Fire Safety Regulation Violations in Toronto: Analysis by Property Type and Inspection Duration for Inspections Closed in 2024*

My subtitle if needed

Yunkai Gu

November 26, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

Table of contents

---

*Code and data are available at: https://github.com/Kylie309/Toronto-Fire-Inspection.

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2….

2 Data

2.1 Overview

This paper uses Highrise Residential Fire Inspection Results dataset obtained from Open Data Toronto as the main source of data (City of Toronto 2024b). The portal is launched by City of Toronto to meet growing demand for open data. The raw dataset presents information of properties where Inspectors from Toronto Fire Services (TFS) have examined on, as well as whether the properties violated fire safety regulations and the details of violations if so.

We use the statistical programming language R (R Core Team 2023)….

Following Alexander (2023), this paper uses Bayesian modeling to predict the pass or fail status for safety checks of the properties in Toronto. The factors of region, property type and inspection time intervals from the dataset are mainly chosen and considered to examine their effects on the likelihood of violations for the properties.

2.2 Measurement

The dataset captures data related to the results of fire safety inspections conducted by Toronto Fire Services (TFS) within highrise residential buildings, which is guided by the Fire Protection and Prevention Act, 1997 (FPPA) and regulations under the FPPA, like the Ontario Fire Code (City of Toronto 2024a). It indicates properties where violations have been found by an Inspector which are required to be fixed for compliance at the time with the Ontario Fire Code as well as properties with no observable violations (City of Toronto 2024b).

The data collection process was conducted by TFS Fire Inspectors. Inspectors assess tangible aspects of fire safety based on observable and measurable attributes of the physical environment within the properties, such as proper storage of combustible materials, functionality of fire alarms and fire pumps, maintenance records for fire safety equipment and others in reference to the regulations under FPPA.

After the inspections were completed, the observations have been translated into records and written into collected data. The raw data entry presented by Open Data Toronto includes three main aspects: property identifier (such as address, property type), inspection information (such as inspection start and end dates), and inspection outcomes (pass or fail status for safety checks, and details of violations).

Note that this dataset only includes 'closed' cases, which are cases whose inspection processes have ended. The ongoing inspections were not available in the dataset, and therefore would lead to potential loss of information. This would be further evaluated in detail in the discussion section.

2.3 Outcome Variable

2.3.1 Violation

The outcome variable for the model is the violation status of the properties. In other words, it is the variable that represents whether the properties inspected have been found violating fire safety regulations by an Inspector or not.

In the raw dataset, columns 'VIOLATION_FIRE_CODE', 'VIOLATION_DESCRIPTION' and 'VIOLATIONS_ITEM_NUMBER' indicates the outcome of the inspections. The previous two provides detailed information on fire code under which violation was noted. The latter one presents the order number of violations by code, and shows 0 if no violations observed.

Therefore, a new column 'violation' is mutated to make the analysis procedure easier. It contains numeric values of 1 and 0, which indicates violation observed and no violation observed respectively. The outcome variable is thus binary.

Figure 1 presents the visualized counts of properties that were found violations versus those that were not. The height of each bar shows the total number of cases corresponding to each

category. We could see that the orange bar is significantly taller, indicating that among all cases selected for analysis, violations were reported more frequently in the dataset (9829 cases out of 13193). On the other hand, the blue bar is shorter, indicating there were fewer cases without violations (3364 cases out of 13193).

In conclusion, the data contains a higher number of cases with violations (1) compared to cases where no violations were observed (0), suggesting that violations are common.

Figure 1: Counts of Properties that Were Found Violations versus Those that Were Not

2.4 Predictor Variables

2.4.1 Inspection Time Interval

One predictor included in the model is the inspection time interval. It refers to the duration of the fire inspections conducted by the TFS Inspector for each case.

In the raw dataset, there are two columns recording the date: 'INSPECTIONS_OPENDATE' and 'INSPECTIONS_CLOSEDDATE'. Previous one is the date TFS opened the inspection file, whereas the latter one is the date TFS closed inspection file or date or date enforcement proceedings ended (whichever is later).

During the data cleaning process, the date is converted into time interval to enable easier analysis. A new column named 'date_num' is mutated by calculating the time difference between the end time and the start time of the inspection, in the unit of days.

Figure 2 shows the distribution of the inspection time interval. X-axis represents the number of days of each inspection case, and Y-axis represents the counts. The distribution has peak at low intervals, and the highest bar is near 0 days, indicating that a large number of properties have very short inspection time intervals. The right-skewed inspection suggests that as the inspection interval increases, the corresponding counts decrease stably. The long tail of the right-hand side of the distribution means that there are properties with very long intervals (up to 1500 days or more).
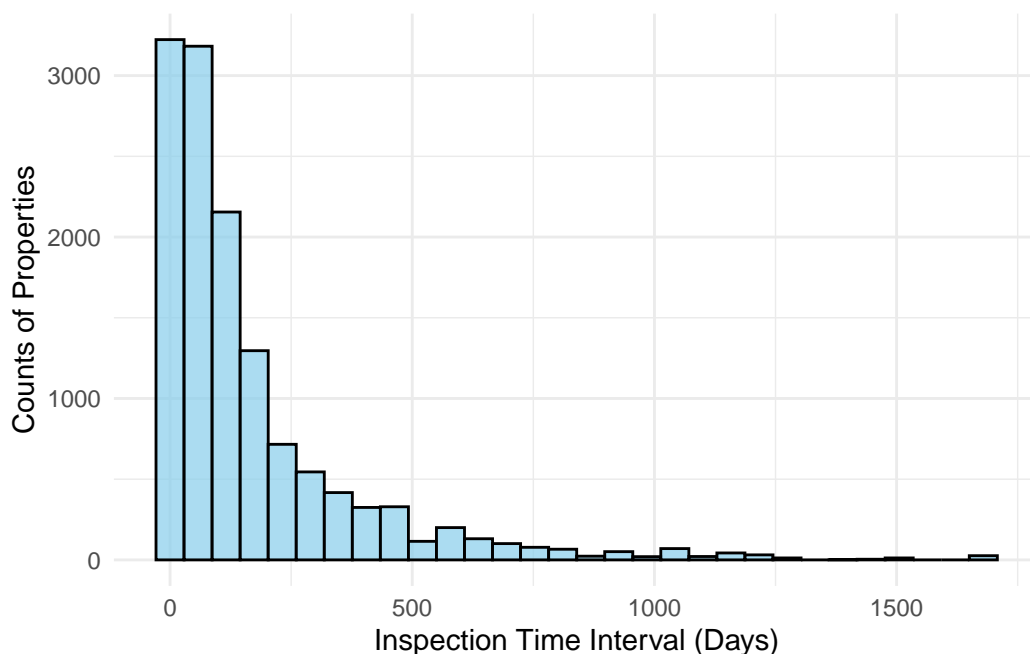

Figure 2: Distribution of Inspection Time Interval

2.4.2 Property Type

The second predictor variable included in the model is the property type. It refers to the occupancy type of the property where the inspection took place, and is directly contains in the raw dataset.

To visualize the variable, two bar plots are created separately in Figure 3. The left panel presents the number of occurrences of "High Rise" and "Low Rise", with the bar representing "High Rise" relatively taller, indicating a higher count compared to low rise. This suggests that the inspection results and data collection focus more heavily on high-rise buildings.

The right panel includes the other categories of properties. Specifically, the types include: "Detention," "Group Home," "Group Home (VO)," "Hospital," "Hotel & Motel," "Nursing Home," "Residential Care," and "Rooming House." We could see that "Rooming House" has the highest count among these property types, whereas categories of "Detention" and "Group Home" have relatively lower counts, indicating that they are less represented in the dataset.

In general, there is a diverse distribution across the property types, with some being far less frequent than others. High-rise properties dominate overall, while rooming houses stand out within other categories. The dataset might reflect an emphasis on inspecting high-rise and rooming houses more than other property types.
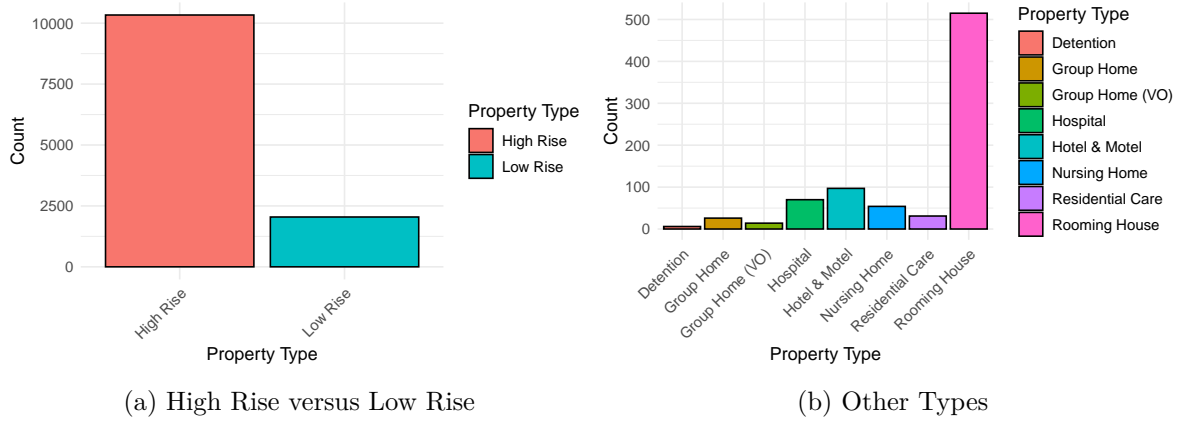
(a) High Rise versus Low Rise

(b) Other Types

Figure 3: Counts of Property Types

3 Model

The goal of the modelling strategy is twofold.

Firstly and most importantly, from the statistical standpoint, the goal is to estimate the the detection of violation from an inspection based on its time interval and the property type.

Secondly, from a more practical standpoint, by predicting which type of property with how long duration is more likely to be observed violations, more efficient inspection process could be constructed. Identifying certain property types that would more frequently violate the fire safety regulations could provide guidelines for interventions and help to establish targeted and focused actions towards them.

The model constructed in the paper is Bayesian logistics regression model designed to predict the violation status (denoted as violation) based on the fixed effect of inspection time interval (denoted as date_num) and property type (denoted as property_type). It is designed to predict the inspection results based on the two predictor variables.

Following sections define, explain and justify each model and the variables, as well as discuss underlying assumptions, potential limitations, software used to implement the model, and evidence of model validation and checking.

Background details and diagnostics of the model are included in Appendix B.

3.1 Model set-up

The Bayesian model was implemented using the R programming language (R Core Team 2023), specifically utilizing the `rstanarm` package of (**citerstanarm?**). This package provides an interface for fitting Bayesian regression models using Stan, and the models are fit using the package.

Define y_i as the binary response variable for observation i. $y_i = 1$ indicates that the case has been observed violations, while $y_i = 0$ indicates that no violation occurs. It follows a Bernoulli distribution:

$$y_i = \text{Bernoulli}(p_i)$$

where p_i is the probability of a property that is observed to have violations of fire safety regulations.

Then define the inspection time interval for observation i as date_num and the property type as property_type.

The Bayesian model could be defined by the following mathematical expressions:

$$\text{logit}(p_i) = \beta_0 + \beta_1 \cdot \text{date}_i + \beta_2 \cdot \text{type}_i$$

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

$$\beta_0 \sim \text{Normal}(0, 2.5)$$

$$\beta_1 \sim \text{Normal}(0, 2.5)$$

$$\beta_2 \sim \text{Normal}(0, 2.5)$$

where:

- $\text{logit}(p_i) = \log \left(\frac{p_i}{1-p_i} \right)$ is the log-odds of p_i.
- β_0 is the intercept term of the Bayesian logistic regression, representing the baseline log-odds of a violation when `date_num = 0` and `property_type` is the reference category. Formula (2) specifies that β_0 follows normal distribution with mean equals 0 and variance equals 2.5^2.

- β_1 is the coefficient for predictor variable `date_num`, representing the change in log-odds of a violation per unit increase in `date_num`, the inspection time interval. Formula (3) specifies that β_1 follows normal distribution with mean equals 0 and variance equals 2.5^2.
- β_2 is the coefficient for predictor variable `property_type`, representing the change in log-odds of a violation per unit increase in `property_type`. Formula (4) specifies that β_2 follows normal distribution with mean equals 0 and variance equals 2.5^2.

Formula (2), (3) and (4) specifies priors for the model. This prior reflects the assumption that coefficients are centered around 0, with a standard deviation of 2.5, allowing moderate variability while discouraging extreme values.

3.2 Model Assumptions and Limitations

The main, underlying assumptions and their limitations for the Bayesian model is discussed below:

Firstly, the model assumes that each observation is independent of others. This assumption suggests that violations at one property should not influence violations at another property unless explicitly modeled (e.g., through random effects).

If observations are clustered and the assumption is violated (for instance, multiple observations were conducted in the same region or by the same inspector), overestimation and underestimation of standard errors and outcomes may occur.

Secondly, the model assumes linearity in the logit. This assumption stated that the log-odds of the response variable ($\text{logit}(p_i)$) are a linear function of the predictors.

If violations change non-linearly over time, the model assumption would be violated, potentially resulting in misfit of the model.

Thirdly, the model assumes priors of $\text{Norma}(0, 2.5)$ for the coefficients. They were constructed on domain knowledge, and poorly chosen priors could make the estimation outcome biased.

Also, if new property_type levels are introduced, the model may not be no longer appropriate for the new situation, as fixed effects could not be feasible to compute estimates. Random effects, then, may be more appropriate.

3.3 Model Justification

Since the response variable is binary for each case and follows a Bernoulli distribution, it would be appropriate to use Bayesian logistic regression model to estimate it.

The predictor variable as 'date_num' is modeled as fixed effect in the model. There are various reasons for this. Firstly, As 'date_num' represents the number of days for the inspection takes place, it is a numeric variable. Modeling it as a fixed effect allows the model to estimate

its direct, continuous influence on the response variable. Secondly, since it applies across all observations, regardless of property type, modeling it as a fixed effect could consider its impact consistently across the entire dataset. It also makes the coefficient of 'date_num' ($beta_1$) be easier to interpret, avoids adding complexity to the model.

The other predictor variable 'property_type' is also modeled as fixed effect. Since it is categorical variable, modeling it as fixed effect allow the model to estimate a separate coefficient for each category, relative to a reference level. This would make it easier to compare between property types, in order to understand which types of properties would be observed to have violations. Moreover, since 'property_type' has only 10 levels, it is feasible to compute it as fixed effect and conduct estimation without overloading the model.

The prior of $Normal(0, 2.5)$ for every coefficients is also reasonable because it reflects prior knowledge that most predictors have small-to-moderate effects on the log-odds scale. This assumption would avoid overfitting to extreme values.

The above final model decision is made after considering several alternative models and variants. One alternative is a simple linear regression model (SLR), which assumes a constant linear relationship between the predictor (violation) and the response (date_num). Although the model is extremely straightforward and easy to interpret, it is far too simple for estimating and neglect the key variable of 'property_type'. More importantly, the response variable is binary, whereas SLR assumes a continuous response variable. MLR also assumes so, and therefore is considered and rejected because of the same reason.

Another alternative considered is the Bayesian model with natural splines (e.g. $f(\text{date}_i)$). While it could model complex, non-linear time trends and is appropriate for 'date_num', the expected model for this study does not primarily focus on exploratory analysis on non-linear relationships between date_num and violations. Non-linear functions would also overfit the data if not regularized properly.

While considering alternative Bayesian logistics model, the model with 'property_type' as fixed effect was also considered. This would capture variability across property types without explicitly estimating coefficients for each type, and does not provide explicit estimates for each property type. The model would be suitable if there are many property types, but the number of types is manageable to include it in the model as fixed effect.

Therefore, after considering many alternatives, the final model of Bayesian logistics model with 'date_num' and 'property_type' as fixed effect is chosen. It could directly provides explicit comparisons between property types without overfitting the model, and meet the focus of the research question.

4 Results

Our results are summarized in Table 1.

9

Table 1: Explanatory models of flight time based on wing width and wing length

	(1)
(Intercept)	−6.80
	(2.25)
date_num	0.02
	(0.00)
property_typeGroup Home	5.93
	(2.34)
property_typeGroup Home (VO)	2.99
	(2.55)
property_typeHigh Rise	6.30
	(2.23)
property_typeHospital	4.76
	(2.29)
property_typeHotel & Motel	5.64
	(2.26)
property_typeLow Rise	5.70
	(2.24)
property_typeNursing Home	1.26
	(2.39)
property_typeResidential Care	4.55
	(2.30)
property_typeRooming House	6.22
	(2.24)
Num.Obs.	13 193
R2	0.341
Log.Lik.	−4947.974
ELPD	−4963.5
ELPD s.e.	105.1
LOOIC	9927.1
LOOIC s.e.	210.2
WAIC	9926.2
RMSE	0.32

5 Discussion

5.1 What is done in this paper?

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 What is something that we learn about the world?

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

5.3 What is another thing that we learn about the world?

5.4 Weaknesses and next steps

Firstly, the limitations in the process of measurement of the original dataset should be considered. Specifically, the limitations include:

- Observer Bias

Although inspectors conduct the violation detection procedure in reference to the same, standardized fire safety regulations (Ontario Fire Code, FPPA and Municipal Code, according to City of Toronto) to ensure consistency, their subjective judgment could still play a significant role in the examination process.

Various interpretations of the guidelines and different decision-making procedures made by the inspectors based on specific situations could impact the outcome of the inspection. Since the examinations have been thoroughly completed by human, the assessments could differ depending on the inspectors' focus and perspectives on certain regulations, influencing the observations recorded, even with clear rules.

- Data Completeness

The dataset does not capture all fire safety-related phenomena in Toronto, as it only contains the properties that have been inspected by the Inspectors from TFS. The properties that are not included in the list fail to provide any information, which would make the dataset not representative enough to conclude the situations of fire safety checks for the whole city.

Moreover, even for the properties that have been inspected by TFS, not all cases are recorded in the dataset. The raw dataset only includes 'closed' cases, which are specific cases where the inspection process has ended, and does not account for cases currently under review or in progress.

In other words, the properties whose initial observations have been made but further investigation or corrective actions are required, are excluded from the dataset. Therefore, the dataset does not represent the full scope of fire safety checks at any given time. The active cases would be neglected in analysis procedure, providing limited insights to identify fire safety concerns in real-time.

Appendix

A Additional data details

B Model details

B.1 Posterior Predictive Check

During the modeling process, posterior predictive checks (PPC) is used.

PPC is the comparison between what the fitted model predicts and the actual observed data, which validates whether the fitted model is compatible with the observed data. The aim is to detect if the model is inadequate to describe the data (**ppcheck?**). This type of diagnostic visualizes the fit of the model by comparing the observed data (y) to the posterior predictive distributions (y_{rep}) generated by the model.

In this PPC plot, x-axis represents predicted probability of the binary outcome p_i, which is the probability of observing violation for a property during inspection; y-axis represents density or frequency of the predicted probabilities or aggregated counts.

The dark black line presents the distribution of the observed outcome (y), which is a smoothed estimate of the distribution of binary response 'violation'. On the other hand, the light blue lines represent the posterior predictive simulations, which are the density of the replicated data (y_{rep}) generated from the posterior predictive distribution of the fitted model.

Whether the black line matches well with the blue lines indicates how well the model fits the observed data. If the black line aligns well with the blue lines, it suggests that the model is performing well in capturing the actual data distribution, and vice versa.

Figure 4 shows the posterior predictive checks for the Bayesian model.

The plot shows that the curves for y (observed data) and $y_{r}ep$ (posterior predictive distribution) align closely with each other, suggesting that the fix effects of two predictor variables capture the overall pattern of the observed data well, and the Bayesian model generates predictions consistent with the observed data. Lack of mismatches between curves mean that the model provides good fit of data. Near extreme probabilities of 0 and 1, two curves tightly follow each other, indicating that the model correctly predicts extreme outcomes.
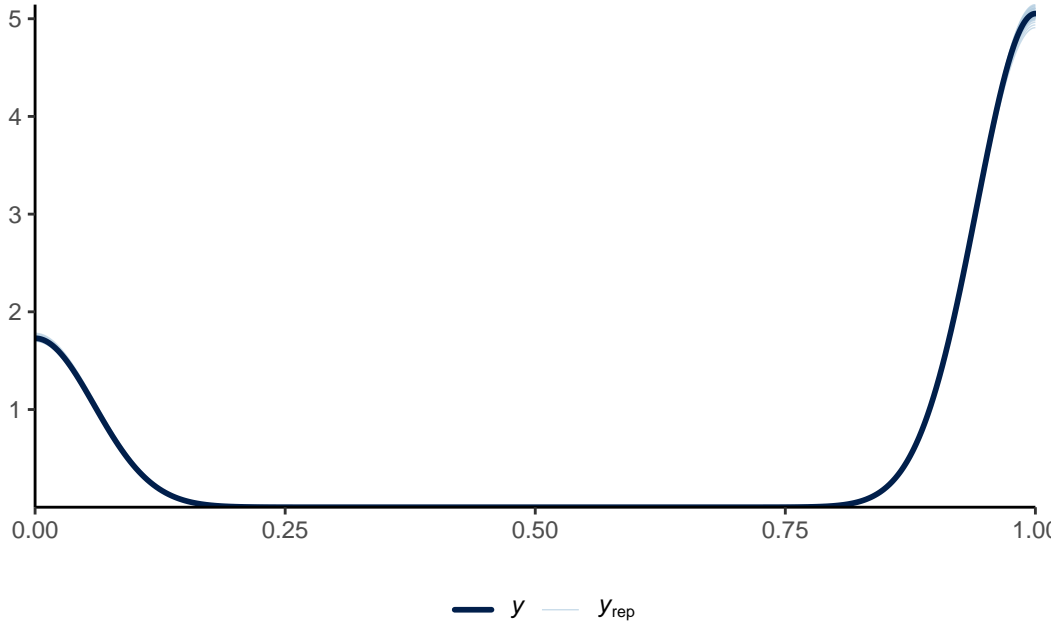
Figure 4: Posterior Prediction Check

B.2 Posterior vs Prior

Comparing the posterior distribution with the prior one is also necessary for model validation. It examines how the model fits and is affected by the data.

Prior distribution represents the beliefs about the parameter values before observing any data, which is constructed by our initial assumptions. Posterior distribution reflects updated beliefs about the parameters after observing the data.

In the plot, the posterior distribution is demonstrated on the left side, and the prior distribution is on the right side. Y-axis shows every parameter included in the model, each colored dot on the graph represents a parameter, and the horizontal line indicates the uncertainty.

?@fig-posteriorvsprior, compares the posterior with the prior distribution of the model, validating whether data has a significant impact on the parameter estimates and whether the model is well-fit.

Shown by the red dot representing "intercept", the narrower posterior compares to the prior one suggests that the model provide good estimate in baseline.

Shown by the blue dot representing predictor variable 'date_num', the posterior mean is close to 0, indicating little effect of it on the probability of violations. On the other hand, the credible interval is much narrower than the prior, reflecting strong eveidence from the data.

The rest of dots representing different levels of predictor variable 'property_type' show various posterior means and credible intervals. For instance, for high-rise buildings, the posterior mean is positive, indicating that high-rise buildings have higher log-odds of violations than the reference category. In general, each level shows a narrower posterior distribution than the prior, reflecting that the uncertainty decreases and the posterior distributions center around data-driven values.

B.3 Diagnostics

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

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

Checking the convergence of the MCMC algorithm

References

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

City of Toronto. 2024a. "Fire Inspection Results." https://www.toronto.ca/city-government/accountability-operations-customer-service/access-city-information-or-records/fire-inspection-results/#listing/eyJxdWVyeVN0cmluZyI6IiIsInRvcCI6MTB9/1.

———. 2024b. "Highrise Residential Fire Inspection Results." https://open.toronto.ca/dataset/highrise-residential-fire-inspection-results/.

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