Contents lists available at ScienceDirect

Accident Analysis and Prevention

Spatial analysis of crashes that ended in injured passengers in urban areas: a case study in Medellin, Colombia.

Mario A. Penaranda-Marquez[a,*,1],

[a] *Universidad Nacional de Colombia, Department of Civil Engineering, Infrastructure and Transportation, Medellín, Colombia.*

1. Introduction

The increasing population and vehicle fleet in countries raises the demand of transportation, so negative externalities such as road crashes must be considered. It is estimated that more than 1.3 million people die each year because of road traffic crashes, with more than 90% of them occurring in low to middle income countries (High-level Advisory Group on Sustainable Transport, 2016). Also, between 20 and 50 million people suffer non-fatal injuries, and this costs most countries 3% of their gross domestic product (GDP) (Santos et al., 2022; World Organization Health, 2022).

For instance, in 2021 Colombia reported 145,921 crashes, that resulted in 79,917 non-fatal injuries, 63,577 injuries and 7,238 deaths. The latter was distributed as follows: 59.7% motorcyclist, 21.4% pedestrians, 7.8% private vehicles and 6.4% cyclists (Agencia Nacional de Seguridad Vial, 2022). The main victims of road crashes that ended in injuries and deaths, are suffered principally by people among 20 and 29 years old and generate economic difficulties to people and their families and push hard over the health system due to the need of expensive treatments, loss of productivity and disabilities (Instituto Nacional de Medicina Legal y Ciencias Forenses, 2022).

It is noteworthy that 88.7% of these reported crashes took place in urban areas and slightly 80% of them occurred in 20 municipalities. An interesting situation occurred in Medellin, the second largest city of the country, which had most of the crashes with 43,482, and places above the capital of the country Bogota with 28,418 crashes (Agencia Nacional de Seguridad Vial, 2022). Another alarming figure is the fact that Medellin's records for 2021 places above the average registered between 2016 and 2020 (39,319) (Agencia Nacional de Seguridad Vial, 2022).

Traffic safety research has developed exponentially in the last decades (International Transport Forum, 2017) and principal efforts are focused on diminishing road crashes through the well-known 4E's countermeasures, which are safety enhancement programs that considers Enforcement, Education, Emergency response, and Engineering (Li et al., 2021). Local municipalities define their own road safety management programs to assess the behavior of the network to identify zones with higher rates of crashes. The first step in this process considers macro-level studies to detect the localities with priorities of investigation aggregated by *traffic analysis zones* (TAZ). These basic geographic units are used for inventorying demographic data and land use within a study area, and it is fundamental for transportation planners (Soltani & Askari, 2017).

For this purpose, geostatistical analysis has been gaining a spotlight as the primary tool to conduct macro-level analysis and it is based on two main approaches: spatial autocorrelation method (Getis & Ord, 1992) or density estimation methods (Sabel et al., 2005). The former incorporates spatial factors to identify location-specific influences on crash occurrence (Mitra & California Polytechnic State University, 2009), and the latter, is commonly used to detect patterns of crashes and analyzes the location without considering their attributes (Khan et al., 2023).

It is important to mention that there are two main categories of spatial autocorrelation: the global spatial analysis, which measures and tests the overall spatial phenomenon and identifies if the feature pattern is clustered, dispersed, or randomly distributed (Getis & Ord, 1992), and the local spatial analysis which measures the level of spatial association at the local scale (Ord & Getis, 1995).

Regarding spatial autocorrelation and density estimation methods, several studies were found using these techniques in the analysis of road crashes. For instance, Blazquez & Celis (2013) identified critical areas prone to high child pedestrian crash occurrence in Santiago, Chile. For this purpose, they used KDE and Moran's I index to assess the clustering pattern regarding some attributes such as age, gender, time of day, road type, among others. The results showed that zones with a high children population

and middle to low socioeconomic level who travel alone to and from school using public transportation are more exposed to pedestrian crashes. Also, there was a clustered pattern of the data at a 95% confidence level, with thresholds of 60 m, regarding the significant attributes like the imprudence of the pedestrian, drivers' violations, and others.

Soltani & Askari (2017) analyzed crashes hotspots considering land use, road network density, population density and the type of injure to determine clusters in Shiraz, Iran. To do so, they computed Moran's I and Getis-Ord indicator to measure spatial autocorrelation. The main results suggest that hotspots appear on arterial roadways and in urban activity centers, such as the central business district (CBD). Also, these zones are located where there is a considerable presence of large trip generators, such as hospitals, universities, shopping centers, among others.

Bassani et al. (2020) evaluated the effects of road network in Torino, considering injured and fatalities in crashes that involved vulnerable road users (VRU) such as pedestrians, bikers, and motorcyclists. They computed the NN index and the KDE function with a bandwidth of 100 m, which is nearly the average length of the arcs in the road network. The main results suggest a correlation between collisions and network structure, due to the concentration of records at the intersections of the main corridors of the city because of the higher speeds. Also, roads with more than two lanes per direction have a higher exposure to the interaction between vehicles and VRU. Finally, the NN indexes suggested that the crash data is clustered.

Khan et al. (2023) applied hotspot techniques to identify single-vehicle lane departure crash clusters across a road network based on injury severity in North Dakota. For this purpose, they estimated the Global Moran's I, the Local Ord-Gi and the KDE to crash records. The results drop that the global spatial autocorrelation was positive suggesting clustering of injured and property damage crashes in the study area. In relation to the KDE, the authors found the presence of clusters in the road network, specifically on curves, junctions, and intersections. Some crashes tend to cluster along the straight road segments that might be related with the circulation speed.

In this broad area, the United Nations has played a key role in recognizing that transportation and mobility are central to sustainable development, so they stated as a priority goal to improve road safety by halving the traffic road deaths by 2030 (High-level Advisory Group on Sustainable Transport, 2016). For this reason, the primary objective of this paper is to contribute to the state of the art in macro-level analysis considering crash records in the urban area of Medellin, Colombia by computing geostatistical analysis. Also, another goal is to model crash frequency to assess the association between risk factors and the road crash frequency on behalf of the contribution to transportation policy and road safety management programs.

2. Data & Methodology

2.1. Data collection and pre-processing

The source of data for the current study was a dataset containing crash records in the Medellin Metropolitan Area (MMA), and it was requested to the Department of Transportation (DOT) of the city. A pre-processing of the dataset was performed considering processes such as standardizing the information, filtering the data, filling missing values, removing rows with inconsistent information, transforming all the categorical features using dummy variables, and removing duplicate records.

Then, to compute a spatial analysis it was necessary to geocode the crash record addresses. To do so, the MapQuest API was used to determine the coordinates of the data. Using the Open Street Map API, the bounding-box coordinates were used to filter those crash records that fell outside the study area. Also, for the purposes of the current research, information gathered from the SIATA, a risk management authority in the MMA, was used to associate daily records of precipitation to the crashes within the dataset.

On the other hand, to conduct the macro-level analysis, it was necessary to define the TAZ. However, it is important to mention that although recent efforts have been made, the municipality of Medellin has not defined such spatial units. For this reason, the aggregation zones are defined as the 265 territorial administrative zones of Medellin, i.e., the neighborhoods of the city.

After the pre-processing, a segmentation of the information was defined considering the following criteria: for further analysis, the records correspond to crashes that resulted in injured people that involved vulnerable road users (VRU) such as bicycles and motorcycles. To have a better representation of the transportation pattern, only were considered Tuesday, Wednesday, Thursday between 5:30 and 17:30 hours. This criterion is usually used by the DOTs, to represent homogeneous conditions.

Finally, 13 characteristics related to demographic, date and time, location, weather, driver, accident type, type of infrastructure, number of vehicles involved in the crash, VRU or auto involved of 23,079 records were considered for the analysis of traffic collisions between 2016 to 2019.

2.2. Kernel Density Estimation (KDE)

It is the key tool for density-based methods, and it is a non-parametric method used for creating smooth maps of density values, i.e., the concentration of points at each location within the neighboring area (Aristizabal, n.d.) and has been broadly used in road safety (Newaz et al., 2017; Shafabakhsh et al., 2017). This function assumes that every individual event has an impact on the density in its spatial neighborhood (bandwidth), making use of a continuous, symmetrical, and decreasing probabilistic function through

a regression factor that depends on the type of function and the interpolation space surrounding in each point (Bassani et al., 2020). The KDE can be expressed as shown in equation 1:

$$\hat{f}(u,v) = \frac{1}{n \cdot h} \cdot \sum_{i=1}^{n} K\left(\frac{d_i}{h}\right) \quad (1)$$

Where $\hat{f}(u,v)$ is the crash density estimated at the location (u,v), n is the number of events, h is the bandwidth, K is the kernel function used for the computations.

2.3. Ripley's K Functions

It is a spatial analysis method used in point pattern analysis that aims to estimate the average number of events located within a radius of any typical event, but it is normalized for the density of the events over the same field of view (Amgad et al., 2015). In other words, Ripley's K-function compares the distribution of events in a field of view to determine if the point pattern is rather a Complete Spatial Randomness (CSR), or a "homogeneous Poisson process" (Amgad et al., 2015).

Some functions are important for current research. First, Ripley's G which tracks the proportion of points for which the nearest neighbor falls within a given threshold, and its plot illustrate the cumulative percentage against the increasing distance radiuses (Aristizabal, n.d.).

$$\hat{G}(x) = \frac{(d_{ik} \leq x, \forall i)}{n} \quad (2)$$

Second, Ripley's F analyses the distance to points in the pattern from locations in empty space, i.e., it characterizes the typical distance from arbitrary points in empty space to the point pattern (paezha, n.d.).

$$\hat{F} = \frac{(d_{ik} \leq x, \forall i)}{n} \quad (3)$$

Where d_{ik} is the distance from the point at i to its nearest neighboring event at location k.

2.4. Global Spatial Autocorrelation

It seeks to summarize the spatial distribution of events through statistical techniques to detect spatial patterns by considering both the location and the features values. These methods can determine if the pattern is clustered, random, or dispersed (Khan et al., 2023). A well-known indicator of spatial autocorrelation is the Global Moran's I, that is expressed as follows:

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{S_o \sum_{i=1}^{n} (x_i - \bar{x})^2} \; \forall i, j = 1, \ldots, n \quad (4)$$

Where x_i is the value of feature on location i, $\bar{x}$ is the feature mean, n is the total number of zones or locations, w_{ij} is the spatial weight representing the contiguity relationships between zones, and S_o is the sum of all spatial weights (Khan et al., 2023).

Moran's I value ranges from -1 to 1. Greater positive values show greater spatial dependence due to similar values, and lower values show spatial spreading. Finally, values near zero show a random distributed pattern (Soltani & Askari, 2017).

2.5. Local Spatial Autocorrelation

It is a clustering method to analyze sub-zones to determine whether a zone is a cluster of low values (a cold spot) or high values (a hot spot) (Soltani & Askari, 2017). A useful tool for this purpose is the Local Moran's I index, and is computed as follows:

$$I_i = \frac{z_i - \bar{z}}{\sigma^2} \sum_{j=1,\ j \neq i}^{n} \left[w_{ij}\left(z_j - \bar{z}\right) \right] \quad (5)$$

Where I_i is the local Moran's I index, z_i the value of the feature at location i, $\bar{z}$ is the feature mean, z_j is the value at all other locations considering that $j \neq i$, and σ^2 is the variance of z (Khan et al., 2023).

A high positive value of local Moran's index indicates the presence of spatial clusters, and high negative values represent spatial outliers (Khan et al., 2023). Moreover, it is important to mention that the local Moran's index generates four types of outcomes: 1) high-high clusters, zones with high values surrounded by high values, 2) low-low clusters, zones with low values surrounded by low values, 3) high – low clusters, zones with high values surrounded by low values, and 4) low-high clusters, zones with low values surrounded by high values (Khan et al., 2023).

2.6. Voronoi Polygons

To associate the records of precipitation to the crashes in the database, it was necessary to estimate the influence area of each pluviometer within the study area. For this purpose, the Voronoi polygons were used. This method aims to partition a plane containing n points into convex polygons, considering that each polygon only contains just one generating point and every point in a polygon is closer to its generating point in that another (Wolfram Mathworld, n.d.).

2.7. Spatial Regression

It is important to acknowledge that processes are not the same everywhere and geographical information is very useful to forecast outcomes of interest. In this sense, predictive methods such as spatial regressions explicitly consider these aspects considering statistical frameworks (Rey et al., n.d.).

There are two main approaches: s*patial heterogeneity* and s*patial dependence.* The former considers the geographic variations through spatial fixed effects (FE), so we let the constant α vary accordingly to different zone's characteristics (Rey et al., n.d.). This approach considers the

effect of each spatial unit instead of all the events in the study area and is known as *spatial fixed effects*. This is computed considering Equation 6:

$$C_i = \alpha_r + \sum_k X_{ki} \cdot \beta_k + \in_i \qquad (6)$$

Where C_i is the dependent variable, X is the set of covariates used to explain the dependent variable, β is a vector of the estimators of the regression, α the constant term, and $\in_i$ is the error of the regression (Rey et al., n.d.). It can also consider the spatial regimes, which assumes that the dependent variable varies according to the geographical pattern. Not only does the constant term vary but also other explanatory variables, also known as *spatial regimes* (Rey et al., n.d.).

$$C = \alpha_r + \sum_k X_{ki} \cdot \beta_{k-r} + \in_i \qquad (7)$$

Where β_{k-r} varies from each spatial unit. The latter approach is focused on the effect that neighbors have in the observation, so it is relevant the spatial configuration of the observations, and the extent to which that influences the outcome we are considering (Rey et al., n.d.). Spatial dependence can be introduced in several ways. One way is through *spatially lagged exogenous effects*, in which spatial lag is considered as an explanatory factor of the dependent variable in relation to its neighbors (Rey et al., n.d.):

$$C = \alpha_r + \sum_{k=1}^{p} X_{ij} \cdot \beta_j + \sum_{k=1}^{p} \left(\sum_{j=1}^{N} w_{ij} \cdot x_{jk}\right) \cdot y_k + \in_i \quad (8)$$

Where $\sum_{j=1}^{N} w_{ij} \cdot x_{jk}$ is the spatial lag of the kth exploratory variable. Another way, is to include it as a *spatial error model*, in which the spatial lag is included in the error term of the equation (Rey et al., n.d.):

$$C = \alpha + \sum_k \beta_k \cdot X_{ki} + u_i \qquad (9)$$

Where $u_i = \sum_j w_{ij} \cdot \in_i$. And finally, the *spatial lag model* in which the spatial lag is introduced in the exploratory variables (Rey et al., n.d.):

$$C = \alpha_r + \rho \cdot C_{lag-i} + \sum_k \beta_k \cdot X_{ki} + \in_i \qquad (10)$$

3. Results

This section provides the findings for the point pattern analysis, hotspot technique identification, and the spatial regression model for crashes that involved VRU and resulted in injured people in Medellin. To illustrate the spatial distribution of the crash records, the Kernel Density Estimation (KDE) was determined as shown in Figure 1. The blue and purple zones show a high concentration of crashes, which in other words represents the probability density estimation of a crash occurrence. The higher values are concentrated in the downtown area and across the Medellin River.

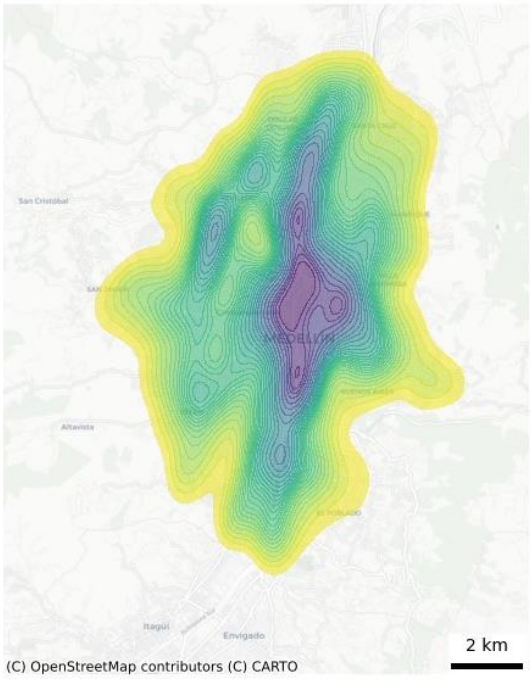

Figure 1. Kernel Density Estimation with 30 levels for the crashes in Medellin.

Ripley's Functions were also estimated. It is important to mention that due to the heavy computational requirements for this analysis, the functions were computed considering a sample of 1,000 crashes. In Figure 2, Ripley's G function is depicted, and the behavior of the observed data show a rapidly increase, so the point pattern tends to be clustered. This function suggests that more than 40% of the crashes have their nearest neighbor within a radius of 50 m and more than 80% less or equal to 250 m. This result is interesting considering that the average length of a block in Medellin is nearly 80 m, suggesting that crashes tend to cluster among blocks.

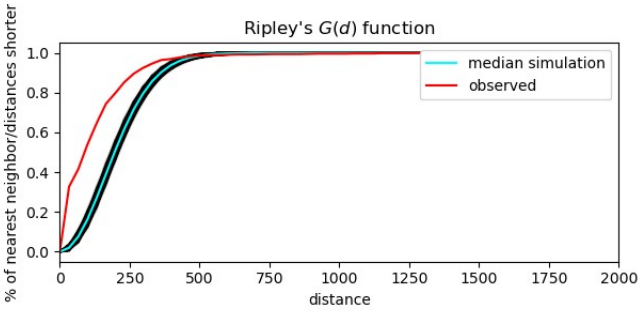

Figure 2. Ripley's G function.

On the other hand, Ripley's F function is depicted in Figure 3. For this case, the curve from the observed crashes decreases rapidly below the curve of the simulation, so the point pattern has large empty areas. This information supports the fact of clusters within the data.

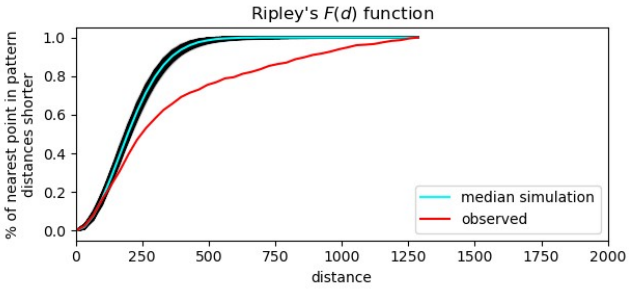

Figure 3. Ripley's F function.

As mentioned before, the current macro-level analysis was conducted considering the neighborhood division in Medellin. The main attributes used in this process considered the number of crashes, average number of vehicles involved, and number of crashes with VRU involved. Some choropleth maps were estimated to illustrate the spatial distribution of the crashes.

Figure 4 illustrates the distribution of the total number of crashes in Medellin. It is noteworthy that most of the crashes are in the center zone and following the alignment of the city. In the top ten of the neighborhoods with more crashes are Perpetuo Socorro, La Candelaria, San Benito (La Candelaria), and Campo Amor, Santa Fé (Guayabal). The former group is in the downtown of the city with multiple land use, while the latter is in the south zone of Medellin, which is mainly an industrial land use.

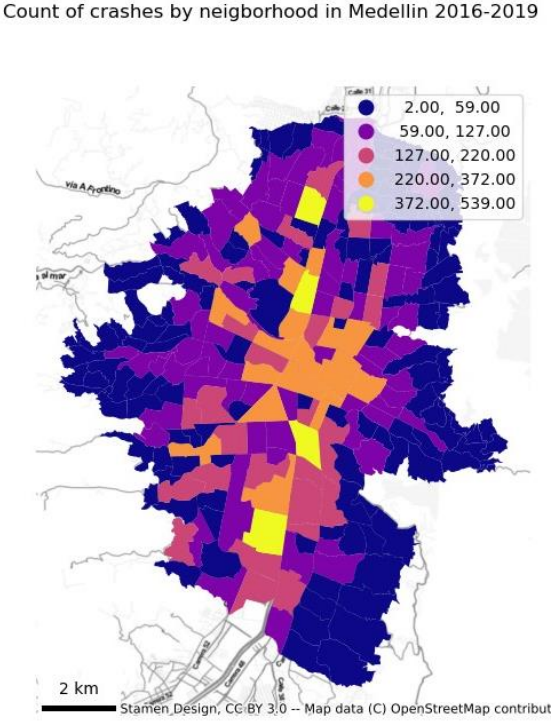

Figure 4. Spatial distribution of the number of crashes in Medellin.

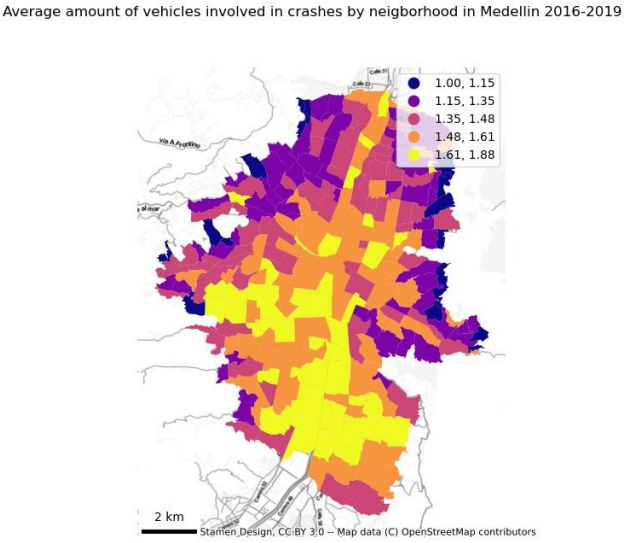

Figure 5. Spatial distribution of average vehicles involved in crashes in Medellin.

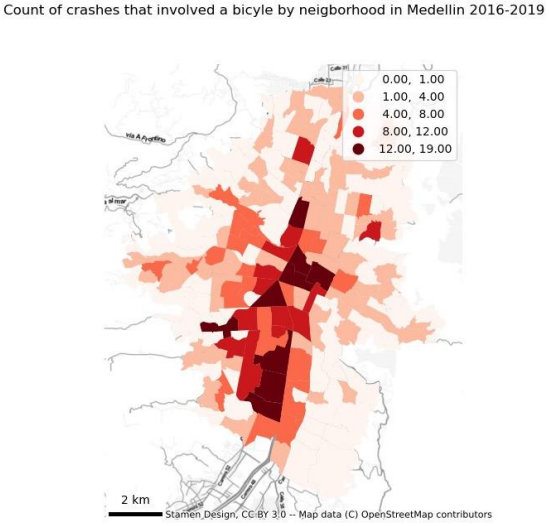

Figure 6. Spatial distribution of crashes that involved bicycles in Medellin.

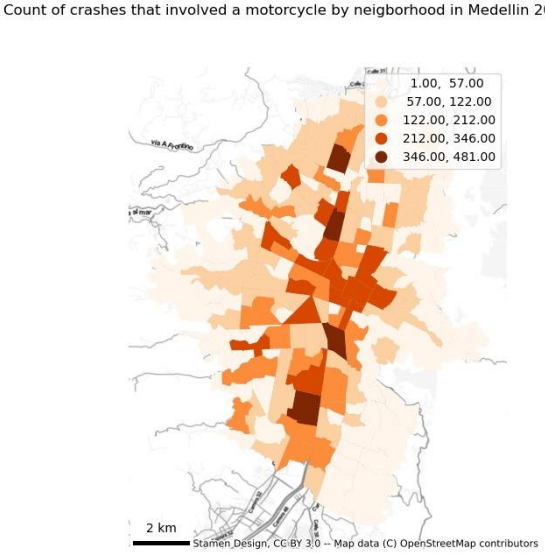

Figure 7. Spatial distribution of crashes that involved motorcycles in Medellin.

Figure 5 depicts the average amount of vehicles involved in crashes. The spatial distribution changes because the downtown does not have the highest average. For this case, the highest averages tend to be in the southern zone of the city, with a decreasing tendency to the northern side. The neighborhoods with the higher average of vehicles involved in crashes are Miravalle and Rosales (Belén), La Castellana and San Joaquín (Laureles-Estadio), Alejandría and Villa Carlota (El Poblado), and La Alpujarra (La Candelaria).

Figure 6 illustrates the number of crashes that involved bicycles. Here, it is important to note the tendency of the data, especially over the southern zones and the center of the city. The highest number of crashes is in Los Conquistadores (Laureles-Estadio) and San Benito, Guayaquil, Corazón de Jesús, La Candelaria (La Candelaria).

Figure 7 shows the number of crashes that involved motorcycles. It is important to mention that the value in this map is like the values shown in Figure 4, suggesting that most of the crashes analyzed involved a motorcycle as a main actor. Also, the occurrence of the crashes tends to be in the downtown area. The neighborhoods with the higher crashes that involved motorcycles are Perpetuo Socorro (La Candelaria), Campo Amor and Santa Fé (Guayabal), Castilla and Caribe (Castilla), San Benito and La Candelaria (La Candelaria), and Los Conquistadores, Los Colores, and Carlos E. Restrepo (Laureles-Estadio). An interesting situation occurs in the neighborhoods located in commune Laureles-Estadio, because its main land use is devoted to residential areas.

Although choropleths draw relevant information, it is also important to conduct a deeper analysis with statistical tools. In this sense, we call the global and local spatial autocorrelation methods. For this purpose, it is necessary to estimate the spatial weight matrix to consider the vicinity among the spatial units. In the current research, it is only considered first-order neighborhood, i.e., only adjacent neighbors have an interaction in the occurrence of crashes. In this sense, it is enough to have a common edge or border to influence the neighbor, so the queen contiguity was adopted for the analysis. For consistency, it was necessary to drop neighborhoods such as La Avanzada (Popular), Calazans Parte Alta (La América), El Rincón (Belén) and Aures No. 2 (Robledo) because they are “islands” and do not have a common boundary.

To estimate the effect of the neighbors onto the spatial unit analyzed (spatial lag), the spatial weight matrix is then standardized and the proportion of the feature of each neighbor is computed. The global indicator of spatial autocorrelation (GISA) and local indicators for spatial autocorrelation (LISA) were calculated for the number of crashes, averages vehicles involved, and VRU involved in crashes using Global and Local Moran’s I statistic, respectively. The results are shown in Table 1 and Figure 8, Figure 9, Figure 10, and Figure 11.

For the GISA, the results in Table 1 show that for all the features considered the Moran’s I is positive. This situation suggests a clustering tendency for the crashes in Medellin with a confidence of 95%. The lower statistic values are for the features number of crashes and crashes that involved motorcycles, with 0.419 and 0.413 respectively. On the other hand, the average number of vehicles involved and crashes that involved bicycles have a higher Moran’s I, with 0.510 and 0.517 respectively. This means that the pattern for the latter variables is better defined and a meso or micro level analysis is necessary to gain insights.

Table 1. Moran's I GISA for the variables considered.

Variable	Moran's I	p-value
number_crash	0.419	0.001
aver_veh	0.510	0.001
bici_involved	0.517	0.001
moto_involved	0.413	0.001

For the LISA, it is noteworthy that the analyzed features tend to cluster high-high values (shown in red) in the center and southern zone of Medellin. In other words, high values surrounded by neighbors with high values are in communes such as La Candelaria, Laureles-Estadio, and Guayabal. On the other hand, the low-low values (shown in blue) tend to cluster in the outer zones of the city. The location of these crashes corresponds to the outer zones of Medellin, which are characterized by the high slopes and are usually isolated communities in which there is a lack of public control, and the report of crashes might be scarce.

Other important results are obtained by analyzing the low-high values (shown in light blue), which are zones with low values that are surrounded by high values. A particular case is depicted in Figure 8, where low-high values are found in Ecoparque Cerro El Volador and Parque Cerro Nutibara, which are two of the seven tutelary hills of Medellin. For obvious reasons, these zones have a low number of crashes within the city. In Figure 10 it is particularly curious that the low-high values are found in U.P.B, Carlos E. Restrepo, and Suramericana neighborhoods. The interesting thing about them is that they have the largest cyclist infrastructure in the city. So, this result suggests that, at least during the analysis period, the reports of crashes involving cyclists were low and the use of bicycles lanes is useful for road safety.

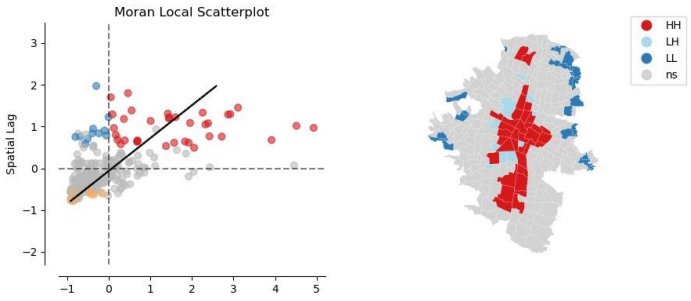

Figure 8. LISA for the number of crashes in Medellin.

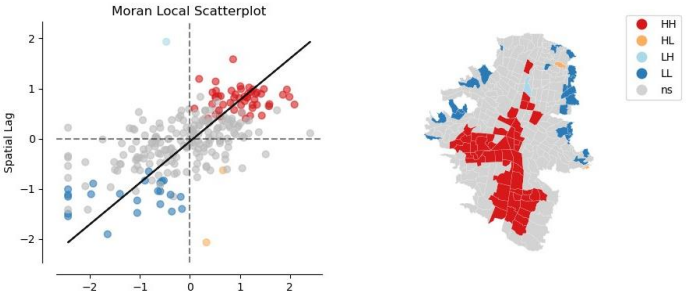

Figure 9. LISA for the average number of vehicles involved in crashes in Medellin.

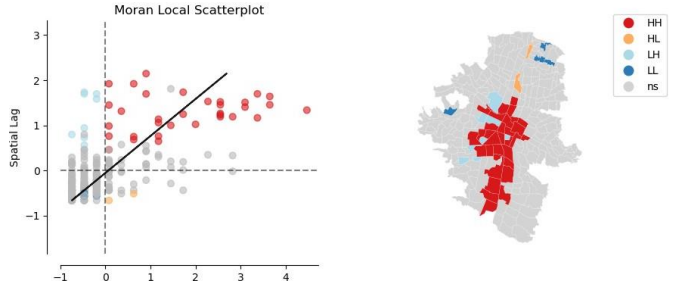

Figure 10. LISA for the crashes with bicycles involved.

Moto involved

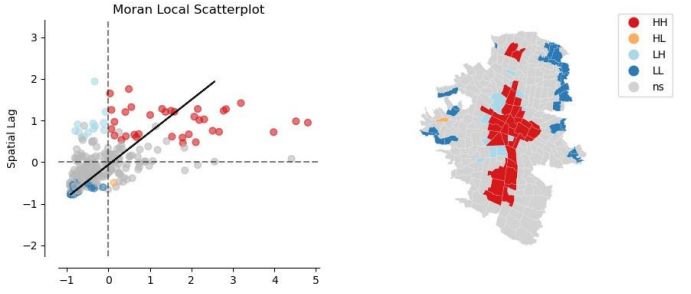

Figure 11. LISA for the crashes with motorcycles involved.

Finally, as a predictive tool to model the crash frequency in the neighborhoods of Medellin, spatial regressions were determined. For this purpose, additional information was gathered from open sources of Medellin's town hall. To enhance the variables included in the model, geographic information regarding the socioeconomic status, the number of signalized intersections, and the number of control cameras. Also, daily precipitation records were requested from SIATA to associate the climate variable temporally to the crash dataset. This information contained more than 70 pluviometers that had records, but it was necessary to assign these values to the neighborhoods in Medellin. To do so, Voronoi polygons were determined to identify the zones for which a pluviometer has a range of representativity. Then, a temporal association was computed to consider the daily variability within regions.

As shown above, all the association methods suggest that the crash records tend to cluster, so the approach to the spatial regression should consider the influence of neighbors. For this reason, the modeling process is done considering spatial dependence. Several models were computed, and the results showed that the number of vehicles involved, and the average age of the drivers were not significant for the linear model. Consequently, the crash frequency expected for a day in Medellin can be forecasted with the socioeconomic status, precipitation, number of signalized intersections, and number of controls diapositives in the specific neighborhood. The correlation matrix was computed for the numerical variables in the database and are shown in **Error! Reference source not found.**. The model with the best performance is shown in Table 2.

Figure 12. Heatmap correlation matrix for the independent variables.

Table 2. Estimators of the spatial error model for the number of crashes per neighborhood.

Variable	Std. Error	z-Statistic	Probability
α	0.124	92.783	0.000
X_E	0.003	-6.796	0.000
X_P	0.002	3.187	0.001
X_S	0.009	9.442	0.000
X_D	0.034	0.003	0.000
u_i	0.029	0.007	0.000

The spatial regression computed is based on a spatial error model, and as can be seen in Table 2, all the variables are statistically significant with a 95% confidence. It is important to note that the constant term, α, has an estimator of 1.157, which means that if all the variables

were zero, one could expect at least 1 crash per neighborhood in Medellin. Moreover, a particular situation occurs with the social class variable because the results show that socioeconomic influences the crash frequency in Medellin. One can say that the estimator of this variable reduces the crash frequency if the analyzed neighborhood is classified as high-income. On the other hand, the model shows that the precipitation has a negative impact on the crash frequency. Although its estimator is small, the increase of precipitation favors the crash frequency in Medellin.

Finally, a counter expected result is obtained from the number of signalized intersections and the number of control dispositive. According to the spatial error model, if the number of these variables increase, the number of crashes should also increase. One can say that the fact that these variables are prone to guarantee road safety, but in Medellin it seems to have an opposite effect. However, it is important to mention that more accurate models should include endogenous variables to capture the behavior of drivers in the road network.

4. Discussion

The 2017 passenger origin-destination study found nearly 6 million trips take place in the MMA (Área Metropolitana del Valle de Aburrá, 2017). The rush hour in the morning is between 6-7 and in the evening from 17-19 (Área Metropolitana del Valle de Aburrá, 2017). The major reasons to use transportation in the MMA are due to commuters and students and the modal split is distributed as follows: 16.2% use transit transportation, 12.3% motorcycle, 8.1% metro line, 13.6% automobile, 6.5% taxi, 1% bicycle, and 3.7% others (Área Metropolitana del Valle de Aburrá, 2017). It is also important to note that in 64% of households there is at least one motorcycle (Área Metropolitana del Valle de Aburrá, 2017).

This study also found that La Candelaria was the most visited commune in Medellin, accounting for 723,544 daily trips (11.8%), followed by El Poblado with 447,616 (7.3%), Laureles-Estadio with 361,772 (5.9%), and Belén 318,850 (5.2%) (Área Metropolitana del Valle de Aburrá, 2017). The importance of La Candelaria is due to it is in the downtown area of the city, which is the Central Business District of the city. To show the importance of this commune, it is important to mention that nearly 33 tons of freight enter this area daily which accounts for nearly 6,600 freight trips (Universidad Nacional de Colombia & Area Metropolitana del Valle de Aburrá, 2019).

Regarding the Global Moran's I, showed in Table 1, it suggests a clear pattern of clustering of the crash records in Medellin. It is noteworthy that the tendency is in the downtown area and through the southern zones, and it is important to mention that this behavior is due to several aspects. First, as mentioned before, the downtown area is an important zone for business and commuters, so the conflict between road users favors crashes that ended in injuries in the zone. Second, the alignment of the crashes matches the Medellin River, where the Southern Freeway and Regional Highway are located. These major corridors are crucial for the regional connectivity between Medellin and the Coffee Region located in the south-western zone of Colombia. Due to its importance, many large freight trip generators, such as factories, warehouses, and marketplaces, are located in the vicinity of these roads, generating a complex interaction in the entrance and exits in the road (Universidad Nacional de Colombia & Area Metropolitana del Valle de Aburrá, 2019). Added to the fact that these corridors have a maximum speed of 80 kilometers per hour.

On the other hand, Medellin has promoted the use of sustainable transportation modes, such as the bicycle. Recent studies have suggested that the use of bicycles have increased by 3% and the trips went up from 60,000 daily trips on average, to 210,000 (AMVA). This figure is considerable because of the huge investment by the local authorities. In Table 3 are shown the neighborhoods in Medellin with the most cyclist infrastructure.

Table 3. Bikes paths by neighborhood in Medellin.

Commune	Neighborhood	Bike path length (m)
Belen	Fátima	6564.20
Guayabal	Parque Juan Pablo II	5486.18
Laureles	Suramericana	4671.61
Laureles	Los Conquistadores	4668.44
Guayabal	El Rodeo	4558.86
Aranjuez	Jardín Botánico	4379.51
Laureles	San Joaquín	4301.21
Laureles	Carlos E. Restrepo	4148.88
La Candelaria	Jesús Nazareno	4115.92
Aranjuez	Sevilla	4101.40

As mentioned in the LISA for the crashes that involved bicycles, Suramericana, Carlos E. Restrepo, and U.P.B were found to have low-high records. In other words, those zones registered low crashes, but their neighbors registered a high number of crashes. As can be seen in Table 3, these neighborhoods are in the top 10 spatial units with the largest amount of bike paths in the city and are in residential areas, but they have common edges with the major corridor of the city, Autopista Sur. This analysis suggests that when cyclist infrastructure is guaranteed, users can travel with high standards of road safety. However, when the users travel in a neighboring zone, the lack of infrastructure derivates in crashes. So, it is important to guarantee connectivity of the cyclist infrastructure to reduce the crashes.

In relation to the spatial regression, the precipitation variable was found to be statistically significant for the model. However, it is important to mention that the

estimator is almost zero, so the influence of it is very small for the crash frequency model. Although this seems to be a counter intuitive result because we expect the crash frequency to increase due to the reduction of the friction factor due to wet pavement, several authors have found that precipitation positively influences the reduction of crashes because drivers are more aware of the dangers of the road (Omranian et al., 2018; Theofilatos, 2017; Zeng et al., 2020).

5. Conclusions

Spatial analysis provides powerful tools to draw conclusions regarding the spatial distribution of crashes in an area of study. These methods are very useful in the first steps of a road safety management program, in order to identify the zones with the higher number of crashes to develop meso and micro-level analysis.

The CBD of Medellin is an important zone to develop economic activities and to study, for this reason further analysis should be done to enhance road safety. Other initiatives such as the Urban Air Protected Zones, which diminish the accessibility to vehicles to specific blocks within the downtown of Medellin, are necessary to reduce the number of crashes.

As could be seen in Figure 7, most of the crashes involved motorcycles. For this reason, it is important to pay attention to the public policy related to this mode of transportation in order to reduce injured people and its associated costs.

The local authority has encouraged the use of sustainable modes of transportation, such as the bicycle. This has been seen through the investment in bike paths in Medellin. However, as was shown in Figure 10 in the LISA, the low-high values are neighborhoods with considerable length of bike paths that registered low number of crashes and have neighboors with high records. In this sense, it is important to prioritize the investment in new bike paths considering the existing infrastructure, because the lack of connectivity seems to affect road safety. However, this macro-level analysis only puts in the spotlight interesting points, and more research is needed to assess this situation.

For further research, an additional segmentation should be considered by ranges of hours, weeks, or months in order to gain deeper insights into the behavior of road crashes in Medellin. Also, the integration of

6. References

Agencia Nacional de Seguridad Vial. (2022). *Anuario Nacional de Siniestralidad Vial - Colombia 2021*. https://ansv.gov.co/sites/default/files/2022-07/Anuario_Nacional__2021_Vfinal.pdf

Amgad, M., Itoh, A., & Tsui, M. M. K. (2015). Extending Ripley's K-function to quantify aggregation in 2-D grayscale images. *PLoS ONE*, *10*(12). https://doi.org/10.1371/journal.pone.0144404

Área Metropolitana del Valle de Aburrá. (2017). *Encuesta Origen Destino*. https://www.metropol.gov.co/observatorio/Paginas/encuestaorigendestino.aspx#header-web

Aristizabal, E. (n.d.). *Point Pattern Analysis*. Retrieved June 25, 2023, from https://github.com/edieraristizabal/AnalisisGeoespacial/blob/master/Notebooks/10_PointPatternAnalysis.ipynb

Bassani, M., Rossetti, L., & Catani, L. (2020). Spatial analysis of road crashes involving vulnerable road users in support of road safety management strategies. *Transportation Research Procedia*, *45*, 394–401. https://doi.org/10.1016/j.trpro.2020.03.031

Blazquez, C. A., & Celis, M. S. (2013). A spatial and temporal analysis of child pedestrian crashes in Santiago, Chile. *Accident; Analysis and Prevention*, *50*, 304–311. https://doi.org/10.1016/J.AAP.2012.05.001

Getis, A., & Ord, J. K. (1992). The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis*, *24*(3), 189–206. https://doi.org/10.1111/J.1538-4632.1992.TB00261.X

High-level Advisory Group on Sustainable Transport. (2016). *Mobilizing Sustainable Transport for Development*. https://sustainabledevelopment.un.org/content/documents/2375Mobilizing%20Sustainable%20Transport.pdf

Instituto Nacional de Medicina Legal y Ciencias Forenses. (2022). *Forensis 2020 - Datos para la vida*.

International Transport Forum. (2017, December 15). *Benchmarking Road Safety in Latin America*. www.itf-oecd.org

Khan, I. U., Vachal, K., Ebrahimi, S., & Wadhwa, S. S. (2023). Hotspot analysis of single-vehicle lane departure crashes in North Dakota. *IATSS Research*, *47*(1), 25–34. https://doi.org/10.1016/J.IATSSR.2022.12.003

Li, Y., Karim, M. M., Qin, R., Sun, Z., Wang, Z., & Yin, Z. (2021). Crash report data analysis for creating scenario-wise, spatio-temporal attention guidance to support computer vision-based perception of fatal crash risks. *Accident Analysis and Prevention*, *151*. https://doi.org/10.1016/j.aap.2020.105962

Mitra, S., & California Polytechnic State University, S. L. Obispo. C. of Engineering. Dept. of C. & E. E. (2009). *Enhancing road traffic safety : a GIS based methodology to identify potential areas of improvement.* https://doi.org/10.21949/1503647

Newaz, K. M. S., Hasanat-E-Rabbi, S., & Miaji, S. (2017). Spatio-temporal study of road traffic crash on a national highway of Bangladesh. *2017 4th International Conference on Transportation Information and Safety, ICTIS 2017 - Proceedings*, 60–66. https://doi.org/10.1109/ICTIS.2017.8047743

Omranian, E., Sharif, H., Dessouky, S., & Weissmann, J. (2018). Exploring rainfall impacts on the crash risk on

Texas roadways: A crash-based matched-pairs analysis approach. *Accident Analysis & Prevention*, *117*, 10–20. https://doi.org/10.1016/j.aap.2018.03.030

Ord, J. K., & Getis, A. (1995). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis*, *27*(4), 286–306. https://doi.org/10.1111/J.1538-4632.1995.TB00912.X

paezha. (n.d.). *Chapter 15: Point Pattern Analysis IV*. Retrieved June 25, 2023, from https://paezha.github.io/spatial-analysis-r/point-pattern-analysis-iv.html#f-function

Rey, S., Arribas-Bel, D., & Wolf, L. (n.d.). *Spatial Regression*. Retrieved June 25, 2023, from https://geographicdata.science/book/notebooks/11_regression.html

Sabel, C. E., Kingham, S., Nicholson, A., & Bartie, P. (2005). *Road Traffic Accident Simulation Modelling-A Kernel Estimation Approach*.

Santos, K., Dias, J. P., & Amado, C. (2022). A literature review of machine learning algorithms for crash injury severity prediction. *Journal of Safety Research*, *80*, 254–269. https://doi.org/10.1016/j.jsr.2021.12.007

Shafabakhsh, G. A., Famili, A., & Bahadori, M. S. (2017). GIS-based spatial analysis of urban traffic accidents: Case study in Mashhad, Iran. *Journal of Traffic and Transportation Engineering (English Edition)*, *4*(3), 290–299. https://doi.org/10.1016/j.jtte.2017.05.005

Soltani, A., & Askari, S. (2017). Exploring spatial autocorrelation of traffic crashes based on severity. *Injury*, *48*(3), 637–647. https://doi.org/10.1016/j.injury.2017.01.032

Theofilatos, A. (2017). Incorporating real-time traffic and weather data to explore road accident likelihood and severity in urban arterials. *Journal of Safety Research*, *61*, 9–21. https://doi.org/10.1016/j.jsr.2017.02.003

Universidad Nacional de Colombia, & Area Metropolitana del Valle de Aburrá. (2019). *Estudio de transporte de carga en el Area Metropolitana del Valle de Aburra*. https://www.metropol.gov.co/movilidad/Documents/Estudio-de-transporte-de-carga-en-el-Area-Metropolitana-del-Valle-de-Aburra.pdf

Wolfram Mathworld. (n.d.). *Voronoi Diagram*. Retrieved June 25, 2023, from https://mathworld.wolfram.com/VoronoiDiagram.html

World Organization Health. (2022, June). *Road Traffic Injuries*. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries#:~:text=Approximately%201.3%20million%20people%20die,pedestrians%2C%20cyclists%2C%20and%20motorcyclists.

Zeng, Q., Hao, W., Lee, J., & Chen, F. (2020). Investigating the Impacts of Real-Time Weather Conditions on Freeway Crash Severity: A Bayesian Spatial Analysis. *International Journal of Environmental Research and Public Health*, *17*(8), 2768. https://doi.org/10.3390/ijerph17082768