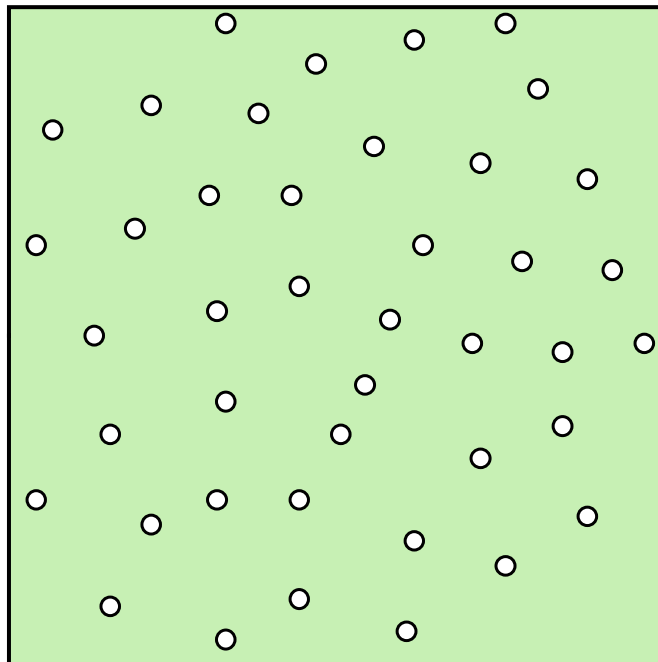


Assignment 2

(1) Cell Centers

1. Problem Statement. The goal of this problem is to analyze a data set that contains the locations of cell centers. Because of the highly structured nature of the cells, this data set is highly uniform. The purpose of analyzing this data set is therefore to verify the results of a test for uniformity. The locations of the cell centers can be seen in the map below.



2. Clark-Evans Test. The data was first analyzed using the Clark-Evans test. This test involves calculating the nearest-neighbor distances for each point, and analyzing them with respect to their distribution under the hypothesis of Complete Spatial Randomness (CSR). If the nearest-neighbor distances are particularly small, this is an indication of clustering. Large nearest-neighbor distances are an indication of uniformity.

Because nearest-neighbor distances have a tendency to come in pairs, the nearest-neighbor data is not independent. Therefore, a subset of the available points will be analyzed, to reduce artifacts of this dependency. In this set, out of the 42 available points, 21 will be analyzed.

If the m points are indeed randomly distributed, then the expected value of the average nearest neighbor distance (*sample mean*),

$$(1) D_m = 1/m \sum_{i=1}^m D_i$$

should depend only on the intensity of the points, λ . This expected value under CSR is given by:

$$(2) E(D_m) = 1 / 2\sqrt{\lambda}$$

Using JMPIN, we find that the theoretical mean for this data set is $E(D_m) = 7.715$

Again using JMPIN, we find that the observed sample mean is $d_m = 13.157$. Because the sample mean is so much higher than the theoretical mean, this is an indication that the data shows strong uniformity. This data can be more rigorously analyzed. First we will construct the standardized sample mean:

$$(3) z_m = (d_m - \mu) / \sigma$$

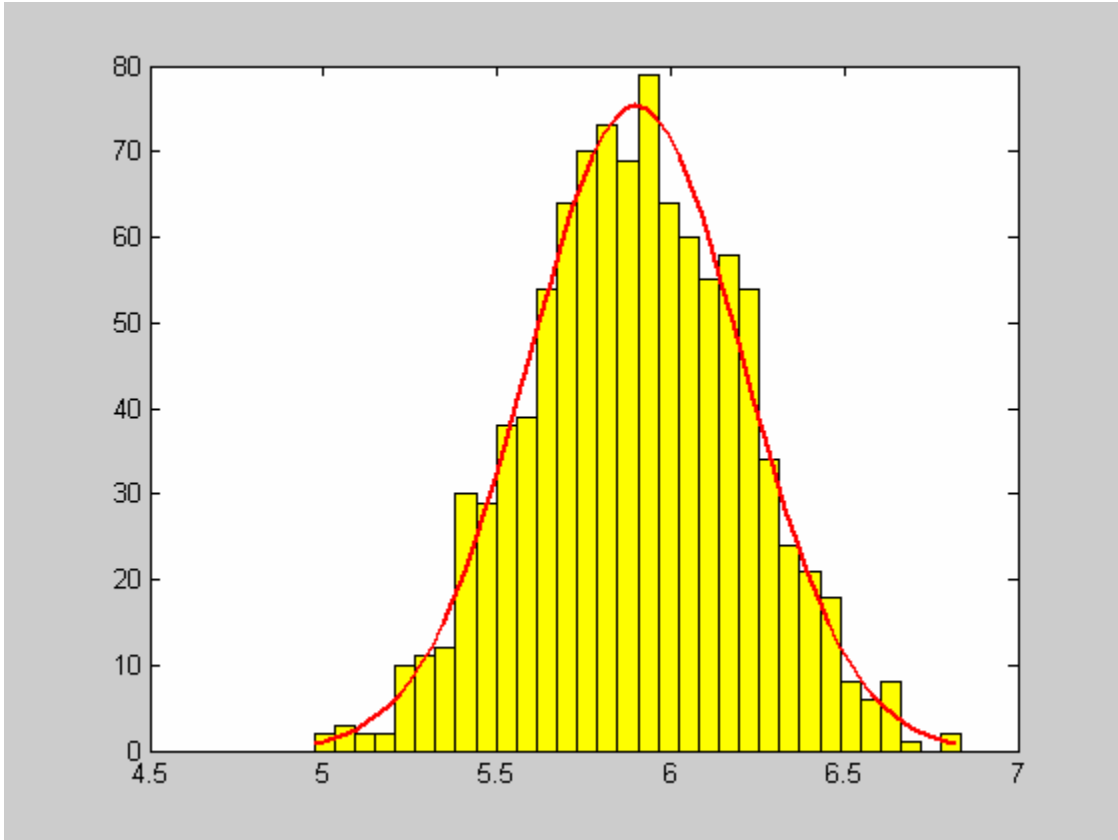
The Clark-Evans test then uses a normal approximation to the distribution of z_m under the CSR hypothesis. To determine the possibility of uniformity, we will compute the P-value using a one-tailed test. Under the Clark-Evans test, the P-value is the probability of observing a standardized value as large as z_m , i.e.,

$$(4) P(Z \geq z_m) = 1 - \Phi(z_m)$$

Again using JMPIN, we find the P-value for the one-tailed test is 0. This means that the probability that this data is completely random is very close to zero. It is likely that the P-value is too small for JMPIN to display, so we may not be seeing the exact value. In any case, this further supports our previous observation that this data is highly uniform.

This can be confirmed using the MATLAB program cluster.m, which repeats the same process. Using this program, we find that the Z-value is 5.516, and the P-value for a one-tailed test of uniformity is 1.734×10^{-8} , which confirms that our earlier observations were correct, and the data set is indeed uniform.

Finally, we will use the MATLAB program clust_distr.m, which analyzes the distribution of Z-values over several simulations. The mean \bar{Z} -value is 5.8994, and the P-value of the mean Z is 1.8244×10^{-9} , further confirming uniformity. The histogram of the values of Z is shown in the chart below.

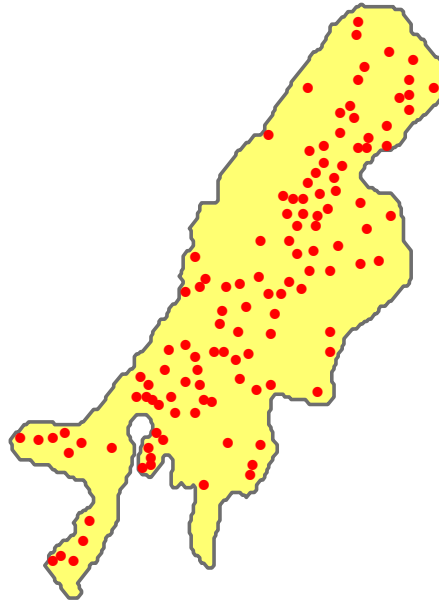


Histogram of Z-values

All the tests showed a high degree of uniformity in this data set. As was mentioned above, this was the expected result, because the centers of cells are spread in a uniform pattern due to the physical structure of the cells. This demonstrates that all the tests used here are accurate when used to determine uniformity.

(2) Ugandan Volcanoes

1. Problem Statement. The objective of this problem is to analyze the locations of volcanoes in a region of Uganda, and determine if they show uniformity or clustering at different scales. The map of the locations can be seen below:

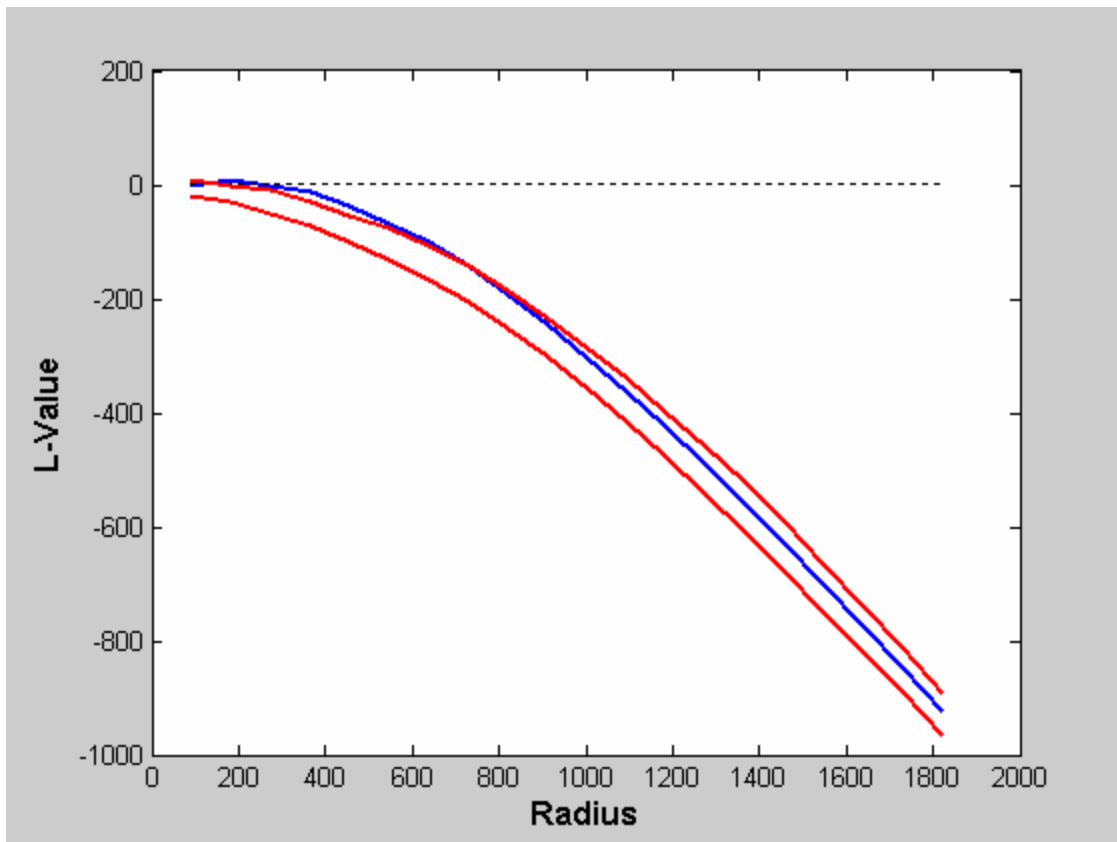


2. Simulation. The locations of volcanoes will be analyzed using Monte-Carlo simulations. Random locations will be selected for volcanoes, and the statistics of a number of trial random patterns will be compared to that of the actual pattern. The advantage of using simulations for such a data set is that artifacts such as edge effects will not be a problem, because anomalies caused by the boundary of the region will be accounted for in the random trials.

First, we will create simulation envelopes. The deviation from the null hypothesis of Complete Spatial Randomness (CSR), is calculated by the following formula:

$$(1) L(h) = \sqrt{(K(h)/\pi) - h}$$

Where $K(h)$, or the K-function, denotes the expected number of points inside a circle of radius h centered on a test point, divided by the overall point density. The values of $L(h)$ are calculated for 99 random pattern trials for several values of h , and the upper and lower values are the bounds of the envelope at each value of h . If the value of $L(h)$ for the actual data is above the envelope, we can conclude that there is relative clustering at scale h . If $L(h)$ is below the envelope, we can conclude relative sparseness at scale h . The envelope and actual values of $L(h)$ are shown in the chart below, which was generated using the MATLAB program `k_function_sim.m`.



Envelopes

From this chart, we can conclude relative clustering from a scale of about 200 to about 800 units. For all other values, $L(h)$ falls within the envelope. The P-value for this test is dependent on the number of trials, N :

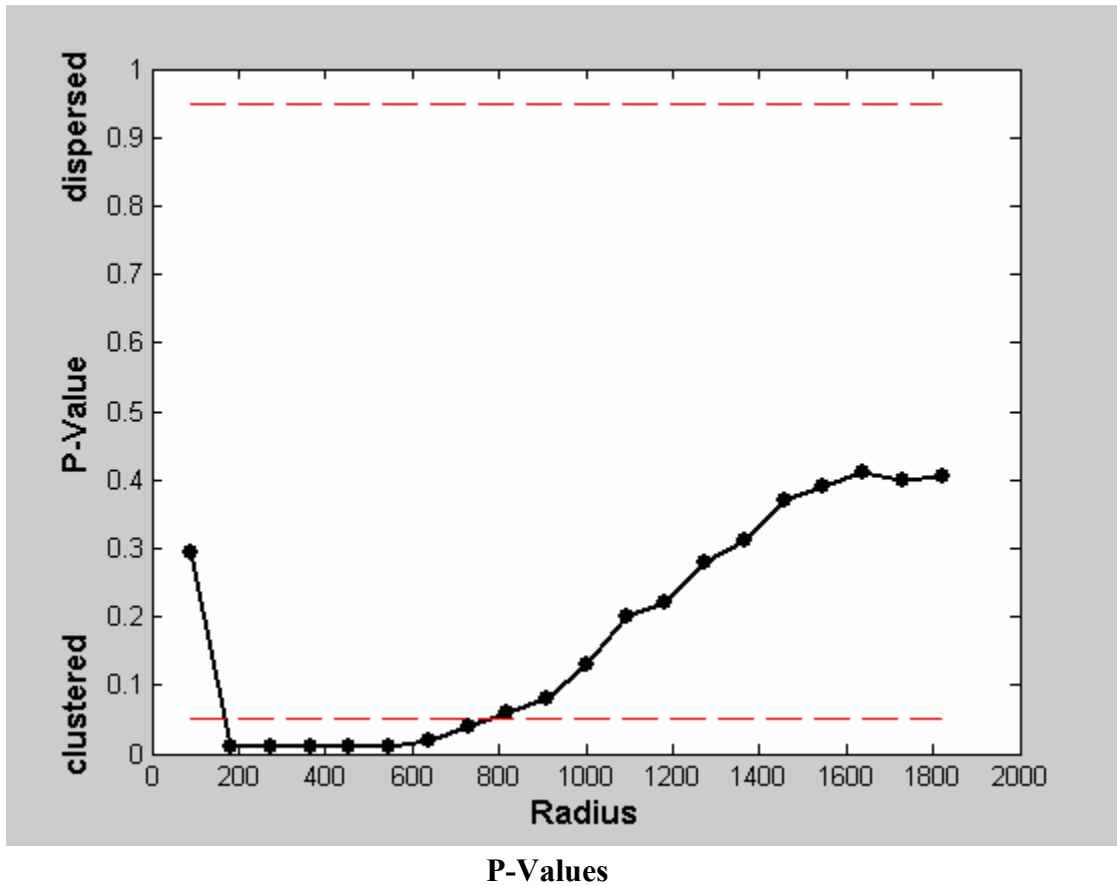
$$(2) P = 1/(N+1)$$

So in this case, we have $P = 1/100$. This means that there is clustering from 200-800, with a significance level of $\alpha = 0.01$.

We can also use Monte-Carlo procedures to find the P-values of the data set at different scales. We do this by again generating random patterns, and counting the number of pairwise distances in several “bins”, or intervals of radii. We then rank the counts for each random pattern for each interval, and define n_i to be the number of counts at least as large as the count for the actual pattern. The P-value is then defined as:

$$(3) P_i = n_i / (N+1)$$

This is calculated in MATLAB using the program `k_count_plot.m`, where we divide the radii into 20 bins, and perform 99 simulations. The chart generated by this process is shown below:

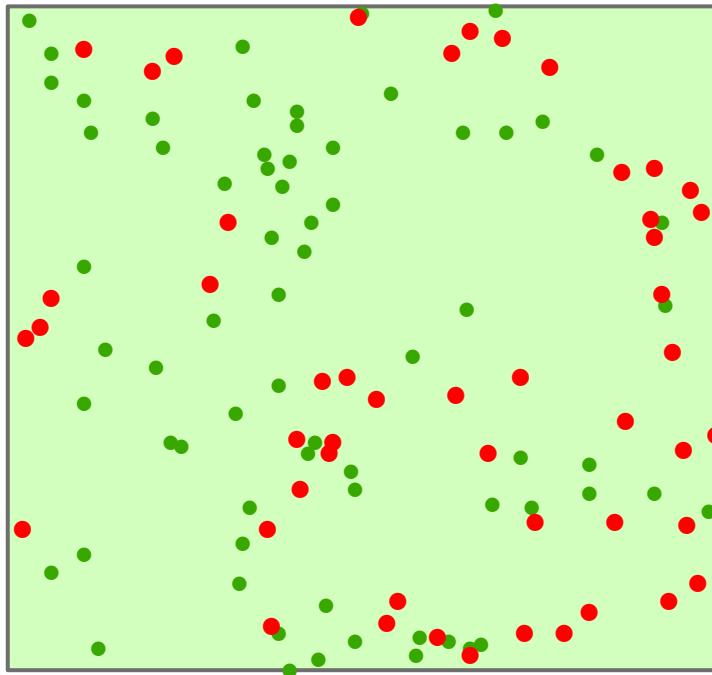


The analysis of P-values confirms our earlier observation in the envelope test that the pattern shows significant clustering from a radius of 200 to 800 units, but does not show clustering or uniformity elsewhere. Furthermore, we can see that the significance is very high from 200-600 units, at around 0.01, but it increases steadily from 600 to 800, at which point the P-value is about 0.05. We can therefore conclude that clustering is extremely significant at 200-600 units.

The evidence of clustering in volcanoes is not surprising. Volcanoes are products of geological processes, and occur at plate boundaries or other volcanic hot spots. A visual inspection of the map confirms that the volcanoes appear to occur along a line of sorts, which makes a great deal of sense if this map shows part of a tectonic plate boundary.

(3) Myrtle Disease

1. Problem Statement. The data set examined here contains spatial locations of healthy and diseased trees. The goal is to evaluate evidence of contagion of this particular disease. If contagion is present, we would expect to see diseased trees appearing in clusters, and some degree of “repulsion” between healthy and diseased trees. The null hypothesis here is that the two populations are independent. The map showing the locations of healthy and diseased trees is shown below:



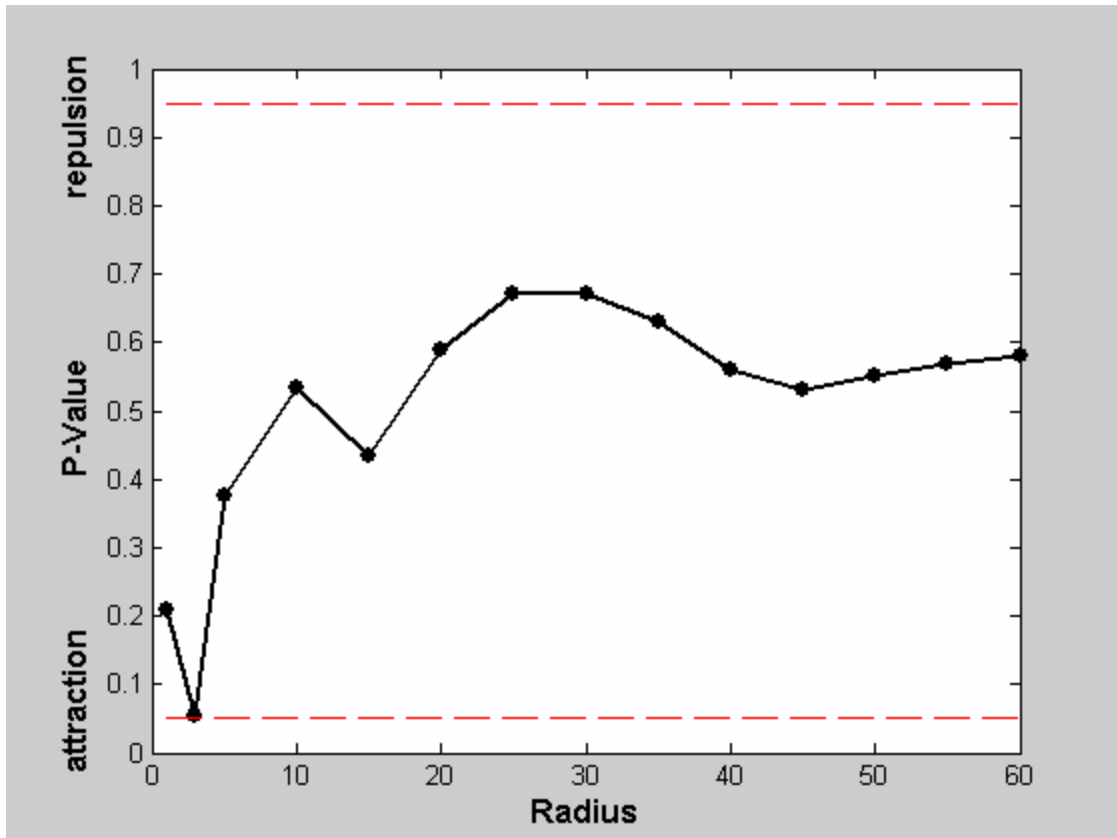
2. Random Shift. The theory of contagion can be tested using the random shift approach. If the locations of the healthy and diseased trees are independent stationary processes, the joint process should not be affected by random shifts in R . This is tested using the mathematical equivalent of wrapping the map around a torus, and randomly shifting it. The bivariate K -Function helps us analyze these shifts, given two populations i and j , with densities λ_1 and λ_2 .

$$(1) K_{12}(h) = 1/\lambda_2 * E(\text{ number of } j\text{-events } \leq h \text{ from an } i\text{-event})$$

We can determine the P -values at different radii, using the formula:

$$(2) P(h) = C_{12}(h)/(1 + N)$$

where $C_{12}(h)$ denotes the number of shifted patterns with higher $K_{12}(h)$ values than that of the observed pattern, and N denotes the number of simulations. This analysis is performed using the MATLAB program `k12_shift_plot.m`. The table generated using the myrtle data and 99 simulations are shown below.

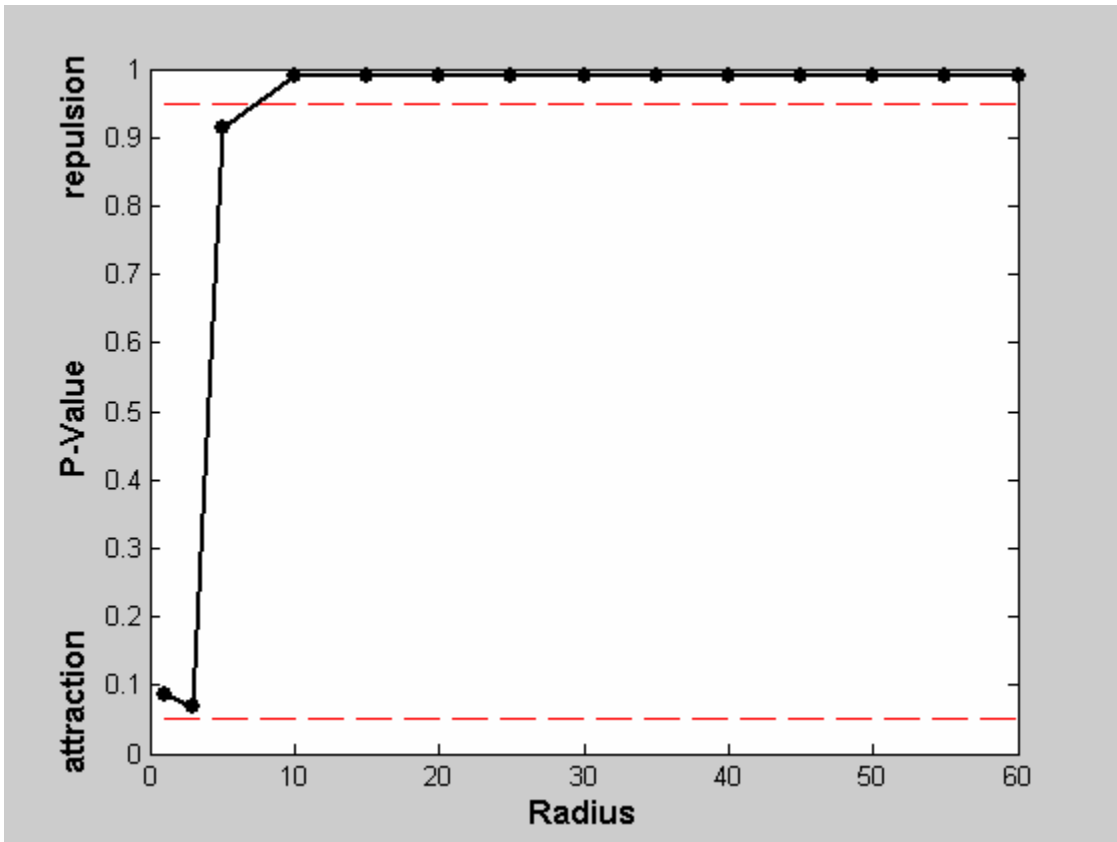


Random Shift Test

As we can see, this chart shows slight attraction at a radius of about 2.5 units, but otherwise not much attraction or repulsion. This indicates that the contagion hypothesis may not be true, and the two populations may be independent.

3. Random Permutations. This test involves randomly relabeling the points as healthy and diseased by permuting the labels over a number of trials. This is not testing exactly the same hypothesis as the random shifts test did, that the populations are independent, but is instead testing that the labels of diseased vs. healthy are randomly distributed. The P-values are calculated using equation (2). The MATLAB program `k12_perm_plot.m` was used to perform this analysis. The chart is shown on the following page.

The chart shows attraction at a radius of about 2.5 units, and significant repulsion between the two populations from a radius of 10 to 60 units. This contradicts the evidence from the random shifts test, and indicates that there is contagion of the myrtle disease. The differences between the two tests may be caused by artifacts created by the random shifts test, such as artificial clusters at the edges. Furthermore, as mentioned above, the two tests are looking at slightly different hypotheses.



Random Permutations Test

4. Attraction Between Populations. In both the Random Shifts and Random Relabelling tests, there was evidence of some attraction at very small distances. To further examine this phenomenon, all diseased trees with neighbors within 3 units were examined. The table on the following page shows the number on neighbors each tree has that is within 3 units, and also the number of healthy neighbors within 3 units. If the distributions were indeed random, we would expect the percentage of healthy trees in this sub-population to be the same as the percentage of overall healthy trees. Overall, 58.5% of the trees are healthy. However, of in this population, 15/21, or 71.4% are healthy. This discrepancy between the expected healthy percentage and the actual healthy percentage is consistent with the evidence of small-scale attraction in the previous tests.

Diseased Tree Number	Neighbors Within 3 Units	Healthy Neighbors
8	2	2
13	1	1
14	1	1
19	2	1
20	3	2
22	1	0
23	1	0
26	2	2
27	2	1
30	2	1
34	1	1
40	2	2
43	1	1
total	21	15

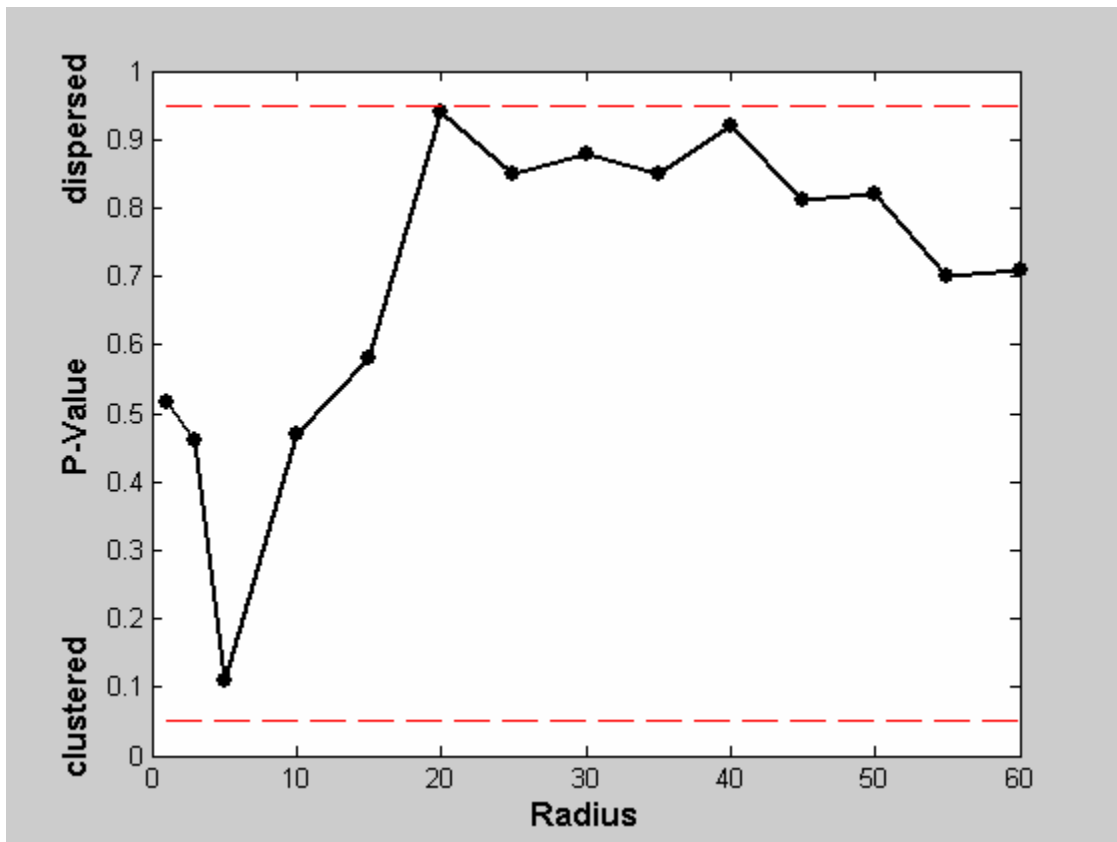
5. Spatial Relations Within Diseased Population. Finally, we will examine similarities between the populations of diseased and healthy trees. If the contagion hypothesis is correct, we would expect diseased trees to be clustered relative to the overall population. The null hypothesis here is that the two populations are drawn from the same point process, and therefore their K-functions should be identical. We will test this by again randomly permuting the labels. We find the difference between the K-functions,

$$(3) D(h) = K_1(h) - K_2(h)$$

and define $C_1(h)$ to be the number of times $D(h)$ of the random populations is at least as large as that of the actual population. We can therefore find the P-value:

$$(4) P(h) = C_1(h) / (N + 1)$$

Where N is the number of simulations. This analysis was performed on the myrtle data using the MATLAB program `k2_diff_plot`, and the chart is shown below:



Difference of Diseased Population and Healthy Population

Although this graph does not show highly significant clustering or dispersion, we can see that there is a general trend for clustering at small distances, and dispersion at large distances. We will therefore conclude that there is a small amount of clustering of diseased trees relative to the amount of clustering of healthy trees.

Overall, the evidence for contagion of the myrtle disease is uncertain. The random shifts test did not show overall repulsion, although the random permutation test did. The evidence of attraction between healthy and diseased trees at small distances contradicts the contagion hypothesis; we would expect repulsion, especially at small distances. There is some evidence of diseased trees clustering, but it is not particularly strong. The tests certainly indicated that there were more than random processes occurring, but contagion as was defined above was not strongly substantiated.