Report 10/04

Boyang Zhang

Introduction

The data used in this research is obtained from the Fatality Analysis Reporting System of the National Highway Traffic Safety Administration from years 2011 to 2015. Population estimates of each county from 2011 to 2015 is obtained from U.S. Census Bureau based on most recent decennial census data (http://www.census.gov/popest/data/counties/totals/2015/CO-EST2015-01.html). Median household income of each county from 2011 to 2014 is also obtained from U.S.Census Bureau (https://www.census.gov/did/www/saipe/data/statecounty/data/index.html).

Goal of this study

1. To identify communities that might be at a higher risk for fatal crashes, either in county-level or state-level
2. Since Maryland continues to be ranked as the richest state in the U.S., we want to analyze relationship between economic conditions (i.e. median household income) and fatality rate in MD and District of Columbia area. Also, we could broaden our analysis to other states in U.S.

Accident file(2011-2015):

Among 2336 accidents and 2524 fatalities, several crash-level related factors are listed as follows. Here, we merged data based on common columns shared by accident data files from 2011 to 2015.

1. Time-varying variables

- Seasonality (month): there is no extreme difference across months
- Day of week: on Mondays and Sundays, accidents are more likely to happen in MD and DC area.

2. Geographical factors

- Intersection type: most of the accidents do not occur in intersection area and remaining accidents are mostly occurred in four-way intersection and T intersection.
- National highway system: among all accidents, a quarter of accidents are in NHS.
- Trafficway: identify specific roads

3. Enviornmental factors

- Light condition: indicate type/level of light that existed at the time of the crash(day or night).
- Weather: summary of weather condition. Most of the accidents do not occur during adverse weather conditions.

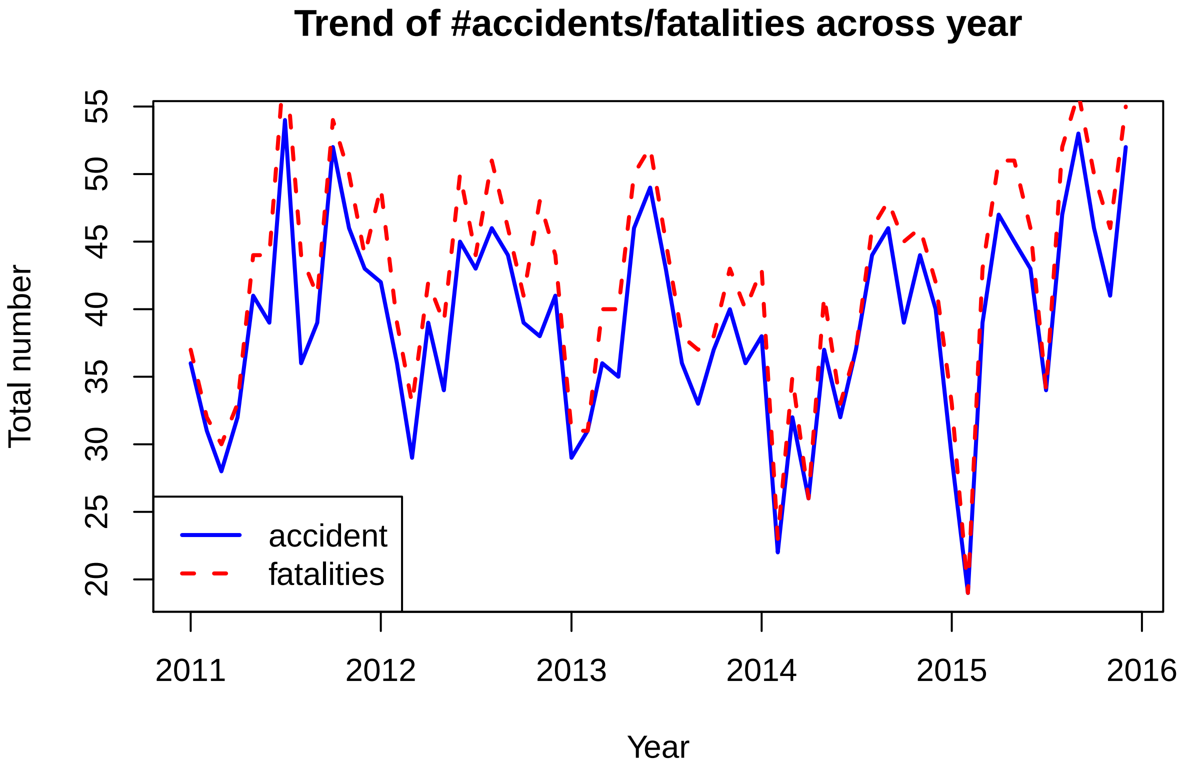
4. Information about crash

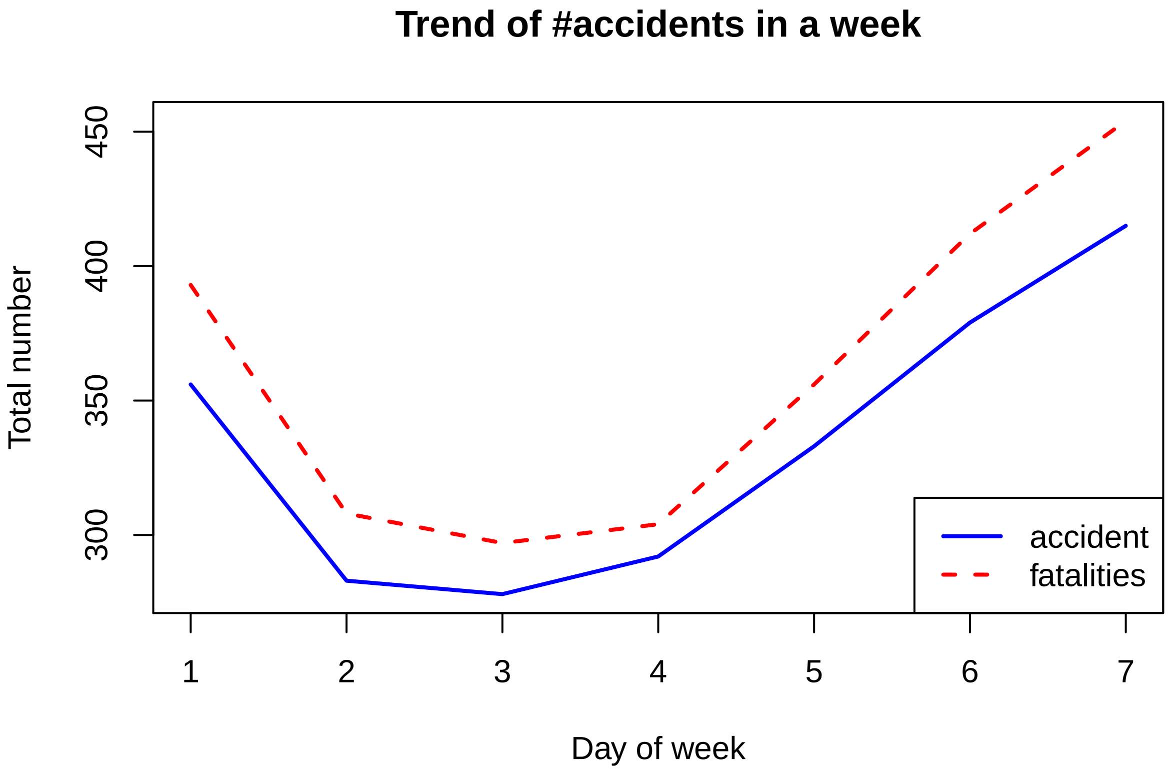
- Number of persons not in vehicles in transport: most of the accidents do not involve other person outside transporting vehicle, but still a quarter involves other people outside the car.
- Number of persons in transporting vehicle ranging from 1 to 18.
- First harmful events(accidents): first cause of crash and the top 5 harmful events are namely, vehicle in transport, pedestrian, tree, curb and guardrail face.

- Crash related factors: This contains too much missingness. Similar information can also be obtained from vehicle files and person files (vehicle-level related factors, driver-level related factors and person-level related factors in the Person data file).
- Drunk drivers: There are 737 drunk driving cases.

Explorartory analysis

Time varying factors



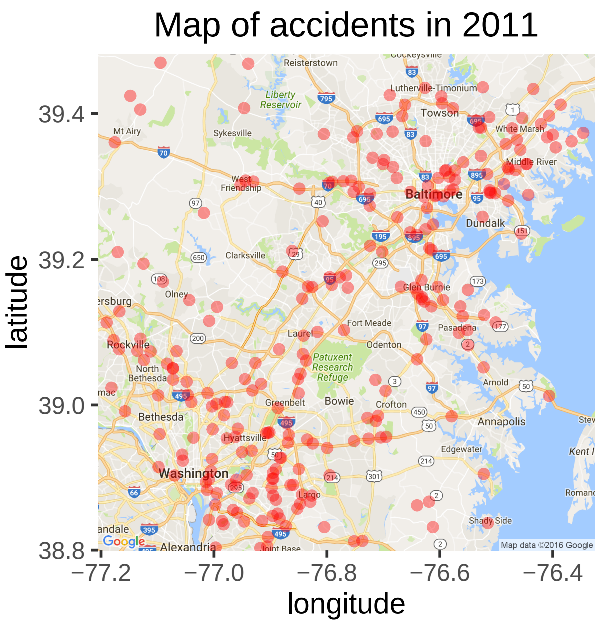


From the above plots, we notice that the seasonality of traffic fatality or accidents across years may not be evident, especially for year 2015. But when we have a closer look at day of week, we observe that a strong "weekend effect", with comparably higher accident rate across years.

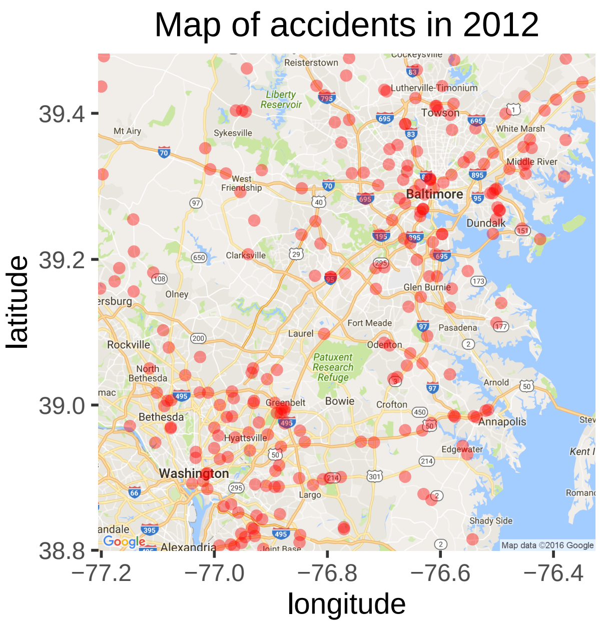
```
## Joining, by = c("year", "geo_id")
```

Geographical factors

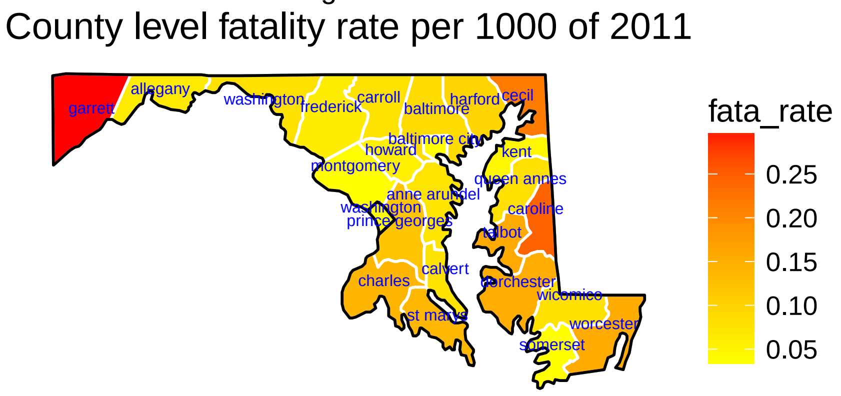
To simplify the question, we narrow down to analyze data from Maryland state and District of Columbia area. We calculate county-level fatality rate per 1000 people each year based on population and fatalities. Also, we try to make a comparison between each county's fatality rate and its median household income.



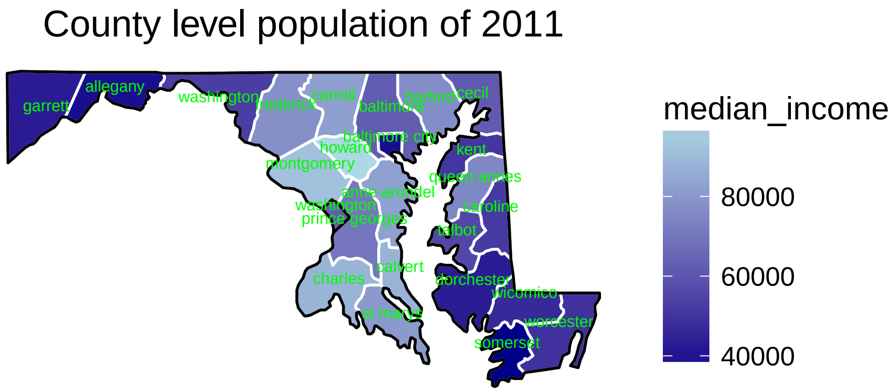
Map of accidents in 2011



Map of accidents in 2012



County level fatality rate per 1000 of 2011



County level population of 2011

As we could see, there exists negative correlation between regional median household income and regional traffic fatality rate in Maryland state and District of Columbia area in year 2011.

Step by step analysis plan

- Exploratory analysis on state-level (i.e. merging state-level data, plotting heatmaps of state-level fatality rate and median household income)
- Find crash-related factors, not only from accident file, but also from vehicle and person files.

  - The main issue for finding related factors is that vehicle-level and person-level data files are multiple responses per accident.
  - Thoughts: conduct principal component analysis on vehicle and person files and extract first 2 principal components from each file as vehicle-related and person-related factors respectively.

- Build prediction models for both state-level and county-level
  - Linear regression/ Poisson regression/ Truncated poisson regression
  - Comparison of methods

- Write reports