Examining the case behind Acer Platanoide's plight on Toronto's Street Trees*

Veyasan Ragulan

December 3, 2024

Acer plantiodes is an invasive species that has taken over the streets of Toronto. Using the Street Tree data from OpenData Toronto, this paper examines the patterns and trends behind this spread. A linear model was created to predict species based on ward, street and cross streets. Uncovered was a abundance of a non-native tree species, and a lack of trees on the streets of Toronto's core

1 Introduction

Urban spaces can be void of vast expanses of nature, whether it be grass fields, blooming flowers, or teaming and diverse natural wildlife. One simple way to tackle this is to incorporate nature within the urban landscape, such as with trees. Trees are a vital plant in the world, they take in carbon dioxide and provide oxygen, a key component in the air we breath. Urban areas are known for the excess in carbon dioxide, which only makes it seem natural that trees can live and perhaps thrive in such environments. One way cities have incorporated trees into the city are roadside trees. Trees line many of the cities highways, arterials, collectors, and small roads. Lines of trees lined alongside highways can act as a natural sound barrier by reflecting, scattering and absorbing nearby noises (Dobson and Ryan (2000)). Tree-lined streets can also alter the way wind affects pedestrians below them, creating a more hospitable environment to promote active transportation such as walking or biking (Krayenhoff et al. (2020)). It should be no surprise then, that many cities around the world, including Toronto, have incorporated trees into their landscape, in busy downtown centres, to sleepy suburbs, and everywhere in between.

Acer platanoides is a well known tree used in urban environments, especially after the demise of elm trees in the 1970s (Sandberg, Bardekjian, and Butt (2014)). These trees are know for 'a vigorous early growth rate, desirable form and size, the capacity to withstand many urban

*Code and data are available at: https://github.com/Veyasan1/STA3014-Paper1

impacts (e.g. pavement, moderate levels of pollution, dusts, and dry soils) and the abilities to transplant well, grow on a wide variety of soils, and withstand ice and snow damage better than other maples' (Nowak and Rowntree (1990)). Being planted in cities, it's seeds can travel by wind across to more wooded suburbs, where the seeds can tolerate shade for extended periods of time (Sandberg, Bardekjian, and Butt (2014)). However, as it's common name suggests, it is not a native species to Canada, instead brought here by European explorers and settlers. While not a very invasive species, taking over just 9% of Southern Ontario's forests, it does so by harming other native plants and trees ((**urbanforests?**)). More attention should be cast on the spread of this tree, as it has the potential to devastate Toronto's and Southern Ontario's forests in the years to come.

In this paper we look at street tree data from Toronto's roadside trees, and discern any patterns or observations behind the rise of Acer Platanoides. To do this, a linear model was created to predict botanical_name. Predictors include ward, streetname, and crosstreet1 and crossstreet2. The linear model uncovered TO BE COMPLETED

Section 2 will outline the source of this data, and highlight key varaibles to be used in the model. Section 3 covers the construction and rationale of the model. Section 4 is where the model results will be outlined. Finally, section 5 discusses key takeaways from the data and the model, and addresses weaknesses and limitations that can be considered for another report.

2 Data

2.1 Measurement

All collection and analysis of data was done using R, a free programming language designed for data scientists (R Core Team (2023)). Additionally, tidycerse (Wickham et al. (2019)), dplyr (Wickham et al. (2023)), and janitor (Firke (2023)) have been used as packages to read, clean, and present data. Finally OpenData Toronto is an online database containing thousands of datasets pertaining to the City of Toronto, which is where the data for this paper was initially accessed (Gelfand (2022)).

Rohan Alexander's book, *Telling Stories with Data* was referenced for troubleshooting and general ideas for data analysis (Alexander (2023))

The data comes from Urban Forestry, an organization dedicated to Toronto's urban forests. They work towards planting more trees, and protecting existing trees from damage due to individuals, private entities, and public entities ("Donate to Urban Forestry" (2024)). This particular dataset is called Street Tree Data, and focuses on city-owned trees located on roads. This information would be most likely used by city planners and road maintenance, ensuring snow plows and street dusters don't impede on any trees that line the roads. It is important to note that there may be some privately owned trees listed among the municipal trees, as they may be of interest with regards to road maintenance or other city services.

Street Tree Data contains around 32000 observations, each observation being an individual tree. Each observation has a general id for data analysis, as well as an structure id, telling us if the tree is part of an existing structure or building. Location is covered in 4 ways: first the address number (address) and streetname (streetname) of nearest postal delivery address, next is the street name and the nearest 2 cross streets (crosstreet1 and crossstreet2), third being the ward number (ward), and finally geometry containain a tuple of general latitude and longitude. Street Tree contains both the species name (botanical_name) and the common English name common_name), however we have decided to use only the species name to avoid confusion between similar sounding trees. Common English names will be provided for species of key interest. There is also measurement called dbh_trunk. This is the standard method forestry experts use to measure the diameter of a tree. Depending on where the measurement is taking placce, the breast height changes. For example, in Canada it is 1.3 meters, while the US uses 4.5 feet. This methodology is comprised of old conventions used to measure trees used throughout the world (Snyder (2006)). The result is a complicated mess where breast_height is a vague term that changes by location. Additionally, the convention breaks down when the tree is anything but vertically upright and straight. As a result, the DBH is more of an estimate rather than a sturdy measurement.

2.2 `botanical_name`

To see why Acer platanoides is of interest in this paper, refer to Figure 1. The most planted street tree in Toronto is Acer platanoides, which has at least a 2x lead over any other tree in the data. This trend doesn't seem to be coincidental, and from prelimiary research, the prevalance might be worrisome. Therefore, the paper will use `botanical_name` as the target variable in a linear model, which can be filtered down to Acer platanoides later.

2.3 `ward`

Ward is one of several location varaibles included in the Street Trees dataset. It lists the ward location of each observation (tree) in the city. Toronto is broken up into municipal wards, who is represented in City Hall by a councillor. Wards vary in size and population, which may have an impact in the spread of Acer platanoides. For example, one of the largest Wards is Etobicoke North, which forms the north-western boundary of Toronto. Given its size and distance from downtown Toronto, it has ample room to host large swaths of forests, and have its roads lined with trees, increasing the possibility of Acer platanoides finding its way here. Toronto-Centre is the smallest ward by far, and being situated in downtown Toronto, is more dense with human infrastructure. There is less space for Acer platanoides to grow here, but the few green spots here could be suceptible to Acer platanoides.

According to Figure 2, Etobicoke North (Ward 2) hosts the most trees of the species Acer platanoides, and Toronto Centre (Ward 13) hosts the least. This falls in line with what was said before.

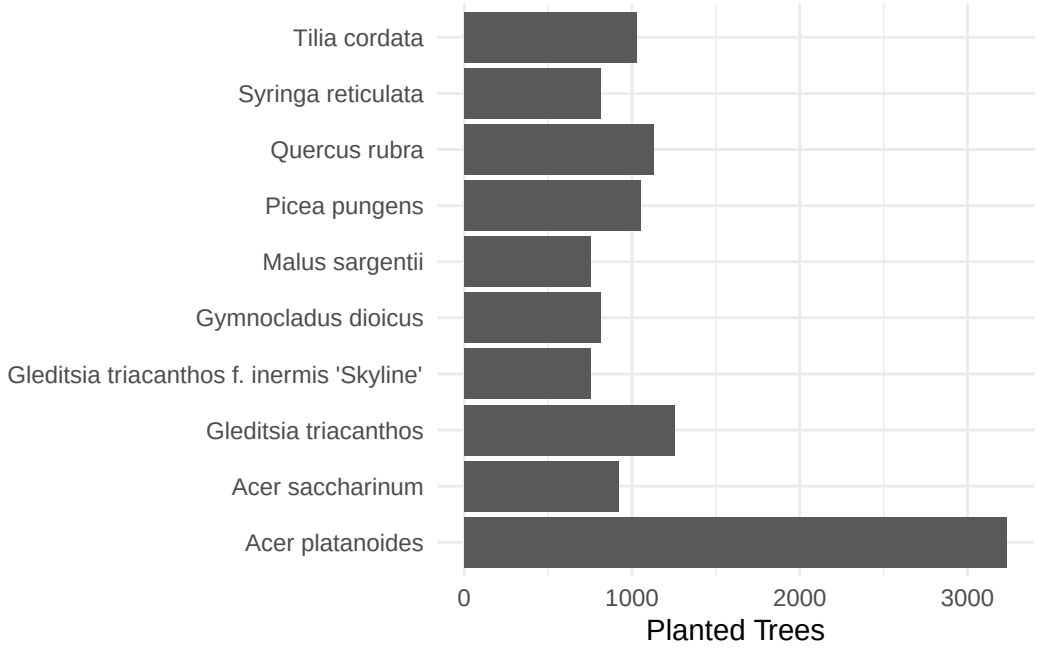

Figure 1: 10 most populous street tree species in Toronto, Ontario

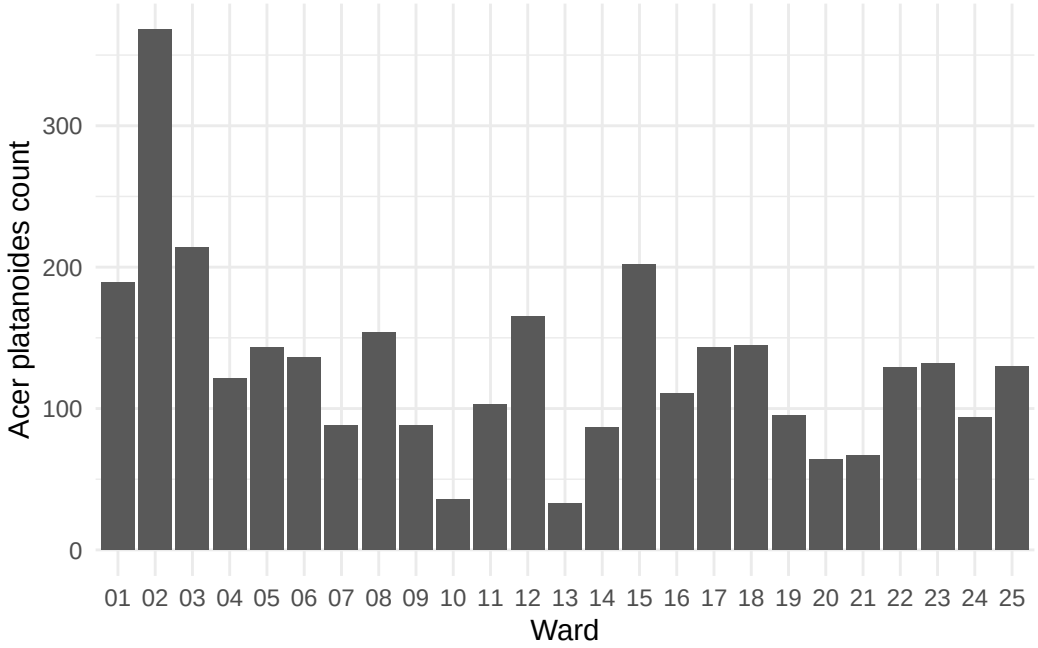

Figure 2: Acer plantiodes distribution across Toronto wards up to September 2024

2.4 `streetname`

Anoter location variable in the Street Trees dataset, `streetname` records the street each observation(tree) lies on. Streets come in many sizes and lengths across Toronto. Arterial streets such as Bloor st / Danforth Ave might be able to hold more trees due to their length, but given their capacity, raod designers may have opted to not leave space for tree growth alon certain stretches. On the other end of the spectrum, small roads such as Eireann Quay, will have less street length for trees, but they may be quiet roads which prioritze natural elements such as trees.

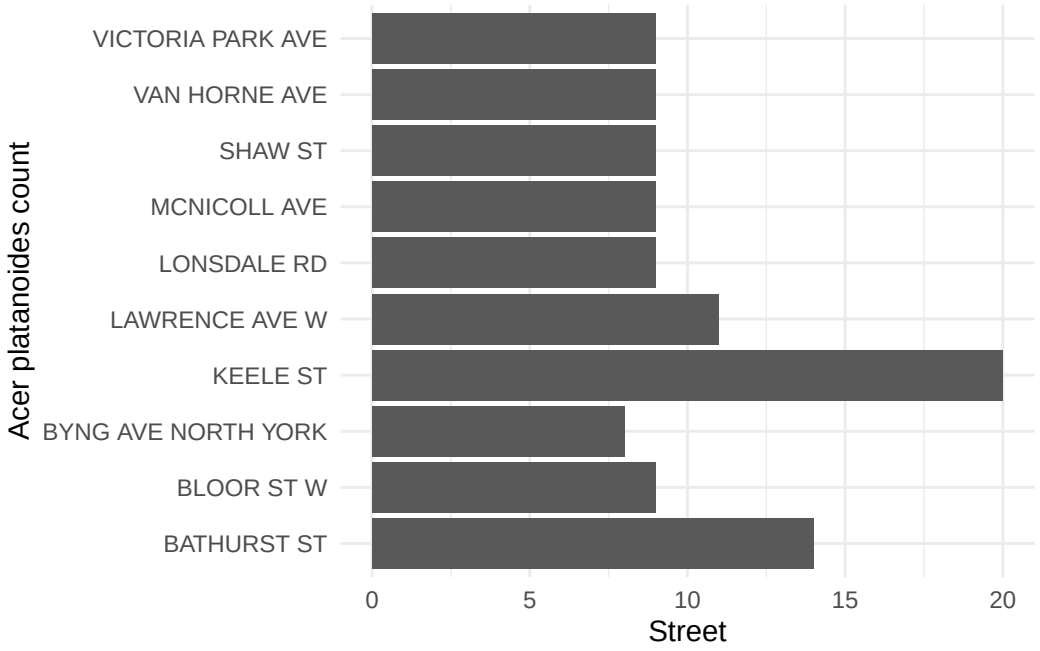

Figure 3: Acer plantiodes distribution across Toronto streets up to September 2024

Figure 3 shows that Keele street holds the most amount of Acer platanoides in Toronto. Keele St is a North-South arterial road that starts at the north-eastern tip of High Park. This suggests that proximity to large parkscapes such as High Park, may have an influence in the spread of Acer platanoides.

2.5 `crossstreet1` & `crossstreet2`

While `streetname` covers the street a tree lies upon, `crossstreet1` & `crossstreet2` tells us what streets (if any) are in close proximity to the tree. This augments `streetname` by narrowing down the general location of the tree to specific intersections. Similarly to `streetname`, small streets intersection with other small streets may may be places where street trees are

planted. Depending on the size and configuration of the intersection, there may be more, or less space for trees

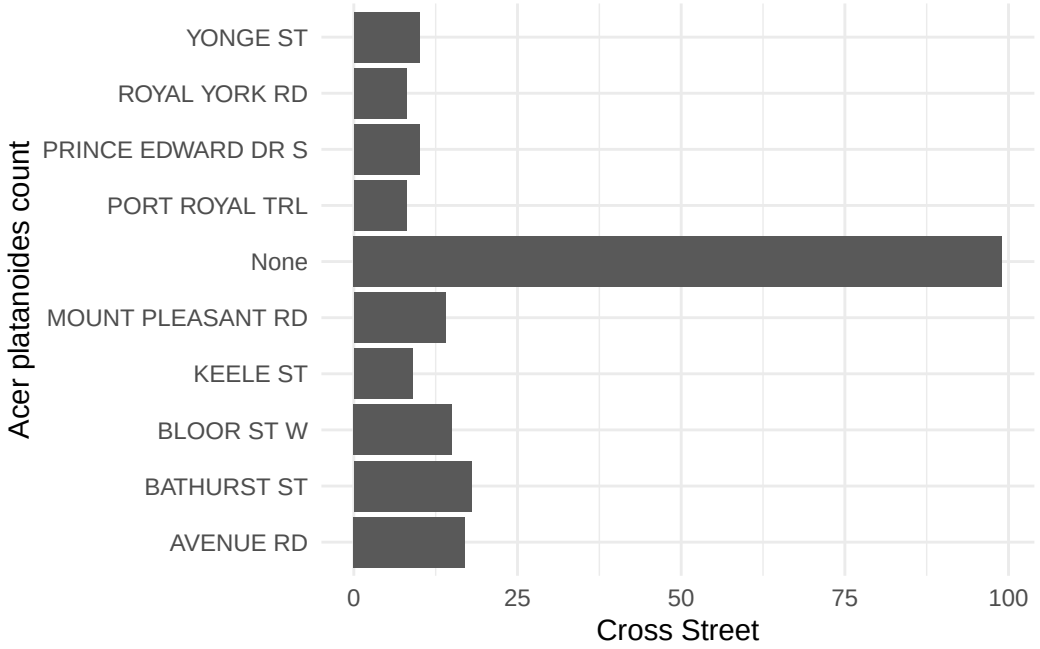

Figure 4: Acer plantiodes distribution across Toronto crossstreets up to September 2024

Comparing Figure 4 & Figure 5, most street trees are placed well clear of intersections. Of the streets that are located in proximity to an Acer platanoides, Bathurst St and Mount Pleasant Rd stick out. Both are prominent in Figure 4 and Figure 5. Augmenting this with Figure 3, we can see that Bathurst St is seen in all three graphs, suggesting than many of Acer platanoides' trees grow along that corridor, and perhaps this is how they spread across the city.

3 Model

Our modeling strategy estimates street trees by botanical name. To achieve this, I created a linear model using location variables as predictors. The goal is to see if Acer plnatoides has a distinct spread by location

The model created uses multiple linear regression, predicting botanical name (`botanical_name`) asa function of 4 location varaibles, `ward` (city ward), `streetname`(tree street is on/nearby), and `crossstreet1` & `crossstreet2` (nearest intersecting streets, if applicable).

Model 1: Linear Model by Date, Transparency Score, and Grade

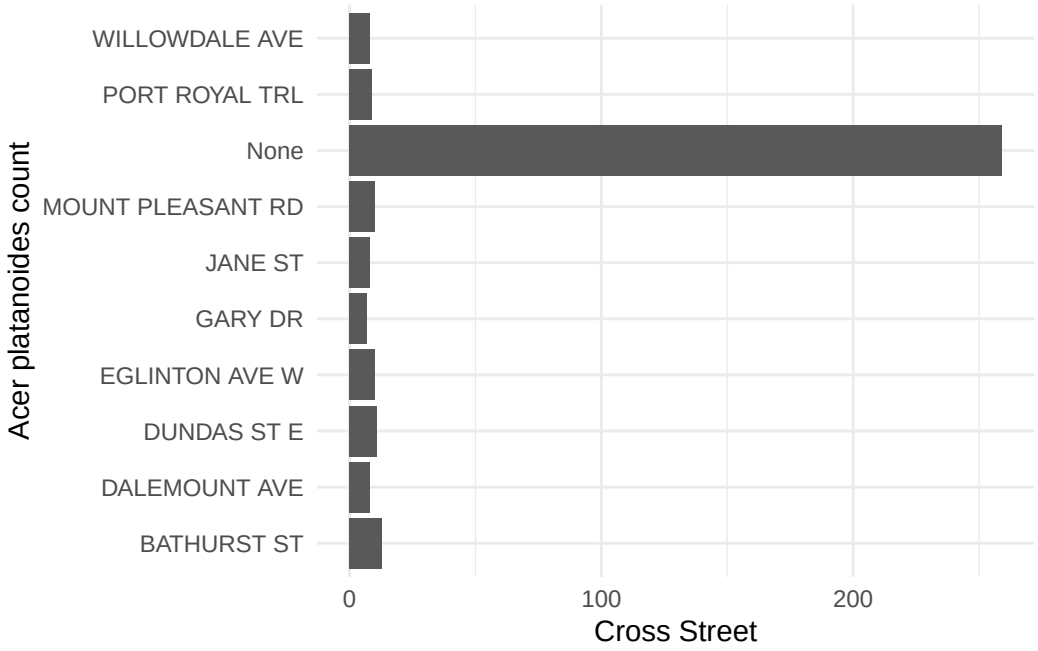

Figure 5: Acer plantiodes distribution across Toronto crossstreets up to September 2024

$$y_i = \beta_0 + \beta_1 \cdot \text{end\_date}_i + \beta_2 \cdot \text{transparency\_score}_i + \beta_3 \cdot \text{numeric\_grade}_i + \epsilon_i \quad (1)$$

$$\epsilon_i \sim \text{Normal}(0, \sigma^2) \quad (2)$$

where:

- y_i represents the botanical name of the tree i.
- $beta_0$ is the intercept.
- $beta_1$, β_2, β_3 , β_4 represent the coefficients for the predictors: `ward`, `streetname`, `crosstreet1`, and `crossstreet2`, respectively.
- $epsilon_i$ is the error term, which is assumed to be normally distributed with variance σ^2.

This model helps capture the general trend tree species across an array of location factors in Toronto.

Model Predictions and Augmentation

The models were then used to predict botanical names, which were added to the original datasets as new columns. The augmented dataset allowed for a comprehensive comparison of predicted versus observed botanical names, and facilitated visualization of how widespread certain species are across Toronto.

3.1 Model set-up

To model the species of street trees, we used multiple linear regression to estimate the relationship between tree species (`botanical_name`) and several lcoation predictors, including the ward, street name, and cross street names. This section provides a detailed breakdown of how we constructed the models, including the selection of predictors and the data processing steps taken to ensure that the models capture meaningful patterns.

3.2 Feature Selection

We used the following predictors to model the polling percentage (pct) for both candidates:

1.City Ward (ward): - The most general location metric in the dataset. We are looking for relationship between species and location, so starting with 25 sections of the city is a good start. The city ward gives a general location for a street tree in Toronto. Given unique characteristics for each of the wards, we can use those to postulate reasons behind a specific species' prominence

2.Nearest Street (streetname): - These next highest location data is the street name. This can help pinpoint specific roads that certain species grow along. Since many roads pass through wards, it can help paint a picture for trends seen in wards data. While wards provide a general idea of location, they come at the cost of precision. Streets can narrow down a location of a tree to a simple line or curve, as opposed to an area. Streets can also cross between wards, which provides potential reasoning for relationships between wards.

3.Nearest Cross Street(s) (crossstreet1 & crossstreet2): - Because streets can span across a ward, or multiple wards, knowing where slong the street a tree is located is helpful. Cross streets define a tree's nearest intersection. Streets are often long lines, which is a problem for precision. Referencing cross-streets for our model can help pinpoint where exactly on a street, a tree is located. This brings the precision down to plots of land of a few square meters.

3.3 Model Training and Prediction

- The model was fitted to predict botanical_name as a function of ward, streetname, crossstreet1 and crossstreet2. This model helps in understanding the general trends in tree species, across the City of Toronto.
- The model was trained using the lm() function in R, which estimates the parameters by minimizing the sum of squared residuals.
- After fitting the model, predictions were made and added to the dataset. This allowed us to compare actual botanical name with those predicted by the models.

3.4 Model Assumptions

The following assumptions underlie the multiple linear regression models used:

- **Linearity:** We assume that the relationship between the predictors (ward, streetname, crossstreet1, crossstreet2) and the outcome (botanical_name) is linear.
- **Independence:** Each poll is treated as an independent observation. We assume that polling percentages from different polls are not influenced by one another.
- **Homoscedasticity:** The variance of the residuals is assumed to be constant across all levels of the independent variables.
- **Normality of Residuals:** The error term (_i) is assumed to be normally distributed, which is required for the validity of hypothesis testing.
- **No Multicollinearity:** We checked that the predictors are not highly correlated, to ensure that the coefficients estimated by the models are reliable and that multicollinearity does not bias the results.

These assumptions are crucial for ensuring that the model coefficients are unbiased and that the predictions are meaningful.

3.5 Cross-Validation and Overfitting Prevention

To evaluate the robustness of our models, a train-test split was applied. The training data was used to fit the model, while the test data (which the model had not seen during training) was used to evaluate model performance. This approach helps in identifying overfitting, ensuring that the model can generalize beyond the specific dataset used for training.

For future improvements, we could explore the use of cross-validation for further robustness checks, ensuring that the model's performance is not overly dependent on a particular train-test split.

3.5.1 Model justification

The choice of multiple linear regression models for predicting the speceis of street trees in Toronto is driven by several key considerations regarding the nature of the data and the objectives of the analysis. Below, we provide a detailed justification for the chosen models and predictors, discuss the reasoning behind the modeling choices, and address limitations and alternatives considered during the process.

3.6 Why Multiple Linear Regression?

Multiple linear regression (MLR) was selected as the modeling framework due to its ability to estimate relationships between multiple predictors and a outcome variable—in this case, the names of tree species (botnaical_name). MLR provides a straightforward approach to quantify the impact of each predictor on polling percentages, allowing us to make inferences about the strength and direction of these relationships.

The primary motivations for using multiple linear regression in this context include:

1.Simplicity and Interpretability: - Linear models offer a high degree of interpretability. Each coefficient in the model provides a clear indication of the expected change in voter support given a one-unit change in the corresponding predictor, holding all other predictors constant. This is crucial for understanding how time, poll quality, and pollster contribute to support dynamics.

2.Ability to Control for Multiple Predictors: - MLR allows us to control for several influential factors simultaneously, such as time (end_date), pollster quality (transparency_score and numeric_grade), and the polling organization (pollster). By accounting for these variables, we can isolate the individual impact of each predictor and better understand the underlying factors that affect polling outcomes.

3.Comparison Between Candidates: - We built identical models for both Kamala Harris and Donald Trump, allowing for a side-by-side comparison of their polling trends over time. This consistency enables a clearer understanding of differences in support and trends across the two candidates.

4 Results

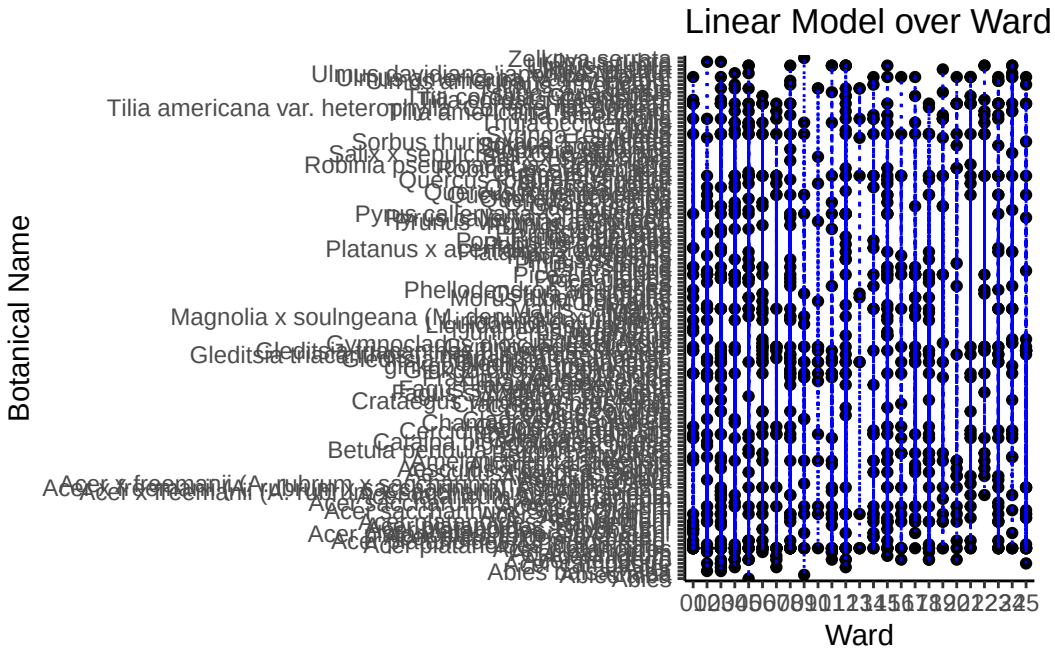

Figure 6: Plot of ward by tree species. The data used is generated from the linear model predicting botanical_name from ward, streetname, crossstreet1, and crossstreet2.

5 Discussion

5.1 Limitations of the Model

1.Assumptions of Linearity and Normality: - The model assumes that the relationship between predictors and the outcome is linear, and that the residuals are normally distributed. In reality, polling data may not always satisfy these assumptions, especially if there are non-linear trends in voter support or heavy-tailed distributions in the errors.

2.Potential for Omitted Variable Bias: - While our models include several important predictors, there may still be unobserved factors influencing voter support that are not captured in the models. Examples include specific campaign events, candidate debates, or sudden shifts in voter sentiment.

3.Multicollinearity: - The predictors transparency_score and numeric_grade both relate to poll quality and may be correlated. This could introduce multicollinearity, which makes it difficult to determine the individual effect of each predictor. However, we inspected the variance inflation factors (VIFs) to ensure that multicollinearity was not excessively high.

4.Non-Independence of Observations: - Polls conducted by the same pollster over time might not be entirely independent, leading to autocorrelation in the data. While Model 2

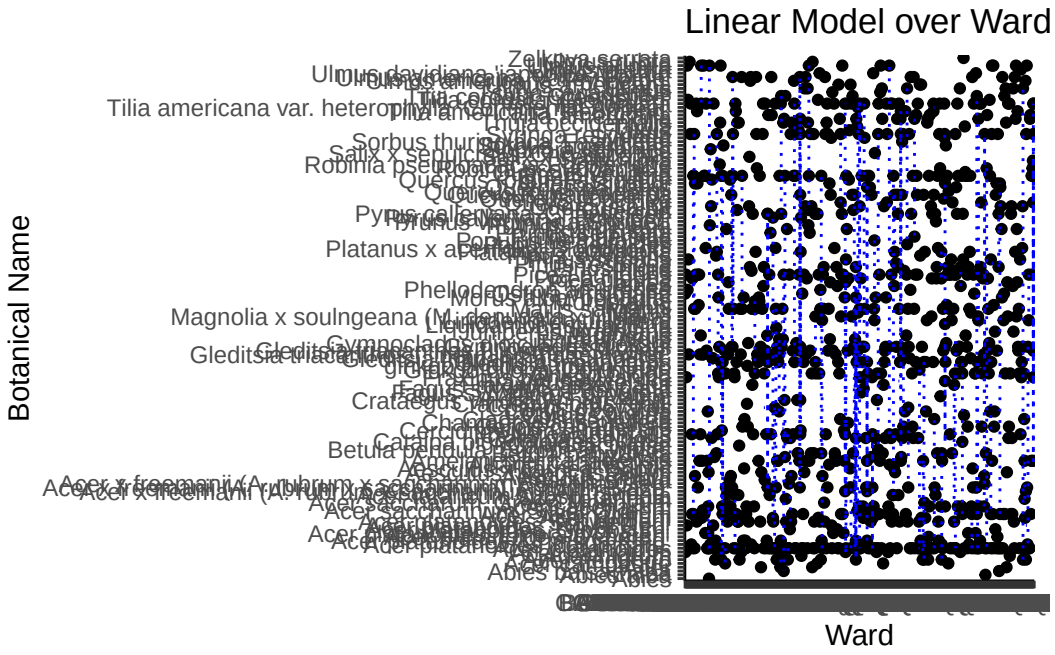

Figure 7: Plot of ward by tree species. The data used is generated from the linear model predicting botanical_name from ward, streetname, crossstreet1, and crossstreet2.

partially addresses this by including pollster as a fixed effect, a more complex model (e.g., hierarchical or Bayesian) could better capture these dependencies.

5.2 Weaknesses and next steps

Street Tree data is cumulative, meaning Urban Forestry has to manually update the status of each tree, which is an arduous task for over 30000 trees. As a result the data may be mismatched, with some trees having only been checked months or even years prior. This may be why Figure 4 is so skewed. There may have been errors in the way the data was inputted for certain trees, leading to misplaced decimals. A similar occurrence happened when observing the raw data, which can be seen in the Appendix. No additional information was provided by OpenData Toronto when viewing the dataset on their database, apart from who provided the data, Urban Forestry. Perhaps more detail on their part, including units of measurement for dbh_trunk, would have helped make this analysis much easier.Another limitation mentioned earlier is the inclusion of some privately-owned trees.

Next steps would be augmenting this data with other data from Toronto's roads, or other trees, to see if any comparisions or correlations can be drawn there. Does vehicle traffic impact the trees lining a certain road, and if so, how long would it take for the effects to be noticeable? Does the distribution of street trees tell us anything about income disparity, fire risk, or quality

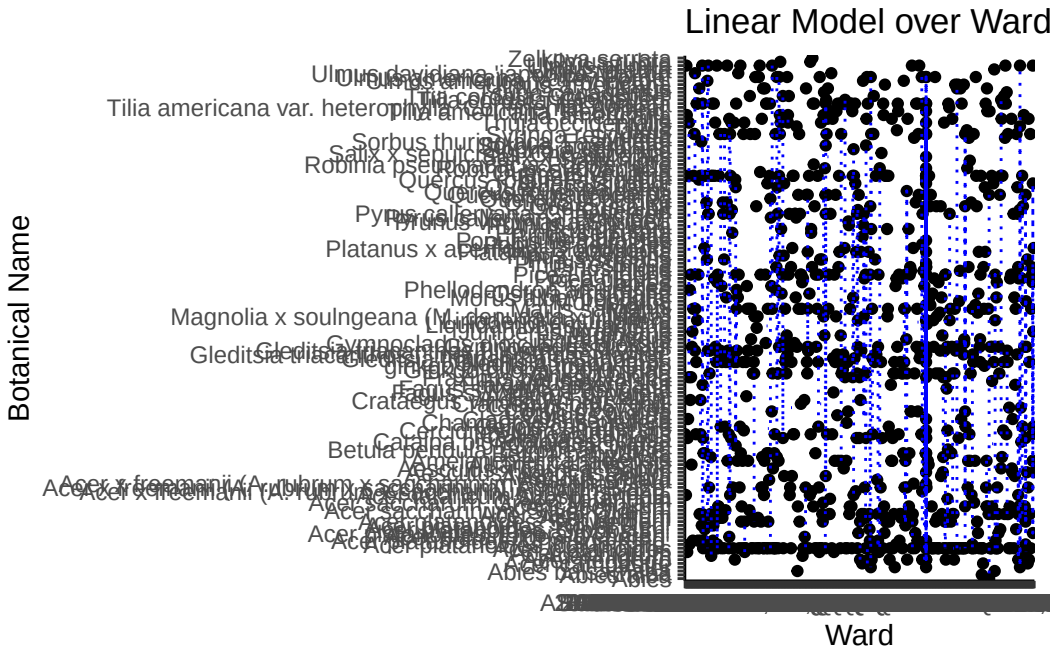

Figure 8: Plot of ward by tree species. The data used is generated from the linear model predicting botanical_name from ward, streetname, crossstreet1, and crossstreet2.

of life within Toronto? This report can be used as a launching pad for further investigations, using key takeways from this report as drivers for further questions.

With newer report, pressure will be put upon Urban Forestry to better curtail the data they have collected thus far. It is imperative that their hard work be recognized and any feedback on the data they have collected be sent to them as well, ensuring better datasets for analysts, leading to more informative conlusions for all.

6 Appendix

6.1 Data Cleaning

While cleaning that dataset, some inconsisten values were discovered. A particular species of tree, Salix x sepulcralis, had it's name misspelled in the dataset. This may have come from using special characters that could not be saved onto the dataset, resulting in a garbled name. An additional step was taken to restore the correct botanical name, using the common name as reference. Similarly, one observation did not have a ward location allocated. This was remedied by reading the globval coordinates from geometry, a varaible in the raw data file that shares the latitude and longtitude of the tree individual. Using these coordinates, and the address number, the general location for this tree was determined, along with its ward location. The ward entry for this observataion was updated accordingly.

References

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

Dobson, Martin, and Jo Ryan. 2000. *Trees and Shrubs for Noise Control*. Arboricultural Advisory; Information Service.

"Donate to Urban Forestry." 2024. City of Toronto.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Gelfand, Sharla. 2022. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://CRAN.R-project.org/package=opendatatoronto.

Krayenhoff, E. Scott, Timothy Jiang, Andreas Christen, Alberto Martilli, Timothy R. Oke, Brian N. Bailey, Negin Nazarian, et al. 2020. "A Multi-Layer Urban Canopy Meteorological Model with Trees (BEP-Tree): Street Tree Impacts on Pedestrian-Level Climate." *Urban Climate* 32: 100590. https://doi.org/10.1016/j.uclim.2020.100590.

Nowak, David J, and Rowan A Rowntree. 1990. "History and Range of Norway Maple." *Arboriculture & Urban Forestry (AUF)* 16 (11): 291–96.

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

Sandberg, L Anders, Adrina Bardekjian, and Sadia Butt. 2014. *Urban Forests, Trees, and Greenspace: A Political Ecology Perspective*. Routledge.

Snyder, Michael. 2006. "What Is DBH?" *Northern Woodlands*.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.