Spatial Prediction and Risk Factor Analysis of Bicycle Crashes in Washington, D.C.

Zhonghua Yang

Department of Urban Spatial Analytics University of Pennsylvania, Philadelphia, USA

Abstract

This study aims to predict and analyze the impact of bicycle crashes in Washington, DC, taking into account the city's promotion of bicycling and geographic accessibility. The paper is organized into four main sections: 1) an overview of the current state and background of bicycle promotion policies and geographic bicycle access in Washington, DC; 2) the purpose of statistical calculations and variable analysis of bicycle crash data in Washington, DC; 3) the use of mathematical statistical models to derive visual results and spatial variable analysis; 4) targeted summaries and policy recommendations based on the results. Factors examined include road categories, bicycle lane distribution, traffic signs, traffic signals, and socio-demographic variables. Furthermore, the study explores the relationships between the aforementioned factors and the frequency of bicycle crashes, aiming to identify potential areas for intervention and improvement. This data-driven analysis contributes to a more comprehensive understanding of the factors influencing bicycle crashes in Washington, DC, ultimately supporting the creation of a safer and more sustainable urban transportation environment for all.

Key Words: Washington,D.C., Bicycle crashes, Analysis of variables, Data Visualization, Transportation Policy

Problem

Bicycle traffic safety has always been a primary focus of our attention. Although the construction of bike lanes and the services provided have seen significant improvements, bicycle accidents still occur from time to time. Many of these accidents are caused by the sharing of bike lanes with pedestrians and vehicles. According to data released by the U.S. Department of Transportation, Washington D.C. has a relatively high bicycle accident rate, ranking among the highest in the nation. Specifically, in 2019, over 800 bicycle accidents occurred in the city, resulting in 14 fatalities and more than 800 injuries. There are numerous factors contributing to bicycle accidents, including road conditions, enforcement of traffic regulations, cyclist education, and awareness.

Purpose

The purpose of this study is to investigate factors and impacts associated with bicycle crashes in Washington, DC, to address potential challenges in promoting cycling in the US. By analyzing bicycle accidents, the study aims to provide data and guidance for developing and improving bicycle safety policies. Additionally, it seeks to understand the risks and influential factors in cycling by examining the nature, causes, locations, and timing of accidents. This research will also help identify safety disparities among different social groups, regions, and road types, supporting targeted safety measures and promoting a safer urban cycling environment.

Method

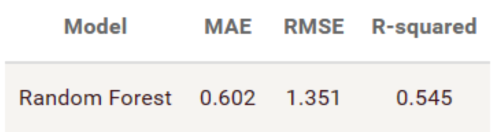
This study employs a logistic regression model to predict bicycle accidents, considering factors such as road type, traffic volume, road infrastructure, and community demographics. Road types, including major and minor roads, are transformed into dummy variables in the model. Traffic volume dimensions, such as vehicle flow, speed, and density, are considered as they impact the likelihood of bicycle accidents. Road infrastructure features, including lane width, traffic signs, and street lighting, are also analyzed. Finally, community demographic information, such as population density, age distribution, and education levels, is incorporated as input variables to improve the predictive accuracy of the model.

Result

In this study, I utilized various forms of statistical graphs to present the data characteristics and the relationships across multiple dimensions. From mathematical charts to maps, these diverse visualization methods provided a comprehensive and enriching representation of bicycle-related information.

Model Accuracy

The following table shows the performance of my random forest regression model in predicting bicycle accidents. We used three evaluation metrics: mean absolute error (MAE), root mean square error (RMSE), and R-squared to assess the accuracy of the model.



Model	MAE	RMSE	R-squared
Random Forest	0.602	1.351	0.545

Variable Importance

We generated a number of visualizations to visualize the contribution of each predictor variable to predicting bicycle accidents. The five most important variables in our results are: *percentage of commutes walking to work*, *percentage of commutes by public transportation*, *percentage of minors*, *number of traffic signals*, and *average speed of traffic*.

Prediction Error

We calculated the absolute value of the difference between predicted and observed values in each grid: 80.5% for 0, 14.8% for 1, 2.63% for 2, 1.03% for 3, 0.426% for 4, 0.308% for 5, and 0.352% for greater than 5. This indicates that 80.5% of our predictions are completely correct, which is an ideal result and shows that our model can be trusted to some extent.

Conclusion

We used a random forest model training algorithm, K-fold cross-validation and spatial cross-validation to evaluate model performance and accuracy. Our models performed well, demonstrating high predictive accuracy and robustness.

Limitations: 1. Considered only limited predictor variables. 2. Our study focuses only on Washington, DC, and may not be applicable to other cities or regions.

E-mail: yazh@upenn.edu Github Repository Personal Website

Content

1. Background and Current Status of Bicycle in Washington, D.C.

Bicycle is a highly advantageous mode of transportation in modern urban areas, offering a multitude of benefits. Firstly, cycling can significantly reduce environmental pollution and lower carbon emissions, thereby promoting a healthier urban environment. Secondly, cycling provides an economical mode of transportation that can help reduce travel costs for residents. Moreover, cycling can also alleviate urban traffic congestion, mitigating the negative impact of vehicular congestion on residents' daily lives. Globally, there is a growing emphasis on promoting and investing in bicycle transportation. Many countries and cities have recognized the importance of cycling as an integral component of urban transportation planning and have developed comprehensive bicycle lane networks and rental systems. For instance, the Netherlands and Denmark have achieved cycling rates exceeding 40%, with widespread bicycle lanes and parking facilities throughout their urban areas. [1] In Asia, some cities have also begun to actively promote cycling through the implementation of bicycle lanes and bike-sharing programs, encouraging more residents to choose cycling as a mode of transportation.

Compared to other countries, the United States still lags behind in terms of bicycle transportation. Although some cities have begun to promote bicycle transportation by constructing bike lanes and bike-sharing systems, bicycle commuting still accounts for a relatively small proportion in the US. [2] Moreover, bicycle transportation safety is also a prominent issue in the US, as the mixed traffic mode of bike lanes and vehicular traffic can easily cause accidents and casualties. This is a problem and focus that we are deeply concerned about and will discuss in the following sections. [2] Therefore, the US government needs to further strengthen support and management of bicycle transportation, improve bicycle transportation safety, and provide residents with a more convenient, healthy, and eco-friendly way of commuting.

In recent years, there has been a growing movement to promote cycling as a viable and sustainable mode of transportation in the United States. Many cities, including Washington D.C., have implemented policies and programs to encourage cycling, with varying degrees of success. Next, I will discuss the policies and programs related to cycling in Washington D.C., their impact, as well as how they compare to other cities in the United States. Washington D.C. has implemented several policies and programs aimed at promoting cycling as a mode of transportation. One such policy is the Bicycle Master Plan, which was first introduced in 2005 and has since been updated multiple times. The plan sets out a comprehensive strategy to increase cycling infrastructure, including the addition of bike lanes and the creation of a city-wide network of cycling routes. Another significant policy is the Capital Bikeshare program, which was launched in 2010 and has since become one of the largest bike-sharing programs in the United States, with over 4,000 bicycles and 500 stations throughout the city.

The impact of these policies has been significant. Since the implementation of the Bicycle Master Plan, Washington D.C. has seen a 142% increase in bicycle commuting. Capital Bikeshare has also been incredibly successful, with over 24 million trips taken since its launch. These policies have had a positive impact on the city, including reducing congestion, promoting healthier lifestyles, and lowering transportation costs for residents. [3][4]

However, there have also been some negative consequences of these policies. One of the major concerns is the safety of cyclists on the road. Despite the addition of bike lanes and the creation of cycling routes, accidents involving cyclists still occur. There have also been concerns about the impact of Capital Bikeshare on the city's transportation infrastructure, including the placement of bike stations and the use of public space.

When compared to other cities in the United States, Washington D.C. is considered to be one of the most bike-friendly cities. According to the League of American Bicyclists, Washington D.C. is ranked as a silver-level Bicycle Friendly Community, placing it among the top 10 cities in the United States for cycling. Other cities, such as Portland, Oregon and Minneapolis, Minnesota, are often cited as leaders in promoting cycling as a mode of transportation.

In conclusion, Washington D.C. has implemented several policies and programs aimed at promoting cycling as a mode of transportation, including the Bicycle Master Plan and Capital Bikeshare. These policies have had a positive impact on the city, including reducing congestion, promoting healthier lifestyles, and lowering transportation costs for residents. However, there have also been some negative consequences, such as concerns about the safety of cyclists and the impact on public space. When compared to other cities in the United States, Washington D.C. is considered to be a leader in promoting cycling, but there is still room for improvement.

Bicycle traffic safety has always been our focus. Although the construction of bicycle lanes and the services provided have been significantly improved, bicycle accidents still occur, many of which are caused by the sharing of bicycle lanes with pedestrians and vehicles. According to data released by the US

Department of Transportation, Washington DC has a high bicycle accident rate and ranks among the top in the country. Specifically, in 2019, there were over 1,000 bicycle accidents in the city, resulting in 14 deaths and over 800 injuries. There are many factors that contribute to bicycle accidents, including road conditions, enforcement of traffic rules, cyclist education and awareness, among others. For example, in Washington DC, some bicycle lanes are narrow and in poor condition, and there is no separation between bicycle lanes and motor vehicle lanes, which increases the likelihood of collisions between bicycles and motor vehicles. In addition, some cyclists have inadequate understanding and compliance with traffic rules, such as failing to stop at traffic lights when crossing the road, which also increases the probability of accidents.[5]

The management of the shared bike system in Washington, D.C. has some issues, such as disorderly parking and damages. Some studies have suggested that digital monitoring technologies, such as smart locks and location tracking, can effectively address these problems.[6] In addition, the equity of bike-sharing services needs further attention. Some studies have found that bike-sharing services receive relatively less support and service in some low-income and minority communities, and targeted measures are needed to improve their service levels. [7] These issues will also be discussed in the third section.

2. Purpose of the Study

Relevant studies have shown that the implementation of bicycle promotion policies can achieve good results. In Washington D.C., the construction of bicycle lanes and the promotion of bicycle sharing services have achieved significant results. A study found that the promotion of bicycle sharing systems has had a positive impact on the travel of residents in Washington D.C., increasing travel diversity while reducing the cost of cycling [1]. Nevertheless, reducing the probability of bicycle accidents remains a serious issue. Exploring the factors and impacts of bicycle accidents can help to clarify the many issues that bicycle promotion may face in the United States.

1. **Studying bicycle crashes in Washington DC can help understand traffic safety issues and provide basic data and guidance for developing and improving bicycle safety policies.** By analyzing bicycle accidents in Washington DC, the types and locations of bicycle accidents, as well as the main factors that cause accidents, such as vehicle speed, compliance with traffic rules, road conditions, and cyclist behavior, can be identified. This information can help policy makers identify traffic safety issues and take appropriate measures to reduce the occurrence of accidents.
2. **Studying bicycle accidents can provide a better understanding of the danger and risk factors associated with cycling by examining the nature, causes, locations, and times of these accidents.** A study conducted by the National Highway Traffic Safety Administration in the United States found that bicycle accidents are more likely to result in injury or death to the cyclist compared to other vehicle accidents. The study showed that the proportion of injuries or deaths in bicycle accidents is much higher than in other vehicle accidents [8]. In addition, the severity of injuries in bicycle accidents is often more serious, particularly in the head and neck areas. Researchers found that head and neck injuries accounted for one-third of all injuries in bicycle crashes.
3. **Studying bicycle accidents also helps to understand the safety differences among different social groups, regions, and road types, and provides support and guidance for targeted safety measures.** According to a study conducted in New York City,[9] the following characteristics were observed in bicycle accidents in the city:
    a. The main cause of accidents was traffic violations by drivers and cyclists, such as failure to yield, running red lights, and driving under the influence.
    b. Accidents occurred more frequently on busy main roads and intersections with many traffic lights, especially in areas lacking bicycle lanes.
    c. The severity of injuries and fatalities in accidents was related to the age and gender of the cyclist, as well as the income and racial composition of the community where the accident occurred.

Similarly, non-white cyclists had higher accident rates in bicycle accidents in Washington DC, indicating that the needs and safety issues of minority cyclists need to be taken into account in bicycle safety policies in Washington DC. In addition, there is an unequal distribution of bicycle road safety facilities, such as bicycle lanes and bicycle traffic signals, between affluent and poor communities in Washington DC,[10] which is also an important factor leading to bicycle crashes.

3. Brief Overview of Analysis Model

In order to predict bicycle accidents, some mathematical statistical models have been widely applied. I plan to use the Random Forest Regression model for analysis. Random Forest Regression is a supervised machine learning algorithm used for regression tasks, which involve predicting a continuous target variable. It is an ensemble learning method, which means it combines multiple individual models to create a more accurate and robust model. In the case of Random Forest, the individual models are decision trees.

Here's an overview of how the Random Forest Regression Model works:

1. **Bootstrapping:** A random forest regression model starts by creating multiple bootstrapped datasets from the original dataset. Bootstrapping is the process of randomly sampling the dataset with replacement, creating multiple new datasets of the same size as the original.
2. **Building Decision Trees:** For each bootstrapped dataset, a decision tree is constructed. During the construction of each decision tree, only a random subset of features is considered for splitting at each node. This introduces diversity among the trees and helps reduce overfitting.
3. **Aggregating Predictions:** When a new data point needs to be predicted, it is passed through all the decision trees in the random forest. Each tree makes a prediction, and the final prediction is calculated by taking the average of all individual tree predictions.

The primary benefits of using a Random Forest Regression Model include:

1. **High accuracy:** By combining the outputs of multiple decision trees, random forests generally achieve better accuracy than individual trees.
2. **Robustness:** Random forests are less prone to overfitting because they take into account the predictions from multiple trees, reducing the impact of noise in the data.
3. **Feature importance:** The algorithm can provide insights into which features are the most important for making accurate predictions.

In bicycle crash prediction, the dependent variable is usually whether the accident occurs or not, while the independent variables include multiple factors and features related to bicycle accidents. According to the survey, I selected some environmental characteristics that may be related to bicycle accidents and included them in the model for prediction. At the same time, I collected data on various features. Next, I will briefly explain the reasons for collecting this data and describe the internal structure and content of the data.

3.1. Road type data

Road type is a crucial factor that affects the prediction model of bicycle accidents. Studies have shown that road type is a key factor that influences the risk of bicycle accidents. In the prediction model, road type can be used as an independent variable or as a control variable for other variables to better understand and predict the occurrence of bicycle accidents. A study on bicycle accidents in different Canadian cities found that different types of roads have different effects on the occurrence of bicycle accidents. In particular, there are more bicycle accidents on main roads in the downtown and commercial areas compared to residential areas and smaller roads.

Road types include many different types, with major roads and minor roads having the greatest impact. A study of bicycle accidents in Washington, D.C. showed that major roads are the main places where bicycle accidents occur, followed by minor roads. Major roads are usually the main thoroughfares in city center and commercial areas, with heavy traffic and dense traffic signals, increasing the probability of conflicts between bicycles and vehicles. [11] The definition of road types is based on the U.S. Department of Transportation's highway classification system, which includes different types such as highways, freeways, main roads, minor roads, and residential streets. In the logistic regression model, road types are converted into dummy variables, with residential streets as the baseline value. By analyzing the coefficients of each road type, it was found that highways and freeways are significantly positively correlated with the probability of bicycle accidents, while main roads and minor roads show a negative correlation trend, indicating a lower probability of bicycle accidents on these types of roads.

Targeted data selection for roads and intersections also requires the use of GIS tools for filtering.

3.2. Traffic volume

Traffic volume refers to the amount of traffic passing through a road or street during a certain period, usually measured in terms of the number of vehicles passing per hour or per day. Traffic volume can reflect the congestion level and vehicle speed of the road traffic, thus affecting the safety of bicycle travel. Many studies have considered the impact of traffic volume on bicycle accidents. For example, a study on bicycle accidents in New York City found that traffic volume is one of the main factors leading to bicycle accidents [12]. Another study on bicycle accidents in Brussels, Belgium also found that traffic volume is one of the important factors affecting the incidence of bicycle accidents [13].

Traffic volume includes many different dimensions, such as vehicle volume, speed, and traffic density, all of which may have an impact on bicycle accidents. For instance, traffic congestion may result in vehicles and bicycles being in closer proximity, thereby increasing the risk of collisions with bicycles. Moreover, vehicles traveling at high speeds may pose a threat to bicycle safety, making vehicle speed another important factor in traffic flow. An increase in traffic volume during peak hours can lead to a significant increase in bicycle accidents, so considering the time factor is also important.

3.3. Road infrastructure

Road infrastructure refers to the facilities on and along the road, such as lane width, road curvature, traffic signs, streetlights, etc. Different road facilities can have different impacts on the speed, visibility, and operability of bicycles, directly or indirectly affecting the occurrence of bicycle accidents. Studies have shown that road infrastructure factors have a significant impact on bicycle accidents, with lane width, traffic signs, and streetlights having the greatest influence. Specifically, lane width is one of the key factors affecting the incidence of bicycle accidents, as wider lanes can provide riders with more freedom and safety while riding [10]. Traffic signs and streetlights are particularly important for preventing nighttime bicycle accidents, as they can increase the visibility of cyclists and the transmission of traffic information.

In urban areas, the placement of traffic signals and crosswalks may be more important as they can regulate the flow of pedestrians and vehicles and ensure the safety of cyclists. On the other hand, on non-urban roads such as highways, road width and shoulder placement may be more crucial, as these facilities can provide a safer driving space and a refuge area in emergency situations.

3.4. Population information in communities

Population information in communities plays a certain role in predicting bicycle accident models and can serve as one of the input variables in the prediction model, thereby improving the accuracy of the prediction. Community population information can reflect the population characteristics of the community, such as population density, age distribution, and educational level, which may have a certain correlation with bicycle accidents. Therefore, adding this information can help improve the accuracy of the prediction model. Taking the study in [12] as an example, the study used census data from Washington, D.C. as one of the input variables and predicted bicycle accidents using a logistic regression model. The results showed that census data had a significant impact on predicting bicycle accidents and could improve the accuracy of the prediction.

In population information in communities, characteristics such as population density, age distribution, and educational level may all have an impact on the prediction model of bicycle accidents. For example, in many studies on the behavioral patterns of cycling, variables such as population density, age distribution, and educational level in census data are often used, and it has been found that population density and educational level have a significant impact on predicting bicycle accidents. In [13], the study used variables such as population density, per capita income, and racial proportions in census data and found that population density and per capita income played an important role in predicting bicycle crashes.

This information also has important practical implications for transportation planning, which aligns well with the purpose of this study.

Firstly, transportation planning needs to take into account the population characteristics of different areas and develop more suitable bicycle transportation plans accordingly. For example, in areas with high population density, it is necessary to increase the construction of sidewalks and bike lanes to alleviate traffic congestion and improve traffic safety. In areas with an older population, efforts should be made to strengthen traffic safety education and management for elderly cyclists. In areas with a higher level of education, measures such as improving road traffic facilities and increasing awareness of traffic safety can be implemented to reduce the occurrence of bicycle accidents.

Secondly, community population information can also provide important reference data for transportation planning. By analyzing community population information, the population characteristics of different communities can be understood, providing more comprehensive and accurate data support for transportation planning. For example, when planning bike lanes, information such as the age and occupation distribution of local cyclists can be obtained through community population information to determine the focus and direction of bike lane construction. When planning traffic signals, the time interval for setting the signals can be reasonably determined based on factors such as population density and pedestrian flow in different areas.

4. Data Wrangling and Visualization

4.1. Crash Data exploration

In this section, our goal is to analyze the percentage of bicycle crashes in total crashes per year to identify the share of bicycle crashes among all crashes. We will examine the proportion of bicycle crashes in all crashes for each year. To do this, we first summarize both the total crash data and bicycle crash data by year. Next, we merge these two sets of data into a single data frame and calculate the proportion of bicycle crashes in all crashes for each year. This can provide a clear representation of the importance of bicycle crashes in traffic accidents in daily life.

We create a stacked bar chart to display the proportion of bicycle crashes in all crashes per year. In chart 1, the blue portion represents bicycle crashes, and the red portion represents other types of crashes. Additionally, we added percentage labels to each bar to provide a more intuitive understanding of the proportion of bicycle crashes each year. Through this chart, we can visually comprehend the proportion of bicycle crashes in all crashes per year, laying the foundation for further analysis and modeling.

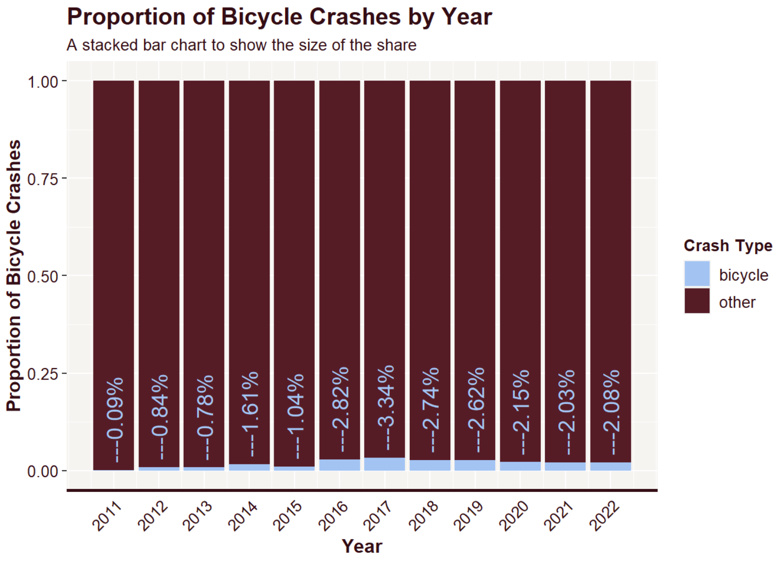


chart 1. Percentage of Bicycle Crashes Per Year

In next step, we delve deeper into the analysis of bicycle accidents, with a focus on the proportion of accidents involving injuries each year. To achieve this goal, we first extract records from the original bicycle accident dataset that involve at least one injured person (including cyclists, drivers, or pedestrians). Next, we summarize the bicycle accidents involving injuries by year and calculate the proportion of these accidents compared to the total number of bicycle accidents each year.

To effectively display this data, we create a stacked bar chart that shows the proportion of bicycle crashes with injuries in relation to all bicycle accidents per year. In chart 2, the blue portion represents bicycle accidents involving injuries, while the purple portion represents accidents without injuries. To enhance the chart’s comprehensibility, we add percentage labels to each bar, providing a clearer understanding of the proportion of bicycle accidents with injuries each year. This analysis not only reveals safety concerns related to bicycle accidents but also highlights trends and potential areas for improvement in road safety and infrastructure planning.

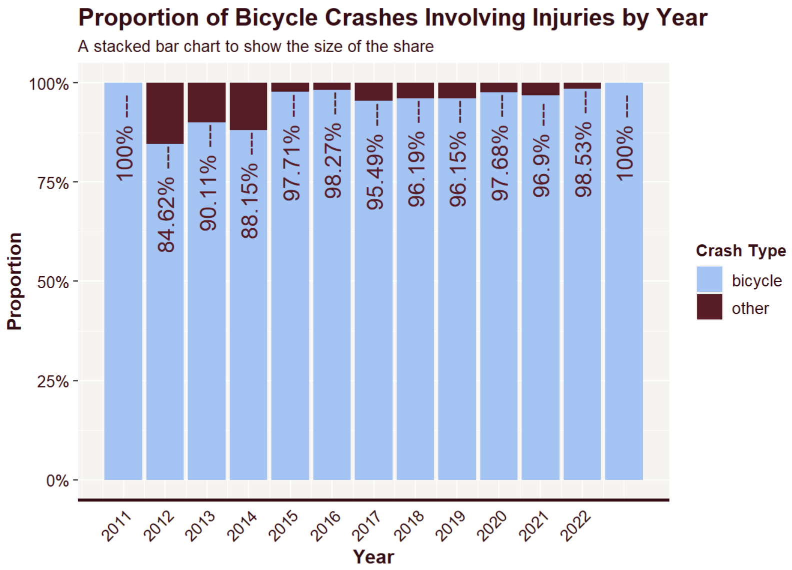


chart 2. Proportion of Bicycle Crashes Involving Injuries by Year

Based on the above two bar charts and data displaying crash proportions, we can observe the following key points:

1. Although bicycle crashes make up a relatively small proportion of all crashes, they do pose significant safety risks. This suggests that bicycle safety should be given adequate attention in research and policy development.
2. From the second bar chart, we can see that the proportion of bicycle crashes involving injuries is very high, nearly 98%. This indicates that the consequences of bicycle crashes are often severe, potentially leading to injuries or even fatalities. Therefore, improving bicycle road safety is crucial for protecting the lives of cyclists and other road users.

These observations are closely related to our research. Our goal is to establish a predictive model to understand the likelihood and associated environmental factors of bicycle crashes occurring in Washington D.C. By analyzing this data, we can better comprehend the severity and urgency of bicycle crashes, providing strong support for policymakers. Furthermore, by identifying the key factors affecting the occurrence of bicycle crashes, our model can help develop targeted policies and measures to reduce the incidence of bicycle crashes, improve road safety, and minimize the harm caused by bicycle crashes to cyclists and other road users.

4.2. Create Fishnet

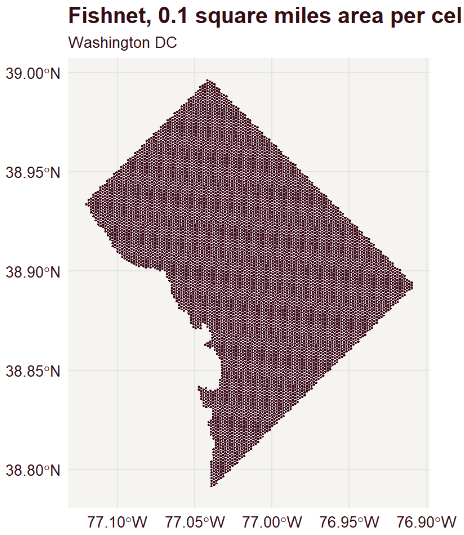


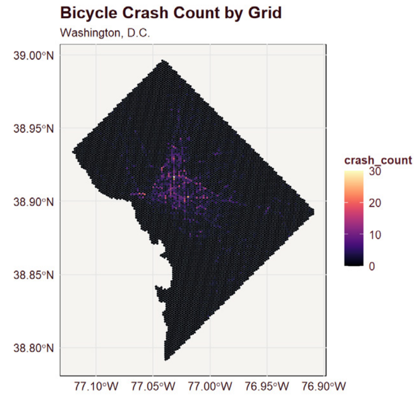
chart 3. Fishnet of Washington, D.C.

We created a fishnet grid covering the Washington D.C. area, with each grid cell having a side length of 500 meters, forming a rectangular grid. Fishnet is a common method of organizing spatial data, dividing geographic space into regular grids, which is convenient for spatial analysis and modeling. Finally, we plotted the generated grid as chart 3. By setting fill, color, and stroke parameters, we can adjust the grid's fill color, border color, and border thickness. The resulting map displays a rectangular grid covering the Washington D.C. area, with each grid cell representing an area of 0.1 square miles.

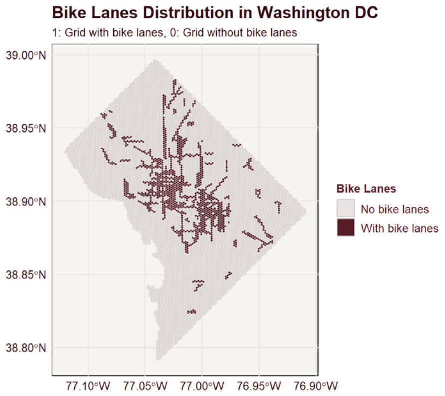
4.3. Integrate Data

In this study, creating grids and integrating data into them is of great significance. First, it allows us to combine various spatial datasets such as traffic signs, traffic signals, street information, turning movement information, and census data into a unified spatial reference for subsequent spatial analysis and modeling. This helps to investigate the relationships between different factors and bicycle crashes, providing a basis for the government to formulate more effective traffic policies. When building the random forest model, integrating data into grids helps us better process and analyze these spatial data. The random forest model requires inputting multiple feature variables to predict the target variable. By integrating various spatial datasets into the grids, we can use these grids as observation samples, with the various variables within the grid as feature variables input into the model. In this way, the model can learn the relationships between different spatial features and bicycle crashes, thereby predicting the probability of bicycle crashes occurring in specific areas.

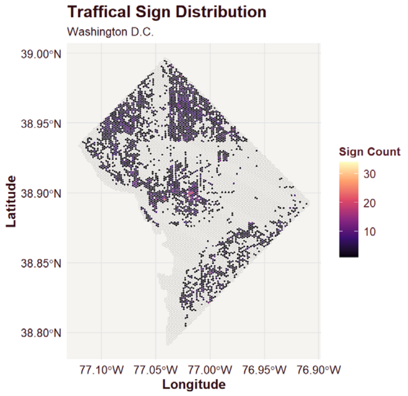
In summary, creating grids and integrating data into them helps to combine various spatial datasets into a complete dataset, facilitating spatial analysis and modeling. Through the random forest model, we can discover the relationships between different spatial features and bicycle crashes, providing the government with recommendations regarding road design, traffic management, and improvements to bicycle facilities, in order to reduce the risk of bicycle crashes. I showed four maps, the distribution of the number of bicycle crashes in the grid data, the distribution of bicycle lanes, the number of traffic signs, and the number of traffic signals.



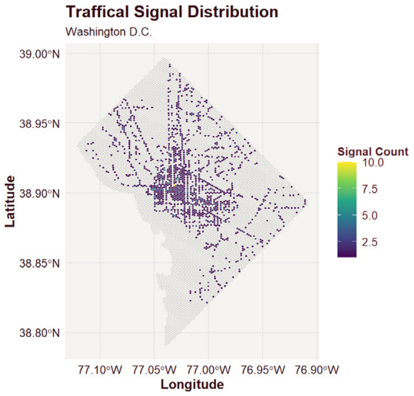
map 1. Connect Bicycle Crash Data to Fishnet



map 2. Connect Bicycle Lanes Data to Fishnet



map 3. Connect Traffical Sign Data to Fishnet



map 4. Connect Traffical Signal Data to Fishnet

4.4. Correlation Matrix

In this part, we analyze the correlation between different variables. First, we select a group of numerical predictor variables and a response variable, and then calculate their correlation matrix. This step helps us understand the relationship between different variables, thus evaluating their roles in the accident prediction model.

Next, we visualize the correlation matrix to show the correlation degree between each variable. In this heatmap, blue indicates negative correlation, red indicates positive correlation, and white indicates no correlation. By observing the heatmap, we can quickly identify which variables have strong correlations, which will be focused on in further analysis and modeling.

By analyzing the correlation matrix, we can better understand the interrelationships between different variables, thus adopting more effective feature selection and feature engineering strategies in the accident prediction model. This will help us establish an accurate and stable predictive model.

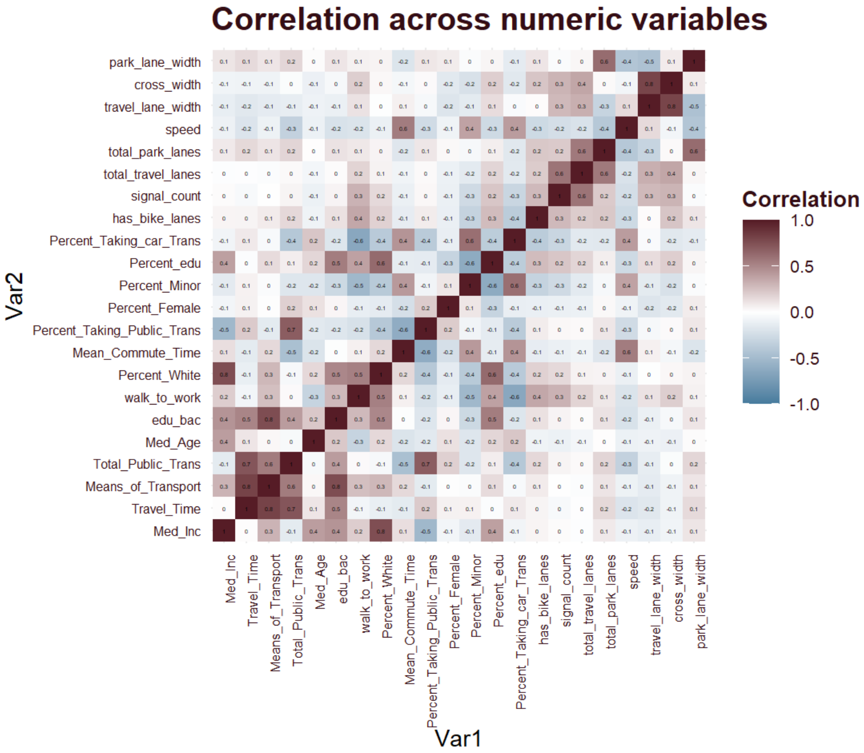


chart 4. Correlation Matrix

5. Modeling for Prediction

We will use a random forest model to predict bicycle crashes in Washington D.C. The main purpose of building this model is to predict the risk of bicycle crashes in different areas of the city based on the collected data. By doing so, we can identify which areas are more prone to bicycle accidents and take targeted preventive measures to improve road safety.

The role of the random forest model is crucial throughout the entire project. By analyzing and modeling various factors, we can identify the factors that significantly affect the risk of bicycle crashes. This will help the government and relevant agencies make more informed decisions on road planning and improvement, traffic safety campaigns, and more. In addition, the model results can guide cyclists to be more cautious in certain areas and reduce the likelihood of bicycle crashes.

Random forest models have unique advantages in this project, as they can handle complex non-linear relationships and high-dimensional data. The random forest model achieves higher prediction accuracy by building multiple decision trees and combining their prediction results, while reducing the risk of overfitting. In this project, we will use the random forest model to predict bicycle crashes in Washington D.C., thus making reasonable predictions for bicycle crashes.

we showed the results of the random forest model, especially the importance of each predictor variable. First, we extract the variable importance from the model and create a table to present these importance values.

Next, we create chart 5 to visually display the contribution of each predictor variable to the prediction of bicycle accidents. In the chart, variable importance is represented by the height of the bars, and color gradient indicates the magnitude of importance, with darker colors indicating higher importance.

By analyzing the variable importance, we can better understand which factors play a crucial role in predicting bicycle accidents in Washington DC.

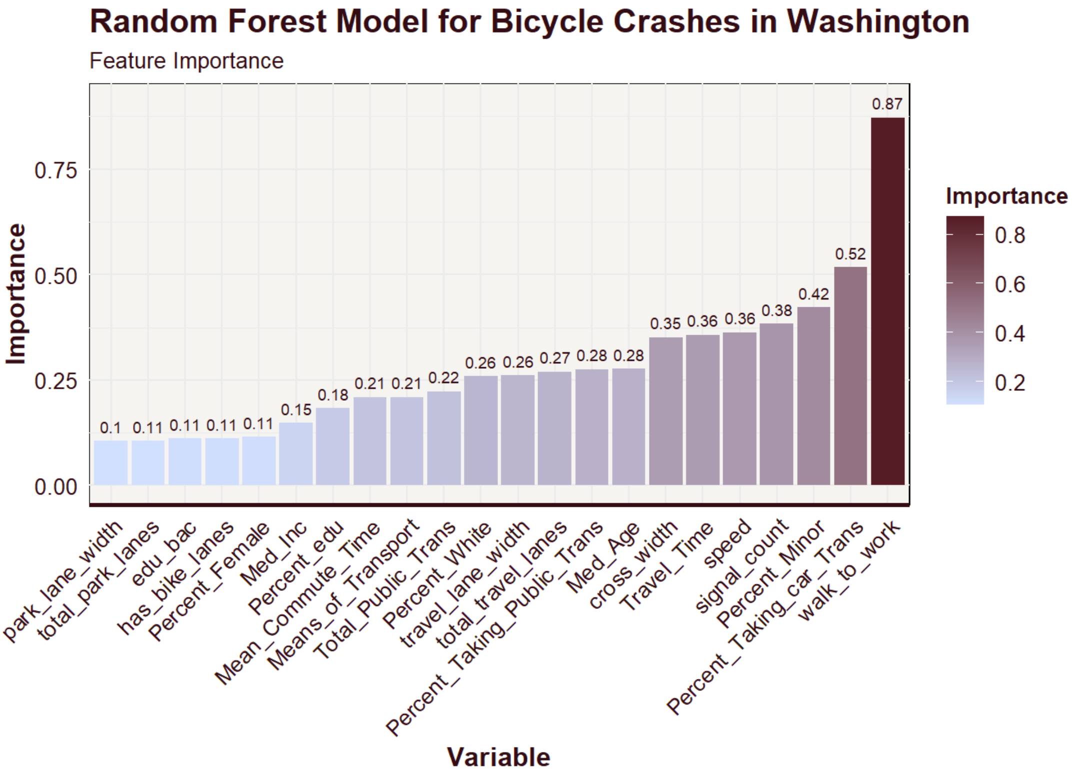


chart 5. Results of the Randomforest Model

Based on the importance results from the Random Forest model, we can draw the following insights:

1. walk_to_work (walking to work) is the most important predictor with an importance of 0.87. This suggests a strong association between the proportion of people walking to work and bicycle accidents. A possible explanation is that walking and cycling share similarities as modes of transportation, and more walkers might imply more cyclists, thus influencing bicycle accident occurrences.
2. Percent_Taking_car_Trans (percentage of car commuters) has an importance of 0.52, indicating a relatively high correlation. This may suggest that areas with a higher percentage of car commuters could have negative impacts on bicycle safety, possibly due to increased traffic congestion and a higher risk of accidents from more car commuters.
3. Other variables with higher importance include Percent_Minor (percentage of minors, 0.42), signal_ count (number of traffic signals, 0.38), Travel_Time (travel time, 0.36), and speed (speed, 0.36). These variables could have direct or indirect influences on bicycle accident occurrences.
4. Some variables have lower importance, such as park_lane_width (parking lane width, 0.10), has_ bike_lanes (presence of bike lanes, 0.11), and total_park_lanes (total number of parking lanes, 0.11). These variables play a relatively smaller role in predicting bicycle accidents but might still affect bicycle safety to some extent.

Overall, these variable importance results reveal the influence of various socioeconomic, transportation, and infrastructure factors on bicycle accidents in Washington D.C. By considering these factors, we can better understand the causes of bicycle accidents and provide insights for policymakers on how to improve cycling safety.

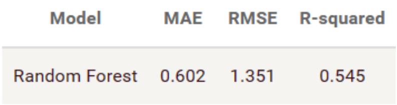
5.1. Calculation of evaluation indicators

This section presents the performance of the random forest regression model in predicting bicycle accidents. We used three evaluation metrics: mean absolute error (**MAE**), root mean square error (**RMSE**), and **R-squared** to assess the accuracy of the model. These metrics help us understand the performance of the model in predicting bicycle accidents and evaluate its effectiveness.

The results show that the random forest regression model has a certain degree of accuracy in predicting bicycle accidents in Washington DC. The MAE is 0.6, indicating an average absolute difference of 0.6 between predicted and actual values. The relatively small MAE value suggests that the model has high predictive accuracy.

RMSE measures the square root of the mean of the squared differences between the observed values and predicted values. RMSE is more sensitive to large errors than MAE, making it more effective in excluding the influence of noise data. In this case, a small RMSE value also indicates that the model has high accuracy.

The R-squared value is around 0.56, indicating that the model explains the variation in bicycle accidents. The closer the R-squared value is to 1, the higher the degree of fitting of the model. In this study, the R-squared value indicates that the model has a certain explanatory power for bicycle accidents prediction, but there is still room for improvement.



Model	MAE	RMSE	R-squared
Random Forest	0.602	1.351	0.545

5.2. K-fold Cross-validation

In this section, we performed K-fold cross-validation to evaluate the performance of the random forest regression model on different subsets of data. By dividing the data into 25 subsets, this method helps to evaluate the model's generalization ability on different data subsets. By evaluating each subset, we can observe the performance differences of the model between each subset, and thereby understand the stability and reliability of the model.

After that, we plotted 3 charts for **RMSE**, **R-squared**, and **MAE** separately. These charts show the results of each evaluation metric on the 25 different subsets. This helps us to understand the performance of the model on different subsets and estimate its prediction ability on new data. Additionally, these charts also display the average value(yellow dashed line) of each performance metric, allowing us to have a better understanding of the overall performance of the model.

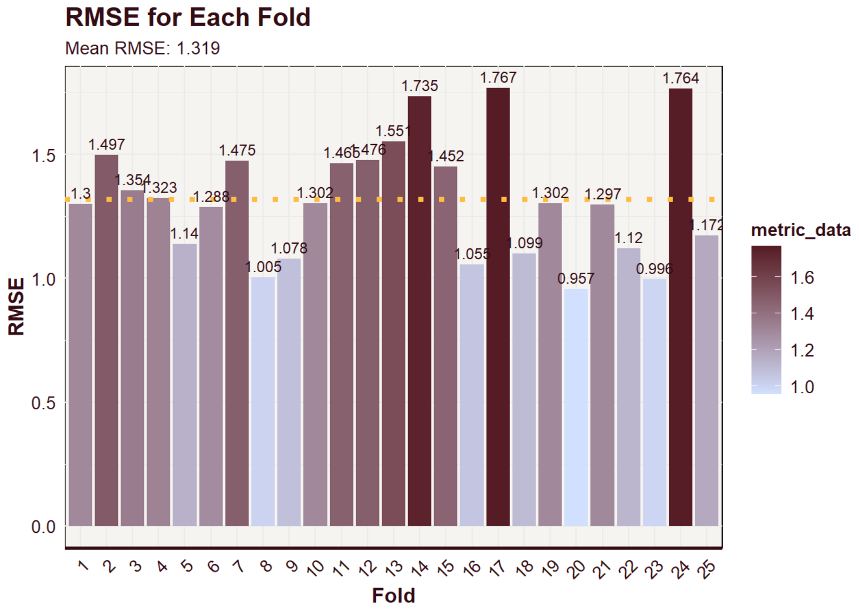


chart 6. Results of RMSE for Every Cross Validation

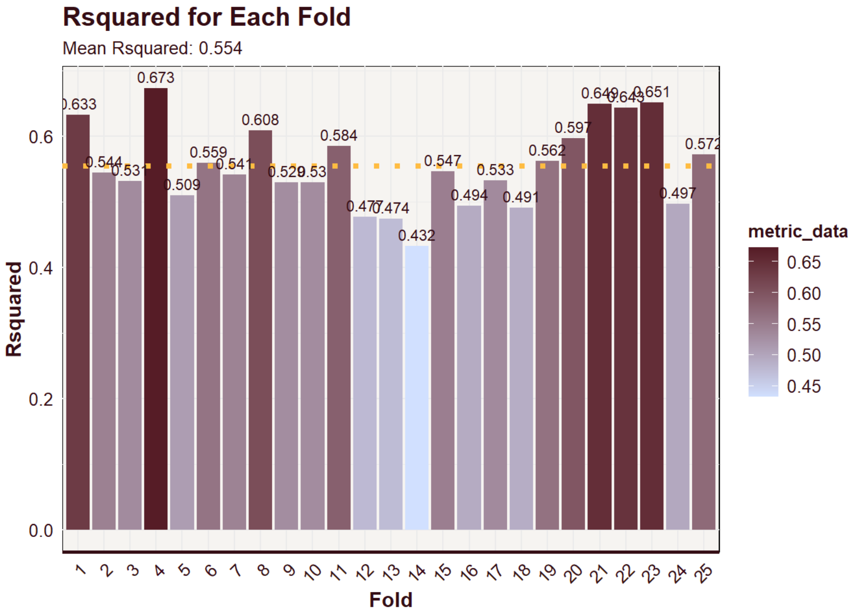


chart 7. Results of R-squared for Every Cross Validation

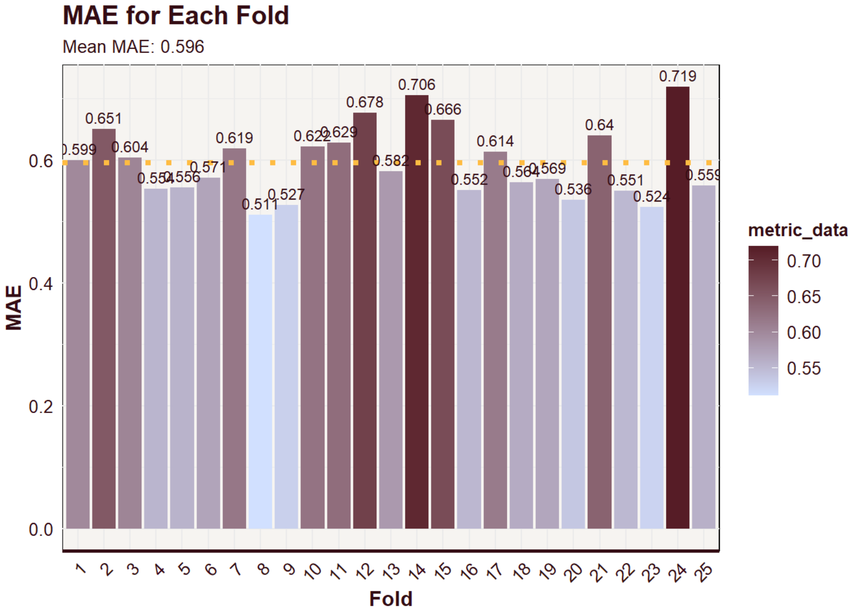


chart 8. Results of MAE for Every Cross Validation

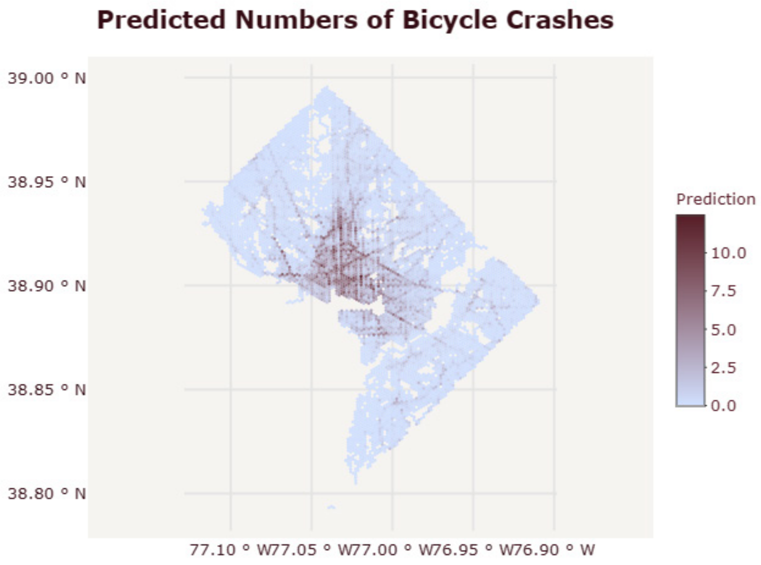
5.3. Spatial Cross-validation

Spatial validation is based on the previous k-fold cross validation by adding the geographic characteristics of the cell, taking the neighborhood where the observation point is located as the unit of calculation, and using the spatial characteristics as an important basis for validating the model, so that the results are more consistent with events with obvious geographic location and prominent spatial characteristics. The prediction of the bicycle crashes is precisely a geographic prediction event, so we choose to use spatial validation in the calculation.

In predicting bicycle accidents, spatial prediction is crucial due to the nature of the event being a geospatial event. Therefore, we chose to use spatial validation in our calculations. We first defined the variables to be included in the prediction, and then executed a spatial cross-validation function on the dataset. Through this method, we can obtain more information about the model's performance on different spatial subsets and improve our confidence in the model's spatial generalization ability. This helps ensure that our model has good predictive capabilities for new data, thereby improving and optimizing the prediction of bicycle crashes.

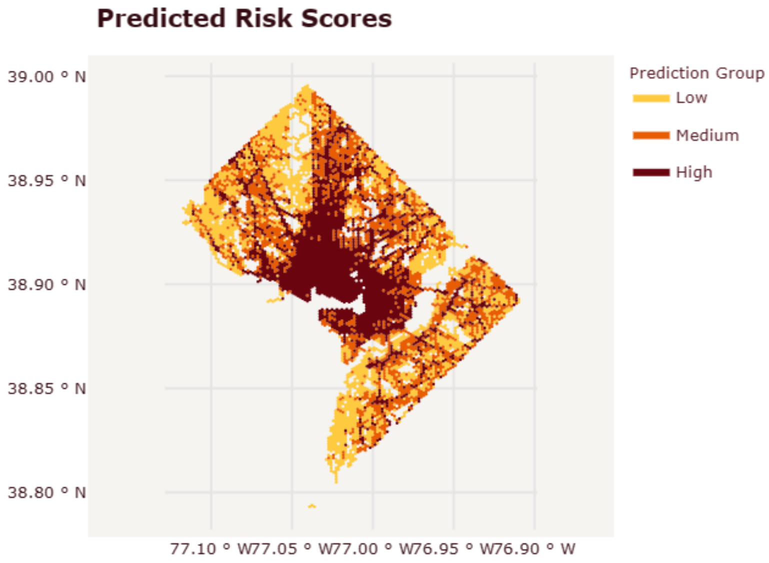
We visualize the results of spatial cross-validation to better understand the distribution of predicted bicycle crashes in Washington D.C. First, we extract relevant information from the spatial cross-validation results and round the predicted values to the nearest integer. Then, we create a choropleth map to represent the predicted bicycle crash counts with a continuous color scale. Finally, we convert the static map into an interactive map for users to explore the predicted results in different areas more conveniently.

Through this interactive map, we can better showcase the spatial distribution of predicted results in the project and identify high-risk areas. This can help to identify the factors related to bicycle crashes and improve the safety of city bike lanes, ultimately reducing the incidence of bicycle accidents. Additionally, the interactive map visualization can further validate the accuracy and generalizability of the model predictions, making the entire research project more convincing and applicable.


Predicted Numbers of Bicycle Crashes
39.00 ° N
38.95 ° N
38.90 ° N
38.85 ° N
38.80 ° N
77.10 ° W77.05 ° W77.00 ° W76.95 ° W76.90 ° W
Prediction
10.0
7.5
5.0
2.5
0.0

map 5. Visualization of Spatial Cross-validation Results

By grouping the predicted results into low, medium, and high-risk categories, we created a map to display the bicycle accident risk in each grid of Washington D.C. The color code indicates the predicted risk level, where yellow represents low-risk, orange represents medium-risk, and red represents high-risk. This visualization method helps to understand the spatial distribution of bicycle accidents and take targeted measures to improve road safety.



map 6. Predicted Risk Scores of Spatial Cross-validation

6. Final Results

we compared the difference between the predicted number of bicycle crashes (Nums) and the actual number of bicycle crashes (Crash Count). We rounded the differences to the nearest integer and calculated the frequency of each difference category.

Then, we used a pie chart to visualize these differences and show the distribution of differences between predicted and actual values. To make the pie chart more appealing, we used a custom color scheme. On the pie chart, we displayed labels and percentages for each difference category to provide a more intuitive understanding of the accuracy of the predicted results.

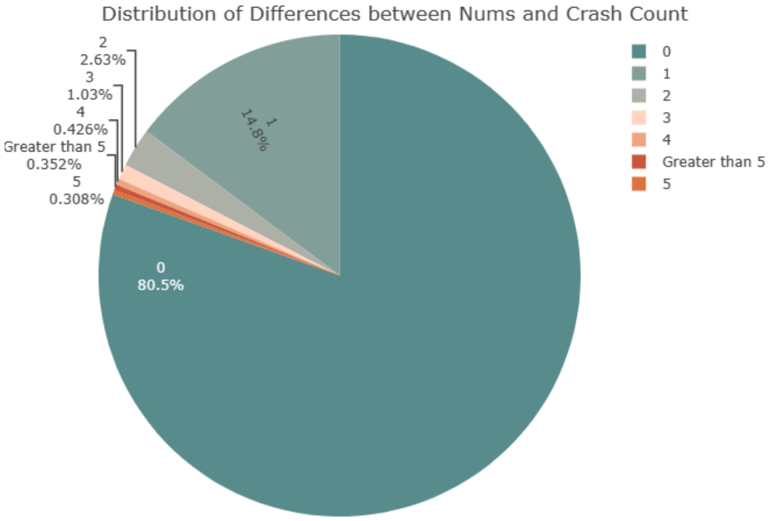
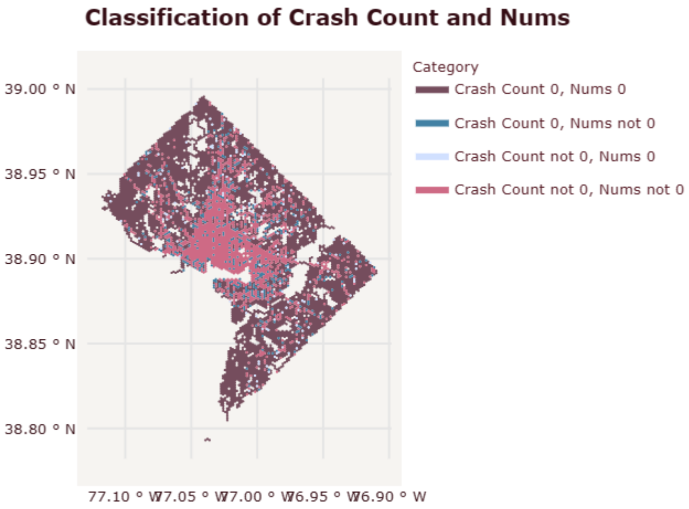


chart 9. Pie Chart of Prediction Error

we compared the actual number of bicycle accidents (Crash Count) with the predicted number of bicycle accidents (Nums) and classified them into four categories:

1. Crash Count is 0 and Nums is 0
2. Crash Count is 0 and Nums is not 0
3. Crash Count is not 0 and Nums is not 0
4. Crash Count is not 0 and Nums is 0

Then, we visualized these categories on a map to gain a more intuitive understanding of the relationship between the actual and predicted values. We used different colors to represent different categories, making the map easier to read. Through this visualization method, we can intuitively understand the areas where the model predicts correctly and the areas where there may be deviations.


Classification of Crash Count and Nums
Category
Crash Count 0, Nums 0
Crash Count 0, Nums not 0
Crash Count not 0, Nums 0
Crash Count not 0, Nums not 0
39.00 ° N
38.95 ° N
38.90 ° N
38.85 ° N
38.80 ° N
77.10 ° W7.05 ° W7.00 ° W6.95 ° W6.90 ° W

map 7. Classification of Crash Count and Nums

7. Conclusion

Our research aims to explore the severity and impact of bicycle crashes in Washington DC by utilizing machine learning methods to investigate multiple demographic, transportation, and environmental variables. Through this research, we have developed an effective predictive model that allows us to more accurately assess the risk of bicycle crashes in different areas of Washington DC and aid in the development of more targeted and effective traffic safety policies.

In terms of mathematical modeling, we used the random forest algorithm for model training and utilized cross-validation and spatial cross-validation to evaluate model performance and accuracy. Our model performed well in predicting the number of bicycle crashes, demonstrating high predictive accuracy and robustness. Additionally, our model considered spatial variables, allowing for a more precise prediction of bicycle crash risk in different areas.

However, our research also has some limitations.

1. Due to data collection and processing constraints, our model only considered a limited number of variables and may not encompass all factors that affect bicycle crashes.
2. Our study only focused on bicycle crash risk in Washington DC and may not be applicable to other cities or regions.
3. Our model can only predict future bicycle crash risk and may not provide effective solutions for bicycle crashes that have already occurred.

In conclusion, our research provides a new method for assessing bicycle crash risk in Washington DC, which has important practical significance. Although there is still room for improvement in our research, it provides useful guidance and inspiration for us to better understand and predict bicycle crash risk.

References

[1] Pucher, J., & Buehler, R. (2008). Making cycling irresistible: lessons from the Netherlands, Denmark and Germany. Transport Reviews, 28(4), 495-528.

[2] Buehler, R., & Pucher, J. (2012). Cycling to work in 90 large American cities: new evidence on the role of bike paths and lanes. Transportation, 39(2), 409-432.

[3] National Association of City Transportation Officials. (2019). Urban Bikeway Design Guide, Second Edition. Retrieved from https://nacto.org/publication/urban-bikeway-design-guide/

[4] Washington Area Bicyclist Association. (2022). Advocacy. Retrieved from https://waba.org/advocacy/

[5] Chen, H., Li, D., Zhang, H., Wang, D., & Chen, Q. (2018). Independent bicycle lane design for improving bicycle traffic safety. Journal of Traffic and Transportation Engineering, 5(3), 233-240.

[6] Zhang, K., Ma, Y., Liu, J., & Lu, J. (2020). A smart bike-sharing management system using blockchain and NB-IoT technologies. Sustainable Cities and Society, 61, 102345.

[7] Borzouei, M., Golrokhian, R., & Harati-Mokhtari, A. (2021). Social equity and accessibility analysis of bike-sharing systems: A case study of the Capital Bikeshare system in Washington, DC. Cities, 108, 103100.

[8] Gkritza, K., & Noland, R. B. (2013). Spatial analysis of bicycle accidents in Washington, DC. Journal of Safety Research, 46, 157-167.

[9] Chang, H.-L., & Yeh, T. (2018). Factors contributing to bicycle accidents in New York City. Journal of Urban Planning and Development, 144(1), 04017017. https://doi.org/10.1061/(asce)up.1943-5444.0000422

[10] Chen, X., & Shen, Y. (2019). Who benefits from bicycle lanes? A machine learning approach to understand the impact of bicycle infrastructure on bike crashes in Washington

[11] Molnar, L. J., & Eby, D. W. (2008). Predicting bicycle accidents in a mid-sized city using GIS methods. Transportation research part A: policy and practice, 42(2), 218-227.

[12] Kim, J.K., Kim, S., & Ulfarsson, G.F. (2007). Analysis of bicycle crashes in Washington State using a Bayesian multinomial logistic regression model. Accident Analysis & Prevention, 39(2), 251-257.

[13] Thomas, T., Wang, Y., & Savolainen, P. T. (2016). A spatial analysis of bicycle accidents in Washington State, 2005–2012. Accident Analysis & Prevention, 92, 229-238.