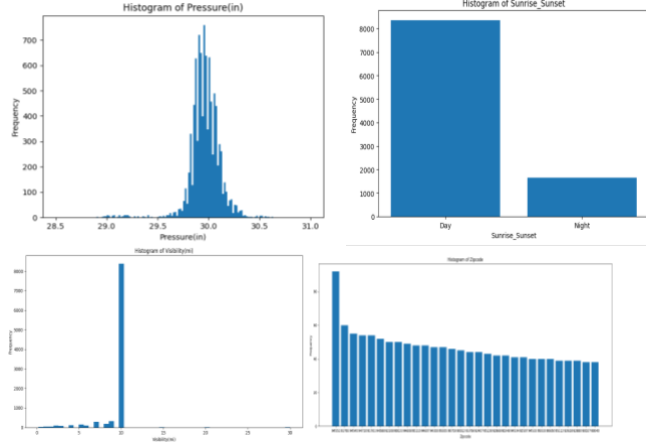
### A. Five Weather-Related Features

Those Five histograms are the frequency of weather-related features, including humidity, wind speed, temperature, precipitation, and weather condition.



### B. Four other Features

These are the other four critical factor histograms. The frequency of pressure, visibility, daytime, and, the top 30 with the highest frequency of zip code.



### C. Networks based on two thresholds for common factors

We built two networks based on seven and eight common factors as the thresholds.

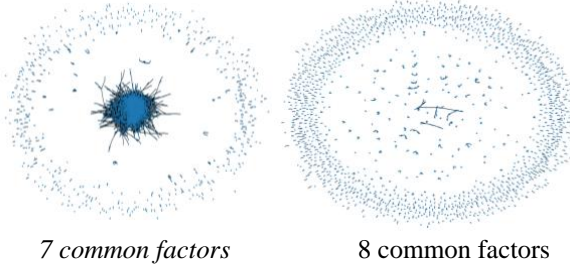
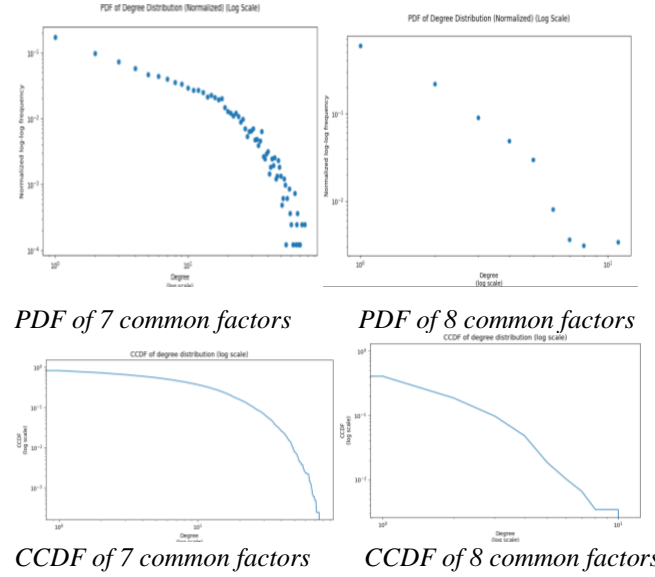


Table 2 Networks Information

Table		
Thresholds	Seven	Eight
Total Nodes	10000	10000
Number of Valid Nodes	8221	3821
Number of Edges	43765	3407
Avg Degree	10.625	1.783
Diameter	23	10
Average Path Length	5.811	1.719
Density	0.001	0.0
Average Clustering Coefficient	0.431	0.84
Number of Components	508	1441
Number of Nodes in the Largest Component	6937	25

### D. PDF and CCDF with log-log scale

The network with seven common factors is similar to the exponential distribution. The other network is more like a power law distribution.



## V. CONCLUSION AND SHORT-TERM PLANS

We classified the natural conditions into ten factors and conducted data cleaning upon broken records with insufficient information.

We also examined the connections between car crash records by establishing edges connecting similar accidents. The preliminary result indicated that most car crashes happened under relatively mild natural conditions. We inferred from this result that natural factors might not directly affect the happening of accidents but instead help or bolster them.

Our plan for Milestone 2 includes two parts:

- Apply weight to different natural factors in that we found some factors like the visibility are almost the same in most records;
- Leverage machine learning methods to make predictions on the severity of car accidents to verify if our analysis from the network itself is correct;

Regarding the contribution of team members, Chong was in charge of the coding part, and Xiangjie did the analysis work. In addition, we shared the work of collecting and cleaning data before Milestone 1.

## REFERENCES

- [1] YANNIS, GEORGE, AND MATTHEW G. KARLAFTIS. "WEATHER EFFECTS ON DAILY TRAFFIC ACCIDENTS AND FATALITIES: A TIME SERIES COUNT DATA APPROACH." PROCEEDINGS OF THE 89TH ANNUAL MEETING OF THE TRANSPORTATION RESEARCH BOARD. VOL. 10. 2010.