The Distribution of Large Trees in Mystic Vale

Magdalen Thot

2023-10-16

1 Introduction

1.1 Background

The main objective of this study is to collect observations of the locations of large trees and conduct multiple point pattern analyses to determine whether the spatial distribution of these trees significantly deviates from randomness. More specifically, our goal is to determine if there is a discernible pattern in the tree locations. Point pattern analysis is important because it can reveal previously unknown underlying relationships about the phenomena under investigation and enables the detection and quantification of spatial patterns (Boots et al., 1988).

These patterns can manifest as random, clustered, or dispersed arrangements, offering valuable insights about the study area. For example, if an area exhibits a clustering of very large trees in a specific region, it may suggest that the soil quality in that area is particularly conducive to tree growth. In conclusion, point pattern analysis provides a straightforward method for assessing the degree of spatial association and autocorrelation among trees in the Mystic Vale region (Getis et al., 1992).

2 Analysis

2.1 Study Site

The study site for the report is Mystic Vale, a forested ravine located southeast of the University of Victoria campus, outside of the ring road. Its canopy contains trees such as Douglas-fir, grand fir, western red cedar, conifers, bigleaf maple, black cottonwood, willow trees, and more. The study site covers 11.6 acres and is regularly used by students and members of surrounding communities for recreational activities. The study site contains a portion of a parking lot, a large, forested area, and a portion of a grassy field.

To collect the data, measuring tapes were used by groups of students and wrapped around the base of trees to determine if their circumference exceeded 250cm. If a tree met the criteria, its latitude, longitude and ID number were recorded. Then the data was transferred to a csv file and imported into R. One limitation of this data collection method was that there was not enough time to measure every tree. As a solution, we used visual observation to estimate whether a tree had a circumference greater than 250cm. This estimation method can introduce errors into the data because human sight isn't always accurate and can result in including a tree that did not meet the criteria. Another limitation is that not every tree was safely accessible, so one would need to estimate its location or exclude it from the dataset.

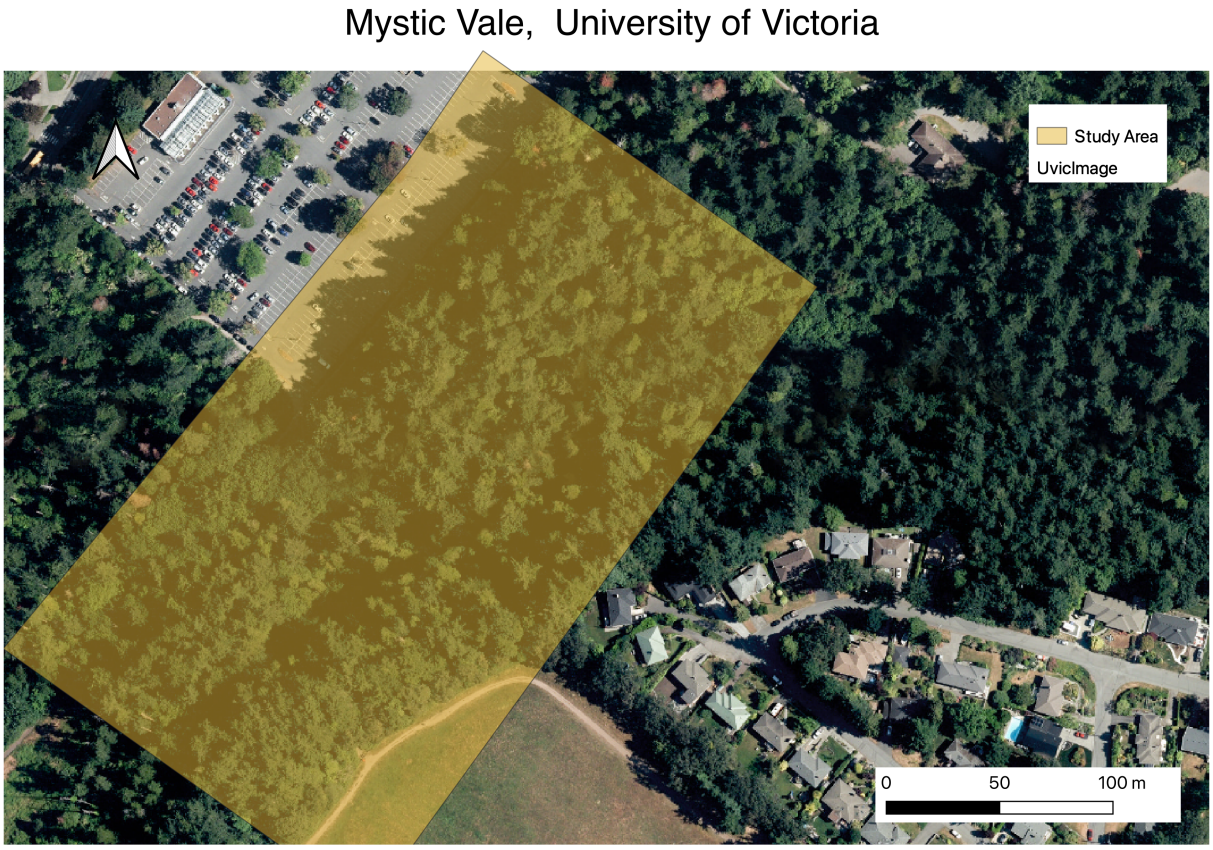

Figure 1: This figure shows a map of the study site.

2.2 Kernel Density Estimation (KDE)

2.2.1 KDE Methods

For the KDE methods, the initial step in the analysis involved visualizing the distribution of large trees in Mystic Vale. This was achieved by creating a KDE surface in R. The primary purpose of employing cross-validation in KDE is to determine the optimal bandwidth that results in the most accurate density estimate. Cross-validation aids in finding the right balance between over-smoothing (bandwidth that is too large) and under-smoothing (bandwidth that is too samll) the density estimate. Various values of the sigma parameter (the bandwidth) are modified and tested to determine which one is the most optimal. Subsequently, the results were visualized using the *spplot* function.

```
# This code tests two bandwidths 500 and 250 meters
kde.500 <- density(shpPPP, sigma = 500, at = "pixels", eps = c(30, 30))
kde.SG <- as(kde.500, "SpatialGridDataFrame")

kde.250 <- density(shpPPP, sigma = 250, at = "pixels", eps = c(30, 30))
kde.SG <- cbind(kde.SG, as(kde.250, "SpatialGridDataFrame"))
```

After conducting some tests with the code above, you may want to determine the optimal bandwidth for the kernel. You can obtain the ideal bandwidth using the *bw.diggle* function from the *spatstat* library (Bw.diggle: Cross validated bandwidth selection for kernel density). This function selects an appropriate bandwidth (sigma) for the kernel using the code below.

```r
# Optimal bandwidth code
bw.d1 <- bw.diggle(shpPPP)
```

After selecting the appropriate bandwidth, it's crucial to choose the right cell resolution. Using an excessively large cell resolution can lead to oversimplification, resulting in the loss of important features. On the other hand, a very small resolution can lead to unnecessary processing time and space consumption, capturing insignificant small fluctuations in the data. Multiple resolutions were tested using a for loop and the *gridExtra* library.

```r
# Vector of resolution sizes to test

resolutions <- c(100, 13, 50, 250)

# empty list to store the  objects
plots <- list()

for (i in resolutions) {
  kde.cell <- density(shpPPP, sigma = bw.d1, at = "pixels", eps = c(i, i))
  kde.cell <- as.data.frame(as(as(kde.cell, "SpatialGridDataFrame"), "SpatialPixelsDataFrame"))
  colnames(kde.cell) <- c("Density", "X", "Y")

  cell <- ggplot() +
    geom_tile(data = kde.cell, aes(x = X, y = Y, fill = Density), alpha = 0.8) +
    ggtitle(paste(i, "m Resolution")) +
    scale_fill_viridis() +
    coord_equal() +
    theme_map() +
    theme(legend.position = "right") +
    theme(legend.key.height = unit(0.5, "cm"),
          plot.title = element_text(hjust = 0.5))
  plots[[as.character(i)]] <- cell
}

grid.arrange(grobs = plots, ncol = 2, nrow = 2)
```

Once you have the optimal bandwidth and cell size, you can combine them into a final KDE map. Where, *bw.d1* as seen above represents the optimal bandwidth, and *eps = c(30, 30)* denotes the cell size.

2.2.2 KDE Results

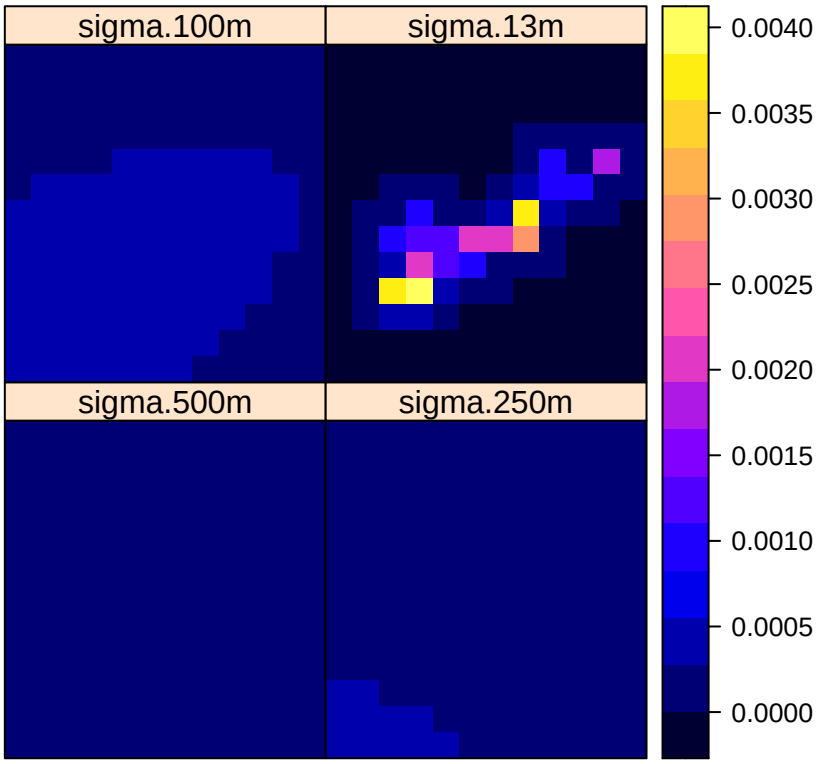

Figure 2: This figure displays different KDE bandwidths and their effects on the surface. As the sigma value increases, you gradually lose detail..

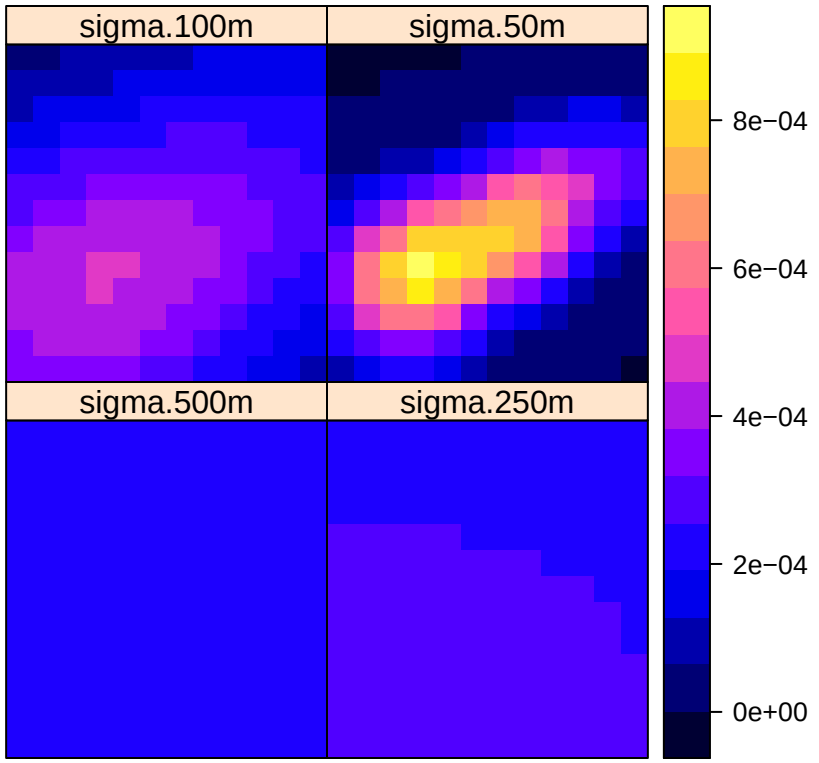

Figure 3: This figure also illustrates various KDE bandwidths and their impact on the surface. However, the smallest value presented here is 50 meters. As the sigma value increases, you progressively lose detail.

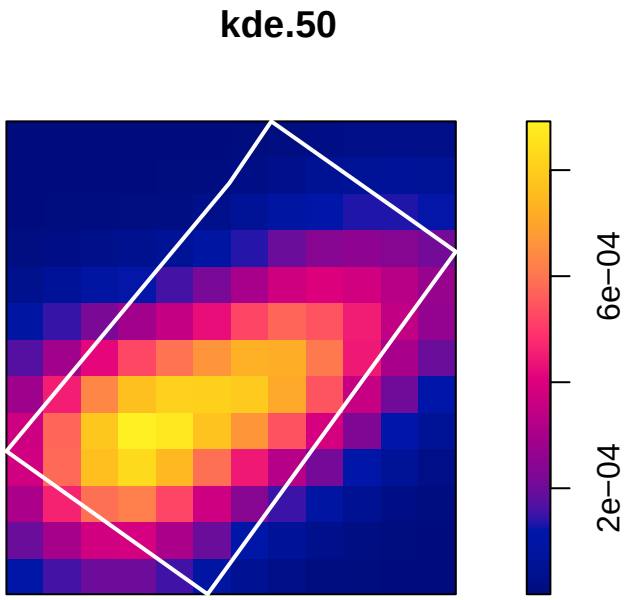

Figure 4: This figure depicts a KDE bandwidth of 50 meters with the study area superimposed onto it. It reveals that some areas have much higher density than others.

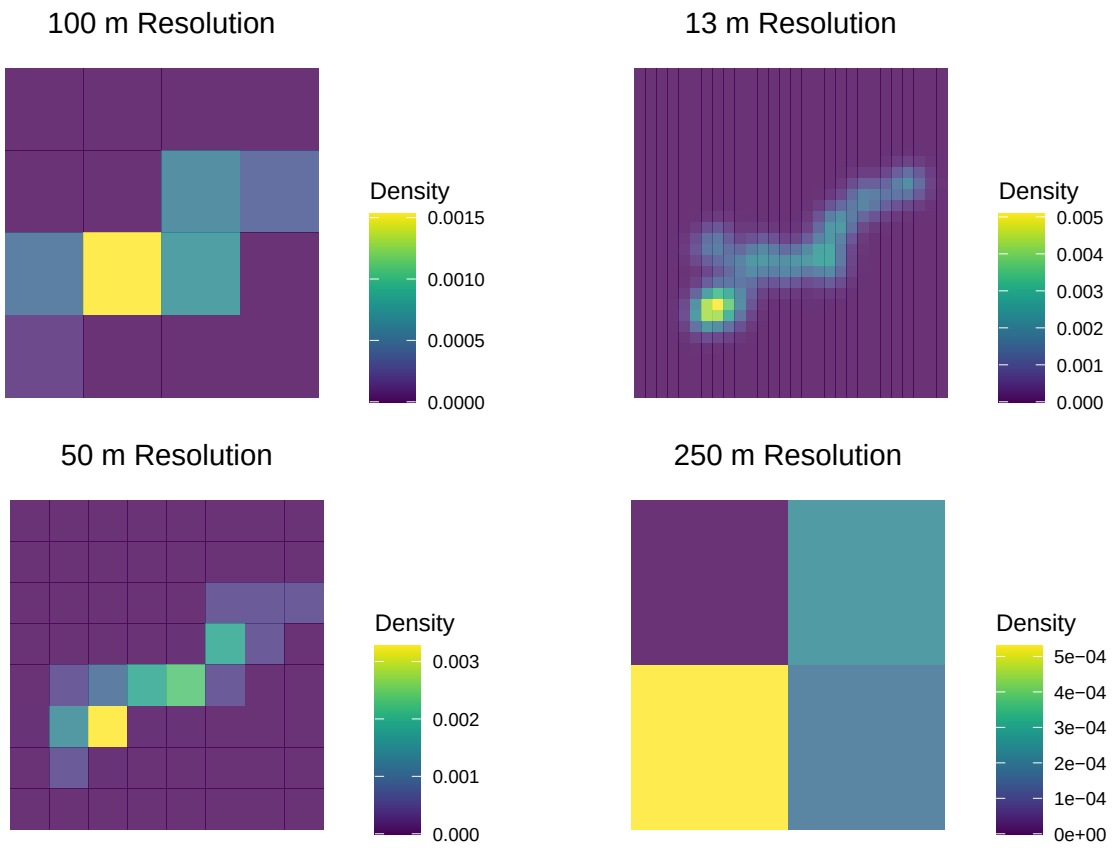

Figure 5: This figure illustrates the effects of different spatial resolutions on the KDE. As you can see, the larger the resolution, the more challenging it is to interpret. Therefore, finding a balance, such as with a 50m resolution, is important.

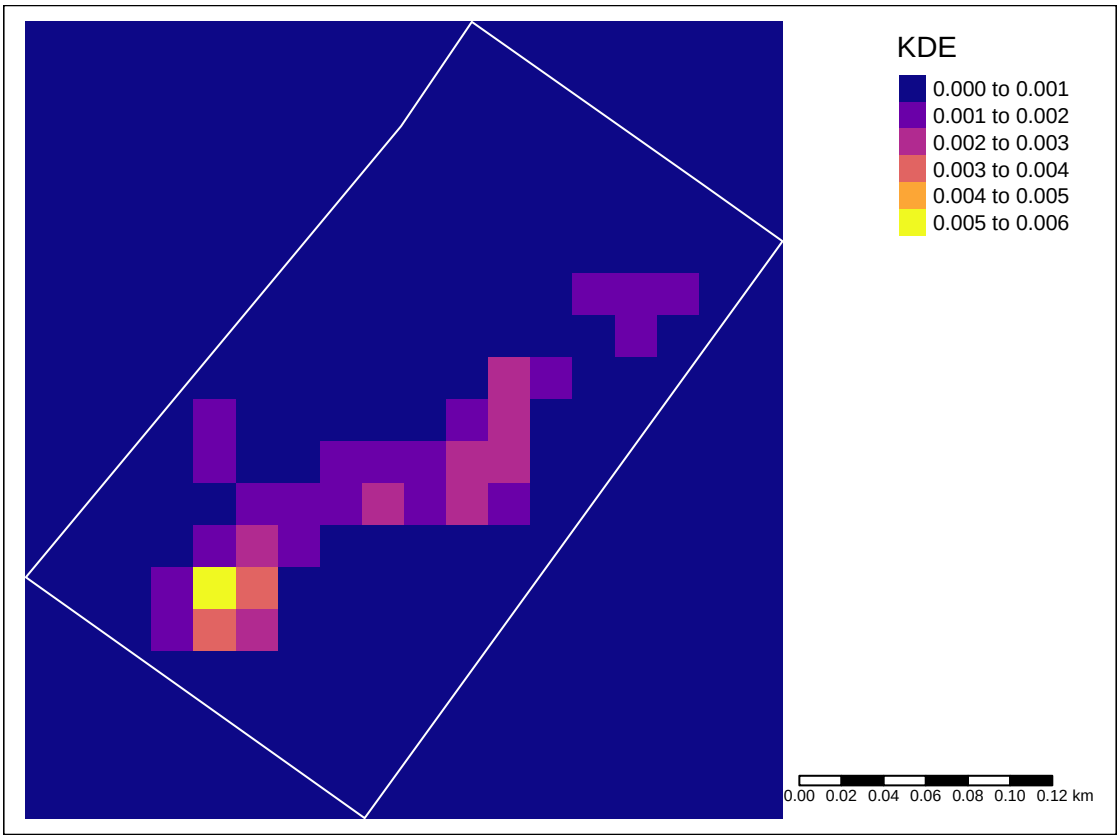

Figure 6: This map displays the final KDE surface after selecting the optimal bandwidth (13 meters) and the optimal spatial resolution (20 meters).

2.2.3 KDE Results Description

Figure 2 and 3 illustrate various KDE bandwidths arranged in a 4 by 4 grid. The bandwidth, a critical parameter in Kernel Density Estimation (KDE), controls the degree of smoothing applied to the data. A larger bandwidth results in a smoother estimate, while a smaller one captures finer details. This sensitivity underscores the significance of selecting appropriate parameters for meaningful and accurate results in KDE analysis.The wrong choice of parameters can lead to two undesirable outcomes: over-smoothed or under smoothed estimates. In the former case, essential features in the data may be lost, and in the latter, spurious patterns which do not accurately represent the data can arise. Moreover, conducting sensitivity analysis by testing different parameter settings and assessing their impact on results is essential to ensure the reliability of the KDE model's outputs. In Figure 2, the outputs are particularly sensitive to changes in the sigma value. For instance, a sigma value of 13 yields a narrower kernel and less smoothing, highlighting smaller-scale patterns and small data variations while reducing the emphasis on larger-scale features. Conversely, with bandwidths of 500m and 250m, the excessive smoothing results in a lack of differentiation between objects on the surface.

The resolution figure (Figure 5), illustrates the contrasting effects of lower versus higher cell resolutions. As shown, with larger cell sizes, small variations may be averaged out or inadequately represented in the analysis, making interpretation nearly impossible. This is clearly evident in the case of the 250m cell resolution, which results in just four generalized blocks of colors. Conversely, if the cell size is too small, it can make differences in the data challenging to interpret because the changes in the data are so small.

Lastly, after conducting sensitivity analyses that involved testing different parameter settings for resolution and bandwidth and assessing their impact on the results, the parameters of 13 meters for bandwidth and 15

meters for spatial resolution are selected. These chosen parameters are then combined and applied to create the final KDE map.

2.3 Quadrat Analysis

2.3.1 Quadrat Analysis Methods

Quadrat analysis is a method used to assess whether the spatial pattern significantly deviates from randomness. For instance, if each cell contains an equal number of points, the pattern would be considered dispersed. The first step in quadrat analysis for this study involved calculating the number of quadrants. Subsequently, after selecting the appropriate number of quadrats, the output was plotted.

```
# This value controls the number of quadrats used, in this case it will be a 2x2 grid
quads <- 2
qcount <- quadratcount(shpPPP, nx = quads, ny = quads)
```

Next, create a data frame from the qcount variable so that the quadrat analysis can be performed.

```
colnames(qcount.df) <- c("x","f")
```

The following steps involve performing a series of calculations to aid in learning more about the dataset via quadrant analysis. This section includes calculating the variance, the mean number of points per quadrant, the variance-mean ratio (VMR), and the Chi-square value. These formulas are to be subsequently transformed into R code.

This formula is used to calculate the variance

$$VAR = \frac{\sum f_i x_i^2 - \left[\frac{(\sum_i f_i x_i)^2}{m}\right]}{m-1}$$

This formal is used to calculate the mean

$$Mean = \frac{Number of Points}{Quadrats}$$

This formal is used to calculate the the variance-mean ratio (VMR)

$$VMR = \frac{VAR}{MEAN}$$

Lastly, the formula below is used to calculate The Chi-square value. The Chi-square value helps us understand if there is a significant difference between the expected and observed values in a dataset

$$\chi^2 = VMR(M-1)$$

Lastly, once the formulas have been calculated in R, they are stored in variables and placed in a data frame for presentation.

2.3.2 Quadrat Analysis Results

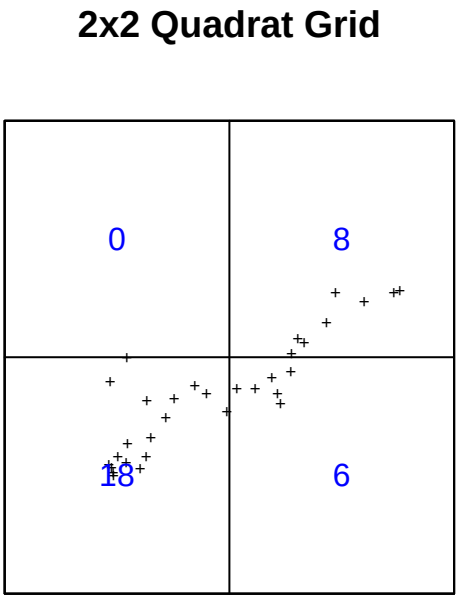

Figure 7: This figure illustrates a 2x2 quadrat grid with the sample locations of large trees in the study area. The numbers in each quadrant represent the count of points that fall inside that quadrant.

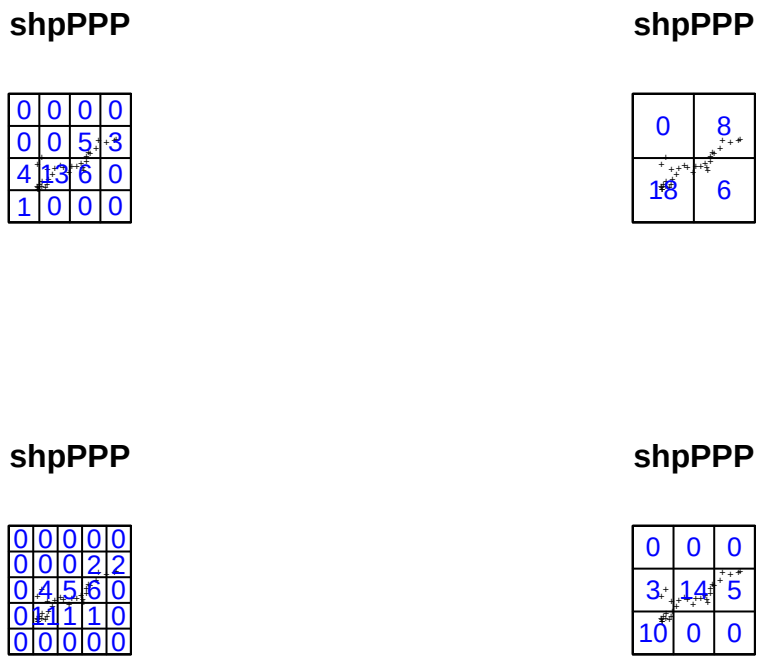

Figure 8: This figure presents multiple quadrat grids with different quad values. The objective is to demonstrate how the ouput changes when you alter the number of quadrats. As with the previous figure, the numbers in each quadrant represent the count of points that fall inside that quadrant.

This table presents the number of Quadrats, Variance, Mean, Variance-Mean Ratio, and Chi-square value of the dataset. These statistics aim to describe the distribution of large trees in the study area. (Table 1).

Table 1: Quadrat Statistics

Quadrats	Variance	Mean	VMR	Chisquare
4	56	8	7	21

2.3.3 Quadrat Results Description

The results of the Quadrat analysis with four quadrats indicates a clustered spatial pattern that significantly deviates from randomness. This is evident as each cell within the quadrats contains very different values. For instance, if the quadrats had a similar number of values in each cell, it would suggest a dispersed pattern. Conversely, if one or a couple of cells had the vast majority of points, it would also indicate a dispersed pattern (Figure 7 and 8). The second quadrat figure illustrates the effects of varying the number of quadrats on the spatial distributions. The results demonstrate that even when the number of quadrats changes, the overall spatial pattern remains consistently clustered (Figure 8).

Regarding the statistics calculated for this part of the analysis, several conclusions can be drawn. In a quadrant analysis with four quadrants, the data exhibits significant variability, as reflected by a variance

of 56 (Table 1). This variance suggests that the distribution of points is not uniform, indicating spatial heterogeneity within the dataset, in other words the data variable. The mean of 8, represents the expected average number of points in each quadrant. However, the Variance-Mean Ratio (VMR) of 7 reveals a substantial departure from randomness. This high VMR value points to spatial clustering, signifying that certain areas in the quadrat contain more points than expected in a random distribution. Furthermore, a relatively large Chi-square value of 21 indicates a larger deviation from the normal distribution, indicating a non-random association or pattern within the data. To conclude,these findings strongly suggest the presence of a clustered spatial trend within the dataset.

2.4 Nearest Neighbor Distance

2.4.1 Nearest Neighbor Distance Methods

The purpose of NND analysis is to determine the average distance between each point and its nearest neighbor. More specifically, we aim to investigate whether the spatial distribution of large trees significantly differs from a random distribution. The initial step in the NND analysis involved calculating the distance from each point to its nearest neighbor, a process facilitated by the **nndist** function from the **spatstat** library. Subsequently, the output of the 'nndist' function was converted into a dataframe, which will simplify the NND calculations.

Before performing the calculations, it is important to declare variables for the study area and density, which will be used in most of the calculations as shown in the code below.

```
# study Area and Density declarations
studyArea <- as.numeric(st_area(sa))
pointDensity <- n / studyArea
```

After those variables are declared, use the formulas below to calculate various statistics related to the NND.

Firstly, the mean NND is calculated, which represents the average distance between each point in the dataset and its nearest neighbor

$$\overline{NND} = \frac{\sum NND_i}{n}$$

Next, the NNDr, or the mean NND under a random distribution, which represents the average NND you would expect in a completely random point distribution, is calculated using the formula below.

$$\overline{NND_r} = \frac{1}{2\sqrt{Density}}$$

Next, the NNDd, or the mean NND under a dispersed distribution, which represents the average NND you would expect in a completely dispersed point distribution, is calculated using the formula below.

$$NND_d = \frac{1.07453}{\sqrt{Density}}$$

Moreover, calculate the Znnd which assess the significance of a difference between sample means and population means

$$Z_{nnd} = \frac{nnd - nnd_r}{\sigma nnd}$$

Next, calculate the standard deviation of the nnd which quantifies the variability or spread in NND values

$$\sigma_{nnd} = \frac{0.26136}{\sqrt{n(Density)}}$$

Finally, compute the standardization index (R-value)

$$R = \frac{\overline{nnd}}{\overline{nnd_r}}$$

After computing the statistics, they were converted into the R programming language, and the results were organized in a data frame for presentation.

2.4.2 Nearest Neighbor Distance Results

This table presents the results from the NND analysis, showing the calculations for NND, NNDd, NNDr, Z-score, Standardization Index, Ratio, and the study area. The purpose is to understand whether the data is clustered, dispersed, or random. (Table 2).

Table 2: Nearest Neighbour Distance Calculations

StudyArea	NNDd	NNDr	NND	Zscore	Ratio
62946.43	47.66	22.18	12.49	1.67	0.56

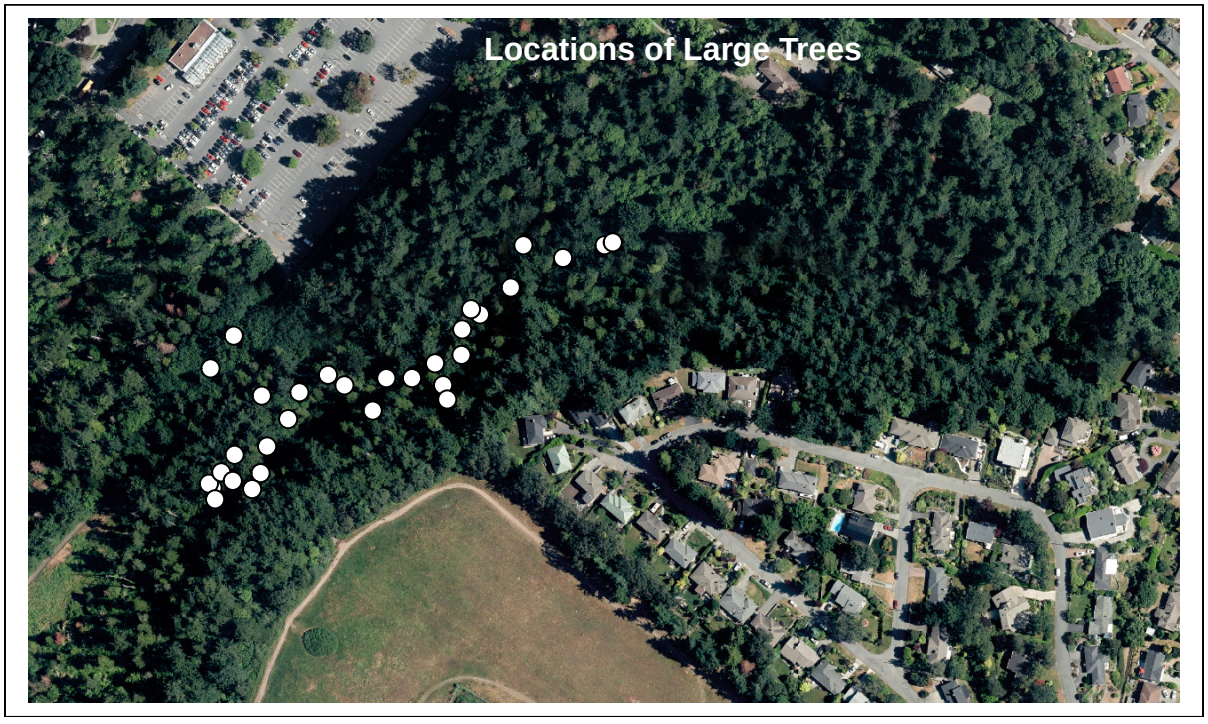

Figure 9: This maps shows the location of large trees in Mystic Vale

2.4.3 Nearest Neighbor Distance Results Description

The Nearest Neighbor Distance (NND) calculations provide valuable insights into the distribution of large trees in Mystic Vale. To better understand the spatial arrangement of these trees, we use a hypothesis test.

Hypotheses

Ho (Null Hypothesis): NND = NNDr

- In this scenario, it assumes that the point pattern is random.

Ha (Alternative Hypothesis 1): NND not equal to NNDr

- This implies the point pattern is not random.

Ha (Alternative Hypothesis 2): NND > NNDr

- In this case, it suggests the point pattern is more dispersed than random.

Ha (Alternative Hypothesis 3): NND < NNDr

- Here, it indicates the point pattern is more clustered than random.

To begin, a *significance level* of 90% is set, which represents the probability of committing a Type I error (rejecting the null hypothesis when it is true). Subsequently, a *Z-test* is conducted, utilizing a Z-score of 1.67 and a 90% confidence level, resulting in a critical range of -1.645 to 1.645. The rejection region is then defined. For the detection of greater dispersion, the rejection region is positioned in the upper tail of the distribution. Conversely, for identifying more clustering, the rejection region is placed in the lower tail of the distribution. In conclusion, the critical range for a Z-score of 1.67 and a 90% confidence level is approximately -1.645 to 1.645. As the Z-score falls outside this range, the null hypothesis is rejected. Consequently, it can be concluded that the spatial distribution of large trees in Mystic Vale is not random.

Additionaly, The specifics of the distribution become evident when examining the individual values in the NND analysis. Since NND (12.49) is less than NNDr (22.18), we can accept alternative hypothesis 3, suggesting that the spatial distribution of large trees in Mystic Vale is clustered. This pattern is evident visually in Figure 9, were many locations are significantly clustered.

2.5 K-Function

2.5.1 K-Function Methods

The K-function, employed to investigate whether points are clustered, evenly dispersed, or randomly distributed in relation to each other, is analyzed in the following steps. Initially, the Kest() function is utilized to compute the K-function. Furthermore, the Kest function is employed to calculate Ripley's K-function for the data in shpPPP. The "Ripley" correction parameter is used to address points near the study area's boundary (Kest: K-function). Following the computation of the K-function, it is essential to visualize it, a task accomplished through the plot() function.

```
# Step 1
k.fun <- Kest(shpPPP, correction = "Ripley")
plot(k.fun)
```

Next ,calculate the envelope of the K-function for a spatial point pattern using the Kest(). The envelope serves as a reference for assessing the statistical significance of the K-function values (Kest: K-function). It employs 99 simulations to generate the random envelope and accounts for edge effects using Ripley correction. The plot function is used to visualize the K-function and its envelope.

```
k.fun.e <- envelope(shpPPP, Kest, nsim = 99, correction = "Ripley", verbose = FALSE)
plot(k.fun.e, main = "K-Function Plot")
```

2.5.2 K-Function Results

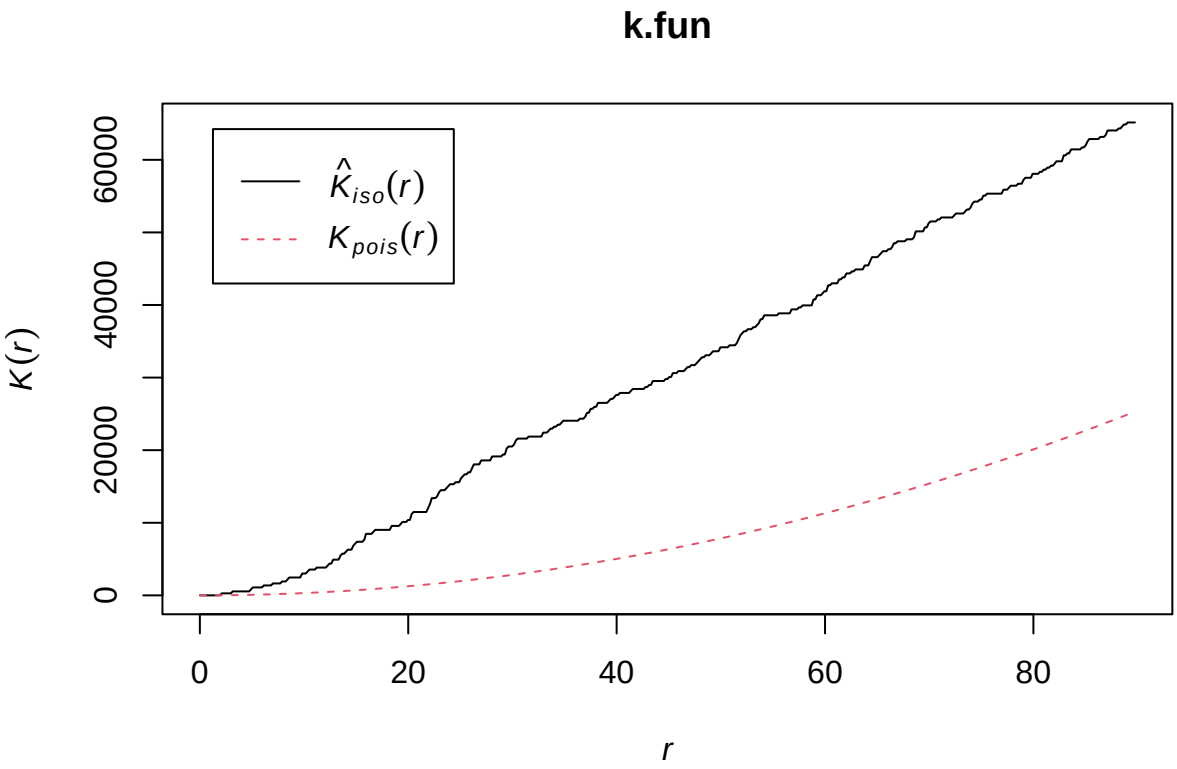

Figure 10: This depecits a K-function plot where the observed K-function as a solid black line, while the red dashed line represents the average of simulated K-functions

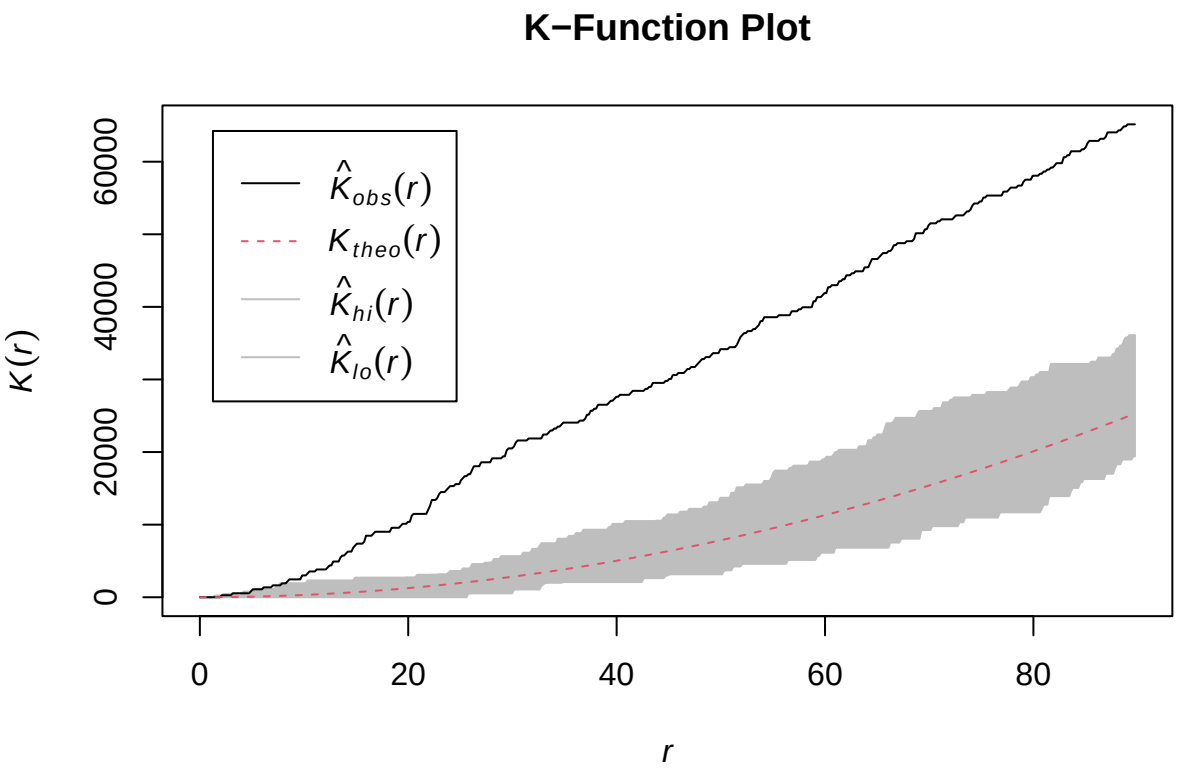

Figure 11: This figure shows a K-function plot. The solid black line represents the observed spatial pattern, the dashed red line represents the expected random spatial pattern, and Khi is the higher confidence envelope, while Klo is the lower confidence envelope. The purpose of the K-function is to assess how much your distribution differs from a random one.

2.5.3 K-Function Results Description

The K-function results provide significant insights into the spatial distribution of large trees in Mystic Vale (Figure 10). The first figure presents the observed K-function as a solid black line, while the red dashed line represents the average of simulated K-functions (Choiruddin et al., 2023). For instance, in the context of Mystic Vale, the black solid line in the graph illustrates the actual distribution of trees, while the red dashed line signifies the average distribution of large trees in completely random distribution.

The second K-function graph (Figure 11) is similar to the first figure, but it includes the interval of the simulated envelope (Choiruddin et al., 2023). This graph strongly suggests that the data is clustered because the observed K-function curve consistently resides above the envelope. This implies that the points in the dataset are closer to each other than would be expected by random chance. Furthermore, the graph indicates a significant deviation from randomness, as the envelope deviates from the observed values.

2.6 Conclusion

2.6.1 Summary of Findings

The objective of this study was to collect event-based observations of large trees in Mystic Value and conduct multiple point pattern analyses to determine whether the spatial distribution of these trees significantly deviates from randomness.

Firstly, the **KDE analysis**, which is useful for understanding the distribution and density of points in the study area, revealed that the points were clustered in the same region of the study area. This clustering is evident in the final KDE map, as well as in the cell resolution and bandwidth visualizations. The analyses found that optimal bandwidth for this study was 13 meters, and the spatial resolution was 20 meters.

Regarding **quadrat analysis**, using four quadrats, reveals a clear clustered spatial pattern that differs from randomness. Each cell within the quadrats contains notably different values, indicating non-uniformity. This is in contrast to a dispersed pattern, where each cells would have a similar amount of points. Moreover, adjusting the number of quadrats doesn't change the overall spatial pattern. The statistics support this conclusion, a high variance (56) signifies non-uniform distribution, and the Variance-Mean Ratio (VMR) of 7 suggests clustering, as certain areas have more points than expected in a random distribution. A chi-square value of 21 further confirms the non-random spatial trend, strongly suggesting a clustered pattern within the dataset.

Concerning the **nearest neighbor distance** analysis, a hypothesis test uncovered several insights about the data, particularly revealing that the observed spatial arrangement of the data points demonstrates a non-random, clustered pattern. This is evident as the null hypothesis was rejected. Lastly, the **K-function** analysis employed a K-function graph, and the results suggest that the locations of large trees are clustered, as evidenced by the observed K-function curve consistently residing above the envelope. In conclusion, the various analysis methods in this study suggest that the spatial distribution of large trees in Mystic Vale significantly deviates from randomness due to clustering.

2.6.2 Description of Cause

To summarize, the objective of this study was to collect event-based observations of large trees in Mystic Vale and conduct several point pattern analyses to determine whether the spatial distribution of these trees significantly deviates from a random pattern. Point pattern analysis is an important aspect of spatial statistics, and research on its applications span various domains. For example, it has been employed to understand the growth patterns of Mayan cities (Adams & Jones, 1981). Additionally, it is used in epidemiology, such as the analysis of the spatial patterns of Covid-19 infections or the spatial clustering of Hodgkin's disease in the San Francisco Bay area (Glaser, 1990). In summary, point pattern analysis offers an efficient means to visualize and interpret spatial patterns.

3 References

Boots, B.N., & Getis, A. (1988). Point Pattern Analysis. Reprint. Edited by Grant Ian Thrall. WVU Research Repository, 2020

Getis, Arthur & Ord, Keith. (1992). The Analysis of Spatial Association by Use of Distance Statistics. Geographical Analysis. 24. 189 - 206. 10.1111/j.1538-4632.1992.tb00261.x.

Bw.diggle: Cross validated bandwidth selection for kernel density. RDocumentation. (n.d.). https://www.rdocumentation.org/packages/spatstat/versions/1.64-1/topics/bw.diggle

Choiruddin, Achmad & Hannanu, Firdaus & Mateu, Jorge & Fitriyanah, Vanda. (2023). COVID-19 transmission risk in Surabaya and Sidoarjo: an inhomogeneous marked Poisson point process approach. Stochastic Environmental Research and Risk Assessment. 37. 1-12. 10.1007/s00477-023-02393-5.

Adams, R. E. W., & Jones, R. C. (1981, April). Spatial Patterns and Regional Growth among Classic Maya Cities. American Antiquity, 46(2), 301–322. https://doi.org/10.2307/280210

GLASER, S. L. (1990, July 1). SPATIAL CLUSTERING OF HODGKIN'S DISEASE IN THE SAN FRANCISCO BAY AREA. American Journal of Epidemiology, 132(supp1), 167–177. https://doi.org/10.1093/oxfordjournals.aje.a115779