Reducing Traffic Accidents in the IoT Age

A data driven approach to traffic analysis and harm reduction

THOMAS ALEXANDER, RYAN KINNEAR, PRANAV BAROT, and RICHARD WU

Additional Key Words and Phrases: Data Science, Traffic Analysis, Internet of Things, Citadel, Correlation One, Datathon

1 INTRODUCTION

The rapid urbanization of the world has been driven by the access to cheap energy and rapid transportation of goods and people. The percentage of automobile ownership in North America has skyrocketed in the last century with Americans and Canadians possessing 797 and 607 automobiles per thousand people respectively (All [n. d.]). Consequently there has been a dramatic increase in the number of severe road traffic accidents with the world wide disease burden of motor vehicle accidents being ranked ninth with a projected rise to third by 2020 (Moh [n. d.]; Krug et al. 2000). In 2014 it was estimated that the economic costs of traffic accidents in the United States approaches $900 billion annually. While the safety of motor vehicles have continued to improve it is not sufficient to rely purely on automobile engineering solutions (Hosaka and Mizutani 2000). Enormous improvements in road safety may be realized with proper traffic engineering and identification of accident hotspots saving providing large reductions in human mortality and material costs(Hauer 1997). In addition to the cost in mortality and injury, there is an enormous economic cost associated with roadway inefficiency. It is estimated that the direct and indirect ecnomic cost of traffic congestion in America resulted in $124 billion dollars in direct and indirect costs in 2014 (Guerrini [n. d.]). Reducing roadway inefficiency can result in enormous savings to the North American public, and the costs of such projects are easily justified (Cos 2008).

2 TOPIC QUESTION

The rapid urbanization of the world has been driven by the access to cheap energy and rapid transportation of goods and people. The percentage of automobile ownership in North America has skyrocketed in the last century with Americans and Canadians possessing 797 and 607 automobiles per thousand people respectively (All [n. d.]). Consequently there has been a dramatic increase in the number of severe road traffic accidents with the world wide disease burden of motor vehicle accidents being ranked ninth with a projected rise to third by 2020 (Moh [n. d.]; Krug et al. 2000). In 2014 it was estimated that the economic costs of traffic accidents in the United States approaches $900 billion annually. While the safety of motor vehicles have continued to improve it is not sufficient to rely purely on automobile engineering solutions (Hosaka and Mizutani 2000). Enormous improvements in road safety may be realized with proper traffic engineering and identification of accident hotspots saving providing large reductions in human mortality and material costs(Hauer 1997). In addition to the cost in mortality and injury, there is an enormous economic cost associated with roadway inefficiency. It is estimated that the direct and indirect ecnomic cost of traffic congestion in America resulted in $124 billion dollars in direct and indirect costs in 2014 (Guerrini [n. d.]). Reducing roadway inefficiency can result in enormous savings to the North American public, and the costs of such projects are easily justified (Cos 2008).

With the advent of Internet of Things (IoT) devices the amount of available data for making transportation decisions has grown dramatically (Fle [n. d.]). The use of traffic aware navigation devices has become ubiquitous in society (Maier et al. 2010). This has resulted in the ability to provide point to point optimal navigation for large traffic segments at an individual level. With the coming advent of self driving cars the ability and requirements of routing policies are expected to continue to grow(Fagnant and Kockelman 2015). The demands on North America's roadways will consequently be refocused not on the needs of the human, but the artificial intelligence pilot. Prior to this restructuring it is crucial that we learn important lessons from the data available today so that we may design the safer and quicker roadways of tomorrow.

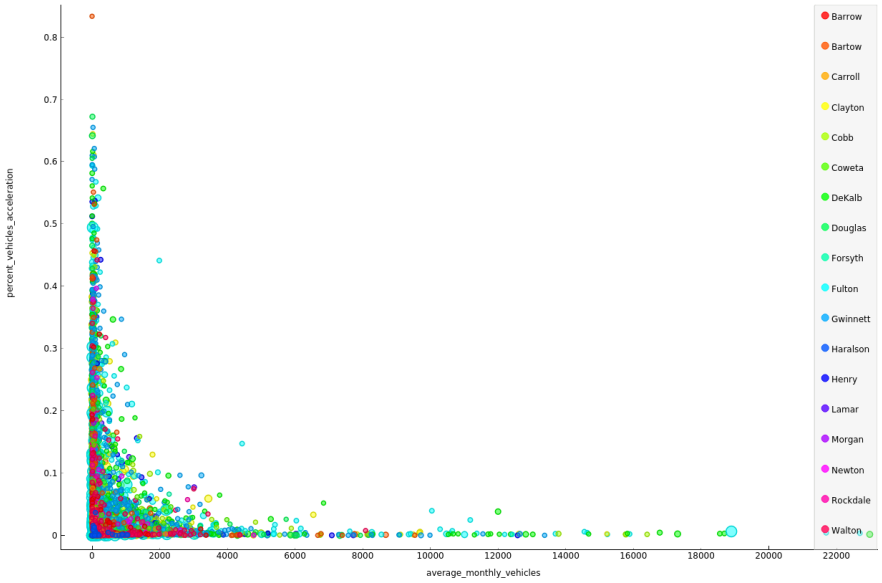


Fig. 1. Average vehicle acceleration compared against number of vehicles.

3 TOPIC QUESTION AND GOALS

With available traffic impediment and hazard data we will locate geographic areas of interest associated with large congestion and high accident severity. We will attempt to discern hot spot high impact areas and their correlation with local driving characteristics, landmarks and government funding. With our findings we will suggest key modifications to roadway design that may be made and generate substantial reductions in economic costs and human suffering associated with traffic accidents and congestion.

4 DATA PREPARATION

We perform all analysis in the Pylab numerical stack (?). Data tables are imported to Pandas, and then indexed based on their geographic locations. The *geohash* key provides an index to join tables together and obtain latitude and longitude coordinates from multiple tables. In particular we join the center *hazardous_driving_areas* with nearest neighbor *road_impediments*,*cell_coverage_dark_spots*and *fuel_station_metrics* by finding

$$\underset{lat_{t_i}, long_{t_i} \in \text{table}}{\arg\min} \sqrt{(lat_{t_i} - lat_{imp})^2 + (long_{t_i} - long_{imp})^2} \tag{1}$$

Where table is the table of interest to find the nearest neighbor, $(lat/long)_{imp}$ are the coordinates of the impediment of interest. We acknowledge that this linearized distance metric is inaccurate over long distances, but should be relatively accurate on the small scales of nearest neighbor distances.

We plot local features by their latitude and longitude with the Python data analysis library Orange (?). This software also interprets FIPs codes. This software allowed us to determine the roles of specific features and traffic incidents such as Cloverleaf Intersections which we focus heavily one.

For initial analysis and exploration of the data we utilized scatter plots to quickly search for parameter relationships by eye. figure 1 demonstrates that areas of high traffic have very few high acceleration events, while the majority of high acceleration events occur on low traffic roads.

A scatter plot of the number of incidents by State/Province is shown in figure 2 on the next page.

Many of the tables were joined on their geographic coordinates truncated to the third decimal place, resulting in a more comprehensive and actionable dataset.

We plot local features by their latitude and longitude with the Python data analysis library Orange (?). This software also interprets FIPs codes. For nation wide and state level plotting we utilize the plotting library Vincent (?).

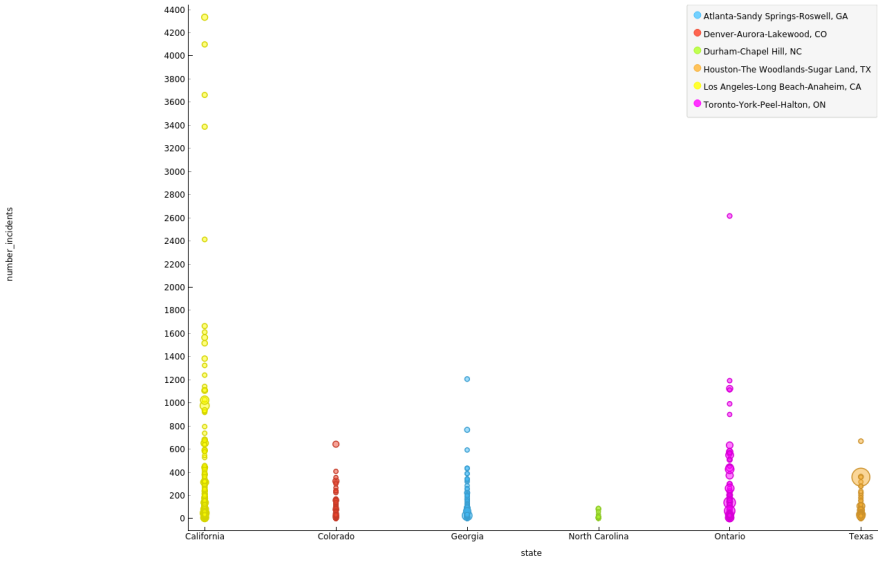
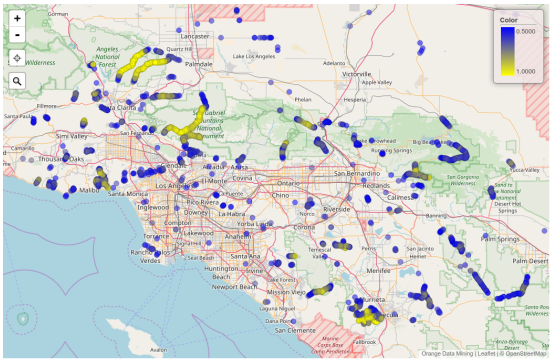


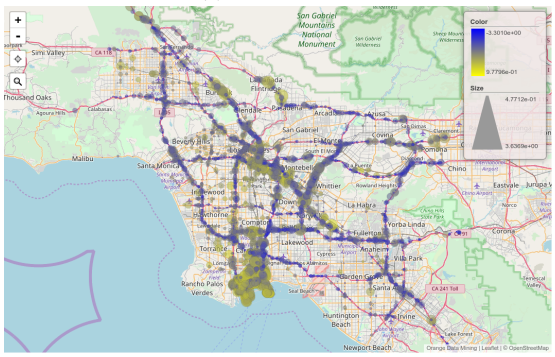
Fig. 2. Number of incidents by state/province. The size of the circle corresponds to the severity.

Fig. 3. Cellular Deadzones and Traffic Incidents are Disjoint

(a) Cellular Dead Zones



(b) Traffic Incidents



5 ANALYSIS

5.1 Cellular Dead Zones

Initially we hypothesized that cellular dead-zones might contribute to accident impediments due to distracted driving and loss of navigation data, or have a close relationship to idle times (where people are waiting for their route to load). However, geographic plots (e.g. in Los Angeles, figure 3) demonstrate that either data on cellular deadspots is disjoint from our other datasets, or that there is just no deadzones in high traffic regions. We are therefore lead to neglect cellular coverage throughout our analysis.

Hypothesis 1 (Cellular Dead Zones)**.** Cellular Deadzones lead to higher accident rates, caused by drivers distracted due to losses of navigation data.

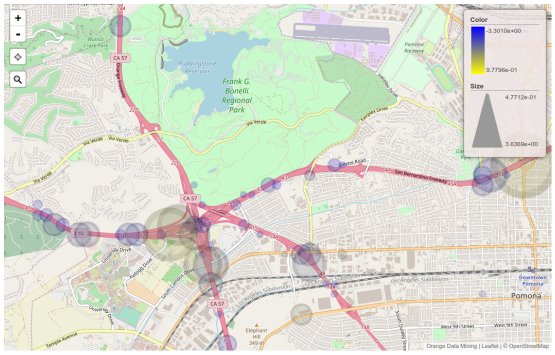
Conclusion 1 (Cellular Dead Zones)**.** The lat-lon coordinates for cellular deadzones and traffic accidents are more or less entirely disjoint, making it infeasible to provide a conclusive answer to hypothesis (1).

Recommendation 1 (Cellular Dead Zones)**.** City planners may benefit from additional data comparing traffic accidents and cell coverage in order to quantify any connections between these two variables.

Fig. 5. Dangerous Cloverleaves – Marker Size indicating number of incidents, Severity from blue (low) to yellow (high)

(a) Toronto Danger Zone Near Pearson

(b) A Clover Interchange in LA



We end with the remark that, while drivers should certainly not be using their devices while they are driving, the reality is that they do – if driving accidents are closely related to loss of cell services, we could save lives by improving cell coverage – there is therefore value in collecting additional data.

5.2 Cloverleaf Highway Interchanges

It seems intuitively clear that cloverleaf highway interchanges would be extremely dangerous, due to the aggregate dangers imposed by the complexity of the interchange, the speed of the vehicles, and the high traffic throughput.

Hypothesis 2 (Cloverleaf Intersections)**.** Highway coverleaf intersections are dangerous.

We can qualitatively verify these dangers via a few geographic plots, illustrative of the general trend. For instance, we show in figure 5 an example of a number of interchanges in downtown Toronto near Pearson airport, as well as a more upclose example from Los Angeles. These images should provide some qualitative credibility to the idea that these cloverleaves are naturally dangerous.

From the available data it is difficult to perform a more rigorous assesment of these interchanges, particularly because the interchange types would need to be hand labeled. Furthermore, these interchanges provide a very high throughput for vehicles (in comparison to, for example, light-controlled intersections), and there is therefore a natural tradeoff between safety and traffic throughput.

Conclusion 2 (Cloverleaf Intersections)**.** We make a *preliminary and qualitative* conclusion that indeed, highway cloverleaf intersections are the most dangerous locations for traffic.

In the context of smart cities, this brief analysis suggests that self-driving car manufacturers need to pay particular attention to building algorithms capable of safely navigating these complex interchanges. Furthermore, it is important for infrastructure engineers and city planners to ensure the requisite infrastructure (e.g. road side cell towers) are available for smart vehicles.

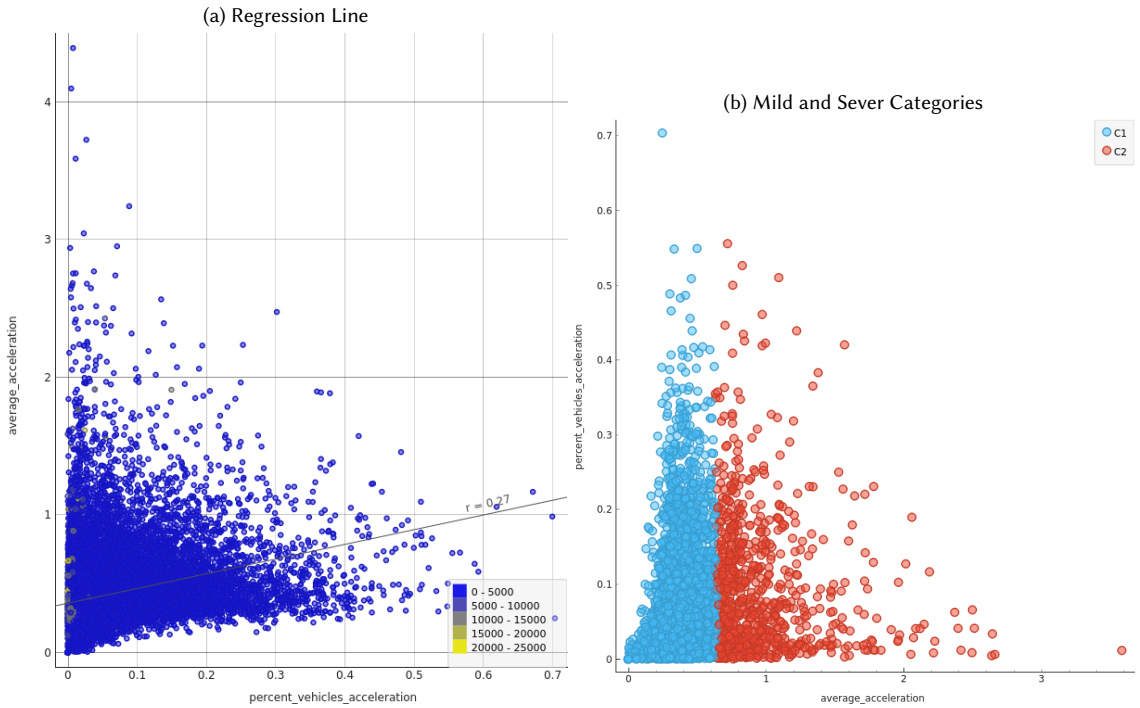
Recommendation 2 (Cloverleaf Intersections)**.** Self driving car manufacturers should focus effort on safe navigation of cloverleaf intersections. Furthermore, city planners have a role to play in ensuring that all of the requisite infrastructure is available and well maintained at these dangerous junctures.

We conclude this section by remarking that a long term solution may actually be to remove or replace these intersections entirely. However, our data is not capable of providing us any information about this possibility.

5.3 Connecting Road Impediments and Traffic Incidents via Generalized Linear Models

General Linear Models (GLMs) provide an interpretable and easy to use framework for attempting to quantify the response of an exogenous variable given information about multiple other (endogenous) variables. Our goal here is to connect both road impediments, and car types to the severity and frequency of driving incidents.

Fig. 7. Acceleration Magnitude vs Acceleration Frequency



For instance, we can model the number of traffic incidents (count data) via a Poisson or NegativeBinomial GLM.

5.4 Predicting the Severity of Incidents

A natural desire of traffic planners is to identify highly dangerous locations of road in a consistent and automatic manner.

Hypothesis 3 (Incident Severity)**.** We may predict traffic incident severity given the traffic type.

We attempt to predict incident severity at a given location with both K nearest neighbors and linear regression on fields *average_acceleration,percent_vehicles_accelerations,average_monthly_vehicles, percent_car,percent_mpv,percent_ldt,percent_mdt,percent_hdt,percent_other,car_inc mpv_incidents,ldt_incidents,mdt_incidents,hdt_incidents,other_incidents,number_incidents.* We utilize a random subset of 2/3 of the data as a training dataset and the remaining 1/3 as a test set. Results are given in table 2 on the following page and demonstrate an ability to accurately predict incident severity.

Conclusion 3 (Incident Severity)**.** kNN is capable of providing accurate predictions of incident severity.

kNN improved performance over linear regression hints at a nonlinear relationship between the features and incident severity. It is likely that more advanced techniques could demonstrate improved prediction performance.

Recommendation 3 (Incident Severity)**.** Incorporate traffic severity predictions into the workflow of traffic engineers. Locations of high incident severity should be flagged, studied and action should be taken to reduce the number and severity of incidents.

5.5 Vehicle type and severity score via Lognormal Linear Models

It is intuitive to think that the type of a vehicle involved in an incident has a non-trivial relationship with the severity of the impact. This comes from our natural instinct and Newton's second law and momentum: a larger mass imposes a great force.

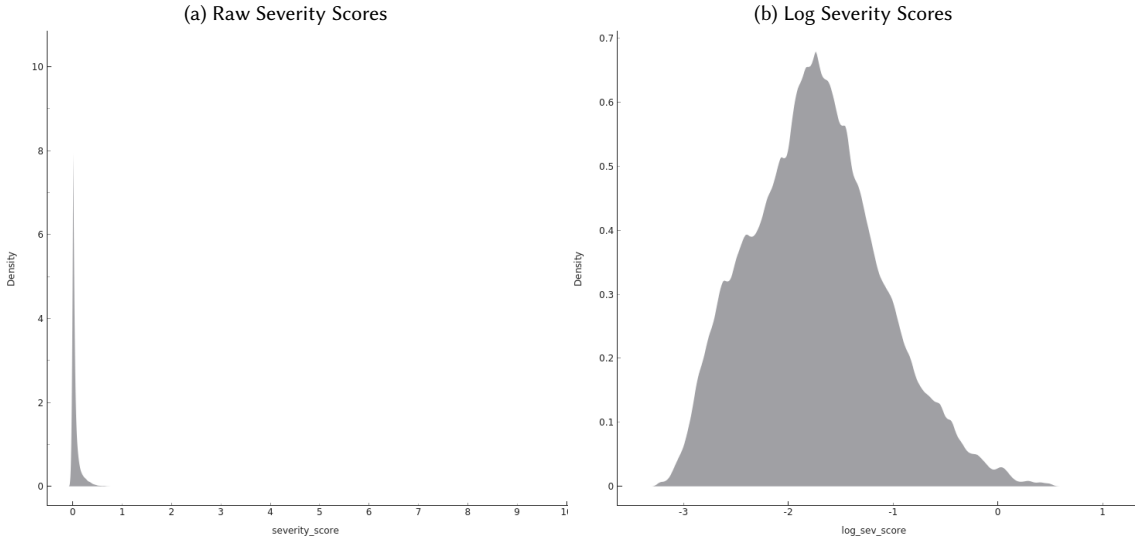
We thus attempt to establish a statistical significance between vehicles involved in an incident and the severity of the incident. Firstly, it is important to note that severity scores are distributed lognormal. That is: we cannot simply perform a simple linear regression on the raw severity scores.

Table 1. My caption

Method	MSE	RMSE
kNN	0.015	0.123
Linear Regression	0.034	0.185

Table 2. Results of kNN and Linear Regression estimates of accident severity.

Fig. 9. Severity Score log transformation



We thus take the lognormal of severity scores and regress the percentage of cars, multi-purpose vehicles, light, medium and heavy delivery trucks.

```
Call:
lm(formula = log(haz$severity_score) ~ car + mpv + ldt + mdt +
    hdt)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9707 -0.9468  0.0226  0.9325  5.5468

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.57066    0.07118 -50.165  < 2e-16 ***
car         -1.55557    0.09874 -15.754  < 2e-16 ***
mpv         -1.24392    0.11313 -10.995  < 2e-16 ***
ldt         -1.84861    0.09561 -19.335  < 2e-16 ***
mdt          0.14200    0.10942   1.298  0.19439
hdt          0.27689    0.08442   3.280  0.00104 **
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 1.378 on 13333 degrees of freedom
```

```
Multiple R-squared:  0.1271,    Adjusted R-squared:  0.1268
F-statistic: 388.2 on 5 and 13333 DF,  p-value: < 2.2e-16
```

To verify our parameters to produce a proper inference, we check the **Variance Inflation Factor (VIF)**:

```
> vif(sev.lm)
      car       mpv       ldt       mdt       hdt
2.063022 1.612827 2.184010 1.751789 3.903792
```

Our VIF scores are well below the general threshold of 10, thus we may proceed with verifying the residuals for normality.

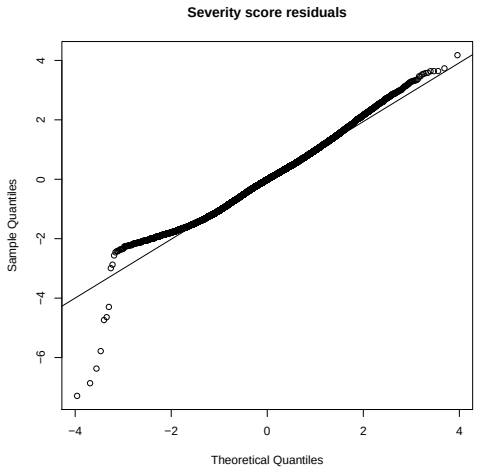


Fig. 11. The QQ-plot for the residuals from the lognormal linear regression.

Our residuals are acceptably normal from the Q-Q plot therefore we can proceed with inferencing. Note we must first map our estimations of our parameters back to some interpretable value for the severity score, that is we take we take

```
> 10^abs(coef) * coef / abs(coef) / 100
[1] -0.35939332 -0.17535575 -0.70568356  0.01386756  0.01891864
```

(we divide by 100 since our original values were in the domain $[0, 1]$). We can interpret these as the estimated change in severity score for every 1% increase in the proportion of each vehicle involved in the crash.

Our statistically significant variates are from the **percentage of cars, multi-purpose vehicles, light-delivery trucks, and heavy-delivery trucks**. As predicted, the lighter vehicles such as cars, multi-purpose vehicles and light trucks are negatively correlated with the severity score, and heavy delivery trucks and unsurpisingly positively correlate with the severity score.

We can therefore **conclude that incidents between heavy trucks and heavy vehicles incur a far greater cost and significant efforts and attention must be made to avoid incidents between these heavy vehicles**.

5.6 Connecting Road Impediments and Traffic Incidents via Generalized Linear Models

General Linear Models (GLMs) provide an interpretable and easy to use framework for attempting to quantify the response of an exogenous variable given information about multiple other (endogenous) variables. Our goal here is to connect both road impediments, and car types to the severity and frequency of driving incidents. For instance, we can model the number of traffic incidents (count data) via a Poisson or NegativeBinomial GLM.

We want to take information from various road impediments such as the percentage of each type of vehicle and the average acceleration and regress that onto the number of incidents.

We first randomly sampled 10000 road impediments from our dataset to so that we can match each road impediment with each hazardous driving area data point. We then took all data points below the 95th percentile for number of incidents to remove any extreme outliers that will skew our regression.

```
> quantile(hazroad$number_incidents, c(0, 0.25, 0.5, 0.75, 0.95, 1))
  0%  25%  50%  75%  95% 100%
   3    3    5    6   35 1565
```

Note that percent_other was not included in the model since it is exactly linear dependent to the other percentages (i.e. it is $1 - \sum percent_{type}$). (Aside: a negative binomial GLM may be more appropriate here if we see that the majority of areas have 0 incidents; it's possible the data source omitted 0-valued number of incidents data points which could bias our data).

```
Call:
glm(formula = haz$number_incidents ~ haz$average_monthly_vehicles +
    haz$percent_vehicles_acceleration + haz$average_acceleration +
    haz$percent_car + haz$percent_mpv + haz$percent_ldt + haz$percent_mdt +
    haz$percent_hdt, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2574  -1.4032  -0.6999   0.0809   6.9374

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                       1.923e+00  8.580e-02  22.408  < 2e-16 ***
haz$average_monthly_vehicles     -4.780e-06  6.499e-06  -0.736  0.46203
haz$percent_vehicles_acceleration 5.165e-02  1.550e-01   0.333  0.73903
haz$average_acceleration          1.002e-01  3.730e-02   2.688  0.00719 **
haz$percent_car                  -7.310e-01  1.732e-01  -4.220 2.45e-05 ***
haz$percent_mpv                   1.198e-01  1.195e-01   1.003  0.31584
haz$percent_ldt                  -1.417e-01  1.251e-01  -1.132  0.25769
haz$percent_mdt                   1.019e-01  1.572e-01   0.648  0.51679
haz$percent_hdt                  -2.090e-01  1.037e-01  -2.015  0.04392 *
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2999.0  on 9951  degrees of freedom
Residual deviance: 2953.4  on 9943  degrees of freedom
AIC: 6299.8
```

To ensure we can properly interpret the parameter values, we check our **Variance Inflation Factor (VIF)**

```
> vif(num_inc.glm)
     haz$average_monthly_vehicles haz$percent_vehicles_acceleration
                         1.074890                          1.094782
          haz$average_acceleration                   haz$percent_car
                         1.073778                          1.331572
                  haz$percent_mpv                   haz$percent_ldt
```

```
                    1.887499                    2.214344
          haz$percent_mdt             haz$percent_hdt
                    1.400960                    2.377371
```

The VIFs are well below the general threshold of 10, thus our paremeters are statistically valid.

We see that with a signficance level of $\alpha = 0.05$, only three covariates stand out from the crowd: **average acceleration, percentage of cars, and percentage of heavy delvery trucks**.

We see that the number of incidents is significant positively related to the average acceleration after correcting for all other variates, where a unit change in acceleration (m/s^2) results in 0.1 mean response change in in the number of incidents. **This affirms the intuition that high magnitudes in acceleration (whether it be positive or negative) results in more incidents**.

We see that the percentage of cars after correcting for all other variates has a significant negative relationship with number of incidents. The actual parameter value is small (a 1% change in the percentage of cars admits a decrease in 0.0073 number of incidents) so there is no meaningful conclusion we can draw in terms of the percentage of each type of vehicle and its effect on the number of incidents, *at least from the impediments data* (this insignificance of the type of vehicle with respect to the number of incidents is further bolstered by the fact that the other percentages of vehicle have no statistically sigificant relationship).

6 CONCLUSION AND RECOMMENDATIONS

From the analyses above, our team spent a fair bit of time thinking of potential solutions and next steps as well. Data analysis and processing lacks purpose, after all, if nothing worthwhile and actionable comes of it.

One potential solution was the introduction of a synthetic impediment in order to slow down vehicles, thereby reducing the frequency and intensity of any accidents. Further data to help confirm this axiom would involve labels of the type of road impediment, and extra analysis on the results the impediment has on acceleration as well as accident data in the impediment's vicinity.

Since it was concluded that clover roadways (entrances to highways) have a high concentration of accidents, we postulate that the introduction of speed bumps as road impediments just at the entrance of clover leaf interchanges would enforce safe merging into a highway and drastically improve safety statistics at these locations. It is intuitively clear as well as anecdotally confirmed that trucks are awful at entering highways safely, especially due to their large frames which makes it very difficult to turn properly at high speeds. The introduction of speed bumps will force not only these large trucks, but any vehicle, to slow down and safely merge into a highway. The expected result is that the frequency of accidents at highway clover interchanges will drastically reduce.

When analyzing any dataset it is important to understand who, how and what data was taken. Data for this competition was supplied by GeoTab(**?**). Geotab's primary market is companies with fleets of trucks. These companies insert tracking units in their trucks and data is acquired from these units. Although not mentioned in the competition handout this is almost certain to bias the data towards a focus on truck related incidents. Consequently many of our results concerned the role of trucks in traffic incidents. It would have been preferable to have a more varied dataset to draw stronger conclusions from other vehicle based traffic.

REFERENCES

[n. d.]. All Countries Compared for Transport > Road > Motor Vehicles per 1000 People. http://www.nationmaster.com/country-info/stats/Transport/Road/Motor-vehicles-per-1000-people. ([n. d.]).

[n. d.]. Fleet Management - GPS Vehicle Tracking. https://www.geotab.com/. ([n. d.]).

[n. d.]. Mohan D. Road Traffic Injuries—a Neglected Pandemic. Bull World Health O - Google Search. https://www.google.ca/search?q=Mohan+D.+Road+traffic+injuries%E2%80%94a+neglected+pan 8. ([n. d.]).

2008. *Costs of Road Congestion in the Greater Toronto and Hamilton Area: Impact and Cost Benefit Analysis of the Metrolinx Draft Regional Transportation Plan.* Greater Toronto Transportation Authority.

Daniel J. Fagnant and Kara Kockelman. 2015. Preparing a Nation for Autonomous Vehicles: Opportunities, Barriers and Policy Recommendations. *Transportation Research Part A: Policy and Practice* 77 (July 2015), 167–181. https://doi.org/10.1016/j.tra.2015.04.003

Federico Guerrini. [n. d.]. Traffic Congestion Costs Americans $124 Billion A Year, Report Says. https://www.forbes.com/sites/federicoguerrini/2014/10/14/traffic-congestion-costs-americans-124-billion-a-year-report-says/. ([n. d.]).

Ezra Hauer. 1997. *Observational Before/After Studies in Road Safety.* Emerald Group Publishing Limited.

Akio Hosaka and Hiroyuki Mizutani. 2000. IMPROVEMENT OF TRAFFIC SAFETY BY ROAD-VEHICLE COOPERATIVE SMART CRUISE SYSTEMS. *IATSS Research* 24, 2 (Jan. 2000), 34–42. https://doi.org/10.1016/S0386-1112(14)60027-3

E. G. Krug, G. K. Sharma, and R. Lozano. 2000. The Global Burden of Injuries. *American Journal of Public Health* 90, 4 (April 2000), 523–526.

Gregor Maier, Fabian Schneider, and Anja Feldmann. 2010. A First Look at Mobile Hand-Held Device Traffic. In *Passive and Active Measurement (Lecture Notes in Computer Science)*. Springer, Berlin, Heidelberg, 161–170. https://doi.org/10.1007/978-3-642-12334-4_17

7 APPENDIX

7.1 Predictive Modelling