Letter to the Editor

Two-Dimensional Distance Class Analysis of Disease-Incidence Data: Problems and Possible Solutions

Francis J. Ferrandino

Associate scientist, Department of Plant Pathology and Ecology, The Connecticut Agricultural Experiment Station, P.O. Box 1106, New Haven 06504.

I thank V. L. Smith and J. S. Ward for repeatedly asking the right questions and S. L. Anagnostakis, S. Sanogo, and D. E. Aylor for repeatedly questioning my answers.

Accepted for publication 18 April 1996.

Many recent papers on disease incidence have used a maplike representation of deviations from random behavior (11–15,23,24, 25,28). The basic procedure is to compare observed infected pair counts with expected pair counts, assuming a random distribution. For the large number of multiple comparisons made in these analyses, the level of confidence used is not appropriate. This letter discusses the problem and proposes a new method of analysis, along with directions for its use.

The spatial distribution of infected plants within a field is an important characteristic of a disease epidemic. The degree of spatial aggregation of disease may depend on the distance to the source of inoculum, as well as variation in the physical and cultural conditions occurring within a field. The spatial aggregation of plant damage due to disease is important in evaluating yield loss (6,7,9,17,21), and the changes in aggregation over time may provide a vital clue to the underlying mechanism of inoculum dispersal and the scales of distance over which it operates (8).

Attempts to quantify nonrandom patterns in terms of a single scalar statistic (2,32) or the fitting of a theoretical distribution (1,18,33) give no information on the physical scale of aggregation. Quadrat methods (16,19,22) and more sophisticated spatial autocorrelation techniques (3,20) can give some information on length scale but, usually, no directional information is obtained. All of the above methods are based on the assumption of a continuously varying disease severity and, as such, are not directly applicable to disease-incidence data, which is binary in nature (the only possible conditions are diseased or healthy). The overall departure of a binary data set from random behavior can be determined by Ripley's second order technique (26,27). This method is based on the difference between the observed and expected number of infected pairs less than a certain distance apart. Unfortunately, this approach yields no directional information.

Since many agricultural crops are planted in rows, there is every reason to believe that contagion is direction dependent. In addition, wind dispersal, rain splash dispersal, and surface water dispersal of spores are all apt to be strongly dependent on direction. Gray et al. (15) suggested a method that incorporates both interplant orientation and distance into the analysis. This procedure was later formalized in the computer program 2DCLASS (25), and then extended to include temporal information in the computer program STCLASS (23). The method categorizes pairs of infected plants into distance-orientation classes depending on the x-distance and y-distance between the infected plants within a pair. Infected plant pairs within each class are counted. The expected number of pairs and the confidence limits about this expectation value are calculated stochastically using Monte-Carlo simulations. The results are summarized in a matrix of probabilities representing deviations from random behavior for each of the x-y distance classes, which are then judged significantly nonrandom at some α-level of probability. In recent years, there has been a preponderance of papers using this technique to characterize disease-incidence data on a two-dimensional grid (11–15,23, 24,25,28). Conclusions drawn from the above analysis are often ambiguous and, for this reason, I would like to reexamine the method on a more formal basis.

There are two interrelated problems with the above distance-class methodology. First, for a large field with many infected plants, these calculations can become unwieldy. With the advent of faster personal computers, the unwieldy nature of these calculations is not as serious a problem as it once was. However, for a large field with 1,000 plants, a single 2DCLASS run with 400 Markov simulations using a compiled program takes about 7 min on a 60 MHz Pentium. I will show that the appropriate number of simulations is closer to 4×10^5. Such a calculation will take almost 5 days. Second, because of the many distances and angular orientations tested, this approach suffers from the problems associated with all multiple comparison tests that can only be corrected by employing a more conservative test of statistical significance, requiring an increase in the number of Monte-Carlo simulations (4,30,31).

These two shortcomings of the Monte-Carlo enumeration of confidence limits can be overcome by an alternative method of analysis. The purpose of this letter is to present an exact analytical method for calculating expectation values and confidence limits for the observed number of infected plant pairs within every possible distance and angle class. In addition, the analytical method provides confidence limits to any level of probability without increasing the time of the calculation. This allows a logical framework within which the multiple comparison tests may be applied conservatively using the Bonferroni method (4).

METHOD

Plant pairs. Consider a field containing n plants arranged in a rectangular array, W plants wide in the x-direction and L plants long in the y-direction (Fig. 1). If a total of n_m of the lattice sites are not occupied (missing), then:

$$n = L \cdot W - n_m \tag{1}$$

Taking these plants two at a time, we can also consider pairs of plants. The total number of unique plant pairs, N_T, is:

$$N_T = n(n-1)/2 \tag{2}$$

Corresponding author: F. J. Ferrandino; E-mail address: fjferr@caes.state.ct.us

Publication no. P-1996-0523-01O

Equation 2 is obtained by realizing that there are n possible choices for the first plant in a pair and $(n - 1)$ choices for the second member of the pair. The division by two is necessary since each pair is counted twice, depending on the order in which plants were chosen.

These pairs may be grouped into distance-orientation classes depending on the number of plants in the x-direction and y-direction between the plants in the pair. For example, pairs within the (1,3) distance-orientation class are constructed by starting at a reference plant (tail of arrow, Fig. 1), moving +1 plants to the right, +3 plants up, and finally arriving at the target plant (head of arrow, Fig. 1). The reference and target plants are then associated as a pair and characterized by their separation vector, $\vec{v}$. In Figure 1, $\vec{v} = (+1,+3)$. Note that reference plants in such a pair must lie within the lower left rectangle in Figure 1 (solid outline) and the target plants are drawn from a different sample, namely the upper right rectangle in Figure 1 (dashed outline). In a rectangular array, the farthest two plants can be apart in the x-direction is $W - 1$. Depending on whether the reference plant is to the left or to the right of the target plant, this maximum distance can be positive or negative ($\pm(W - 1)$). Likewise, the separation in the y-direction ranges from $-(L - 1)$ to $+(L - 1)$. Letting (j,k) represent an arbitrary distance-orientation class, the allowed values of j and k are given by:

$$-W + 1 < j < W - 1 \text{ and } -L + 1 < k < L - 1 \tag{3}$$

Of the total number of plant pairs, N_T, only a portion, N_{jk}, belong to the (j,k) distance class defined above. Note that for $j = 0$ and $k = 0$, the value of N is n. If there are no missing plants ($n_m = 0$), then N_{jk} is given by:

$$N_{jk} = (W - |j|)\,(L - |k|) \tag{4}$$

Equation 4 represents the number of plants included within either rectangle in Figure 1.

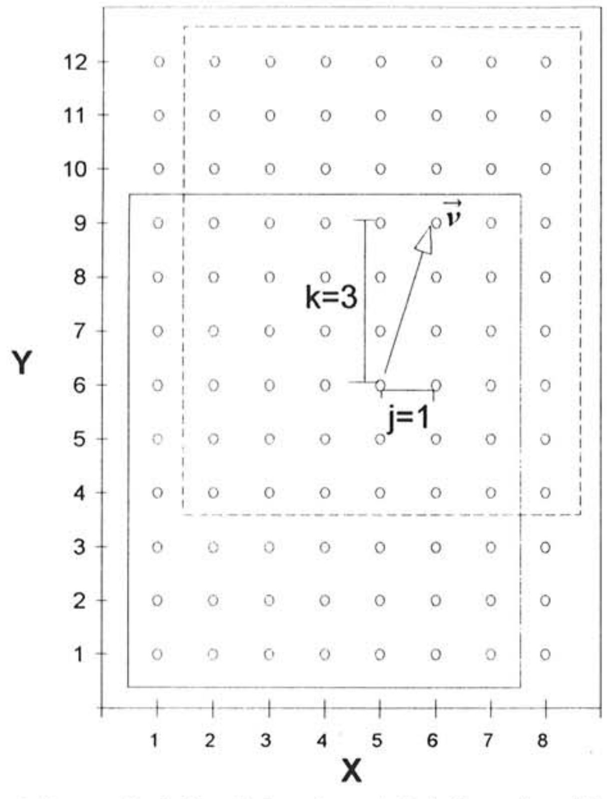


Fig. 1. Rectangular lattice of plants (open circles). Plant pairs within the distance-orientation class (j,k) are defined such that the vector, $\vec{v} = (j,k)$, goes from one plant to the other within the plant pair.

Due to reflective symmetry, however, $N_{jk} = N_{-j-k}$. This redundancy is the reason for the division by two in equations 2 and 5. Thus, the pair illustrated in Figure 1 also can be drawn with the vector going from the target plant (arrow head) to the reference plant (arrow tail). If there are missing plants ($n_m > 0$), the expression for N_{jk} is a bit more complicated, and the details of its evaluation are shown in the Appendix (equation A5).

Assume a number, i, of the plants are infected, leaving $n - i$ plants healthy. In this case, pairs of plants within the plot also can be characterized according to the disease status of each plant in the pair, in addition to interplant distance and direction. This involves counting all possible combinations of diseased, D, and healthy, H, plants (i.e., H-H, H-D, D-H, and D-D) within each distance-orientation class. In analogy with Equation 2, the number, I_T, of the plant pairs containing two infected plants (D-D), is given by:

$$I_T = i(i - 1)/2 \tag{5}$$

A number of these infected pairs, I_{jk}, lie within the (j,k) distance class. If infection is a random process, then we would expect the fraction of N_{jk} that are infected pairs will be independent of interplant distance and orientation (i.e., I_{jk}/N_{jk} = constant). If, however, contagion is present, then proximity to a diseased plant would enhance the probability of infection, and one would find a larger number of infected pairs within smaller distance classes.

Monte-Carlo simulations. Gray et al. (14) suggested a three-step process for evaluating and interpreting the distribution of infected plant pairs among the possible distance-orientation classes. These steps are:

Pair enumeration. Because of the different possible number of pairs in each distance-orientation class, Gray et al. (14) suggested the calculation of the standardized count frequency (SCF) for each distance-orientation class ([X,Y]). This is accomplished by dividing the observed infected pair count by the total number of possible pairs in each distance-orientation class (i.e., I_{jk}/N_{jk}). In assigning distance classes, Gray et al. (14) did not differentiate between positive and negative values for the x-displacement or the y-displacement ($\mathrm{X} = |j|$ and $\mathrm{Y} = |k|$). The number of infected, I_{XY}, and total, N_{XY}, plant pairs in Gray's [X,Y] distance classes can be related to I_{jk} and N_{jk} defined above, as follows:

$$\begin{aligned} I_{\mathrm{XY}} &= I_{jk} + I_{-jk} \\ N_{\mathrm{XY}} &= N_{jk} + N_{-jk} \quad ; j,k \neq 0 \\ &\text{and} \\ I_{\mathrm{XY}} &= I_{jk} \\ N_{\mathrm{XY}} &= N_{jk} \qquad ; j \text{ or } k = 0 \end{aligned} \tag{6}$$

It then follows that $\mathrm{SCF[X,Y]} = I_{\mathrm{XY}}/N_{\mathrm{XY}}$. The algorithm used in the computer program 2DCLASS (25) calculates the observed value for SCF[X,Y] in each distance-orientation class by simply counting all plant pairs and infected plant pairs and assigning them to the appropriate distance class.

Monte-Carlo estimation of the distributions. To estimate the expected distribution of pair counts, 400 numerical simulations are run using a pseudo-random number generator to place i infected plants within the lattice. SCF[X,Y] is calculated for each simulation in the same way as for the observed data. The mean value and confidence limits are then estimated for each SCF[X,Y] from the results of the numerical simulations.

The multiple comparison matrix. The results of this analysis are usually summarized as a maplike matrix, within which SCFs that are calculated to be significantly different from expected ($P < 0.05$) are flagged by a plus sign, "+", or a minus sign, "–", depending on whether the observed value is larger or smaller than expected (11–15,23,24,25,28).

In what follows, I will show that step 2, the evaluation of the expected probability distributions, can be achieved analytically. In addition, I found that the interpretation of the probability matrix obtained in step 3 can be troublesome because of the high probability of many false positives (Type I errors).

THEORY

I propose a new method, based on combinatorial theory, to calculate the mean and confidence limits of the expected number of infected plant pairs. First, I calculate the probability, p, that a randomly chosen pair is infected. If all outcomes are equally likely, then the probability of success is the ratio of successful events (number of infected pairs, I_T) to the total number of events (number of pairs, N_T) so that $p = I_T/N_T$. Note that p is equivalent to what Gray calls 'the standardized number of expected infected pairs of plants' in his Table 1 (14) and applies to all distance-orientation classes.

In like manner, one can calculate the probability that exactly I infected pairs are included in a sample of N pairs drawn from a total population of size N_T that contains I_T infected pairs. This probability distribution is derived in the Appendix and shown to be given by the Hypergeometric function, $P_{\mathrm{HG}}\ (I,I_T,N,N_T)$, defined by equation A2 (5,30):

$$P_{\mathrm{HG}}\left(I,I_T,N,N_T\right)=\frac{I_T!\cdot N!\cdot\left(N_T-I_T\right)!\cdot\left(N_T-N\right)!}{I!\cdot\left(I_T-I\right)!\cdot\left(N-I\right)!\cdot\left(N_T-I_T-N+I\right)!\cdot N_T!} \qquad \text{(A2)}$$

Fisher's exact test (10,30) is a statistical method that was specifically designed to calculate confidence limits on population distributions described by the Hypergeometric function (equation A2). The method involves construction of a 2 × 2 contingency table (Table 1) for each distance-orientation class. Deviation from random behavior is tested using a chi-square (χ^2) test of independence with one degree of freedom, in which, for a sample containing N pairs of which I are infected, χ^2 is given by:

$$\chi^2=\frac{\left[\left|I\cdot\left(N_T-I_T-N+I\right)-\left(N-I\right)\cdot\left(I_T-I\right)\right|-\left\{\frac{N_T}{2}\right\}\right]^2 N_T}{I_T\cdot\left(N_T-I_T\right)\cdot N\cdot\left(N_T-N\right)}$$

which simplifies upon expansion to: (7)

$$\chi^2=\frac{\left[\left|I\cdot N_T-N\cdot I_T\right|-\left\{\frac{N_T}{2}\right\}\right]^2 N_T}{I_T\cdot\left(N_T-I_T\right)\cdot N\cdot\left(N_T-N\right)}$$

Equation 7 includes Yate's adjustment for continuity (term in curly brackets; 10,30). The above method is very accurate, as long as the expected number of pairs [$\lambda = (NI)/N_T$, Appendix] is greater than five (10). At smaller expectation values, the skewness of the distribution is not as well described by the symmetrical χ^2 distribution, and the Hypergeometric function (equation A2) can be used to calculate the exact probability.

Multiple comparisons as a binomial distribution. As the number of experimental comparisons increases, spurious results become more likely (4,31), as a simple example will show. Suppose there are 20 Ping-Pong balls in a paper bag. One of these balls is marked with a plus sign, "+", and one is marked with a minus sign, "–". The remainder of the balls are unmarked. If a ball is chosen at random from the bag, then the probability of choosing an unmarked ball is 0.9 (successes/total = 18/20), the probability of choosing the ball marked "+" is 0.05, and the probability of choosing the ball marked "–" is 0.05. The drawn ball is now returned to the bag and the process is repeated. The probability of choosing unmarked balls twice in a row is 0.81 (0.9 × 0.9). As the process continues, it becomes less and less likely to repeatedly choose unmarked balls. After 10 draws, the probability is less than 0.35, and it falls to less than 0.05 after 29 draws. This process is analogous to repeatedly comparing observed with expected pair counts at the 0.05 level of confidence. Of course, when the ball marked "+" or "–" is drawn from the bag, this corresponds to an SCF that is greater or less than expected ($P \le 0.05$), respectively. Typical applications of the 2DCLASS method involve many comparisons (11–15,23,24,25,28), so that spurious results are a serious problem.

Suppose that c comparisons are made between observed and expected values, and that a number, s, of them are judged significant at the α-level of probability. By definition, the probability of falsely claiming significant deviation from random behavior is α for any one of the individual comparisons. The probability, $P_{\mathrm{B}}(f,c,\alpha)$, that a number, f, of the c comparisons are falsely judged significant is, then, given by the binomial distribution:

$$P_{\mathrm{B}}(f,c,\alpha)=\frac{c!}{f!(c-f)!}\alpha^f\cdot(1-\alpha)^{c-f} \qquad (8)$$

in which the mean value of f is α × c. This mean value is the expected number of "significant" ($P < \alpha$) SCFs that would be, mistakenly, reported by 2DCLASS for a random distribution. Equation 8 can, of course, be used to describe the above Ping-Pong ball experiment. In this case, f = 0, the probability of choosing a marked ball, is $\alpha = 0.1$, and the probability of choosing an unmarked ball is $(1 - \alpha) = 0.9$, so that the probability of choosing c unmarked balls in a row is $P_{\mathrm{B}}(0,c,\alpha) = (1 - \alpha)^c = 0.9^c$.

Binomial evaluation of the overall deviation from a random distribution. The number of distance-orientation classes, s, for which the SCFs as calculated either by 2DCLASS or through the use of equation A2 or 7 are found to be significantly different from expected ($P < 0.05$), can be used to determine the overall deviation from a random distribution. The procedure is to perform a summation over equation 8 to evaluate the probability that s or more of the comparisons are mistakenly judged significant:

$$P_{\mathrm{B}}(f\ge s,c,a)=\sum_{f=s}^{f=c}\frac{c!}{f!\cdot(c-f)!}\alpha^c\cdot(1-\alpha)^{c-f} \qquad (9)$$

in which the total number of comparisons is $c = L \times W - 1$, $\alpha = 0.1$, because both tails of the distribution (high and low) are tested at the 5% level and f is a summation variable. Before any interpretation of the results of a 2DCLASS analysis is made, equation 9 should first be applied to the distance-orientation class probability matrix to establish a significant deviation from random behavior. This protocol is analogous to the "protected" approach for multiple comparisons in which one does not proceed with the analysis unless there is a significant treatment effect (4,31).

The Bonferroni confidence limit. Even if the application of equation 9 has demonstrated that the data set deviates significantly ($P \le 0.05$) from a random distribution, the probability matrix may still be littered with spurious results. To ensure that there

TABLE 1. The application of Fisher's exact test (2 × 2 contingency table; 10) to estimate confidence limits using a chi-square (χ^2) test of independence with one degree of freedom

	Infected	Healthy	Total
Sample	I	$N - I$	N
Remainder	$I_T - I$	$N_T - I_T - N + I$	$N_T - N$
Total	I_T	$N_T - I_T$	N_T

is less than a 5% chance of even one spurious value, a more conservative confidence limit, β, must be applied (30), such that:

$$P_B(f \geq 1, c, \beta) = 1 - P_B(f = 0, c, \beta) = 0.05$$

which can be solved for β using equation 8 to yield: (10)

$$\beta = 1 - (1 - 0.05)^{1/c} = 1 - (0.95)^{1/c}$$

For the previous example of a field with 1,000 plants, $c = 1{,}000$ and equation 10 can be evaluated to yield $\beta = 0.0000513$, approximately equal to $0.05/c$. For the Monte-Carlo method employed by 2DCLASS to detect deviations from random at such a small level of probability, the number of numerical simulations must be increased by about a factor of c, so that on the order of 4×10^5 rather than 400 simulations must be run.

APPLICATION

An illustrative example. The hypothetical data set proposed by Gray et al. (14) is reproduced here as Figure 2, with $L = 12$ and $W = 8$. As a simple count of infected plants (filled symbols) will show, $i = 21$, and, since there are no missing plants, $n = 96$ (equation 1). Thus, the total number of pairs, N_T, is 4,560 (96 × 95/2 using equation 2), and the total number of infected pairs, I_T, is 210 (21 × 20/2 using equation 4). Direct pair counts give the distribution of these 210 infected pairs among the (j,k) (I_{jk}, Fig. 3A) and [X,Y] (I_{XY}, Fig. 3B) distance-orientation classes. Stochastic determination of the significant SCFs, as reported by Gray et al. (14), is shown in Figure 4. In constructing this probability matrix, a total of 95 ($L \times W - 1$) comparisons are made. Since both tails of the distribution (high and low) are tested at the 5% level ($\alpha = 0.05$), the probability of falsely claiming significant deviation from random behavior is 0.1 (2 × α) for each comparison (distance-orientation class). Thus, the expected number of false positives (Type I error) is 9.5 (0.1 × 95). This is to be compared with the 41 reported significant SCFs (21 larger than expected, "+" or "⊕", and 20 smaller than expected, "–", Fig. 4). In addition, using equation 9 (4,31), there is a 5% chance that as many as 15 of the 41 SCFs that Gray et al. (14) calculated to be significantly different from expected may be spurious. The question as to which 15 may be spurious casts serious aspersions on any geometrical interpretation of the pattern shown in Figure 4. On the other hand, the probability of obtaining 41 flagged SCFs from a random distribution of plants is less than one in a billion (equation 9). Thus, the disease pattern is definitely not random.

	1	2	3	4	5	6	7	8
1	○	○	○	○	○	○	○	○
2	○	○	○	○	○	●	●	●
3	○	○	○	○	○	●	●	●
4	○	○	○	○	○	○	●	●
5	○	○	○	○	○	○	○	○
6	○	○	○	○	○	○	○	○
7	●	●	○	○	○	○	○	○
8	●	●	○	○	○	○	○	○
9	●	●	○	○	○	○	○	○
10	○	○	○	○	○	●	○	●
11	○	○	○	○	○	●	●	●
12	○	○	○	○	○	●	●	○

Y

Fig. 2. Hypothetical data set on an 8 × 12-rectangular lattice, first introduced by Gray et al. (14). Infected plants are denoted by filled symbols healthy plants as open circles.

Using equation 10 with $c = 95$, I obtain $\beta = 0.00054$. Using equation A2 to reexamine the significance of the observed SCFs at this more conservative level of probability yields only 13 significant SCFs ("⊕", Fig. 4). By definition of β (equation 10), there is less than a 5% probability that even one of these 13 is spurious and, using equation 9, there is less than one chance in 10^{27} that 13 or more significant SCFs at this confidence level could be drawn from a random distribution.

It is interesting to note that none of the SCFs that had significantly less than the expected number of pairs at $\alpha = 0.05$ level of confidence ("–", Fig. 4), were significant at the more conservative $\beta = 0.00054$ level of confidence. A comparison of Figures 3B and 4 reveals that 19 of 20 minus signs in Figure 4 correspond to zero observed pairs (one has two pairs). The low number of counts makes it impossible to have significant differences on the low side with the more conservative confidence level (β).

A *j*

		-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
	0	.	.	.	.	.	4	11		11	4	.	.	.	.	.
	1	.	.	.	.	.	3	8	13	8	3	.	.	.	.	.
	2	1	1	1	1	.	1	4	5	3	1	.	.	1	2	1
	3	2	3	3	2	.	.	.	.	.	.	.	1	3	4	2
	4	2	4	5	3	.	.	.	.	.	.	.	2	5	6	3
	5	1	3	4	2	.	.	.	.	.	.	.	2	4	4	2
k	6	.	1	2	1	.	.	1	1	1	1	.	1	2	2	1
	7	.	.	.	.	.	1	2	4	3	2	.	.	.	.	.
	8	.	.	.	.	.	2	3	6	5	3	.	.	.	.	.
	9	.	.	.	.	.	1	3	5	4	2	.	.	.	.	.
	10	.	.	.	.	.	.	1	2	2	1	.	.	.	.	.
	11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

B *X*

		0	1	2	3	4	5	6	7
	0		11	4	.	.	.	.	.
	1	13	16	6	.	.	.	.	.
	2	5	7	2	.	1	2	3	2
	3	.	.	.	.	3	6	7	4
	4	.	.	.	.	5	10	10	5
	5	.	.	.	.	4	8	7	3
Y	6	1	2	1	.	2	4	3	1
	7	4	5	3	.	.	.	.	.
	8	6	8	5	.	.	.	.	.
	9	5	7	3	.	.	.	.	.
	10	2	3	1	.	.	.	.	.
	11	.	.	.	.	.	.	.	.

Fig. 3. Tabular representation of the infected pair counts in **A,** (j,k) notation (I_{jk}) and **B,** the [X,Y]-distance classes (I_{XY}) introduced by Gray et al. (14).

DISCUSSION AND CONCLUSIONS

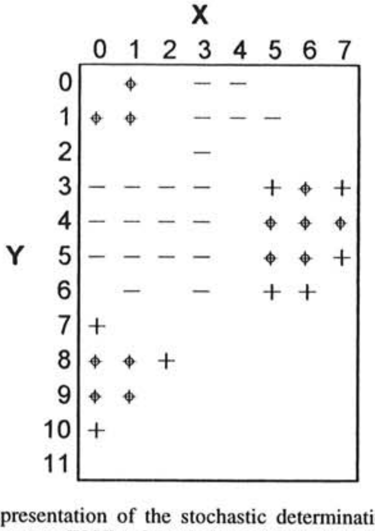
For disease-incidence data, the statistical determination of anomalously high or low pair counts within certain distance-orientation classes is a powerful and promising technique in the study of the two-dimensional distribution of disease. However, although I have presented an analytical approach to calculating expected pair counts and confidence limits, there are still some remaining problems with the method that future work must address.

Difficulties in the interpretation of the "core cluster" size and "reflected clusters" still remain (14,15). In particular, it appears that the "core cluster" size is a function of the size of the field (G. Hughes, *personal communication*, 29). This result is not too surprising given that the 2DCLASS method defines "core cluster" size in terms of a probabilistic limit that is intimately tied to the number of infected plants in the sample. Thus, both the size of the field, as well as the incidence of disease, would be expected to affect the "core cluster" size. It would be preferable to define a contagious area scale in some sample-size independent fashion.

In addition, because of the inherent spatial displacement between the regions where reference plants and target plants are located, each member of an infected plant pair is drawn from a different population with possibly different levels of infection. This problem is especially severe when anomalous behavior, either high density clumps of infected plants or areas with very low disease incidence, are in close proximity to the boundary. The manner by which this boundary effect confounds the detection of nonrandom behavior at differing length scales must also be investigated.

Because of these remaining difficulties, one must flavor interpretation of point patterns with caution. This involves the use of conservative confidence limits and, of course, well-planned and adequately replicated experiments that pay heed to the effects of the plot boundary.

I present here the mathematical tools needed to examine the results of a two-dimensional distance-orientation class analysis of disease-incidence data. These include a method to ascertain the overall deviation from random behavior (equation 9) and the means to calculate more conservative confidence limits on infected pair counts without an undue penalty in increased computer time (equations A2 and 7). The use of this technique will reduce spurious results and extend the applicability of the method over a larger range of disease incidence.



X

Y	0	1	2	3	4	5	6	7
0		⌽		−	−			
1	⌽	⌽		−	−	−		
2				−				
3	−	−	−	−		+	⌽	+
4	−	−	−	−		⌽	⌽	⌽
5	−	−	−	−		⌽	⌽	+
6		−		−		+	+	
7	+							
8	⌽	⌽	+					
9	⌽	⌽						
10	+							
11								

Fig. 4. Tabular representation of the stochastic determination of the standardized count frequency [X,Y] values that are significantly larger ($P < 0.05$, "+" and "⌽") or significantly smaller ($P < 0.05$, "–") than expected, as calculated using 2DCLASS (14). Values significant at the more conservative $\beta < 0.00054$ are also shown (calculated using equations 2–5, "⌽").

APPENDIX

The Hypergeometric distribution. The question to be answered is: How many of the total number of infected pairs, I_T, are to be found in a certain subset of size N, selected from an equally probably total number of plant pairs, N_T? Let I be the number of infected pairs meeting this requirement.

In general, the number of ways to choose n indistinguishable objects from a population of N such objects is given by the combinatorial or binomial coefficient, $\mathcal{C}(n,N)$:

$$\mathcal{C}(n,N)=\binom{N}{n}=\frac{N!}{n!\cdot(N-n)!}$$

in which

$$N!=\prod_{m=1}^{N}m=N\cdot(N-1)\cdot(N-2)\cdot\ \ldots\ \cdot 1$$

in which m is a dummy integer variable. If all outcomes are equally likely, then the probability of an event is the ratio of number of successful events to the total number of events. In this case, a successful event has I infected pairs within subset N. The total number of successful events is equal to the product of the number of ways ($W_I = \mathcal{C}(I,I_T)$) to choose I infected pairs from a population of size I_T and the number of ways ($W_{NI} = \mathcal{C}[(N - I),(N_T - I_T)]$) to choose the remaining ($N - I$) noninfected pairs from a population of size ($N_T - I_T$). The total number of ways ($W_T = \mathcal{C}(N,N_T)$) to choose the N pairs of plants from the total number of pairs, N_T, is the total number of possible events. Therefore, the probability that exactly I pairs fulfill the above requirements is given by:

$$P_{\text{HG}}\left(I,I_T,N,N_T\right)=\frac{W_I W_{NI}}{W_T}=\frac{\binom{I_T}{I}\binom{N_T-I_T}{N-I}}{\binom{N_T}{N}} \tag{A1}$$

$$\text{in which } \binom{N}{n}=\frac{N!}{n!(N-n)!} \text{ so that}$$

$$P_{\text{HG}}\left(I,I_T,N,N_T\right)=\frac{I_T!\cdot N!\cdot\left(N_T-I_T\right)!\cdot\left(N_T-N\right)!}{I!\cdot\left(I_T-I\right)!\cdot\left(N-I\right)!\cdot\left(N_T-I_T-N+I\right)!\cdot N_T!} \tag{A2}$$

in which N_T,I_T are given by equations 2 and 4, respectively. Equation A2 defines the Hypergeometric distribution, ($P_{\text{HG}}(I,I_T,N,N_T)$; 5,30), to be interpreted as the probability of exactly I successful events within a subpopulation of size N drawn from a population of size N_T containing I_T successful events. Although formidable in appearance, equation A2 is easily evaluated and the Hypergeometric function is supplied with all major spreadsheet packages (e.g., in QuattroPro: @HYPGEOMDIST(a,b,c,d)). It is interesting to note that equation A2 gives the same numerical answer if I_T and N are interchanged. This reflects the fact that we can judge a successful event either by whether or not the pair is infected or by whether or not the pair belongs to the appropriate subsample.

Under the condition that population values are much larger than the sample value (i.e., N_T,I_T >> N,I), equation A2 reduces to the binomial distribution ($P_{\text{B}}(N,I,p)$):

$$P_{\mathrm{B}}(N,I,p)=\binom{N}{I}p^{I}(1-p)^{(N-I)} \qquad \text{(A3)}$$

in which $p = I_T/N_T$ is the probability that a randomly chosen pair is infected. If infected pairs are also a small fraction of the total pair count (i.e., $N_T >> I_T$ and $N >> I$), then equation A3 reduces further to the Poisson distribution ($P_{\mathrm{Poisson}}(I,\lambda)$):

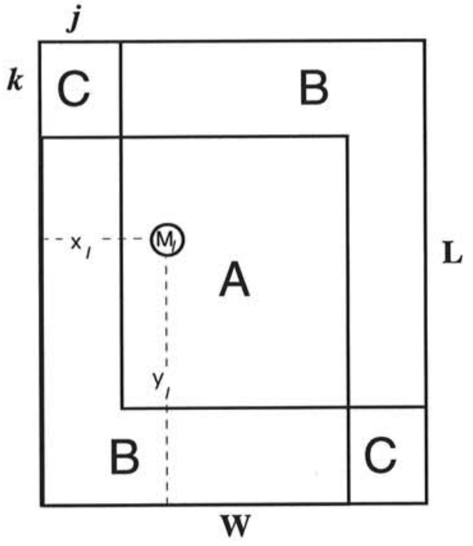


Fig. 5. Schematic representation of the effect of missing plants on pair counts corresponding to distance-orientation class (j,k) (equation A5). Missing plants in region A can reduce possible pair count by two, since missing plant can act as either a reference or a target plant. Missing plants in regions B can, at most, reduce possible pair count by one, since the missing plant can act as either a reference or a target plant, but not both. Missing plants in regions C do not affect pair counts.

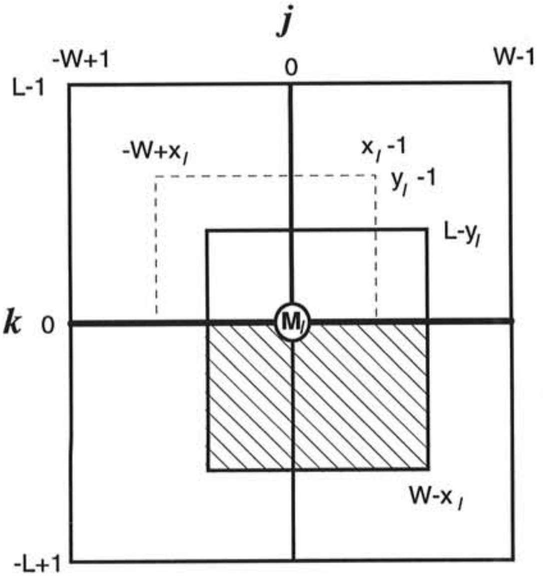


Fig. 6. Schematic representation of distance-orientation classes affected by a missing plant at position (x_l,y_l) (circled "M_l", Fig. 5) in the plant-centered view. Dashed rectangle represents the reflection of hatched area into the upper half of $j - k$ plane. The number of pairs within distance classes lying within both the dashed rectangle and the solid rectangle in the upper half plane are reduced by two. Whereas, the number of pairs within distance classes lying within either the dashed rectangle or the solid rectangle, but not both, are reduced by one.

$$P_{\mathrm{Poisson}}(I,\lambda)=\frac{\lambda^{I}}{I!}e^{-\lambda} \qquad \text{(A4)}$$

in which $\lambda = (NI_T)/N_T$ is the value of the mean for all of the above probability distributions (equations A2, A3, and A4). The variances for the three distributions are as follows: $\lambda(1 - p)(N_T - N)/(N_T - 1)$ for equation A1b, $\lambda(1 - p)$ for equation A2, and λ for equation A3. Thus, the distribution becomes more dispersed (variance gets larger) in going from Hypergeometric distribution to the Binomial and Poisson approximations.

Pair enumeration. As discussed above, if there are no missing plants ($n_m = 0$), then the total number of pairs of plants within distance-orientation class (j,k), N_{jk}, is given by equation 4. Missing plants will reduce this number. In Figure 5, the rectangular lattice is conceptually divided into three regions (A, B, and C; Fig. 5). If a missing plant is located within area C, there is no effect on the total number of pairs. A missing plant located within area B, however, can reduce N_{jk} by, at most, one pair, and a missing plant located within area A can reduce N_{jk} by as much as two plant pairs. Basically, plants in region A can act as both a reference and a target plant, since they lie within both reference and target rectangles (Fig. 1). Plants in region B can act as either reference or target plants, but not both. To calculate the total number of plant pairs, I, first define $n_{m\mathrm{A}}$, $n_{m\mathrm{B}}$, and $n_{m\mathrm{C}}$ to be the number of missing plants in rectangular regions A, B, and C, respectively (Fig. 5). Potentially, the total number of pairs can be reduced by as much as $2n_{m\mathrm{A}} + n_{m\mathrm{B}}$. However, some of the missing plants may pair with other missing plants. The procedure is, then, to count the number of missing-missing plant pairs that fall into distance-orientation class (j,k), M_{jk}. Then the total number of pairs of plants within distance-orientation class (j,k), N_{jk}, is given by:

$$N_{jk} = (W - |j|)\cdot(L - |k|) - 2n_{m\mathrm{A}} - n_{m\mathrm{B}} + M_{jk} \qquad \text{(A5)}$$

This calculation (equation A5) must be carried out for each distance-orientation class (j,k).

Because there are usually many more distance-orientation classes than missing plants, there is a more convenient algorithm than equation A5 to calculate N_{jk} on a computer. The first step is to construct the N_{jk} matrix ($-W + 1 < j < W + 1$ and $-L + 1 < k < L - 1$) and fill it with the appropriate values (equation 4) for the case in which there are no missing plants. Then a list containing the x-y coordinates of each of the missing plants is constructed. Let x_l and y_l be the x and y coordinates, respectively, of the lth missing plant (circled M_l, Fig. 5). In the coordinate system centered on this missing plant, the field appears as an off-centered rectangle (Fig. 6). For each of the missing plants, two passes through the N_{jk} matrix are then made. On the first pass, matrix elements such that $-W + x_l \le j \le x_l - 1$ and $0 \le k \le y_l - 1$ are reduced by one (dotted rectangle, Fig. 6). This accounts for the loss of possible pairs for which the missing plant is the target plant. On the second pass, matrix elements such that $-x_l +1 \le j \le W - x_l$ and $0 \le k \le L - y_l$ are reduced by one (solid unhatched rectangle, Fig. 6). This accounts for the loss of possible pairs for which the missing plant is the reference plant. This stepwise reduction of the appropriate N_{jk} matrix elements continues until all the missing plants have made their contribution. Finally, M_{jk} is evaluated by direct counting over all possible pairs of missing plants and added element-wise to the N_{jk} matrix.

LITERATURE CITED

1. Anscombe, F. J. 1950. Sampling theory of the negative binomial and logarithmic series distributions. Biometrika 37:358-382.
2. Chapham, A. R. 1936. Over-dispersion in grassland communities and use of statistical methods in plant ecology. J. Ecol. 24:232-251.
3. Cohen, W. B., Spies, T. A., and Bradshaw, G. A. 1990. Semivariograms of digital imagery for analysis of conifer canopy structure. Remote Sens.

Environ. 34:167-178.

4. Day, R. W., and Quinn, G. P. 1989. Comparisons of treatments after an analysis of variance in ecology. Ecol. Monogr. 59:433-463.
5. Feller, W. 1968. Elements of combinatorial analysis. Pages 26-53 in: An Introduction to Probability Theory and Its Applications. Vol. 1, 3rd ed. John Wiley & Sons, Inc., New York.
6. Ferrandino, F. J. 1989. Spatial and temporal variation of a defoliating plant disease and reduction in yield. Agric. For. Meteorol. 47:273-290.
7. Ferrandino, F. J. 1989. A distribution-free method for estimating the effect of aggregated plant damage on crop yield. Phytopathology 79:1229-1232.
8. Ferrandino, F. J. 1993. Dispersive epidemic waves: I. Focus expansion within a linear planting. Phytopathology 83:795-802.
9. Ferrandino, F. J., and Elmer, W. H. 1992. Reduction in tomato yield due to Septoria leaf spot. Plant Dis. 76:208-211.
10. Fisher, R. A., and Yates, F. 1948. Statistical Tables. 3rd ed. Hafner Publishing Co., Inc., New York.
11. Gottwald, T. R. 1995. Spatio-temporal analysis and isopath dynamics of citrus scab in nursery plots. Phytopathology 85:1082-1092.
12. Gottwald, T. R., Avinent, L., Llácer, G., Hermoso de Mendoza, A., and Cambra, M. 1995. Analysis of the spatial spread of sharka (plum pox virus) in apricot and peach orchards in eastern Spain. Plant Dis. 79:266-278.
13. Gottwald, T. R., Cambra, M., Moreno, P., Camarasa, E., and Piquer, J. 1996. Spatial and temporal analyses of citrus tristeza virus in eastern Spain. Phytopathology 86:45-55.
14. Gray, S. M., Moyer, J. W., and Bloomfield, P. 1986. Two-dimensional distance class model for quantitative description of virus-infected plant distribution lattices. Phytopathology 76:243-248.
15. Gray, S. M., Moyer, J. W., Kennedy, G. G., and Campbell, C. L. 1986. Virus-suppression and aphid resistance effects on spatial and temporal spread of watermelon mosaic virus 2. Phytopathology 76:1254-1259.
16. Greig-Smith, P. 1952. The use of random and contiguous quadrats in the study of the structure of plant communities. Ann. Bot. (Lond.) 16:293-316.
17. Hughes, G. 1988. Spatial heterogeneity in crop loss assessment models. Phytopathology 78:883-884.
18. Hughes, G., and Madden, L. V. 1993. Using the beta-binomial distribution to describe aggregated patterns of disease incidence. Phytopathology 83:759-763.
19. Hurlbert, S. H. 1990. Spatial distribution of the montane unicorn. Oikos 58:257-271.
20. Legendre, P., and Fortin, M. J. 1989. Spatial pattern and ecological analysis. Vegetatio 80:107-138.
21. Madden, L. V., and Hughes, G. 1995. Plant disease incidence: Distributions, heterogeneity and temporal analysis. Annu. Rev. Phytopathol. 33:529-564.
22. Morisita, M. 1959. Measuring of the dispersion of individuals and the distribution patterns. Mem. Fac. Sci. Kyushu Univ. Ser. E Biol. 2:215-235.
23. Nelson, S. C. 1995. Spatiotemporal distance class analysis of plant disease epidemics. Phytopathology 85:37-43.
24. Nelson, S. C., and Campbell, C. L. 1993. Comparative spatial analysis of foliar epidemics on white clover caused by viruses, fungi, and a bacterium. Phytopathology 83:288-301.
25. Nelson, S. C., Marsh, P. L., and Campbell, C. L. 1992. 2DCLASS, a two-dimensional distance class analysis software for the personal computer. Plant Dis. 76:427-432.
26. Ripley, B. D. 1979. Modeling spatial patterns. J. R. Stat. Soc. B 39:172-212.
27. Ripley, B. D. 1981. Spatial Statistics. John Wiley & Sons, Inc., New York.
28. Ristaino, J. B., Larkin, R. P., and Campbell, C. L. 1993. Spatial and temporal dynamics of *Phytophthora* epidemics in commercial bell pepper fields. Phytopathology 83:1312-1320.
29. Samita, S. 1995. Analysis of aggregated plant disease incidence data. Ph.D. dissertation. University of Edinburgh, Scotland.
30. Sokal, R. R., and Rohlf, F. J. 1981. Tests of independence: Two-way tables. Pages 731-744 in: Biometry, 2nd ed. W. H. Freeman & Co., San Francisco.
31. Stewart-Oaten, A. 1995. Rules and judgments in statistics: Three examples. Ecology 76:2001-2009.
32. Taylor, L. R. 1961. Aggregation, variance and the mean. Nature 189:732-735.
33. Wagonner, P. E., and Rich, S. 1981. Lesion distribution, multiple infection, and the logistic increase of plant disease. Proc. Natl. Acad. Sci. U.S.A. 78:3292-3295.