# Predicting Driving Risks in Fulton County
## Final Project

Paul Horton, MacMillan Jacobson, Candice Djorno

April 2023

ISYE 6416 Computational Statistics
Professor Yao Xie

Georgia Institute of Technology
School of Industrial and Systems Engineering

# 1 Abstract

Large cities require robust transportation infrastructure to operate efficiently. In many cities, such as Atlanta, public transportation is limited and residents rely on driving for daily transportation. Driving brings safety risks for the passengers as well as potential unexpected delays due to traffic accidents. In the following paper, we utilize traffic accident data with 95,000 observations for Fulton County, GA from 2020 to 2021 with variables such as time of day, road condition, and location. Our goal is to accurately predict the severity of an accident based on regularly available data. We consider the value of this approach to be two sided: individuals will better understand the risks driving in certain conditions and the Department of Transportation could better manage resources with the additional information. We take a novel approach by implementing an EM-GMM algorithm to identify accident hot-spots and consider the proximity of these hot-spots as factors in our model. Additionally, we explore methods to manage the imbalanced classes for accident severity. We use a random forest model as our classification algorithm after finding other researchers were successful with it. Our results were disappointing as we find the predictive quality too low to be useful in a practical setting.

# 2 Introduction

Traffic accidents have a high cost for the individuals involved and a secondary impact on those around as traffic congests and commutes lengthen. Atlanta is especially susceptible to traffic disruptions as the city lacks the public transit infrastructure and adoption of other large cities. In a survey from 2016, Atlanta had the second highest average commute time out of the 20 largest cities in the US [Parker(2019)] Another more recent survey tracked the average commute for Atlanta drivers to be 32 minutes [Dulcio(2023)]. In this paper, we use data from the Georgia Department of Transportation (GDOT) [GDOT(2021)] to predict the severity of a traffic accident based on factors such as road condition, lighting, time of day, and location. The goal is that this can provide information to guide the resource allocation of the GDOT to better prepare for and manage accidents to reduce the impact on other drivers. We additionally take the perspective of an individual driver to analyze the risk associated with driving in hazardous conditions. To accomplish this, we use a combination of supervised and unsupervised learning. Predicting accident severity and identifying contributing factors has been done for other cities using datasets with other variables [Abellán et al.(2013)], [Vanishkorn and Supanich(2022)]. However, our approach is novel because we consider accident hotspots by estimating the distribution of location with a EM-GMM model. We use the proximity of the accident location to these hot spots as an independent variable in our predictive models.

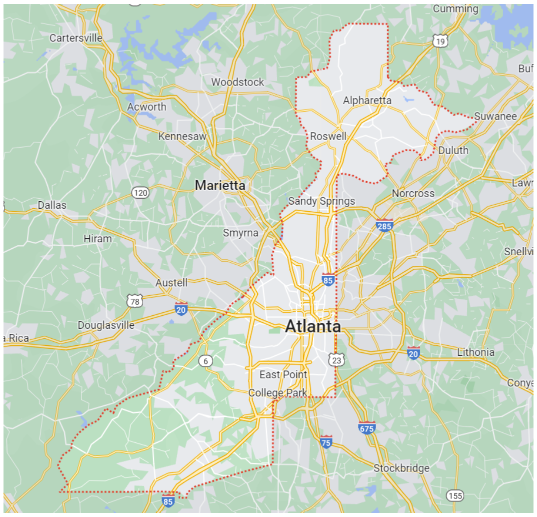


Figure 1: Map of Fulton County, GA

# 3 Related Work

Several studies focus on predicting the risk of traffic accidents as well as the degree of injury in car crashes. The analyzed data is from different countries, for diverse age groups and using a variety of machine learning techniques. [Alkheder et al.(2016)] used an artificial neural network and k-means clustering to predict the severity of injury of traffic accidents in Abu Dhabi. [Chen and Chen(2020)] compared logistic regression, classification and regression tree, and random forest to model road accident severity in Taiwan. They concluded that random forest gives the most accurate predictions. In fact, according to [Santos et al.(2022)], random forest seems to be the best approach to model car crash injury compared to support vector machine, decision tree and k-nearest neighbor, based on 56 studies. [Hasan et al.(2022)] studied the performance of support vector machine, random forest, Catboost, light GBM, and XGBoost to predict the severity of injury of workzone accidents in New Jersey, and found that random forest and Catboost resulted in the best predictive models. We conclude from these studies that random forest is likely to give the best results, and hence we decide to use it to predict the severity of outcome in our paper.

A number of publications analyzed data from the Georgia Department of Transportation (GDOT). [Daniel et al.(2000)] studied the relationship between fatal car accidents and work zones, using statistical tests. [Dai(2012)] applied spatio-temporal clustering on GDOT data to determine zones of risk of injury for pedestrians involved in road crashes. [Dai et al.(2010)] identified clusters of crashes based on spatial clustering of data from the area of Georgia State University (downtown Atlanta). In our project, we will only consider locations to identify hotspots of road accidents, using the Expectation-Maximization algorithm, which we believe hasn't been applied to this type of problem. However, this technique has been used in combination with k-means to detect crime hotspots [Appiah et al.(2022)], which shows this is a feasible approach.

# 4 Data

## 4.1 Data Collection

We collected data from the Georgia Department of transportation, which consists of approximately 95,000 observed traffic accidents from 2020-2021 in Fulton county in Georgia. The raw data contained descriptive attributes about the accidents not all of which were useful for our project. It would be ideal to have an expert review the data to remove unrelated variables however we could confidently keep a subset of them without losing valuable information. Therefore, the independent variables that we will use in our analysis are weather, surface and lighting conditions, season, time of day, day of week, holiday, daylight saving, latitude and longitude of accidents. The dependent variable is the severity of injury of which there are 5 values: Suspected Major Injury (A), Suspected Minor/Visible Injury (B), Possible Injury / Complaint (C), Fatal Injury (K), No Injury (O), Unknown (U).

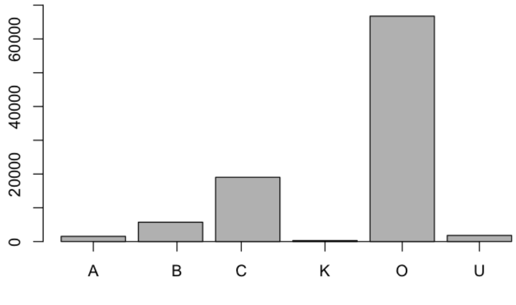


Figure 2: Distribution of Observations by Outcome

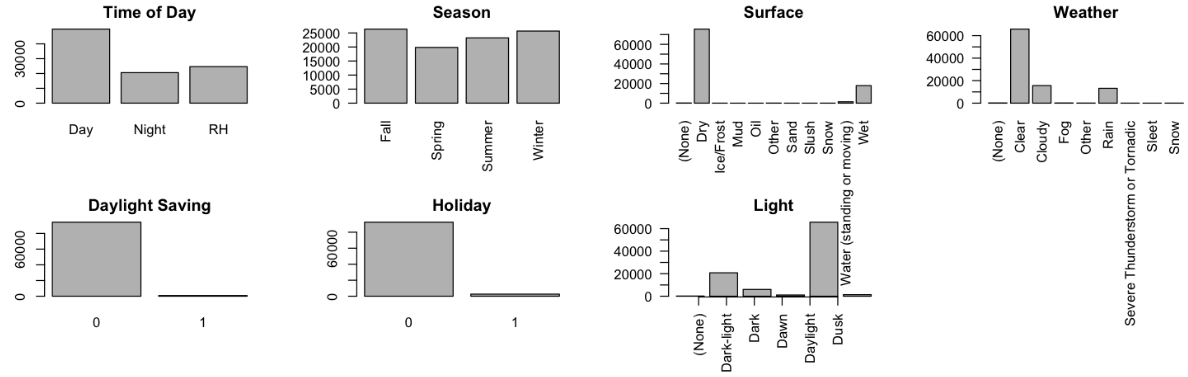
The raw data contained the roadway/intersection name, the city and county. However, it was insufficient for determining a distance for the EM-GMM algorithm so we used the ArcGIS Python API to map the intersection to a longitude and latitude.

For the categorical variables, we kept the road conditions values as they were given in the raw data. From the date and time data, we created several categorical variables including one variable with 7 values for the day of week, and another two binary variables for the daylight savings and holidays, serving as indicator variables. We reduced the months to its season. Finally, since the time of day is not a continuous variable, but it contains too many discrete values, we created three buckets, namely "day", "night" and "rush hour".

## 4.2 Data Exploration

To further understand the variables that we would work with, we conducted some data exploration. The distribution of observations by outcome can be seen in Figure 2. Approximately 70 percent of our observations resulted in no injury so we will consider methods for managing class imbalance, in the modeling part. We can see in Figure 3 the distributions of the observations for the categorical variables. Notice that more accidents happen during the day than at night or rush hour. The number of accidents are approximately the same for each season. The road conditions that are most prone to accidents are: dry and wet surfaces; clear, cloudy and rainy weather; daylight, dark-light and dark lighting.

For the location data, we noticed outliers where latitudes and longitudes had been assigned outside of Fulton County. Therefore, we removed those observations by setting horizontal and vertical thresholds at the boundaries of the area considered in this project. Figure 4 shows the accidents displayed on the map, based on their coordinates.



(a) Time and Day

(b) Road Conditions

Figure 3: Barplots of Categorical Variables

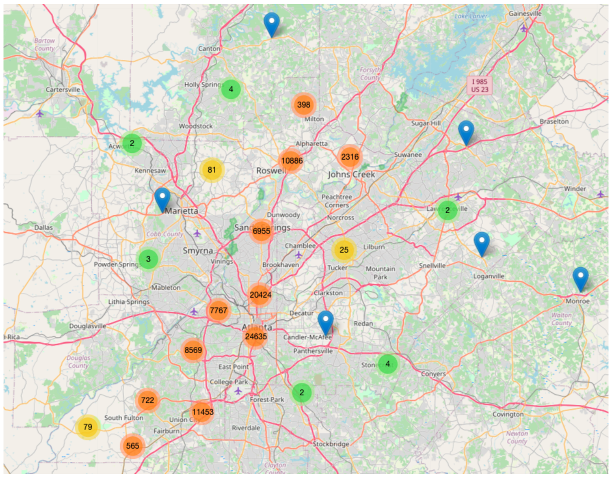


Figure 4: Map of Fulton County, GA with Coordinates of Accidents

# 5 Methodology

## 5.1 Hot Spot Assignment

It is generally understood that increased traffic is associated with an increase in traffic accidents. Thus, being closer to these hot spots increases the likelihood of being involved in a traffic accident. In their paper, [Zahran et al.(2021)] utilize four different methods with the severity and number of accidents to generate accident hot spots. Our approach is unique in that we use the coordinates of the accident as inputs to the EM-GMM algorithm to find the location and distribution of hotpots. This enables us to assign a probability of an accident to belong to a cluster. Without any prior knowledge of the number of clusters, we consider this as a tuning parameter for our model. To find the optimal number of clusters for an EM-GMM algorithm, [Fraley and Raftery(1998)] propose using the BIC metric given by the equation:

$$BIC = 2 \log \mathcal{L} - m \log (n)$$

In the above equation, $\mathcal{L}$ is the likelihood of our observations given the clusters, $m$ is the number of parameters or in this case clusters, while $n$ is the number of observations in the dataset. We evaluated the number of clusters ranging from 1 to 20 over 5 random initializations while taking the average log-likelihood for the input to our BIC calculation. With this method, we found four as the optimal number of clusters. Figure 5 shows the plot of all accidents in blue along with the cluster means in red. Intuitively, we see all the clusters within the boundary of Fulton county and roughly evenly dispersed.

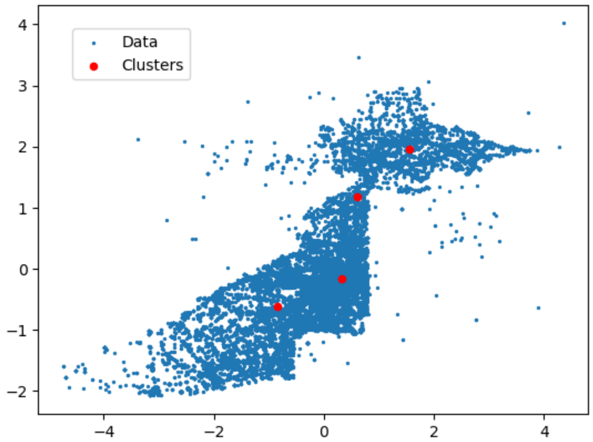


Figure 5: Traffic Accident Data with Cluster Means

We encountered scenarios as we increased the number of clusters where a cluster would converge to an isolated, tightly packed group of observations. As the points were dense, the covariance of this local cluster would be relatively small. Combine this with the oval shape of the data distribution and we had points that were many standard deviations away which resulted in probabilities of distant points belonging to the tightly packed clusters as zero due to the precision limitations of how computers store numbers. This caused an error when calculating the log-likelihood. To remedy this, we added a $0.1I$ buffer to our covariance matrix where $I$ is the identify matrix.

We used the predicted probability of belonging to a cluster along with a threshold to assign a value for a binary indicator variable for each cluster. In this scenario, if the probability of assignment was above the threshold, the cluster variable received a ”1”, and a ”0” otherwise.

We confirm our predicted clusters using the visualization from Kernel Density Estimation. We used the standard Gaussian kernel with the following rule of thumb for the optimal bandwith $h \approx 1.06\hat{\sigma}n^{-1/5}$. In Figure 6, we have one cluster centered at the densest location for accidents while others are dispersed along the length of the county shape.

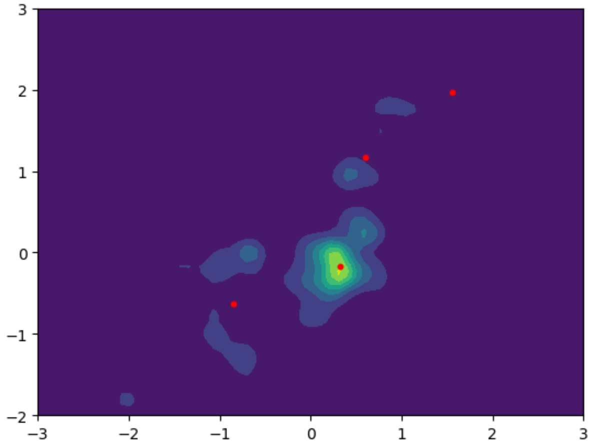


Figure 6: KDE Plot Overlayed with Cluster Means

## 5.2 Classification

After identifying our clusters and assigning accidents to their corresponding clusters, we built a model that can, based on several variables, predict accident severity. The variables in question are all categorical variables and include the weather conditions, surface conditions, light conditions, season, time of day bucket, day of the week, daylight savings boolean, holiday boolean and a cluster assignment if one is appropriate.

We decided to use a random forest to build our model since in recent years random forest's popularity has grown as it has proven to be an excellent algorithm for classification problems. In building our model, however, we needed to address the unbalanced class set in our outcome variable as described above in the data section of this report. It is clear from the bar chart above that there is a heavy skew to the "(O) No Injury" class in our dataset. As such, we employed synthetic minority oversampling technique (SMOTE) to assuage this issue. SMOTE essentially builds new, synthetic tuples that are similar to the real data found in the data set. By doing this, it builds a new, larger, data set with a balanced class set. We then trained the model on this new synthetic data set, but when assessing the model's performance, checked it against the original data set. We did this because we wanted the model to perform well on real data, not synthetic data. Finally, we trained our hyperparameters by using a 5-fold cross validation with random search scheme with 5 iterations.

# 6 Results

We first built a random forest model as described above without using SMOTE. The resulting model's hyperparameters, which were found by the tuning scheme described above, were a maximum depth of 6 levels and 435 estimators for the model. By not using SMOTE, we were essentially seeing what the random forest would do without balancing our class set. The SMOTE-less model achieved an accuracy of 72%, however this was meaningless as the model also predicted "(O) No Injury" for every prediction. The confusion matrix, displayed below in Figure 7 illustrates why SMOTE must be employed in cases like ours.

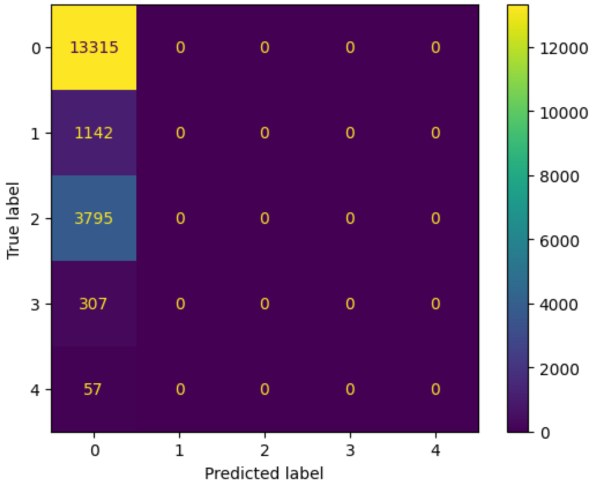


Figure 7: Confusion Matrix for RF without SMOTE

Next, we applied SMOTE to our data set and re-trained our model. Using the same hyperparameter tuning scheme as above, we found a maximum depth of 19 levels and 242 estimators for the random forest classifier. We had hoped that in using SMOTE we would be able to maintain a relatively high accuracy while minimizing false negatives in our predictions. However, this proved to be unfruitful as our accuracy dropped to 36%. The confusion matrix for our new SMOTE assisted model is displayed below.

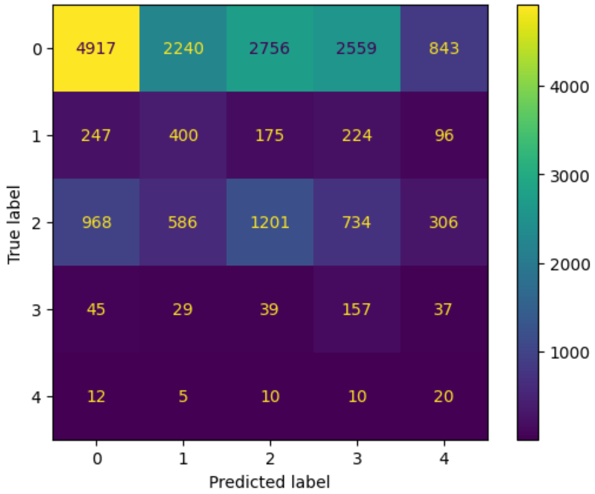


Figure 8: Confusion Matrix for RF with SMOTE

Clearly, our model did not do well in predicting the severity of an accident based on the variables defined above. This can be explained when looking at the correlations between our independent variables and the response. The heatmap in Figure 9 shows that none of the explanatory variables are correlated with the severity of outcome, denoted by ”outcome”.

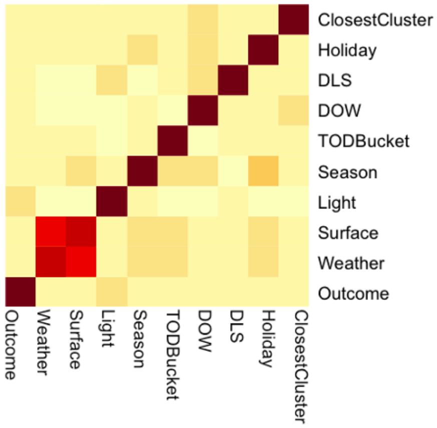


Figure 9: Heatmap of Correlations Between Variables

# 7 Conclusion

Despite early hopes of coming up with a classifier that would predict the severity of an accident based on road, time and location conditions, our final model failed to do so with high accuracy. Unfortunately, we did not find conclusive evidence that our independent variables were highly associated with severity of accident. Though we failed to build a predictive model, this report is not ultimately a failure. We learned quite a bit in conducting this project. Specifically, we learned that the variables we used are likely not predictive of the severity of an accident. Additionally, we built a novel way to assign accidents to hot spots. This way of assignment would be quite valuable in any future analysis done on this data set. For instance, in a follow-up report we may look at these same variables in the context of predicting whether or not an accident will happen rather than the severity of one.

Finally, our clustering has helped us locate the hot spots in Fulton County, as shown in Figure 10. Specifically, they are (from left to right on the image):

- The intersection of Harbin Road SW and Landrum Drive SW, in Atlanta
- Near Emory University Hospital Midtown at the intersection of Pine Street NE and Peachtree Street NE, in Atlanta
- The intersection of Glenridge Drive NE and Mabry Road NE, in Sandy Springs
- Near Haynes Bridge Road on Rock Mill Road, in Alpharetta

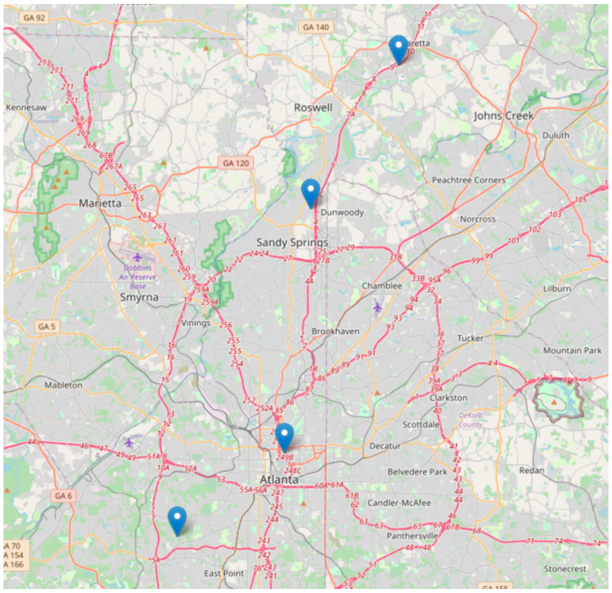


Figure 10: Hot Spots Locations in Fulton County

## 8 Work Breakdown

All three team members contributed equally to this project.

## References

[Abellán et al.(2013)] Joaquín Abellán, Griselda López, and Juan de Oña. 2013. Analysis of traffic accident severity using Decision Rules via Decision Trees. *Expert Systems with Applications* 40, 15 (2013), 6047–6054. https://doi.org/10.1016/j.eswa.2013.05.027

[Alkheder et al.(2016)] Sharaf Alkheder, Madhar Taamneh, and Salah Taamneh. 2016. Severity Prediction of Traffic Accident Using an Artificial Neural Network. *Journal of Forecasting* 36 (2016), 100–108. https://doi.org/10.1002/for.2425

[Appiah et al.(2022)] Simon Kojo Appiah, Kingsley Wirekoh, Eric Nimako Aidoo, Samuel Dua Oduro, and Yarhands Dissou Arthur. 2022. A model-based clustering of expectation-maximization and K-means algorithms in crime hotspot analysis. *Research in Mathematics* 9, 1 (2022), 2073662. https://doi.org/10.1080/27684830.2022.2073662

[Chen and Chen(2020)] Mu-Ming Chen and Mu-Chen Chen. 2020. Modeling Road Accident Severity with Comparisons of Logistic Regression, Decision Tree and Random Forest. *Information* 11 (2020). https://doi.org/10.3390/info11050270

[Dai(2012)] Dajun Dai. 2012. Identifying clusters and risk factors of injuries in pedestrian-vehicle crashes in a GIS environment. *Journal of Transport Geography* 24 (2012), 206–214. https://doi.org/10.1016/j.jtrangeo.2012.02.005

[Dai et al.(2010)] Dajun Dai, Emily Taquechel, John Steward, and Sheryl Strasser. 2010. The Impact of Built Environment on Pedestrian Crashes and the Identification of Crash Clusters on an Urban University Campus. *The Western Journal of Emergency Medicine* 11 (08 2010), 294–301.

[Daniel et al.(2000)] Janice Daniel, Karen Dixon, and David Jared. 2000. Analysis of Fatal Crashes in Georgia Work Zones. *Transportation Research Record* 1715, 1 (2000), 18–23. https://doi.org/10.3141/1715-03

[Dulcio(2023)] Jenne Dulcio. 2023. How long will it take me to commute to Downtown Atlanta? https://roughdraftatlanta.com/2023/01/25/how-long-will-it-take-me-to-commute-to-downtown-atlanta/

[Fraley and Raftery(1998)] C. Fraley and A. E. Raftery. 1998. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *Comput. J.* 41, 8 (01 1998), 578–588. https://doi.org/10.1093/comjnl/41.8.578 arXiv:https://academic.oup.com/comjnl/article-pdf/41/8/578/1032918/410578.pdf

[GDOT(2021)] GDOT. 2021. Crash Reporting Data. http://www.dot.ga.gov/GDOT/Pages/CrashReporting.aspx

[Hasan et al.(2022)] Ahmed Sajid Hasan, Md Asif Bin Kabir, Mohammad Jalayer, and Subasish Das. 2022. Severity modeling of work zone crashes in New Jersey using machine learning models. *Journal of Transportation Safety & Security* 0, 0 (2022), 1–32. https://doi.org/10.1080/19439962.2022.2098442

[Parker(2019)] Najja Parker. 2019. Atlanta -yet again- is named one of the worst places to commute by transit or car in recent ranking. https://www.ajc.com/news/world/atlanta-yet-again-named-one-the-worst-places-commute-transit-car-recent-ranking/Yfv86up06oBlC6y9YEpWKJ/

[Santos et al.(2022)] Kenny Santos, João P. Dias, and Conceição Amado. 2022. A literature review of machine learning algorithms for crash injury severity prediction. *Journal of Safety Research* 80 (2022), 254–269. https://doi.org/10.1016/j.jsr.2021.12.007

[Vanishkorn and Supanich(2022)] Buddhaporn Vanishkorn and Weeriya Supanich. 2022. Crash Severity Classification Prediction and Factors Affecting Analysis of Highway Accidents. In *2022 9th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*. 1–6. https://doi.org/10.1109/ICAICTA56449.2022.9932998

[Zahran et al.(2021)] El-Said Mamdouh Mahmoud Zahran, Soon Jiann Tan, Eng Hie Angel Tan, Nurul Amirah 'Atiqah Binti Mohamad 'Asri Putra, Yok Hoe Yap, and Ena Kartina Abdul Rahman. 2021. Spatial analysis of road traffic accident hotspots: evaluation and validation of recent approaches using road safety audit. *Journal of Transportation Safety & Security* 13, 6 (2021), 575–604. https://doi.org/10.1080/19439962.2019.1658673 arXiv:https://doi.org/10.1080/19439962.2019.1658673