Exploring Temporal and Spatial Patterns of Major Crime Indicators*

An analysis for for Toronto Urban Safety and Policy Design

Yuechen Zhang

November 26, 2024

This paper analyses spatial and temporal trends in major crime in Toronto using data on key crime indicators from 2014 to the present. Linear regression models were used to examine crime categories, temporal patterns, and their prevalence in the community. The findings in the paper suggest that certain crime categories, such as auto theft and robbery, exhibit significant fluctuations over time, with higher frequencies in urban areas compared to suburban neighbourhoods. In addition, the findings highlight that year-to-year changes in the number of offences are influenced by the prevailing socio-economic and other factors. These insights underscore the importance of data-driven strategies for crime prevention and urban policy planning, thereby ensuring a trend towards safer community and social development

Table of contents

*Code and data are available at: https://github.com/Yuechen-Zhang603/Major-Crime-Indicators

1 Introduction

Despite numerous initiatives aimed at reducing crime in urban areas, major crimes such as assaults, robberies and auto thefts continue to be a significant challenge for communities throughout Toronto. Crime not only jeopardises public safety, but also affects socio-economic stability and the quality of life of residents. Within the past year, the government announced $390 million to help stop crime and violence (**Canada_2023?**). While the state has invested heavily in intervening in the crime situation, major crimes like assaults, robberies, and auto thefts have been prevalent over the past decade, with significant increases in certain categories.In 2014, there were 32,461 major crimes reported in Toronto, and in 2024, this number increases to 35,950, a 10.7 per cent increase in ten years.The number of major crimes reported in Toronto has risen from 1.7 per cent in 2014 to 1.3 per cent in 2024, which is a significant increase. These statistics highlight the need for a deeper understanding of the factors that contribute to crime in order to develop more effective policies to enhance public safety.

Research into crime trends often leverages data from law enforcement agencies to identify the causes and predictors of crime rates. The dataset used in this study, the Toronto Major Crime Indicators dataset, provides detailed records of reported crimes across multiple categories, including assault, robbery, break and enter, auto theft, and theft over. By analyzing this dataset, we can uncover temporal and spatial crime patterns and understand how crime has evolved over the years in Toronto's neighborhoods.

This paper builds models to analyze Toronto's crime patterns, focusing on the years 2014 and 2024. Using linear regression models, we examine the relationships between major crime categories, annual trends, and neighborhood-specific data. These models help us explore questions such as whether certain crime categories are more likely to increase over time or whether specific neighborhoods experience higher crime rates consistently.

Our findings reveal that major crimes such as auto theft and robbery show disproportionate increases, with particular neighborhoods seeing higher crime prevalence. These insights emphasize the importance of targeted policies for crime reduction and resource allocation. Additionally, by comparing 2014 and 2024 data, we highlight the growing severity of crime in Toronto and suggest data-driven approaches to mitigate these issues.

The remainder of this paper is structured into different sections. Section 2 demonstrates the data used for our report and includes some tables and graphs to illustrate the different groups of people in our data. Section 3 builds the model and discusses its justification and explanation. Section 4 highlights the results of the predictions using tables and graphs. Section 5 contains discussions that conducted based on the findings, which addresses the poverty status results based on mortgage states and income levels. Statistical programming language R (R Core Team 2023) is used in this report, with packages `tidyverse` (**citeTidyverse?**), `here` (**citeHere?**), `rstanarm` (**citeRstanarm?**), `modelsummary` (**citeModleSummary?**), `ggplot2` (**citeGgplot2?**), `knitr` (**citeKnitr?**), `marginaleffects` (**citeMarginalEffects?**), plotly (**citePlotly?**), tibble (**citeTibble?**), margins (**citeMargins?**), testthat (**citetestthat?**) and kableExtra (**citeKableExtra?**).

1.1 Estimand

2 Data

2.1 Data Overview

This report uses the Key Crime Indicators dataset from the website open data toronto to provide comprehensive data on reported criminal activity as our primary data source. These data include all MCI incidents reported to the Toronto Police Service, including those for which the location could not be verified. (**opentorontodata?**). The authenticity of the data can be guaranteed because the publisher clearly states that the data excludes incidents that are considered unfounded. Following a police investigation, it was confirmed that the reported crime did not occur and that there were no attempts to be recognized as unfounded incidents by Statistics Canada.

For the data section, this paper will focus on analyzing trends over time (e.g., monthly, yearly) based on different variables in the dataset to identify patterns or spikes in certain types of crime. And using geographic data to map crime hotspots and understand spatial distribution so that rates and types of crime can be compared across neighborhoods.

After loading the dataset using the R programming language (R Core Team 2023) and the here package (**citehere?**), the tidyverse (**citetidyverse?**) package was used to generate graphs. In doing so, R code was adapted from (**tellingstorieswithdata?**).

2.2 Data Table

2.3 Features

The original SPM 2019 dataset, which shows in **?@tbl-raw** in Appendix Section A, contains 157959 data entries and many variables. Since it is difficult to observe such a large dataset, this report will only explore and analyze through several data features. We first chose these 17 variables: h_seq, spm_poor, spm_tenmortstatus, spm_totval, spm_snapsub, spm_caphousesub, spm_schlunch, spm_engval, spm_wicval, spm_fedtax, spm_eitc, spm_actc, spm_fica, spm_sttax, spm_childsuppd, spm_capwkccxpns, and spm_medxpns. Their explanation are in **?@tbl-feature**. We can use those features (other than h_seq, spm_poor, spm_tenmortstatus) to calculate the income after tax, by adding all the subsidies and benefits and substracting all the expenses (just like how the SPM does it (**sgp2019?**)). This way the income is able to reflect the financial situation on households more accurately.

Table 1: Top 10 Neighborhoods by Total Crime Incidents

Neighborhood	Assaults	Break & Enters	Drug Arrests	Thefts	Total Incident
Waterfront Communities-The Island (77)	9978	3012	0	0	1548
Church-Yonge Corridor (75)	8927	2138	0	0	1366
West Humber-Clairville (1)	3103	1522	0	0	1132
Bay Street Corridor (76)	6713	1761	0	0	1035
Moss Park (73)	5847	1605	0	0	921
Downsview-Roding-CFB (26)	4660	971	0	0	812
York University Heights (27)	3776	1345	0	0	798
Kensington-Chinatown (78)	4531	1551	0	0	739
Islington-City Centre West (14)	2797	1326	0	0	719
Woburn (137)	4131	927	0	0	693

Figure 1: Toronto Major Crime Indicators

2.4 Data Measurement

The data was mainly collected by surveys conducted by USCB. Portions of SPM resources, such as values of taxes and some non-cash benefits, were not asked of respondents and were estimated using models. However, the estimated amounts are not perfect. For example, they found that the estimated total benefits from both the Earned Income Tax Credit and the child tax credit are lower than those found when examining federal income tax returns (**spm?**).

The unit of measurement for numerical data are in US Dollars. For other data, they are categorical features with no units.

2.4.1 Data Consideration

The data are estimates based on a sample of the population, so it cannot represent the whole population of the United States. This creates margins of error. Also, since the SPM poverty estimates are from household survey data, the data might be affected by issues such as under-reporting, over-reporting, or making up data. Some people might not remember the exact value of some data, for example, not all people have or know that they receive income from government benefit programs, so they might be reporting it inaccurately. In addition, some SPM resources, such as taxes and non-cash benefits, are estimated using models. The models use administrative data, but the estimated amounts are not perfect.

2.5 Methodology

The original dataset includes duplicated responses from

After cleaning, 62990 rows of data with 3 data features remain. **?@tbl-clean-data** shows a preview of the cleaned dataset.

Table 2: Preview of the cleaned crime analysis dataset

Event ID	Year	Month	Crime Type	Neighborhood	Latitude	Longitude	Crime Count	NA	NA
1	GO-20141261609	2014/1/1	2014/1/1	2014	January	1	1	Wednesday	9
2	GO-20141260033	2014/1/1	2013/12/31	2014	January	1	1	Wednesday	2
3	GO-20141260127	2014/1/1	2014/1/1	2014	January	1	1	Wednesday	1
4	GO-20141260597	2014/1/1	2014/1/1	2014	January	1	1	Wednesday	2
5	GO-20141260618	2014/1/1	2014/1/1	2014	January	1	1	Wednesday	5
6	GO-20141260618	2014/1/1	2014/1/1	2014	January	1	1	Wednesday	5
7	GO-20141260056	2014/1/1	2014/1/1	2014	January	1	1	Wednesday	1
8	GO-20141263672	2014/1/1	2014/1/1	2014	January	1	1	Wednesday	18
9	GO-20141262608	2014/1/1	2014/1/1	2014	January	1	1	Wednesday	14
10	GO-20141263706	2014/1/1	2014/1/1	2014	January	1	1	Wednesday	18

2.6 Data Visualization

Figure 2 provides a detailed overview of the total reported crimes in Toronto from 2014 to 2024. It illustrates significant fluctuations in crime rates over the years, with key trends emerging from the data. Starting in 2014, the total reported crimes were relatively stable at approximately 35,000 incidents. A gradual increase is observed from 2015 to 2018, reaching around 40,000 incidents by 2018.

A notable peak occurs in 2019, with total reported crimes surging to over 45,000, suggesting a significant spike in criminal activity during that year. This could potentially be linked to socio-economic changes or other external factors influencing crime rates during this period. Following this peak, a decline is observed in 2020 and 2021, where crime rates drop back to approximately 35,000, potentially reflecting changes during the COVID-19 pandemic, including lockdowns and reduced public mobility.

However, in 2022, crime rates sharply rise again, exceeding 45,000 incidents, marking the highest point in the observed period. This sudden increase may reflect a post-pandemic resurgence of activity and mobility in urban areas. By 2024, the total reported crimes decline significantly, falling back to below 40,000 incidents, indicating potential effectiveness of recent crime prevention strategies or other mitigating factors.

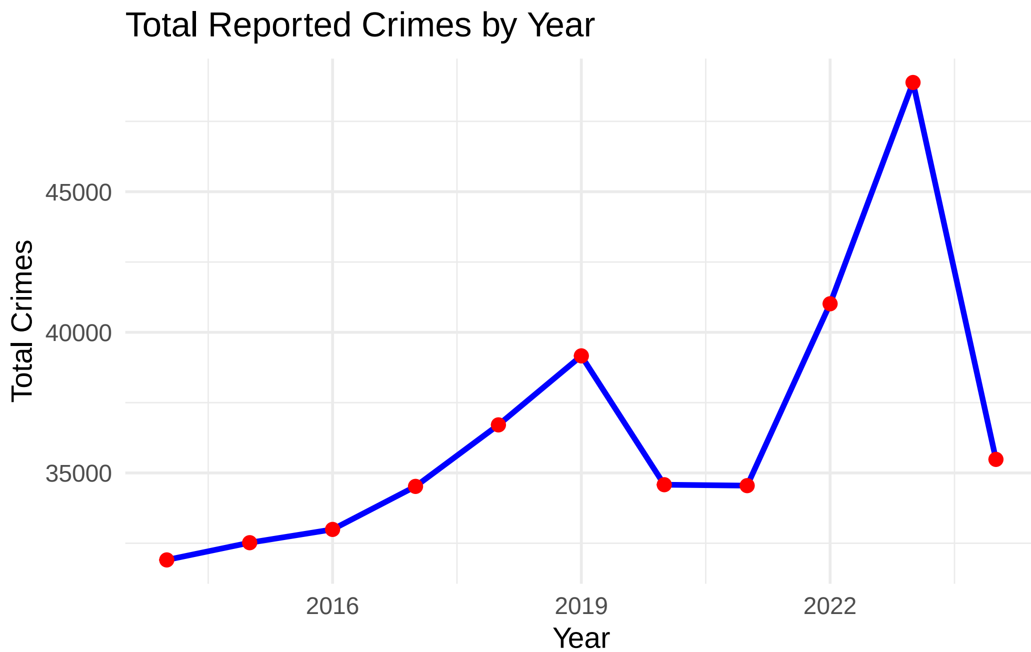

Figure 2: Time-Series Plot of Total Reported Crimes by Year in Toronto

Figure 3 illustrates the distribution of reported crimes in Toronto categorized by major crime types. The visualization highlights the relative prevalence of each crime type, providing insights into which offenses contribute most significantly to the overall crime burden in the city.

From the graph, it is evident that Assault represents the largest category of reported crimes, far surpassing other crime types. This indicates that violent interactions are a key area of concern for law enforcement and policymakers. Following Assault, Break and Enter and Auto Theft are the next most frequent crime categories, reflecting issues related to property security and vehicle safety. Robbery and Theft Over are less frequent but remain notable contributors to the overall crime landscape.

Figure 4 visualizes the spatial distribution of reported crimes across Toronto. Each point

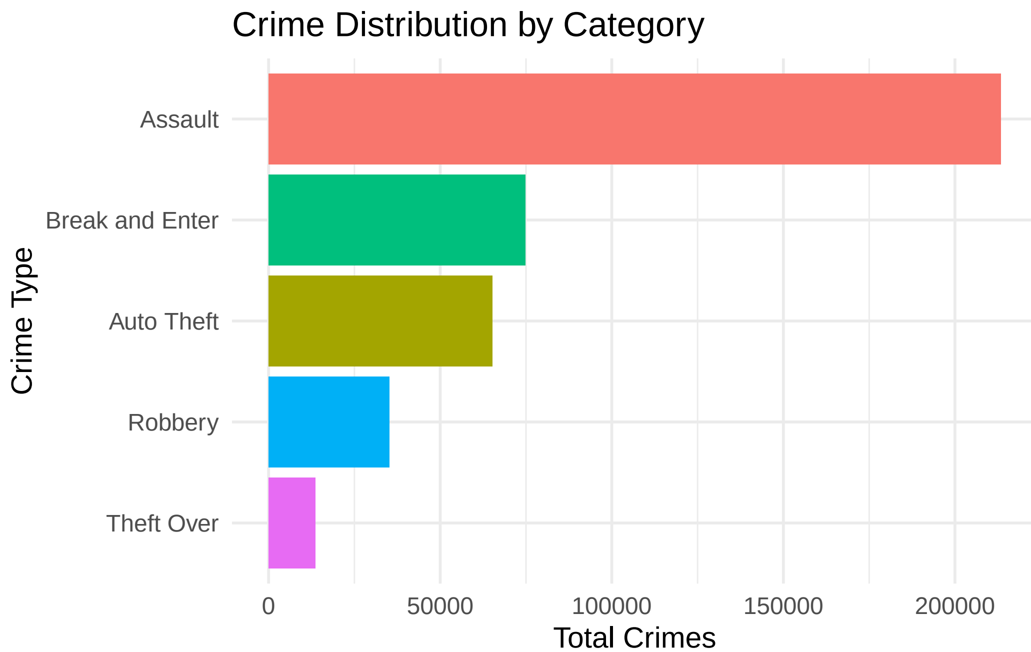

Figure 3: Bar Plot Showing the Distribution of Crimes by Category in Toronto

on the map represents a reported crime, plotted using geographic coordinates (latitude and longitude) over the city's boundaries. The visualization provides a comprehensive view of how crime incidents are geographically dispersed throughout Toronto.

From the map, it is evident that crimes are concentrated in specific urban areas, with a higher density in downtown Toronto compared to suburban and outlying neighborhoods. This spatial clustering indicates potential hotspots of criminal activity, often associated with higher population density, commercial zones, or socio-economic disparities. Conversely, areas with fewer points suggest lower reported crime rates, possibly reflecting suburban or less populated regions.

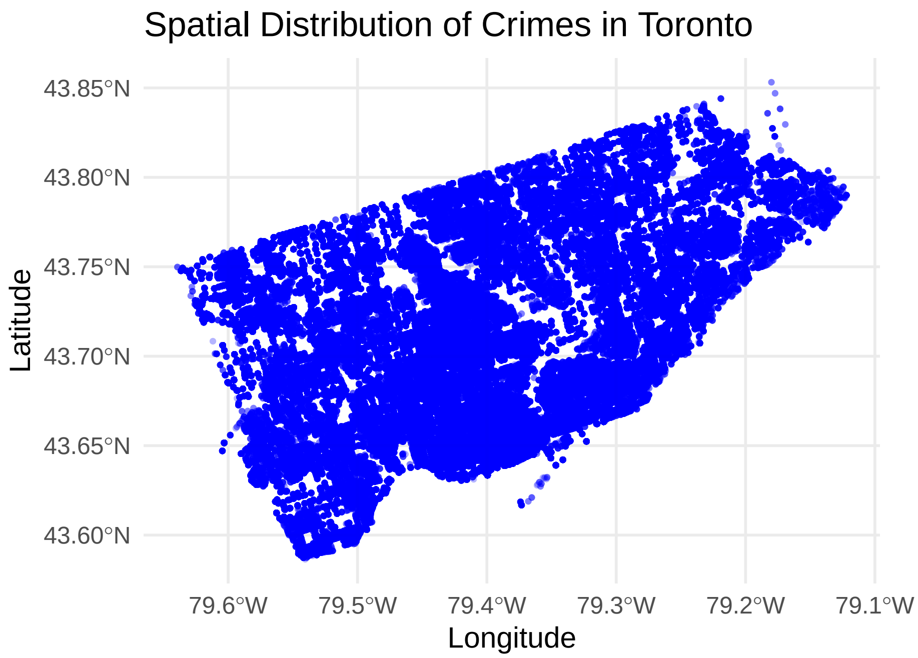

Figure 4: Spatial Distribution of Crimes in Toronto Visualized on a Geographic Map

3 Model

In our analysis, we utilized a Linear Regression model to examine the relationship between crime rates and a variety of factors, including temporal indicators (e.g., year, time of day), spatial variables (e.g., geographic coordinates, neighborhood), and crime-specific characteristics (e.g., crime type, division). Linear regression provides a straightforward and interpretable framework to quantify how these predictors influence crime rates, making it ideal for understanding patterns and informing policy decisions.

3.1 Model set-up

The model is formulated as follows:

3.2 Model Equation

The model is formulated as follows:

$$y_i = \beta_0 + \beta_1 \cdot \text{Year}_i + \beta_2 \cdot \text{Hour}_i + \beta_3 \cdot \text{Longitude}_i + \beta_4 \cdot \text{Latitude}_i + \sum_{j=1}^{k} \gamma_j \cdot \text{CrimeType}_{ij} + \sum_{l=1}^{m} \delta_l \cdot \text{Division}_{il} + \epsilon_i$$

Where: - y_i: The predicted number of crimes in observation (i). - β_0: Intercept term, representing the baseline level of crime when all predictors are at their reference levels. - β_1, β_2, β_3, β_4: Coefficients for temporal and spatial variables. - γ_j: Coefficients for categorical crime type variables. - δ_l: Coefficients for categorical police division variables. - ϵ_i: Residual errors, assumed to be normally distributed.

The target variable y_i in our model represents the observed crime count for a specific instance in the dataset. The predictors include:

Temporal Variables · Year (Year): Captures changes in crime rates over time, reflecting long-term trends. · Hour (Hour): Accounts for daily crime patterns, such as higher incidences during peak hours.

Spatial Variables · Longitude and Latitude (Longitude,Latitude): Represent geographic locations of crimes, enabling spatial analysis to identify high-crime areas. · Neighborhood (Division): Encoded as dummy variables, these capture the effects of police divisions or regions on crime rates.

Crime-Specific Variables .Crime Type (CrimeType): Also encoded as dummy variables, these distinguish betwe categories such as theft, assault, and robbery.

The linear regression model assumes that the crime count y_i is a linear combination of these predictors, with coefficients β quantifying the impact of each variable.

We selected linear regression for this analysis because of its interpretability and ability to provide clear, quantitative relationships between predictors and the target variable. This simplicity is particularly valuable for policymakers and urban planners, who can leverage these insights to target specific variables (e.g., certain neighborhoods or times of day) in crime prevention strategies. Linear regression is also computationally efficient, allowing us to process the large dataset (n=402326) with ease.

While alternative models, such as Random Forest or Bayesian Regression, could potentially capture more complex interactions or provide uncertainty quantification, linear regression remains a robust starting point for exploratory analysis. The assumption of linearity in relationships, though somewhat restrictive, is appropriate for identifying general trends in crime data.

3.3 Model justification

Regarding the relationship between crime rates and temporal factors, we anticipate that time-related variables such as the year and hour of the day significantly influence crime trends. The year ((Year_i)) captures long-term changes, reflecting broader societal and economic shifts. For example, as cities grow and technology advances, the frequency and type of crimes may change. Additionally, social policies and law enforcement strategies implemented in specific years could directly impact crime rates. The hour ((Hour_i)) of the day is equally critical, as certain hours, such as late evenings or early mornings, are associated with increased crime activity due to decreased surveillance and higher social activities like nightlife.

Spatial factors, such as geographic coordinates ((Longitude_i) and (Latitude*i)) and neighborhood divisions ((Division*{il})), play a vital role in understanding where crimes are concentrated. High-crime areas often have specific characteristics, such as socioeconomic challenges, poor infrastructure, or limited law enforcement presence. Geographic locations help identify patterns such as clusters of crimes near transit hubs, parks, or commercial areas. Neighborhood divisions encoded as categorical variables further highlight disparities across regions, allowing for a deeper understanding of how local characteristics, like economic conditions or population density, influence crime rates.

Crime-specific variables, such as the type of crime ((CrimeType_{ij})), are essential for distinguishing between different categories of offenses. For instance, thefts might show different spatial and temporal patterns compared to violent crimes like assaults. By including these categories as predictors, the model can tailor insights to specific crime types, enabling law enforcement to address each issue effectively.

4 Results

Our results are summarized in

4.1 Model Validation

For posterior predictive checks,

5 Discussion

5.1 Relationship between Time-based and Crime Indicators

The model highlights a clear relationship between time-based variables and crime rates. The year variable…….

5.2 Relationship between Geographical distribution and Crime Indicators

High-crime areas are often associated with specific characteristics, such as economic deprivation, limited access to education, or inadequate policing……

5.3 Crime-Specific Insights

as a predictor allows the model to uncover unique patterns associated with different categories of offenses.....

5.4 Government Policy Implications

The findings underscore the importance of data-driven policy-making in crime prevention.....

5.5 Limitations

5.5.1 Simplified Representation of Temporal Trends

5.5.2 Lack of Detailed Socioeconomic Context

5.5.3 Limitations of Linear Assumptions

5.6 Future Steps

Future research should focus on......

Appendix

A Data

A.1 Raw data

A.2 Data feature

A.3 Data Visualization

B Model details

B.1 Posterior predictive check

References

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.