Assignment 02
Zixi Liu

Exploring Whether the County Seats in Iowa Are More Dispersed Than Random

Introduction

Every county in Iowa has a county seat, in which a city or town was designated as the county's center for local government. With few exceptions, the county seat was the town in the county with the largest population. In a few cases, when the largest town lacked a central location, the county seat went to a smaller town nearer the county center. The regularity of the grid-like pattern of county boundaries in Iowa strongly contributes to the spatial regularity of its county seats. It is therefore expected that the locations of county seats in Iowa is dispersed over the state. However, whether the locations of these county seats are more dispersed than would be expected by chance alone is controversial and will be examined in this study.

August Lösch, a German economist, theorized the most advantageous regional shape for market areas with the “Central Place Theory”, in which he suggested a honeycomb network that covered the whole area is the most favorable as it realized the same total demand as other regional shapes but with the least requirement for land. He used the beer-producing market areas in South Germany as an example to underline that demand would fall rapidly with the distance from the product site and regular hexagons, fully utilizing the corners in the economic region, has the advantage of maximizing the demand per unit of area. Therefore, an idealized agricultural economy should build upon a hierarchy of towns that are regularly spaced on a hexagonal grid of small farms. Peter Haggett made further contributions in analyzing the distribution of towns in his book *Locational Analysis in Human Geography*, in which he argued on the assumption that all settlement must have equal access to resources so that “the centers of the hexagons, the nodal points, must form a regular triangular lattice to conform with the same minimum energy requirements” (Haggett, 88). Towns would develop at the centers of a “honeycomb” because locations at a center of gravity has the advantage of being almost equally near to all boundaries and the uniform pattern of locations would facilitate communication between settlements.

We found this hexagonal shape of economic regions is quite similar to the case of Iowa when we examine the voronoi diagram of county seats in Iowa (Fig 1b). Voronoi tessellation is a partition of the region, in which each point in space is placed in the cell defined by its neatest county seat. When first we looked at the map of Iowa counties seats (Fig 1a), we noticed that especially for interior countries (away from Missouri and Mississippi Rivers), each seat is near the center of its county. The voronoi cells for county seats is also close to hexagonal shapes for interior counties. Comparing the two figures, we found that the voronoi cells for county seats reflects drawbacks Haggett pointed out in the Löschian system of regular hexagons such that actual settlement arrangements could be distorted by other relevant considerations such as agglomeration.

Drawing upon Michael Dacey’s county seat model — an alternative probabilistic model of “place” patterns that are more regular than Poisson randomness, we assumed that county seats

were of sufficient prominence in each county so that their distribution are assumed to be more dispersed than random (Darcey, 54:559-565). Darcey's model of dispersion assumed that each county are required by law to have a county seat so that each county must have at least one place. The frequency of place in this shifted Poisson distribution started from one place instead of zero and then we compare this new distribution with complete spatial randomness. Darcey's model of dispersion differs from Lösch's model in the way that Darcey's country-seat model lacks generality as it built upon the case of Iowa, whereas Lösch made references to several market areas in order to theorize regular hexagons as the most favorable economic region.

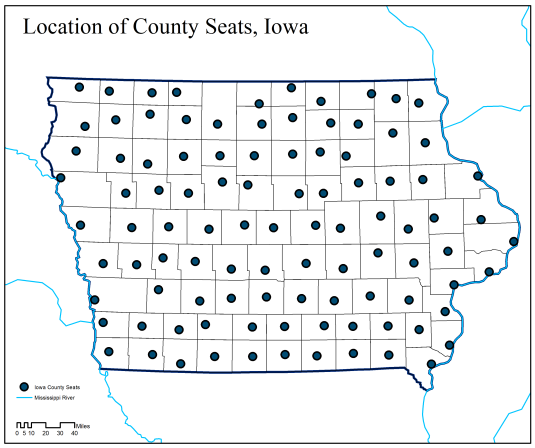

Figure 1a: Location of County Seats in Iowa

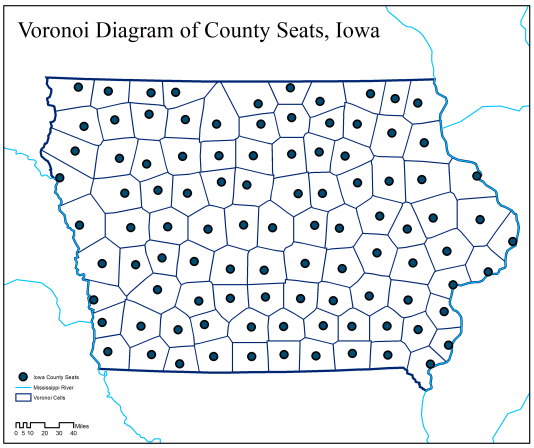

Figure 1b: Voronoi Diagram of County Seats in Iowa

Methodology and Results

From the above figures, it is obvious that county seats have a dispersed distribution. In the next steps, we want to run statistical tests to reject the Complete Spatial Randomness hypothesis in favor of the alternative hypothesis of dispersion. Complete Spatial Randomness (CSR) is the hypothesis that assumes n points are located randomly in a region R and the locations of these points has no influence on each other, which suggests they are statistically independent for the region. Under the Complete Spatial Randomness, we have the cell count distribution that the sum of independent Bernoulli variables which describes the event Xi(C) that point i is located in region C is by definition a Binomial random variable with distribution given by

$$\Pr[N(C)=k\,|\,n,R]=\frac{n!}{k!(n-k)!}\left(\frac{a(C)}{a(R)}\right)^{k}\left(1-\frac{a(C)}{a(R)}\right)^{n-k},\ k=0,1,..,n \quad \text{, where} \quad N(C)=\sum_{i=1}^{n}X_i(C)$$

Clark-Evans Test is a test of Complete Spatial Randomness hypothesis using nearest neighbor distances. Under the Central Limit Theorem, independent sums of identically distributed random variables are approximately normally distributed. Hence the most common test of the CSR Hypothesis based on nearest neighbors involves a normal approximation to the sample mean of D , as defined by $\bar{D}_m=\frac{1}{m}\sum_{i=1}^{m}D_i$.

From the Central Limit Theorem, it follows that for sufficiently large sample sizes, $\bar{D}_m$must be approximately normally distributed under the CSR Hypothesis with mean and variance given by

$$\bar{D}_m \sim N\left[\frac{1}{2\sqrt{\lambda}}, \frac{4-\pi}{m(4\lambda\pi)}\right]$$

The null hypothesis of CSR states that the point events occur within a study area is completely random. We want to test on this hypothesis by standardizing the sample mean, $\bar{D}_m$, in order to use the standard normal tables. Hence, if we denote the standardized sample mean under the CSR Hypothesis by

$$Z_m = \frac{\bar{D}_m - E(\bar{D}_m)}{\sigma(\bar{D}_m)} = \frac{\bar{D}_m - [1/(2\sqrt{\lambda})]}{\sqrt{(4-\pi)/(m4\pi\lambda)}}$$

then it follows from that under CSR, $Z_m \sim N(0,1)$

If we use all the distances to construct the sample-mean statistic, we would violate the assumed independence of nn-distances on which this distribution theory is based. Therefore, we must select a subset of nn-distance values that contained no common points. Then we construct a sample-mean value $\bar{d}_m = \frac{1}{m}\sum_{i=1}^{m} d_i$ and use this to construct tests of CSR.

Running tests on CSR hypothesis has certain drawbacks as boundary issues (which occur because of the loss of neighbors in analyses that depend on the values of the neighbors) might introduce considerable bias into parameter estimates for that region. For instance, in our case, county seats would seem to be more dispersed than they actually are because the nearest neighbor statistic are influenced by the shapes of study regions and the expected values of nn-distances are increased for points near the boundary. To perform the Clark-Evans Test in Matlab and JMP, we first calculated the nearest neighbor distances from the location of all 99 county seats. Then we randomly select 30 nn-distances as our sample and run the Clark-Evans Test. Below is the table of the Clark-Evans Test Results.

area	n	lam	mu	sig	s-mean	Z	P-Val CSR	P-Val Clust	P-Val Disp	S	chi-square	.05-quantile
56269	99	0.001759	11.9203	1.137622	20.64786	7.671755	1.70E-14	1	8.48E-15	145.8158	1	43.18796

Table 2: Clark-Evans Test Results Based on a Sample of 30 Nearest Neighbor Distances in JMP

We pick a sub-sample of 30 because under the Central Limit Theorem, the sample size of a sufficiently large sample should be at least 30 and we cannot violate the assumed independence of nn-distances on which this distribution theory is based. We can observe from the table that the value “mu” gives the theoretical mean of nearest-neighbor distance predicted by the CSR theory and the sample mean is calculated in the column “s_mean”. From the table above, we observe that the value “mu” is 11.9203 and “s_mean” is 20.65. The sample mean of the nearest neighbor distance is larger than the theoretical mean of nearest neighbor distance where points are complete random. This comparison result suggests that the points in our sample is more dispersed than random.

Now we want to further examine the p-values of the Clark-Evans Test to determine whether this pattern is significantly more dispersed than random. First, we use the one-tail test of “dispersion”

versus the null hypothesis of CSR and examine the p-value to evaluate how likely it would be to obtain a standardized sample mean this large if the CSR Hypothesis were true. Here the appropriate p-value is given by “P-Val Disp”. Drawing upon the table above, the p value of dispersion is 8.48E-15, which is so low that we could reject the null hypothesis of CSR and support the alternative hypothesis that the pattern of county seats are more dispersed than random. Then we look at the p-values of the two-tail CSR test in which both the possibility of “significantly small” values of $\bar{d}_m$(clustering) and “significantly large” values of $\bar{d}_m$(dispersion) are considered. The p-value evaluates how likely would it be to obtain a value as far from zero as our standardized sample mean if the CSR Hypothesis were true. The p-value for the two-tail CSR is 1.70E-14, which is also so low that it yields strong evidence against the CSR Hypothesis.

We also run the Clark-Evans Test in Matlab with our sample of nn-distances and arrived at a p-value of 2.1383E-14 in the one-tail test for dispersion. This p-value is also very low so that we could reject the null hypothesis of CSR and support the alternative hypothesis that the locations of county seats in Iowa are more dispersed than random. Then, since the results of this Clark-Evans test depend on the particular sample chosen, we want to run the Clark-Evans Test 1000 times with simulations. Here we use Monte Carlo simulations, which rely on repeated random sampling to obtain results and allows us to consider different outcomes in the process in order to deal with uncertainty. The p-value of one-tail test for dispersion this time is 7.8715E-14, which is also very low so that we could reject the null hypothesis and favor the alternative hypothesis. It is also interesting to note that the p-value we arrived at through simulations is higher than the p-value we initially get, which indicates we are more likely to obtain a standardized sample mean this large if the CSR Hypothesis were true and therefore less likely to reject the CSR hypothesis.

We also create the Vororoi Diagram of county seats in Matlab and compare it with the Vororoi Diagram in ARCMAP. It is obvious that the Vororoi Diagram in ARCMAP deals better with border effects than the Vororoi Diagram in Matlab and is hence more favorable.

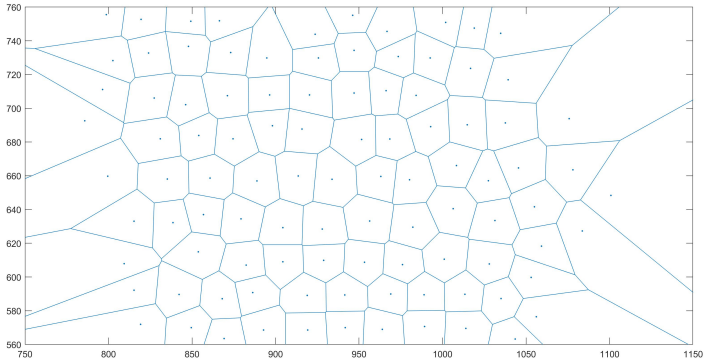

Figure 3: Vororoi Diagram of County Seats in Matlab

Discussions and Conclusions

We cannot safely reject the CSR Hypothesis because in order to perform the Clark-Evans Test, we want to make sure the location of one county seat in our sample is independent of the location of another county seat. However, in reality the location of a county seat might be influenced by some other relevant considerations such as distortion of agglomerations suggested by Darcey in his findings. Also we first perform the Clark-Evans Test on a sub-sample of 30 and then run the test 1000 times. We should further test the CSR Hypothesis using different sub-sample sizes and simulate more than 1000 times to see if the test results change.

In *The Economics of Location*, Lösch introduced the "Central Place Theory" with a system of lowest-order (self-sufficient) farms in South Germany, which were regularly distributed in a triangular-hexagonal pattern. Then he pointed out that a "production site" such as towns should be located sometimes in a settlement and sometimes at a center of gravity and these centers are all separated from one another by a minimum distance, which resonates with our findings that the county seats near Mississippi River are close to the river rather than at the geographical center and the location of county seats are more dispersed than random (Lösch, 121). The counties near the river has county seats close to the Mississippi River probably because of the distortion of resource localization as port cities may have more access to resources. Then, according to Dacey's "county-seat model", Darcey assumed county seats are by definition of sufficient prominence to be included among the placed in each county and if we assume other places are consistent with Complete Spatial Randomness, then the distribution of county seats are more dispersed than random. After running his tests and using his modified Poisson model, he has arrived at a shifted Poisson distribution where variance is smaller than the mean. Intuitively, we regard a frequency distribution as more regular than random if its variance is smaller than its mean. Therefore he conclude his tests provides substantial evidence that the arrangement of county seats in Iowa is more dispersed than random, which again resonates with our test results (Darcey, 564).

Tectonic Patterns in Areal Volcanism, Uganda

Introduction

Areal volcanism is a phenomenon in which a number of volcanic vents with associated tephra are found in the same area, and which are clearly fed from the same magmatic source (Tinkler, 335). It is considered as volcanic clusters and various studies has been conducted to analyze the point pattern of these craters. In this study, we will investigate the findings by Keith J. Tinkler in his article *Statistical Analysis of Tectonic Patterns in Areal Volcanism: The Bunyaruguru Volcanic Field in West Uganda* and perform a statistical analysis of the crater patterns in Uganda. Uganda, our area of interest, has eight active and extinct volcanoes in total, among which the Bunyaruguru field lies in western branch (Uganda) of the East African rift system (Figure 1). In western branch (Uganda), there are unique potassic alkaline rocks which could be found from North to South, including Fort Portal, Katwe-Kikorongo, Bunyaruguru fields and the Bufumbira field shown in figure 2, which implies these volcanic fields may have some structural patterns.

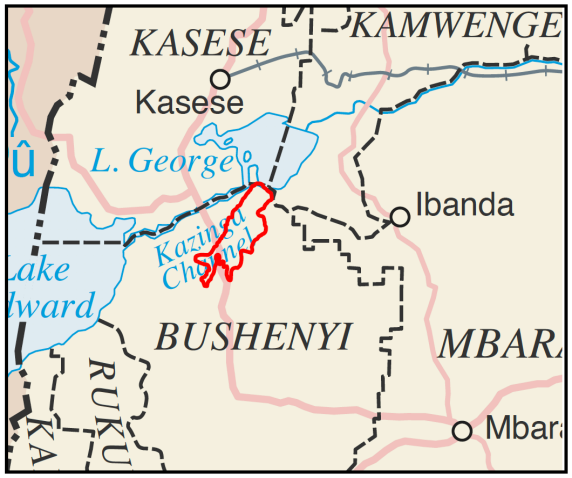

Figure 1: Map of Uganda and Neighbor Countries

Tinkler has conducted a quantitive analysis to analyze the spatial arrangement of 149 explosion craters in the western rift of Uganda and he found that the spatial pattern of the craters reveals significant structural patterns that guided volcanism and the clustering of craters. In this analysis, we will try to replicate Tinkler's findings using K-functions and L-functions to analyze the point patterns of volcanoes in Uganda.

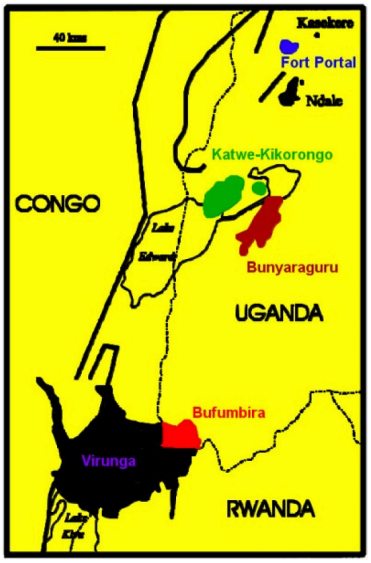

Figure 2: Map of Volcanic Fields

Methods and Results

First, we introduce the K-functions, which is a method to test the CSR hypothesis by incorporating "scale" as a variable in the point pattern analysis. K-functions can be used to test for clustering and dispersion at various scales of analysis with assumptions of spatial stationarity and isotropic underlying point process. K-functions are different from the Clark-Evans Test in our previous study with regard to the edge effect such that unlike the Clark-Evans Test, where the expected values of nn-distances are increased for points near the boundary edge effects, the expected value of point counts in K-functions are reduced for these points at border.

Intuitively, K-function takes the form $K(h) = \frac{1}{\lambda} E(\text{number of } \textit{additional} \text{ events within distance, } h, \text{ of an arbitrary event})$ to take into account point density and the varying scale h in order to yield information about clustering or dispersion. To apply the K-function under CSR Hypothesis, we will first ignore the

edge effects and use the formula $K(h) = \frac{1}{\lambda}(\lambda\pi h^2) = \pi h^2$ where λ denoted the point density and K-function reduces simply to area. In this case, $K(h) > \pi h^2$ implies a mean point count higher than would be expected under CSR, and hence indicates some degree of clustering at scale h. Similarly, $K(h) < \pi h^2$ implies a mean point count lower than would be expected under CSR, and hence indicates some degree of dispersion at scale h.

We want to investigate the estimation of K-functions values and we use $d_{ij} = d(s_i, s_j)$ to denote the Euclidean distance between a pair of points. Then we define the indicator function for point patterns by

$$I_h(d_{ij}) = I_h[d(s_i, s_j)] = \begin{cases} 1 & , d_{ij} \le h \\ 0 & , d_{ij} > h \end{cases}$$

As the total number of additional points within distance h is given by the sum $\Sigma_{j \neq i} I_h(d_{ij})$, we have an estimate of the K-function, designated as the sample K-function, is given by

$$\hat{K}(h) = \frac{1}{\hat{\lambda} n} \sum_{i=1}^{n} \sum_{j \neq i} I_h(d_{ij})$$

where $\hat{\lambda} = n / a(R)$. If the underlying process were truly stationary (and edge effects were small) then this sample K-function would be approximately unbiased as an estimator of the common expected point count.

However, when we use K-functions to test the CSR Hypothesis, area values are difficult to interpret directly so we standardize K-functions for analysis. For each h, we have $L(h) = \sqrt{\frac{K(h)}{\pi}} - h$ so under CSR, we have L(h) = 0. Since L(h) is an increasing function of K(h) , it follows that L(h) is positive exactly when $K(h) > \pi h^2$ and is negative exactly when $K(h) < \pi h^2$. Therefore when L(h) is greater than 0, we have clustering at scale h and when L(h) is less than 0, we have dispersion at scale h. With the estimation of K(h), we can estimate L(h) by

$$\hat{L}(h) = \sqrt{\frac{\hat{K}(h)}{\pi}} - h$$

As shown in figure 3, our study region is very irregular and edge effects are very important. The expected value of point counts are reduced for these points at border, which results in a downward bias.

In order to analyze the volcanic point pattern in Matlab, we use Monte Carlo simulations and plot a simulation envelope of 99 random patterns (figure 4). The essential idea of the simulation envelopes is to simulate N random patterns as above and to compare observed estimate L(h) with the range of estimates $\hat{L}_i(h)$ obtained from this simulation. Lower-envelope is the minimum of $\hat{L}_i(h)$ and upper-envelope is the maximum of $\hat{L}_i(h)$. We define the lower-envelope and upper-envelope functions respectively by $L_N(h) = \min\{\hat{L}_i(h) : i = 1,..,N\}$ and $U_N(h) = \max\{\hat{L}_i(h) : i = 1,..,N\}$. Then $\hat{L}_i(h)$ is compared with lower envelope and upper envelope for each h.

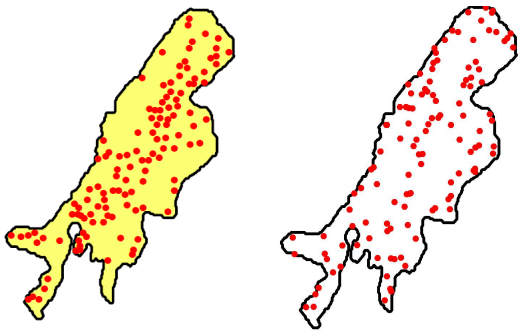

Figure 3: Map of Volcanic Fields and 120 Random Points in Uganda

We then compute the k-Function for a point pattern plus random point patterns for a single polygon and plots the normalized L-Function plus upper and lower envelopes. From figure 4, we observe that there is some clustering at radius 0.1-0.2 $\times 10^4$ and there is some degree of dispersion at larger radii. Then we want to also examine the p-value plot where we have clustered point pattern at a radius of 0.1-0.8$\times 10^4$. These results are more or less consistent with ones above for p-values that are around 0.1, but it also give us information on other p-values. From figure 5, we can observe that we have clustered points at the radius 0.1-0.8$\times 10^4$.

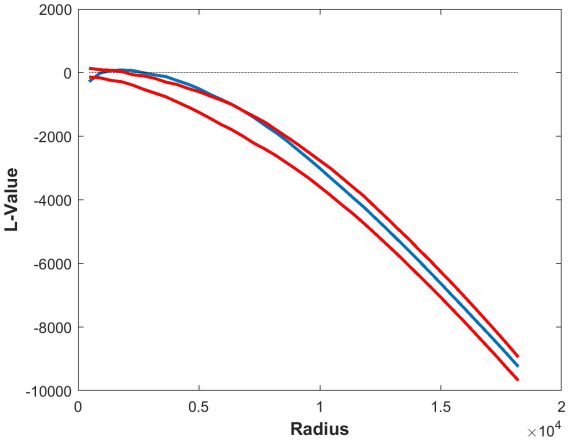

Figure 4: Plot of Simulation Envelope of 99 Random Patterns

This result resonates with what we observe in figure 3 such that the map on the left is the volcanic fields in Uganda and the map on the right is 120 randomly generated points in Uganda. The actual volcanic fields shown in the left map obviously are more clustered than the random points on the right.

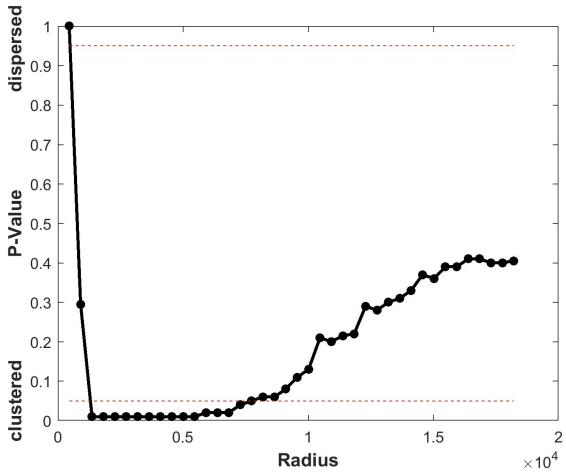

Figure 5: P-value Plot

Discussions and conclusions

In East Africa, spreading processes have already torn Saudi Arabia away from the rest of the African continent, forming the Red Sea. The East African Rift Zone is a new spreading center that may be developing under Africa. Our test results resonates with Tinkler's findings that the craters of volcanos are clustered. According to Tinkler, volcanos are clustered because of tectonic patterns in the East African rift system.

Exploring The Contagion Hypothesis of Myrtle Wilt in Tasmania

Introduction

Myrtle wilt is a disease which attacks Myrtle Beech trees, an evergreen tree native to Tasmania, and has caused considerable concern in many of the rainforest communities in Tasmania, Australia. Myrtle wilt occurs when a fungus named *Chalara* enters a Myrtle Beech tree through an exposed flesh wound and the fungus can also infect neighboring trees through root graft, which was mentioned by Jillian M. Packham in his *Studies on Myrtle Wilt* as a major driver of the spread of myrtle wilt. Various studies have been conducted to investigate the dynamics of the disease, especially the pattern of attack and its mortality rate. Previous surveys by Elliott et al. has shown that diseased trees were clumped, with the degree of clumping being dependent on the nearest neighbor distances within sites (Packham, 20). According to Packham, it is important to study the disease spread in different trees by rainforest type because the rate of spread are meant to be permanent and representative of their respective rainforest types.

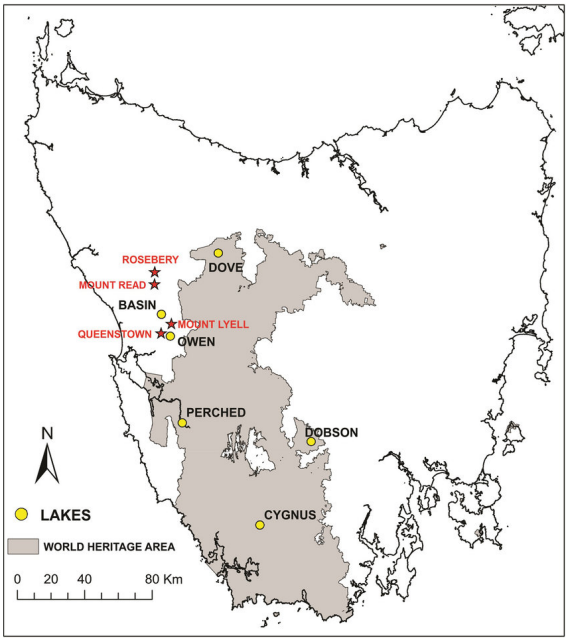

Figure 1: Map of World Heritage Area (Forests), Tasmania

He also points out that clumping of diseased trees is evident when a grid size greater than 10m was adopted, whereas a local spread of disease due to root grafts with a maximum distance of 8-10m between trees also occurs. In this study, we will analyze the spatial relations between diseased and health myrtles. In particular, we determine whether the spatial relations are consistent with contagion of myrtle wilt, or whether they are more consistent with random infections. We will use two methods to test the contagion hypothesis, which is whether the diseased trees appear in “clumps” or “patches”. We first use the random shift approach to assess the spatial dependencies between diseased and healthy trees. That is, we test the Spatial Independence Hypothesis and at scales where clumsiness is observed, healthy and diseased trees should show some degree of “repulsion”. Then we use the random permutations (relabeling) method to examine the spatial relations between the healthy and the diseased population. Within the diseased trees population, there should be some degree of “clustering”. From figure 2, we observe that there are some patches of diseased myrtles and we want to further test the contagion hypothesis in our next steps.

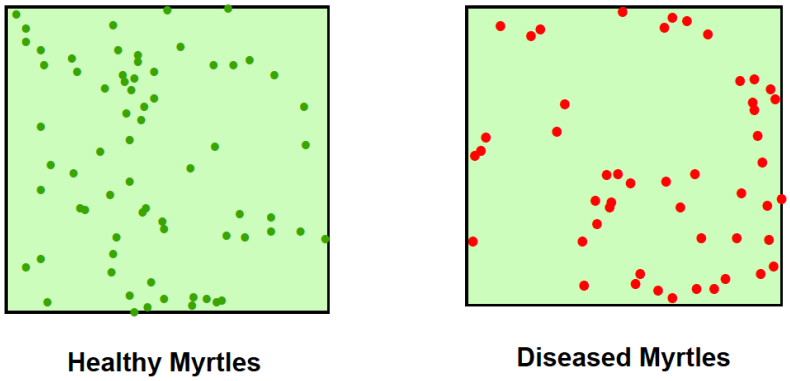

Figure 2: Maps of Diseased and Healthy Myrtles

Methods and Results

First, we have point data including 49 diseased myrtles and 69 healthy myrtles and the x and y dimensions of the bounding box, which is our grove in interest. To avoid neighborhoods overlapping themselves on the torus in our random shift test, the maximum distance for shifts cannot exceed half the minimum of the width and height of the bounding box, which is 46m in this case. Now we have a set of 11 representative distances between 1m and 46m to be used for our analysis.

We want to use the cross k-function to conduct a comparative analysis of two populations. Rather than looking at the expected number of population 1 within distance h of population 1, we want to look at the expected number of population 2 within distance h of population 1. Here we use λ_1 and λ_2 to denote the intensities of population 1 and 2. The cross k-function is therefore given by $K_{12}(h) = \frac{1}{\lambda_2} E(\text{number of } j\text{-events within distance } h \text{ of an arbitrary } i\text{-event})$

Then we want to estimate the cross k-function and observed intensities in population 1 and 2 are estimated by $\hat{\lambda}_k = \frac{n_k}{a(R)}$, $k = 1,2$ where n represents the sample size from the population.

Then we use an indicator function $I_h(d_{ij}) = I_h[d(s_i, s_j)] = \begin{cases} 1, & d_{ij} \le h \\ 0, & d_{ij} > h \end{cases}$ to indicate whether a member j of population 2 is within distance h of a given member i of population 1. Therefore our estimation of cross k-function is given by the sample cross k-function: $\hat{K}_{12}(h) = \frac{1}{\hat{\lambda}_2 n_1} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I_h(d_{ij})$

Alternatively, when we use the cross k-function, we are interested in the spatial pattern of one population compared with another population. For instance, under the Spatial Independence Hypothesis, the cross K-function yielded by random shifting process should be the same as for the original process.

We first conducted the Random Shift test to examine whether healthy and diseased trees have repulsion. If diseased trees (N1) and healthy trees (N2) are independent stationary processes, then neither process should be affected by random shifts of the other within region R, which is the grove where myrtles grow. In Matlab, we generate 999 random shifts of this grove and compute the unadjusted k12 function for each random shift of healthy trees. In particular, we compare two point populations, diseased trees and healthy trees, in a bounding box by simulating 999 random shifts of healthy trees (with diseased trees fixed) and compute k12-values for a range of distance values. In this case, as we mentioned before, we have a set of 11 representative distances between 1m and 46m, which we used for random shifting. Then attraction p-values are computed, which is the estimated probability of obtaining a value as large as $\hat{K}_{12}^0(h_w)$ (observed sample cross K-function) under this spatial independence hypothesis.

The formula of attraction p-value is given by $\hat{P}_{attraction}(h_w) = \frac{M_+^0 + 1}{N+1}$ where M_+^0 denotes the number of simulated random shifts with $\hat{K}_{12}^m(h_w) \ge \hat{K}_{12}^0(h_w)$ and N is the number of random shifts. Here small values of attraction p-values can be interpreted as implying significant attraction between populations 1 and 2 at a random shift distance (scale) of h_w.

Similarly, we calculate the repulsion p-values, which is the estimated probability of obtaining a value as small as $\hat{K}_{12}^0(h_w)$ (observed sample cross K-function) under this spatial independence hypothesis. The formula of repulsion p-value is given by $\hat{P}_{repulsion}(h_w) = \frac{M_-^0 + 1}{N+1}$ where M_-^0 denotes the

number of simulated random shifts with and N is the $\hat{K}_{12}^{m}(h_w) \le \hat{K}_{12}^{0}(h_w)$number of random shifts. Here small values of repulsion p-values can be interpreted as implying significant repulsion between populations 1 and 2 at a random shift distance (scale) of h_w.

After running the Random Shift test in Matlab, we generated a plot of the p-values for the distances in the set of distances D based on 99 random shifts of healthy trees relative to the diseased trees. By convention, low p-values here correspond to significant attraction between patterns and high p-values indicate significant repulsion between patterns. From the plot below, we can observe that there appears to be some degree of attraction between the two populations at very small distances [D(2) = 3] and some degree of repulsion at relatively greater distance [D(7) = 25, D(8) = 30, D(9) = 35].

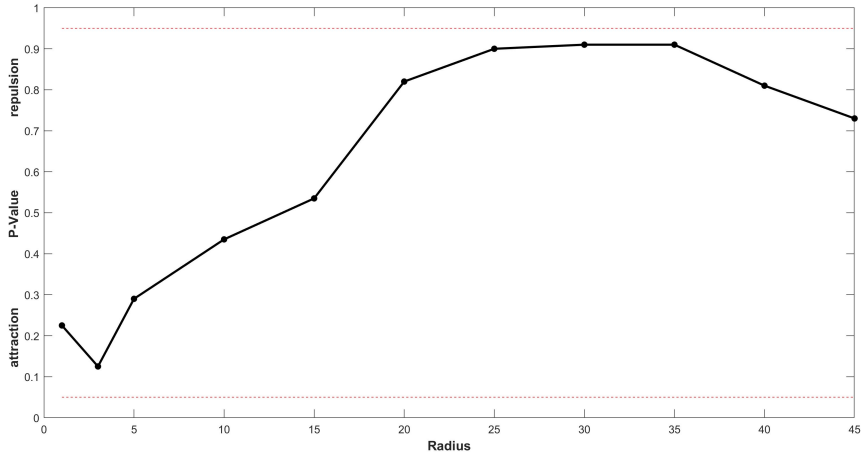

Figure 3: Plot of P-values Based on 999 Random Shifts in Random Shift Test

This result shows that in terms of attraction versus repulsion, we can not reject the null hypothesis that cell counts for healthy trees are not influenced by the locations of diseased trees. We also note that if h is almost 3m, there seems to be (barely significant) attraction between healthy and diseased trees. Therefore, all in all, we cannot reject the null independence hypothesis to support the contagion hypothesis.

Then we perform the random re-labeling approach to test for repulsion between the healthy trees and the diseased trees. In this context, we compare two point populations, healthy trees and diseased trees, by randomly permuting their labels and computing k12-values for a range D of distance values. This time we generate join patterns in a different way as we use a marked point process and generate (s_i, m_i) where $s_i \in R$is the location of event i in our grove and $m_i \in \{1,2\}$is a marker to denote whether a point is healthy tree or diseased tree. The key advantage of this approach is that it allows the location process to be separated from the distribution of event types. Then we want to test the spatial independence for population comparisons in the marked point process, which is to test whether event labels are influenced by their locations. We test on the Spatial Indistinguishability Hypothesis, which is for any observed tree location could be occupied by either health tree or diseased tree, the point processes generating populations 1 and 2 are essentially indistinguishable. Finally, in Matlab, we again compute the p-values, which reflect whether there is significant attraction or repulsion between these patterns and the resulting plot is as follows.

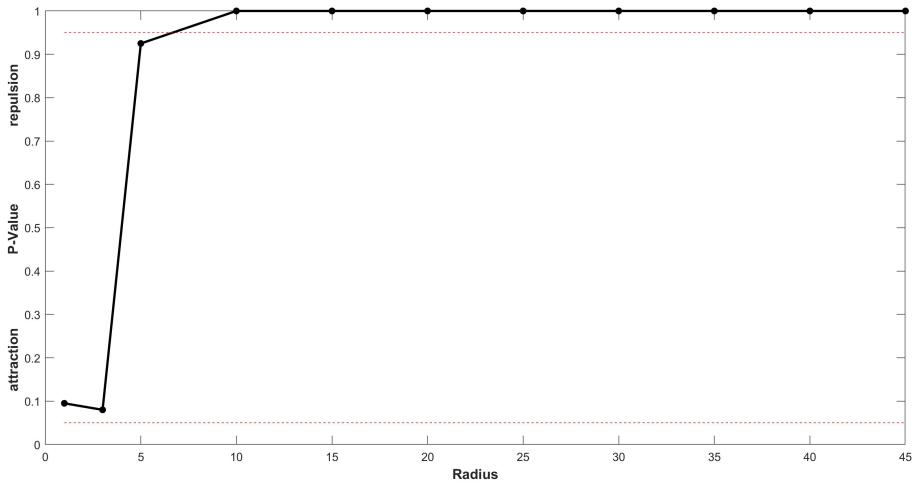

Figure 4: Plot of P-values in Random Permutation Test

In terms of p-values, the formula for attraction p-values and repulsion p-values in Random Permutation test are the same as the formula we used in Ransom Shift test. Now we let M_{+}^{0} denote the number of simulated random relabelings with $\hat{K}_{12}^{\tau}(h_w) \geq \hat{K}_{12}^{0}(h_w)$ where r is the number of random permutations. Then attraction p-value is the estimated probability of obtaining a value as large as $\hat{K}_{12}^{0}(h_w)$under this hypothesis of spatial indistinguishability. Similarly we let M_{-}^{0} denote the number of simulated random relabelings with $\hat{K}_{12}^{\tau}(h_w) \leq \hat{K}_{12}^{0}(h_w)$where r is the number of random permutations. Then repulsion p-value is the estimated probability of obtaining a value as small as $\hat{K}_{12}^{0}(h_w)$ under this spatial indistinguishability hypothesis.

From figure 4, we observe that we reject the null hypothesis of spatial indistinguishability as we have significant repulsion at distances radii larger than 10m. In terms of the contagion hypothesis, this means that healthy trees and diseased trees has significant repulsion when the distance radii is larger than 10m. This Random Re-labeling test gives us different results from the Random Shift test results because the re-labeling method is independent of boundaries, whereas the random shifts method is not. Another difference in the two tests is that Random Re-labeling test allows the location process to be separated from the distribution of event types.

We also noticed that both methods show that there may be significant attraction between diseased and healthy trees when h is about 3 units. In order to check whether there is attraction between the two populations at h = 3, we calculated the percentage of healthy trees from all trees that are in the neighbor of diseased trees at h = 3. As we have 15 healthy trees in all trees neighboring diseased trees at h = 3, the percentage of healthy trees from all neighboring trees is calculates by 15/21 * 100% = 71.42%. If the distribution were random, we would expect about 69/(49+69) * 100% = 58.5% of these neighbors to be healthy trees. Here we arrived at a higher percentage of healthy trees within all neighboring trees, which supports the finding that there is certain degree of attraction between healthy trees and diseased trees at h = 3.

Finally, we also consider possible spatial relations within the diseased population relative to the healthy population by means of random permutations.

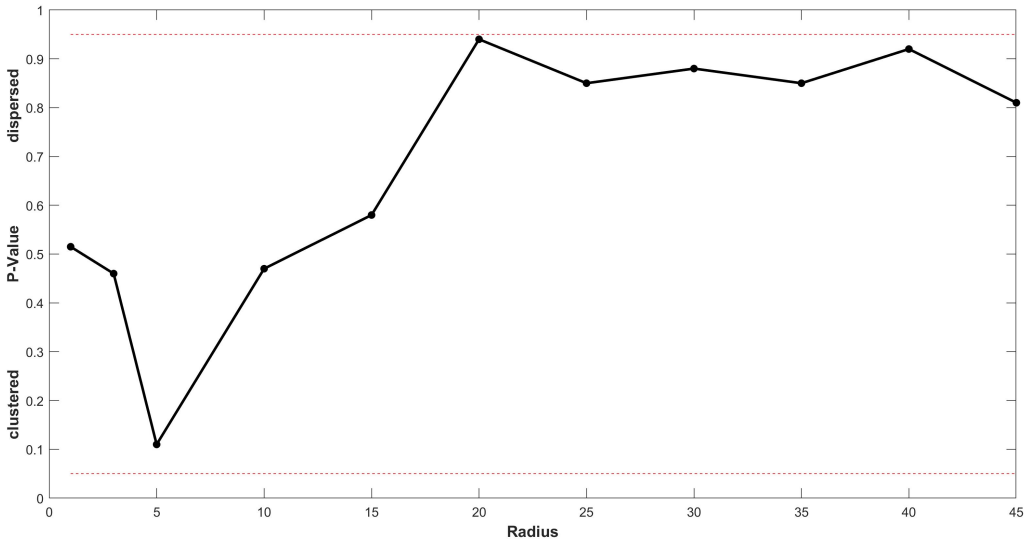

Figure 5: Plot of P-values in Random Permutation Test

If we find clustering within the diseased population relative to the healthy population of myrtles, this will give us additional support to the contagion hypothesis. Rather than characterizing the relative clustering as repulsion between the two populations, we might simply ask whether the pattern of diseased trees is more clustered than the pattern of healthy trees. Hence we want to know how similar are the spatial point patterns of healthy trees and diseased trees. Here we want to measure the "complete similarity" between two populations which means both point patterns are generated by the same spatial point process. If the healthy pattern isn't significantly different from diseased pattern, then the healthy and diseased trees might be samples from the same point process.In that case, the K-functions of both patterns should be similar. Therefore, our null hypothesis is to treat the combined process as a marked point process and both healthy trees and diseased trees come from the same point process, $H_0 : K_1(h) = K_2(h)$.

If we have $K_{diseased}(h) > K_{healthy}(h)$, then we say that diseased trees are significantly more clustered than healthy trees at radius h. If we have$K_{diseased}(h) < K_{healthy}(h)$, then we say diseased trees are significantly more dispersed than healthy trees at radius h. Now we use the 999 random permutations again and compute the p-values. The relative clustering p-value is the probability of obtaining a value as large as$\Delta^0(h) = \hat{K}_1^0(h) - \hat{K}_2^0(h)$ under this hypothesis of complete similarity. It is calculated as $\hat{P}^{12}_{r\text{-}clustered}(h) = \frac{m_+^0 + 1}{N+1}$ where m_+^0 denote the number of simulated random relabelings with $\Delta^r(h_w) \le \Delta^0(h_w)$. Similarly, the relative dispersion p-value for population 1 versus population 2 is the probability of obtaining a value as small as $\Delta^0(h) = \hat{K}_1^0(h) - \hat{K}_2^0(h)$ under the complete similarity hypothesis. The relative dispersion p-value is calculated as $\hat{P}^{12}_{r\text{-}dispersed}(h) = \frac{m_-^0 + 1}{N+1}$ where m_-^0 denote the number of simulated random relabelings with $\Delta^r(h_w) \le \Delta^0(h_w)$.

Drawing upon the p-values plot in figure 5, we cannot reject the null hypothesis that the trees come from the same point process. It is interesting to note that at h = 5, we have some degree of relative clustering and at h = 20, we have some degree of relative dispersion. This confirms our previous findings that diseased trees are more clustered at smaller scales (h = 5) relative to healthy trees. This phenomenon is probably because of the root grafting, which was

evidenced as a major driver for local spread of myrtle wilt. Further, diseased trees are also relatively more dispersed at larger scales (h = 20).

Discussion and Conclusion

In summarizing our test results, the Random Permutation test supports the contagion hypothesis that we have significant repulsion between the healthy trees and the diseased trees at distances radii larger than 10m and there may be significant attraction between diseased and healthy trees when h is about 3 units. Intuitively, this attraction at a smaller scale may be due to root grafting, which agrees with Packman's dissertation that root grafting explained the clumped pattern of diseased trees on a scale of 2.5-14m and gave rise to patches of dead and diseased myrtles (Packman, 20). Then, our finding in testing the complete similarity hypothesis resonates with Elliott's survey in that she claimed that diseased trees were shown to be clumped and the degree of clumping is dependent on the nearest distances within sites.

Citation Page

Lösch, August, and William H. Woglom. *The Economics of Location*. UMI, 1990.

Kolars, John, and Peter Haggett. "Locational Analysis in Human Geography." *Economic Geography*, vol. 43, no. 3, 1967, p. 276., doi:10.2307/143300.

Tinkler, Keith J. "Statistical Analysis of Tectonic Patterns in Areal Volcanism: The Bunyaruguru Volcanic Field in West Uganda." *Journal of the International Association for Mathematical Geology*, vol. 3, no. 4, 1971, pp. 335–355.

Packham, J. M., et al. "Rate of Spread of Myrtle Wilt Disease in Undisturbed Tasmanian Rainforests." *Australian Forestry*, vol. 71, no. 1, 2008, pp. 64–69.