Vision over Incidents and Claims Research Proposal

Simardeep Kaur, Merve Sahin, Brayden Tang, Xugang (Kirk) Zhong

Summary

In 2019/2020, TransLink has paid around 20 million in expenses for insurance claims to ICBC. Additionally, their premium to ICBC has increased by 200% during the last five years. To investigate the potential factors related to the increasing insurance cost and identify the pattern among these predictors the research is proposing a four-phase research plan, using statistical methods such as zero-truncated regression, time series modeling, and mixed-effect models. In a reproducible and interactive form, the research results will be delivered to TransLink at the end.

Introduction

With the largest transit service area in Canada, TransLink is operating more than 245 bus routes and 79 kilometers of rapid transit to meet the transportation needs of 2.5 million people in Metro Vancouver as of the end of 2018 (TransLink 2018). Legislation requires TransLink to carry a $1 million per occurrence liability policy on each of its revenue vehicles and a $200,000 per occurrence liability policy on each of its non-revenue vehicles. Since 2014/2015, the premium paid to ICBC has increased by over 200% to cover onboard passenger injuries, cyclist injuries, pedestrian injuries, and losses from collisions with third party vehicles. For at-fault physical damage losses to its vehicles, the premium paid to its own captive insurance company has increased by 33%. In response to soaring insurance costs and road safety concerns, this research will investigate variables of interest that may influence the frequency of transit incidents and claims. The research objectives are to:

- Characterize current patterns of incident frequency in terms of variables related to drivers and vehicles
- With the help of external data such as weather conditions and geographic data, examine how predictive these variables are of incidents/claims
- Recommend actions for TransLink to reduce the incident frequency strategically
- Identify research needs for future study

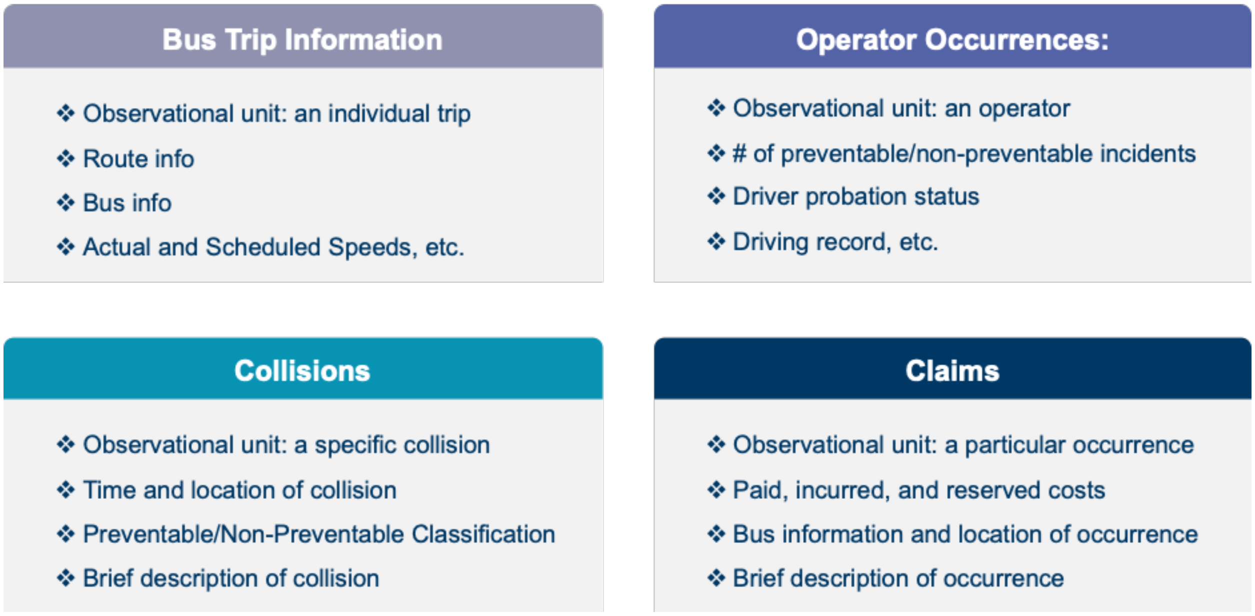
Data Product

A fully reproducible and interactive report will be delivered at the end of the project. The report will give a visual representation of relationships between the frequency of claims and specific variables interactively. The project will use a fully reproducible data pipeline so that the user can run the entire analysis using simple and understandable make commands. A docker container for reproducibility on any operating system will also be included.

Data Description

Below is a summary of the data sets that this project will be built on.

The key variables in each dataset that have the highest potential to answer the core business problems are: - Bus model and length from the "*Bus Trip Information*" dataset. - Experience and probation status of the drivers from the “*Operator Occurrences*” dataset. - Loss location and claim description from the “*Collisions*” dataset. - Loss date and more granular information about the claim amount from the “*Claims*” dataset.



Bus Trip Information

- ❖ Observational unit: an individual trip
- ❖ Route info
- ❖ Bus info
- ❖ Actual and Scheduled Speeds, etc.

Operator Occurrences:

- ❖ Observational unit: an operator
- ❖ # of preventable/non-preventable incidents
- ❖ Driver probation status
- ❖ Driving record, etc.

Collisions

- ❖ Observational unit: a specific collision
- ❖ Time and location of collision
- ❖ Preventable/Non-Preventable Classification
- ❖ Brief description of collision

Claims

- ❖ Observational unit: a particular occurrence
- ❖ Paid, incurred, and reserved costs
- ❖ Bus information and location of occurrence
- ❖ Brief description of occurrence

Methodology

The research will focus on preventable, rather than non-preventable, claims incidents. Non-preventable incidents are less likely to lend themselves to easy interventions that could reduce insurance-related costs. Preliminary analysis will focus on the incident occurrence rather than incidence cost. This is because claims can remain open for an extended time. which means costs cannot be assessed accurately.

Multiple separate analyses will be used to determine variables of interest in predicting incident occurrence. Individual variables, such as time, operator, location, and bus type, may require different statistical or machine learning approaches to relate them to the variable of interest. Through this analysis, features can then be created that can be used to fit one composite model incorporating all predictors simultaneously. To assess the predictive power of time and weather on occurrences, the research will involve a time series analysis, incorporating weather as an exogenous variable. In addition, the predictive power of bus type and location are to be analyzed using mixed-effect model approaches, due to the natural grouping of the data (region and bus category). For simplicity, a linear mixed model will be investigated first, using region and potentially bus category as a random effect.

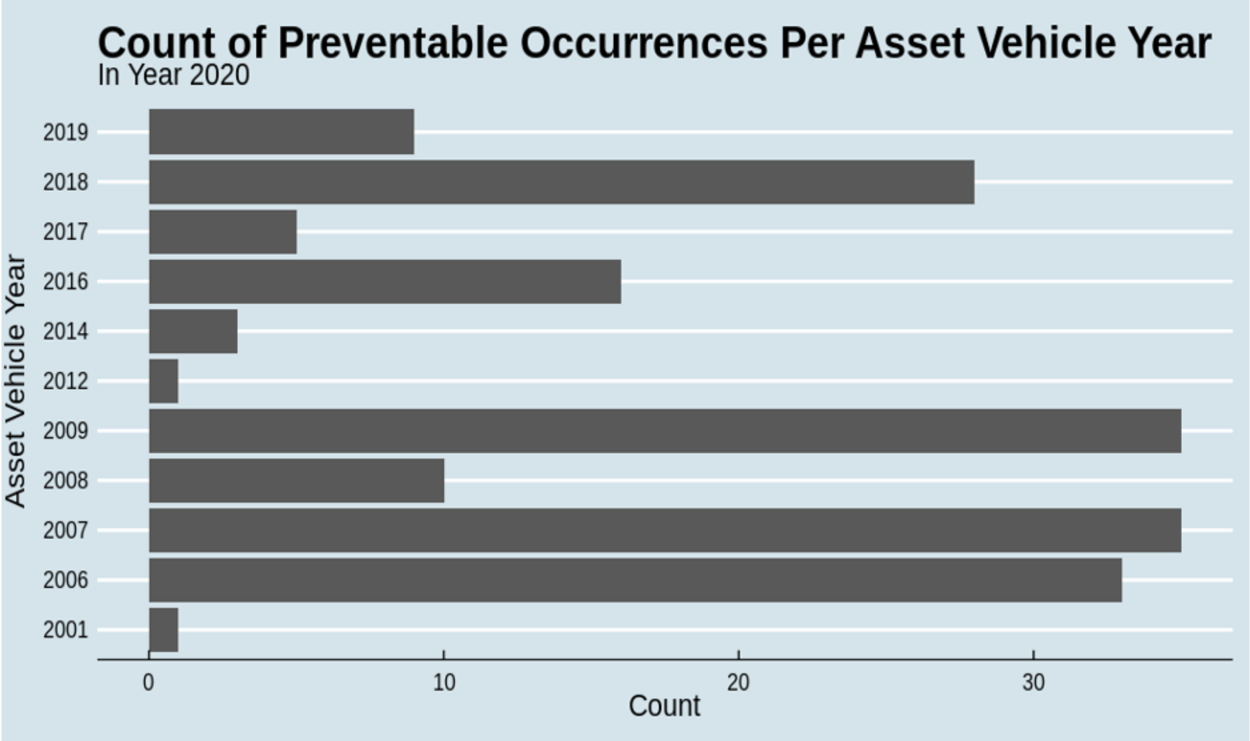


Figure 1: Count of preventable occurrences per asset vehicle year. Older vehicles appear to show a greater number of incident occurrences. However, it is essential that the number of each type of bus in use (in each year) is obtained to get a fairer comparison so that rates can be calculated instead.

The operator dataset only contains operators with at least one occurrence. Therefore, to account for this bias the research will examine methods such as zero-truncated models or Bayesian regression models. Finally, to find out whether certain descriptions and/or codes exhibit similar loss experience, methods such as standard cluster analyses will be deployed to identify potential groupings. Furthermore, topic modeling for the manually written accident descriptions will be used to further determine potential groupings or patterns in claim behavior.

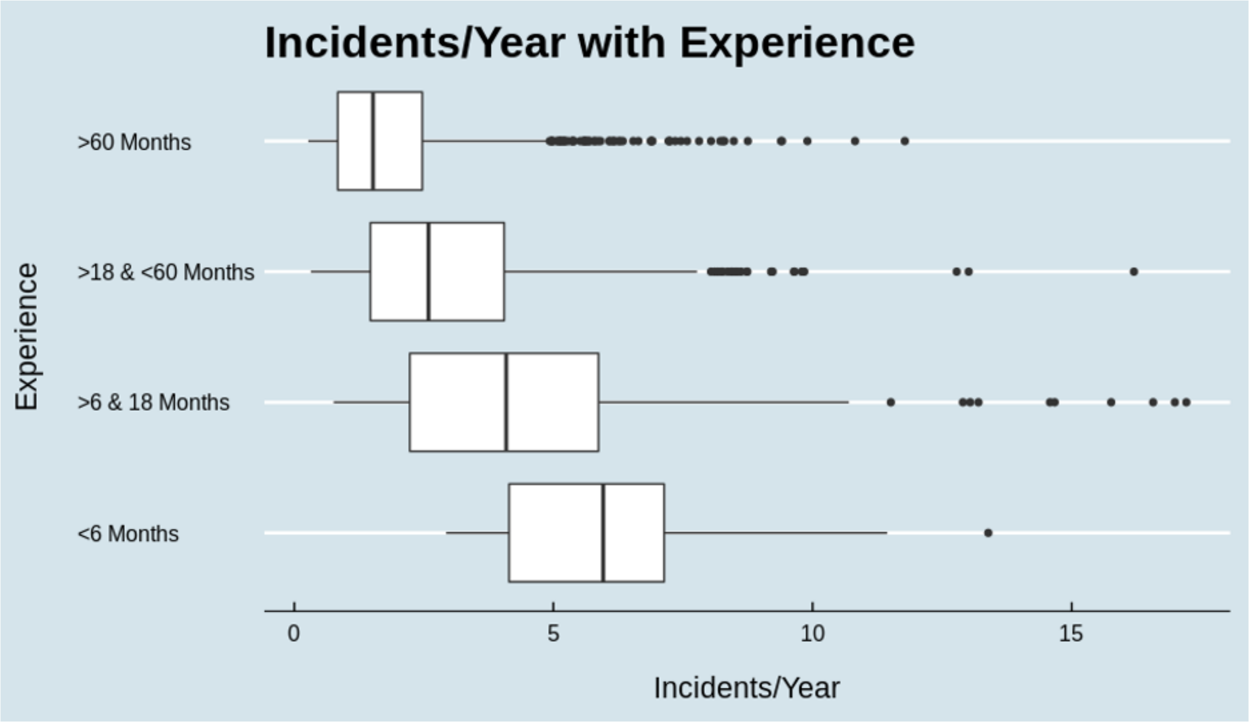
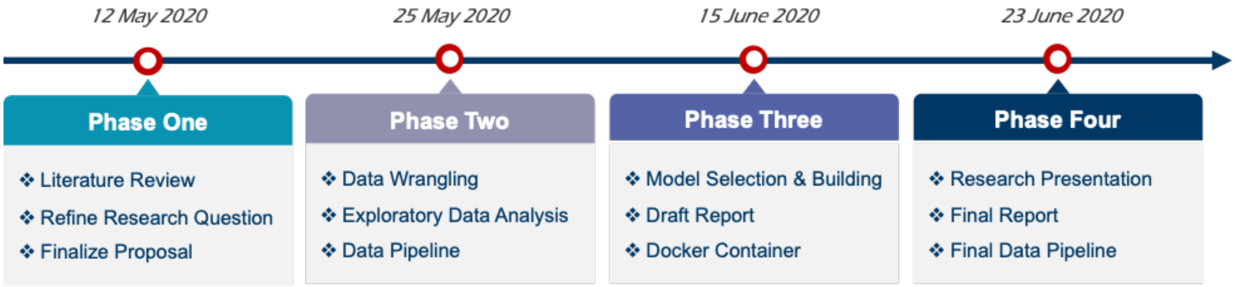


Figure 2: The incidents/year against operator experience. There is a very clear correlation between operator experience and incident rate.

Timeline

The project will be completed with a four-phase research plan. After refining the research objectives, the proposal is expected to be composed during the first phrase. The second phase consists of three tasks: data preprocessing and getting some preliminary results via exploratory data analysis. Findings will be communicated directly to TransLink to further explore and implement any possible feedback to improve the analysis. Meanwhile, a Makefile is expected throughout the project to establish a data pipeline. In the third phase, the final predictive model will be developed, and the final interactive report will be generated. With a docker container, the data pipeline will be completed to make the project fully reproducible. The fourth phase is dedicated to final reviews and completion of the project.



Reference