Network-Based Spatiotemporal Predictive Model for Car Crash Risk Assessment

Junyi Li, Qingyuan Feng, Jiuling Zhong, Morris Gao, Linode Ye

Abstract

Road traffic accidents are the primary cause of mortality, resulting in the death of around 1.19 million individuals annually. Road crash prediction models are helpful in enhancing urban road safety. Various studies have progressively investigated the spatiotemporal patterns of vehicular crashes and identified hotspots.However, few researchers consider the inherent characteristics of road networks, with most studies focusing solely on the Euclidean metric rather than the city block distance. This study explores the connectedness of roadways and quantifies the spatial and temporal relationships between events using an index derived from the gravity model, weighted by the time delay. Our aim is to develop a predictive model for traffic crashes in New York City, aiding policymakers in understanding the risks associated with demographic and road network effects, facilitating the implementation of targeted improvement policies.

Introduction

In the pursuit of enhancing urban road safety, the study of traffic dynamics and the prediction of car crashes has evolved significantly. Traditional traffic safety models, often static and linear, have struggled to capture the complex, interconnected nature of urban traffic systems. These conventional methods fail to accommodate the dynamic interactions among various factors like traffic volume, road conditions, and weather variations, which frequently influence traffic behavior.With the advent of network-based spatiotemporal predictive models, a new paradigm has emerged in the field of traffic safety analysis. These models leverage the interconnected nature of urban road networks and employ advanced analytics to integrate diverse data across different times and locations. This approach allows for a holistic view of traffic systems, pinpointing high-risk areas and times for targeted interventions. The ongoing study titled "Network-Based Spatiotemporal Predictive Model for Car Crash Risk Assessment" exemplifies this shift. It focuses on single vehicle collisions in New York City—a critical area of study given these collisions' frequent association with severe outcomes such as high speeds and driver distraction.

This research aims to dissect these incidents across various dimensions—temporal, spatial, and network-based—to unearth underlying patterns and causative factors. By doing so, it

intends to offer actionable insights that could guide traffic management strategies and policy-making, ultimately contributing to the reduction of traffic-related fatalities and injuries. Additionally, the predictive capability of this model holds potential as a cornerstone for further research in traffic safety and urban planning, adapting to the continuously evolving challenges of urban landscapes.

literature review

Spatial autocorrelation measures the degree to which the occurrence of one feature is influenced by the distribution of similar features in adjacent areas, and it provides a useful measure of spatial patterns. Accordingly, if attraction among similar entities acts as a driving force, then positive spatial autocorrelation exists, and the distribution would be characterized by the clustering of similar entities or vice-versa (Chou, 1995). The spatiotemporal pattern of vehicular crashes and identification of hotspots have progressively been investigated in several researches. Analyzing spatial patterns of motorcycle crashes in Honolulu, Hawaii in 1990, Levine et al., found that most crashes occurred in the vicinity of working sites instead of residential areas (Ned Levine et al., 1995). In fact, commercial and business areas are positively related with more fatal and injury crashes (Karl Kim et al., 2010). In contrast, Kim and Yamashita, discovered that the residential districts were more risky areas, especially during the peak hours (Kim Karl, Yamashit Eric,2002). Wang and Abdel-Aty's study of longitudinal crash data and intersection clustering for approximately 500 signalized intersections along the Florida corridor showed a significant correlation between crash rates in 2000 and 2002(Abdel-Aty et al., 2006). Huang et al. conducted a similar study using a Bayesian spatial model in Florida to explain regional differences in the risk of collision after controlling for variables such as travel distance and population density (Abdel-Aty et al., 2010). The results confirm that areas with greater traffic volume, population and activity density are associated with higher crash records. Higher truck traffic volumes are associated with more severe crashes, while time spent traveling to work is inversely related to various crash risks.

According to the World Health Organization (WHO), 1.35 million people die in road traffic accidents every year, with an average of 3,287 deaths every day. It is estimated that up to 50 million people are injured in road traffic accidents globally every year (World Health Organization) The critical reason for most of the traffic collisions is assigned to driver, vehicle, or environmental factors. The National Motor Vehicle Crash Causation Survey indicates that 94 percent of the crashes are caused due to driver behavior (National Highway Traffic Safety Administration,2018). However, underlying factors that could affect driver behavior resulting in a crash involvement have not been widely discussed. The socioeconomic and demographic characteristics of the driver and their environment may influence their driving conduct and eventual crash involvement. Blatt and Furman also reached the same conclusion through an examination of the correlation between socioeconomic characteristics of the driver residence and

crash occurrence (Blatt, J,1998). They demonstrated that fatal crashes are more likely to take place on rural roads, while drivers who reside in rural areas or small towns have significant involvement in such crashes. The demographic, socioeconomic, and traffic characteristics of the accident location are important factors that influence the occurrence of accidents. However, these factors do not provide any information about the type of driver involved in the accident. Therefore, it makes more sense to focus on characteristics related to where the accident drivers live rather than just examining location characteristics, because the drivers involved are more likely to come from a different area than the accident location.

Data

Cenpop - Population in NYC based on census tract

In the Cenpop database, New York City is segmented into 5,411 geographic areas. Each census tract that contains at least one household is assigned a unique geographic identifier (geo ID), and its geographical details are represented as polygons using geopandas. Additionally, each tract includes associated population data. Initially, the Cenpop database lacked detailed geographic information, so it will be difficult to future spatial join for analysis. To address this, we enriched it by merging it with the ESP:4326 geographic information table using the geo ID as a key. This enhancement allows for the visualization of population density, calculated as the area of the tract divided by its population. This visualization aids in analyzing the correlation between population density and car collisions, contributing to a broader demographic analysis.

Gpd_med_income - Median income by census tract with geometry information

As part of our demographic analysis, median income was sourced from the NYU Furman Center. Like the Cenpop dataset, the dataset containing median income by census tract also assigns a unique geographic identifier (geo ID) to each census tract. This dataset similarly lacks geographical details necessary for geopandas analysis. Consequently, we employ the same method used with Cenpop to integrate EPSG:4326 geographic information. This integration facilitates demographic analyses and prepares the data for spatial join with road geographic information in subsequent stages.

Crash_data - Car collision happens in NYC with longitude and latitude

Crash_data, unlike the previously mentioned demographics datasets, is central to our analysis and constitutes the most critical parameter in all the models we will develop. Each collision in the dataset is uniquely identified by a 'collision ID.' Additionally, the dataset records the crash date and time, specific location within the borough, zip code, and geographic coordinates (latitude and longitude), which are represented as scatter points on a map. It also details the

number of vehicles involved in each incident as contributing factors, which is crucial for statistical analysis of collision severity. Due to the necessity of spatial joins with other datasets, rows missing geometry data, particularly latitude and longitude, have been excluded. The dataset has been converted into a pandas-compatible format using longitude and latitude and subsequently transformed to the EPSG:4326 coordinate system, aligning with our two primary research objectives.

Exploratory Analysis

Spatial visualization:

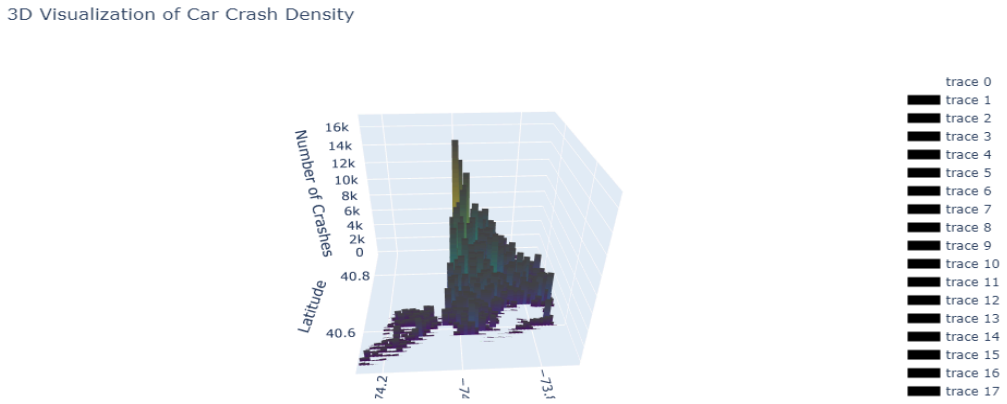

Figure 1.1: 3D bar plot of car collision

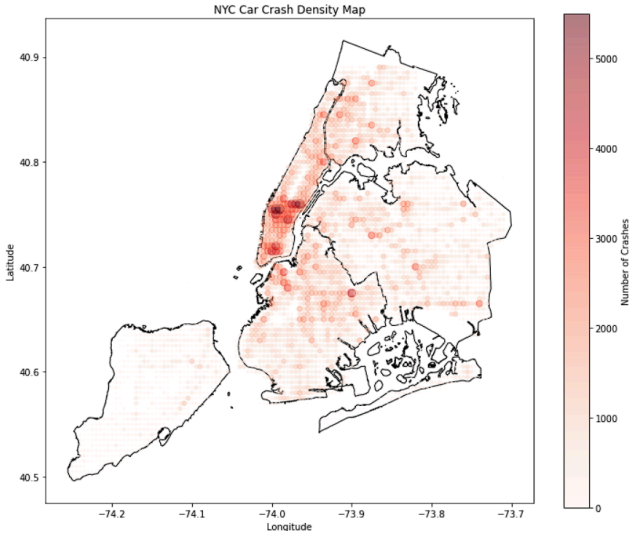

Figure 1.2: heatmap visualization of car collision

For analytics purposes, we aggregate the entire incidence-based dataset into square grids, each point representing a square area around them.

From the spatial visualization, with all the expectation of our intuition, midtown manhattan and downtown manhattan are 2 areas with the highest car accident rate. However, There is also a lot of high car accident areas that seem to be randomly scattered around. If we take a closer look, all those scattered car collision hotspots actually form some band-like pattern.

Time-Series:

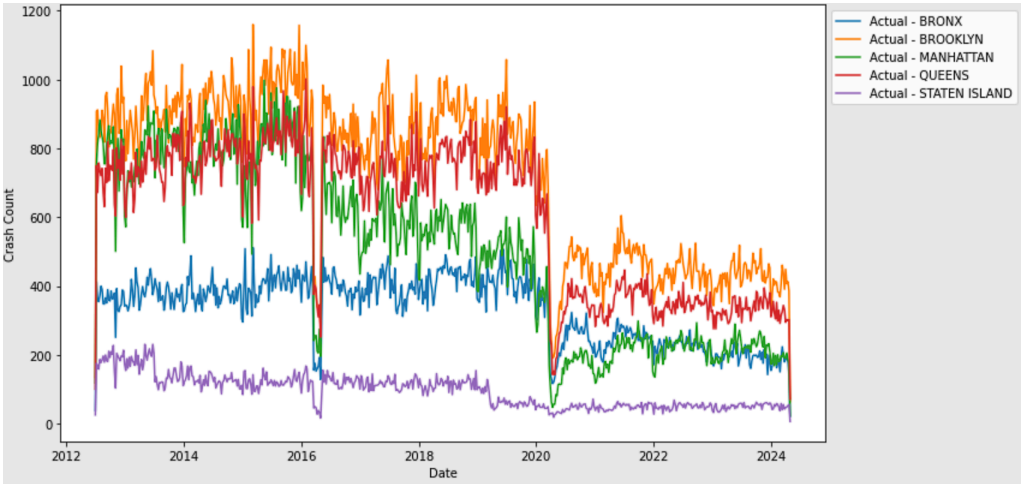

Figure 1.3: time series by 5 boroughs

The dataset includes the timing of accidents, allowing for the analysis of temporal patterns in NYC car crashes through time series plots. Observations reveal a notable decline in accidents in 2020, coinciding with the onset of the COVID-19 pandemic. Additionally, an anomaly was detected in 2016; however, after researching relevant news sources, this appears to be due to issues with data collection rather than a true anomaly in accident occurrences.

Methodology

GCN (Graph Convolution Network)

Graph Convolutional Networks (GCNs) are a type of neural network designed to process data structured as graphs, making them particularly suitable for tasks where data points have explicit or implicit relationships.

In the context of analyzing car accidents in New York City, GCNs can leverage the spatial correlations and topological structure of the city's road network. The dataset consists of precise location coordinates of car accidents, which are aggregated into square bins on the map to form the nodes of the graph. Adjacency between nodes is defined based on proximity, with each node connected to its five nearest neighbors, reflecting the dense and interconnected nature of NYC's transportation network. This setup captures local spatial relationships, allowing the GCN to learn patterns not just based on individual locations but also their context within the larger urban fabric. Additional inputs like the time of day and day of the week of each accident are integrated to enhance the model's predictive power, accounting for temporal variations in traffic and accident likelihood. This approach enables the GCN to generate insights into accident hotspots and predict future occurrences by understanding both spatial and temporal patterns within the data.

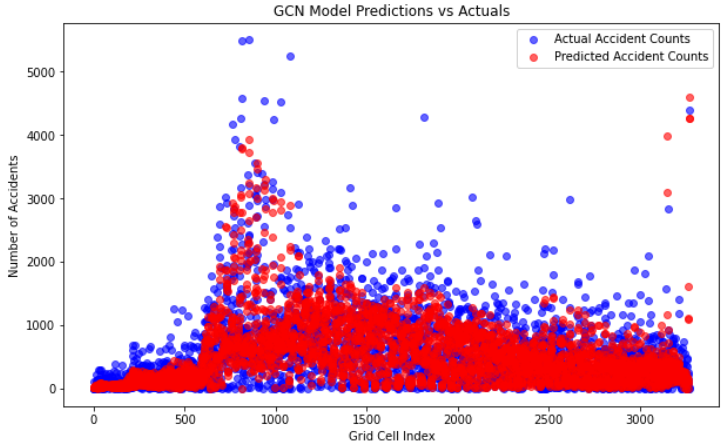

Figure 2.1: GCN model prediction vs real value on each bin area

The predictive result generally captured the spatial pattern like hotspots as well as hotspots of car crashes. However, the prediction distribution has lower variance than actual data, that might be because of the limitation of the dataset.

However, in order to help with policy making, we might want more interpretability. We will introduce an approach developed from the classical gravity model and apply it to car accidents.

Spatiotemporal Index

Spatiotemporal Index is measured by Gravity model. Gravity index assumes that accessibility at origin i is proportional to the attractiveness (weight) of destinations j, and

inversely proportional to the distance or travel cost between i and j. The Gravity index at Origin i within graph G at Search Radius r, is defined as follows:

$$Gravity[i]^r = \sum_{j \in G-\{i\}, d[i,j] \leq r}^{n} W[j]/e^{\beta * d[i,j] - plateau}$$

where W[j] is the weight of destination j, measured in hours in hours of time interval. The function of travel distance between i and j in the denominator represents an exponential distance decay rate that is controlled by parameter beta = 0.001. Plateau radius is a baseline distance from origin i in which no distance decay is applied and destinations contribute their full weights towards the index. We set it zero to ensure that distance decay is always included. For each crash after 2022, we calculate the Gravity Index at Search Radius 3 kilometers, focusing on destinations j where a crash happened within one month.

Logit Models

Logit models can integrate many details relating to the crash occurrence (i.e. such as the Spatiotemporal of crash event, the demographic of location and the degree of street).The dependent variable (vulnerability of crash) is measure by crash severity, which has four outcome categories (fatal-1, injury-0.6, property damage-0.4, no crash-0).

$$\mathrm{Log}it\left(p\right) = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k$$

p: the probability of presence of an outcome of interest,
Xk: the vector of k independent variables,
b0: the regression coefficient on the constant term (intercept),
bk: the vector of regression coefficients on the independent variables Xk

Results

Figure 3.1: Map of real and predicted severity of traffic collision.

	coefficients	std error	t	P>|t|	[0.025	0.975]
gravity	5.241e-06	4.59e-07	11.412	0.000	4.34e-06	6.14e-06
median_income	1.149e-06	6.89e-08	16.665	0.000	1.01e-06	1.28e-06
population	4.17e-05	1.34e-06	31.012	0.000	3.91e-05	4.43e-05
degree	0.2792	0.002	158.626	0.000	0.276	0.283

Table 1. Summary of OLS linear regression result.

The regression model has a strong R-squared value of 0.774, indicating that 77.4% of the variance in 'severity' is explained by the predictors. Despite a high condition number suggesting multicollinearity, the correlation matrix and Variance Inflation Factor (VIF) analysis show only mild concerns. The VIF for 'degree' is slightly above the threshold at 10.18, but its statistical significance and substantial contribution to the model justify retaining it. Notable coefficients include 'gravity_to' and 'median_income', which have small yet significant positive effects on

'severity', and 'population', which has a larger positive impact. The 'degree' variable also shows a strong positive influence.Overall, the model is robust and statistically significant.

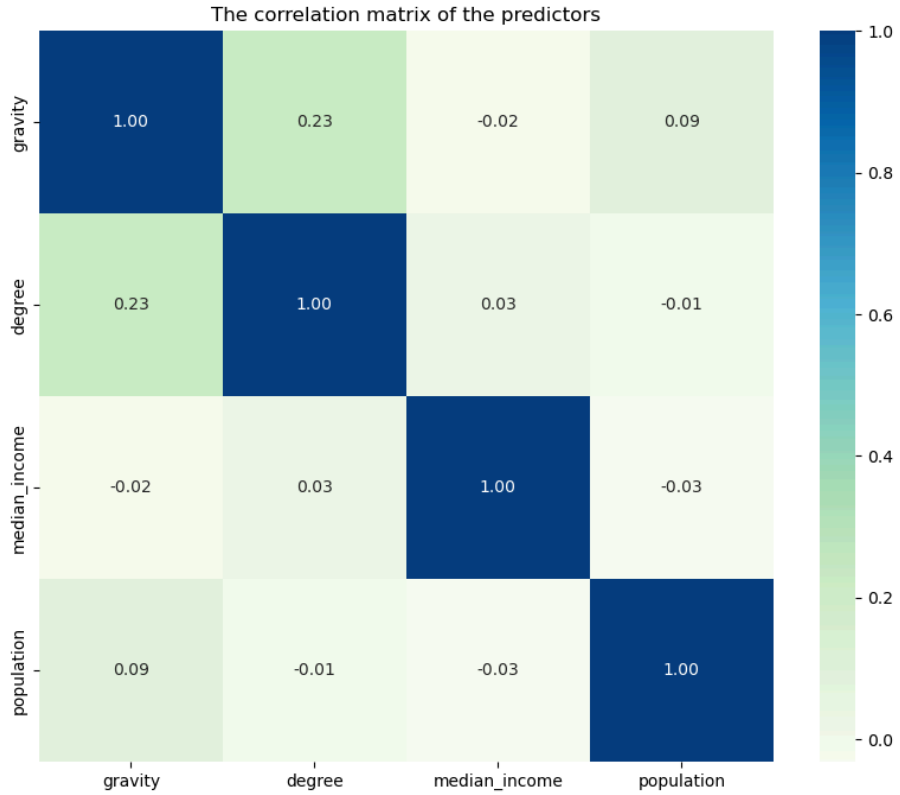

Figure 3.2: The correlation matrix of predictors.

Future Improvements

The current GCN model effectively captures spatial relationships but lacks the ability to process time-series data, treating the dataset as a static snapshot. To better predict accident patterns, future improvements could integrate LSTM networks or similar techniques to account for temporal dynamics alongside spatial dependencies. For the spatiotemporal index, instead of using hard boundaries of each accident, we may introduce weight soft boundaries and develop a measurement of spatial-temporal distance. A flexible soft boundary may also allow us to capture the heterogeneity of the transportation system. There is also a room of improvement for input features: explicit causation of each accident will be a valuable input with predictive models.

Conclusions

This report explores New York City car crash data through predictive modeling, emphasizing the potential interconnectedness of individual accidents. We use Graph Convolutional Networks (GCN) to capitalize on the spatial relationships inherent in the

grid-based count of car accidents, enhancing the model's ability to predict accidents by considering the geographic distribution and density of accidents across the city.The GCN model effectively incorporates these spatial dependencies, offering some suggestive insights into accident temporal-spatial patterns.

To get a more interpretable model, we built a spatiotemporal index based on the gravity model to measure the influence of nearby accidents adjusted for distance and time, though this aspect serves as a supplementary analysis to the primary GCN model. Regression results indicate strong statistical significance for key variables including 'gravity', 'median_income', 'population', and 'degree', each having statistical significant influences on crash severity. This robust modeling approach provides policymakers with explainable and actionable insights.

References

Abdulhafedh, Azad. “Road crash prediction models: Different statistical modeling approaches.” *Journal of Transportation Technologies*, vol. 07, no. 02, 2017, pp. 190–205, https://doi.org/10.4236/jtts.2017.72014.

Alhassan, Abdulaziz, and Andres Sevtsuk. *Madina Python Package: Scalable Urban Network Analysis for Modeling Pedestrian and Bicycle Trips in Cities*, 2024, https://doi.org/10.2139/ssrn.4748255.

Blatt, J., and Furman, S. M. Residence Location of Drivers Involved in Fatal Crashes. Accident Analysis and Prevention, 1998. 30(6), 705-711. Doi:

Chou, Yue -Hong. “Spatial pattern and spatial autocorrelation.” *Lecture Notes in Computer Science*, 1995, pp. 365–376, https://doi.org/10.1007/3-540-60392-1_24.

Choi, Seongjin, et al. “Network-wide vehicle trajectory prediction in urban traffic networks using Deep Learning.” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2672, no. 45, 7 Sept. 2018, pp. 173–184, https://doi.org/10.1177/0361198118794735.

Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey. National Highway Traffic Safety Administration, U.S. Department of Transportation, Washington, DC. 2018.

Levine, Ned, et al. “Spatial Analysis of Honolulu motor vehicle crashes: II. zonal generators.” *Accident Analysis & Prevention*, vol. 27, no. 5, Oct. 1995, pp. 675–685, https://doi.org/10.1016/0001-4575(95)00018-u.

Kim, Karl, et al. “Accidents and accessibility: Measuring influences of demographic and land use variables in Honolulu, Hawaii.” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2147, no. 1, Jan. 2010, pp. 9–17, https://doi.org/10.3141/2147-02.

Kim, Karl, and Eric Yamashita. “Motor vehicle crashes and land use: Empirical analysis from Hawaii.” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1784, no. 1, Jan. 2002, pp. 73–79, https://doi.org/10.3141/1784-10.

“Road Traffic Injuries.” *World Health Organization*, World Health Organization, www.who.int/news-room/fact-sheets/detail/road-traffic-injuries. Accessed 5 May 2024.

Wang, Xuesong, and Mohamed Abdel-Aty. “Temporal and spatial analyses of rear-end crashes at signalized intersections.” *Accident Analysis & Prevention*, vol. 38, no. 6, Nov. 2006, pp. 1137–1150, https://doi.org/10.1016/j.aap.2006.04.022.

Wang, Xuesong, and Mohamed Abdel-Aty. “Temporal and spatial analyses of rear-end crashes at signalized intersections.” *Accident Analysis & Prevention*, vol. 38, no. 6, Nov. 2006, pp. 1137–1150, https://doi.org/10.1016/j.aap.2006.04.022.

Yuan, Haitao, and Guoliang Li. “A Survey of Traffic Prediction: From Spatio-Temporal Data to Intelligent Transportation - Data Science and Engineering.” *SpringerLink*, Springer Singapore, 23 Jan. 2021, link.springer.com/article/10.1007/s41019-020-00151-z.

Appendixes A Individual Roles and Contributions

	All members contribute to report and video record
Junyi Li	Data processing, Spatiotemporal Index, Logit Model
Linuode Ye	Data processing, Exploratory analysis, GCN Model
Qingyuan feng	Data collection, Data pre processiong
Jiuling zhong	Data collection, Data pre processiong
Morris Gao	Data collection, Data pre processiong, vedio recording edit