

SOCC70: Models of the Social World

How Close is Too Close? Exploring the Relationship Between Bike Theft and Police Facility

Distance in Toronto

by

Zehui Yin

Supervised by

Prof. Ethan Fosse

Department of Sociology

University of Toronto Scarborough

Abstract

Bike thefts pose a significant deterrent to urban cycling. This research delves into the influence of police presence on bike thefts and the geographical scope of its deterrent effect. Leveraging bike theft records from 2014 to 2023 in Toronto, alongside the most recent Canadian census data, I employ two-way fixed-effect Zero-inflated Negative Binomial Regression models to scrutinize the impact of network walking distance to the nearest police facility on the number of bike thefts within each census tract. To ensure robustness, I present a variety of model specifications, including OLS multiple regressions, Poisson regressions, and standard Negative Binomial Regression. The findings indicate that a police facility within a 500-metre radius notably curtails bike thefts, with statistical significance at the 0.05 level. Yet, this deterrent effect wanes with increasing distance and becomes inconsequential beyond 500 metres. Furthermore, at distances exceeding 2 kilometres, the presence of police facilities does not influence the number of bike thefts. The study infers that police facilities exert a localized effect on bike theft prevention in urban areas. Consequently, merely increasing the number of police facilities city-wide is not an efficacious strategy for reducing bike theft. Policymakers are thus encouraged to consider alternative methods or strategies for enhancing crime deterrence.

1 Introduction

Cycling is a key component of sustainable urban development, as it reduces environmental impacts and provides alternative mobility options that can decrease automobile dependence in cities. Previous studies have demonstrated that cycling has various health, economic and social benefits (Blondiau et al., 2016; Oja et al., 2011). Toronto promotes physical activity through the ActiveTO program that expands the cycling network (City of Toronto, n.d.). However, a higher rate of cycling also entails a higher risk of bike theft, which discourages people from using this sustainable mode

of transportation (Van Liero et al., 2015). In this project, I will explore the relationship between the occurrence of bike thefts and the proximity of the nearest police facility, using Toronto bike theft data from 2014 to 2023. My research will be guided by the following 2 research questions:

RQ1. What causal influence does the presence of police facilities have on the frequency of bike thefts in Toronto?

RQ2. Upon observing a deterrent effect, to what extent does the proximity of police facilities causally mitigate bike theft incidents in Toronto?

2 Data and Methods

2.1 Data Cleaning and Creating Dataset for Regression Analysis

I created a panel dataset for the regression analysis from 5 data sources, as shown in Table 1. The study area is the City of Toronto. Figure 1 below displays the spatial distribution of the police facilities, bike theft incidents, and census tracts in Toronto. A large number of bike thefts are concentrated in and around the downtown core, whereas police stations are distributed more uniformly throughout the city. I filtered out the bike theft records that lacked geographical location or occurred outside of Toronto. I also excluded those that happened before 2014, as the data quality improved after 2014 and few incidents were recorded before then.

Table 1: Data Sources

| Name | Description | Format | Source |
|---------------------------|--|---------------------|--|
| Bicycle Thefts | All bicycle theft occurrences reported to the Toronto Police Service, including those where the location has not been able to be verified. | GeoPackage | Toronto Open Data |
| Police Boundaries | Geographical areas in the city are associated with each Toronto Police Division. | GeoPackage | Toronto Open Data |
| Police Facility Locations | A geographical point shape file that depicts the police facility locations. | GeoPackage | Toronto Open Data |
| 2021 Census Data | 2021 Canadian Census Data at Census Tract (CT) Level. | <i>sf</i> dataframe | Statistics Canada through <i>cancensus</i> package (von Bergmann et al., 2021) |
| Road Network | OpenStreetMap road network data for Toronto. | PBF | OpenStreetMap through Geofabrik |

I aggregated the bicycle theft data to the 2021 census data by the census tract of each theft event fell within. Then I assigned the police boundaries to each census tract by the largest area of intersection. I calculated the network distance in metres from each census tract centroid to the closest police facility location using the *r5r* package (Pereira et al., 2021) in *R* and the OpenStreetMap road network. For the Toronto Islands census tract, which has no road network connected to any police station, I used the Euclidean distance from its centroid to the closest police station. Finally, I transformed the sociodemographic variables from the census data into percentages. The final panel dataset has $585 \times 12 \times 7 = 49,140$ observations and 11 variables.



Figure 1: Study Area Map (Left) and Conceptual Directed Acyclic Graph (Right)

2.2 Modelling Methodology

The unit of analysis for this project is a census tract, month, and day of the week combination. I used two-way fixed-effect Negative Binomial Regression with log-link to model the count of bike thefts in Toronto. I included two fixed effect variables: month, $Month_i$, and day of the week, $DayOfWeek_t$. I used maximum likelihood with package *fixest* (Bergé, 2018) in *R* to estimate the model. Two model specifications are used in this project.

$$\log[E(BikeTheft_{it})] = Month_i + DayOfWeek_t + \beta_1 \log(Distance_{it}) + \beta_2 \log(Distance_{it})^2 + \mathbf{X}_{it}\boldsymbol{\beta} \quad (1)$$

$$\begin{aligned}
\log[E(\text{BikeTheft}_{it})] = & \text{Month}_i + \text{DayOfWeek}_t + \beta_1 \text{Within500To1000Metres}_{it} + \\
& \beta_2 \text{Within1000To2000Metres}_{it} + \beta_3 \text{Above2000Metres}_{it} + \\
& \beta_4 \text{Within500To1000Metres}_{it} \times \text{Distance}_{it} + \\
& \beta_5 \text{Within1000To2000Metres}_{it} \times \text{Distance}_{it} + \\
& \beta_6 \text{Above2000Metres}_{it} \times \text{Distance}_{it} + \beta_7 \text{Distance}_{it} + \mathbf{X}_{it}\boldsymbol{\beta}
\end{aligned} \tag{2}$$

Where

BikeTheft_{it} = the number of bike thefts in month i and day of week t at a specific census tract.

Distance_{it} = the network distances in metres of the census tract to the nearest police facility.

\mathbf{X}_{it} = a set of control variables that vary by census tract.

$\text{Within500To1000Metres}_{it}$, $\text{Within1000To2000Metres}_{it}$, and $\text{Above2000Metres}_{it}$ = a set of dummy variables indicating whether Distance_{it} belongs to some specific ranges, using Distance_{it} below 500 metres as the reference category.

The treatment variable is the network walking distance to the closest police facility, Distance_{it} .

The coefficients of interest are those β s associated with Distance_{it} and interaction terms of Distance_{it} . The two specifications both capture the nonlinear causal effect of the treatment on the outcome, one by using a squared term and the other by using interaction terms with dummy variables. I exclude the census tract fixed effect because it would cause perfect collinearity with the distance to the nearest police facility. However, I include a set of dummy variables for the police division in the control variables to account for unobserved heterogeneities among different police divisions. The two fixed effects control for time shocks that may confound the treatment and the outcome variable. As a robustness check, I also estimate these model specifications using OLS multiple linear regression and Poisson regression. Figure 1 above shows the hypothesized conceptual directed acyclic graph of the treatment on the outcome. If the graph holds, both model specifications can identify the casual total effect. However, the strong assumption is that no unobserved spatial dependent variables directly affect bike thefts.

To address the potential zero inflation in the response variable of bike theft, where approximately 72% of the observations in the panel dataset report zero incidents, I further estimated a zero-inflated negative binomial regression using maximum likelihood with the *pscl* package (Zeileis et al., 2008). The mathematical formulation of the model is presented as Equation 3 below (NCSS Statistical Software, n.d.).

$$Pr(BikeTheft_i = j) = \begin{cases} \pi_i + (1 - \pi_i)g(BikeTheft_i = 0) & \text{if } j = 0 \\ (1 - \pi_i)g(BikeTheft_i) & \text{if } j > 0 \end{cases} \quad (3)$$

Where

$$\text{logit}(\pi_i) = \widetilde{\mathbf{X}}_i \boldsymbol{\beta}$$

$\widetilde{\mathbf{X}}_i$ = a set of control variables that vary by census tract excluding the police division dummy.

$g(BikeTheft_i)$ is the negative binomial distribution.

$$\begin{aligned} & \log[E(g(BikeTheft_{it}))] \\ &= Month_i + DayOfWeek_t + \sum \gamma_k PoliceDivision_{kit} \\ &+ \beta_1 Within500To1000Metres_{it} + \beta_2 Within1000To2000Metres_{it} \\ &+ \beta_3 Above2000Metres_{it} + \beta_4 Within500To1000Metres_{it} \times Distance_{it} \\ &+ \beta_5 Within1000To2000Metres_{it} \times Distance_{it} \\ &+ \beta_6 Above2000Metres_{it} \times Distance_{it} + \beta_7 Distance_{it} \end{aligned}$$

Given that the treatment variable, $Distance_{it}$, remains constant over time and varies only across units or census tracts, it exhibits limited exogenous variation. To mitigate multicollinearity and enhance the precision of the estimates, independent variables are included exclusively in either the zero-inflated or the negative binomial portion of the model, but not both. After experimenting with

various combinations of independent variables in each portion, the final model specification shown in Equation 3 was chosen for its sensible estimates and superior model fit.

3 Results

3.1 Descriptive Statistics

Table 2: Descriptive Statistics of the Panel Dataset Generated with summarytools Package (Comtois, 2022)










| Name | Type | Statistics | Graph | Missing |
|--|-------------|------------------------------|---|------------|
| Number of Bike Thefts | Count | Mean (sd): 0.70 (1.84) |  | 0 (0.0%) |
| Network Distance in Metres to the Nearest Police Facility | Continuous | Mean (sd): 2881.26 (1567.83) |  | 0 (0.0%) |
| Police Division | Categorical | 16 Unique Levels |  | 0 (0.0%) |
| Percentage of Female Residents | Continuous | Mean (sd): 51.66 (2.46) |  | 0 (0.0%) |
| Percentage of White Residents (not Visible Minority) | Continuous | Mean (sd): 46.33 (24.02) |  | 168 (0.3%) |
| Percentage of Residents with Household Income below \$40,000 | Continuous | Mean (sd): 20.02 (8.33) |  | 168 (0.3%) |
| Percentage of Residents with Age below 19 | Continuous | Mean (sd): 18.70 (5.49) |  | 0 (0.0%) |
| Percentage of Residents aged between 20 to 24 | Continuous | Mean (sd): 6.46 (2.38) |  | 0 (0.0%) |
| Population Density per Square Kilometre | Continuous | Mean (sd): 8446.46 (9184.19) |  | 0 (0.0%) |

Table 2 shows the descriptive statistics for the independent and dependent variables used in this project, except for the two fixed effect variables, month and day of the week. Both variables percentages of female and low-income people have about 0.3% missing values, as data are suppressed for areas below certain population thresholds set by Statistics Canada to protect individual responses' confidentiality. This is more likely when the census question is in the long-form questionnaire and the census geography is small. The response variable number of bike thefts is a count variable, as it is positive and many observations have a value of 0. Variable population density is right-skewed and thus is log-transformed in the regression analysis. The mean network

distance of a census tract in Toronto to the nearest police facility is about 2,900 metres with a right-skewed distribution, which means that most census tracts have a police facility within 3 kilometres.

2.2 Regression Results

Table 3 displays the estimation results. I estimated two types of specifications for distance: one using squared terms (Equation 1) and the other using dummy variables and interactions (Equation 2), using both Poisson and Negative Binomial regression with and without control variables and fixed effects. The zero-inflated Negative Binomial regression results, which utilize a modified version of dummy variables and interactions as depicted in Equation 3, are shown in the last column. Equation 2 specification has a lower BIC value than Equation 1 specification for the same type of model. Also, for the same specification in Negative Binomial models, adding control variables and fixed effects lowers the BIC value considerably. Moreover, based on simulation-based tests for over/underdispersion using the *R* package *DHARMA* (Hartig, 2022), the p-value is almost 0, indicating an overdispersion problem for the Poisson regressions. Therefore, Negative Binomial regression using dummy variables with interactions, control variables and fixed effects (Equation 2), shown in the second last column, is the best model among the first six models.

Given the potential issue of zero inflation, the zero-inflated negative binomial model may provide a better fit than the standard negative binomial model, leading to a focus on the interpretations derived from the two final columns. In both models, the dummy variables representing distance ranges of 500 to 1000 metres and 1000 to 2000 metres are positive and statistically significant at the 0.05 level or lower. With the reference category being within 500 metres, the findings suggest that census tracts located 500 to 2000 metres from a police facility experience a higher incidence of bike thefts. However, the coefficients for distance and its interaction terms are generally small and lack statistical significance. The sole exception is the interaction term for the 500 to 1000

metres category and distance, which is significantly negative at the 0.001 level, suggesting a downward trend in the rate of bike thefts within this specific distance interval.

Table 3: Regression Results

| Variable | Model | | | | | | |
|--|--------------------------------------|--------------------------------|--------------------------------|------------------------|--------------------------------|--------------------------------|--|
| | Poisson ^c | Negative Binomial ^d | Negative Binomial ^d | Poisson ^c | Negative Binomial ^d | Negative Binomial ^d | Zero-inflated Negative Binomial ^e |
| Constant | -2.2548**** (0.4738) ^b | -6.1914**** (0.9822) | | 0.3461* (0.1671) | 2.1498*** (0.2647) | | -0.1668 (0.1630) |
| <i>Distance</i> | | | | 0.0006** (0.0002) | -0.0038*** (0.0007) | 0.0006 (0.0004) | 0.0002 (0.0004) |
| <i>log(Distance)</i> | 1.1072*** (0.1267) | 2.8217*** (0.2651) | 0.6479*** (0.1438) | | | | |
| <i>log(Distance)</i> ² | -0.0881*** (0.0089) | -0.2693*** (0.0178) | -0.0562*** (0.0086) | | | | |
| <i>Within500To1000Metres</i> | | | | 1.5926*** (0.1014) | -0.3121 (0.3013) | 1.1804*** (0.1711) | 1.0236*** (0.1872) |
| <i>Within1000To2000Metres</i> | | | | 0.9730*** (0.0963) | -0.0690 (0.2804) | 0.4657** (0.1577) | 0.4676* (0.1957) |
| <i>Above2000Metres</i> | | | | 0.3192*** (0.0896) | -1.8086*** (0.2682) | 0.1708 (0.1501) | -0.0168 (0.1585) |
| <i>Within500To1000</i> interacts with <i>Distance</i> | | | | -0.0022*** (0.0002) | 0.0022*** (0.0007) | -0.0018*** (0.0004) | -0.0014*** (0.0004) |
| <i>Within1000To2000</i> interacts with <i>Distance</i> | | | | -0.0010*** (0.0002) | 0.0024*** (0.0007) | -0.0007* (0.0003) | -0.0005 (0.0004) |
| <i>Above2000Metres</i> interacts with <i>Distance</i> | | | | -0.0007** (0.0002) | 0.0034*** (0.0007) | -0.0006 (0.0004) | -0.0003 (0.0004) |
| Control Variables ^f | ✓ | | ✓ | ✓ | | ✓ | ✓ |
| Month Fixed Effect | ✓ | | ✓ | ✓ | | ✓ | ✓ |
| Day of Week Fixed Effect | ✓ | | ✓ | ✓ | | ✓ | ✓ |
| Sample Size | 48,972 | 49,140 | 48,972 | 48,972 | 49,140 | 48,972 | 48,972 |
| BIC | 92,329 | 97,957 | 84,632 | 91,842 | 97,822 | 84,612 | 84,880 |
| Over-dispersion Parameter | | 0.3581 | 1.1802 | | 0.3639 | 1.2036 | 1.7529 |

Notes:

a. Significance codes: 0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1.

b. Standard errors are in brackets.

c. The models are estimated with iteratively reweighted least squares and a dummy variable approach for fixed effects using the *stats* package (R Core Team, 2013). The coefficient standard errors for these models are not clustered.

d. The models are estimated with maximum likelihood using the *fixest* package (Bergé, 2018), which applies a demeaning procedure and, thus, does not estimate an intercept. The coefficient standard errors are clustered by the two fixed effects when they are in the model.

e. The model is estimated with maximum likelihood and a dummy variable approach for fixed effects using the *pscl* package (Zeileis et al., 2008). The coefficient standard errors for this model are not clustered.

f. The control variables are the percentage of female, white, low-income (before tax annual household income < 40,000), young (age < 19), and young-adult (age between 20–24) residents, log of population density, and police division dummies.

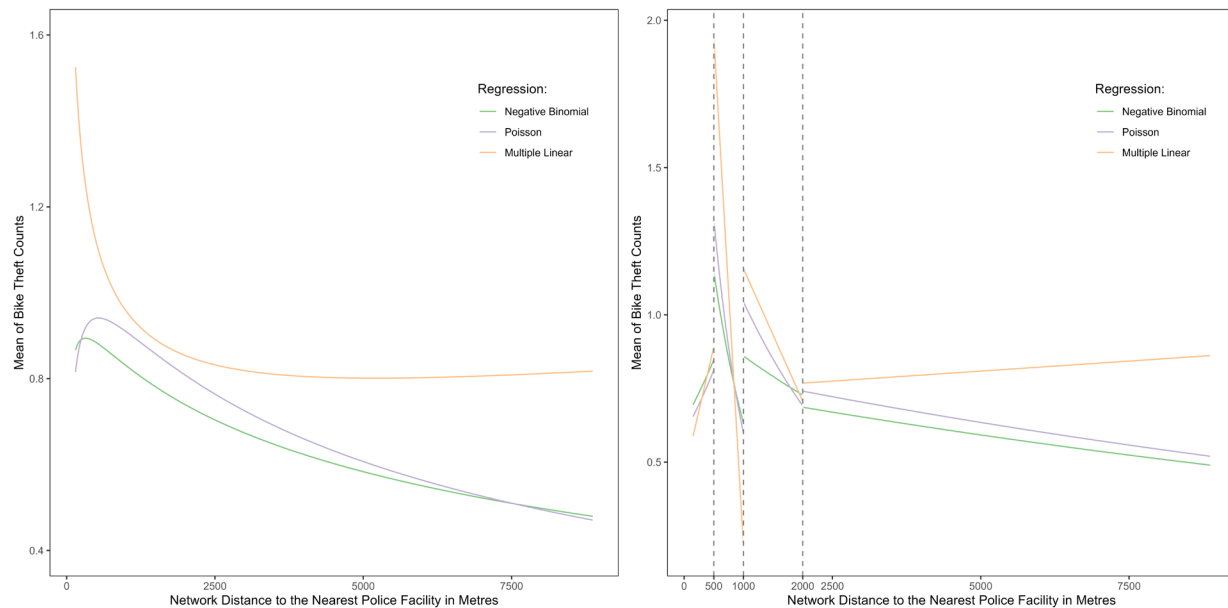


Figure 2: Partial Effect Plots for Squared Term Specification (Left) and Dummy Variable Specification (Right)

Figure 2 above shows the partial effects of network walking distance on the mean number of bike thefts in the census tract for both model specifications outlined in Equations 1 and 2 with control variables and fixed effects estimated with OLS multiple linear regression, Poisson and Negative Binomial regressions. All other independent variables are maintained at their typical values—means for continuous variables and modes for categorical or dummy variables. Notably, the green curve in the right panel corresponds to the most preferred not zero-inflated model in the second last column of Table 3. All the models have generally similar results, except the OLS multiple regression produced unreliable prediction in the left panel. The bike thefts are low within 500 metres, high between 500 to 2000 metres, and flat beyond 2000 metres, suggesting that police station presence beyond 2 kilometres has no effect on bike thefts.

Figure 3 displays the most preferred not zero-inflated model's partial effect plot with a 95% confidence interval and its deviance residual plot. Based on the residual plot, the prediction is

fairly good for response variables around the sample mean. There is some evidence of zero inflation (Zuur et al., 2009), as the number of zeros in the response variable is too large for the negative binomial distribution. The model overall tends to overestimate the response and the residuals have noticeable patterns at the two tails. Consequently, the zero-inflated negative binomial model emerges as the most suitable among all models discussed in this paper.

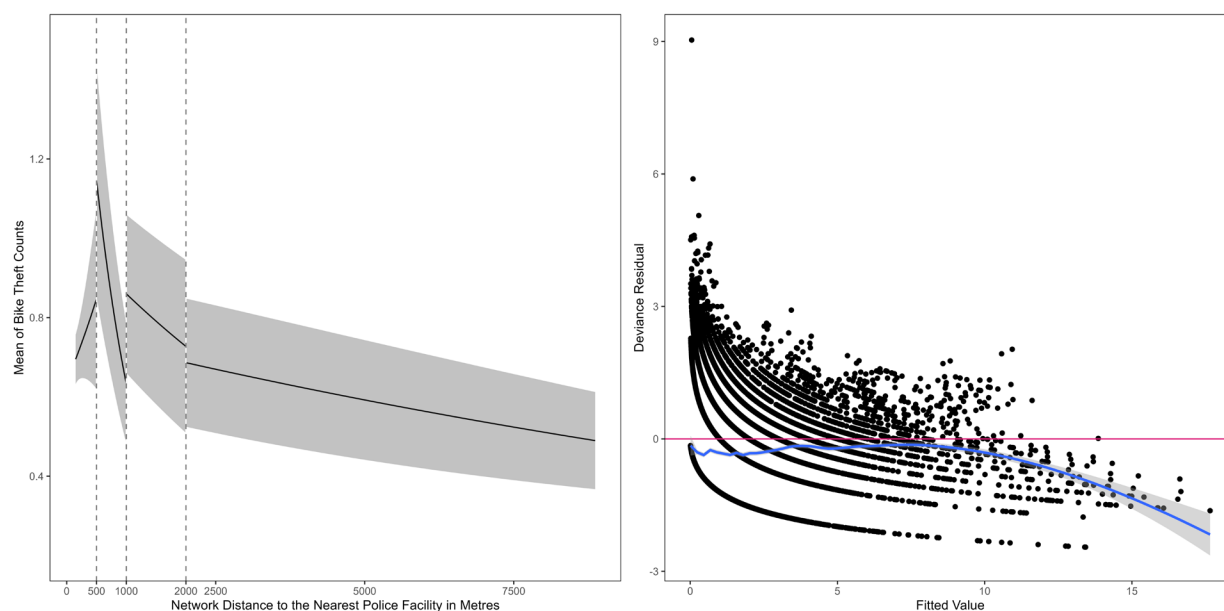


Figure 3: Partial Effect Plot with Confidence Interval (Left) and Residual Plot (Right) of Negative Binomial Regression with Dummy Variable Specification

Figure 4 illustrates the partial effect plot of the preferred model. The transition from the non-zero-inflated to the zero-inflated model shows minimal change in the partial effect plot, which aligns with expectations given the closely aligned estimated coefficients in the negative binomial component of both models. Notably, the predicted mean bike theft counts exhibit a positive slope only within the 0 to 500-metre range, with noticeable jumps in predicted bike thefts as the distance surpasses the two thresholds of 500 and 1000 metres. This pattern underscores the effectiveness of police deterrence primarily within the immediate vicinity.

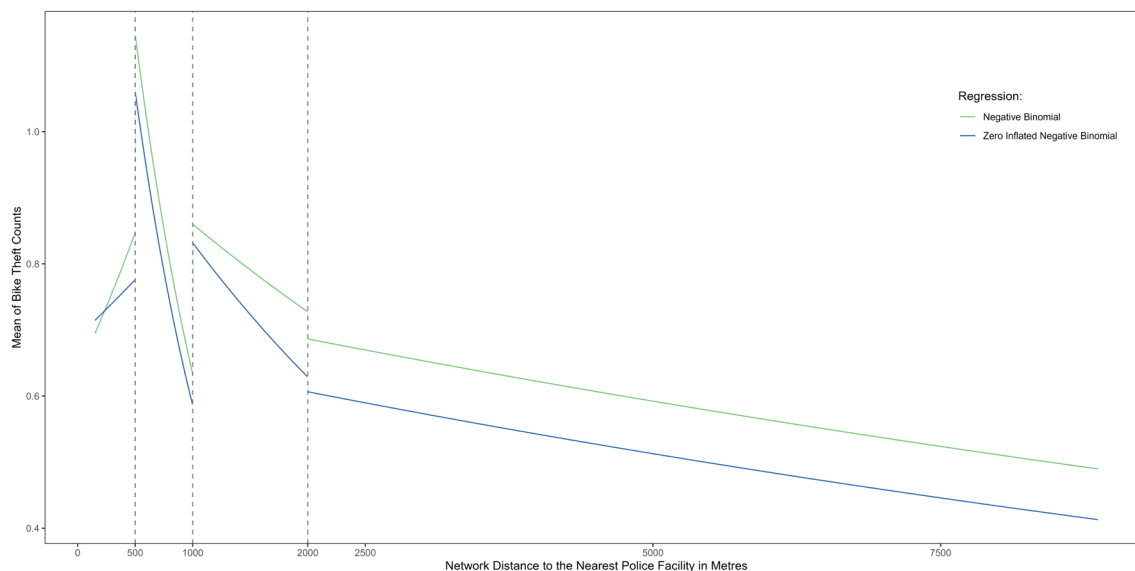


Figure 4: Partial Effect Plot for Dummy Variable Specification for the Zero-inflated Negative Binomial Regression

4 Conclusion

This project demonstrates that police facilities within 500 metres of network walking distance from the census tract centroid lower bike thefts in that census tract significantly, while police facilities beyond 2 kilometres have no impact on bike thefts. These results are reasonable in a complex urban environment like Toronto, where cycling is quicker than driving in short distances and 500 metres of network walking distance takes more than 5 minutes on foot. Beyond this distance, police officers would find it hard to pursue offenders on foot, particularly when offenders are likely cycling on the stolen bike. The main limitation of the analysis is the strong assumption that unobserved spatial-dependent factors such as housing density and traffic volume do not influence bike theft directly. Should this assumption be invalidated, these confounding variables would bias the coefficient estimates, thereby undermining the purported causal relationship between police presence and crime deterrence. Additionally, the study faces a data limitation due to the unavailability of historical data on police station locations, preventing the models from leveraging temporal variations in distance.

References

- Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. *Journal of statistical software*, 27(8), 1-25. <https://doi.org/10.18637/jss.v027.i08>
- Bergé, L. (2018). Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm. *CREA Discussion Papers*. <https://cran.r-project.org/web/packages/fixest/index.html>
- Blondiau, T., Van Zeebroeck, B., & Haubold, H. (2016). Economic benefits of increased cycling. *Transportation research procedia*, 14, 2306-2313. <https://doi.org/10.1016/j.trpro.2016.05.247>
- City of Toronto. (n.d.). ActiveTO. <https://www.toronto.ca/explore-enjoy/recreation/activeto/>
- Comtois, D. (2022). *summarytools: Tools to Quickly and Neatly Summarize Data*. <https://cran.r-project.org/web/packages/summarytools/index.html>
- Hartig, F. (2022). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. <https://CRAN.R-project.org/package=DHARMa>
- NCSS Statistical Software. (n.d.). Chapter 328 Zero-Inflated Negative Binomial Regression. https://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Zero-Inflated_Negative_Binomial_Regression.pdf
- Oja, P., Titze, S., Bauman, A., De Geus, B., Krenn, P., Reger-Nash, B., & Kohlberger, T. (2011). Health benefits of cycling: a systematic review. *Scandinavian journal of medicine & science in sports*, 21(4), 496-509. <https://doi.org/10.1111/j.1600-0838.2011.01299.x>
- Pereira, R. H. M., Saraiva, M., Herszenhut, D., Braga, C. K. V., & Conway, M. W. (2021). r5r: Rapid Realistic Routing on Multimodal Transport Networks with R5 in R. *Findings*, 21262. <https://doi.org/10.32866/001c.21262>

R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <http://www.R-project.org/>

Van Lierop, D., Grimsrud, M., & El-Geneidy, A. (2015). Breaking into bicycle theft: Insights from Montreal, Canada. *International Journal of Sustainable Transportation*, 9(7), 490-501. <https://doi.org/10.1080/15568318.2013.811332>

von Bergmann, J., Shkolnik, D., & Jacobs, A. (2021). *cancensus: R package to access, retrieve, and work with Canadian Census data and geography*. <https://mountainmath.github.io/cancensus/>

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. Springer.