Executive Summary - Impact of the Built Environment on Severity of MVCs

Amanda Curtis, Arthur Diep-Nguyen, Olti Myrtaj, Brandon Owens, Fabbio Ricci

Introduction - There are more than 5 million motor vehicle crashes (MVCs) in the United States every year. While some contributing factors are out of the hands of policy makers, others are not. In our investigation, we seek to better understand the relationship between the built environment of a community and the number and severity of crashes in that area. Features of the built environment (which include roads, crosswalks, bike paths, and traffic signals) can be augmented or replaced. By modeling the relationship between built environment features and frequency and severity of MVCs, we can better anticipate the number of severe MVCs a community may expect in certain areas. Such knowledge could help local governments and city planning organizations better anticipate and respond to MVCs with emergency services and other resources, or consider ways to lower the crash density of existing and future regions, thus making their communities safer.

Data sets - We used two data sets, one for motor vehicle crashes, and another for features of the built environment.

The motor vehicle crash data comes from US Accidents (2016-2023) on Kaggle. The data set contains approximately 7.7 million accident records spanning the contiguous United States from 2016 to 2023, scraped from Bing and MapQuest Traffic APIs and augmented with data from various entities, such as government agencies. The data set has 46 columns, with variables including accident location, time, and severity, as well as many categorical variables describing weather conditions, visibility conditions, and proximity to points-of-interest like junctions, traffic stops, roundabouts, etc.

The data about the built environment comes from the Smart Location Database from the EPA. The data set contains data for every census block group in the United States, with over 200,000 rows. The data set has over 90 variables across various categories such as housing density, diversity of land use, neighborhood design, destination accessibility, transit service, employment, demographics, etc. Notably, the data set comes not as a CSV file, but as a file geodatabase.

Data augmentation and processing - Before we could begin, we needed to combine our two data sets. One possible approach would be to add built environment variables from the EPA's Smart Location Database to each motor vehicle crash in the Kaggle data set. Such an approach would mean that our model would be observing features of crashes, then predicting the severity of the crash. However, this approach does not address the likelihood of a crash happening; it addresses only how the built environment attenuates severity, given that a crash already happened.

Thus, we adopted a different approach: Instead of attaching variables from the Smart Location Database to each crash from the Kaggle data set, we aggregated the crashes by census block group. With this approach, our model would observe features of a census block group, then predict the number of severity-weighted crashes that occurred during the timeframe when the crash data was collected.

To engineer our target variable, we took the severity variable from Kaggle to create a "severity-weighted" crash, so that our prediction model would weigh severe crashes more heavily than light crashes. We then divided by population density.

To merge the two data sets, we used GeoPandas to take the latitude-longitude data from each motor vehicle crash and convert those coordinates to the same Coordinate Reference System used by the Smart Location Database. We could then determine the census block group in which each crash occurred by using spatial join in GeoPandas.

We cleaned the data of rows with nonsensical or extreme data (e.g. census block groups with zero population, zero land area, zero roads, etc.). We chose to exclude census block groups with fewer than 4 crashes, since a census block group with fewer than 4 crashes from 2016 to 2023 probably has negligible motor vehicle activity. We also excluded crashes from the year 2020, since driving and pedestrian activity during that year was impacted by the COVID-19 pandemic lockdown.

After cleaning our data, we performed log transformations on highly skewed variables, which increased their interpretability.

Next, we sought to reduce our number of features to fewer than 20. We began by highlighting features that had a moderate correlation (roughly $0.3 < |R^2| < 0.8$) with our target variable, as indicated by correlation heat maps. We eliminated features by selecting between highly correlated variables to reduce multicollinearity.

Model selection and results - After performing an 80/20 train-test split of our data set and spending some time on exploratory data analysis, we began to train some models and compare their performances using root mean squared error (RMSE) as a performance metric. We started with Multiple Linear Regression as a baseline model and went on to consider Lasso and Ridge regression, Random Forest Regression and XGBoost Regression. For each of these models we used 5-fold cross validation while tuning hyperparameters in an attempt to minimize overfitting. In the end, XGBoost performed the best, scoring a RMSE of 0.552 on the training data. We chose to use this tuned XGBoost model as our final model, and found that it scored a RMSE of 0.547 on the test data.