APPENDIX F1:

ADDITIONAL COMMENTS ON SPATIAL POINT PATTERN ANALYSIS

1. INTRODUCTION

This appendix supplements the review provided in Appendix F, in the theory of inter-point distance analysis. It is focused on mathematical analysis and does not change the conclusions reached in Appendix F in terms of validity of the results. Two topics are discussed:

1. Equations 1 to 5 on Page 364 of published article
2. Confidence interval of a CSR pattern

The presentation of equations 1 to 5 on page 364 is difficult to follow. The subheading for this section is "Evaluation of the spatial point pattern by modified Ripley's K-function method. If we skip over these five equations, and read on, it is learned that these equations represent a new outgrowth to the method presented by Ward and Ferrandino in reference 15. Figure 4 might look like a Ripley K-function analysis, but there is no definition for Kexp and K'obs as presented in Figure 4 on page 372. The y-axis is labeled, K(Di) but no where has variable Di been defined. The only link between Figure 4 to these five equations is the note accompanying Figure 4 that "Calculations are described in equations 1 to 5." So, it is a confusing presentation. The article states this is an extension of Ward and Ferrandino's article, but this is incorrect.

If the discussion is read carefully, the article states that the equivalence of this analysis to Ripley K-function occurs when there are infinite number of trees uniformly covering the site (CSR pattern). Obviously, epidemiology problems are unlikely to find "infinite trees" within a study site.

Continuing on with page 364 in the first new paragraph, there certainly an impression that a Kolmogorov-Smirnov (K-S) test was to be conducted between two empirical distribution as given in equations 1 and 2. But mysteriously no results were presented. It was obviously a good question, where the empirical distribution formed by all citrus trees (equation 1) is statistically different from the one which included only infected trees.

So, this supplemental appendix begins by what the equations are not: (a) an extension of Ripley K-function and (b) procedures used to conduct a K-S test.

But, were these equations used to define the confidence intervals of the Kexp curve in Figure 4? If not, then why are they presented as part of the spatial point procedure.

2. Inter-Point Distances, Equations 1 and 2

As discussed in Appendix F, if there are *N* points, and they are interconnected by lines, this results in N(N-1)/2 unique distances. In equation 1, and empirical cumulative distribution of inter-point distances of all trees is created using quantiles. Thus, distances from infected to healthy tree, healthy to healthy tree and infected to infected trees are part of this distribution. The random variable, for the discussion here, will be denoted as *D*. Thus, equation 1 provides estimates for the probability of any randomly selected distance from the collection of calculated distances, will be less or equal to d is cdf_T(d). Similarly, equation 2 is the empirical cumulative distribution of inter-point distances of all infected trees, with *I(I-1)/2* unique distances.

To simplify the notation, the number of distances as calculated from a set of trees will be denoted as A if all trees are used and B if only infected trees are used.

3. Equations 2 to 5, Page 364

Equation 3 is explained on page 364 of published article as follows :

> For a particular distance (*d*), the probability of selecting infected pairs in a sample of *N*(*N* – 1)/2 $cdf_T(d)$ tree pairs chosen randomly from a population of size *N*(*N* – 1)/2 of which *I*(*I* – 1)/2 are infected is given by the hyper- geometric distribution: ...

Equation 3 which follows is the hyper-geometric distribution. The statement is in the form of a sampling problem of a discrete variable. The distribution as given in equation 1 can easily be approximately transformed into a discrete variable distribution (probability mass function) of distance d = {1, 2, 3 ... } . So, the distribution of random variable *D*, representing discrete inter-point distances of all trees, is a probability mass function, *P(D = d)*, which can be expressed as *P(d).*

As with sampling problems, we have the parent (or collective) and the sample population. For a specified value of d, $N_a \cdot P(d)$ would be the parent number of all tree pairs with distances equal to the value *d,* as correctly stated in the above excerpt.

But as we examine the equations 3 and 4, the cumulative distribution for infected trees as given in equation 2 is not included. It is presumed that the infected trees are considered to be randomly located within the population of all trees. If the overall population is considered to be a CSR pattern, the trees randomly assigned as infected, would in theory, have the same pattern, hence the distribution related to inter-point distances would be the same.

Thus, equations 3- 5 in the published paper form a sampling problem where: (1) All trees are of a CSR pattern and (2) Infected trees are randomly assigned within this pattern. As will later be reviewed, this analysis seems extraneous to the determination of confidence intervals for the Kobs curve in Figure 4.

The parameters of the hypergeometric distribution are correctly stated in equation 3 of the published article. It is given below in an expanded form.

$$P\{X=i\} \;=\; f(i|n,A,B) \;=\; \frac{\binom{B}{i}\binom{A-B}{n-i}}{\binom{A}{n}}$$

Or P = Hyper(i, A, B, n)

The parameters are defined as follows:

n = Sample size of distances.

A = All distances calculated from healthy and infected trees

B = All distances calculated from infected trees only .

i = Number of distances within the random sample based on infected tree pairs only.

Sample size is equal to $A \cdot F(d)$ The cumulative distribution is denoted as *F(d)* instead of $cdf_T(d)$ for convenience.

Distribution mean and variance are calculated as follows:

$$\mu \;=\; n \cdot B/A \;=\; B \cdot F(d)$$

$$\sigma^2 \;=\; n\left(\frac{B}{A}\right)\frac{(A-B)}{A}\frac{(A-n)}{A-1}$$

Substituting $n = A \cdot F$(d) and grouping/ canceling like terms*:*

$$\sigma^2 = \frac{[BA \cdot P(d) - B^2 \cdot F(d)]}{(A-1)}\,(1-F(d))$$

Since $\mu \;=\; B \cdot F(d),$ based on the assumptions as discussed above, the expression becomes:

$$\sigma^2 \ = B \cdot F(d)\left[1-\frac{B}{A}\right](1-F(d))\left[\frac{A}{(A-1)}\right]$$

For variance, a slightly different equation is presented in the published article. The equation presented in the article, Equation 4 is:

$$\sigma_{exp}^2 = \ i_{exp}^2[1-cdf_\tau(d)]\left(1-\frac{I(I-1)}{N(N-1)}\right)$$

with mean value,

$$i_{exp} = \frac{I(I-1)\cdot cdf_T(d)}{2}$$

In attempting to reconcile these differences, our equation based on the definition of variance with the notation used in the published article is:

$$\sigma^2 = \ i_{exp}[1-cdf_T(d)]\left(1-\frac{I(I-1)}{N(N-1)}\right)\left(\frac{N(N-1)}{N(N-1)-1}\right)$$

The last term is approximately one for large *N*, yielding:

$$\sigma^2 = \ i_{exp}[1-cdf_T(d)]\left(1-\frac{I(I-1)}{N(N-1)}\right)$$

This the exponent of i_{exp} should be 1 and not 2 as given in the published article. This could easily be a typographical error.

The authors state the normal distribution can be used to approximate the calculation of the hypo-geometric. The conditions for this approximation appear to be satisfied.

The following example is used to further explain equations 1 to 5 as provided in the published article.

4. Example Problem

A total of 1000 trees are distributed by a CSR process in a square area, Infected trees are randomly assigned with 500 of the trees infected. A sample of IP distances is taken with the size calculated = $A \cdot F(d)$. Calculate the mass and cumulative probability that the sample contains 62237 values calculated only based on infected tree pairs. The distribution of F(d) could be calculated either analytically or by simulation but for this problem, *F(d) = 0.50.*

A = *N(N-1)/2* = 1000* 999/2 = 499,500 distances based on all trees

B = I(I-1)/2 = 500 * 499/2 = 124,500 distances based on infected trees.

n = 499,500 * 0.5 = 249,750 distances

Excel Program used for calculations:

P{X = 62250} = Hypgeom.dist(62250, 249750, 124450, 49950, False) = 0.0026

P{X < 62250} = Hypgeom.dist(62250, 249750, 124450, 49950, True) = 0.5013

mean = 124,000 * 0.5 = 62,250

$$\sigma^2 = n\left(\frac{B}{A}\right)\frac{(A-B)}{A}\frac{(A-n)}{A-1}$$

Variance = 23,367

$$\sigma^2 = i_{exp}[1 - cdf_T(d)]\left(1 - \frac{I(I-1)}{N(N-1)}\right)$$

Variance - 62250 * 0.5 * (1 - 124500/499500) = 23,367

Standard deviation = 15286

Normal approximation:

P(X = 62250) = P(X < 62251) - P(X < 62250) = 0.0026

P(X < 62500) = 0.50

5. Confidence Envelope

While equations 1 to 5 are correct, except where noted on the variance equation, the authors do not explain how either the normal or the hyper-geometric distributions were used to calculate the confidence intervals. This lack of documentation is unfortunate. The only possible means of reviewing the confidence limits was by comparing the article's results with those obtain by simulation.

A MATLAB routine was used to: (a) generate 71 independent and random points in a 3.0 x 1.2 km area, (2) calculate IP distances for *n* points and (3) rank these distances from lowest to highest and (4) assign probabilities based on quartiles, so a full set of probabilities ranging from p1 to p99 is calculated. A series of 100 realizations were performed, and for each probability, ranging from 0.10 to 0.99, probability limits of 0.05 and 0.95 were calculated based on quartiles.

The simulation area and number of points were chosen for comparison with the published article, Figure 4. In Table 1 of the published article, time period 4 has 40 previously infected trees and 31 newly infected trees for a total of 71 infected trees. The area dimensions of 3 x 1.2 km were used because these dimensions fit the expected curve in Figure 4.

Figure 1: Simulation Results with Mean, Lower Limit and Upper Limit Curves

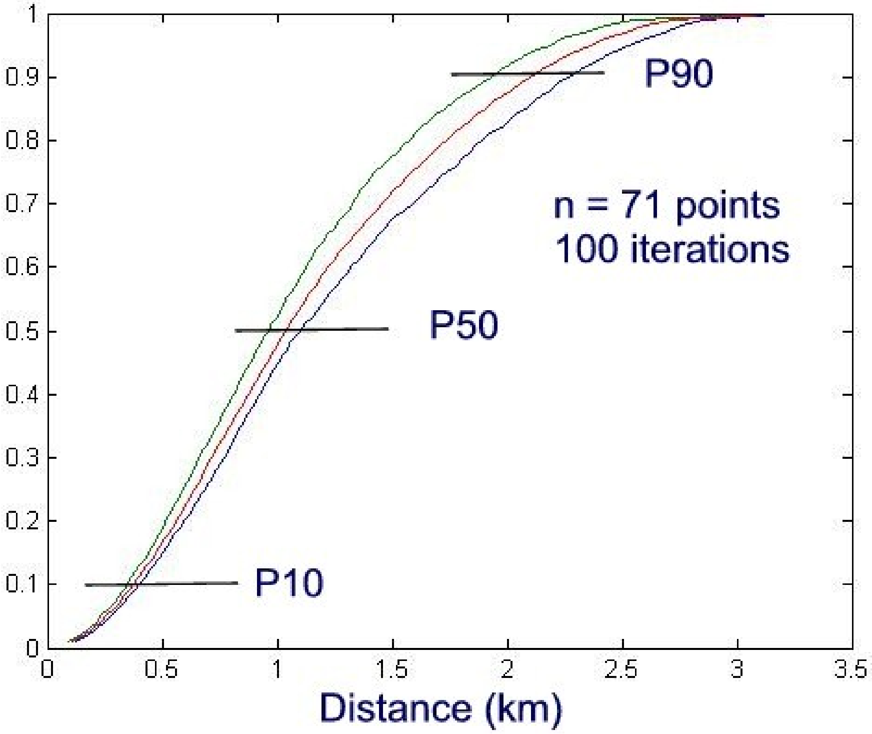
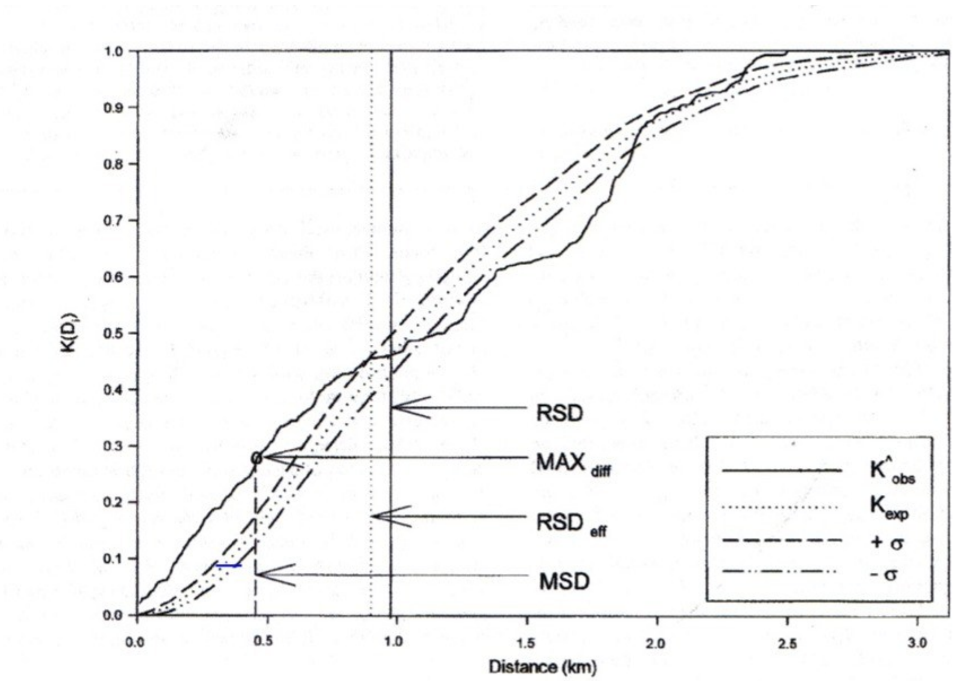


Figure 2: Diagram from Published Article, showing K_{exp} Curve



Confidence interval is shown as $+\sigma$ or $-\sigma$.

COMPARISON OF FIGURE 4 OF THE PUBLISHED ARTICLE WITH SIMULATION

Examining both Figure 1 and 2, it is apparent that the simulated results show a tighter bounds at the lower end of the curve and a wide bounds at the upper end, while Figure 4 in the article shows more uniform bounds. The comparison of results at selected probability values is shown in Tables 1 and 2.

Table 1: Lower Limit Curve Comparison of distances (km)

Probability Level	Lower Limit distance (km) from article, Fig 4	Lower Limit distance (km) simulation	% Difference
P = 0.10 (P10)	0.317	0.374.	15.2
P = 0.50 (P50)	0.991	0.964	(2.8)
P = 0.90 (P90)	2.008	1.912	(10.8)

% Difference = (Article distance - Simulation distance)/ Article distance. Article distances were obtain by digitizing the graph.

Table 2: Upper limit curve comparison of Distances

Probability Level	Upper Limit distance (km) from article, Fig 4	Upper Limit distance (km) simulation	% Difference
P = 0.10 (P10)	0.415	0.401	(3.5)
P = 0.50 (P50)	1.127	1.111	(1.4)
P = 0.90 (P90)	2.209	2.297	3.8

6. Conclusions

1. Equations 1 - 5 were reviewed in detail. There is a minor error in the variance calculation. .

2. Appendix F concluded the expected curve was based on simulation of inter-point distances of a CSR pattern. Similarly, it is considered possible that the confidence intervals (probability limits) are based on simulation. There is insufficient theoretical detail to know the manner of calculation.

3. In this review, confidence intervals were generated using simulation. When Figure 4 confidence intervals were compared with simulated results, there were significant differences in the shape of the curves. This was most apparent in the lower end of the distribution (P10) with the lower bound with 15.2% difference in distance.

References

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.

Matlab Program Listing

```
 %
%  Program to calculate CI for IP distances in a
%  CRS Pattern
%
%  nt = number of points,
%  ni = number of iterations
%  xsize, ysize = x and y coodinates
%
%
clear;
ni = 4000;
nt= 71;
xsize = 3; ysize = 1.2;
for niter= 1:ni;
    x = rand(1,nt)*3;
    y = rand(1,nt)*1.2;
    k = 1;
    for i = 1:nt;
        for j = 1:nt;
            if i > j;
                d(k)  = ((x(i)-x(j))^2+(y(i)-y(j))^2)^0.5;
                k = k + 1;
            end;
        end;
    end;
    d = sort(d);
     ncount = size(d,2);
     p = 1:1:ncount;
     p = p/ncount;
     px = 0.01:.01:.99;
     kk  = ceil(px*ncount);
     % data we use to form confidence intervals
     cdata(niter,1:99) = d(kk);
    end;
    cdata = sort(cdata);
    ilower = ceil(niter*0.05);
   iupper = ceil(niter*0.95);
   xmean = mean(cdata)
   clower(1:99) = cdata(ilower,1:99);
   cupper (1:99) = cdata(iupper,1:99);
   clower (100) = sqrt(xsize^2+ysize^2); cupper(100) = clower(100);
   xmean(100) = cupper(100)
   px(100) = 1;
    plot(cupper,px, clower,px, xmean,px);
    % comparison to gottwald's work
    delta = cupper-clower;
    err1 = delta(10)-0.096
    err2 = delta(50) - 0.151
    err3 = delta(90) - 0.237
    pct_err1 = err1/delta(10)*100
    pct_err2 = err2/delta(50)*100
    pct_err3 = err3/delta(90)*100
```