Past Nonrandomness and Aggregation to Spatial Correlation: 2DCORR, a New Approach for Discrete Data

Francis J. Ferrandino

The Connecticut Agricultural Experiment Station, P.O. Box 1106, New Haven 06504.

Accepted for publication 14 November 1997.

The spatial disposition of disease within a field has major consequences on both the rate of development and the extent of spatial spread of the ensuing epidemic, and, consequently, on the ultimate impact of disease on yield (8–12,23). Thus, it is useful to describe spatial patterns of disease incidence in an objective and, hopefully, descriptive manner. During the last 50 years a great deal of time and effort has been expended to quantify and interrelate measurements of the spatial properties of points distributed on a plane (1–4,21,22,27–29,34–36,41,42). These points can represent diseased plants, individuals of a wildebeest herd, or trees in a forest. Unfortunately, the biological meaning represented by proposed indices of nonrandomness, aggregation, patchiness, intensity, or covariance may not be as easy to specify.

The starting point of most of these analyses of spatial structure is the examination of some measure of deviation from spatially random behavior. This kind of negative approach is not necessarily informative. The ultimate goal is to reject the hypothesis of random behavior and triumphantly declare the data set to be nonrandom. However, what is meant by a nonrandom pattern?

By definition, a random process is one in which every possible outcome is equally likely to occur (7). Therefore, a nonrandom process must favor certain types of outcomes. The characterization of these more frequent outcomes in a descriptive and biologically meaningful way has proven difficult. The common property that favored patterns possess may be given a name (e.g., aggregation), but ultimately, the nature of their commonality is deeply intertwined with the mathematical form of the index used to establish deviation from random behavior. Contrary to conventional wisdom (2,39), the myriad of possible spatial patterns of disease that can occur within a 2D field cannot be restricted by a simple 1D classification scheme proceeding from “regular” to “clustered” (i.e., aggregated or contagious), with “random” somewhere in between (26). For example, consider a regular array of point clusters (Fig. 1). This pattern is both regular and aggregated to some extent and is certainly not random. Spatial structure can occur in many forms, and it is unreasonable to believe that a single measure can sort it all out. Pielou (34) said it best, “the phrase ‘degree of aggregation’ describes a vague, undefined notion that is open to several interpretations.... ways of measuring aggregation are not different methods of measuring the same thing: they measure different things.” The question requiring an answer may not be how a calculation is to be done, but rather, exactly what property is to be measured?

The spatial properties of the point pattern illustrated in Figure 1 suggest a two-pronged description of spatial structure, in which the notions of aggregation and regularity are examined independently. The first of these concepts deals with the variability of point density within fixed areas of a field. Consider circular area *a* in Figure 1. The number of points within circle *a*, for this example, can vary from zero to seven, depending on its location. If the point pattern is clustered in little groups, then we would expect to obtain both large numbers of points within the circle when it is aligned with clusters and few or zero points in the circle when it is between clusters. Thus, clustering of points increases the variability of the expected number of points in an arbitrary circle. The more tightly packed the clusters are, the more extreme the variation. On the other hand, if the points are spread homogeneously throughout the field, then the number of points in the circle tends to hover around the mean value, with very little variation. The variability in the number of points within a randomly placed circle measures what Pielou (34) termed the “intensity of the spatial pattern.” By definition, this variability is explicitly dependent on the physical size of the quadrat. Areas with less than an average number of points are just as important in increasing the variability as areas that are densely populated.

However, there is a problem with the method described above. Consider a field consisting of infected and healthy plants. If this field is subdivided conceptually into quadrats, then the frequency distribution of the number of infected plants within each quadrat can be analyzed. This analysis may result in the variance to mean ratio (3,21,34), Morisita’s index (29), or the parameters of a beta-binomial distribution (24). If one pictures the field as a slide puzzle consisting of movable quadrats, then the results of any one of these analyses remains unchanged if the quadrats are rearranged in space. In particular, if the quadrats are sorted so the quadrats with the least disease are at the north end and those with the most disease are at the south end, then the resultant spatial arrangement is clearly ordered. In addition, one can rearrange infected plants within each quadrat so as many infected plants as possible are adjacent to another infected individual. Once again, this is a clear case of ordered behavior that is undetectable in quadrat analysis.

In summary, quadrat methods are insensitive to disease density changes on a length scale larger or smaller than the quadrat size and to spatial alignment of infected plants within quadrats. For all of the above reasons, I prefer the term “scale-dependent spatial heterogeneity” to describe this property of a 2D pattern of points. There are many examples of this type of analysis in the literature (21,24,25,28,34), and the method is particularly useful when there is a natural sampling unit. For example, this unit is a single plant when one wants to examine the effect of the spatial variability of damage due to disease on yield (9,12,23).

The second concept, regularity, has to do with the relative position of points within the field irrespective of actual location. Consider a point-sampling apparatus consisting of a stick of a certain

Additional keywords: 2DCLASS, disease incidence, distance class analysis, spatial pattern, STCLASS.

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

Publication no. P-1997-1223-01O

© 1998 The American Phytopathological Society

length with a small loop attached to either end (b in Fig. 1). We move this device over the entire field, keeping its orientation constant, and count how many times points simultaneously appear in both sampling loops. The pair count obtained is a function of the length of the stick and the angle at which it is held (distance sampling sensu Pielou [34]). I contend that the behavior of this property is much more closely related to the dispersal of inoculum than are measures of spatial heterogeneity. Thus, the orientation and distance between infected plants may be a more important consideration for the construction of epidemiological models. For this reason, I wish to concentrate on this aspect of spatial behavior in this letter.

One plant-centered approach based on the observed number of infected-infected plant pairs used for discrete data is 2D spatio-temporal distance class analyses (2DCLASS and STCLASS, respectively [19,20,30–32]). The pair counts are arranged in a number of distance-orientation classes characterized by the distance and spatial orientation between plants (the length of the stick and the angle at which it is held; b in Fig. 1). Such a vector-based approach has the promise of providing geometric information on the relative position of infected individuals. Although it is not at first obvious, plant pair-counting methods also make use of quadrats. For each distance-orientation class, infected plant pairs are separated by a certain distance vector. Not all the plants in the field can act as either the head or tail of such a vector, and some plants can act as both. Because of this geometric fact, the field is implicitly subdivided into possible reference-plant sites, possible target-plant sites, sites that can act as either a reference or a target, and sites that cannot contribute for each distance-orientation class under consideration (11). Many recent papers on disease incidence have used this map-like representation of deviations from random behavior (16–18,37); however, this method is very sensitive to boundary effects and often gives spurious “reflected clusters” (sensu Nelson and Campbell [32]) that can be difficult to interpret (11). The above method also yields a spatial measure of the scale over which contagion is important, the “core cluster size.” Unfortunately, this length scale is defined on a probabilistic basis and, as such, may depend on plot size and disease incidence (38).

In what follows, I will show that both reflected clusters and edge effects are the result of an interaction between a large-scale structure within the field and the assumed equivalence of reference- and target-plant populations. Because the reference- and target-plant populations are different, the actual disease incidence of either of these populations may differ from the overall average and from each other. Such heterogeneity will affect the expected number of infected pairs. Rather than attempting to quantify the degree of spatial heterogeneity at any length scale, I have chosen to correct for the bias it introduces into the infected pair counts. Here, I define a new pair-count method, 2DCORR, which accounts for differences between reference- and target-plant populations. I show that this approach is equivalent to a total reckoning of plant-plant pairs into healthy-healthy, infected-healthy, healthy-infected, and infected-infected categories. This approach is the discrete analog of correlation analysis applied to a 2D lattice of plants. The method properly handles problems associated with anomalous densities near the plot boundary and totally eliminates the reflected clusters that plague previous pair-count methods. In addition, I present a Kolmogorov-Smirnov-type analysis based on the cumulative probability density function for the total number of infected-infected plant pairs within a given distance (5). This method independently yields an estimate for the length scale over which disease is correlated and the probability of deviation from a random spatial distribution.

A SIMPLE EXAMPLE

I present a simple example to illustrate the confounding effect of the unequal distribution of disease on the observed number of

infected plant pairs. Consider a field containing a total of 60 plants, N_T , consisting of two side-by-side 30-plant rows. Assume that the total number of infected plants in the field, I_T , is 28, of which 20 are in row A and 8 are in row B, I_A and I_B , respectively (Fig. 2). A direct comparison of the two rows indicates that 8 of the 30 side-by-side plant pairs consist of 2 infected plants ($I_{AB} = 8$). At this point, one might ask whether these eight aligned pairs could have resulted from pure happenstance (2DCLASS [19,20,30,32,33]). A certain fraction of all the possible ways to arrange the 28 infected plants among the 60 sites will result in 8 or more infected side-by-side plant pairs. If this fraction (probability) is large enough, then one is forced to accept the possibility of random behavior. For this simple case, the number of possibilities is directly enumerated by combinatorial analysis in the appendix. From this analysis, one obtains the probability frequency distribution for obtaining I_{AB} infected side-by-side plant pairs for two plant rows containing a total of N_T plants, of which I_T are infected ($p(I_{AB}, I_T, N_T)$ [equation A1, appendix]). A plot of p versus I_{AB} for $N_T = 60$ and $I_T = 28$ is illustrated in Figure 3 (top panel, bar graph). Also shown is the cumulative frequency distribution (Fig. 3, top panel, dashed line with filled circles). The expected number of infected pairs is 6.41, and the probability of obtaining eight or more infected pairs is 0.21 (Fig. 3, top panel, filled bars and arrow). Therefore, there is about a 21% chance that eight or more infected pairs resulted from pure serendipity. This is much too large a probability to ignore, and one is forced to accept the fact that the observed number of infected pairs could have been drawn from a random distribution.

I suggest an alternative step-wise approach to this analysis. The first question asked is, does the apportionment of disease between rows suggest spatial heterogeneity? Given the answer to the first question, one asks a second question: is there a tendency for infected plants to be aligned?

In what follows, I will show that the above simple examination of the number of infected pairs (equation A1, appendix) does not address either of these questions directly but rather combines both tendencies (heterogeneity and alignment) in a rather convoluted fashion.

First consider the apportionment of infected plants between the two rows. Because half of the 60 plants are in each row, one

Fig. 1. Hypothetical spatial pattern of points that is both “regular” and “clustered.” Also shown are schematics for quadrat sampling (a) and point pair sampling (b; described in text). The results of a quadrat sampling-based analysis explicitly depend on the relative size of the quadrats compared to the cluster size and the separation between clusters in the spatial point pattern.

Now on to question 2. Because there are only eight infected plants in row B (Fig. 2), it is impossible to have more than eight infected plant pairs. This must have a major impact on the probability of obtaining eight or more infected plant pairs. To address this effect, I pose a new question: given the fact that there are I_A and I_B infected plants in rows A and B, respectively, could the observed number of infected pairs ($I_{AB} = 8$) have resulted from pure chance? A certain fraction of all the possible ways to arrange the infected plants in row A ($I_A = 20$, $N = 30$) and the infected plants in row B ($I_B = 8$, $N = 30$) will result in $I_{AB} = 8$ infected side-by-side plant pairs. These possibilities are enumerated in the appendix. The fraction obtained represents the conditional probability distribution function ($\rho'(I_{AB}, I_A, I_B, N)$ [equation A3, appendix])

In the above example, the unevenness of disease between the two rows tended to reduce the observed number of infected pairs of plants, and the alignment tended to increase this value. Thus, the simple comparison of the observed number of infected pair counts to the expected value when infected plants are randomly placed within the field is governed by both heterogeneity and alignment. It is only when the bias introduced by heterogeneity is included in the analysis that the tendency for like plants (infected or healthy) to be side by side is significant.

Figure 1 consists of two vertically stacked plots. Both plots share the same x-axis, 'Number of Infected Pairs', ranging from 0 to 14. The left y-axis for both is 'Probability Density Function', ranging from 0 to 0.3. The right y-axis for both is 'Cumulative Distribution Function', ranging from 0 to 1. The top plot is labeled 'Equation A1' and shows a distribution with a peak at 6, indicated by a value of 0.21. The bottom plot is labeled 'Equation A3' and shows a distribution with a peak at 5, indicated by a value of 0.022. Both plots include a dashed line representing the cumulative distribution function and a solid line representing the probability density function.

Fig. 3. Probability density function (histogram, left y-axis) and cumulative distribution function (dashed line, right y-axis) for the number of infected pairs (I_{AB}), calculated by equation A1 in the appendix (top panel, 2DCLASS), and for the number of infected pairs, calculated by equation A3 in the appendix (bottom panel, 2DCORR), for the data shown in Figure 2. Equation A3 accounts for the unevenness in the distribution of disease (described in text). Filled bars represent pair counts greater than or equal to observed ($I_{AB} = 8$). Total fractional area of the filled bars are numerically labeled and shown by arrows on the cumulative curves.

of infected reference and target plants, respectively. The probability that a randomly chosen target is infected is given by the ratio of I_B to N . Because each of the I_A infected reference plants have an equal probability (I_B/N) of being part of an infected pair, the expected value for the number of infected plant pairs ($\langle I_{AB} \rangle$) is given by

$$\langle I_{AB} \rangle = \frac{I_A \cdot I_B}{N} \quad (1)$$

(see equation A3, appendix). Equation 1 directly includes the observed disease densities in the field and, thus, explicitly accounts for the bias caused by an uneven distribution of infected plants. The 2DCORR procedure is used to compare the observed number of infected pairs to the value predicted by equation 1. The significance of deviation from the expected is evaluated by Fisher's exact test (11,13,39). This is done by performing a chi-square comparison test (Fig. 3) with 1 df, in which the value of chi-square (χ^2) is given by

$$\chi^2 = \frac{\left(\left| I_{AB} \cdot N - I_A \cdot I_B \right| - N/2 \right)^2 \cdot N}{I_A \cdot I_B \cdot (N - I_A) \cdot (N - I_B)} \quad (2)$$

The $N/2$ term is Yates' correction for continuity, which accounts for the fact that the distribution is discrete (13). The four terms in the denominator are the marginal sums of the appropriate 2×2 contingency table in which infected-infected pairs (I_{AB}), infected-healthy pairs ($I_A - I_{AB}$), healthy-infected pairs ($I_B - I_{AB}$), and healthy-healthy pairs ($N - I_A - I_B + I_{AB}$) are enumerated (Fig. 4). When applying equation 2, the infected reference (I_A) and target (I_B) plant populations must be counted in addition to the number of infected pairs (I_{AB}) and the total number of pairs (N) for each distance-orientation class examined. Missing plants are accounted for as prescribed by Ferrandino (11). Please contact the author to obtain software for the above analysis.

To illustrate the analogy between the chi-square analysis outlined above (equation 2) and correlation analysis, consider the product moment correlation (R_{AB}) between two vectors, A and B , which by definition (39) is

$$R_{AB} = \frac{\sum_{i=1}^N (A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_{i=1}^N (A_i - \bar{A})^2 \sum_{i=1}^N (B_i - \bar{B})^2}} \quad (3)$$

where

$$\bar{A} = \frac{\sum_{i=1}^N A_i}{N} \quad \text{and} \quad \bar{B} = \frac{\sum_{i=1}^N B_i}{N}$$

Consider the special case where A_i and B_i take on the discrete values of either 0 or 1. If the reference or target plant of the i th plant pair is healthy or infected, respectively, then equation 3 becomes

		Totals
	I_{AB}	$I_B - I_{AB}$
	$I_A - I_{AB}$	$N - I_A - I_B + I_{AB}$
Totals	I_A	$N - I_A$
		N

Fig. 4. The 2×2 contingency table for the four possible outcomes of plant pairing, i.e., infected-infected (I_{AB}), infected-healthy ($I_A - I_{AB}$), healthy-infected ($I_B - I_{AB}$), healthy-healthy ($N - I_A - I_B + I_{AB}$) pairs.

$$R_{AB} = \frac{\sum_{i=1}^N A_i B_i - N \bar{A} \bar{B}}{\sqrt{\left(\sum_{i=1}^N A_i^2 - N \bar{A}^2 \right) \left(\sum_{i=1}^N B_i^2 - N \bar{B}^2 \right)}} = \frac{N I_{AB} - I_A I_B}{\sqrt{I_A (N - I_A) I_B (N - I_B)}} \quad (4)$$

where

$$I_{AB} = \sum_{i=1}^N A_i B_i, \quad I_A = N \bar{A} = \sum_{i=1}^N A_i = \sum_{i=1}^N A_i^2 \quad \text{and} \quad I_B = N \bar{B} = \sum_{i=1}^N B_i = \sum_{i=1}^N B_i^2$$

A comparison of the square on the far right side of equation 4 with the chi-square value defined above (equation 2) shows that aside from the Yates' continuity correction $\chi^2 = N \cdot R_{AB}^2$.

APPLICATION OF THE METHOD

As an illustration, I will apply the 2DCORR method to the nettle-head diseased hop plant field (Fig. 5) (14) first analyzed by Gray et al. (19) using 2DCLASS. This field consisted of an 11×30 rectangular lattice of plant sites. Of the sites, 12 were empty and 121, I_T , of the remaining 318 plants, N_T , were infected (Fig. 5). In the application of a 2DCORR or 2DCLASS analysis, one first constructs a matrix consisting of the predicted number of infected pairs for all the possible distance-orientation classes. This matrix

Fig. 5. Illustration of the 11×30 nettle-head diseased hop field used in the text (14,19,20). Healthy plants are denoted by open circles, infected plants are denoted by circles with included dots, and missing plants have no symbol. The reference and target populations for distance-orientation class correspond to target plants 20 plant units south (down) and 9 plant units east (right) of the reference plant.

is compared, element by element, to the observed pair count by a chi-square analysis (2DCORR [equation 2]; 2DCLASS [equation 7 in Ferrandino (11)]).

For illustrative purposes, I will consider the chi-square analyses for plant-plant pairs such that the target plant is 20 plant units

south and 9 plant units east of the reference plant (Fig. 5, vector). The reference plants are constrained to the 20-member 2×10 rectangle in the upper left corner of Figure 5, whereas target plants must lie in a similar rectangle in the lower right corner of the figure. A simple count reveals that $N = 20$, $I_A = 14$, $I_B = 11$, and $I_{AB} = 10$.

Fig. 6. Illustration of both **A** and **B**, 2DCLASS (upper panels) and **C** and **D**, 2DCORR (lower panels) analyses of the nettle-head diseased hop field (14) shown in Figure 3. **A** and **C** are 3D plots of the probability (z-axis) of deviation from a random arrangement as a function of the x and y distances between infected plant pairs calculated by **A**, 2DCLASS and **C**, 2DCORR. **B** and **D** schematically represent deviations at the 0.05 level of probability (+ or – signify observed pair counts are greater or less than expected, respectively). **B** represents significance at the Bonferroni level of probability (described in text).

The expected numbers of infected pairs using 2DCLASS (11) and 2DCORR are

$$\langle I_{AB} \rangle_{2DCLASS} = N \cdot SCF = N \cdot \frac{I_T(I_T - 1)}{N_T(N_T - 1)} = 20 \cdot \frac{121(120)}{318(317)} = 2.881$$

$$\langle I_{AB} \rangle_{2DCORR} = \frac{I_A \cdot I_B}{N} = \frac{14 \cdot 11}{20} = 7.7$$

where the standardized count frequency (SCF) is the ratio of the expected number of infected pairs to the total plant pairs. SCF was determined stochastically by Gray et al. (19,20) and Nelson (30), Nelson and Campbell (32), and Nelson et al. (33). However, in the above calculation, I used an analytically defined value for SCF (11). The great disparity between the two predictions for I_{AB} is due to the overly high incidence of disease in these two rectangles. The average incidence over the entire field is 38% ($100\% \cdot I_T/N_T$), whereas the incidence for the reference and target rectangles are 70% ($100\% \cdot I_A/N$) and 55% ($100\% \cdot I_B/N$), respectively (Fig. 5). Comparing the observed value (10 pairs) to the predicted values for I_{AB} using 2DCORR analysis (7.7 pairs) results in a value of $\chi^2_{2DCORR} = 3.117$ ($P < 0.0775$ [equation 1]), whereas 2DCLASS analysis (2.881 pairs) results in a value of $\chi^2_{2DCLASS} = 17.78$ ($P < 0.000025$ [equation 7 in Ferrandino (11)]).

Similar analyses are conducted for all possible interplant distances and orientations. The results of these comparisons are usually arranged in a map-like presentation of distance-orientation classes whose observed pair counts differ significantly from expected (Fig. 6). The complete results of both 2DCLASS and 2DCORR analysis are illustrated in Figure 6B and D, where significant deviations from expected at the 0.05 level of probability (one tailed) are shown (+ or – signify that observed pair counts are greater or less than expected, respectively). In examining the 2DCLASS results (Fig. 6A and B), one immediately notices two reflected clusters that are absent from the 2DCORR analysis (Fig. 6C and D). Both of these clusters are manifestations of the spatial heterogeneity of disease in the field (Fig. 5). The sharp positive peak in the lower right corner of Figure 6A and B (+ values: $X = 6$ to 10 , $Y = 12$ to 24) corresponds to the fact that both the upper left and the lower right quadrants of the field have a higher than average disease incidence (Fig. 5). In a similar fashion, the rather extended region of less than predicted infected pairs (– values in the lower left corner of Figure 6B) correspond to the fact that the lower left quadrant in Figure 3 has very little disease, whereas the upper right quadrant is heavily infected. When these variations in the population are taken into account (2DCORR; Fig. 6C and D), the spurious peaks are no longer present.

In addition to eliminating reflected clusters, the 2DCORR analysis considerably enhances the core cluster (the peak around the origin: $X = 0$, $Y = 0$ [30,32,33]) (Fig. 6). The reason for this increased sensitivity is the decreased error in the expected number of infected pairs. In a 2DCLASS analysis, the probability distribution of infected pairs is calculated by including all possible allotments of infected plants between the reference and target populations. This estimate has much more variability than the 2DCORR prediction in which the observed reference- and target-plant population numbers are used (Fig. 3).

In an earlier letter (11), I questioned the validity of the multiple comparisons made in a 2DCLASS- or 2DCORR-type analysis. For this reason, an extremely conservative application of the Bonferroni inequality was suggested (6,11). The basic logic is that, because so many comparisons (n) are being made ($n = 629 \chi^2$ tests in Fig. 6), allowances must be made for serendipitous false positives. For this reason, the tests of significance are applied at a higher level of confidence ($\alpha = 1 - (1 - 0.05)^{1/n} = 0.0000815 \sim 0.05/n$) to guarantee there is less than a 5% chance of even one false positive. Tests that pass this conservative test are labeled B for Bonferroni in Figure 6B and D.

THE CUMULATIVE PAIR COUNT

If the results of a 2DCORR analysis are symmetrical about the origin, it may be meaningful to construct a radial cumulative probability distribution function ($P(r)$) from observation (35,36). This function represents the fraction of the total number of observed infected pairs for which the component plants are less than a distance, r , apart (Fig. 7, +). The 2DCORR predictions for each distance-orientation class can be used to obtain a prediction for $P(r)$ (Fig. 7, line). The deviation, Δ , between observed and predicted (Fig. 7, inset) provides information on the distance range (λ , the distance at which Δ is maximum) over which spatial correlation is important. In addition, the maximum value of Δ can be used as a Komogorov-Smirnov test statistic (5) to estimate overall significant deviation from random behavior. Because the distribution is discrete rather than continuous, the application of this test will always be conservative (5). In Figure 7 the maximum value of Δ is 0.045 at a distance of $r = 8.6$ plant separation units. The total sample size of infected pairs is 7,260 ($IP_T = I_T(I_T - 1)/2 = 121(120)/2 = 7,260$), so $\Delta_{max} > 0.012$ corresponds to significantly ($P < 0.05$ [5,39]) more pairs than expected (Fig. 7 inset, line). With the above analyses, one comes to the following conclusions: (i) there is significant spatial correlation of like pairs of plants (infected-infected and healthy-healthy) in the nettle-head diseased hop field (14) illustrated in Figure 5, and (ii) this spatial correlation extends to a distance of 8.6 plant lengths in all directions.

If the results of a 2DCORR analysis are not symmetrical about the origin, it may be meaningful to construct a cumulative probability distribution function based on ellipses instead of circles. However, the fitting of the eccentricity and orientation of the appropriate ellipses is beyond the scope of this letter.

DISCUSSION

When one performs a product moment correlation on continuous variables, each function is centered and normalized. The functions are centered by subtracting the mean value and normalized by dividing by the standard deviation. In this way, the product moment correlation is a measure of similarity in shape, irrespective of magnitude. Thus, the difficulty encountered in comparing two rows of plants (Fig. 2) with different levels of disease incidence would not have happened if we were considering continuous variates. I suggest that the 2DCORR algorithm outlined above provides us with the same functionality as correlation analysis, measuring similarity in shape, irrespective of magnitude, in a form appropriate to discrete data sets.

Fig. 7. Plot of the fractional number of infected plant pairs less than a distance, r , apart for the field shown in Figure 5A (Observed: +, 2DCORR prediction, line). Inset shows the deviation, Δ , between the two curves. The distance at which maximum deviation occurs ($\lambda = 8.87$) is a measure of the spatial range of correlation. Horizontal line indicates deviation from random behavior at the 0.05 alpha level (described in text).

The Bonferroni correction as applied in the 2DCORR analysis outlined above (6,11) may be too conservative. The protection against type I errors (false positives) may create a problem with type II errors (missed significance) in certain cases. The correction method is based on the assumption that all comparisons will be considered with no a priori assumption about spatial behavior. In other words, the Bonferroni correction is a penalty for going on a “fishing expedition” and blindly seeking out any significant comparison. If, on the other hand, one wants to test the validity of the a priori assumption that like plants tend to be near each other, then one may be able to relax the Bonferroni condition (40). For instance, if one makes the a priori proximity hypothesis that adjacent plants will tend to be both infected or both healthy, then there are only two distance-orientation classes to consider ($[X = 0, Y = 1]$ and $[X = 1, Y = 0]$). In this case, the n in the Bonferroni correction would be 2 ($\alpha = 1 - (1 - 0.05)^{1/2} = 0.0253$). The proximity hypothesis can be extended to larger distances that will encompass more distance-orientation class comparisons, as long as the n in the Bonferroni correction is increased accordingly. This stepwise approach to the many multiple comparisons made in a 2DCORR analysis will provide protection from both types of error.

CONCLUSIONS

I have presented a method of analyzing infected pair counts that focuses on the alignment of like pairs of plants independent of the effects of spatial heterogeneity. This method eliminates reflected clusters and enhances the detection of short-range correlation. In addition, the procedure is extended by a Kolmogorov-Smirnov-type analysis based on the properties of the radial cumulative probability distribution function. This approach allows independent evaluation of the overall significance of the observed deviation from random behavior, as well as the length scale, within which distance, disease is correlated.

Although an improvement on previous approaches, 2DCORR is still basically a descriptive analysis. Future work should incorporate the effect of heterogeneity on spatial correlation for a mechanistic model (15). The spatial arrangement of disease within a field remains a virtually untapped reservoir of potentially useful information. To access this treasure trove, one must first refrain from disproving what spatial patterns are not present and concentrate on describing what pertinent structure, if any, is there in a clear and unambiguous manner. Pertinence, in this case, depends on the objective. For yield analysis, the distribution of damage due to disease averaged over whole plant quadrats may be appropriate. For comparison of different epidemics, a distance-sampling approach may be more likely to reveal the properties of the underlying inoculum dispersal mechanisms. Thus, if the various mathematical methods of pattern evaluation are applied to real-world problems in a straightforward manner, Pielou's (34) warning concerning the disparity between what different indices measure may cause less confusion.

APPENDIX

Consider two separate fields (Fig. 2, rows A and B), each containing N plants. The total number of sites in both fields is N_T ($N_T/2 = N$). We wish to consider all the possible permutations of the placement of infected plants within the two-row field illustrated in Figure 2. Even for this limited example, the total number of possible arrangements is forbiddingly large. Because there are 60 choices for placing the first infected plant, 59 for the second, and so on to 33 choices in placing the 28th infected plant, the total number of possible arrangements is the product of all these numbers. Once the infected plants are positioned, the order in which they were placed is unimportant. Therefore, one must divide the overall number by the number of ways to order 28 plants ($28!$).

Using this logic, the number of ways to place 28 indistinguishable infected plants (I_T) into 60 (N_T) possible sites is given by

$$\frac{N_T!}{I_T!(N_T - I_T)!} = \frac{60!}{28!32!} = \frac{60 \cdot 59 \cdot \dots \cdot 33}{1 \cdot 2 \cdot \dots \cdot 28} > 10^{17}$$

This is indeed a large number. If 1 million arrangements could be generated per second, the complete calculation would still take more than 3,000 years. Fortunately, the exact probability distribution can be derived by combinatorial analysis.

The probability that there are exactly I_{AB} side-by-side infected pairs is given by the ratio of the number of possible arrangements that give this result to the overall total number of ways to position I_T infected plants among N_T sites. The number of ways to place the I_{AB} infected pairs among the $N_T/2$ possible sites is given by the following combinatorial:

$$\frac{(N_T/2)!}{I_{AB}![(N_T/2) - I_{AB}]!}$$

Because each infected pair takes two infected plants, there remain a total of $I_T - 2I_{AB}$ infected plants to place among the remaining $N_T/2 - I_{AB}$ pair sites. These infected plants can only enter singly into these sites, because all infected pairs have already been counted. Thus, the number of ways to place these $I_T - 2I_{AB}$ infected plants among the $N_T/2 - I_{AB}$ pair sites is

$$2^{(I_T - 2I_{AB})} \frac{(N_T/2 - I_{AB})!}{(I_T - 2I_{AB})![(N_T/2) - I_T + I_{AB}]!}$$

where the premultiplier accounts for the possibility that each of the $I_T - 2I_{AB}$ singlet infected plants can be placed in either row A or B. Thus, the probability of obtaining I_{AB} side-by-side infected plant pairs within two rows with a total of N_T sites populated by I_T infected plants, $\rho(I_{AB}, I_T, N_T)$, is given by

$$\rho(I_{AB}, I_T, N_T) = 2^{(I_T - 2I_{AB})} \frac{(N_T/2)!}{I_{AB}![(N_T/2) - I_{AB}]!} \frac{(N_T/2 - I_{AB})!}{(I_T - 2I_{AB})![(N_T/2) - I_T + I_{AB}]!} \frac{N_T!}{(I_T)!(N_T - I_T)!}$$

which can be simplified to yield

$$\rho(I_{AB}, I_T, N_T) = 2^{(I_T - 2I_{AB})} \frac{(I_T)!(N_T - I_T)!(N_T/2)!}{I_{AB}!(I_T - 2I_{AB})![(N_T/2) - I_T + I_{AB}]!N_T!} \quad (A1)$$

For the two-row field illustrated in Figure 2 ($I_T = 28$ and $N = 30$), equation A1 is plotted versus I_{AB} in Figure 3 (top panel). The mean of the distribution is 6.41, and the probability of obtaining 8 or more infected plant pairs is 0.21.

There is an alternative way to derive equation A1 that makes explicit the dependence of the infected pair count on the apportionment of the infected plants between the two rows. First, one must determine the fraction of the total possible ways to place I_T plants in N_T sites, which results in I_A infected plants in field A and I_B infected plants in field B, such that $I_A + I_B = I_T$. In analogy with the combinatorials above, the number of ways to place I_A plants in an N plant row is

$$\frac{N!}{I_A!(N - I_A)!}$$

and likewise, the number of ways to place I_B plants in an N plant row is

$$\frac{N!}{I_B!(N - I_B)!}$$

The total number of possibilities is the product of these two combinatorials. Thus, the probability of obtaining I_A infected plants in field A and I_B infected plants in field B when I_T infected plants are placed in N_T sites is

$$\frac{N!}{I_A!(N - I_A)!} \frac{N!}{I_B!(N - I_B)!} \frac{N_T!}{I_T!(N_T - I_T)!}$$

Recalling that N , the number of sites in a row, is one-half the number of total sites, N_T , the above expression can be rearranged to yield the following hypergeometric distribution (HGD(I_A , N , I_T , $2N$) [7,39])

$$\text{HGD}(I_A, N, I_T, 2N) = \frac{I_T!(N!)(N!)(2N - I_T)!}{I_A!I_B!(N - I_A)!(N - I_B)!(2N)!} \quad (\text{A2})$$

Given the apportionment of infected plants in each row (I_A , I_B , such that $I_A + I_B = I_T$), we wish to examine whether the I_{AB} infected pairs could have resulted from a random draw. If one conceptually overlays maps of the above two fields, a certain number, I_{AB} , of the infected plants will align. If infected plants are randomly distributed in each field, the probability of I_{AB} alignments (infected plant pairs) once again is given by the fraction of the total possible arrangements that give a positive result. We now consider the field as a set on N side-by-side pairs of plants divided into four categories: infected-infected pairs (I_{AB}), infected-healthy pairs ($I_A - I_{AB}$), healthy-infected pairs ($I_B - I_{AB}$), and healthy-healthy pairs ($N - I_A - I_B + I_{AB}$) (Fig. 3). The number of ways to distribute these four groups in N positions is (7)

$$\frac{N!}{I_{AB}!(I_A - I_{AB})!(I_B - I_{AB})!(N - I_A - I_B + I_{AB})!}$$

which when divided by the number of ways to obtain I_A infected plants in field A and I_B infected plants in field B is

$$\frac{N!}{I_A!(N - I_A)!} \frac{N!}{I_B!(N - I_B)!}$$

which yields the following HGD ($\rho'(I_{AB}, I_A, I_B, N)$) [7,11,39]):

$$\rho'(I_{AB}, I_A, I_B, N) = \frac{I_A!I_B!(N - I_A)!(N - I_B)!}{I_{AB}!(I_A - I_{AB})!(I_B - I_{AB})!(N - I_A - I_B + I_{AB})!N!} \quad (\text{A3})$$

Equation A3 represents the conditional probability of obtaining I_{AB} infected plant pairs given the fact that I_A and I_B infected plants are in fields A and B, respectively.

As a check, I take the product of equations A2 and A3 and sum the result over all possible disease distributions that result in I_{AB} infected plant pairs to obtain

$$\rho(I_{AB}, I_T, N) = \frac{I_T!N!(2N - I_T)!}{(2N)!I_{AB}!(N - I_T + I_{AB})!} \sum_{I_A=I_{AB}}^{I_T-I_{AB}} \frac{1}{I_A!(I_A - I_{AB})!(I_T - I_{AB} - I_A)!} \quad (\text{A4})$$

where $\rho(I_{AB}, I_T, N)$ is the overall probability of obtaining I_{AB} infected pairs when conceptually overlaying two fields, each containing N plants, of which a total of I_T is infected. The limits of the summation in equation A4 are due to the simple physical reality that in order to have I_{AB} infected pairs there must be at least this number of infected plants in each field, i.e., $I_A \geq I_{AB}$ and $I_B = (I_T - I_A) \geq I_{AB}$, which is equivalent to $I_A \leq (I_T - I_{AB})$.

Equation A4 can be simplified. If the summation on the right side of equation A4 is multiplied by $(I_T - 2I_{AB})!$ and the premultiplier is divided by $(I_T - 2I_{AB})!$, then the summation can be recognized as the sum of the binomial coefficients of order $(I_T - 2I_{AB})$, which is equal to $2^{(I_T - 2I_{AB})}$ (7,39). Thus, equation A4 is equivalent to equation A1.

ACKNOWLEDGMENTS

I thank S. Sanogo, V. L. Smith, J. Ward, and D. Dingman for helpful discussion and D. E. Aylor, S. L. Anagnostakis, W. H. Elmer, and two anonymous reviewers for critical reviews.

LITERATURE CITED

- Bradshaw, G. A., and Spies, T. A. 1992. Characteristic canopy gap structure in forests using wavelet analysis. *J. Ecol.* 80:205-215.
- Campbell, C. L., and Madden, L. V. 1990. Introduction to Plant Disease Epidemiology. John Wiley & Sons, New York.
- Chapham, A. R. 1936. Over-dispersion in grassland communities and use of statistical methods in plant ecology. *J. Ecol.* 24:232-251.
- 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.
- Conover, W. J. 1980. Practical Nonparametric Statistics. 2nd ed. John

Wiley & Sons, New York.

- Day, R. W., and Quinn, G. P. 1989. Comparisons of treatments after an analysis of variance in ecology. *Ecol. Monogr.* 59:433-463.
- Feller, W. 1968. Chapter 2. Pages 26-53 in: *An Introduction to Probability Theory and Its Applications*. Vol. 1. 3rd ed. John Wiley & Sons, New York.
- Ferrandino, F. J. 1989. Spatial and temporal variation of a defoliating plant disease and reduction in yield. *Agric. For. Meteorol.* 47:273-290.
- Ferrandino, F. J. 1989. A distribution-free method for estimating the effect of aggregated plant damage on crop yield. *Phytopathology* 79:1229-1232.
- Ferrandino, F. J. 1993. Dispersive epidemic waves. I. Focus expansion within a linear planting. *Phytopathology* 83:795-802.
- Ferrandino, F. J. 1996. Two-dimensional distance class analysis of disease-incidence data: Problems and possible solutions. *Phytopathology* 86:685-691.
- Ferrandino, F. J., and Elmer, W. H. 1991. Reduction in tomato yield due to Septoria leaf spot. *Plant Dis.* 76:208-211.
- Fisher, R. A., and Yates, F. 1948. *Statistical Tables*. 3rd ed. Hafner Publishing Co., New York.
- Freeman, G. H. 1953. Spread of diseases in a rectangular pattern with vacancies. *Biometrika* 40:287-296.
- Gibson, G. J. 1997. Investigating mechanisms of spatiotemporal epidemic spread using stochastic models. *Phytopathology* 87:139-145.
- Gottwald, T. R. 1995. Spatio-temporal analysis and isopath dynamics of citrus scab in nursery plots. *Phytopathology* 85:1082-1092.
- Gottwald, T. R., Avinent, L., Llcer, G., Hermosa 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.
- 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.
- 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.
- 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.
- 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.
- Hill, M. O. 1973. The intensity of spatial pattern in plant communities. *J. Ecol.* 61:225-235.
- Hughes, G. 1987. Spatial heterogeneity in crop loss assessment models. *Phytopathology* 78:883-884.
- Hughes, G., and Madden, L. V. 1993. Using the beta-binomial distribution to describe aggregated patterns of disease incidence. *Phytopathology* 83:759-763.
- Hughes, G., Madden, L. V., and Munkvold, G. P. 1996. Cluster sampling for disease incidence data. *Phytopathology* 86:132-137.
- Hurlbert, S. H. 1990. Spatial distribution of the montane unicorn. *Oikos* 58:257-271.
- Legendre, P., and Fortin, M. J. 1989. Spatial pattern and ecological analysis. *Vegetatio* 80:107-138.
- Madden, L. V., and Hughes, G. 1995. Plant disease incidence: Distributions, heterogeneity and temporal analysis. *Annu. Rev. Phytopathol.* 33:529-64.
- Morisita, M. 1959. Measuring of the dispersion of individuals and the distribution patterns. *Mem. Fac. Sci. Kyushu Univ. Ser. E* 2:215-235.
- Nelson, S. C. 1995. Spatiotemporal distance class analysis of plant disease epidemics. *Phytopathology* 85:37-43.
- Nelson, S. C. 1996. A simple analysis of disease foci. *Phytopathology* 86:332-339.
- 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.
- 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.
- Pielou, E. C. 1977. *Mathematical Ecology*. John Wiley & Sons, New York.
- Ripley, B. D. 1979. Modeling spatial patterns. *J. Roy. Stat. Soc. B.* 39:172-212.
- Ripley, B. D. 1981. *Spatial Statistics*. John Wiley & Sons, New York.
- 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.
- Samita, S. 1995. Analysis of aggregated plant disease incidence data. Ph.D. dissertation. University of Edinburgh, Edinburgh, Scotland, UK.
- 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.
- Stewart-Oaten, A. 1995. Rules and judgments in statistics: Three examples. *Ecology* 76:2001-2009.
- Taylor, L. R. 1961. Aggregation, variance and the mean. *Nature (Lond.)* 189:732-735.
- Ver Hoef, J. M., Noel, A. C., and Glenn-Lewin, D. C. 1993. Spatial models for spatial statistics: Some unification. *J. Veg. Sci.* 4:441-452.