When Risk Hits the Road: Does Danger Drive a Return?*

Motor Theft and Collisions in Toronto Are Driven by [Neighborhood Clustering] and [Environmental Conditions]

Yingke He

December 1, 2024

This study examines motor vehicle theft and traffic collisions in Toronto, focusing on spatial and temporal patterns, recovery outcomes, and contributing factors. [The analysis reveals that thefts are geographically clustered, with recovery rates differing by location type, while collisions are more frequent under poor visibility and adverse road conditions.] These findings underscore the need for targeted interventions, including enhanced surveillance in high-theft areas, infrastructure improvements for road safety, and optimized strategies for vehicle recovery. Such measures are critical for fostering safer and more secure urban mobility in Toronto.

Table of contents

*Code and data are available at: https://github.com/ohyykk/Toronto_Motor_Viehicle/tree/main.

1 Introduction

Decisions surrounding motorbike ownership and usage carry significant implications for personal safety and financial liability. Recent statistics highlight the increased risks associated with owning and riding motorbikes, including a heightened likelihood of theft and collisions compared to other vehicles (Yasmin and Eluru 2016). These risks are influenced by various factors such as geographic location, road conditions, time of day, and type of motorbike, making it essential to develop tools that effectively assess and mitigate these dangers.[can add one more sentence + citations]

This study introduces a composite risk score model to assess the risks associated with owning and riding a motorbike. Using data from Open Data Toronto, the model evaluates two critical events: motorbike theft and collisions. Logistic regression is employed to estimate the probabilities of these events, which are then combined into a single, interpretable composite risk score. This metric, incorporating factors such as neighborhood characteristics, road and lighting conditions, and time of day, is designed to guide motorbike users, insurers, and policymakers in risk assessment and decision-making, while informing strategies to mitigate these risks.

The primary estimand of the analysis is the composite risk score, derived from the individual probabilities of motorbike theft and collision. This score is calculated using predictor variables such as neighborhood characteristics, road and lighting conditions, time of day, and other contextual factors, which were selected for their documented relevance to motorbike-related risks.

This analysis confirms and extends three key findings: (1) material properties significantly influence motor efficiency, with high-conductivity materials yielding higher energy savings; (2) geometric optimization of motor components enhances torque generation and reduces energy losses; and (3) operational conditions, such as load and temperature, exhibit nonlinear effects, underscoring the need for adaptive motor control strategies. Furthermore, our projections suggest that integrating advanced materials and control systems could improve motor efficiency by up to 15% by 2050, aligning with global energy transition goals [To be updated..].

The structure of the paper is organized as follows: following Section 1, Section 2 outlines the data collection and preprocessing procedures, along with a detailed description of the outcome variable and the predictor variables used in the analysis. Section 3 introduces the logistic regression models applied to estimate the probabilities of theft and collision, as well as the method used to combine these probabilities into a composite risk score. Section 4 then presents the main findings, including insights into how different factors contribute to the risks of owning and riding a motorbike. Finally, Section 5 interprets the results, highlighting significant trends and implications for motorbike risk assessment, and concludes with a discussion on the limitations of the analysis and future research directions.

2 Data

We use the statistical programming language R (R Core Team 2023)…. Our data (**shelter?**)…. Following Alexander (2023), we consider…

[Libraries To be updated…]

Details about the data cleaning process and the criteria for variable selection are provided in Appendix A.

2.1 Source

This study utilized two datasets published by the Toronto Police Service. The first dataset focuses on motor vehicle collisions involving killed or seriously injured persons (KSI), while the second examines thefts from motor vehicles.

The Motor Vehicle Collisions dataset includes all reported incidents in which a person was either killed or seriously injured since 2006. It offers detailed information about each collision, such as the type of incident, the severity of injuries, and the location of the event, when available. Additionally, the dataset includes fields for both the old 140 and new 158 neighborhood structures in Toronto, allowing for flexible neighborhood-level analysis across different definitions.

The Theft from Motor Vehicle dataset contains all reported occurrences of thefts from vehicles, categorized by reported date. These offences are classified based on the value of the stolen items, distinguishing between theft under and theft over thresholds. Each occurrence number may include multiple rows, representing the various offences associated with a single event. The dataset excludes "unfounded" occurrences, adhering to Statistics Canada's definition that these events were determined not to have occurred or been attempted. Like the KSI dataset, this dataset includes fields for both the old and new neighborhood structures, enabling comprehensive geographic analyses of theft trends.

[Add variable types and trends]

2.2 Data Measurement and Limitations

The process of translating real-world phenomena into entries in the dataset involves several stages. When a traffic collision or theft occurs, it is reported to law enforcement through various channels, such as emergency calls, online submissions, or in-person reports. Police officers or administrative personnel document the event details, including date, location, type of incident, and additional attributes such as severity or value of stolen items. These records are then digitized and aggregated into structured datasets, with fields organized to support analysis and reporting. However, during this process, certain changes or context-specific information may be lost, and the data ultimately reflects a structured summary of the events rather than their full complexity.

The datasets from Open Data Toronto did not specify the exact methods used for data collection, which may introduce some uncertainty regarding the consistency and reliability of the recorded events. Additionally, for privacy reasons, the locations of crime occurrences have been deliberately offset to the nearest road intersection node. This may result in discrepancies when analyzing counts by division or neighborhood, as the reported locations may not reflect the exact sites of the occurrences.

Some coordinate information in the datasets appears as "0, 0," indicating that the specific location was either not validated or could not be geocoded. In such cases, a general division or neighborhood association may still be provided, but for invalid or external locations, the designation "NSA" ("Not Specified Area") is used. Furthermore, the Toronto Police Service does not guarantee the accuracy, completeness, or timeliness of the data, which may lead to potential misinterpretations or incomplete analyses.

Additional details about the dataset are available in the datasheet, accessible through the repository linked to this paper.

2.3 Outcome Variables

The outcome variable of this analysis is the Risk Index, a composite metric that integrates the probabilities of two underlying events: theft and collision.

2.4 Predictor Variables

The predictor variables in this study—**Neighborhood ID**, **Premises Type**, **Report Hour**, **Location Type**, and **Report Day**—each play a critical role in understanding patterns in motor vehicle thefts. These variables collectively provide a structured framework for analyzing temporal, spatial, and contextual factors influencing theft occurrences, offering insights into how and where such incidents are most likely to happen.

The temporal variables, **Report Hour** and **Report Day**, capture the timing of reported incidents, enabling the identification of peak periods and weekly patterns in theft occurrences. **Neighborhood ID** provides a spatial dimension, facilitating the analysis of geographic hotspots and regional variations in theft rates. Contextual variables like **Premises Type** and **Location Type** offer insights into the environments where thefts are most likely to occur, helping to understand the situational factors that may contribute to these incidents. Together, these variables enable a comprehensive analysis of motor vehicle theft trends and inform strategies to mitigate risks.

For motor vehicle collisions, predictor variables like **Neighborhood ID**, **Road Surface Condition**, **Lighting Condition**, and **Traffic Control** provide critical insights into environmental and infrastructural factors affecting collision risks. These variables collectively allow the model to assess how location-specific, weather-related, and regulatory conditions contribute to the likelihood and severity of collisions, offering a structured view of collision dynamics.

2.4.1 Collision Probability

draw the distribution of each predictor variable #### Neighborhood ID #### Premises Type #### Report Hour #### Location Type #### Report Day

2.4.2 Theft Probability

2.4.2.1 Neighborhood ID

2.4.2.2 Road Surface Condition

2.4.2.3 Lighting Condition

2.4.2.4 Traffic Control

3 Model

The main idea of this composite risk score model is to calculate an index of risk for owning and riding a motorbike. The modeling strategy has two primary objectives. The first objective is to estimate the likelihood of theft (P(Theft)) and collision (P(Collision)) under diverse environmental and situational conditions using logistic regression models. The second objective is to combine these probabilities into a unified Risk Index, providing actionable insights into the risk for owning and riding a motorbike

The models were run in R (R Core Team 2023) using the `stats` package to implement logistic regression models predicting theft and collision probabilities. The theft model incorporates predictor variables such as neighborhood ID, premises type, report hour, location type, and report day, while the collision model includes neighborhood ID, road surface condition, lighting condition, and traffic control.

Both models estimate the log-odds of their respective probabilities using the predictors, enabling reliable predictions under varying environmental and situational conditions. Model diagnostics and background details, including variable descriptions and performance metrics, are included in Appendix B.

3.1 Model Set-up

The logistic regression model for theft probability takes the form of the following equation:

$$\log\left(\frac{P(\text{Theft})}{1-P(\text{Theft})}\right) = \beta_0 + \beta_1 \cdot \text{HOOD}_{158} + \beta_2 \cdot \text{Premises Type} + \beta_3 \cdot \text{Report Hour} \\ + \beta_4 \cdot \text{Location Type} + \beta_5 \cdot \text{Report Day} + \epsilon$$

This model utilizes the following predictor variables:

- **Neighborhood ID** (`hood_158`): Unique identifier for the neighborhood.
- **Premises Type** (`premises_type`): Type of premises where the theft occurred like parking lot, garage
- **Report Hour** (`REPORT_HOUR`): Hour of the day the theft was reported.
- **Location Type** (`location`):Location of the theft like indoors and outdoors.
- **Report Day** (`REPORT_DOW`): Day of the Week Offence was Reported.

The logistic regression model for collision probability takes the form of the following equation:

$$\log\left(\frac{P(\text{Collision})}{1-P(\text{Collision})}\right) = \beta_0 + \beta_1 \cdot \text{HOOD}_{158} + \beta_2 \cdot \text{Road Surface Condition} + \beta_3 \cdot \text{Light} \\ + \beta_4 \cdot \text{Traffic Control} + \epsilon$$

This model utilizes the following predictor variables:

- **Neighborhood ID** (`hood_158`): Unique identifier for the neighborhood.
- **Road Surface Condition** (`road_conditions`): Road surface condition like wet and dry.
- **Lighting Condition** (`lighting_conditions`): Lighting conditions at the time of the collision like dark or clear.
- **Traffic Control** (`traffic_control`):Type of traffic control present like stop signs or traffic signals After fitting both models, combine the probabilities into a weighted risk index:

$$\text{Risk Index} = w_1 \cdot P(\text{Theft}) + w_2 \cdot P(\text{Collision})$$

where weights $w_1$and w_2 are adjust based on the relative importance of theft and collision risks.

3.2 Model justification [edit]

The analysis adopts a hybrid approach that combines sub-index calculations for theft risk with logistic regression for collision probability. This design ensures that the model reflects the specific data characteristics and practical considerations when assessing motorbike-related risks. Logistic regression is not utilized for theft risk due to the absence of negative (non-theft) cases in the dataset. Instead, theft risk is represented through sub-index calculations for three critical factors: hour of the day, day of the week, and premises type. Each factor contributes equally to the total theft component. The Hour of the Day Index captures temporal variations in theft risk across 24 hours, while the Day of the Week Index accounts for weekly patterns of theft. The Premises Type Index reflects variations in risk based on location type, such as houses, outdoor spaces, or commercial premises. These sub-indexes are normalized so their contributions to the theft component are proportional and balanced. This approach ensures an accurate representation of theft risk patterns while providing actionable insights into temporal and spatial risk factors. Linearity is assessed using smoothed scatter plots of the predicted logit values:

$$\text{logit} = \log\left(\frac{P}{1-P}\right)$$

Logistic regression is utilized for collision risk due to its effectiveness in estimating probabilities for binary outcomes and its statistical rigor in modeling relationships between predictors and outcomes. By modeling the log-odds of collision probability as a linear combination of neighborhood-specific characteristics, road surface conditions, lighting conditions, and traffic control measures, the method provides clear insights into the factors contributing to collision risks. Logistic regression's capacity to estimate odds ratios enables intuitive interpretation of each predictor's impact on collision likelihood. Its ability to handle both categorical and continuous predictors, include interaction terms, and manage multicollinearity further enhances its robustness in capturing complex relationships.

The final Risk Index integrates the theft and collision components using a weighted formula. The collision probability component is weighted at 0.7, reflecting its higher immediate impact on safety, while the theft component is weighted at 0.3. This weighting scheme prioritizes collision risk while ensuring that theft risk is not overlooked. By combining these components, the Risk Index provides a unified measure of motorbike-related risks, enabling stakeholders to assess and compare safety conditions across different contexts.

This modeling approach is justified by its statistical efficiency and interpretability. The subindex method for theft risk is tailored to the dataset's characteristics, avoiding assumptions about unobserved cases, while ensuring proportional representation of key risk factors. Logistic regression's flexibility and reliability in predicting binary outcomes make it an ideal choice for collision risk modeling. Together, these components enable the model to deliver statistically robust and actionable risk estimates, providing a comprehensive framework for evaluating motorbike ownership and usage risks. By combining statistical rigor with practical insights, this approach addresses both immediate safety concerns and long-term theft risks.

3.3 Model Assumtions and Validations

To validate the use of logistic regression models, the four main assumptions: binary nature of the outcome, linearity of the logit, indipendence of observations, and lack of multicollinearity will be assessed (Kononen, Flannagan, and Wang 2011). To validate the use of a composite Risk Index, the correlation of the index with observed outcomes will be assess using a Pearson correlation test.

3.3.1 Logistic Regression Models

3.3.1.1 Binary Nature of the Outcome

A fundamental assumption of logistic regression is that the response variable is binary or dichotomous, meaning it can take on only two possible outcomes (Nick and Campbell 2007). This assumption is satisfied in both models. In the theft model, the response variable indicates whether a theft occurred, with a value of 1 representing the occurrence of theft and 0 representing its absence. Similarly, in the collision model, the response variable denotes whether a collision occurred, where 1 represents the occurrence of a collision and 0 indicates no collision.

The logistic regression model does not directly predict 1 or 0. Instead, it estimates the probability that the response variable equals 1, given the predictor variables. For instance, the theft model outputs the probability of theft occurring (P(Theft)), and the collision model outputs the probability of a collision occurring (P(Collision)). These probabilities range between 0 and 1, allowing for detailed predictions of the likelihood of the respective events.

3.3.1.2 Linear Relationship in the Log-Odds

[Identifying Continuous Variables in Your Models: For LR Model 1 (Theft Model): Continuous Variable: REPORT_HOUR (Hour of the day the theft was reported For LR Model 2 (Collision Model): No Continuous Variables]

[Conclusion: Total Number of Plots: 1 Draw a binning plot for REPORT_HOUR in LR Model 1 to assess the linear relationship]

3.3.1.3 Absence of Multicollinearity

[The easiest way to check for multicollinearity in your logistic regression models is to calculate the Variance Inflation Factor (VIF) for your predictors. This provides a straightforward and interpretable measure of multicollinearity.]

3.3.2 Composite Risk Index

3.3.2.1 Pearson Correlation Test

4 Results

Our results are summarized in **?@tbl-modelresults**.

4.1 First Result Point

4.2 Second Result Point

4.3 Third Result Point

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows...

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows... This suggests...

?@fig-stanareyouokay-2 is a Rhat plot. It shows... This suggests...

References

Alexander, Rohan. 2023. *Telling Stories with Data.* Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Kononen, Douglas W., Carol AC Flannagan, and Stewart C. Wang. 2011. "Identification and Validation of a Logistic Regression Model for Predicting Serious Injuries Associated with Motor Vehicle Crashes." *Accident Analysis & Prevention* 43 (1): 112–22.

Nick, Todd G., and Kathleen M. Campbell. 2007. "Logistic Regression." In *Topics in Biostatistics*, 273–301. Springer.

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

Yasmin, Shamsunnahar, and Naveen Eluru. 2016. "Latent Segmentation Based Count Models: Analysis of Bicycle Safety in Montreal and Toronto." *Accident Analysis & Prevention* 95: 157–71. https://doi.org/10.1016/j.aap.2016.06.015.