sta465 project

Jiawei Gong

January 2025

1 Introduction

1.1 Background

From the data published by Toronto Transportation Services, there are 695 motor vehicle collisions with killed or seriously injured persons within City of Toronto in 2023, which means 2 serious collision occur per day occur in average. It's not surprised that collision may happen in cluster in some neighborhoods due to specific traffic factors. These reasons can be road condition, driver condition, and so on. City of Toronto government has put efforts to prevent collision by building more road safety infrastructure such as traffic camera and street light poles. However, the budget is limited. It's necessary to take full advantage of infrastructure by making sure they are built in the the places where they are needed most, in other words, where collision occurs frequently.

1.2 Research Question

Does these road safety infrastructure effectively help to reduce the occurrence of vehicle collisions in Toronto?

1.3 Datasets

There are seven datasets. (1) Five multipoint shapefiles consist of coordinates of different road safety infrastructure, including red light camera, traffic signal pole, street light pole, traffic calming devices, and traffic cameras. (2)Another shapefile contains the details including coordinates of every motor vehicle collisions with killed or seriously injured persons from 2006 to 2023. The associated circumstances including driver condition, road condition, and premise type in the details are used to categorize collisions for further study of examining the effectiveness of certain type of road safety infrastructure on certain type of circumstance. (3) Finally, the shapefile of the 'Toronto Neighbor' contains 140 neighborhoods in the form of polygons in Toronto, and provides a clear boundary of each neighborhood, so that data in other shapefile can be grouped based on which neighborhood they are located in.

2 Method

2.1 Data cleaning

All dataset are transformed to WGS84 datum and projected to UTM zone 17N when spatial computation is performed. There are 18957 collisions from 2006 to 2023. However, the location of infrastructure can changes with time, and lots of infrastructure are built every year. Therefore, to be up-to-date, we only focus on the collision from 2020 to 2023, and there are 2725 collisions in total. Then, the dataset of collisions and five road safety infrastructure are spatially merged into the neighborhoods. For each neighborhood, the counts of collisions and five types of safety infrastructure are computed for the next step.

To answer the research question in a holistic way, we can decompose it into several parts:

2.2 Spatial data visualization

At the detail level of eyeballing, heatmap of counts of collisions and five types of safety infrastructure are plotted. Besides straightforward visualization of the spatial patterns of clustering of collisions and infrastructure, these heatmap can help to have a rough sense of the effectiveness of infrastructure, by examining whether the location of cold spot of collisions and hot spot of infrastructure are close geographically. If close, it implies such infrastructure is effectively constructed to prevent collisions. Otherwise, such infrastructure is not in the optimal locations.

2.3 Spatial data exploration

Although traffic collision is in the form of point pattern data, testing for complete spatial randomness is discarded because traffic collision mostly occurred in the roads. So inherently the collisions will present a clustering pattern around anywhere with roads. Hence, proving the statistically significant of clustering or regular pattern by K-function, G-function and so on essentially provide no useful information.

Alternatively, focus is put on checking the clustering of neighbors in terms of collision counts. Globally, the evidence of spatial autocorrelation in neighborhoods in Great Toronto Area for the number of motor vehicle collisions is examined, by using Moran's I. Locally, Local Moran's I statistics are used to explore which specific areas have the significant spatial autocorrelation.

Moran's I
Assumption:

- Moran's I, as a global index, quantifies the degree to which a variable is spatially clustered, dispersed, or random, by calculating the weighted average of similarity between areal units
- Spatial data is assumed to be stationary, meaning the mean is constant among areal units (Neighbors) are consistent across the study area (Toronto)
- The spatial pattern is entirely represented in the spatial weight matrix

Formulation:

$$I = \frac{1}{s^2} \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i \sum_j w_{ij}}$$

- $s^2 = \frac{\sum_i (y_i - \bar{y})^2}{n}$
- y_i = counts of collisions in each neighbor
- $\bar{y} =$ average counts of collisions across all neighbors
- w_{ij} is the weight that defines proximity.

Distribution:
Null Hypothesis: No spatial autocorrelation in neighborhoods, i.e. y_i is iid.

$$I \sim Normal(E(I), V(I))$$

- $E(I) = -\frac{1}{n-1}$
- $V(I) = \frac{ns_1 - s_2 + 3\left(\sum_i \sum_j w_{ij}\right)^2}{(n-1)(n+1)\left(\sum_i \sum_j w_{ij}\right)^2} - E(I)^2$

Weight Matrix
As Moran's I is sensitive to the chosen weight matrix w_{ij}, weight matrices by Queen connectivity, 6NN connectivity and 3NN inverse distance will be calculated and applied by row standardization. Then, we will select the weight matrix with highest Moran's I, because such matrix can best capture the spatial autocorrelation.

- Queen Connectivity weight matrix: A single shared boundary point means they are connected. i.e. w_{ij}=1 if $Neighbor\_i$ and $Neighbor\_j$ share any boundary point.
- 6NN Connectivity weight matrix: Top 6 Nearest distance between centroids of areal units. i.e. w_{ij}=1 if the centroid of $Neighbor\_j$ is top 6 nearest to the centroid of $Neighbor\_i$
- 3NN Inverse distance weight matrix: the weight of neighboring areal unit is the inversed squared distance between the centroids of two areal units. '3NN' means only top 3 nearest area units will be regarded as neighbors and given the weights.

$$w_{ij} = \frac{1}{(distance(Neighbor\_i, Neighbor\_j))^2}$$

Row standardization will be applied to create proportional weights in cases where features have an unequal number of connections. Every weight w_{ij} is divided by the number of the connections of areal units.

Next, the value of Global Moran's I for different lags will be examined to check if the statistically significant spatial autocorrelation is mainly a local phenomenon or a broader scope, by Correlogram

Compute Moran's I under Monte Carlo
Mechanism:

1. Observations(counts of collisions) are assigned at random in the areal units (neighbors) for a large number of times
2. For each bootstrap permutation, Moran's I is calculated.
3. Across the simulations, the empirical mean E(I) and variance V(I)is computed based on the empirical distribution
4. Compute the statistic: $\frac{observed-expected}{s.d.(expected)}$. Then compute the p-value by the computing the proportion of simulated values as extreme or more extreme than the statistic observed.

Advantages:

- This method is more robust because it does NOT rely on the assumption of normality of data.

Local Moran's I
Assumption:

- Local Moran's I for each unit gives an indication of the extent of significant spatial clustering of similar observations located around that unit.
- Test statistics are generated under randomization.
- Patterns of low-low, low-high, high-low and high-high can reveal the local spatial autocorrelation

Formulation:

$$I_i = \frac{y_i - \bar{y}}{s} \sum_{j=1}^{n} w_{ij} \frac{(y_j - \bar{y})}{s}$$

- Calculating I for each areal unit

2.4 Spatial data modeling

For each type of infrastructure, we want to know if they prevent collision. In general, we can model the relationship between the count of collisions and the count of each type of traffic infrastructure using Simultaneous Autoregressive (SAR) and Conditional Autoregressive (CAR) models. These two models can illustrate the signifiance of covariates after accounting for the spatial pattern.

SAR Spatial Log-Error Models
Assumption:

- The spatial autocorrelation in both Y and ϵ is accounted. i.e. SAR lag-Error is fitted with the addition of a spatially lagged dependent variable plus the spatially correlated errors,.

Formulation:

$$Y = X\beta + \epsilon$$

- $\epsilon = (I - \lambda W)^{-1}(Y - \rho WY - X\beta)$. ϵ is the residuals adjusted for both the spatial lag and the spatial error.
- W is the weights matrix
- ρ, λ are the spatial autoregressive parameter
- β is the regression coefficients
- X is the covariates. In this case, counts of (1)red light camera (2)street light pole (3)traffic signal pole (4) traffic camera (5)calming device
- Y is the response variable. In this case, the count of collision in each neihgbor

Algorithm:

- The goal is to maximize the likelihood:

$$L(\rho, \lambda, \beta, \sigma^2 \mid Y, X) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\epsilon'\epsilon + \log|I - \rho W| + \log|I - \lambda W|$$

- $\beta, \rho, \sigma, \lambda$ are optimized by Newton-Raphson method.

CAR Models
Assumption:

- Each observation $Y(s_i)$ is conditional on the values of all other observations, $p[Y(s_i) \mid Y_{-i}]$.
- The state of a particular area is influenced only by its neighbors and not neighbors of neighbor. It assumes Markov random field property holds.

Formulation:

- $Y_i \mid Y_{-i} \sim N\left(\sum_{j \in i} \beta_{ij} Y_j, \sigma_i^2\right)$.
- β, ρ, σ are optimized by MLE and Newton-Raphson.

2.5 Spatial data modeling comparison

For two models:

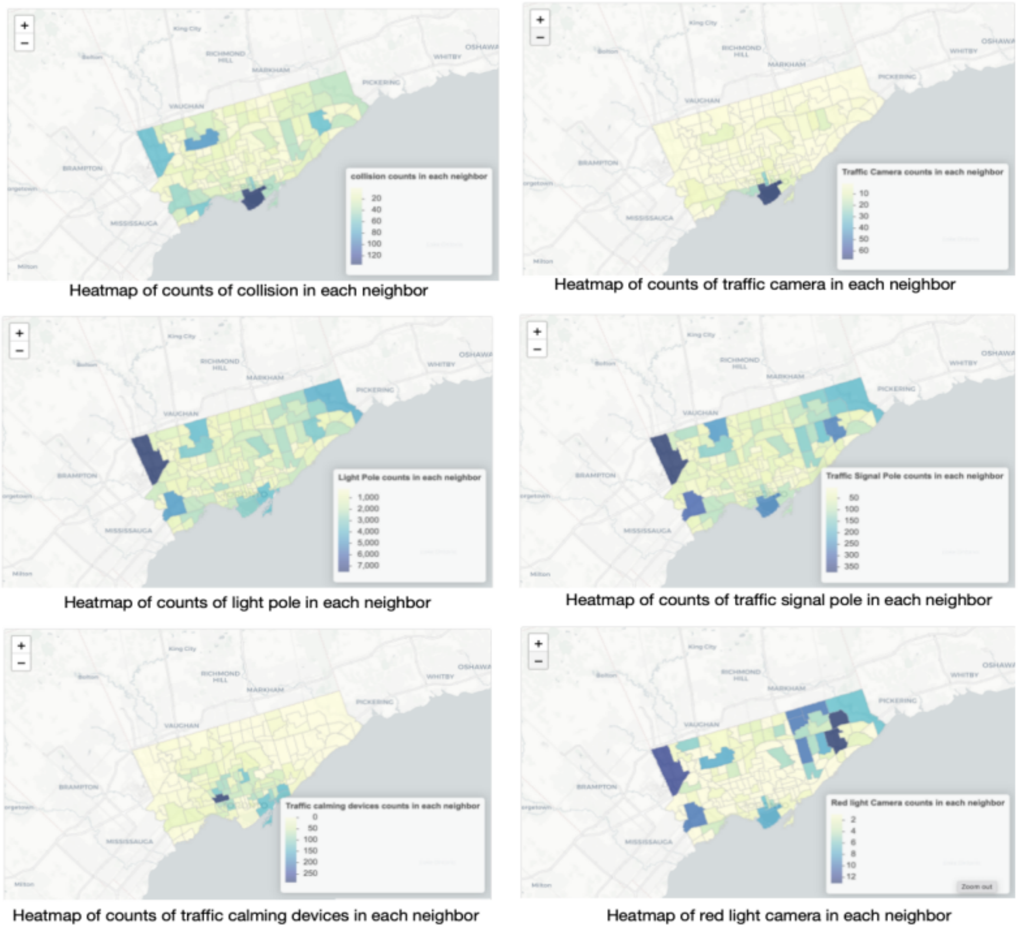
Interpretations: The sign and associated statistical significance of the coefficients before each predictor will be interpreted to illustrate the effectiveness of traffic infrastructure. If applicable, the spatial lag parameter and spatial error parameter will be interpreted.

Diagnostics: Moran's I will be calculated to test the spatial autocorrelation in the residuals. The p-value will be analyzed to show if the model does fully capture the spatial structure of the data. In other words, if it captures the effects of the predictors and the spatially structured unobserved factors.

Goodness of fit: AIC, SSE and likelihood ratio test value will be calculated and compared mutually to find the best model fit.

3 Result

3.1 Spatial data visualization


+
−
King City
Bolton
RICHMOND HILL
MARKHAM
WHITBY
OSHAWA
PICKERING
VAUGHAN
BRAMPTON
MISSISSAUGA
Milton
collision counts in each neighbor
20
40
60
80
100
120
Traffic Camera counts in each neighbor
10
20
30
40
50
60
Light Pole counts in each neighbor
1,000
2,000
3,000
4,000
5,000
6,000
7,000
Traffic Signal Pole counts in each neighbor
50
100
150
200
250
300
350
Traffic calming devices counts in each neighbor
0
50
100
150
200
250
Red light Camera counts in each neighbor
2
4
6
8
10
12
Zoom out

Heatmap of counts of collision in each neighbor

Heatmap of counts of traffic camera in each neighbor

Heatmap of counts of light pole in each neighbor

Heatmap of counts of traffic signal pole in each neighbor

Heatmap of counts of traffic calming devices in each neighbor

Heatmap of red light camera in each neighbor

Figure 1: Heatmaps of collisions and five types of traffic safety infrastructure in each neighbor

Firstly, from the heatmap of counts of collisions in each neighbor, a general spatial pattern of the location of the hotspot, which refers to the clusters of neighbors with high counts of collisions, can be observed. The waterfront community neighbor has the most count of collisions, while the east and west sides of Toronto show clusters of less higher counts of collisions. Secondly, by eyeballing the heatmaps of five traffic safety infrastructure, a preliminary analysis of the effectiveness of five infrastructure in preventing collisions can be conducted, and there is a basic assumption about interpretation will be followed: If higher counts of infrastructure and lower counts of collisions are in the same region, this means this type of infrastructure might effectively prevent collisions in this region. For two examples, (1) there is only one hotspot around the waterfront community in the heatmap of traffic camera, meaning traffic cameras do not effectively prevent collisions. (2) There are higher counts of light pole in the east and west sides, and less light pole in the waterfront community. This pattern is the opposite of the pattern of collisions, meaning the light pole might effectively prevent the collisions. To be concise, a table of the results of analysis of the effectiveness of five type of infrastructure is presented below.

Table 1: Comparison of pattern of counts of collision and infrastructure

	Waterfront Community	**East Side**	**West Side**	**Potential Effectiveness**
Collisions	Higher Counts	Lower Counts	Lower Counts	/
Traffic Camera	Higher Counts	Lower Counts	Lower Counts	No
Light Pole	Lower Counts	Higher Counts	Higher Counts	Yes
Traffic signal Pole	Lower Counts	Higher Counts	Higher Counts	Yes
Traffic Calming Device	Higher Counts	Lower Counts	Lower Counts	No
Read light Camera	Lower Counts	Higher Counts	Higher Counts	Yes

The choice of using kernel density map is discarded because the collisions and infrastructure concentrate on the roads in every neighbors, making the density map output less informative than heatmap of counts.

3.2 Data description

Table 2: One example row in the dataset. There are 140 rows.

Neighbor	**#Collisions**	**#Traffic Camera**	**#Street Light Pole**	**#Traffic signal Pole**	**#Traffic Calming Device**	**#Read light Camera**	**Geometry**
Yonge-St.Clair	11	1	595	16	82	1	POLYGON ((-79.39119 43.6810...

Note traffic calming device refers to the implementation of speed cushion, speed bump and traffic island in Toronto. Traffic calming device is designed for slowing down traffic and improving road safety.

3.3 Spatial data exploration

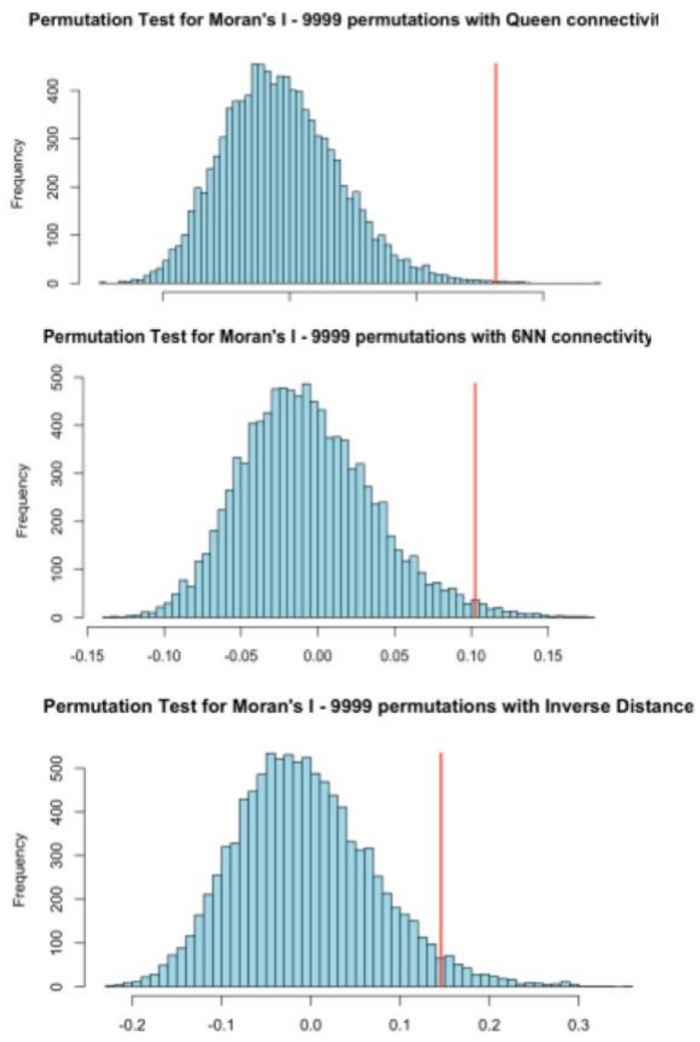
Firstly, construct a queen, 6-NN, and inverse distance based adjacency matrix. Row standardization is applied, so the sum of weights in a row is one.

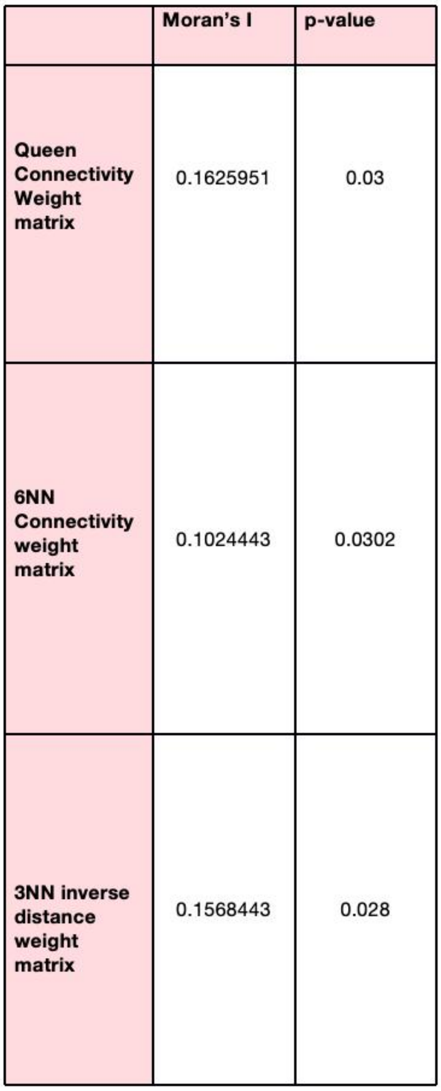
Table 3: Comparison of weights and number of links of three types of weight matrices

	Average number of links	**Min Weight**	**Mean Weight**	**Max weight**
Queen Connectivity weight matrix	5.97	0.09091	0.16746	0.33333
6NN Connectivity weight matrix	6	0.1667	0.1667	0.1667
3NN Inverse distance weight matrix	3	0.06957	0.33333	0.71199

For Queen Connectivity weight matrix, max weight as 0.33333 means that the number of connection by 'Queen' is at least 3. max weight as 0.09091 means that the number of connection by 'Queen' is at most 11. For 6NN Connectivity weight matrix, all the weights is 0.1667, because every areal unit has 6 connections by 6NN method. The average numbers of links of Queen and 6NN connectivity are both approximately 6. For 3NN inverse distance weight matrix, larger distance will lead to a smaller inverse distance weight, and smaller distance will lead to a larger inverse distance weight. Therefore, closer neighors are given a higher weight. Note that the value of inverse distance weight can not reflect the number of connections like Queen and 6NN connectivity do.

Using Monte Carlo with 9999 permutation, Global Moran's I and associated p-values are calculated and compared to find the largest Global Moran's I.





	Moran's I	p-value
Queen Connectivity Weight matrix	0.1625951	0.03
6NN Connectivity weight matrix	0.1024443	0.0302
3NN inverse distance weight matrix	0.1568443	0.028

Figure 2: Plots of 9999 permutations and observed Morans's I in red line; Table of Moran's I value and associated p-value for three types of weight matrix

The test results indicate a statistically significant positive spatial autocorrelation for counts of collisions across neighbors in Toronto, when using the row standardized weight matrix by Queen connectivity, 6NN connectivity, and 3NN inverse distance. The p-values (0.03, 0.0302, 0.028) with 'two.sided' are all less than 0.05, meaning the result is statistically significant. The Moran I statistic (0.1625951, 0.1024443, 0.1568443) are all positive, meaning neighboring areal units have similar values than distant units. In other words, neighborhoods in Toronto with similar collision concurrence (either high or low counts) are more likely to be located near each other.

Queen weight matrix is chosen for the further analysis, because it shows a larger Moran's I, meaning it can capture the spatial autocorrelation better.

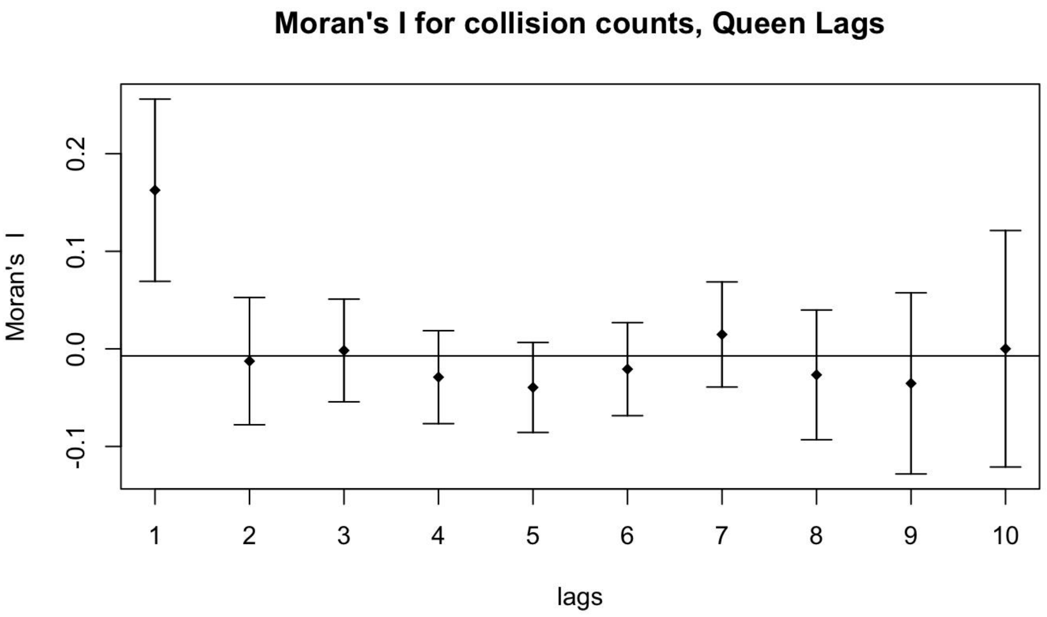


Figure 3: Moran’s I for collision counts, Queen Lags

From the plot of the value of Moran’s I for collision counts for from 1to 10 Queens Lags, Moran's I is largest when lag=1, and then Moran's I decrease to negative value, and then go back to the value around zero. This shows that strong positive spatial correlation is mainly a local phenomenon. In other word, similarities of collision counts are primarily among immediate neighbors, instead of distant neighbors.

To explore the specific location of clusters of neighbors with similar collision counts, Local Moran’s I is examined.

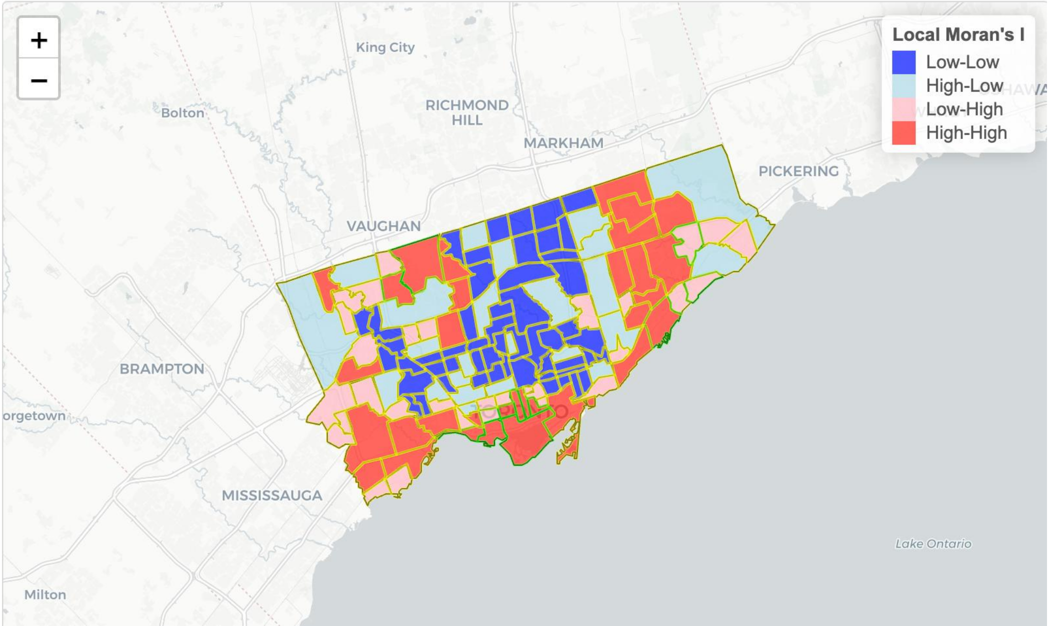


Figure 4: Map of the local Moran’s I with low-low, high-high, low-high, high-low regions annotated

From the map, there are two huge High-High clusters at the east to the middle lake shore, and the west side of Toronto. Also, there is a small cluster in the northeast part. For Low-Low pattern, there is a huge cluster in the middle part of Toronto. In contrast, High-Low and Low-High do not form clusters. Green boundary indicates the significance of local Moran's I, while yellow boundary indicates insignificance. After doing p-value correction by bonferroni adjustment, the proportion of significant local Moran's I is very small, meaning the spatial autocorrelation is a global phenomenon rather than local phenomenon.

In this context, High-High cluster represent urban centers (e.g. waterfront community) where high traffic density increases the likelihood of collisions. High-High cluster also includes the regions such as Scarborough, which have many highways. The high speed of vehicles can serve as a potential reason for the clustering of high counts of collisions. Low-Low regions might represent regions with lower traffic and fewer residents, or areas with effective safety infrastructure in place. High-Low and Low-High regions could be caused by the local safety infrastructure is well placed in only one neighbor instead of implementation on large areas.

3.4 Spatial data modeling

Since a significant positive Moran's I has indicated the presence of a clustered spatial pattern, the linear regression model is expected to perform poorly and is therefore discarded. Instead, models that can accounts for spatial pattern in the dataset are built. SAR Lag-Error and CAR models are fitted with covariates of counts of five traffic safety infrastructure and response of counts of collisions. Due to the right-skewed distribution of counts of collisions and five infrastructure, log transformation is applied to both the response and predictors to make distribution closer to normal for a better model fit. To handling large amount of zero values, a constant 1 is added, i.e. Y' = log(Y + 1).

Table 4: Description of the estimated coefficients of five infrastructure and associated significance, and spatial parameters

	Traffic camera	**Light pole**	**Traffic signal pole**	**Traffic calming device**	**Red light camera**	**Rho**	**Lambda**
SAR Lag-Error	0.2454*	0.2471	-0.3826*	0.0037	-0.3274*	-0.0663	0.3572*
CAR	0.2528*	0.2739	-0.4482*	0.0371	-0.2590	/	0.0888

Note: (*) means this estimated parameter is statistically significant. Otherwise, ()

SAR Lag-Error Model Interpretation:

Only counts of (1)traffic camera, (2)traffic signal pole, and (3)red light camera are statistically significant, after accounting for the spatial effect. The coefficient estimates of (2)traffic signal pole and (3)red light camera are negative, while the coefficient of (1)traffic camera is surprisingly positive. For significant negative coefficients, the statistical interpretation is that the number of collisions decreases as the number of these significant infrastructure, which indicates (2)traffic signal pole and

(3)red light camera effectively prevent vehicle collisions. However, positive value actuality implies that the number of collisions increases as the number of these significant infrastructure. One possible explanation is that Toronto city government decided to build more (1)traffic cameras in the neighbors to prevent collisions, after they discover collisions occurred more frequent there. However, the effect is delayed. Additionally, this results match the comparison result of heatmaps where the hotspot of traffic cameras and collisions is in the similar location.

The spatial lag parameter (rho) is -0.0663, and is statistically insignificant (p-value > 0.005). This negative value and insignificance indicates no significant spatial autocorrelation in the response variable. This suggests that the spatial lag effect of neighboring regions' collision counts on the response variable is negligible. This results contradicts our expectation as previously calculated Moran's I is positive. One possible explanation is that observed clustering might be explained by spatial error dependence (lambda) or spatially clustered covariates, rather than direct dependence of collision counts on neighboring values. The spatial error parameter (lambda) is 0.3572, and is statistically significant (p-value < 0.005). This suggests that there is spatial autocorrelation in the residuals. And there are unobserved spatial factors that are causing higher or lower collisions exhibit some spatial clustering across neighborhoods.

CAR Model Interpretations:
All the significance and sign of coefficient estimates are same, except the predictor, (3)red light camera number, turns into insignificant, after accounting for the spatial effect. This suggests number of collisions has no relation with number of (3)red light cameras by CAR model.

The spatial error parameter (lambda) is 0.0888, and is statistically insignificant (p-value > 0.005). This suggests that spatial autocorrelation in the residuals is not strong enough to conclude that there are unobserved spatial factors affecting the response variable.

3.5 Spatial data model comparison

Table 5: Comparison of goodness of fit of models

	AIC	SSE	Likelihood ratio test	Moran's I for residuals
SAR Lag-Error	254.81	43.3970	4.5212*	0.0032
CAR	256.69	45.71	0.6462	0.0672*

AIC & SSE Comparison:
SAR Lag-Error model has smaller AIC and SSE than CAR model, indicating SAR Lag-Error is better fitted.

Likelihood Ratio Test Comparison:
For SAR model, the associated likelihood ratio test is 4.5212 and statistically significant (p-value < 0.05). This indicates the response variable is fitted better by SAR Lag-Error Model than the linear regression model, because it accounts for spatial autocorrelation both in the dependent variable and unobserved factors.

For CAR model, the p-value>0.05 indicates that the inclusion of lambda does not significantly improve the model fit, since there is no spatial autocorrelation in the error term. This suggests included covariates have accounted for the spatial pattern in the response variable, because safety infrastructure are clustered themselves.

Residual Moran's I Comparison:

Residual Moran's I for SAR model is positive, and NOT statistically significant(p-value > 0.05). This suggests spatial autocorrelation is NOT present in the residuals, which means the SAR Lag-Error model has fully captured the spatial structure of the data. In contrast, the statistically significant (p-value < 0.05) Residual Moran's I for CAR model indicates the spatial autocorrelation is present in the residuals. CAR model does not capture the spatial structure of the data. This result matches the previous result that the CAR model is a worse fit.

4 Conclusion

In general, SAR model perform better for a more global spatial autocorrelation, while CAR perform better relatively local spatial autocorrelation. After computing local Moran's I, the proportion of significant statistics turns out small, which indicates the weak local spatial autocorrelation in our data. This explains why SAR model is a better fit according to the comparison result.

Answering the research question, only building more traffic signal pole and red light camera has been statistically proved that they can effectively prevent vehicle collisions. Otherwise, there is no pronounced evidence that building street light pole, traffic camera and calming devices can prevent collisions.

5 Limitation & Potential Solution

Limitation 1: In this project, for a given domain, there is a underlying assumption: decreased counts of collision and increased safety infrastructure at the same time reflect the effectiveness of infrastructure. However, temporal factors, the occurrence time of collisions and completion time of infrastructure, are overlooked. Building more infrastructure can be served as the measure to help the regions where collisions are frequent. This will make the hotpots of collisions and infrastructure be in the same place. The effect can be delayed but we don't take this into consideration. With knowing this, we can not rigorously conclude that significant negative coefficients reflect the effectiveness of infrastructure. Therefore, this bias lowers the reliability of our conclusion.

> **Possible Solution:** One way to improve this analysis is that we can do spatial analysis on the effectiveness of infrastructure after it has been built. For example, we will fit a spatial model with response variable being the collisions in end of 2022, and covariates being the infrastructure completed in start of 2022. In this way, we can examined if the infrastructure effectively prevent collisions in 2022.

Limitation 2: There are latent factors affecting the number of collisions and infrastructure in a neighbor. It's crucial to realize that different neighborhood may have different total length of roads. The total length of road can affect the number of collisions and infrastructure, because more roads lead to higher probability of collision and more space for infrastructure. In other words, for a neighborhood, it's more reasonable to look at the number of collision with knowing the length of roads.

> **Possible solution:** Dataset needs to be normalized. The purpose of calculating the ratio of the number of collisions and infrastructure to the length of roads is to account for the effect of different lengths of roads in neighbors. There is a shapefile containing all the roadway, high-way, and any way that a vehicle can pass in Toronto. Using this dataset to calculate the ratios can inform whether several neighborhood really have more frequent than others given the length of roads, and therefore make the conclusion more rigorous

6 Future direction

The dataset of collision can be explored further. Collision can be categorized by their driver condition, road condition, and premise type.

- For driver condition, there are (1)drunk driver (2)speeding driver (3)careless driver
- For road condition, there is (1)light condition
- For premise type, there are (1)red lights (2)Stop Signs (3)No control

This allows us to examine the effect of one type of infrastructure on one specific type of collision. For example, we can build the autoregression model to assess the effect of the number of street light pole on the number of collisions in dark night , in order to see whether these poles really reduce the number of collision during night.

7 Data Source

All data are obtained from City of Toronto's Open Data:
https://open.toronto.ca/dataset/neighbourhoods/
https://open.toronto.ca/dataset/motor-vehicle-collisions-involving-killed-or-seriously-injured-persons/
https://open.toronto.ca/dataset/traffic-cameras/
https://open.toronto.ca/dataset/traffic-calming-database/
https://open.toronto.ca/dataset/topographic-mapping-poles/
https://open.toronto.ca/dataset/red-light-cameras/