Analysis of Tree Health in Melbourne, Australia

Michael Chen
maichen@ucdavis.edu

Douglas Turner
dcturner@ucdavis.edu

March 15, 2018

Abstract

In this paper, we seek to analyze the effects of tree age, size, location, placement, and other covariates on their useful life expectancy, a measure of health. Data is taken from official records kept by the City of Melbourne on its publicly-maintained trees. Though the original data includes information on the precinct for each tree, we instead use k-means clustering to divide the trees into clusters based on their geographical location. Using regression diagnostics, we find that using twelve clusters via the aforementioned method provides a better fit than using the original twelve precincts. Our final model indicates that tree size, location, placement, and age are important factors in determining useful life expectancy.

1 Introduction

For about 20 years, the City of Melbourne has kept detailed data on almost 70,000 trees. This is done to monitor and ensure the health of Melbourne's urban forest. Maintaining a healthy population of trees within a city is important for carbon reduction, safety, property values, deterring climate change, and citizen health [2].

In addition to carbon dioxide, trees absorb other harmful chemicals such as sulfur dioxide, carbon monoxide, nitrogen oxides, cadmium, nickel, and lead [2]. This makes the air cleaner and the residents healthier. Trees provide valuable oxygen and pollution control, especially when they are located in typically polluted and smog-ridden regions such as large metropolitan centers.

According to recent research, the presence of trees and green space can actually deter crime [1]. The authors found that people living in greener spaces reported lower levels of fear and aggressive/violent behavior. Trees have also been shown to make residents better drivers, since the presence of trees alongside the street can give drivers a better perception of how fast they are driving, which can cause them to slow down [2]. It is also hypothesized that the presence of trees alongside a road causes drivers to relax, decreasing incidents of road rage [2].

Research shows that trees increase the value of adjacent properties [8]. The benefit for property owners is large; a study published in the Journal of Arboriculture showed that good tree cover increases residential property value by almost 6% [8]. Another study found similar results for residential homes in Athens, Georgia [10]. The article estimates that this increase in sale price leads to an increase in property tax revenue for the city, a result that all residents, regardless of their home ownership status, benefit from.

Melbourne (the city at the center of this analysis) has been severely affected by climate change. The yearly state of the climate report, produced by the Australian federal government, shows that Australia's mean temperature has increased steadily since 1910 [6]. With the increase in temperature, the rate of extreme heat events and fires has increased as well. Trees can help mitigate climate change through their ability to absorb carbon dioxide [2], provide shade, and insulate buildings from hot or cold air. Melbourne suffers from a warming effect known as the urban heat island problem, which among other causes, results from insufficient tree cover over dark surfaces that absorb more sunlight [11].

As mentioned above, trees clean air and improve breathability. A USDA Forest Service and Davey Institute study showed that the presence of trees has saved 850 lives and helped with almost 700,000 incidences of acute respiratory symptoms [9]. Benefits extend past increased air quality. A recent study [4] in Toronto revealed that people who lived closer to more vegetated space had better mental health. Other benefits include increased biodiversity and positive effects to runoff and sewer systems [7].

Melbourne's urban forest has faced considerable challenges. Of Australia's capital cities, Melbourne has the least canopy cover [12]. Additionally, many of Melbourne's trees were planted in surges of planting activity and are now aging and need increased maintenance or replacement. Another problem is biodiversity, as Melbourne's urban forest is dominated by only a few genera of trees. This lack of biodiversity makes the urban tree population vulnerable to diseases and pests [13].

Considering these beneficial effects of a healthy and large tree population and the specific challenges of the city, it is in Melbourne's best interest to design an urban forest that maximizes tree quality and survival. Melbourne maintains an Urban Forest Strategy, a plan to maintain and enhance the population of trees in the city. The goals of the strategy include mitigating climate change, designing for livability, creating healthier ecosystems, and positioning Melbourne as a leader in Urban Forestry [7]. Instrumental in achieving these goals are increasing tree health and diversity.

In conjunction with this initiative, detailed data has been collected on many of the trees in Melbourne's urban forest. This data is obtained from Melbourne's official government website `https://data.melbourne.vic.gov.au` and contains 69,275 initial data points. It provides detailed descriptions of tree species, location, health, maturity level, and size.

Important categorical variables include the scientific name of the tree (which includes genus and species), age description (describes the age of the tree as new, juvenile, semi-mature, mature or over-mature), precinct (describes the governmental precinct that the tree is located in), and placement (whether each tree is located by the street or in a park).

Important quantitative variables are useful life expectancy (an arborist's estimate of how long a tree is likely to remain in the landscape based on health, amenity, and risk), diameter breast height (a measure of tree diameter at 1.4 meters above ground, in centimeters), and the latitude and longitude of each tree. A full list of variables and descriptions can be found in Table 1.

This data needs to be explored and analyzed to answer questions such as: Where do Melbourne's trees grow best? How does tree size affect health outcomes? Are trees in the park healthier than trees by the street? How does maturity level relate to health? Answers to these questions and others that result from this analysis would be of interest to city planners, arborists, and biologists. An accurate model of tree health in urban areas would be of great use in creating and maintaining an urban forest that is beneficial for residents in the numerous ways stated above.

2 Methods and Results

2.1 Exploratory Data Analysis

Missing Data The original dataset has 69,275 observations. However, after dropping missing cases, 24,742 observations remain. The authors do not find any indication that there is any pattern in this missing data.

Location: Tree location is described by four variables: `Latitude`, `Longitude`, `Precinct` and `Located.in` (values of `Street` or `Park`). Figure 2 shows the location of almost 25,000 trees colored by whether they are in a park or by the street. It is clear that most trees are by the street in Melbourne. There are many advantages to street-adjacent trees, especially with regards to the urban heat island problem, as discussed in Section 1. While Longitude and Latitude are informative, it would be better to have some way of grouping trees together by location. It is more likely that trees would differ from group to group rather than across somewhat arbitrary geographical directions. Figure 3 shows the location of trees colored by their government precinct (with the exception of the level `Water`, which indicates the tree is over water). While this variable may be useful in determining which government precinct is effectively maintaining the urban forest, it does not describe location particularly well because government lines are often drawn for political reasons rather than geography. To look for a more natural way to group trees together, we utilize a k-means clustering approach to search for a more logical way to group trees together. A plot of tree location/clusters with a cluster size of 6 (just as an example) is shown in Figure 5.

Species, Genus, and Family: There are 111 levels of the variable `Genus`, 49 levels of `Family`, and 323 levels of `Species` in the dataset. However, a few families and genera dominate the dataset. This can be seen in the pie charts in Figure 6 and 7. In Figure 6, the Myrtaceae, Platanaceae, and Ulmaceae families make up the majority of trees. In Figure 7, the Eucalyptus, Corymbia, Platanus, and Ulmus genera make up the majority of trees.

Diameter Breast Height: Diameter Breast Height is a measure of the width of a tree. Although standards differ internationally, in this dataset it is measured from 1.4 meters above ground. The units are in centimeters. A histogram of this variable can be found in Figure 8. Note that if a tree has not reached the height of 1.4 meters, then its Diameter Breast Height is set to zero. It is clear that bigger trees are less common in the city than smaller trees.

This variable is standardized in the following analysis. This is done so that we have an indication of whether a tree is wide for its species. Certain species are bigger than others; we want to investigate whether an bigger tree leads to a healthier tree after accounting for species. This variable is standardized by subtracting the sample mean and dividing by the sample variance for its species. Since some species only have one observation, those observations have been dropped due to having a sample variance of zero.

Useful Life Expectancy Value: Useful Life Expectancy Value is an arborists assessment of the time a tree can remain in the landscape based on health, amenity, and risk [13]. The arborist's assessment is subjective and takes on the values of 1, 5, 10, 20, 30, 60, and 80 years. A useful life expectancy value of more than 20 years indicates a healthy tree, while less than 5 years indicates a dying tree. However, this variable must be considered in the context of the species and other variables, since not all trees have the same lifespan. It is important to note that the city of Melbourne recommends this variable to be used as a health indicator, not as a timeline for removal [13].

We choose to standardize this variable with respect to tree species. This means that for each value of Useful Life Expectancy, we subtract the sample mean and divide by the sample standard deviation within that species. To see why this necessary, consider the London Plane Tree (Platanus acerifolia). This tree is very popular as a street tree due to it's ability to provide shade. It is also a tree that prospers in city environments, especially hot ones [14] and has become quite popular in Melbourne (4,274 occurrences in the initial dataset). Due to its prosperity, a high value of Useful Life Expectancy for the London Plane tree may not be high for another tree. By standardizing the response, we can consider the effects of the independent variables on Useful Life Expectancy regardless of average species health/longevity. A histogram of standardized Useful life expectancy is shown in Figure 9. This variable appears to be approximately normally distributed and there is no indication of a necessary transformation at this point.

Challenges: The first significant challenge was how to account for species. We expect different species to have different diameters and health levels. There are 323 species in the dataset, and controlling on 322 dummy variables would lead to multicollinearity, high-dimensional problems, and a lack of interpretability. We resolve this problem by standardizing the Diameter Breast Height and Useful Life Expectancy variables with respect to species, as illustrated above.

The next challenge is how to quantify location. As previously discussed, there are two potential approaches: using the `Precinct` variable or using the results of a k-means clustering.

2.2 Models

In these models, the response is the standardized useful life expectancy: for each observation, its useful life expectancy minus the sample average useful life expectancy for its species divided by the sample standard deviation for that species. All models are fit with least squares estimation.

Precinct Model: In this initial model, tree location was controlled for using the `Precinct` variable. Initial linear model fits included `Age.Description`, `Standardized Diameter Breast Height`, `Located.in`, and `Precinct` as X variables. Note that `Precinct`, `Age.Description`, and `Located.in` are categorical variables (factors) with 12, 5, and 2 levels, respectively. `Age.Description`, `Standardized Diameter Breast Height`, and `Located.in` are included in this model as well as subsequent models since they are the variables of interest. We wish to examine their significance, sign, and magnitude.

Residuals from initial fits show a heavy-tailed distribution from the normal Q-Q plot. The box-cox plot for this model is displayed in Figure 10. Based on this plot, we choose the following transformation:

$$f(StanULE) := sign(StanULE) * |StanULE|^{3/4}$$

Note that this transformation maintains the sign of the response after transformation. The Q-Q plot is shown in Figure 12. This Q-Q plot of the transformed data, coupled with the normality of the histogram of standardized useful life expectancy (Figure 9) indicates that the transformed response is normally distributed or can be approximated well by a normal distribution.

This model can be summarized by the following equation:

$$\begin{aligned} f(StanULE_i) = \beta_0 + \beta_1 StanBreastHeight_i &+ Located.in_i \alpha \\ &+ Precinct_i \gamma \\ &+ AgeDescription_i \zeta \\ &+ \varepsilon_i \qquad\qquad i = 1, \ldots, n \end{aligned}$$

In this equation, `Located.in`, `Precinct` and `Age.Description` are row vectors of length $n_l - 1$ containing indicator variables for the levels of the corresponding factor, where n_l is the number of levels for the factor. α , γ and ζ are parameter vectors.

A residuals v.s. fitted values plot is shown in Figure 11. This plot indicates homoscedastic errors. Although lower fitted values may appear to have less variance than larger fitted values, there is considerably less data with fitted values of less than -0.5, making the plot appear slightly heteroskedastic. The Cook's distance plot (Figure 13) indicates that a few observations (particularly 63488) may be outliers and/or influential cases. Upon examining these observations, there is nothing to indicate error in entering the data and all the values are consistent with expected ranges of the variables (no negative diameters or life expectancies). Hence, there is no reason to believe these observations should be dropped.

Clustering Model In this Model, we attempt to use an alternative (more geographically motivated) clustering of the trees in Melbourne. A k-means clustering algorithm was used for cluster size values ranging from 3 to 15. The input of this algorithm is the latitude and longitude variables for each tree. Thus, each tree is sent to a cluster based on its location, where the number of clusters is determined by the cluster size. This clustering creates a new factor variable to be included in the regression. A model was fit for each cluster size. The dependent variable was transformed standardized Useful Life Expectancy (as in Precinct Model) and the X variables were `Age.Description`, `Standardized Diameter Breast Height`, `Located.in`, and the cluster factor variable. As cluster size increases, the number of predictors (p) increases. Too large of a cluster size would overfit the data, while too small of a cluster size may cause bias in the results by not capturing important geographical differences. This indicates that a balance between bias and variance needs to be achieved; we will examine this using model selection criteria. Model selection statistics are shown in Figure 14. SSE, R^2 (follows from SSE), R_a^2, AIC, and BIC all choose the model with a cluster size of 12. This is a strong indication that 12 is the optimal cluster size. It is interesting that the model selection happens to choose the same number of clusters as precinct (minus the Water factor level). Thus, the k-means clustering method is choosing different (more relevant) ways to cluster the data than the precinct designation, but with the same number of clusters. The last line in Figure 14 is the model fit using the `Precinct` variable instead of clusters.

The chosen model (Cluster Model) has a dependent variable of transformed standardized Useful Life Expectancy and X variables of `Age.Description`, `Standardized Diameter Breast Height`, `Located.in`, and the cluster factor with a cluster size of 12.

This model can be summarized by the following equation:

$$\begin{aligned} f(StanULE_i) = \beta_0 + \beta_1 StanBreastHeight_i &+ Located.in_i \alpha \\ &+ Cluster_i \gamma \\ &+ Age.Description_i \zeta \\ &+ \varepsilon_i \qquad\qquad i = 1, \ldots, n \end{aligned}$$

As above, `Located.in`, `Cluster`, and `Age.Description` are row vectors of indicator variables indicating the level of the corresponding factor. `Cluster` is a factor of 12 levels with each level corresponding to a cluster in the k-means clustering algorithm. α, γ, and ζ are parameter vectors.

A Q-Q plot is shown in Figure 16. The plot shows a good agreement with normal or approximately normal errors. Additionally, the fitted values vs. residuals plot (Figure 15) does not indicate strong heteroskedasticity. As discussed previously, there is less data at low levels of standardized ULE. Lastly, the Cooks Distance plot is shown in Figure 17; this plot shows a few cases that may be influential outliers. After examining these observations, there is no indication that they should be removed.

With a 75/25 split on the total data, this model has $MSE = 0.6928041$ on the training dataset. Using the validation dataset, the mean squared prediction error is $MSPE = 0.8755101$. While this may be an indication of some small bias, the difference is small and indicates that the model is useful.

2.3 Interpretation

As seen from the final model equation above, our final model takes into account tree size (diameter breast height), location, placement (whether the tree is by a street or in a park), and age description to predict standardized useful life expectancy, which is a measure of the health of the tree. The regression results for the model are tabulated in Figure 1.

	stanule34		
	B	*CI*	*p*
(Intercept)	0.44	0.39 – 0.50	<.001
Age.Description			
Mature	-0.63	-0.67 – -0.59	<.001
New	0.10	0.05 – 0.16	<.001
Over-mature	-1.13	-1.23 – -1.03	<.001
Semi-Mature	-0.51	-0.54 – -0.48	<.001
standbh	-0.14	-0.15 – -0.12	<.001
Located.in (Street)	-0.07	-0.09 – -0.05	<.001
factor.cluster.			
2	-0.28	-0.34 – -0.22	<.001
3	0.14	0.08 – 0.20	<.001
4	0.01	-0.05 – 0.06	.820
5	0.08	0.03 – 0.14	.004
6	0.13	0.08 – 0.19	<.001
7	0.11	0.06 – 0.17	<.001
8	-0.02	-0.08 – 0.03	.416
9	0.18	0.13 – 0.24	<.001
10	-0.02	-0.08 – 0.04	.479
11	0.13	0.06 – 0.21	<.001
12	-0.03	-0.09 – 0.03	.275
Observations	24592		
R^2 / adj. R^2	.161 / .160		

Figure 1: Fitted regression coefficients and p-values for the final model.

As we can see, all the individual regression coefficients are significant in the presence of the other predictors at the $\alpha = 0.05$ level with the exception of the indicator variables for clusters 4, 8, 10, and 12. However, the cluster factor variable as a whole is highly significant (p-value of less than 2×10^{-16}) using the extra sum of squares F-test. The majority of the predictors are highly significant, which indicates that this set of predictors is helpful for explaining the variation in useful life expectancy.

The reference level for `Age.Description` is ”juvenile”. From the results of this analysis, it appears that newly planted trees have slightly better expected useful life expectancy the juvenile trees, while older trees have progressively worse expected useful life expectancy. This is an intuitive result.

The coefficient for standardized `Diameter Breast Height` is negative, meaning that after accounting for the other variables (including age), larger trees have somewhat worse expected useful life expectancy than smaller trees. This is an interesting result; it may be of interest to arborists in Melbourne to figure out the reasoning behind this phenomenon. Perhaps larger trees need more intensive and specialized care, which the city is not properly administering.

Being located next to a street rather than in a park has a negative effect on a tree's useful life expectancy, which is again quite intuitive. Trees near the road may be subject to more wear and tear while also being planted in small lots rather than the open fields of a park. Furthermore, it is possible that trees in parks receive better care and thus are in better condition. With this information, Melbourne city planners may want to devote some attention to the upkeep of trees by the road. As mentioned in the introduction section of this paper, roadside trees have positive effects on drivers and landscape, so their maintenance is important.

The reference level for `Cluster` is cluster 1. It may be worthwhile for city planners to examine whether there are any real geographical reasons for some clusters being worse for tree health than other clusters or whether it is simply because of poorer upkeep in those clusters. If the issue is the latter, city officials may want to commit more resources to tree upkeep in the locations represented by these clusters as they are behind the rest of the city in terms of tree health.

3 Conclusion

Trees are essential to the health and prosperity of an urban area. Based on our analysis, there is a lot that Melbourne city planners and government officials can do to help the urban tree population of their city thrive. Firstly, locating trees in parks or paying more attention to trees located near the street can improve overall tree health. Secondly, attention should be paid to which areas of Melbourne have struggling tree populations. As indicated by the significance of coefficients on cluster factor levels, there is significant variation in tree health by location in the city. While our analysis shows this to be true for Melbourne, this may be true for other cities as well. Officials should examine their jurisdiction policies regarding tree upkeep, maintenance, and planting.

Questions left for later analysis include: Why do some locations lead to better tree health? Do certain types of trees grow better in specific locations? What species thrive alongside streets (a difficult but important area for tree growth)?

The benefits of a vibrant urban forest cannot be understated. Maintaining a healthy population of trees within a city is important for carbon reduction, safety, property values, deterring climate change, and citizen health [2]. The Melbourne government has already made a strong effort towards improving the urban forest with their ”Urban Forest Strategy”. Now, further action and analysis is necessary to begin targeted and calculated improvements to the city’s tree population.

References

[1] Kuo, E.F. and Sullivan W.C. 2001. Environment and Crime in the Inner City: Does Vegetation Reduce Crime? *Environment and Behavior*. Vol 33, Issue 3, pp. 343 - 367.

[2] Benefits of Urban Trees Guide. A guide by Greenblue Urban. 2016. http://www.greenblue.com/wp-content/uploads/2016/05/Benefits-of-Urban-Trees.pdf

[3] Trees of Strength Project, NC State Department of Horticultural Science https://projects.ncsu.edu/project/treesofstrength/benefits.htm

[4] Kardan, Omid and Gozdyra, Peter and Misic, Bratislav and Moola, Faisal and Palmer, Lyle J. and Paus, Tom and Berman, Marc G. 2015. Neighborhood greenspace and health in a large urban center. *Scientific Reports*. 5, Article number: 11610

[5] Center for Public Impact. Responding to climate change: Melbournes urban forest strategy. 2016. https://www.centreforpublicimpact.org/case-study/urban-forest/

[6] State of the Climate Report. 2016. Bureau of Meteorology and CSIRO. https://www.csiro.au/en/Showcase/state-of-the-climate

[7] URBAN FOREST STRATEGY: Making a great city greener 2012-2032, 2012, City of Melbourne. https://www.melbourne.vic.gov.au/SiteCollectionDocuments/urban-forest-strategy.pdf

[8] Morales, D.J. 1980. The contribution of trees to residential property value. *Journal of Aboriculture* 6 [10]: 305-308.

[9] Nowak, D.J. Hirabayashi, S. Bodine, A. and Greenfield, E. 2014. Tree and forest effects on air quality and human health in the United States. *Environmental Pollution*. 193. 119-129.

[10] Anderson, L.M.; Cordell, H.K. 1988. Influence of trees on residential property values in Athens, Georgia: a survey based on actual sales prices. Landscape and Urban Planning. 1988, 15: 1-2,153-164.

[11] Yang, L. Qian, F. Song, D. and Zheng, K. 2016 Procedia Engineering Research on Urban Heat-Island Effect. *Procedia Engineering*. 169. 11-18.

[12] Benchmarking Australias Urban Tree Canopy: An i-Tree Assessment, Final report. 2014. Institute for Sustainable Futures, and University of Technology, Sydney. http://202020vision.com.au/files/where_are_all_the_trees.pdf

[13] URBAN FOREST STRATEGY: Making a great city greener 2012-2032, 2012, City of Melbourne http://melbourneurbanforestvisual.com.au/#mapexplore

[14] Plane Tree. Encyclopedia Britannica. March 19, 2014. Access Date: November 28, 2017.

[15] Mount J and Zumel N (2016). vtreat: A Statistically Sound data.frame Processor/Conditioner. R package version 0.5.27, URL https://CRAN.R-project.org/package=vtreat.

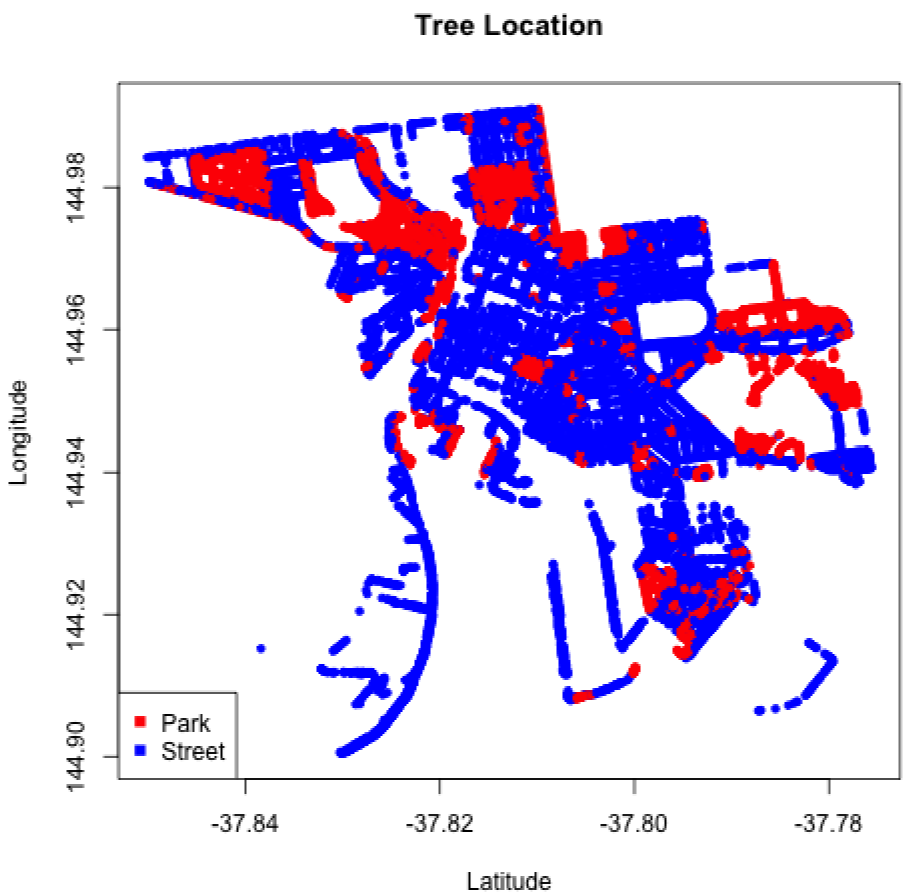

Figure 2: Plot of Trees colored by their location (Street or Park)

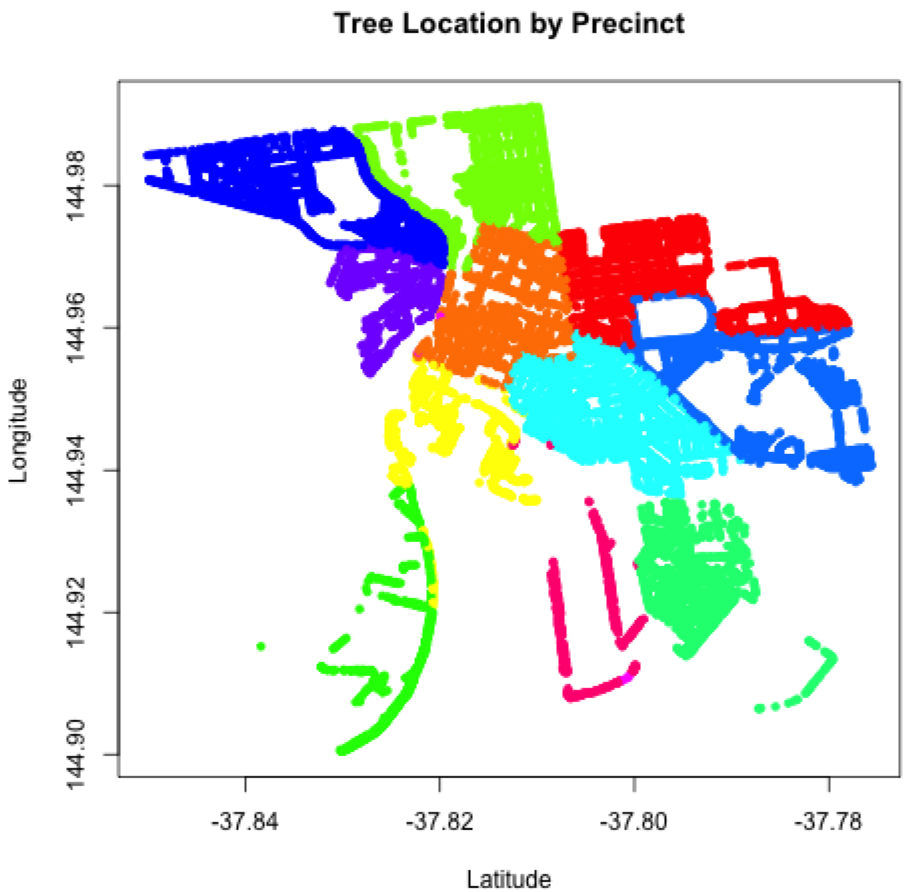

Figure 3: Plot of Trees colored by their Precinct

Figure 4: Plot of Trees colored by their Precinct legend

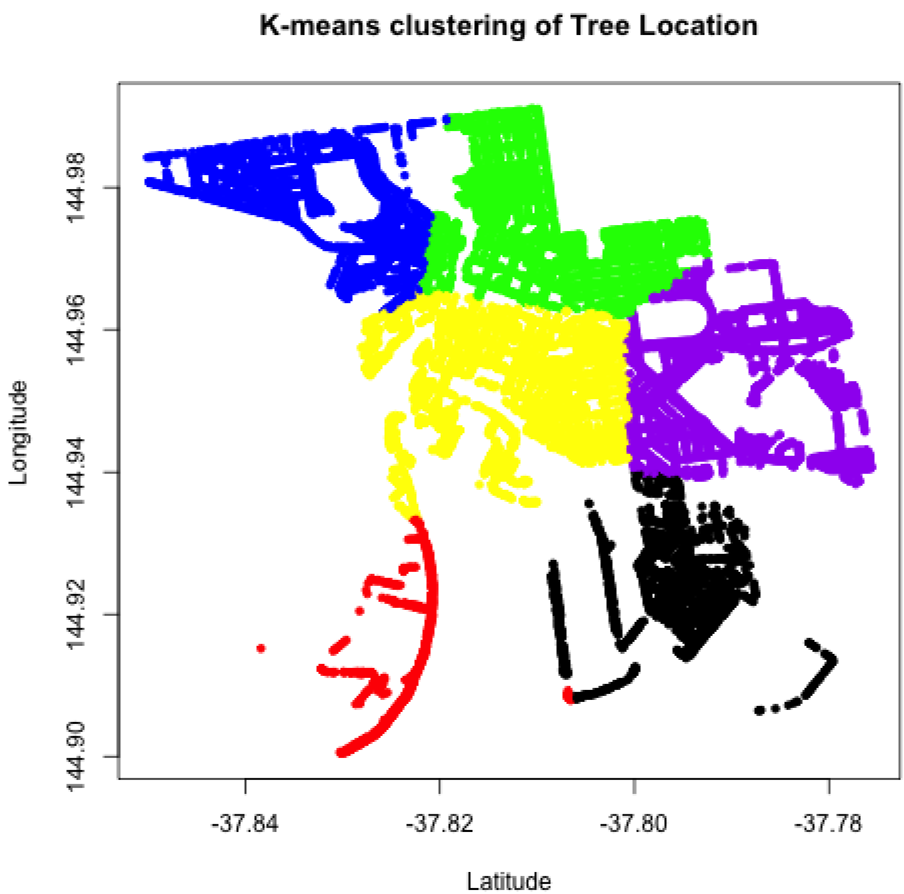

Figure 5: Plot of Trees colored by their cluster

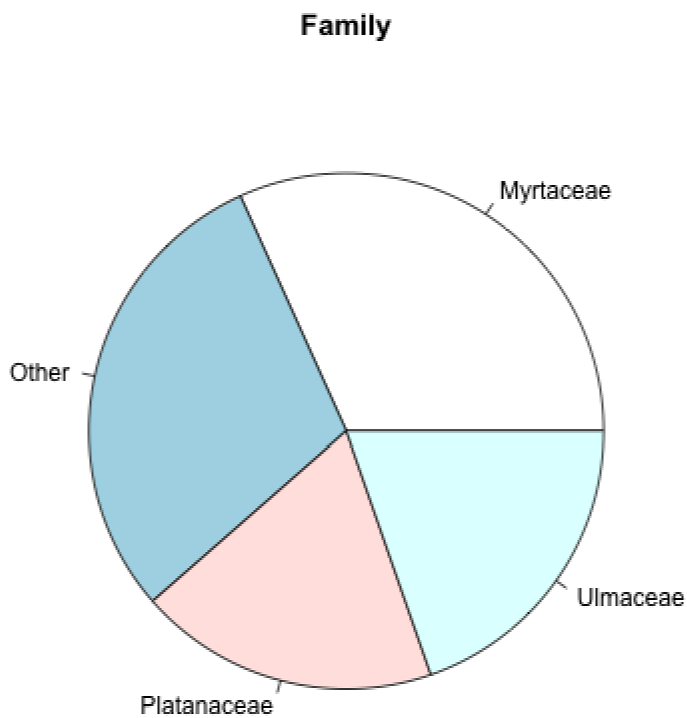

Figure 6: Pie chart for Taxonomic Family

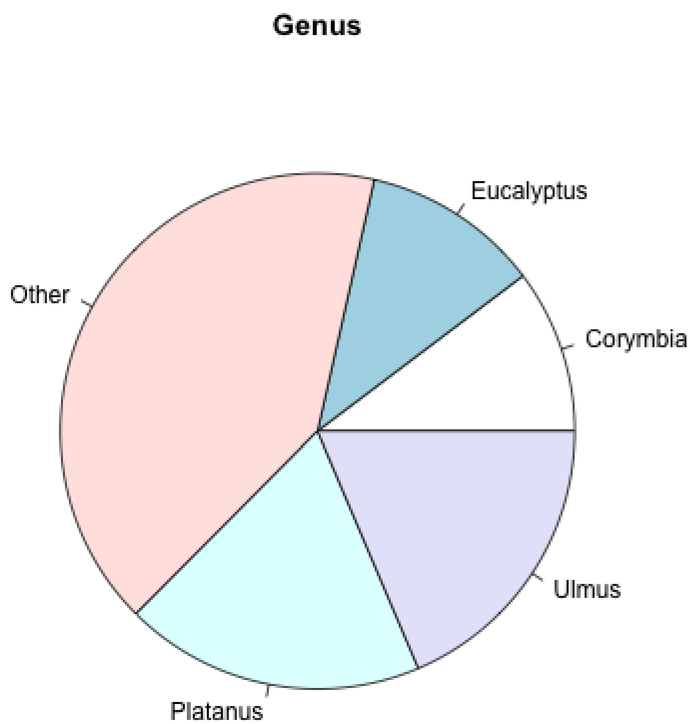

Figure 7: Pie chart for Taxonomic Genus

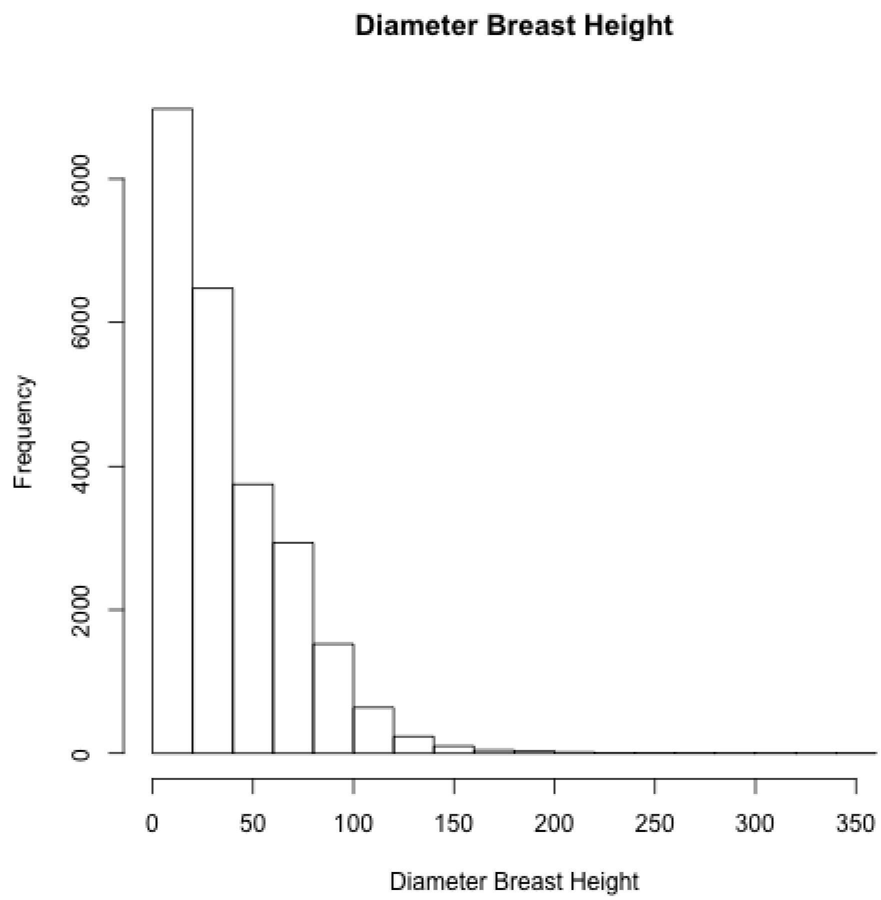

Figure 8: Diameter at Breast Height Histogram

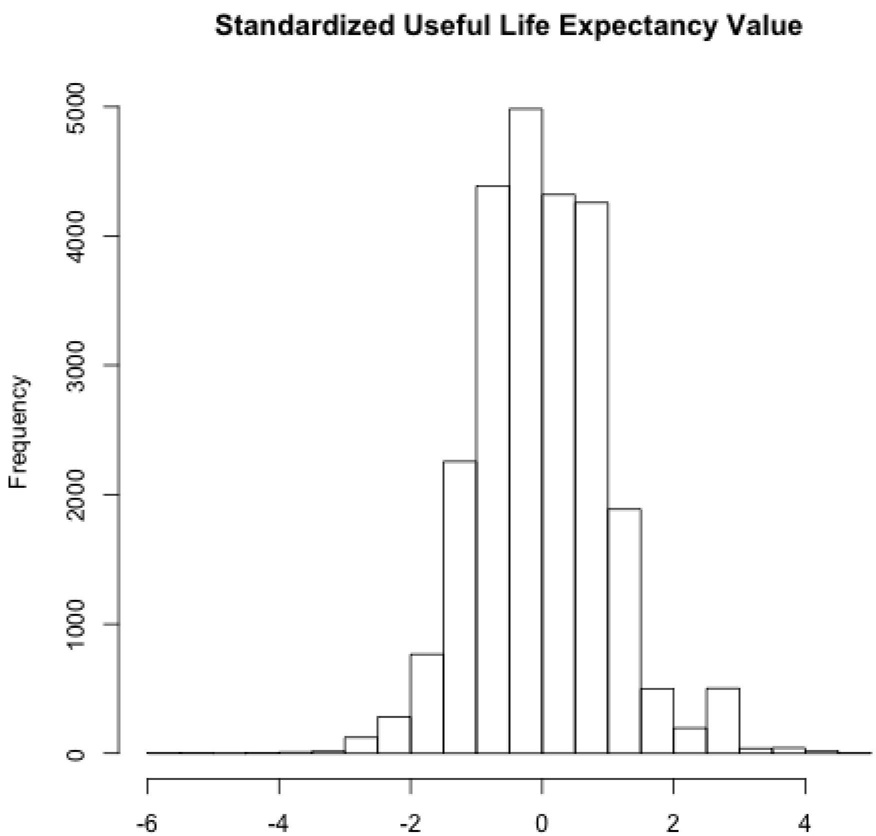

Figure 9: Standardized Useful Life Expectancy

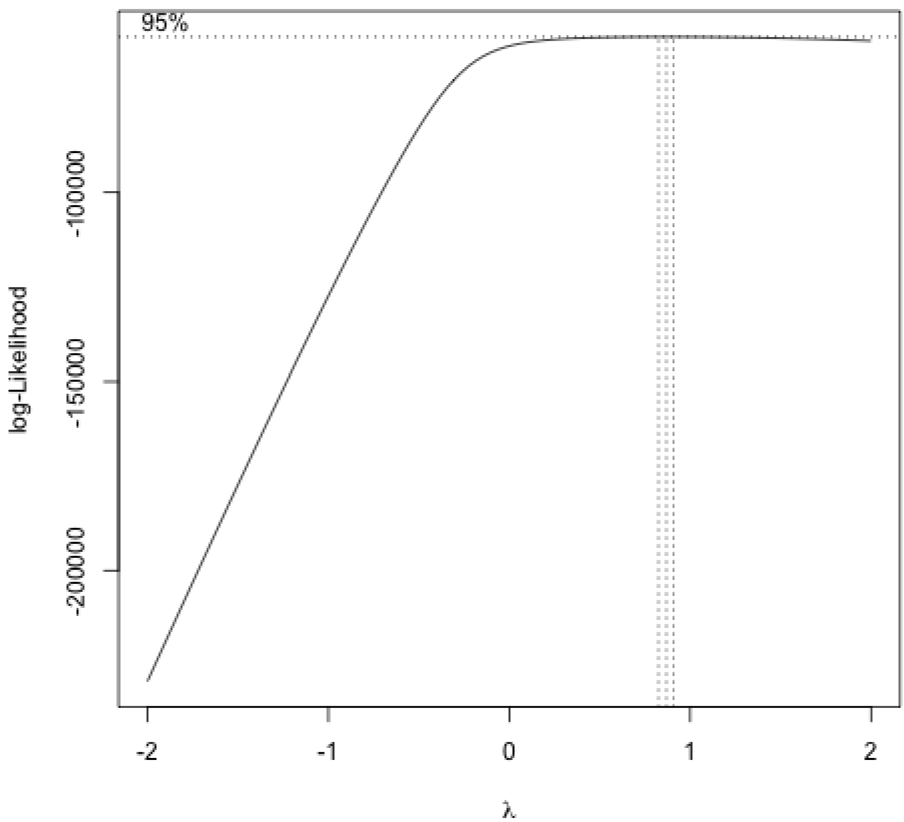

Figure 10: Box-cox Plot for Precinct Model (Standardized Useful Life Expectancy is the response)

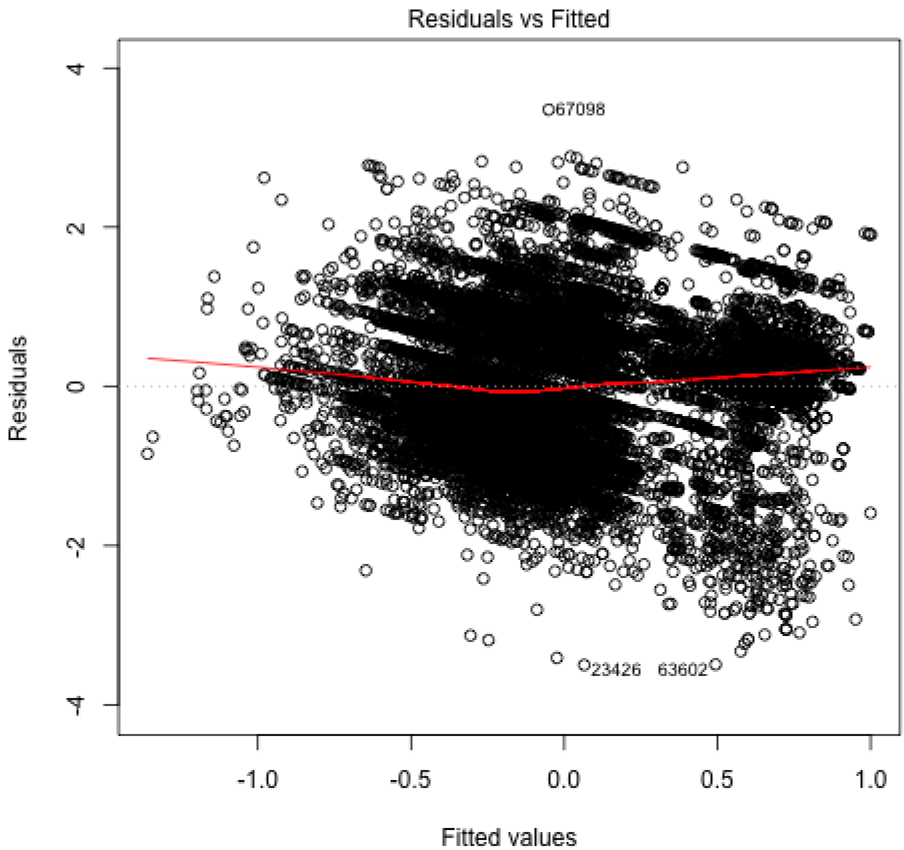

Figure 11: Fitted Values vs Residuals for Precinct Model (Standardized Useful Life Expectancy is the response)

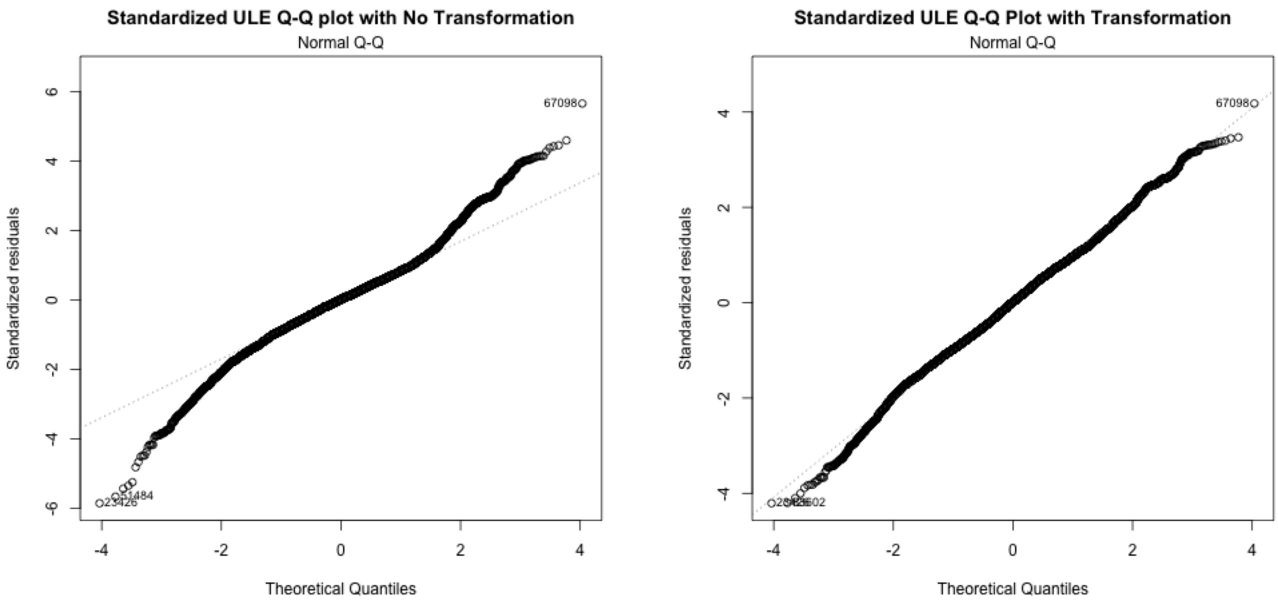

Figure 12: Q-Q Plot for Precinct Model. The left plot is without transformation. The right plot is with transformation. (Standardized Useful Life Expectancy is the response)

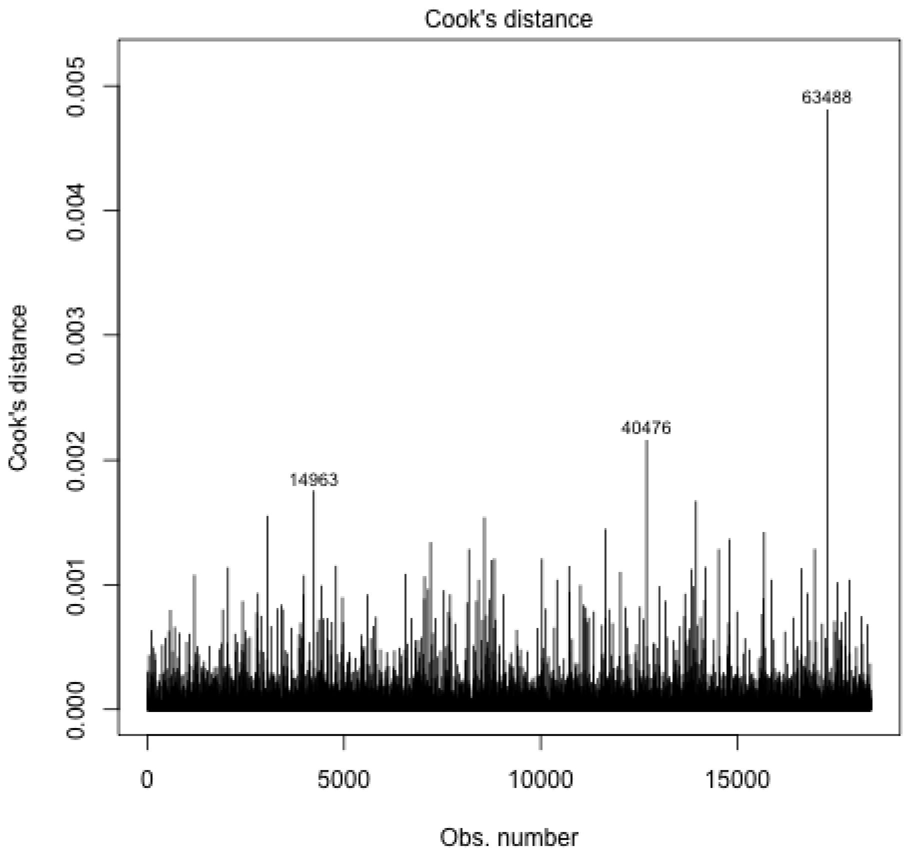

Figure 13: Cooks distance Plot for Precinct Model (Standardized Useful Life Expectancy is the response)

	clustersize	p	SSE	Rsq	R2a	aic	bic
[1,]	”3”	”9”	”12975.3441558592”	”0.150223762530735”	”0.149854133893297”	”−6410.46401624735”	”−6340.08257363784”
[2,]	”4”	”10”	”12968.8757226326”	”0.150647390666034”	”0.150231743149096”	”−6417.63951991068”	”−6339.43791701122”
[3,]	”5”	”11”	”12953.1287043029”	”0.151678688320019”	”0.151217393425141”	”−6437.99588673756”	”−6351.97412354815”
[4,]	”6”	”12”	”12913.0848600582”	”0.154301224330422”	”0.153795341110434”	”−6492.96961926873”	”−6399.12769578937”
[5,]	”7”	”13”	”12918.9226339896”	”0.153918899090561”	”0.153366746969019”	”−6482.65273720169”	”−6380.99065343238”
[6,]	”8”	”14”	”12898.7297915396”	”0.15524136094397”	”0.154644098622345”	”−6509.43680815944”	”−6399.95456410019”
[7,]	”9”	”15”	”12903.5865790131”	”0.154923282091049”	”0.154279799329669”	”−6500.50954271004”	”−6383.20713836085”
[8,]	”10”	”16”	”12803.5306156636”	”0.161476108671223”	”0.160791971691624”	”−6641.74918056785”	”−6516.62661592871”
[9,]	”11”	”17”	”12858.3877486558”	”0.157883426464691”	”0.157150513868054”	”−6561.07792712109”	”−6428.135202192”
[10,]	”12”	”18”	”12735.8179945138*”	”0.165910717552399*”	”0.165139378935111*”	”−6735.32276244066*”	”−6594.55987722162*”
[11,]	”13”	”19”	”12787.1696405022”	”0.162547615349351”	”0.161727566229358”	”−6659.27791107937”	”−6510.69486557039”
[12,]	”14”	”20”	”12770.2778906505”	”0.16365388331888”	”0.162789372344671”	”−6681.60155559837”	”−6525.19834979945”
[13,]	”15”	”21”	”12770.2358940055”	”0.163656633746968”	”0.162746575677052”	”−6679.66206967198”	”−6515.4387035831”
[14,]	NA	”18”	”12736.5942243438”	”0.165859881007611”	”0.16508849537834”	”−6734.20128207334”	”−6593.43839685431”

Figure 14: Model Selection for Clustersize

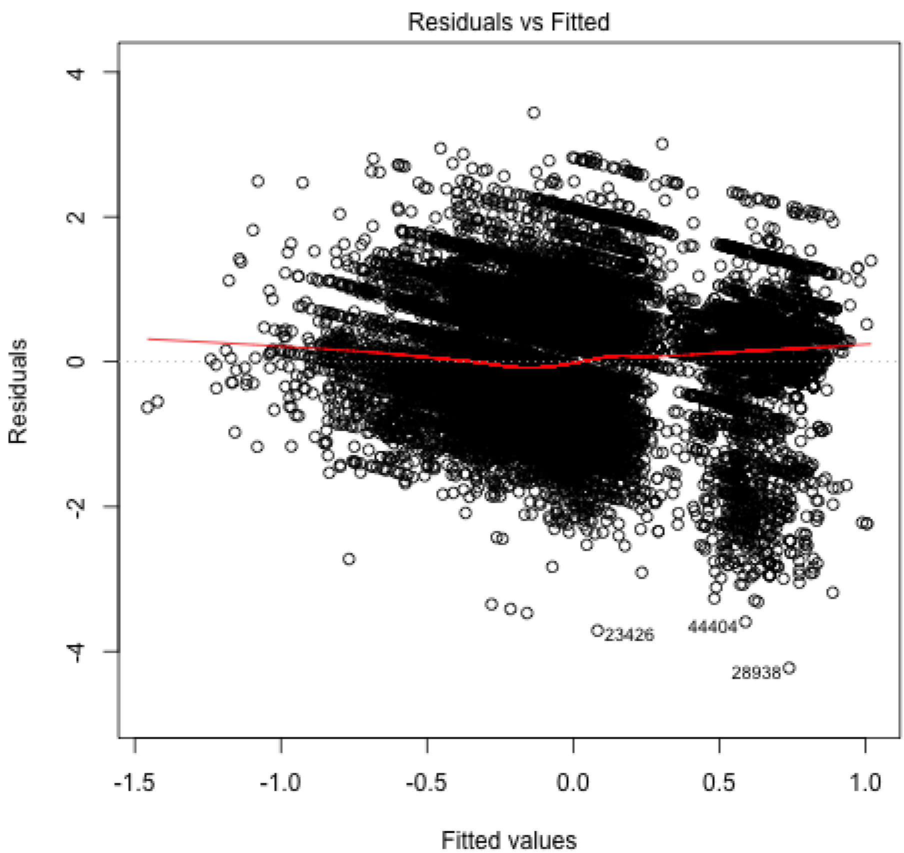

Figure 15: Fitted Values vs Residuals for Cluster Model (Standardized Useful Life Expectancy is the response)

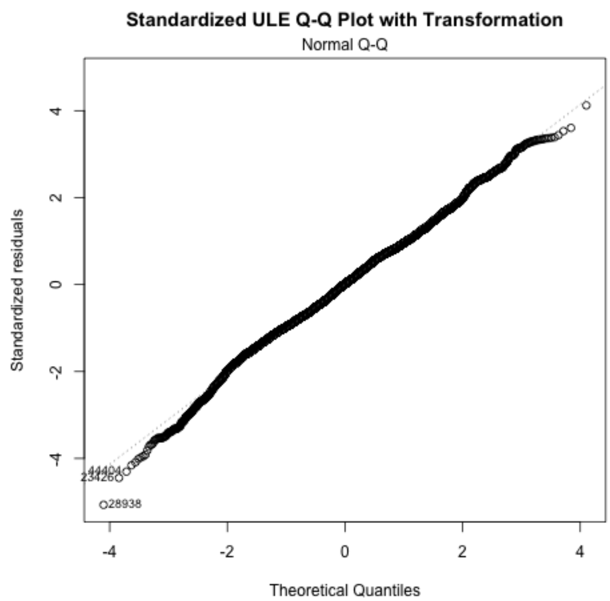

Figure 16: Q-Q Plot for Cluster Model after 3/4 power transformation on the response.

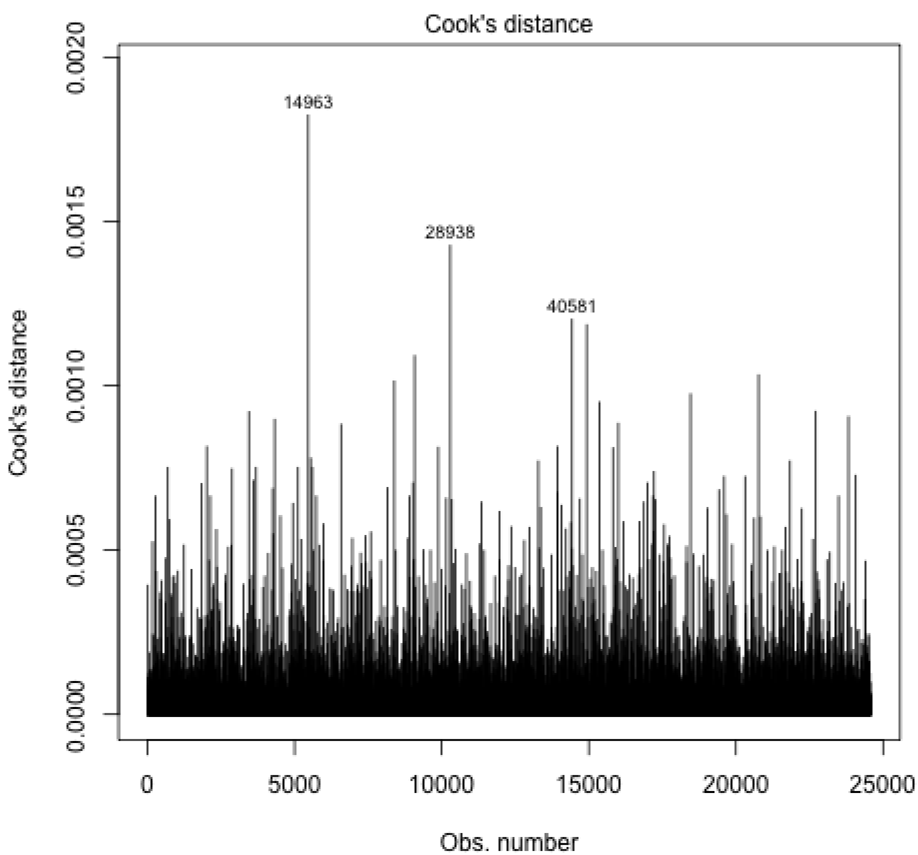

Figure 17: Cooks distance Plot for Cluster Model (Standardized Useful Life Expectancy is the response)

Table 1: Variables and Descriptions

- `CoM ID`: The ID of each tree assigned by the City of Melbourne. (categorical)
- `Common Name`: The common name of each tree. (categorical)
- `Scientific Name`: The scientific name of each tree. (categorical)
- `Genus`: The scientific genus that each tree belongs to. (categorical)
- `Family`: The scientific family that each tree belongs to. (categorical)
- `Diameter Breast Height`: A standard measure of the diameter of a tree's trunk. Measured at 1.4m above ground level. (quantitative)
- `Date Planted`: The date each tree was planted. Accurate from 2003 onwards, otherwise it is the date that the tree was added to the database. (quantitative)
- `Age Description`: Describes the age of the tree as new, juvenile, semi-mature, mature, or over-mature. (categorical)
- `Useful Life Expectancy`: The useful life expectancy of each tree. A health indicator. (quantitative)
- `Precinct`: The geographical precinct that each tree belongs to. (categorical)
- `Located in`: Whether the tree is located in park or along a street. (categorical)
- `Latitude`: The latitude of each tree. (quantitative)
- `Longitude`: The longitude of each tree. (quantitative)