A Visual and Machine Learning Analysis of Municipally-Maintained Trees in the City of Burlington, VT

Isabelle Franke, Caroline Green, Will Guisbond

4/15/2022

1 Introduction

City arborists in Burlington, Vermont have maintained a comprehensive dataset containing various features detailing every tree planted by the city since 1982, including date planted, GPS coordinates, land use type, species, diameter, height, tree condition, and appraisal estimate. These data have important implications in environmental assessments and future sustainability efforts from the city. The creation and curation of the data is part of the Urban Forestry Program by the Department of Parks and Recreation in Burlington, dedicated to expanding and protecting urban forests in the city. Trees, particularly in urban environments, can have massive impacts on the ecosystem. These benefits include increased property value, less drastic temperature fluctuations, reduced erosion, and air pollution control. According to the Burlington Parks and Rec Department, " tree with a 50-year life span provides nearly $60,000 of benefit over its lifetime," not to mention all the other un-quantifiable ways trees add to our lives through recreation, aesthetics, and ecosystem benefits [@OurTreesGreenways].

However, the current state of the data leaves a lot to be desired by those interested in analyzing it. There doesn't appear to be any clear bias, but the dates included in the `modified` column expose a lack of updating to the dataset. The dataset was accessed in early 2022, and the most recent update listed is from December, 2019. Additionally, although the dataset is quite large, it is not comprehensive, and contains vast numbers of empty values or zeros. For example, the `planted` column, meant to inform the user of when the tree was planted, is about 85% zeros.

When assessing what data cleaning steps needed to be taken, we had to first decide whether the large number of zeros were due to a lack of value or a lack of information. Looking closer into the data, the variables containing zeros (`diameter`, `height`, `spread`, `trunks`, `condiition`, `appraise`, `planted`) could not realisticlally be zero and be a living tree. Therefore, we determined that all of the zeros were, in reality, null values and were adjusted accordingly to prevent zeros from skewing downstream analyses. In addition to zeros, we also converted any empty cells to NA as well. Other data cleaning steps, all outlined in our code below, included splitting `Geo.Point`, the coordinates, into longitude and latitude to increase ease of use. We also split up the `species` column into Genus and Species, to allow for better grouping for visualizations (though it should be noted that all genera and species are listed in common names, not scientific). The column also posed another problem: a lack of a common naming format. FOr example, two Maple species were named "mapl,red" and "mapl,sugar" but another one was named "maple free aut bl", meaning that some fine tuning of the spelling and separation of the species names was necessary. We also updated the date format for the `modified` column.

The observational data curated by the city arborists is first and foremost a reference database for the city to keep track of the number and condition of trees in their care. For this to be a valuable asset, however, it requires that the city maintain and keep the information up-to-date. On the other hand, it can also be used by citizens to learn more about the trees on their street or in their neighborhood park. It provides anyone with a computer access to an abundance of information on the trees of Burlington and possesses a lot of potential to better understand the urban landscape of the city. This research project will analyze

the Burlington tree data using visualizations and machine learning algorithms to better understand the abundance, distribution, and makeup of the urban forest in the city of Burlington, Vermont.

2 Visualization & Analysis

2.1 Data Cleaning

```
# read in csv file
trees <- read.csv("./Burlington_Trees.csv")
library("magrittr")

# separate Geo.Point column to latitude and longitude
# and convert to numeric variables
trees <- trees %>%
         separate(Geo.Point, c("lat", "long"), ",") %>%
         mutate(lat = as.numeric(lat),
                long = as.numeric(long))

# separate species column into genus, species
trees <- trees %>%
         separate(species, c("genus", "species"), ",")

# remove the "spp" included as a place holder for genus
trees$genus <- gsub(" spp", "", trees$genus)

# correct the misspelling of mapl to maple
trees$genus <- gsub("mapl", "maple", trees$genus, fixed = TRUE)
trees$genus <- gsub("maplee", "maple", trees$genus)

# convert zeros in numeric columns to NA so they will not be included in graphs
# in this case, zero values are due to lack of information, not lack of value,
# so all were converted to NA values to be filtered out later
# repeat for blank values
trees[trees == 0] <- NA
trees[trees == ""] <- NA

# convert dates to better format
trees <- trees %>%
  mutate(modified = as.yearmon(modified, "%m/%Y"))

# cleaned data frame
head(trees)
```

```
##        lat      long zone_id site_id modified park    landuse          site_typ
## 1 44.44855 -73.22863  ward 5    2933     <NA> <NA>  residence              Tree
## 2 44.44970 -73.22843  ward 5    2936     <NA> <NA>  residence              Tree
## 3 44.44975 -73.22844  ward 5    2937     <NA> <NA>  residence              Tree
## 4 44.45999 -73.22088  ward 5      29     <NA> <NA> apartments Planting Stumpite
## 5 44.45991 -73.22088  ward 5      30     <NA> <NA> apartments Planting Stumpite
## 6 44.45978 -73.22089  ward 5      31     <NA> <NA> apartments              Tree
##        genus    species diameter height spread trunks conditn appraise planted
## 1 arborvitae       <NA>        2      5      5      1      80      240      NA
## 2     linden littleleaf       16     50     30      1      70     5300      NA
## 3     linden littleleaf       13     45     25      1      80     4070      NA
```

```
## 4    unknown       <NA>       NA     NA     NA     NA      NA       NA      NA
## 5    unknown       <NA>       NA     NA     NA     NA      NA       NA      NA
## 6      maple        red       41     45     35      1      70    27800      NA
```

2.2 Number of Trees by Genus

```
# number of trees by species
# histogram
```

2.3 Relationship between Genera and Land Use Type

The Burlington tree dataset includes all trees maintained by the city, which includes areas such as business parks, parking lots, schools, cemeteries, and so on. This led us to the question, are certain genera of trees better suited or more favored by certain land use types? To visualize this, created a frequency table of tree genera by land use, and filtered for those with a frequency greater than 0.1 - to create a more readable visualization. We then plotted this information on a multiple bar chart, showing the proportion of each land use type that was made up by different genera.

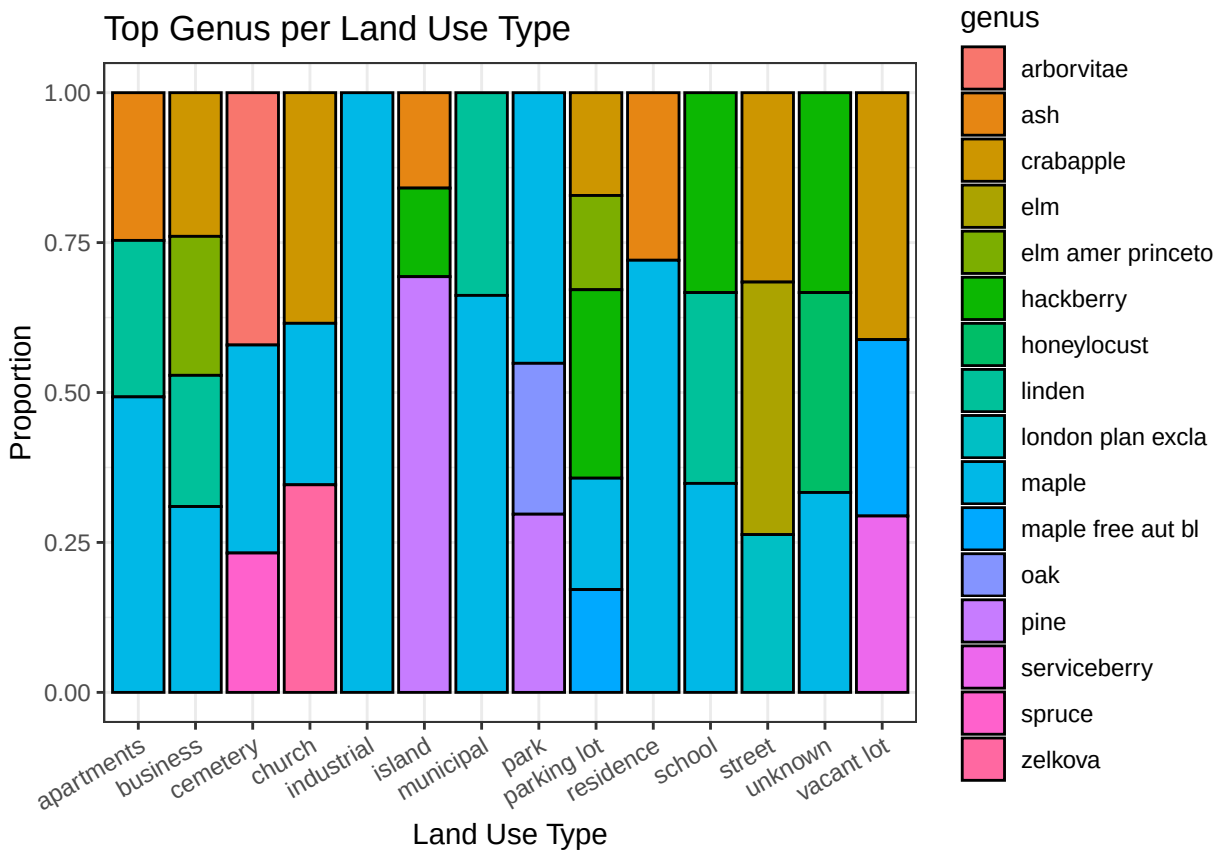
Note: I want the proportions on the left to be how much of the city is that type of land use, but I haven't figured out how to do that yet.

```
# landuse v species
# multiple bar chart
# see if certain types of trees are more common by business, residential, etc.

# format the frequency table into a new data frame
landuse_by_genus <- xtabs(formula = ~ landuse + genus,
                          data = trees) %>%
  prop.table(margin = "landuse") %>%
  data.frame() %>%
  filter(Freq > .1)

head(landuse_by_genus)
```

```
##      landuse      genus      Freq
## 1   cemetery arborvitae 0.2431333
## 2 apartments        ash 0.1107474
## 3     island        ash 0.1138211
## 4  residence        ash 0.1145349
## 5   business  crabapple 0.1250000
## 6     church  crabapple 0.1538462
```

```
# make bar graph with frequency table
landuse_species_bar <- ggplot(data = landuse_by_genus,
       mapping = aes(x = landuse,
                     fill = genus,
                     y = Freq)) +
  geom_bar(color = "black",
           stat = "identity",
           position = "fill") +
  labs(y = "Proportion",
       title = "Top Genus per Land Use Type",
       x = "Land Use Type") +
  theme(axis.text.x = element_text(angle = 30,
                                   hjust = 1))
```

```
landuse_species_bar
```



2.4 The Trees of Burlington

```
# will
# map of burlington with points as trees and colored by land use type
# map of burlington with points as trees and colored by species
# facet wrap them next to each other!
```

2.5 Appraisal Estimates

```
# diameter v appraisal
# scatterplot probably
```

3 Machine Learning