

Identifying High-Risk Areas for Traffic Collisions in Montgomery, Maryland Using KDE and Spatial Autocorrelation Analysis

STANISLAV LIASHKOV

UNIVERSITY OF COLORADO BOULDER

Stanislav.Liashkov@colorado.edu

Despite a global decline in motor vehicle crash fatalities due to improved research and road safety policies, road traffic injuries remain a significant public health concern. The World Health Organization's 2023 report highlights [1] that road traffic injuries are the leading cause of death among individuals aged 5-29, with over half of fatalities involving pedestrians, cyclists, and motorcyclists. This study addresses this critical issue by identifying high-risk areas in Montgomery County, Maryland, contributing to the global goal of halving road traffic deaths and injuries by 2030. Using Kernel Density Estimation (KDE) and spatial autocorrelation analysis, we estimate collision densities and identify hotspots for targeted interventions. Our findings reveal significant spatial clustering of traffic collisions, with distinct patterns in densely populated urban areas and rural regions, offering valuable insights for policymakers to enhance road safety.

Keywords: Road collisions analysis; Spatial Autocorrelation; Geographic Information System; Hotspots detection

1 Introduction

Montgomery County stands as the most populous county in the U.S. state of Maryland, with a population of 1,062,061 as recorded in the 2020 census, reflecting a 9.3% growth since 2010. The county seat is Rockville, while Germantown holds the distinction of being the most populous municipality within the county. Situated adjacent to Washington, D.C., Montgomery County forms a vital part of both the Washington metropolitan area and the Washington–Baltimore combined statistical area. The majority of its residents live in bustling communities such as Silver Spring, Bethesda, Germantown, and the incorporated cities of Rockville and Gaithersburg.

Motor vehicle crashes are a critical public health concern, particularly for young individuals, for whom they are the leading cause of death in the United States. Developing effective strategies to mitigate these fatalities requires a comprehensive understanding of the spatial and temporal patterns of traffic collisions. Traditional statistical approaches often fail to account for the complex spatial relationships inherent in collision data, underscoring the need for advanced spatial analysis methods to identify and prioritize the most hazardous hotspots for policymakers.

In spatial analysis, Moran's I is utilized to evaluate whether the distribution of collision counts deviates from spatial randomness, providing a global measure of spatial dependence. Local Moran's I complements this by identifying statistically significant clusters of spatial autocorrelation and classifying them based on their characteristics.

* **Severe collision** - is defined as an incident where either a driver or a non-motorist (e.g., pedestrian, cyclist) sustains a fatal injury or a suspected serious injury requiring hospitalization.

LISA clustering complements this by identifying localized hotspots and assessing the significance of spatial autocorrelation for specific locations. By calculating Local Moran's I for each census tract, LISA categorizes locations into cluster types, such as high-high (HH) clusters, where the target variable is relatively high and surrounded by similar areas, and low-low (LL) clusters, where the target variable is relatively low and similarly clustered. Importantly, only locations with significant spatial autocorrelation, as determined by pseudo p-values from Local Moran's I, are considered meaningful.

In addition, Kernel Density Estimation (KDE) is employed to visualize collision densities across Montgomery County. KDE generates heat maps that highlight areas with high concentrations of collisions, facilitating the identification of critical zones for effective interventions.

In this work, we apply these techniques to analyze traffic collisions in Montgomery County. Our objective is to identify and highlight unsafe areas prone to collisions that may require targeted interventions to enhance road safety and reduce fatalities.

2 Related Work

The problem of finding hotspots of road crashes has been researched for decades and several approaches have been established. Though there is no unique standard of approaching this kind of problem, spatial statistics methods have been gaining popularity among researchers in comparison to other methods.

Dezman et al. [2] analyzed traffic collisions in Baltimore from 2009 to 2013, utilizing ARIMA modeling to identify temporal patterns and spatial autocorrelation techniques, such as Moran's I [3] and LISA [4] clustering, to pinpoint crash hotspots. Additionally, they employed Kernel Density Estimation (KDE) to illustrate dense hotspots for various types of collisions. Their study revealed that most collisions occurred in high-density urban areas and at intersections with significant pedestrian activity, with distracted driving emerging as the dominant risk factor. These findings underscore the importance of addressing modifiable factors, such as road design and driver behavior, to enhance safety. Building upon their work, our geospatial analysis applies similar methodologies to Maryland's collision data from 2015 to 2024. By utilizing their approach in a different region and larger time frame, we aim to uncover insights into traffic collision patterns within Montgomery country.

Researches Ali Soltani et al. [5] have applied methods from spatial statistics such as Getis Ord Gi* [6] statistics and Moran's I to detect hotspots of road crashes in Shiraz city and found that collisions produced significant clustering. The result of their study confirmed crashes hotspots generally appeared on arterial roadway which tend to have higher car speed/volume and more travel lanes. Md Saiful Alam et. al [7] utilized Getis Ord Gi*, the crash severity index and Moran's I to evaluate distribution of road traffic crashes in Ohio. As result of their research, they discovered significant hotspots in big cities of state: Cleveland, Cincinnati, Toledo, and Columbus and demonstrated the efficiency of spatial analysis applied to identifying dangerous areas prone to vehicle crashes.

3 Data

The dataset was compiled from open data on traffic collisions in Montgomery County, Maryland. [8] These data were collected by the Automated Crash Reporting System (ACRS) [9] of the Maryland State Police and reported by the Montgomery County Police, Gaithersburg Police, Rockville Police, and Maryland-National Capital Park Police. The dataset includes general information about each collision as well as detailed records of all incidents reported within the county. Various characteristics and contributing factors are documented, including location, vehicle type, type of harmful event, driver distraction, alcohol involvement, and more. The dataset can be found here [10].

3.1 Scope of Data

This dataset focuses on Montgomery County in Maryland (figure 1), USA, spanning the years 2015 to 2024. Over this period, around 106,000 road accidents were documented, including 2,547 severe incidents that caused fatalities or serious injuries as result of collision.

Figure 1: Montgomery Country, Maryland - Planning map ([link](#))



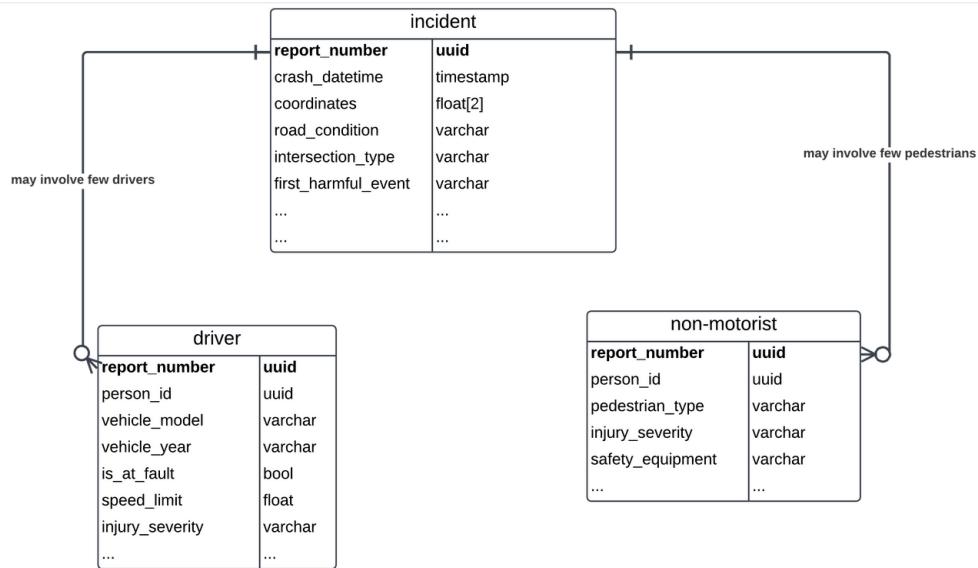
3.2 Schema

The dataset consists of three interconnected tables: incidents, drivers, and non-motorists. The incidents table provides general information about each collision, including time, location, road conditions etc. Incidents may involve multiple drivers, detailed in the drivers table, and occasionally non-motorists, such as pedestrians or cyclists, described in the non-motorists table

Table 1: metadata of tables in Dataset (after normalization)

Table name	Number of rows	Number of columns	Entity
incidents	106k	17	reported collision
drivers	180k	40	driver involved
non-motorist	6k	14	pedestrian/cyclist involved

Figure 2: Entities Relationship Diagram (ERD)



3.3 Limitations of Data

The collision data used in this study is subject to several limitations that should be carefully considered. As specified on the county's data provision website, collision reports are based on preliminary information provided to the Police Department by the reporting parties. This approach may introduce potential challenges, including:

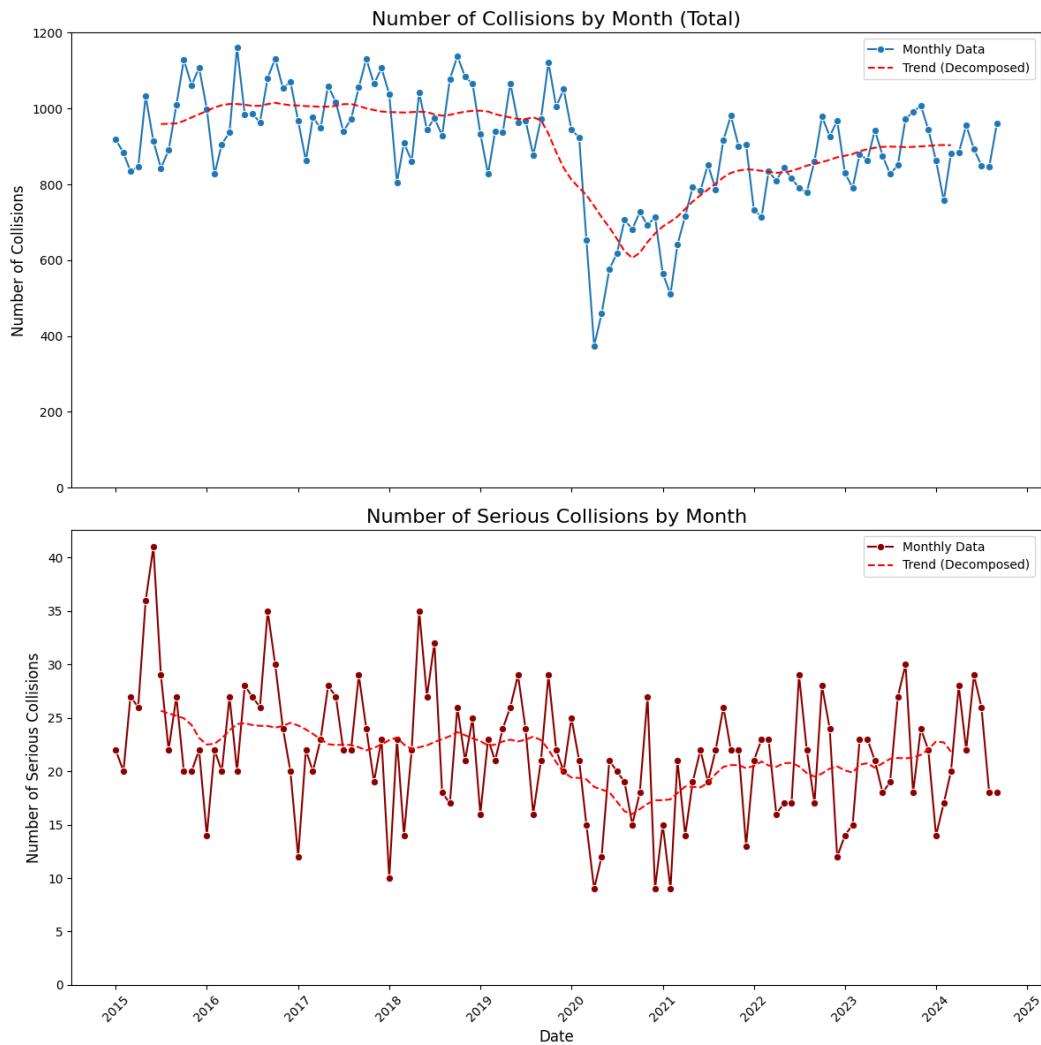
- **Unverified Information:** Initial reports may include details that have not been corroborated through further investigation, potentially leading to inaccuracies.
- **Incomplete Collision Details:** Certain elements of collisions may be missing or inaccurately recorded in the preliminary reports, affecting data completeness.
- **Subsequent Updates:** Collision data may be revised or updated following official investigations, resulting in inconsistencies and potential discrepancies within the dataset.
- **Reporting Errors:** Mechanical malfunctions or human errors during data entry and reporting processes may lead to inaccuracies in the collision records.

4 Results

4.1 Time Series

The time series plot (Figure 2) displays evident seasonality in monthly collision incidents along with an overarching trend. Notably, in the spring of 2020, there is a pronounced structural break and a significant reduction in traffic accidents attributed to the COVID-19 pandemic and associated lockdown measures. This deviation beginning in spring 2020 and extending into subsequent months makes the time series more complex for modeling with traditional approaches such as ARIMA or SARIMA, as it violates the assumptions of stationarity and homoscedasticity. Furthermore, the plot indicates that the overall trend in the total number of collisions is slowly recovering toward pre-pandemic levels. Nevertheless, even several years after COVID-19, the number of collisions remains below pre-pandemic figures.

Figure 3: Time-Series: Number of collisions from 2015 to 2024 (total / severe only)



By overlaying monthly collision data for each year in Figure 3, a distinct cyclic trend emerges. The highest number of collisions occur in May and October, while the fewest are observed in February. We hypothesize that the elevated collision rates in May and October are driven by increased tourism during these months which leads to higher traffic volumes and, consequently, a greater likelihood of accidents. This consistent pattern underscores the significant seasonal variation in traffic accidents. Repost on tourism in Montgomery can be found here [11].

Figure 4: Seasonality of Total Collisions (top) and Severe Collisions (bottom) from 2015 to 2024

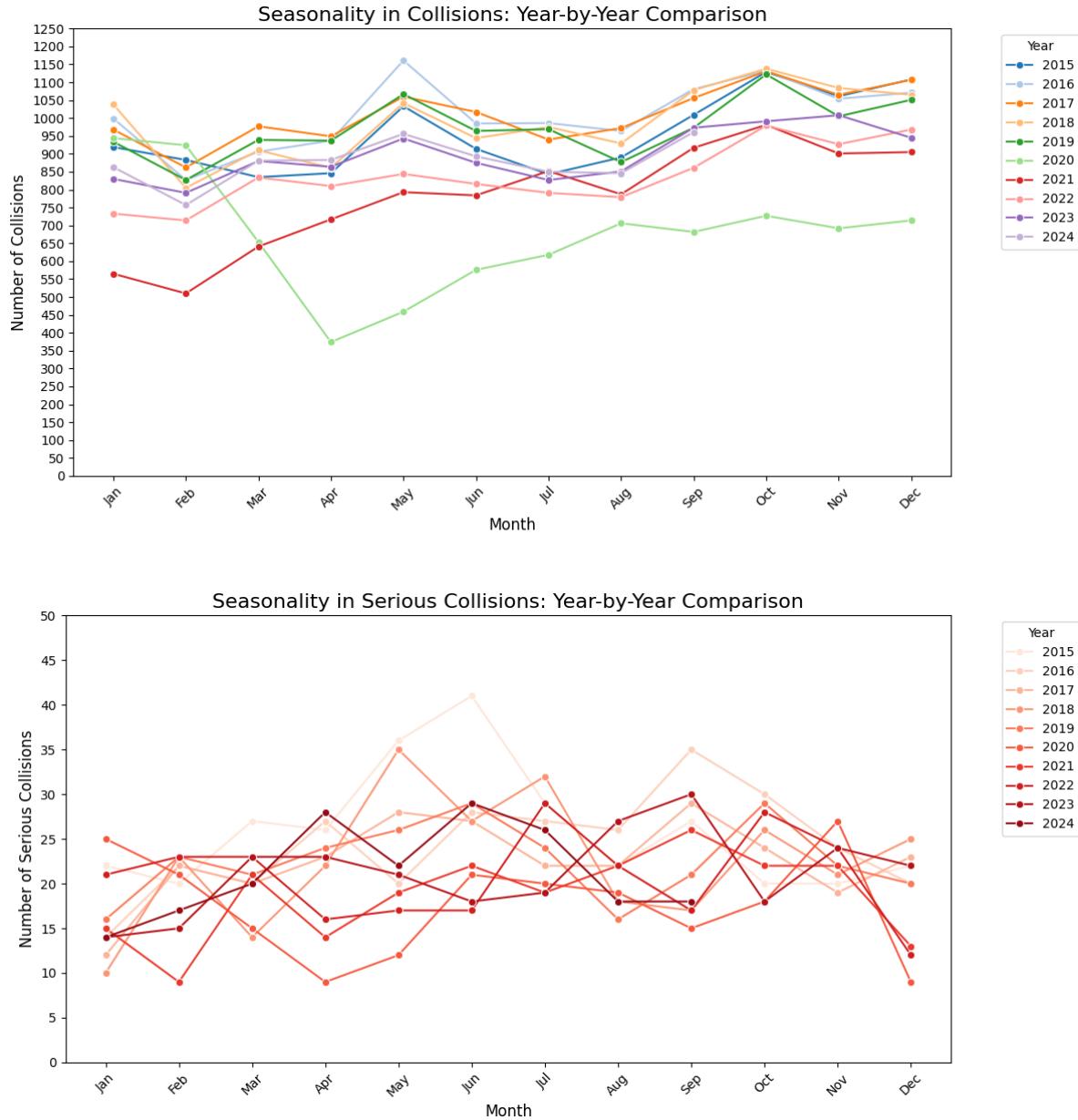
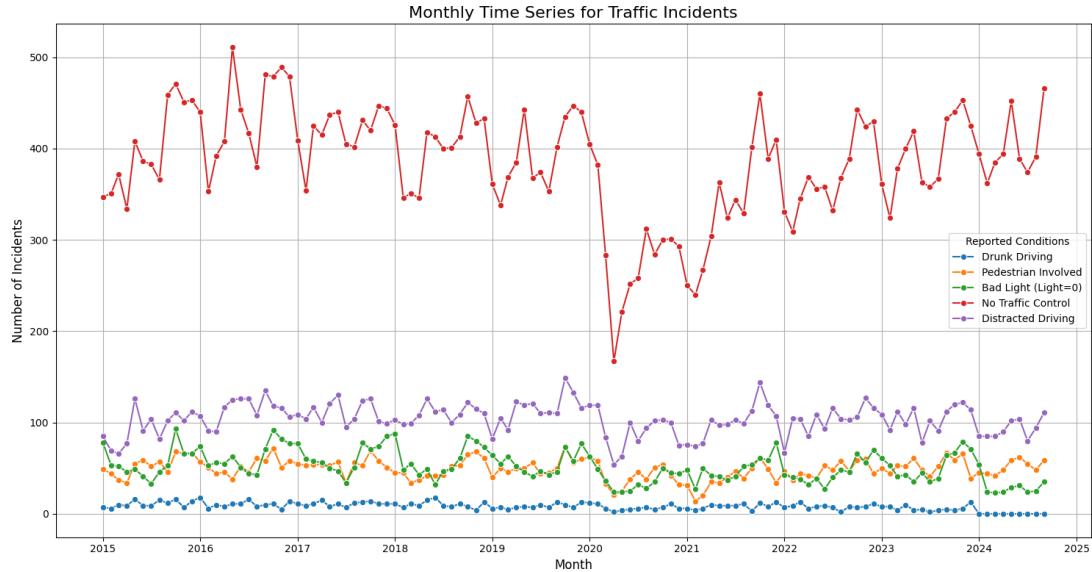


Figure 5: Time series of collisions by conditions



Montgomery County, approximately 106,000 traffic crashes were reported, with 2.17% resulting in fatalities or serious injuries. Of these, 43% occurred at locations without traffic control. Driver distraction accounted for 11% of crashes, poor lighting for 5.7%, 5.5% involved pedestrians, and 0.9% involved drivers under the influence of alcohol. Figure 5 illustrates the temporal variation in collision frequencies under different conditions, showing that the seasonal frequency of collisions under poor lighting conditions is smaller compared to those involving driver distraction.

4.2 Dense areas for different types of crashes

It is reasonable to assume spatial patterns in traffic collisions. To generate a heatmap of collisions, we applied the non-parametric Spatial Kernel Density Estimation (KDE) technique to approximate the density of collisions under various circumstances. As expected, the estimated density of both total and severe crashes is concentrated in densely populated areas of the county, including Wheaton–Glenmont (in the southern part of the county, near Washington, DC) and the Gaithersburg–Germantown area (a densely populated northern region). Another pattern (fig. 6) is that the most dense areas for collisions with pedestrians are located in most populated towns of the county. Figure 7 also reveals interesting spatial patterns on both plots: most collisions with animals occurred in the northern part (near Germantown) as this area is surrounded by forests and national parks.

Figure 6: Spatial KDE map for collisions (total/severe)

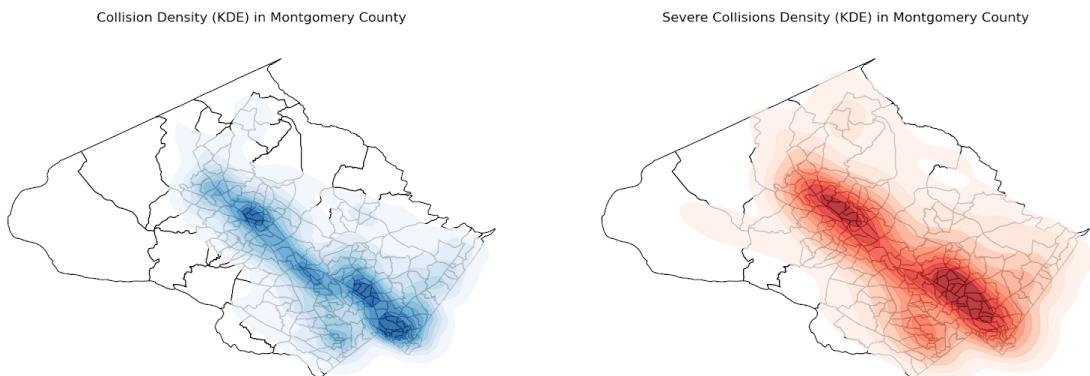


Figure 7: Spatial KDE map for collisions with pedestrians and with fixed objects

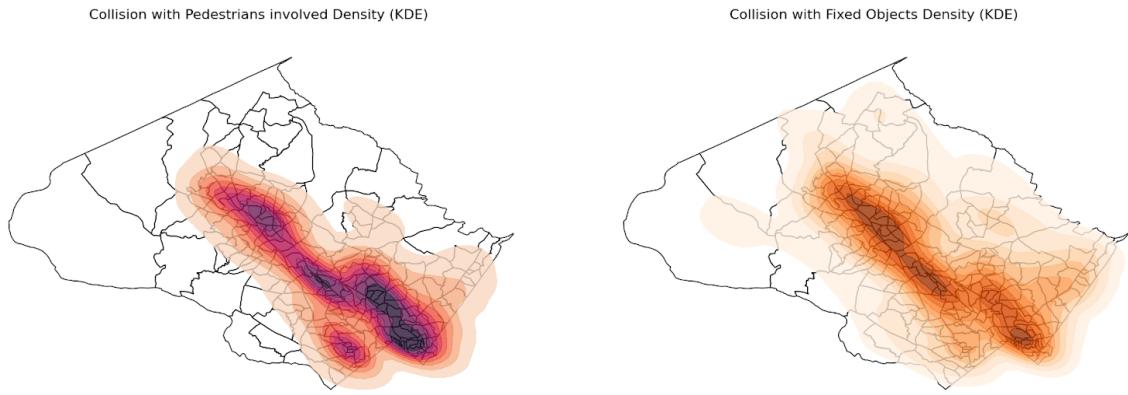


Figure 8: Spatial KDE map for collisions with animals and with parked vehicles

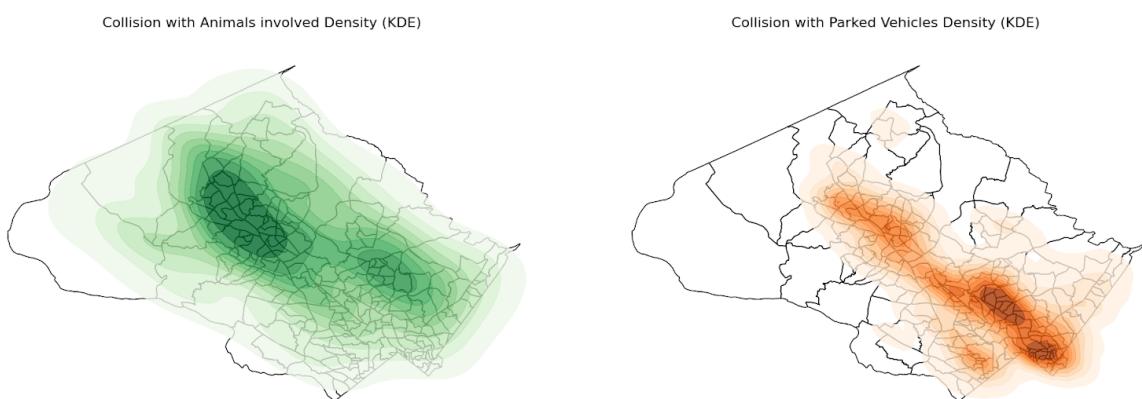
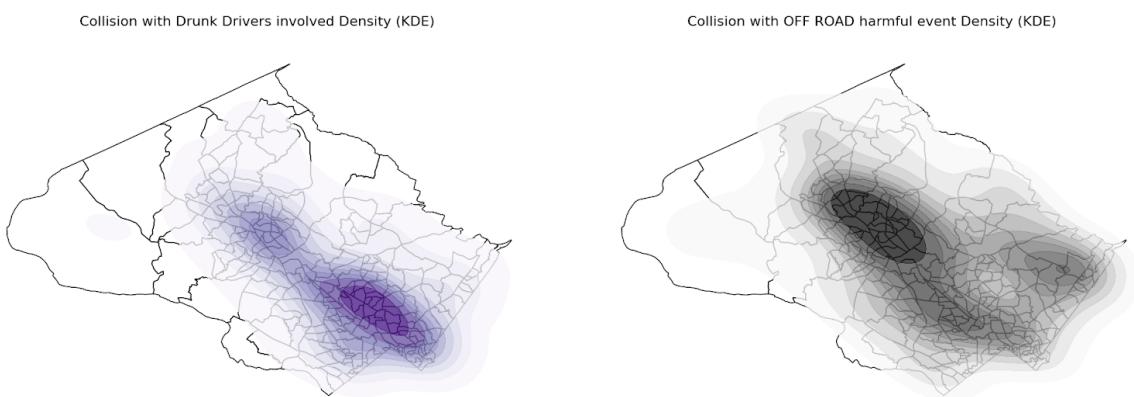


Figure 9: Spatial KDE map for collisions with pedestrians and with fixed objects



4.3 Spatial Analysis with Local Moran and LISA

4.3.1 LISA for counts of collision within census tracts

To analyze spatial autocorrelation in traffic collisions, each collision location was mapped to its corresponding census tract within Montgomery County. All subsequent analyses were conducted at the census tract level. We first computed the global Moran's I statistic to test the null hypothesis of spatial randomness. Utilizing a spatial weights matrix constructed with the k-nearest neighbors (KNN) method ($k = 10$) and a significance level of 0.05, the Moran's I value was 0.11 ($p\text{-value} = 2.4\text{e-}5$). This result rejected the null hypothesis, indicating significant positive spatial autocorrelation, where census tracts with high collision counts tend to cluster spatially.

Figure 10: LISA clusters of collisions counts (significance level = 0.05)

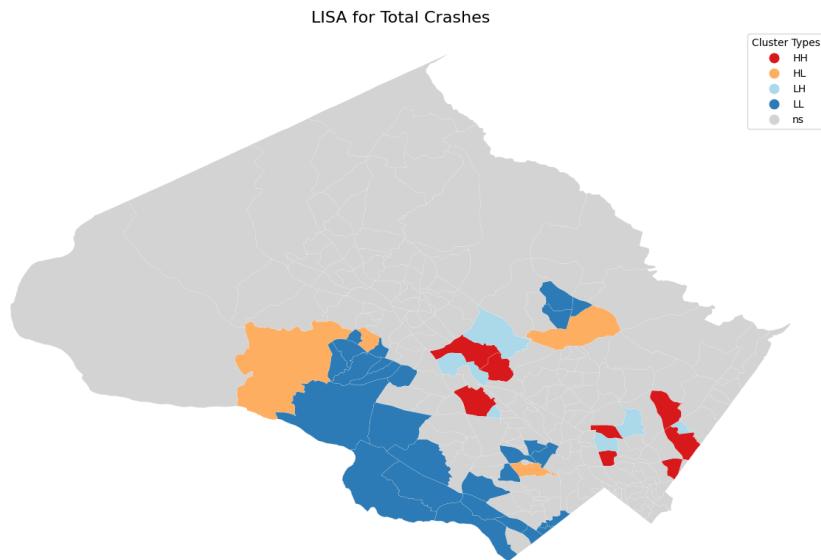
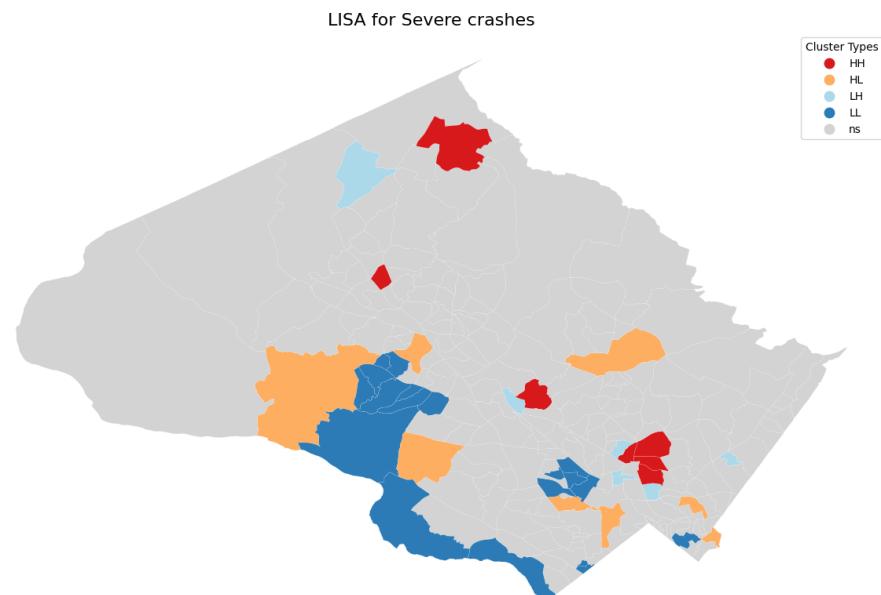


Figure 11: LISA clusters of severe collisions counts (significance level = 0.05)



Once spatial randomness was rejected, the Local Indicators of Spatial Association (LISA) method was employed to examine local spatial patterns of traffic collisions in more detail. LISA provided deeper insights into collision clustering by calculating the Local Moran's I for each census tract, assessing both the significance and the relative value of the target variable. Significant locations were categorized into cluster types: high-high (HH), high-low (HL), low-high (LH), and low-low (LL).

The LISA results for all crashes (Figure 9) and for only severe crashes (Figure 10) indicate that significant cluster locations are similar yet distinct. High-high (HH) clusters for all collisions (Figure 9) are primarily located in densely populated areas of the county, such as Rockville, Wheaton, and Glenmont. In contrast, significant HH clusters for severe collisions (Figure 10) are distributed differently, including areas like Damascus, Germantown, Rockville, Wheaton, and Glenmont. This distinction suggests that while densely populated regions are prone to a higher number of collisions overall, severe collisions also occur in additional areas (Damascus and Germantown) and require further analysis.

4.3.2 LISA for severity rate among tracts

An important aspect of collision spatial analysis is assessing the extent of damage inflicted on the involved parties in different areas. In our study, we define this as the “average severity,” calculated as the rate of severe collisions (the number of severe collisions divided by the total number of collisions in an area). This metric effectively identifies both high-risk areas with elevated severity rates and relatively safer regions with lower severity rates. However, using rates introduces a challenge: the variance of a proportion depends on its expected value. Consequently, different rates exhibit varying variances, violating the homoscedasticity (constant variance) assumption.

To address this issue, we applied the Empirical Bayes Index (EBI) [12] to stabilize the variance of the severity rates. The EBI method adjusts the estimated rates by incorporating the uncertainty associated with each rate, effectively shrinking them toward the global average proportion in proportion to their variance. (figure 12) Additionally, to handle zero rates and prevent division by zero errors, Laplacian smoothing [13] was employed.

Figure 12: Computation of Empirical Bayes Index (EBI) with Laplace Smoothing. EBI addresses the problem of variance instability across rates and shrink rates towards global mean rate based on standard deviation of rate.

$$\text{severity_rate}_i = \frac{\text{severe_crashes}_i + 1}{\text{total_crashes}_i + 2}$$

$$\text{severity_rate_std}_i = \sqrt{\frac{(1 - \text{severity_rate}_i) \times \text{severity_rate}_i}{\text{total_crashes}_i + 2}}$$

$$\text{global_severity_rate} = \frac{\sum_{i=1}^N \text{severe_crashes}_i + 1}{\sum_{i=1}^N \text{total_crashes}_i + 2}$$

$$\text{EBI_severity}_i = \frac{\text{severity_rate}_i - \text{global_severity_rate}}{\text{severity_rate_std}_i}$$

$$\text{EBI_severity_standardized}_i = \frac{\text{EBI_severity}_i - \mu_{\text{EBI}}}{\sigma_{\text{EBI}}}$$

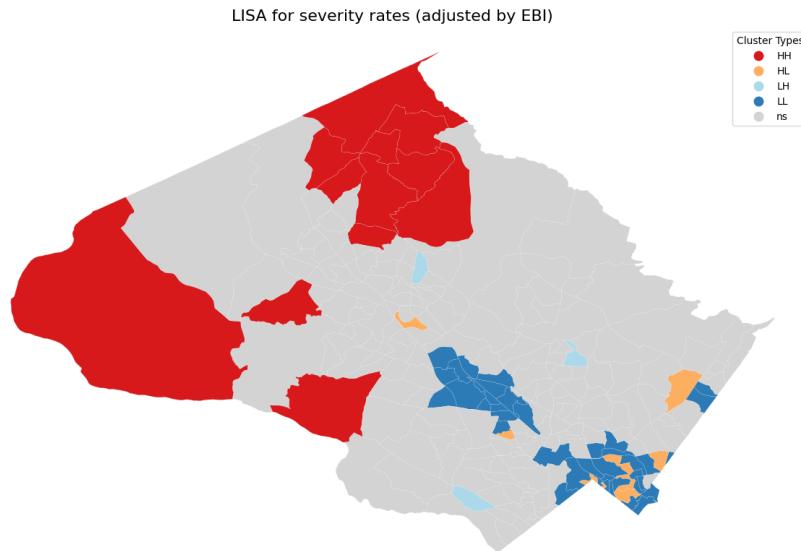


Figure 13: LISA for EBI of severity rate. (significance level = 0.05)

The LISA clusters for the adjusted severity rate, illustrated in Figure 12, reveal that hotspots (HH clusters) are not concentrated around the county's most populated towns. Instead, these hotspots are primarily located in less populated rural areas such as Damascus and Poolesville. In contrast, the most populated areas are predominantly classified within Low-Low clusters, exhibiting lower severity rates. These hotspots signify regions with a higher risk of severe road crashes relative to the county's overall severity rate, while the Low-Low clusters indicate safer, more densely populated regions. This spatial distribution suggests two plausible hypotheses for the observed pattern:

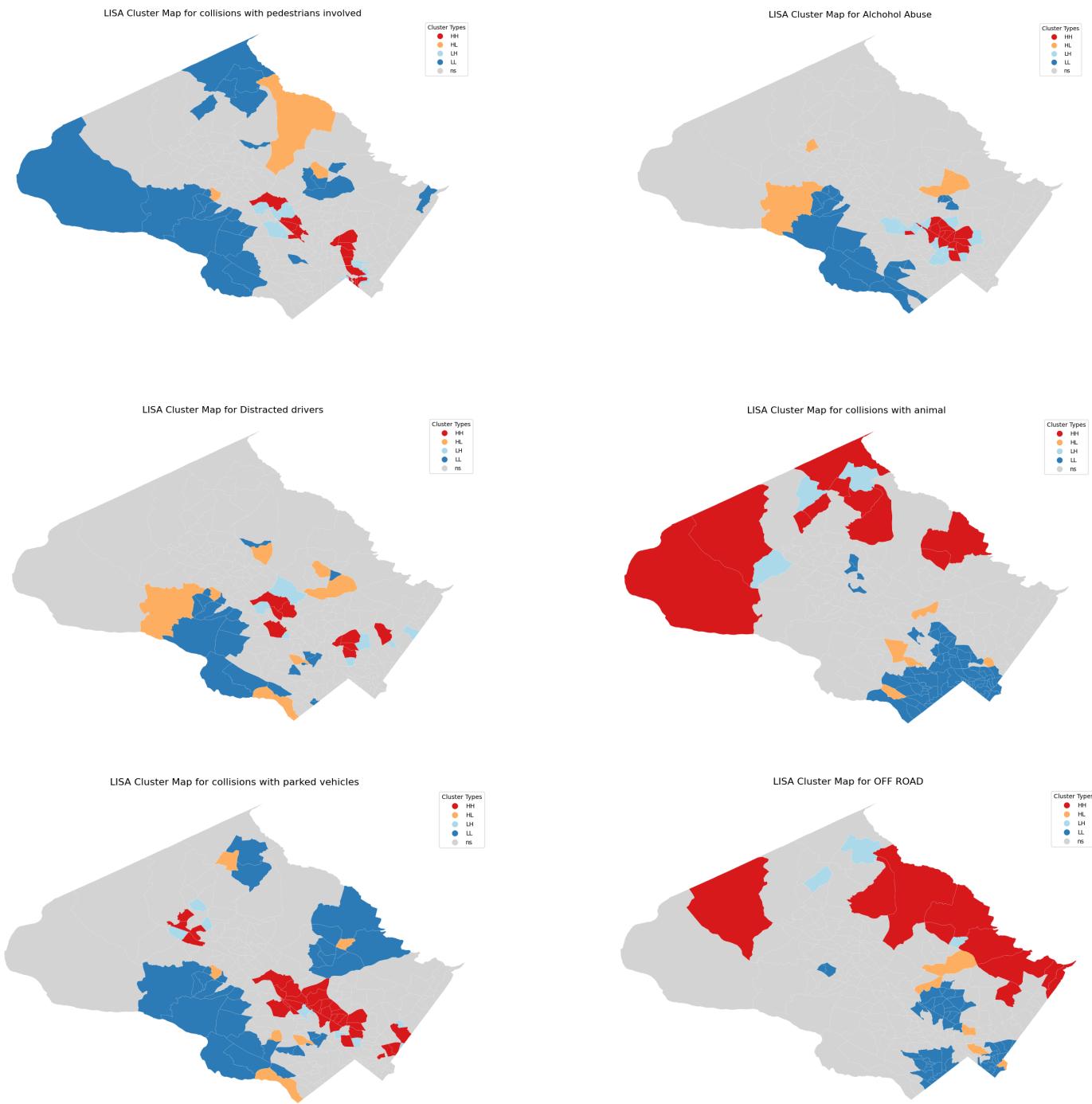
- **Infrastructure and Traffic Control Deficiencies:** These rural areas may be more hazardous due to poorer road conditions, inadequate lighting, insufficient traffic control measures, and more aggressive driving practices.
- **Underreporting of Minor Incidents:** The remoteness of these areas from police stations may result in the underreporting of minor collisions, causing only severe crashes to be recorded and thereby increasing the apparent severity rate.

4.3.3 LISA for different types of collisions

The spatial patterns of traffic collisions vary significantly when analyzed by different types, such as alcohol-related crashes, pedestrian involvement, off-road incidents, and distracted driving. Figure 13 (1 row, 1 col) illustrates that densely populated areas, including Rockville, Wheaton, and Glenmont, form significant high-high (HH) clusters for collisions involving pedestrians. In contrast, rural areas are primarily classified as low-low (LL) clusters, indicating significantly fewer pedestrian-involved collisions.

For alcohol-related crashes (Figure 13, Row 1, Column 2), the LISA analysis identified a single significant HH cluster spanning the Wheaton and Glenmont areas. For collisions involving parked vehicles (Figure 13, Row 3, Column 1), the results are as expected, with significant clusters concentrated in the county's most populated towns. In the case of collisions caused by distracted drivers, significant clusters are observed in tracts spanning Rockville, Wheaton, Glenmont, and also Poolesville, which notably do not appear as significant clusters in other LISA maps.

Figure 14: LISA for crashes with specific factors or circumstance



Crashes involving animals (Figure 13, Row 2, Column 2) show a distinct pattern compared to total collisions, with significant HH clusters located in rural, less populated areas near forests and national parks. A similar spatial pattern is observed for off-road collisions (Figure 13, Row 3, Column 2), which are also concentrated in the county's sparse, rural regions.

4.3.4 Crashes points distribution for significantly dense areas

To identify critical areas for traffic safety analysis, we focused on census tracts within Montgomery County's most populated towns that exhibited both a relatively high number of collisions and significant spatial autocorrelation. As illustrated in Figure 14, crash points within these selected tracts are marked in gray, whereas severe crashes are denoted by red markers. Further examination of these densely collision-prone locations, shown in Figure 15, reveals that most collisions occur at intersections.

Figure 15: Distribution of crash points in significant areas (first row - Rockville area and Wheaton, second row - Montgomery hills and White Oak)



Figure 16: Distribution of crash points in significant areas (first row - Rockville area and Wheaton, second row - Montgomery hills and White Oak)

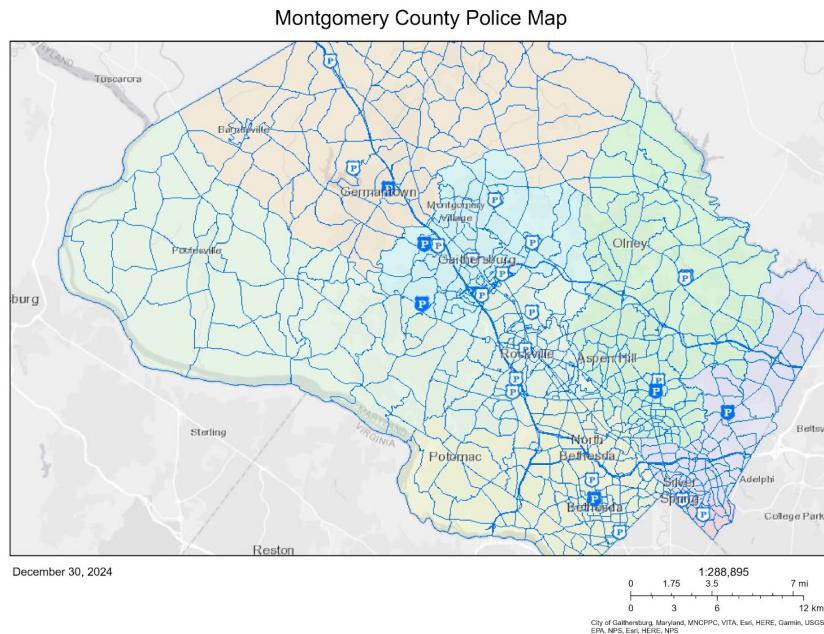


5 Discussion

This study found that there is enough evidence to reject hypotheses about spatial randomness in collisions and that there is a significant positive autocorrelation which means areas with large collision density tend to cluster together close to each other. By looking at a more granular level of individual census tracts and applying LISA, we observed an expected pattern that collisions tend to occur and cluster at most populated areas such as big towns and cities or loaded interstate highways. Similar pattern of clustering near populated towns is observed when considering only severe collisions. Tracts with significant autocorrelation and relatively high number of collisions: 70390, 700903, 70121, 70150, 701001

In contrast, when we considered not the counts of collisions, which depend heavily on the population of a tract, but the adjusted severity rate, LISA uncovered significant positively autocorrelated tracts located in more rural and less populated areas such as Poolesville and Damascus as tracts with relatively high severity rates for collisions. This observation suggests two plausible explanations for the observed pattern: these rural areas may be more hazardous due to poorer road conditions, inadequate lighting, insufficient traffic control measures, and more aggressive driving practices, or the remoteness of these areas from police stations (see Figure 17) may lead to underreporting of minor collisions, resulting in only severe crashes being recorded and thereby inflating the apparent severity rate. Tracts with significant autocorrelation and relatively high severity rate: 700500, 700204, 700207, 700209

Figure 17: Locations of Police Stations in Montgomery, MD



Certain crash conditions or contributing factors are modifiable and may be mitigated through targeted interventions. Analyzing crashes by specific conditions (Figure 14, 1st row, 2nd col) reveals a notable pattern for alcohol-related collisions. The analysis identified a single significant high-high (HH) cluster located in the Wheaton-Glenmont area, surrounded by significant low-low (LL) clusters. The factors contributing to the elevated incidence of alcohol-related collisions in this region remain unclear and require further investigation to inform effective intervention strategies.

Collisions involving animals tend to form significant clusters in relatively rural areas (Figure 14, 2st row, 2nd col) adjacent to forests and national parks. Although it may be impractical to eliminate these incidents entirely, certain interventions can mitigate their occurrence. For example, implementing reduced speed limits in zones where animals are most likely to cross the road can diminish collision risk and enhance overall safety or by constructing green bridges to enable wildlife to cross roads without having to negotiate traffic.

Collisions involving vehicles leaving the roadway (off road type) exhibit a similar spatial pattern of significant clusters, primarily concentrated in rural areas located in the eastern region of the county. This distribution likely corresponds to the presence of high-speed roads and highways in these areas.

Investigating areas with significant crash clustering (Figure 16), we observe that in densely populated regions, collisions frequently occur at intersections and road segments with pedestrian crossings, consistent with findings by Dezman et al. in Baltimore.

6 Limitations

The study is subject to several limitations that warrant consideration. Firstly, the reliance on police-reported data may introduce biases due to human error, misreporting, or omissions, and some collisions might not have been reported, potentially affecting the accuracy of the findings. Secondly, the analysis does not incorporate traffic flow or road network load data, which could enhance the estimation of accident probabilities across different road segments. Lastly, while the study identifies high-risk areas, it does not provide specific intervention strategies to reduce motor vehicle crash damage and fatalities. These limitations should be considered when interpreting the results and planning future research.

7 Conclusion

Identifying dangerous zones prone to motor vehicle crashes is a critical step toward developing a sustainable policy framework for reducing the damage caused by traffic accidents. This study demonstrates how spatial autocorrelation analysis and kernel density estimation can be applied to identify clusters with a significant concentration of crashes and to explore the spatial distribution of various modifiable risk factors, such as alcohol abuse, distracted driving, and collisions with animals. The findings from this research can serve as a guide for targeted interventions by policymakers, traffic planners, and other stakeholders to enhance road safety across the United States.

REFERENCES

- [1] Global Status Report on Road Safety 2023, World Health Organization. [link](#)
- [2] Hotspots and causes of motor vehicle crashes in Baltimore, Maryland: A geospatial analysis of five years of police crash and census data. Z. Dezman et al. 2016, [link](#)
(University of Maryland School of Medicine, Baltimore, Maryland, USA; State University of Maringá, Maringá, Brazil; Duke University, Durham, North Carolina, USA)

- [3] Wikipedia, Moran's I [link](#)
- [4] Local Indicators of Spatial Autocorrelation (LISA), Luc Anselin, 2020 [link](#) (GeoDA software)
- [5] Exploring spatial autocorrelation of traffic crashes based on severity. Ali Soltani, Sajad Askari, 2017, [link](#)
(Department of Urban Planning, Traffic and Transportation Research Centre (TTRC), Shiraz University, Iran)
- [6] Getis–Ord Spatial Statistics to Identify Hot Spots by Using Incident Management Data, Praprut Songchitruksa and Xiaosi Zeng, 2010 [link](#)
- [7] Spatial pattern identification and crash severity analysis of road traffic crash hot spots in Ohio, Md Saiful Alam and Nusrat Jahan Tabassum, 2023, [link](#) (University of Maryland, University of Toledo)
- [8] Montgomery County, Maryland, USA. Wikipedia, [link](#)
- [9] Automated Crash Reporting System (ACRS) Field reference guide, 2017, [link](#)
- [10] Crash Reporting - Incidents Data from 2015 to 2024, data.gov, [link](#)
- [11] Community report on tourism in Montgomery County, Visit Montgomery MD, 2022, [link](#)
- [12] Empirical Bayes Index (mentioned in lecture), Luc Anselin, 2016, [link](#)
(University of Chicago)
- [13] Laplacian Smoothing (Additive Smoothing), Wikipedia, [link](#)