An Analysis on Commercial Real Estate Property Assessment and Property Tax in B.C.

Gian Carlo Di-Luvi   Mallory Flynn   Sophia Li   Vittorio Romaniello
STAT 550 Project Proposal
March 2020

Department of Statistics
University of British Columbia

Summary

Property taxes are the single greatest operating expense for property owners in British Columbia, and depend on two factors: property assessment—published each year in January—and municipal mill rates—published in April. Accurately projecting annual mill rates between January and April and future years assessment values between May and December would allow businesses and individuals to budget for these expenses, a value-add to the client's consulting services. This project will build statistical models to accurately predict mill rates and property assessment values. For this purpose, an expanded exploratory data analysis will be conducted to inform relationships between relevant property-related variables. It will also scrutinize trends uncovered by the STAT 450 analysis. Multiple statistical models—including stepwise mixed effects models, random forest regression, and simultaneous equations model—will then be explored. After the most suitable model is chosen, a Shiny app will be created to provide the client with a straightforward user interface which incorporates model predictions into a property-specific tax assessment.

1 Objectives

This project will address two research questions, building on the analysis of STAT 450 students. First, it will determine how assessment values in a given year, as well as other relevant property-related variables, affect mill rates before they are released in April. Second, it will explore how a given year's assessment values and mill rates affect the next year's assessment values.

In order to address these research questions, our analysis will build predictive models that accurately forecast both mill rates and assessment values for municipalities and properties in British Columbia. Mill rates for three tax classes will be predicted - residential (class 1), light industrial (class 5), and commercial (class 6). An exploratory data analysis (EDA) will also be conducted, which will further investigate aspects of the EDA completed by STAT 450 students.

2 Data

A thorough description of the data can be found on the STAT 450 students' report. The data set used in this report is the same, with the caveat that information from all municipalities—not only Metro Vancouver—is also considered. The variables are summarised in Table 1.

Variable	Type	Included as
Mill rate	Continuous	Dependent variable
Total assessment	Continuous	Independent variable
Total land assessment	Continuous	Independent variable
Total improvement assessment	Continuous	Independent variable
Tax class code	Categorical. One of 1 (residential), 5 (industrial), 6 (commercial)	Independent variable
Municipality	Categorical	Independent variable
Year	Discrete. Values of 2016, 2017, 2018, and 2019	Independent variable. Random slope/intercept

Table 1: Brief description of the variables included in the data set.

3 Analysis outline

The analysis will be divided in two parts: (1) exploratory data analysis (EDA) and (2) modeling and evaluation.

3.1 Exploratory data analysis

Our EDA will be similar to the one conducted by STAT 450 students. However, we will try to gain more insight from visualisations, meeting the client's request to understand relationships between the variables. More specifically, we plan on observing how mill rate changes not only within a municipality but also within tax classes. Figure 6 in the STAT 450 final report shows a difference in mill rate across tax classes, and so we deem necessary to further explore the impact of this variable at the level of municipalities. Furthermore, we have observed a negative relationship between mill rate and average total assessment found in Figure 4 of the STAT 450 report. Such an inverse relationship makes intuitive sense because mill rate and average total assessment for each municipalities should balance each other out in reality.

In addition to informing our analysis through visualisations, we plan on exploring whether transformations help improve understanding of the data and finally, model performance. In particular, a log transformation may help better analyze information related to property assessment, given the high variance and right-skewness present in the data.

3.2 Modeling

To answer the research questions we will explore a variety of models, which we outline below.

Mixed effects models and Generalized Estimating Equation The data contains repeated measurements taken over years for each subject. Based on the EDA, linear mixed models (LME) or nonlinear mixed effect models (NLME) will be fit to the data, with random effects on municipality and/or individual houses. For mill rates, municipalities will be treated as subjects, while for assessment values, individual properties will be treated as subjects.

The NLME and LME models will be compared with General Estimating Equation (GEE) using diagnostics tools. GEE has poor interpretability but good prediction capabilities since it does not have any distributional assumptions on the data. NLME and LME have better interpretability, but assumptions will need to be verified.

Random forest regression Given the clients' interest in prediction accuracy, we plan to exploit the predictive power offered by random forest in this setting. Since the response variable is continuous, we propose random forest regression.

Random forest is a boosted decision tree-based algorithm that is preferred over a single regression tree as it helps prevent overfitting and reduces the variance of predictions. Both problems are likely to appear in our situation, as the data contains a limited number of data points per property.

Simultaneous equations model The mixed effects model and random forest regression can only predict one outcome at the time. To answer the research questions, we will need to build two models under this methodology - one to predict mill rate, and one to predict the next year's assessment values. Using two models, where one takes as input the output of the other, increases prediction variance and can generally lead to highly biased predictions as the error accumulates over the models.

We address this problem proposing a simultaneous equations model. This model, popular in the economics literature, models a set of linear relationships where the response variable appears as explanatory variable for the other response variable (e.g. modeling supply and demand in economics). Such approach allows to answer both research questions with one model.

Initial covariates will be chosen based on previous work by STAT 450 students and client suggestion. Feature selection will be performed using stepwise model selection based on Akaike information criterion (AIC), and multicollinearity will be assessed using the variance inflation factor (VIF). In order to evaluate each model, the data will be randomly split into two data sets: a training set and a test set. Each model will be fitted using the training data. Each model will then be evaluated by its prediction accuracy on observations in the test set. Because the client is interested in high accuracy predictions, the model with the most precise test set predictions will be chosen as the final model. This process will be repeated for both research questions.

4 Shiny app

After the modelling process is over and the final models are selected, they will be implemented in a Shiny app. Once the government publishes the assessments in January, the client will be able to estimate that year's tax payment based on previous years' information and the current years' assessment. After the mill rates are released in April, the client will then be able to estimate next year's assessment. The app will allow the client to export the information into Excel, which will make further analyses easier to carry out.