Letter to the Editor

An Assessment of the "Distance Necessary to Circumscribe" Method using a Stochastic Simulation Approach

David B. Lord

dbl Consultants

6451 SW 73 St, South Miami, FL 33143

ABSTRACT

Gottwald et al. introduced a distance-based spatial analysis method, or "Distance Necessary to Circumscribe" method to calculate statistics on inter-tree disease transmission distances for the pathogen causing the citrus canker. In this study, the measure of importance is the upper probability range of the transmission distance denoted as the 95% percentile or D95 measure. An analytical solution is presented for the sampling distribution of this statistic from a random sample. In all other cases, a Monte Carlo simulation model is used to identify the mean and its confidence interval of the sampling distribution. Simulation suggests that biasing factors can result in under or over estimation of D95, depending on which factors dominate the study. Results from simulation also suggests that the mean D95 may contain a high degree of uncertainty as evident from the confidence interval. The model used is generic in nature and should not be inferred to represent the citrus canker epidemic as studied by Gottwald et al. An alternative means of identifying the upper probability range of calculated distances is suggested by use of an assumed distribution function. This could be used as a check on the results.

Additional keywords: simulation, beta distribution, order statistics

Gottwald et al. (3) introduced a method for the spatial analysis of infected trees, denoted as the "distance necessary to circumscribe" or DNC procedure. The method was used in an epidemiology study to estimate distances of inter-tree movement of the pathogen causing citrus canker within urban residential sites in Florida (3). Gottwald et al. presented the upper probability range of calculated transmission distances, since the statistics were meant to help regulators decide on an appropriate policy for removal of hosts in proximity to an infected tree to prevent future spread of the disease (3).

In this study, the capability of the DNC method to provide measures of the upper probability range of transmission distances is examined. For this objective, various cases are presented which provide a better understanding of the effects of small sample sizes and data flaws on results. In the evaluation, an analytical sampling distribution is used in the most idealized case (Case 1) and sampling distributions based on Monte Carlo simulation are used in the more complex cases (Cases 2 to 6). The model is generic in nature and not considered representative of the transmission behavior of a particular pathogen.

Monte Carlo simulation evaluations of percentiles in the upper probability range have been done by Modarres et al. (6) for log normal and other log transformed distributions used in environmental studies. Modarres et al. identified the potential for under and over estimation of percentiles when the sample size is small and uncertainty exists in

Corresponding Author: D. Lord, Email address: dblord@hotmail.com

the choice of the appropriate probability model. It was concluded that in the extreme probability range ($p > 0.99$), and sample sizes of less than 30 values, all of estimators were unreliable. There was better success with larger samples sizes (100 and 1000 values) in the extreme range.

Distance Necessary to Circumscribe Procedure. The DNC procedural steps as given below are based exclusively on published article (3) with two exceptions: (a) the term "prior infected tree" is used instead of the term "focal tree" in the published article and (b) the terms "percentile" and "Dp" as used in step 4 are not within the published article. However, the calculation methodology of percentiles is consistent with the results presented in the published article and commonly used procedures.

1. Data collection: All infected trees are identified within the site by repeated surveys. The date of discovery and location of infected tree are identified. The age of the oldest lesion is estimated. For each infected tree, an infection initiation date (IID) is calculated equal to the discovery date minus the age of the oldest lesion on the infected tree.

2. Data Parsing: The data parsing creates infection scenarios for discrete time periods. The scenarios consider that PI trees are the sources of the infection for the NI trees. The NI trees are considered infected, but not yet infectious within the time period. All NI trees are re-classified as PI trees in the next time intervals and remain so classified for all successive periods.

3. Tree Associations and Distance Calculations. For each NI tree in a time period, the PI tree that is the closest to the NI tree is used for distance calculation, so each NI tree is associated with a PI tree. For each time period, a set of transmission distances is calculated.

4. Percentile Calculation. An estimate of the upper probability range of distances is determined by calculating the p^{th} percentile, denoted as Dp. The p^{th} percentile of a sample is the smallest value such that at least p percent of the sample is less than or equal to this value (5). Consistent with this definition, for a set of observations in ascending order with rank k ($k = n$ the highest value), $k = \lceil p \cdot n \rceil$ where $\lceil . \rceil$ denotes a ceiling function,. i.e. smallest integer greater or equal to the argument of the function.

Spatial relationships with the DNC Procedure. The DNC procedure requires: (a) the ability to estimate the initial date of infection of each infected tree and (b) the ability to inspect all hosts for the disease. The DNC procedure forms a parent-offspring relationship whereby each offspring (NI tree) must have only one parent, however one parent (PI tree) can be associated with innumerable offspring. For every infected tree, the origin of the disease to a particular tree is determined with the exception of the PI trees in the first period, as the sequence must begin with infected tree(s) of an unknown origin. Thus, there exists at least one unidentified infected tree outside of the site that is responsible for the onset of the disease within the site.

Each discovered tree is placed in a chronological sequence, depending on the age of the oldest lesion and discovery date. Thus, if the latency period is long or inspections are infrequent or incomplete, it is conceivable that the last infected tree discovered has a sufficiently high lesion age that would place the tree chronologically in the first time period as a prior infected tree. In this case, the origins of the disease would be the last information discovered. The unique characteristic of the method is that one can not be certain of any tree association until all information has been gathered.

If source trees exists beyond the site's boundaries, this may affect results. If the initially infected trees are at the center of the site, then those NI trees associated with a PI tree at long distances would be located closer to the edges of the site. If unidentified or removed infected trees were near the boundaries, they could be the sources of infection. The procedure would consequently result in incorrect tree association.

Assessment of DNC Method. The DNC method was evaluated by running simulation cases, where the p^{th} percentile is known, generating a set of distances and then using the DNC procedure to calculate the p^{th} percentile statistic from this set. The calculated percentile is compared with the known one. An unbiased statistic converges, after infinite and independent sampling, to a value which matches the value generating distribution. The mean of the 95% percentile (D95) was used in our assessment for comparison between calculated and true values. In addition, the lower and upper confidence limits, corresponding to 5% and 95% percentiles of D95, are presented to judge the uncertainty of the mean D95. Other statistics include the fraction of NI trees that are incorrectly associated with PI trees compared to all associations (f_{ia}) and the fraction of NI trees that are outside of the site area as compared to those inside the site (f_{out}).

Case 1 considers a single parent within an unlimited area, while Cases 2 to 6 consider multiple parents which are randomly distributed in a square area. The Matlab 6.0 program (Mathworks Inc, MA) was used to simulate the process and calculate relevant measures. A minimum of 5,000 runs were made, which results in mean D95 measures that vary by less than 1%. An exponential probability distribution function (pdf) is used to describe the transmission distance with mean θ. It is one of the more commonly used

93 distribution in plant disease epidemiology (4). All cases considered a low mean
94 transmission distance ($\theta = 10$ m) and a high mean transmission distance ($\theta = 100$ m)
95 denoted as the t10 and t100 sets. It was considered that the number of offspring in a
96 time period will likely be limited, since offspring soon become parents. The simulation
97 cases consider only a single time period with 30 or fewer NI trees.
98 It is recognized that by varying the description of transmission, various assumptions and
99 parameters, innumerable other cases could be studied. The cases here should not be
100 considered to bracket all possibilities.
101 **Case Descriptions and Results.**
102 **- Case 1: Single Parent/ Unlimited Area for Offspring.** For this case, the sampling
103 probability distribution function (pdf) of the D95 percentile can be analytically
104 determined. Representing the transmission distance of the disease from a source tree to
105 an NI tree with a pdf of *f(x)* with a corresponding cumulative distribution function *F(x),*
106 the pdf of the k^{th} order random variable $X_{(k)}$ is given by:

107 $$g(x) = \frac{n!}{(k-1)!(n-k)!} f(x)[F(x)]^{k-1}[1-F(x)]^{n-k} \qquad (1)$$

108 per reference 2. If *f(x)* is an exponential pdf, then $g(x)$ can be expressed in the form of
109 a beta distribution as provided in the appendix. The true value of transmission distances
110 at the 95% percentile, $F^{-1}(0.95)$, equals $2.996 \cdot \theta$, where θ is the mean of the
111 exponential distribution.
112 Results of Case 1 are provided in Table 1 for *n* values of 10, 20 and 30, with
113 corresponding *k* values are 10, 19 and 29. The mean of D95 ranges from 26 to 29 m and
114 260 to 290 m for$\theta = 10$ and 100 m, (denoted as t10 and t100), which is below the true
115 values of 29.9 and 299.6 m, respectively.

Confidence limits and mean of D95 as fraction of the true value for values of *n* up to 300 are shown in Figure 1. Due to the discrete nature of percentile measures, a saw tooth pattern is observed in Figure 1, with downward breaks in mean and confidence intervals occurring at $n = 20, 40, 60, \ldots$ reflecting changes in rank *k*. As the number of calculated transmission distances tends towards infinity, the value of D95 will converge on the true value of 29.96 m.

The utility of this simple model is to identify the impact of limited sample size exclusive of all other factors. Within the model, the point source of the disease (parent) is known and all offspring remain as offspring or exposed state (as infected, but not infectious). There is no ambiguity as to the source of the disease. Information on offspring location is complete and perfect. Thus, the model is probably more representative of a highly controlled environment, such as an experiment in a containment greenhouse where a disease source has been introduced than for a situation encountered in nature.

- Case 2: Effect of Multiple Parents in an Unlimited Observation Area. This case considers that there are multiple parents in a defined area. The area available to offspring is unlimited. The initial *m* parents are randomly distributed in a square area with sides *S*. The PI trees (parents) to produce the NI trees (offspring) are randomly chosen by a uniform discrete distribution.

Results of Cases 2a, 2b and 2c for t10 and t100 with $S = 1000$ m are provided in Table 1 for the number of PI trees ranging from 10 to 30 with the number of NI trees equal to 20. The mean D95 values range from 25 to 26 m for t10 and from 153 to 196 m for t100. The f_{ia} value (fraction of incorrect associations) is between 0.005 to 0.017 for t10 while it is as high as 0.393 for t100. Thus, for t100, nearly 40% of the distances of the

sampled population is based on incorrect associations. The f_{ia} value increases directly with the number of parents and the transmission distance. For this case, all incorrect associations are negative biasing factors leading to an under estimation of the mean D95 value. The upper and lower confidence limits for these cases, are generally lower than Case 1, reflecting this biasing factor. However, as will be shown in Cases 4a and 6, there is also potential for positive biasing factors as well when data error are present.

- **Case 3: Effect of Multiple PI trees in a Limited Observation Area.** This case is identical to Case 2, but with a limited area to observe the offspring of the PI trees. Thus, in the model runs, any NI tree that is located outside of the specified area (a square with sides, *S*, equal to 1000 m) is ignored and an additional NI tree location is generated until a full count of NI trees inside the site as specified is achieved. The area specified in Case 2 for generating PI tree locations is used in this case to spatially limit both the generation of PI and NI tree locations. Thus, the location of an NI tree becomes dependent on its parent's location, θ and *S*.

Results are shown in Table 1 for PI trees of 10 (Case 3a), 20 (Case 3b) and 30 (Case 3c). The mean D95 value is 25 m for the t10 set and ranges from 118 to 162 m for t100 set. The f_{out} value is less than 2% for the t10 set, while it is as high as 39% for the t100 set. This means that if there are 20 NI trees inside the site, an additional 39% or approximately 8 NI trees lie outside the site that are not counted. This is referred to as a censored data set and results in a mean D95 that is below its true value. The censoring of data also causes the upper confidence limit to be below the true mean value of D95 (299.6 m). Also, unless the site is surrounded by non-hosts, there is the potential that the infected trees outside of the area could later be undetected sources.

Cases 4. One PI Tree Outside Site Area. Cases 4, 5, 6 are perturbations of Case 3b designed to assess the impact of minor changes in Case 3b on calculated results. Case 4 considers that an additional PI trees exist in a periphery area outside of the site area. The peripheral area is defined in this case to extend 10 m beyond the site boundaries on all sides. The unobserved parent tree is randomly located in the periphery. Selection of which parent is to generate an offspring is done in the same manner as in all cases, by a uniform discrete distribution. Thus, the unobserved parent only has an affect in runs, where by chance, it is selected to generate an offspring. In all other respects, the case is identical to Case 3b.

Results of Case 4 are shown in Table 1 for $m = 20$. For the t10 set, the upper confidence limit increased from 38 m in Case 3b to 47 m in Case 4. For the t100 set, the upper limit increased from 187 to 200 m, which is still below the true value of 299.6 m. While the additional PI tree is a positive biasing factor, it is not sufficient to offset the other negative biases (closest neighbor assumption and confined area).

An additional runs were made (Case 4a) with two source trees outside of the study area. Results are presented only for t10, where the impact was more significant. All other model parameters were unchanged. In this case, the upper confidence limit increased from 38 to 93 m, which is approximately three times the true value of D95 (29.9 m).

- Case 5. Effect of an NI tree that is incorrectly identified as PI tree. This case considered the impact of an NI tree being misclassified as a PI tree. In this case, the error can occur from an over estimation of the oldest lesion age, resulting in a calculated IID (as described in the DNC procedures section) further back in time. Since discrete time periods are used, an error of one day is sufficient to place a tree into an incorrect

time period. Misclassification in this case results in 21 PI and 19 NI trees instead of the 20 PI and 20 NI trees.

As shown in Table 1, for both t10 and t100 sets, the mean D95 increases by approximately 10% over the unperturbed case, Case 3b. The lower and upper confidence limits are also higher. The addition of additional parent would normally lead to an under estimation of the mean D95 measure. However, there is a second factor involved which is the discrete nature of percentile measures. This results in a saw tooth pattern as shown in Figure 1. The D95 measure (with all other factors equal) has a higher mean and upper bound of the confidence interval for n =19 than n = 20. Additional runs (not shown) confirmed that in general, the effect of incorrectly classifying an NI tree as a PI tree will cause an under estimation in the mean D95 value.

- Case 6. Effect of a PI tree that is incorrectly identified as an NI tree. This case considers the impact of a PI tree being misclassified as a NI tree. This could occur from an under estimation error in the oldest lesion age resulting in the IID (as described in the DNC procedures section) further ahead in time. In this case, the upper confidence limit for t10 increased from 39 to 189 m. This is more than five times the expected value of 29.9 m. A similar increase did not occur for θ = 100 m, as the NI trees which are more dispersed (less aggregated) in the site area, there is less impact of a PI tree which is misclassified as a NI tree.

Discussion. The cases show potential for over and under estimation of D95 measure. Two factors present in Cases 2 and 3 bias the results negatively: (1) the closest neighbor assumption for PI association (2) the limited area for new infected trees. These factors were most apparent in the Case 3c (multiple parents/ limited observation area with

$\theta = 100$ m), where the mean *D95* is 118 m or 39% below its true value. The potential for positive biasing factors affecting D95 is shown in Cases 4a and 6 due to imperfect information. In Case 4a ($\theta = 10$ m), with two PI trees outside of the area, the upper confidence bound is 93 m, or 310% above its true value. In Case 6 ($\theta = 10$ m) the upper confidence bound is 188 m. This means that there is a 5% chance that the D95 estimate could be 631% greater than the true value, due to a single misidentification. As noted, this error is possible if the age of the oldest lesion on a tree is off by as little as one day.

The core of the estimation problem is that the population being sampled contains correct and incorrect distances. The inclusion of the incorrect distances in the Cases 2,3 and Case 6 are negative bias factors, leading to an under estimation of D95. However, the mean values tell only part of the story. If one could repeat an experiment many times under identical circumstances, these mean values would result. However, in situations where the field environment makes this difficult, then the uncertainty of results as reflected in the confidence interval becomes important. Identifying which set of upper bounds is the most representative of an actual study may lie in the ability to identify PI and NI trees in an unambiguous manner, and the ability to know that the phenomena (disease transmission) is solely from trees within the site. It is noted that surveys of trees are vastly better in a grove environment with uniform cultivars and management than a residential area, with various cultivars, varying degrees of care and barriers to entry and search of the premises.

The significance of over and under estimation will be application dependent. Certainly for inspection and quarantine purposes, over estimation may justified as a safety factor.

However, when plant removal is involved, the economic consequences of over estimation (unnecessary removals) will probably be serious. Under estimation is also undesirable as repeated inspections and removals add to the cost of the program.

For a set of collected data, it is also possible to test the robustness of the statistics by adding to the data, additional randomly located source trees, representing the undiscovered trees (through actions of owners or defoliation) inside or outside the area and re-running the analysis. Additionally, lesion ages can be randomly changed by incremental values that reflect the uncertainty inherent in their values and the impact on results identified.

Also, if the calculated statistics are possibly influenced by erroneous associations, it is suggested that an alternative is to determine the mean and variance of a sample, then use these parameters and an assumed distribution to calculate the upper percentile distance such as D95. The results will most likely be sensitive to the selection of distribution. If through experimentation, the best distribution function can be identified, this can improve the predicted distance. Where maximum distances can be identified by experimentation, then truncated forms of these distribution can be employed for improved results.

Identifying the inherent variability of transmission due to many small factors in nature should be the goal of statistical analysis. Uncertainty, often referred to as “epistemic uncertainty”, is the assessor’s lack of knowledge (level of ignorance) about a parameters that characterize the physical system being modeled (7). Good information tends to provide insight, however events perceived to have occurred due to incorrect information increases our lack of understanding or uncertainty of a process. Use of more controlled

254 studies may better identify the distribution form of transmission, allowing checks on
255 calculated results.
256
257 **LITERATURE CITED**
258 1. Balakrishnan, N. and Cohen, A.C. 1991. Order Statistics and Interference, Estimation
259 Methods, Academic Press. Hartcourt Brace Jovanovich, Boston.
260 2. David, H. A. and Nagaraja, H. N. 2003. Order Statistics. Wiley-Interscience, Hoboken.
261 3. Gottwald, T. R., Sun, X., Riley, T. Graham, J. H., Ferrandino, F., Taylor, E. L. 2002.
262 Geo-referenced spatiotemporal analysis of the urban citrus canker epidemic in Florida.
263 Phytopathology, 92: 361-376.
264 4. Jeger, M.J. (ed) 1989. Spatial Components of Plant Disease Epidemics. Prentice-Hall,
265 Englewood Cliffs.
266 5. Larsen, H.J. 1982. Introduction to Probability Theory and Statistical Interference. John
267 Wiley, New York.
268 6. Modarres, R., Nayak, T., Gastwirth, J. 2002. Estimation of upper quantiles under
269 model and parameter uncertainty. Computational Statistics and Data Analysis, 39, 529-
270 554.
271 7. Voss, D. 2000. Risk Analysis. John Wiley, New York.
272
273
274

Appendix: Analytical solution for distribution of the k^{th} order statistic for single point source when transmission distance is exponentially distributed.

For an independent and identically distributed random sample of size *n,* the probability distribution function (pdf) of the k^{th} order statistic, ranked from lowest to highest with $1 \leq k \leq n$ is:

$$g(x) = \frac{n!}{(k-1)!(n-k)!} f(x)[F(x)]^{k-1}[1-F(x)]^{n-k} \quad \text{(A-1)}$$

where *f(x)* and *F(x)* are the pdf and cdf, respectively, of the transmission distance from a single source. Rank *k* can be calculated as $\lceil n \cdot p \rceil$ for the p^{th} percentile of sample size *n,* this results in a step function for *k* over the full range of p. Other percentile calculation methods use linear interpolation between $X_{(k)}$ values (6).

For $f(x) = 1/\theta \cdot \exp(-x/\theta)$ (exponential pdf) with mean θ, *g(x)* can be stated in the form of the beta distribution as:

$$g(x) = \beta(y \mid a,b) \cdot f(x) \quad \text{(A-2)}$$

where $\beta(\cdot)$ is the beta pdf and $y = \exp(-x/\theta)$, $a = n-k+1$ and $b = k$. From Equation. A.2, it follows that the confidence limit (CL) for probability *p* is:

$$\mathrm{CL}(p) = G^{-1}(p) = -\ln[\mathrm{B}^{-1}(q \mid a,b)] \cdot \theta \quad \text{(A.3)}$$

where B^{-1} is the inverse of the beta cumulative distribution, $q = 1-p$ and a, b as previously defined. Probabilities for the lower and upper CL in this study were $p = 0.05$ and 0.95. The mean of *g(x)* as per reference 1 is:

$$\mu_{n:k} = \theta \cdot \sum_{i=n+k-1}^{n} 1/i \quad \text{(A-4)}$$

294 Note per equation A.2, $\beta(x \mid a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{\beta-1}$ for $0 < x < 1$, zero

295 otherwise, where $\Gamma(a) = \int_0^\infty x^{\alpha-1}e^{-x}dx$ (gamma function).

296 Also, per equation. A.3, the inverse beta cdf is available in Microsoft Excel (Microsoft

297 Excel 2000, Redmond, California) as BETAINV function.

298

299

299

300 TABLE 1. Results of analytic solution (case 1) and simulation (cases 2 to 6)*

301

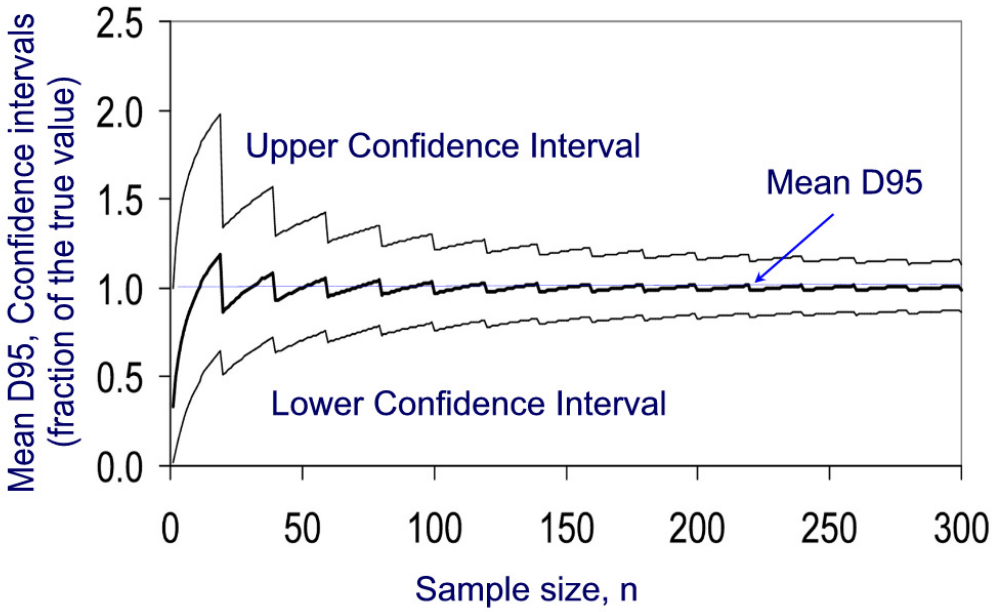
t10 set (mean = 10 m) 95% Probability distance = 29.96 m		Mean "D95" m	# of NI trees	Confidence Interval m	f_{out}**	f_{ia}***
Case 1a	Single PI tree, infinite area	29	10	14 to 53	0	0
Case 1b	Single PI tree, infinite area	26	20	15 to 40	0	0
Case 1c	Single PI tree, infinite area	30	30	19 to 44	0	0
Case 2a	Multiple PI trees, unlimited obs area, m=10 ****	26	20	15 to 40	0	0.005
Case 2b	Multiple PI trees, unlimited obs area, m=20	26	20	15 to 39	0	0.012
Case 2c	Multiple PI trees, unlimited obs area, m=30	25	20	15 to 39	0	0.017
Case 3a	Multiple PI trees, limited obs area, m=10	25	20	15 to 39	0.013	0.005
Case 3b	Multiple PI trees, limited obs area, m=20	25	20	15 to 38	0.013	0.010
Case 3c	Multiple PI trees, limited obs area, m=30	25	20	15 to 38	0.014	0.016
Case 4:	One PI tree outside area	30	20	16 to 47	0.050	0.023
Case 4a:	Two PI trees outside area	36	20	16 to 93	0.088	0.033
Case 5:	One NI misclassified as a PI tree	34	19	19 to 56	0.013	0.024
Case 6:	One PI misclassified as NI tree	30	21	22 to 188	0.013	0.058

t100 set (mean = 100 m) 95% Probability distance = 299.57 m		Mean "D95" m	# of NI trees	Confidence Interval m	f_{out}**	f_{ia}***
Case 1a	Single PI tree, infinite area, exact solution	295	10	138 to 533	0	0
Case 1b	Single PI tree, infinite area, exact solutions	260	20	150 to 401	0	0
Case 1c	Single PI tree, infinite area, exact solutions	301	30	191 to 446	0	0
Case 2a	Multiple PI trees, unlimited obs area, m=10	196	20	117 to 307	0	0.214
Case 2b	Multiple PI trees, unlimited obs area, m=20	169	20	100 to 277	0	0.324
Case 2c	Multiple PI trees, unlimited obs area, m=30	153	20	88 to 262	0	0.393
Case 3a	Multiple PI trees, limited obs area, m=10	162	20	106 to 232	0.141	0.212
Case 3b	Multiple PI trees, limited obs area, m=20	135	20	92 to 187	0.138	0.323
Case 3c	Multiple PI trees, limited obs area, m=30	118	20	82 to 163	0.135	0.390
Case 4:	One PI trees outside area	140	20	94 to 200	0.170	0.336
Case 5:	One NI misclassified as a PI tree	167	19	96 to 239	0.139	0.316
Case 6:	One PI misclassified as NI tree	151	21	104 to 212	0.139	0.334

302

303 * Results of analytical model and stochastic simulation of hypothetical cases to assess
304 the DNC method. Disease transmission distances are exponentially distributed. Mean
305 D95 is the mean 95% percentile of transmission distance. Upper and lower confidence
306 intervals are 5% and 95% percentiles of the D95 sampling distribution. ** f_{out} = fraction

of new infected (NI) trees outside the study area as compared with the number inside the area. *** fia = fraction of incorrect associations as compared with all associations. **** m = # of prior infected (PI) trees.



310

311 **Fig. 1** Mean and Confidence Intervals of the D95 statistic from Case 1 (single parent, *n*

312 offspring, unlimited area) based on an analytical solution. D95 is the 95% percentile of a

313 set of calculated transmission distances. Confidence intervals are 5% and 95%

314 percentiles of sampling distribution of D95.