Testing the Spatial Association of Disease Patterns between Two Dates in Orchards

Gaël Thébaud[1,2], Nathalie Peyrard[2], Sylvie Dallot[1], Agnès Calonnec[3], Gérard Labonne[1] and Joël Chadœuf[2]

Institut National de la Recherche Agronomique (INRA)
[1] UMR BGPI, Campus de Baillarguet, Montpellier, France;
[2] Unité de Biométrie, Domaine Saint-Paul, Site Agroparc, Avignon, France;
[3] UMR Santé Végétale, Villenave d'Ornon, France.

Keywords: distance class, nonparametric, permutation, sharka, statistics, toroidal shift.

Abstract

The analysis of spatiotemporal patterns can provide clues about disease spread by assessing if the spatial pattern of diseased plants at one date is associated with the pattern of previously diseased plants. No generic statistical test was available to answer this question for spatiotemporal maps of binary data (healthy or diseased plants) in regular plantings (e.g., orchards). Here we describe a Monte Carlo test of the hypothesis that the location of newly diseased plants is independent of the location of previously diseased plants, even when the disease is spatially aggregated within each assessment period. This spatiotemporal test is designed to cope with the censorship arising on a lattice when plants are missing or cannot recover between the two dates. Expected patterns are simulated by shifting on a torus the whole pattern at the second date relatively to the pattern at the first date. For each simulation, we discard the censored points from observed and simulated data. In case of a positive association between disease patterns at two dates, the distances between newly and previously diseased trees should be smaller in the observed than in the simulated patterns. As an illustration, we analysed the dependence between patterns of trees showing *Plum pox virus* symptoms at two dates.

INTRODUCTION

One of the aims of epidemiological studies is to understand the biological processes driving the spread of a disease. The analysis of the observed spatiotemporal patterns of disease cases provides an opportunity to address questions that are related to the underlying processes. For a systemic disease affecting an orchard, a classical way to record disease spread consists of marking the disease cases (i.e., diseased trees) on a map at successive dates. If enough information is available on how the disease spreads, it may be possible to build a mathematical model, and to estimate some relevant biological parameters by fitting the model to the available maps. However, when such basic information is unknown, nonparametric tests are often more relevant for addressing simple questions about disease patterns. One of the basic objectives is often to assess if there is a secondary transmission within the orchard. If it is true, the newly diseased trees generally should be more clustered with previously diseased trees than what would happen by chance. Thus, a typical question arising from spatiotemporal disease maps is:

Thébaud G. *et al.* (2008) Acta Horticulturae 781:255-260.

"Is there an association between the spatial pattern of newly diseased trees (t_2 cases) and the spatial pattern of previously diseased trees (t_1 cases)?"

Provided that there is no heterogeneity in the data (e.g., mixed cultivars), the principle of such a test of independence is to perform many simulations (typically 1000) that randomly reallocate the newly diseased trees onto the orchard, and then to test if the observed distances between t_1 cases and t_2 cases are significantly different from the distances obtained in the randomizations. However, this apparently simple principle is complicated by some features of the data that, until now, have prevented the development of the corresponding test: (i) aggregation can exist within both t_1 cases and within t_2 cases (regardless of the relative position of the two groups of cases), and (ii) the observed pattern is partially censored by t_1 cases and by missing trees, because no t_2 case can be observed at the location of a missing tree or a t_1 case (if the diseased trees are removed or if the disease quickly becomes systemic, later potential recontaminations cannot be detected). None of the existing statistical methods (e.g., Nelson, 1995; Mugglestone et al., 1996; Diggle, 2003) considers these two issues. Thus, using these methods when the data have such characteristics would lead to inaccurate *P*-values (generally, misleadingly low), which may in turn suggest incorrect interpretations of the observed spatiotemporal patterns in terms of dispersal processes of the disease. In addition, some of the existing methods group the trees into quadrats rather than using all the information contained in the point pattern of diseased trees (Reynolds and Madden, 1988; Perry and Dixon, 2002), which reduces the power of the test to detect an existing association. Therefore, we developed a method based on point patterns and specifically dedicated to the analysis of aggregated patterns and incompletely observed data (censored by t_1 cases and missing trees). In our Monte Carlo test, the aggregated nature of both patterns is preserved in the simulations of the expected patterns under the null hypothesis.

Here we present this nonparametric test of independence assessing whether the locations of new disease cases depend on the locations of previous disease cases within an orchards or another type of regularly spaced planting. After presenting the test, its accuracy and its power are evaluated on computer-simulated data sets, and it is applied for the analysis of a spatiotemporal disease map drawn during the implementation of roguing against the *Plum pox virus* in an orchard.

MATERIALS AND METHODS

Test of Independence

The spatial pattern of the disease within each group of cases determines the choice of the statistical test to be used. Thus, before testing the independence between the two patterns, a preliminary analysis must be undertaken to assess the spatial pattern at each date separately (and also to check that there is no edge effect). Here we present the test corresponding to the situation where this first step enabled the conclusion that each pattern had a non-random structure (aggregated, generally).

This spatiotemporal test is based on a classical test (Lotwick and Silverman, 1982) that preserves the spatial structure of t_2 cases, modified to cope with the censoring issues (Chadœuf et al., 1997; Chadœuf et al., 2000; Peyrard et al., 2005). An intuitive explanation of this test follows. First, the tested hypothesis (H_0) is: "the newly diseased trees (t_2 cases) are spatially independent of the previously diseased trees (t_1 cases)". A criterion naturally related to this hypothesis is the distribution of distances between diseased trees at t_1 and diseased trees at t_2. If, in the observed data, this distribution is significantly shifted towards small distances in comparison to the expected distribution, it means that t_2 cases are closer to t_1 cases than what would be expected if the two patterns were independent, and hence H_0 should be rejected. The expected distribution of distances under the null hypothesis is obtained by simulating 1000 random t_2 patterns. However, if t_2 cases were reallocated completely at random, we would not be able to differentiate nonrandomness caused by the spatial dependence between t_2 cases and t_1 cases from nonrandomness caused only by the aggregation within t_2 cases. Thus, only the relative location of the patterns is randomized, while the internal spatial structure of each pattern is retained. This is achieved by converting the square orchard into a torus (treating the opposite boundaries as if they were connected) and through randomly shifting the pattern of t_2 cases by a random number of trees along and across rows on this torus (Lotwick and Silverman, 1982). All distances are measured on the torus.

Randomisation tests are based on the principle that, under the null hypothesis, the probability of an event is the same in the real data and in the simulations. However, in the simulations, the location of every disease case is known and a t_2 case can be shifted on the location of a non-shifted t_1 cases or missing tree, whereas in the observed data such potential t_2 case would be censored (i.e., hidden by a t_1 case or a missing tree) and would remain unnoticed (Fig. 1). It is therefore necessary to balance the two situations, through discarding both the simulated t_2 cases that are shifted onto the censoring patterns and the observed t_2 cases on which this censoring pattern is shifted (Chadœuf et al., 1997). Because of this last censoring event, different observed t_2 cases are discarded after each simulation and, as a result, the distribution of distances observed in the real data varies from one simulation to another. This prevents using a classical Monte Carlo test in which a single observed value (test statistic) is compared to a distribution of simulated values (expected test statistic under the null hypothesis). Instead, after each simulation, we compute the frequency distribution of toroidal distances between the observed t_1 cases and either the observed or the simulated t_2 cases; then, the difference between the two distributions is computed (Chadœuf et al., 2000). After rescaling it to improve the graphical display, this difference defines a test statistic with an observed value of 0 (when no shift is applied to the t_2 cases). Thus, the rank of 0 among the whole set of computed differences provides the basis on which to decide if H_0 should be rejected. The test statistic is computed over increasing distances d for all t_1–t_2 pairs of cases closer than d, where d takes discrete values at regular intervals (defining distance classes).

In summary, the test statistic is defined by: $S_c(d) = [N_{1,2}^{\phi}(d) - N_{1,2}(d)]/d$, where $N_{1,2}^{\phi}(d)$ and $N_{1,2}(d)$ are the number of pairs of trees closer than a distance d involving one observed t_1 case and one non-censored t_2 case, from shifted or observed patterns, respectively. When t_2 cases are independent of t_1 cases, the interval between the 2.5% lower and upper values of $S_c(d)$ should include 0, and a P-value is derived as twice the number of simulated values more extreme than or equal to 0 (Manly, 1991). The case of a positive association between disease patterns at the two dates would lead to more pairs at small distances in observed than in expected patterns, and thus to values of $S_c(d)$ significantly below 0.

Numerical Validation

A simulation study was performed to validate this test of independence on well-defined patterns for three levels of disease incidence (3%, 14%, and 25%). The test was applied to 1000 computer-generated disease patterns on a 50 × 20 grid with 2% missing trees, and with an average number of t_2 cases twice that of t_1 cases. A first Neyman-Scott point process was used to simulate the aggregated pattern at t_1. For each simulation, 13 cluster centres on average were located at random, and marked points were created around these centres at a distance following an exponential distribution (mean: 2.5).

The level of type I error of the test corresponds to the proportion of the simulations for which H_0 is wrongly rejected when the t_2 process was simulated independently of the t_1 process (for a significance level $\alpha = 0.05$, only 5% of the simulations should be rejected). In this situation, the point pattern for t_2 cases is obtained as before, and independently of the first one.

The statistical power of the test corresponds to the proportion of the simulations for which H_0 is correctly rejected when the simulated process at t_2 is spatially dependent of the simulated process at t_1. In this situation, the point pattern for t_2 cases is obtained by generating a second Neyman-Scott process from the t_1 cases, with the same properties as the initial process.

Experimental Data Set

This test can help to address several issues, among which is the efficiency of a given control method to prevent further intra-orchard spread of a disease. As an example, we analysed a spatiotemporal disease map (Fig. 2) corresponding to a peach orchard planted in 1988 on a 2 × 5 m lattice, affected by the *Plum pox virus* strain M (PPV-M). In this orchard, the symptomatic trees in 1992-1993 had been removed immediately and, as new trees showed symptoms in 1994, we were interested to know if they could be caused only by infectious vectors coming from outside the orchard (in which case these new cases should be independent of previous diseased trees).

RESULTS AND DISCUSSION

Numerical Validation

Fig. 3A demonstrates that when two independent processes are simulated, the level of type I error is approximately equal to the predefined 5% level, but for low disease incidence, the test is slightly conservative. In addition, Fig. 3B shows that the power of this test is high (above 90% for the three levels of disease incidence) despite some censoring in the data; for a disease incidence of 3%, it also shows an increased power at the second distance class. During our analyses, we frequently noticed this feature, which is caused by the cumulative nature of the test statistic.

Application to the Experimental Data Set

A preliminary analysis of the map indicated no particular border effect, but we found a very significant aggregation of the disease within each of the two groups of diseased trees. The test of independence between these two groups of cases showed that the observed values of $S_c(d)$ were lying far above the 95% band, for the first distance classes (Fig. 4). The corresponding *P*-values indicated a highly significant dependence between diseased trees at t_1 and t_2 up to 10 m (each distance class encompasses 2 m). Thus, as suggested by the map, there is a tight aggregation between new and earlier PPV-symptomatic trees, despite roguing. In this orchard, it appears clearly that removing symptomatic trees did not immediately stop the autonomous intra-orchard dynamics of the disease, highlighting the need for repeated surveys.

Concluding remarks

This method allows testing the independence between two dates in regular plantings with intra-date clustering and between-date censoring. Such a test can provide valuable clues about pathogen dispersal, especially when the epidemiology of a disease is poorly known, because it makes no assumption about the process of disease spread. The method described here in the context of diseases spreading in orchards can be applied in any regular planting, and more generally in any situation that can be formalized as a test of independence between two aggregated patterns where there is both an internal and an external censoring pattern. This test can be simplified if one of the potentially censoring patterns (either internal or external) is lacking. Moreover, when at least one of the two patterns shows no specific structure, it is possible to avoid using toroidal shifts, and slightly different tests based on the same general principle have been built for this situation (Thébaud et al., 2005). Further corrections could also be included to take into account some simple types of heterogeneity (e.g., edge effects, mixed cultivars) that would otherwise prevent any other explanation for the aggregation between two dates. This set of methods is available for the analysis of disease maps in order to study different epidemiological questions such as the secondary transmission of a disease, the efficiency of control methods, or the association between different systemic diseases.

ACKNOWLEDGEMENTS

We thank Etienne Klein and Marie Gosme for helpful discussions, and Dan Haydon for his comments on the paper. We are also grateful to the Fédération Régionale de Défense contre les Organismes Nuisibles (FREDON) of Languedoc-Roussillon for proving the disease map.

Literature Cited

Chadœuf, J., Brix, A., Pierret, A. and Allard, D. 2000. Testing local dependence of spatial structures on images. J. Microsc. 200: 32-41.

Chadœuf, J., Capowiez, Y., Kretzschmar, A. and Dessart, H. 1997. Testing interaction between a random area process and another spatial process: the analysis of spatial patterns of soil sections. Acta Stereol. 16: 251-258.

Diggle, P. J. 2003. Statistical Analysis of Spatial Point Patterns. 2nd edition. Hodder Arnold, London.

Lotwick, H.W. and Silverman, B.W. 1982. Methods for analysing spatial processes of several types of points. J. Roy. Stat. Soc. B 44: 406-413.

Manly, B.F.J. 1991. Randomization and Monte Carlo Methods in Biology. Chapman and Hall, London.

Mugglestone, M.A., Lee, B.Y.Y., Roff, M.N.M., Jones, P., Plumb, R.T. and Deadman, M.L. 1996. Point process modelling of BYMV incidence in lupins. Aspects Appl. Biol. 46: 87-94

Nelson, S.C. 1995. Spatiotemporal distance class analysis of plant disease epidemics. Phytopathology 85: 37-43.

Perry, J.N. and Dixon, P.M. 2002. A new method to measure spatial association for ecological count data. Ecoscience 9: 133-141.

Peyrard, N., Calonnec, A., Bonnot, F. and Chadœuf, J. 2005. Explorer un jeu de données sur grille par tests de permutation. Rev. Stat. Appl. LIII:59-78.

Reynolds, K.M. and Madden, L.V. 1988. Analysis of epidemics using spatio-temporal autocorrelation. Phytopathology 78: 240-246.

Thébaud, G., Peyrard, N., Dallot, S., Calonnec, A. and Labonne, G. 2005. Investigating disease spread between two assessment dates with permutation tests on a lattice. Phytopathology 95: 1453-1461.

Figures

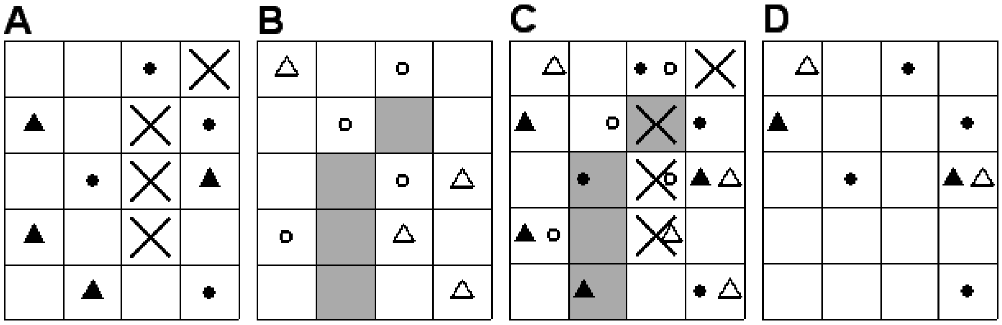

Fig. 1. Handling the censoring patterns in the test of independence between two dates. A: Observed pattern. B: Toroidal shift of the observed pattern (1 unit left and down, here). C: Superimposed observed and shifted patterns. D: The remaining non-censored points used to compute distances between dates. Each square symbolizes one tree. Filled symbols: observed t_1 or t_2 cases; open symbols: shifted t_1 or t_2 cases. Circles: observed or shifted t_1 cases; triangles: observed or shifted t_2 cases. Crosses: observed missing trees; grey squares: shifted missing trees.

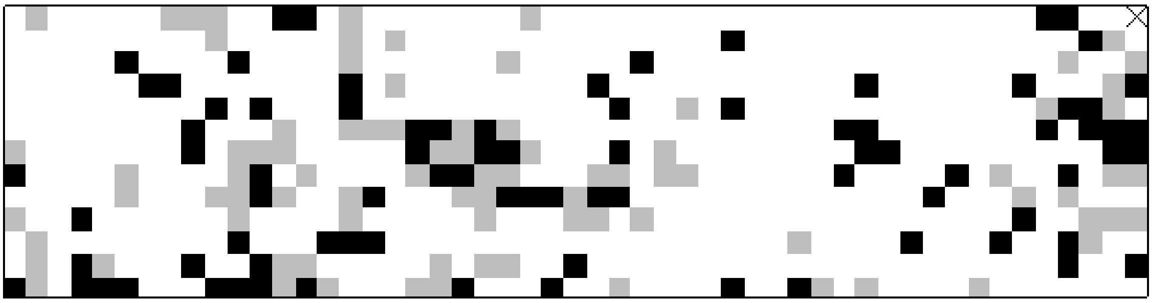

Fig. 2. Map of a PPV-infected orchard (100 × 65 m). Each square symbolizes one tree. Black squares: trees with PPV symptoms in 1992-1993; grey squares: trees with PPV symptoms in 1994; cross: missing tree.

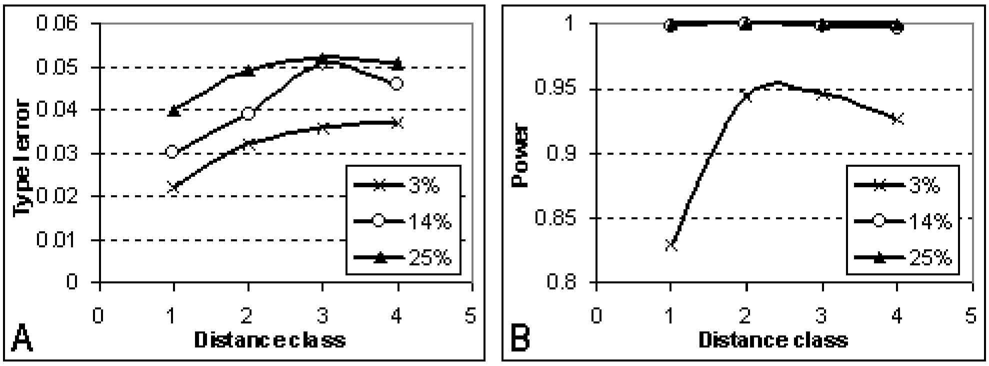

Fig. 3. Validation of the test on simulated patterns for three levels of disease incidence (3, 14, and 25%). A: Level of type I error when aggregated but independent patterns are simulated (the test has a nominal error rate of 5%). B: Power of the test to reject the hypothesis of independence when dependent patterns are simulated.

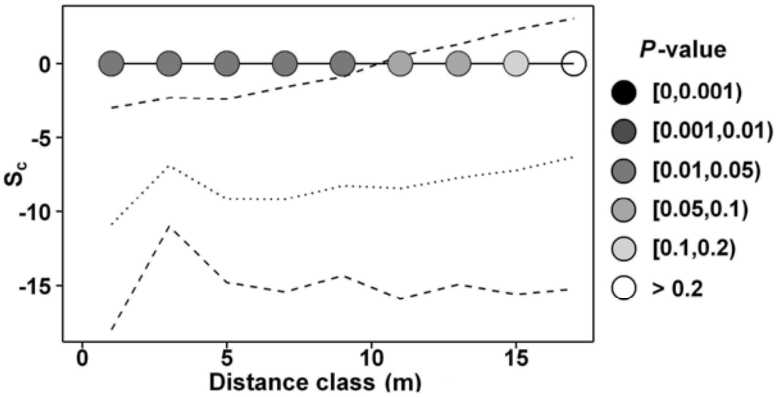

Fig. 4. Spatial dependence between the trees with *Plum pox virus* symptoms removed in 1992-1993 and the symptomatic trees observed in 1994. Dotted line: mean value of the test statistic when the relative position of the two patterns is randomly shifted. The dashed lines delimit the central 95% of the simulated values, in which the observed values (circles) should lie if the patterns were independent. The shaded points indicate the *P*-value of the test for each distance class.