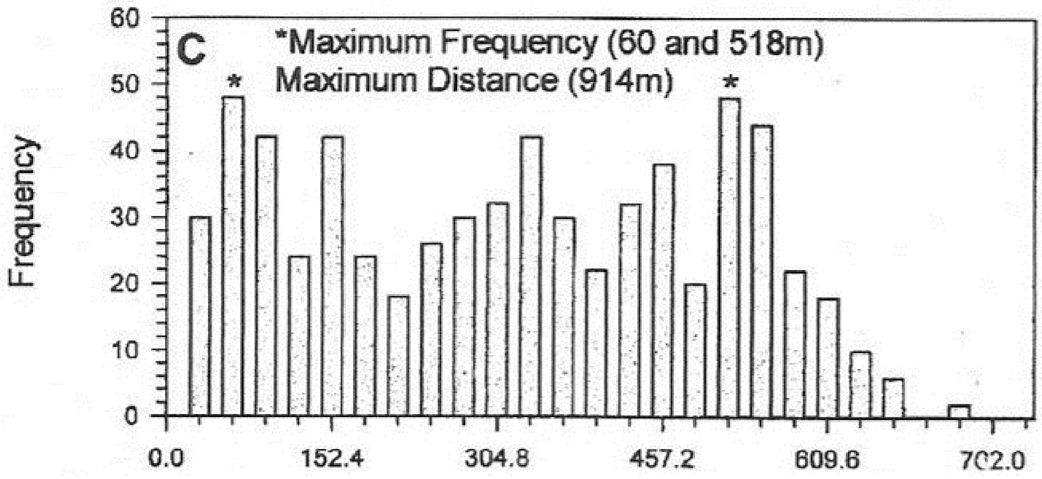
APPENDIX D1: IPD SUPPLEMENTAL INFORMATION



Site D3, Frequency Counts of inter-Tree Distances

INTRODUCTION

This appendix examines four topics:

(1) Verification of IPD Methodology
(2) Site's geometry as related to Inter-Point Distances
(3) Simulation results using CSR pattern
(4) Probability interpretation of IPD analysis

In general, the discussion in this appendix does not alter the review in Appendix D nor any of the conclusions reached. In Appendix D, it was concluded that the maximum frequency or maximum distances are not an "upper bound" of distances of spread, but are related to the geometry and area of the site. If the same analysis was conducted of all infected trees in Florida, the upper bound of distances of spread as interpreted by these two values would be in terms of hundreds of miles.

Wherever possible, this review attempts to verify results are consistent with the stated methodology. The frequency counts in Site D3, as shown above were examined for consistency with the IPD methodology.

DISCUSSION

(1) VERIFICATION OF IPD METHODOLOGY

It was easy to identify the frequency counts for Site D3. Frequency counts for Site D3 is shown below, The figure digitized and the frequency counts were read off the figure as shown in Table 1. A total of 648 frequency count result.

Figure 1: Frequency counts for Site D3

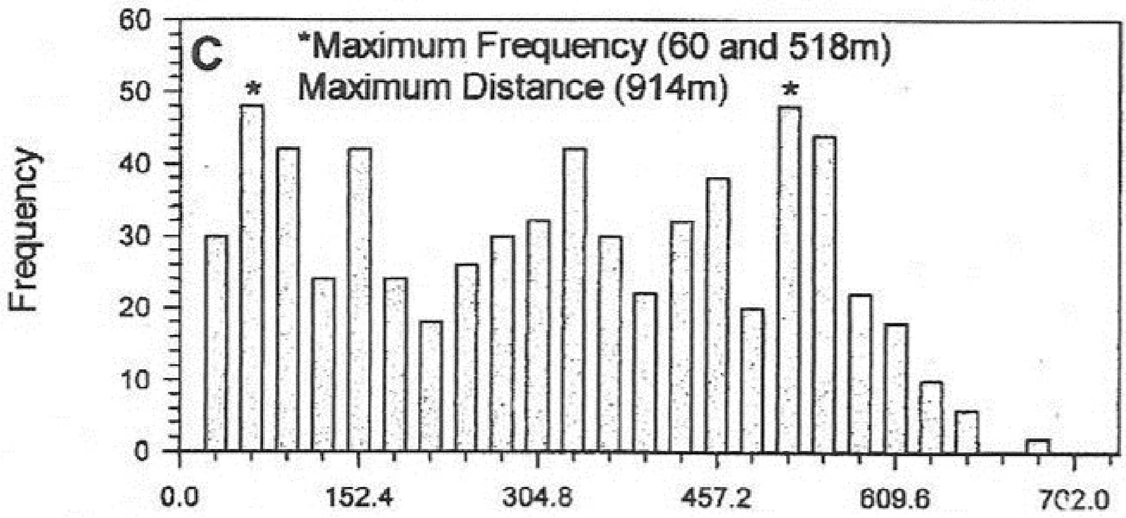


Table 1: Frequency counts for Site D3

Bin #	Distance (m)	Freq counts	Bin #	Distance (m)	Freq counts
1	30.48	30	13	396.24	22
2	60.96	48	14	426.72	32
3	91.44	42	15	457.20	38
4	121.92	24	16	487.68	20
5	152.40	42	17	518.16	48
6	182.88	24	18	548.64	42
7	213.36	18	19	579.12	22
8	243.84	26	20	609.60	18
9	274.32	30	21	640.08	10
10	304.80	32	22	670.56	6
11	335.28	42	23	701.04	0
12	365.76	30	24	731.52	2
			25	762.00	0
					648

The authors state, "A third VBA was used to calculate the distance from focal trees to all possible **subsequent** infected trees." This was interpreted to mean that each NI tree within a time period was associated with all PI trees, and the respective distances calculated. This is shown in Table 2.

Table 2: Inter-tree distances for Site D3 as calculated = 274 distances

Period	# NI Tree	# PI Tree	# Distances
1	2	1	2
2	3	1	3
4	4	1	4
5	5	2	10
12	7	1	7
13	8	6	48
16	14	4	56
17	18	8	144
Total			274

There should be 274 distances in the frequency diagram, and instea there are 648 distances, 2.3 times as many.

Alternatively, the number of distances could be calculated as $n(n-1)/2$ where n is the number of trees at the end of the study. This results in 26 (25)/2 = 325. This will always provide a larger estimate than calculating distances for each time period. Thus, this is the upper limit estimate. Using this upper limit estimate, the ratio of observed to calculated is (648/350) = 1.99.

Other Sites

For the other sites, cumulative frequency counts were estimated by approximating the frequency distributions with straight lines, resulting in triangles. The bin size is 30.5-m, so an estimate of the number of bins was possible, rounded to the next higher integer.

All frequency counts are given in terms of thousands (000's).

Table 4: Comparison of Methods in thousands of distances

	(1) Counts as Estimated from Graph	(2) Upper Limit Counts *	Error ratio (1)/(2)
Site D1	2425	1544	1.57
Site D2	1022	573	1.88
Site B1	159	101	1.57
Site B2	49	26	1.88

* $n(n-1)/2$ where n is the number of infected trees.

It is very difficult to know why there is this discrepancy. Including site D3, the average error ratio is 1.8. It is possible, that the calculation considered the distance from every infected tree to every other one, and included duplicates. The words "subsequent infected trees" would indicate that the distance should be calculated as shown in Table 2, considering the parsing of trees into time periods.

Secondly, it is odd that the authors would include duplicate distances. The authors seem aware that duplicate distances would result if the full matrix of distances are calculated, and correctly include '2' in equations 1 and 2 on page 364. There is even an explanation for this on page 364. It is in the numerator because these equations are calculating the frequency counts at specified distances divided by the total population of infected trees.

In conclusion, it is a mystery why there is this seeming inconsistency between procedure and results. In general, perhaps it is very consequential as the method is not valid for determining the distances of spread as discussed in Appendix D. In the future, if there is a plausible explanation, I will post it on the website.

(2) SITE'S GEOMETRY AS RELATED TO INTER-POINT DISTANCES

- MAXIMUM DISTANCE

Points widely dispersed in large areas will result in many long distances. The maximum possible distance is equal to the diagonal distance of the site. Site D3 is excluded from analysis due to the difficulty in identifying the site's actual boundaries as discussed in Appendix A. Sites B1 and B2 appear to have irregular boundaries, so the aspect ratio is calculated based on the maximum dimensions.

The ratio of the maximum distance to the diagonal distance (range of frequency graph) varies from 0.64 to 0.96.

Table 2: Ratio of max distance to diagonal distance

	(1) Maximum Distance (m)	Area (sq mile)	(2) Diagonal Distance (m)	(1)/ (2) Ratio
Site D1	3444	2.0	3600	0.96
Site D2	2133	1.3	2640	0.81
Site B1	4754	3.0	4998	0.95
Site B2 (3)	1432	0.78	2231	0.64

(1) Maximum distance is obtained from the 2002 published article.

(2) Area, aspect ratio and diagonal distance are discussed in Appendix A.

(3) The geometry of Site B2 appears to be trapezoidal. The eastern boundary is approximately 0.96 miles, and the base is 1.0 miles, so W/L = 1/0.96 = 1.04.

- PEAK FREQUENCY VALUES

The peak frequency value has been normalized, denoted as M*, by dividing by the peak value by the diagonal distance. No real trend is apparent for M*. Simulation shows a trend towards lower M* values with increasing aspect ratios as shown in the next section.

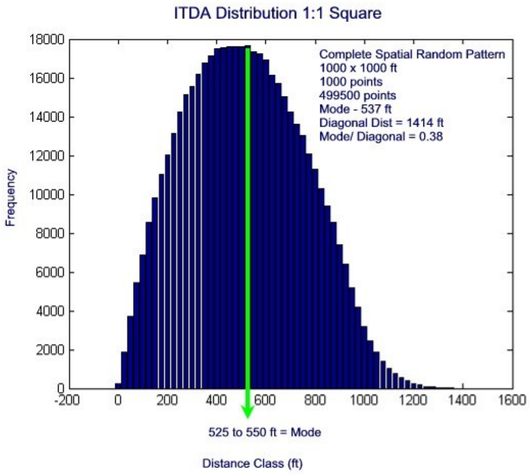
Table 3: Ratio of maximum frequency distance to the diagonal distance

	(1) Max Freq. Distance (m)	Aspect Ratio	(1)/(2) M*
Site D1	840	2.0	0.23
Site D2	457	1.3	0.17
Site B1	243	2.8	0.08
Site B2	335	1.0	0.15

(3) SIMULATION RESULTS

A CSR pattern was generated with MATLAB with varying aspect ratios. A mode to diagonal distance was calculated as M*. As the aspect ratio increase from a 1:1 geometry to a rectangle with a 1:10 ratio, M* declined as shown in Figures 2 to 4, and in Table 8.

Figure 2: Square geometry, M* = 0.38



The range of the distribution is from 0 to the diagonal distance of the site. For a square site, this would be 1.414 x L, where L is the length of one side. The mode divided by the range of the distribution, will be denoted as M*. The mode is estimated as 537 ft, midpoint of the distance bin 525- 550 ft, resulting in M* = 0.38.

Figure 3: 1:2 Rectangle, M* = 0.25

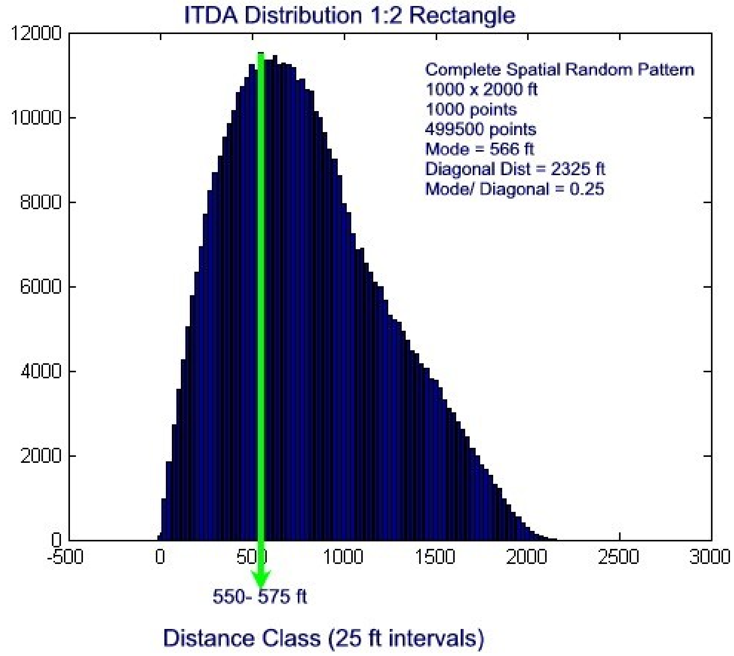


Figure 4: 1:6 Rectangle, M* = 0.12

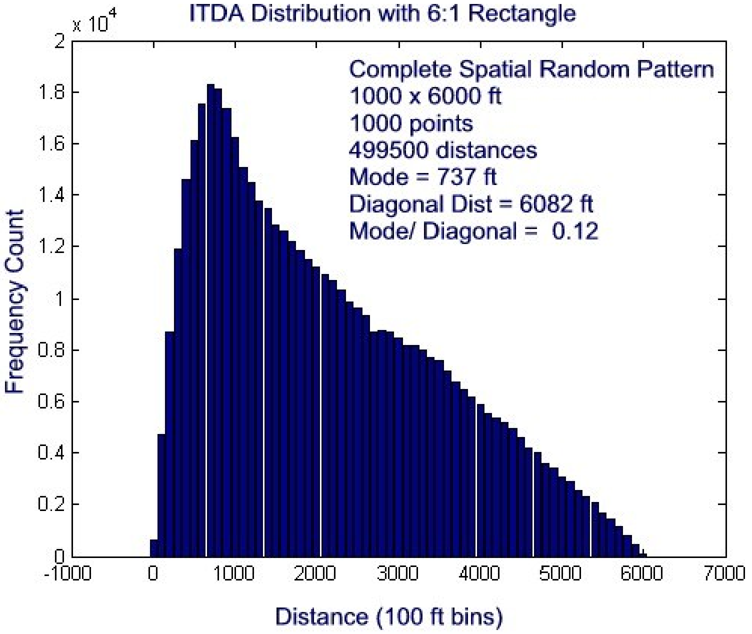


Table 4: Normalized modes (M*) with respect to aspect ratios for CSR patterns.

Aspect Ratio	M*
1:1	0.34
1:2	0.25
1:3	0.21
1:4	0.18
1:5	0.15
1:6	0.12
1:10	0.075

All results above for aspect ratios of 1:2 and higher are from simulation, so the results are approximate. The result for the unit square case was from Barlett's equation as noted in the supporting information section.

The M* as calculated from CSR patterns would not be expected to be the same as calculated for the study sites. The trees within the site are likely not represented by a CSR pattern, because owners tend to plant their trees in their backyards. However, the calculated values of Site D1, D2 and B2 (0.23, 0.17 and 0.15) are similar to the simulated values of 0.34 for a 1:1 square and 0.25 for a 1:2 rectangle area.

(4) INTEPRETATION OF IPD

Using stochastic simulation approach, a CSR or other known patterns within a specific area can be generated and IPD distances calculated. After a large number of replications, a cumulative distribution, *F(d)* can be obtained*,* where *d* is the distance between points can be generated. The shape of *F(d)* would depend only on the geometry of the area. Simulation is used as a more convenient manner to generated *F(d)* for any area, as opposed to an analytical solution, where the mathematics is complex. The distribution provides estimates of the probability, $P\{X \leq d\} = p$, where *X* is the event of selecting two points at random within the site, and *p* is the probability that the distance between the two random points is less than the value *d*. The basis for constructing the distribution is that each of the generated distances from random deviates is equally probable. Accuracy of estimates can always be improved through additional replications.

This interpretation does not seem in any way connected to the presumed challenge of finding an appropriate eradication radius.

A recent textbook on spatial point pattern by Illian et al [3] identifies the "distance method" as one of four methods to test the hypothesis that a spatial arrangement is representative of a complete spatial randomness pattern.

Illian writes, “If the CSR hypothesis is rejected, then the more interesting part of point process statistics begins, in particular the search for spatial correlation in the given pattern.”

Diggle [4] discusses IPD method (he calls it inter-event distances) and expresses similar thoughts in the negative, “... a pattern for which CSR is not rejected scarcely merits any further formal statistical analysis.” Textbooks by Diggle and Illian see hypothesis testing as an exploratory first step in spatial pattern analysis.

In Appendix F, the published article suggested that the distribution of infected trees could be tested against the hypothesis of a CSR pattern, using the Kolmogorov-Smirnov test, although no results were shown. Diggle cites a study in 1977 (page 14) which used a chi-square approach to evaluate the “goodness of fit” between pattern data and theoretical distribution of CSR, based on inter-point distances.

SUMMARY

The supplemental information in this appendix does not change the conclusions reached in Appendix F.

1. Unexplained inconsistencies exist in frequency counts in all sites in comparison with the expected number of distances. On average, there are 1.8 times more distances in the frequency plots in the published article than as calculated.

2. The maximum distance as given in the frequency distributions is approximately 64 to 96% of the diagonal distance of the site. This conclusion is consistent with Appendix F, that for widely dispersed incidences of disease, the maximum distance should be a function of the size and geometry of the site.

3. A ratio of maximum frequency to diagonal distance (M*) was calculated for each site and compared to simulation values. The results should show differences, as infected trees are unlikely to be represented by a CSR pattern. However, similar results were shown for sites D1, D2 and B2.

4. A probability distribution of a specified pattern determined from simulation. IPD may be used for testing the hypothesis that a spatial arrangement is represented by a complete spatial randomness pattern. This interpretation does not seem in anyway connected to the presumed challenge of finding an appropriate eradication radius.

REFERENCES

1. Gottwald, T.R., X. Sun, Riley, T. Graham, J.H., Ferrandino, F. and Taylor, E., (2002), Geo-Referenced Spatiotemporal Analysis of the Urban Citrus Canker Epidemic in Florida, Phytopathology, Vol 92, No. 4.
2. Diggle, P.J., (2003), Statistical Point Patterns, Oxford University Press, NY, ISBN 0 340 74070 1 (bb).
3. Illian, J. Penttinen, A., Stoyan, H., and Stoyan, D. , (2008). Statistical Analysis and Modelling of Spatial Point Patterns, John Wiley, Chichester, ISBN 978-0-470-01491-2.
4. Diggle, P.J., 2003. Statistical Analysis of Spatial Point Patterns, Oxford University Press. IBSN 0-340-74070-1.

SUPPORTING INFORMATION

Simulation Code

Further evaluation of the IPD distribution shapes was made using Monte-Carlo simulation with 1000 points were randomly generated in a square, 1000 ft on a side, with the bin size of 20 ft. The maximum diagonal distance 1414 ft. The simulation was done with the Matlab program. A set of points representative of a completely spatial random (CSR) pattern is generating random number 0 to 1000 ft for both the x and y coordinates. The sides are 1000 units per side, and the number of points is equal to 100. The notation on plots was done with imaging software.

The program listing is given below:

Figure 2: Matlab code

```
clear;
xside = 1000;
yside  = 1000;
bin_size = 25;
max_dist = 1500;
np = 100;
x = rand(1,np)*xside;
y = rand(1,np)*yside;
diag = (xside^2 + yside^2)^0.5
kk = 0;
n = 0:bin_size:max_dist;
for k = 1:np;
    ks = k + 1;
    for k1 = ks:np;
      kk = kk+1;
      d(kk) = ((x(k)-x(k1))^2+(y(k)-y(k1))^2)^0.5;
    end
end
xout=hist(d,n);
bar(n,xout)
```

Mathematical Determination of Mode for Unit Square

The mode can also be determined using mathematics for a unit square. The cumulative distribution of distances (Diggle, 2003) is:

$$F(x) = \begin{cases} \pi x^2 - 8x^3/3 + x^4/2 & 0 \le x \le 1 \\ 1/3 - 2x^2 - x^4/2 + 4(x^4-1)^{1/2}(2x^2+1)/3) \\ \quad + 2x^2 sin^{-1}(2x^{-2}-1) & 1 < x \le \sqrt{2} \end{cases}$$

The mode is defined as $df(x)/dx = 0$, where *f(x)* is the probability density function of *F(x).*

Considering the mode is less than x = 1 based on simulation, only *F(x)* in the domain of $0 \le x \le 1$ needs to be considered. Therefore, for $0 \le x \le 1$,

$$f(x) = 2\pi x - 8x^2 + 2x^3$$

$$f'(x) = 2\pi - 16x + 6x^2$$

Setting *f'(x) = 0*, and using the quadratic formula to solve for x, x = 0.48. This is for a unit square. M* = 0.48/1.414 = 0.34, in relatively close agreement with the simulation results.