Evaluation of Tests for Randomness of Infected Plants

L. V. Madden, Raymond Louie, J. J. Abt, and J. K. Knoke

Systems specialist, Department of Plant Pathology, Ohio Agricultural Research and Development Center; research plant pathologist, agricultural research technician, and research entomologist, Agricultural Research, Science and Education Administration, U.S. Department of Agriculture, Wooster, OH 44691, respectively.

Cooperative investigations of the Ohio Agricultural Research and Development Center (OARDC), and Agricultural Research, Science and Education Administration, U.S. Department of Agriculture. Approved for publication as Journal Article 191-80 of the Ohio Agricultural Research and Development Center, Wooster.

This paper reports the results of research only. Mention of a commercial or proprietary product does not constitute an endorsement by the USDA.

We thank S. D. Simeral for technical assistance and L. R. Nault, D. T. Gordon, and L. J. Herr for reviewing the manuscript.

Accepted for publication 31 March 1981.

ABSTRACT

Madden, L. V., Louie, R., Abt, J. J., and Knoke, J. K. 1982. Evaluation of tests for randomness of infected plants. Phytopathology 72:195-198.

A study was conducted to evaluate ordinary runs, original doublets, and corrected doublets for detecting the pattern of sweet corn plants infected by maize dwarf mosaic virus. Four fields were divided into a total of 636 quadrats and analyzed with each test. When the data were divided into 10% ranges of infected plants, no trend in the agreement between ordinary runs and original doublets or between ordinary runs and corrected doublets was evident. For original and corrected doublets, agreement ranged from 100% at the lowest disease incidence to 0% at the highest level of incidence. In a simulation study, corrected doublets gave unsatisfactory results (>5% misclassifications) when random patterns were generated, original doublets gave unsatisfactory results when nonrandom patterns were generated, whereas ordinary runs did not give unsatisfactory results with random or nonrandom patterns. Thus, ordinary runs was the best test for determining randomness of infected plants.

Additional key words: epidemiology, virus diseases, *Zea mays*.

Identification of the type of disease pattern and spread in a field is critical in epidemiological investigations (1). A random pattern of infected plants suggests that, at the time of observation, the pathogen is not spreading from plant to plant. Conversely, aggregations (clusters) of infected plants suggest that the pathogen is spreading from plant to plant within a field.

Vanderplank (15) proposed doublet analysis as a technique for determining the type of pathogen movement in field plots. Several researchers since then have used this approach (6–8). Converse et al (2) recently claimed that the equation for the standard deviation in Vanderplank's doublet analysis was incorrect. They presented a corrected form of doublet analysis as derived from the work of Freeman (3). More complicated statistical procedures are available for assessing randomness or clustering of infected plants (11,12). Some of these techniques have been used by plant pathologists (6,8), but the statistical properties of these procedures have not been thoroughly explored.

In this paper, we explore ordinary runs (5) as an alternative analysis to doublets for determining the pattern of infected plants in rows and fields. Results of the ordinary runs test are compared with the original and corrected doublets. Maize dwarf mosaic virus (MDMV) in corn (*Zea mays* L.) is used as the pathogen-suscept system. Simulated random and nonrandom data also were generated and used to evaluate some of the properties of these tests.

RANDOMNESS TESTING

Ordinary runs analysis. "In an ordered sequence of some two types of symbols, a run is defined as a succession of one or more identical symbols, which are followed and preceded by a different symbol or no symbol at all." (4,5). For example, consider the following pattern of 10 numbers as representing a corn row with 10 plants: 0 0 1 1 0 1 0 1 1 1. The 0s represent disease-free plants and the 1s represent MDMV-infected plants. There are six runs in the ordered sequence (corn row). Reading from left to right, the ordinary runs are: "00," "11," "0," "1," "0," and "111."

If diseased plants in a row resulted from a pathogen spreading from plant to plant, one would expect an aggregation (clustering) of infected plants and an aggregation of healthy plants. Thus, there would be few runs. If there were no movement of the pathogen from plant to plant one would expect a random mixing of healthy and infected plants and a correspondingly large number of runs. The null hypothesis evaluated in this test is that the ordered sequence of symbols (infected plants) is random. The alternative hypothesis is that the ordered sequence is clustered.

Let m represent the number of infected plants in a row with a total number of plants equal to N. The total number of runs can be represented by the symbol U. Under the null hypothesis of randomness, the expected value (E) of U is given by:

$$E(U) = 1 + 2m(N-m)/N \quad (1)$$

The observed number of runs will be *less* than $E(U)$ if there is a clustering of infected plants (5). The standard deviation of U is given by:

$$s_u = (2m(N-m)[2m(N-m) - N]/[N^2(N-1)])^{1/2} \quad (2)$$

The standardized U is given by:

$$Z_u = [U + 0.5 - E(U)]/s_u \quad (3)$$

The constant 0.5 is the "correction for continuity" (5). The asymptotic sampling distribution of Z_u is the standard normal distribution (4,5). The value of Z_u will be a large negative number if there is clustering. Therefore, the test for nonrandomness (clustering) is one-sided and the left-tail probability is used (5). A row of plants was considered to have a nonrandom sequence of infected and healthy plants if $-Z_u$ was greater than 1.64 ($P = 0.05$).

The Z-statistic (equation 3) does not follow a normal distribution for $N < 20$. Tables are available for determining significance levels

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. § 1734 solely to indicate this fact.

in these small-sample situations (5,13). Use of sample sizes much less than 20 is unwise for determining aggregations of infected plants.

Doublet analysis. "A doublet consists of two adjacent diseased plants (15,16)." Two adjacent diseased plants equals one doublet, three adjacent diseased plants equals two doublets, and so on. In the example given above, there are a total of three doublets. The total number of doublets is represented by D. Under the null hypothesis of randomness, the expected number of doublets is given by:

$$E(D) = m(m-1)/N \quad (4)$$

where m and N are defined as before (15,16). The observed number of doublets will be *greater* than $E(D)$ when there are clusters of infected plants. According to Vanderplank (15), the standard deviation of D is given by:

$$s_D = ([m(m-1)/N][1-2/N])^{1/2} \quad (5)$$

A common practice is to combine adjacent rows for the analyses described in this paper (16). For example, four adjacent rows of 25 plants each would be combined to form an arbitrary row length of 100 plants. An observer in the field would read up one row and down the next, and so on. When data for adjacent rows are combined, the expected number of doublets under the null hypothesis of randomness is represented by:

$$E(Dc) = [N-r][m(m-1)/N(N-1)] \quad (6)$$

where $E(Dc)$ is the "corrected" expected number of doublets and r is the number of combined rows (2). Equation 6 reduces to equation 4 when $r=1$. Converse et al (2) stated that equation 5 is incorrect, even if no rows were combined, and presented a more complicated equation for s_{Dc}.

The standardized D and Dc is given by a Z-statistic with the appropriate expected value and standard deviation substituted for $E(U)$ and s_u. Unlike the Z-statistic for ordinary runs, however, this value for the two types of doublets will be a large positive value if the infected plants are clustered.

TABLE 1. Results of ordinary runs (U), original doublet (D), and corrected doublet (Dc) tests for one quadrat of maize dwarf mosaic virus-infected sweet corn (*Zea mays*) at three times which corresponded to three levels of disease incidence (m)

	Observed	Expected	Standard deviation	Z[a]	P[b]
$m = 14$					
U	22	25.08	2.37	−1.09	0.138
D	3	1.82	1.34	1.26	0.104
Dc	3	1.76	1.16	1.50	0.067
$m = 31$					
U	32	43.78	4.25	−2.66	0.004
D	15	9.30	3.02	2.05	0.020
Dc	15	9.02	2.12	3.05	0.001
$m = 84$					
U	23	27.88	2.65	−1.65	0.049
D	72	69.22	8.27	0.34	0.367
Dc	72	67.61	1.55	3.16	0.001

[a] Standardized variable; large negative values indicate clustering with ordinary runs, whereas large positive values indicate clustering with original and corrected doublets.
[b] Significance level.

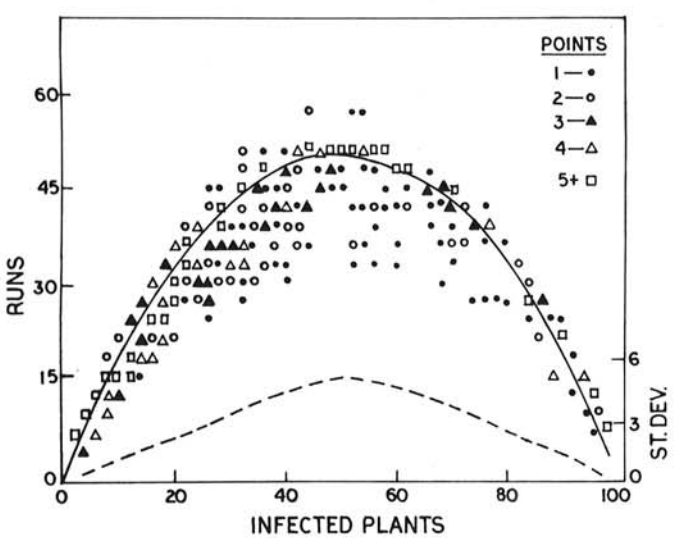

Fig. 1. Observed runs, expected number of runs under the null hypothesis of randomness (solid line), and standard deviation (dashed line) in relation to number of maize dwarf mosaic virus-infected plants. Multiple data points are indicated with the symbols shown in the plot.

MATERIALS AND METHODS

Disease assessment. Epidemics of MDMV were studied in four fields of sweet corn in northern Ohio during 1979. The fields (0.5–5.3 ha) were divided into 48–100 quadrats, each ranging from 0.01–0.03 ha. In the center of each quadrat, four adjacent rows of 25 plants each were visually assessed for MDM symptoms (10). Observations for symptoms began when plants were 15–25 cm tall and ended when at the pollen shed stage. For analysis, the four rows were combined to form an abstract row length of $N=100$. Only data from quadrats with a minimum of two and a maximum of 98 infected plants were analyzed, giving a sample size of 636 quadrats.

Simulation study. A Monte Carlo study was conducted to evaluate some properties of the original doublet, corrected doublet, and the ordinary runs tests with data exhibiting a random pattern. Two-hundred pseudo-random samples were generated by using the linear congruential recurrence method (9,14) for row lengths of 100 plants. The simulation was carried out for the probability of a plant becoming infected (p) equal to 0.25 and also 0.75. On the average, there were 100 p infected plants per row.

A second simulation study was performed to determine the outcome of the three tests when the data exhibited a nonrandom (clustered) pattern. Two-hundred samples were generated as described previously. In this situation, p equaled 0.75 if the previous plant was infected and 0.25 if the previous plant was disease-free. This resulted in the clustering of disease-free and infected plants in the rows of 100.

RESULTS

Several properties of the three tests are exemplified in Table 1. The value of $E(Dc)$ was less than $E(D)$; these two variables would be equal if rows were not combined for analysis. The difference between, and magnitude of, $E(Dc)$ and $E(D)$ increased as m increased. In contrast, the magnitude of $E(U)$ increased as incidence approached 50% ($m=50$) and then decreased as m approached N (Figs. 1 to 3).

The standard deviation of Dc was less than that of D. The value of s_D increased as D (or m) increased (Fig. 2); the values of s_{Dc} and s_u were low when m was small, reached a maximum around the midpoint of the disease incidence scale, and then decreased as m approached N (Figs. 1 and 3).

Tables 2 to 4 contain results of the two-way comparisons of the three tests for randomness. Total percent agreement (ie, number of cases when the two tests indicated randomness plus the number of cases when two tests indicated nonrandomness divided by 636) ranged from 73 to 79%. When the data on infected plants were divided into 10% ranges, no trend was evident in the agreement between ordinary runs and original doublets or between ordinary runs and corrected doublets (Tables 2 and 3); the combined agreement fluctuated around 78 and 73%, respectively. In a comparison between ordinary runs and corrected doublets, almost

all misclassifications resulted from the ordinary runs test indicating randomness when the corrected doublet test indicated clustering. Only three of 173 disagreements resulted from the ordinary runs test indicating nonrandomness and corrected doublets randomness.

For the comparison of original and corrected doublets (Table 4), low levels of disease incidence ($m \leq 10$) gave 100% agreement between the two tests. This agreement decreased to 100% disagreement at the highest level of disease incidence ($m > 90$). All disagreements resulted from the corrected doublets indicating nonrandomness and original doublets randomness.

Results of the random pattern study are presented in Table 5. At both probabilities of infection, corrected doublets had the highest level of misclassification (6.5–7.0%); original doublets had the fewest misclassifications (0.0–0.5%). Both ordinary runs and original doublets misclassified less than 5% of the cases.

Corrected doublets and ordinary runs correctly classified 100% of the generated nonrandom samples in the clustered pattern study. The original doublets misclassified 9% of the nonrandom samples; ie, original doublets indicated a random pattern when the data were nonrandom.

DISCUSSION

The three tests for determining the pattern of diseased plants in a row produced different results. When random patterns were generated, corrected doublets gave unsatisfactory results (>5% misclassifications). When nonrandom patterns were generated, the original doublets gave unsatisfactory results (>5% misclassifications). Only the ordinary runs test gave satisfactory results (<5% misclassifications) in both situations.

For the field data, agreement between original doublets and corrected doublets displayed a clear trend relative to number of infected plants (Table 4). Results of both tests indicated predominantly randomness at the lower levels of infected plants. At the higher levels of infection, doublets indicated randomness, whereas corrected doublets indicated nonrandomness. These discrepencies were due to the differences in their respective standard deviations. The standard deviation for the original doublet test increases continuously as m increases. The Z-statistic, therefore, decreases as m approaches N. Small values of Z indicated that one cannot reject the null hypothesis of randomness. As Converse et al (2) pointed out, the standard deviation for the corrected doublet test goes to zero as m approaches N; thus, the Z-statistic becomes large as m approaches N. At large values of Z, one rejects the null hypothesis of randomness in favor of the

TABLE 2. Agreement between ordinary runs and original doublets at different levels of maize dwarf mosaic incidence (m) for 636 quadrats of sweet corn (*Zea mays*)

	Number of quadrats					Agreements
m	R:R[a]	N:R	R:N	N:N	Total	(%)
2–10	89	0	73	87	249	71
11–20	72	3	15	32	122	85
21–30	47	4	0	18	69	94
31–40	31	10	0	16	57	82
41–50	19	4	0	3	26	85
51–60	14	8	0	2	24	67
61–70	10	4	0	0	14	71
71–80	10	4	0	0	14	71
81–90	21	2	0	0	23	91
91–98	24	14	0	0	38	63
Combined	337	53	88	158	636	78

[a] R = random; N = nonrandom; the first symbol refers to results for the ordinary runs test and the second symbol to those for the original doublet test. R:R and N:N represent agreement between the two tests and N:R and R:N disagreement.

TABLE 3. Agreement between ordinary runs and corrected doublets at different levels of maize dwarf mosaic incidence (m) for 636 quadrats of sweet corn (*Zea mays*)

	Number of quadrats					Agreement
m	R:R[a]	N:R	R:N	N:N	Total	(%)
2–10	89	0	73	87	249	71
11–20	63	3	24	32	122	78
21–30	37	0	10	22	69	86
31–40	25	0	6	26	57	89
41–50	14	0	5	7	26	81
51–60	10	0	4	10	24	83
61–70	7	0	3	4	14	79
71–80	6	0	4	4	14	71
81–90	4	0	17	2	23	26
91–98	0	0	24	14	38	37
Combined	255	3	170	208	636	73

[a] R = random; N = nonrandom; the first symbol refers to results for the ordinary runs test and the second symbol to those for the corrected doublet test. R:R and N:N represent agreement between the two tests and N:R and R:N disagreement.

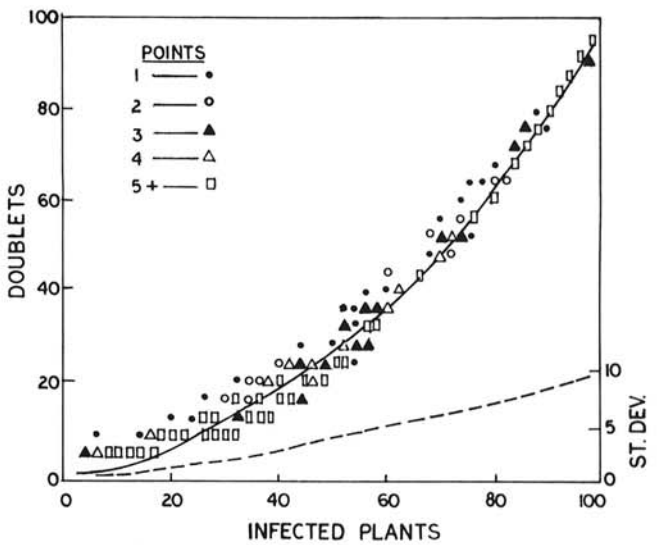

Fig. 2. Observed doublets, expected number of doublets (original) under the null hypothesis of randomness (solid line), and standard deviation (dashed line) in relation to number of maize dwarf mosaic virus-infected plants. Multiple data points are indicated with the symbols shown in the plot.

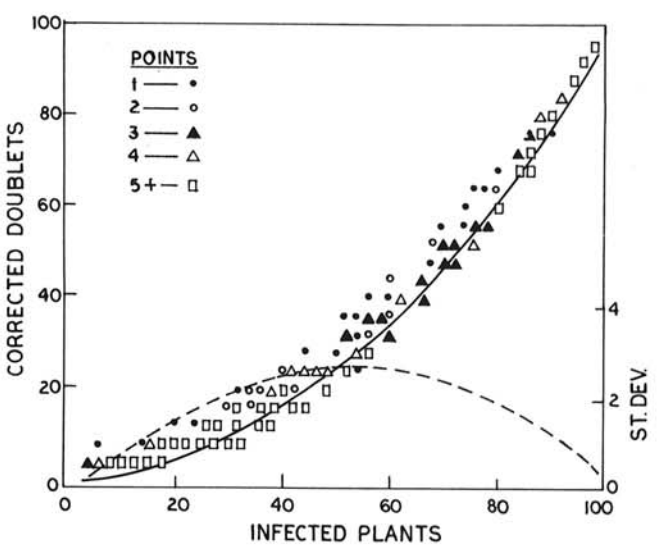

Fig. 3. Observed doublets, expected number of doublets (corrected) under the null hypothesis of randomness (solid line), and standard deviation (dashed line) in relation to number of maize dwarf mosaic virus-infected plants. Multiple data points are indicated with the symbols shown in the plot.

TABLE 4. Agreement between original doublets and corrected doublets at different levels of maize dwarf mosaic incidence (m) for 636 quadrats of sweet corn (*Zea mays*)

m	Number of quadrats					Agreement (%)
	R:R[a]	N:R	R:N	N:N	Total	
2–10	89	0	0	160	249	100
11–20	66	0	9	47	122	93
21–30	37	0	14	18	69	80
31–40	25	0	16	16	57	72
41–50	14	0	9	3	26	65
51–60	10	0	12	2	24	50
61–70	7	0	7	0	14	50
71–80	6	0	8	0	14	43
81–90	4	0	19	0	23	17
91–98	0	0	38	0	38	0
Combined	258	0	132	246	636	79

[a]R = random; N = nonrandom; the first symbol refers to results for the original doublet test and the second symbol to those for the corrected doublet test. R:R and N:N represent agreement between the two tests and N:R and R:N disagreement.

alternative hypothesis of clustering.

Obvious trends were not noted in the agreement between ordinary runs and either form of doublets (Tables 2 and 3). These results, and those under both simulation schemes, suggest that ordinary runs analysis is the best of the three procedures for determining randomness (or nonrandomness) of infected plants in rows.

The individual significance level (P) for rejecting the null hypothesis was 0.05 in this study. No effort was made to control the "family" significance level. All of the data were reanalyzed with P=0.01 (data not shown); no new or conflicting information was revealed. Removal of the "correction for continuity" also did not change the results appreciably, except that corrected doublets misclassified fewer of the simulated random samples.

The usefulness of ordinary runs and doublet tests is based on Vanderplank's (15) postulate that clusters of infected plants indicate that a pathogen is predominately spreading from plant to plant, provided that individual samples lie within homogenous areas. We feel that this is still an acceptable postulate. A random pattern of diseased plants, although indicating no or insignificant spread from plant to plant, does not necessarily indicate the source of inoculum. The source of inoculum may be exogenous (eg, migrating viruliferous aphids) or the pathogen may be seedborne. Hence, information on the distribution of disease levels throughout the quadrats in relation to prevailing wind direction as well as knowledge of the biology of the disease would be needed to determine inoculum source. Statistical analysis and mathematical models will greatly enhance the understanding of virus-vector-environment interactions resulting in disease epidemics. The ordinary runs test was found most suitable for the determination of random or nonrandom distribution of virus-infected plants. The ordinary runs test is currently being used to relate the pattern of MDMV-infected corn plants to time (growth stage), level of disease incidence, and location.

TABLE 5. Number of times of 200 that ordinary runs (U), original doublets (D), and corrected doublets (Dc) indicated clustering when pseudo-random data were generated at two probabilities of infection (p)

Test	$p = 0.25$	$p = 0.75$
U	4 (2.0)[a]	6 (3.0)
D	1 (0.5)	0 (0.0)
Dc	13 (6.5)	14 (7.0)

[a]Numbers in parentheses are percentages.

LITERATURE CITED

1. Campbell, C. L., and Pennypacker, S. P. 1980. Distribution of hypocotyl rot caused in snapbean by *Rhizoctonia solani*. Phytopathology 70:521-525.
2. Converse, R. H., Seely, J., and Martin, L. W. 1979. Evidence for random local spread of aphid-borne mild yellow-edge virus in strawberries. Phytopathology 69:142-144.
3. Freeman, G. H. 1953. Spread of diseases in a rectangular pattern with vacancies. Biometrika 40:287-296.
4. Gibbons, J. D. 1971. Nonparametric Statistical Inference. McGraw-Hill, New York. 306 pp.
5. Gibbons, J. D. 1976. Nonparametric Methods for Quantitative Analysis. Holt, Rinehart, and Winston, New York. 463 pp.
6. Hill, J. H., Lucas, B. S., Benner, H. I., Tachibana, H., Hammond, R. B., and Pedigo, L. P. 1980. Factors associated with the epidemiology of soybean mosaic virus in Iowa. Phytopathology 70:536-540.
7. Huang, H. C., and Hoes, J. A. 1980. Importance of plant spacing and sclerotial position to development of sclerotinia wilt of sunflower. Plant Dis. 64:81-84.
8. Jarvis, W. R., and Hawthorne, B. T. 1972. Sclerotinia minor on lettuce: progress of an epidemic. Ann. Appl. Biol. 70:207-214.
9. Lewis, T. G., and Payne, W. H. 1973. Generalized feedback shift register pseudo-random numer algorithm. J. Am. Comp. Math. 20:456-568.
10. Louie, R. 1977. Maize virus disease identification by symptomatology. Pages 89–91 in: L. E. Williams, D. T. Gordon, and L. R. Nault, eds. Proc. Maize Virus Disease Colloquium and Workshop. 16–19 August. 1976. Ohio Agric. Res. and Dev. Center, Wooster. 145 pp.
11. Pielou, E. C. 1962. Runs of one species with respect to another in transects through plant populations. Biometrics 18:579-593.
12. Roach, S. A. 1968. The Theory of Random Clumping. Methuen, London. 94 pp.
13. Swed, F. S., and Eisenhart, C. 1943. Tables for testing randomness of grouping in a sequence of alternatives. Ann. Math. Stat. 14:66-87.
14. Tootill, J. P. R., Robinson, W. D., and Eagle, D. J. 1973. An asymptotically random. Tousworthe sequence. J. Am. Comp. Math. 20:469-481.
15. Vanderplank, J. E. 1946. A method for estimating the number of random groups of adjacent diseased plants in a homogeneous field. Trans. R. Soc. S. Africa 31:269-278.
16. Vanderplank, J. E. 1960. Analysis of epidemics. Pages 229–289 in: J. G. Horsfall and A. E. Dimond, eds. Plant Pathology, Vol. 3. Academic Press, New York. 675 pp.