Ecological Monographs, 84(3), 2014, pp. 477–489
© 2014 by the Ecological Society of America

On tests of spatial pattern based on simulation envelopes

Adrian Baddeley,[1,2,6] Peter J. Diggle,[3,4] Andrew Hardegen,[5] Thomas Lawrence,[5] Robin K. Milne,[5] and Gopalan Nair[2,5]

[1]*Center for Exploration Targeting (M006), University of Western Australia, 35 Stirling Highway,
Crawley, Western Australia 6009 Australia*
[2]*CSIRO Computational Informatics, 65 Brockway Road, Floreat, Western Australia 6014 Australia*
[3]*Department of Epidemiology and Population Health, University of Liverpool, Liverpool L69 3BX United Kingdom*
[4]*Medical School, Lancaster University, Lancaster LA1 4YB United Kingdom*
[5]*School of Mathematics and Statistics, University of Western Australia, 35 Stirling Highway,
Crawley, Western Australia 6009 Australia*

Abstract. In the analysis of spatial point patterns, an important role is played by statistical tests based on simulation envelopes, such as the envelope of simulations of Ripley's K function. Recent ecological literature has correctly pointed out a common error in the interpretation of simulation envelopes. However, this has led to a widespread belief that the tests themselves are invalid. On the contrary, envelope-based statistical tests are correct statistical procedures, under appropriate conditions. In this paper, we explain the principles of Monte Carlo tests and their correct interpretation, canvas the benefits of graphical procedures, measure the statistical performance of several popular tests, and make practical recommendations. There are several caveats including the under-recognized problem that Monte Carlo tests of goodness of fit are probably conservative if the model parameters have to be estimated from data. Finally, we discuss whether graphs of simulation envelopes can be used to infer the scale of spatial interaction.

Key words: confidence bands; conservative test; deviation test; global test; K function; Monte Carlo test; null model; pair correlation function; pointwise test; scale of interaction; spatial point pattern; variance stabilization.

Introduction

Recent literature in statistical ecology has questioned the validity of a well-established technique for testing hypotheses about spatial point patterns. Pioneered by Ripley (1977) and popularized in statistical ecology by Kenkel (1988), this technique is based on computing a summary function of the point pattern, such as Ripley's K function, and comparing it with the envelope of the same functions obtained from several simulations of the null model.

In an influential article, Loosmore and Ford (2006) correctly pointed out a common error in the interpretation of such simulation envelopes. As an alternative to simulation envelopes, they advocated a test procedure that does not require graphical display, and is based on a numerical index of deviation between the summary functions. However, their findings have been widely misinterpreted as meaning that simulation envelopes do not have any valid interpretation as a significance test. Many writers have agreed, advising that "envelope tests should not be thought of as formal tests of significance" (Law et al. 2009:619) and drawing a distinction between invalid envelope tests and valid deviation tests (Grabarnik et al. 2011).

This opinion is at odds with the viewpoint widely held in statistical science, especially in spatial statistics (Diggle 2003, Illian et al. 2008:455–459) and nonparametric statistics (Bowman and Azzalini 1997, Chandler and Scott 2011), that simulation envelopes do have a valid statistical interpretation. Indeed envelope tests and deviation tests have the same statistical rationale, that of a Monte Carlo test (Barnard 1963, Hope 1968), so it cannot be true that envelope tests are invalid while deviation tests are valid.

It is, therefore, not surprising that the scientific literature has become confusing and contradictory about questions of the validity of statistical inference for spatial point patterns, and about practical recommendations. This paper is intended to clarify these matters. It also contributes some new insights, based partly on new experimental findings.

Envelopes discusses several different kinds of simulation envelopes and their interpretation. *Monte Carlo tests and envelopes* explains the fundamental principles of Monte Carlo tests and their application to simulation envelopes. *Null models with parameters* mentions an important caveat on the validity of Monte Carlo tests that appears to have been overlooked in ecological literature. *Performance* evaluates the performance (power) of these tests and makes recommendations for maximizing performance. *Interaction* warns about the difficulty of using summary functions to infer the scale

Manuscript received 31 October 2013; accepted 5 November 2013. Corresponding Editor: A. Ellison.

[6] E-mail: adrian.baddeley@uwa.edu.au

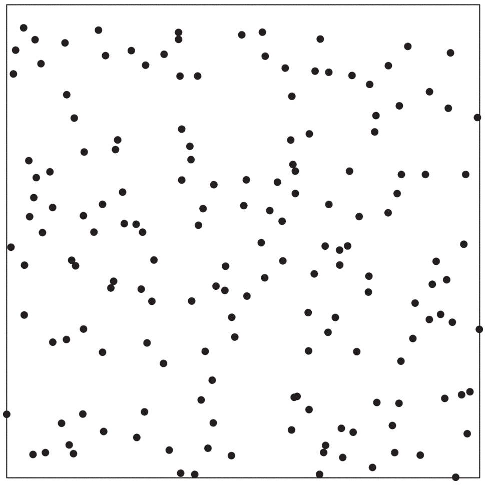

FIG. 1. Synthetic data, illustrating the locations of seedlings in a 100 × 100 m survey plot, exhibiting inhibition between points at distances up to 4 m.

of spatial interaction. Conclusions are summarized in *Conclusions*.

ENVELOPES

In this section, we introduce several different kinds of envelopes and briefly discuss their interpretations.

Fig. 1 shows a typical spatial point pattern that might have been obtained by mapping the locations of seedlings in a survey plot. It was actually generated synthetically, so that the correct answer is known. The points are spatially inhibited at distances up to 4 m, where the survey plot is a square with sides of 100 m.

Pointwise envelopes

Fig. 2 shows the type of graphic discussed by Kenkel (1988) and Loosmore and Ford (2006) and originating from Ripley (1977). Ripley's K function was computed from the data in Fig. 1, and transformed to Besag's L function $L(r) = \sqrt{K(r)/\pi}$. This is plotted as the solid black curve in Fig. 2A. Then 39 simulated, random point patterns were generated according to complete spatial randomness (CSR), the L functions of each simulated pattern were computed, and the maximum and minimum L values were plotted as the limits of the gray-shaded envelope in Fig. 2A. Fig. 2B shows the same information after subtracting the theoretical value for CSR and restricting the distance variable r in $L(r)$ to the range 0–10 m, which highlights the important details. Since the solid black line wanders outside the gray-shaded envelope in some places, there is a suggestion that Fig. 1 is not a completely random pattern.

This procedure can be used for any null hypothesis that can be simulated. In this example, we took the null model to be CSR to keep the exposition simple, and because it is used to illustrate a technical point in *Null models with parameters*. We hasten to add that CSR is usually not the most appropriate null hypothesis in ecological applications (Loosmore and Ford 2006:1929). The null hypothesis should not be chosen naively, but should correspond to a scientifically meaningful scenario in which "nothing is happening" (Strong 1980). In a study of the spatial pattern of plants, the appropriate null hypothesis might involve spatial inhomogeneity due to known environmental factors, and spatial clustering and regularity due to known processes of dispersal and competition. Exceptions occur in the initial investigation (Diggle 2003:12; Baddeley 2010), where CSR serves as a dividing hypothesis that is often known to be false (Cox 1977:51–52).

A fundamental issue raised by Loosmore and Ford (2006) concerns the statistical significance of a simulation envelope. In Fig. 2, the L function for the data point pattern wanders outside the envelope of the L functions of 39 simulated patterns generated under

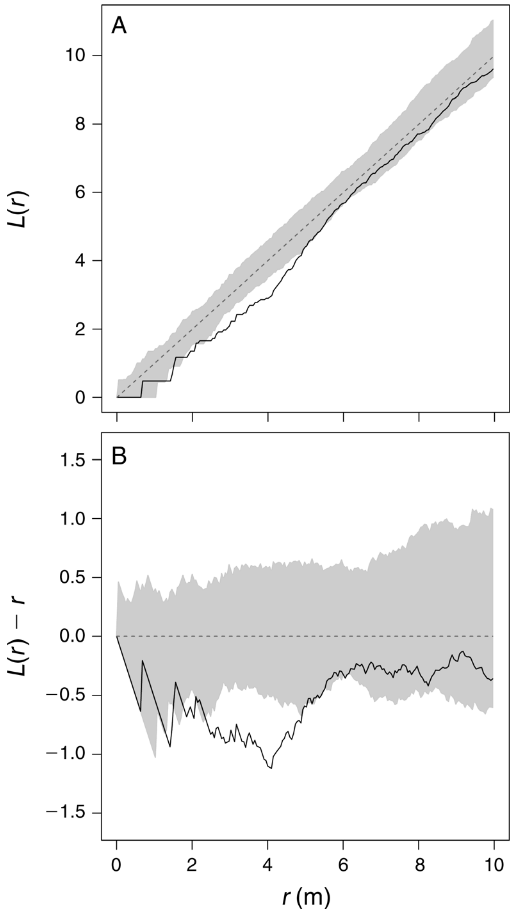

FIG. 2. Analysis of spatial pattern using a pointwise envelope. Panel (A) shows the L function (defined in *Envelopes: Pointwise envelopes*); panel (B) shows the centered L function, $L(r) - r$ (where r is a distance variable). Solid lines represent the value computed from the data pattern from Fig. 1. Shading indicates the envelope of values obtained from 39 simulations of complete spatial randomness (CSR). Dashed lines show the theoretical value for CSR.

CSR. Kenkel (1988) and many subsequent writers attach a statistical significance of $2/(39 + 1) = 0.05$ to this outcome. An intuitively reasonable explanation is that, if the data were also CSR, it would be a chance of 1 in 40 that the data would give an L value smaller than all 39 simulated values, and another 1:40 of giving a value larger than all the simulated values.

Loosmore and Ford (2006:1926) correctly point out that this reasoning would be valid only if we had decided in advance to inspect the L function at one particular interpoint distance (on the horizontal axis). For example in Fig. 2, if we had fixed the distance of 4 m in advance of seeing the data, then the outcome would indeed be statistically significant at the level 0.05. However, it would be very rare to have such accurate knowledge of the scale of spatial patterning. The usual practice is to plot the simulation envelope without any preconceived idea, and look for deviations from the envelope at *any* position. This practice leads to invalid statistical inferences. The probability that the L function will wander outside the envelope somewhere is very much higher than 0.05, so that the results in Fig. 2 cannot then be declared statistically significant at the 0.05 level.

The same caveat was already mentioned in Ripley's pioneering paper (Ripley 1977:181) and has been emphasized by many expositors of spatial statistics (e.g., Diggle 1983:12). It was certainly correct of Loosmore and Ford (2006) to draw attention to this common error.

However, this caveat does not imply that simulation envelopes are statistically invalid. The pointwise envelope does give a valid significance test when it is correctly interpreted (see *Monte Carlo tests and envelopes: Monte Carlo tests based on functional summary statistics*). Moreover, there are other types of simulation envelopes that do not suffer from the same problem (see *Envelopes: Global envelopes*).

One benefit of graphically displaying a summary function like the K function is that it contains information from different spatial scales. However, extracting this information is not straightforward. Many writers estimate the scale of spatial interaction by reading off the position where the observed K function lies furthest outside the simulation envelope. Loosmore and Ford (2006) correctly point out some flaws in this logic. We discuss this further in *Scales of interaction*.

As an alternative to envelope tests, Loosmore and Ford (2006) advocate a different, statistically valid test for spatial pattern, which does not require graphical display. This test is discussed in *Monte Carlo tests and envelopes: Monte Carlo tests based on functional summary statistics*.

Global envelopes

Another statistically valid test based on simulation envelopes is shown in Fig. 3. Distinct from the pointwise envelope in Fig. 2, the global envelope in Fig. 3 is a zone of constant width. The width is determined by finding the most extreme deviation from the theoretical L function that is achieved by any of the 39 simulated L functions, at any distance r along the horizontal axis. Global envelopes were also proposed by Ripley (1977, 1979, 1981), but do not appear to have been used in the ecological literature; in particular, they are not discussed by Kenkel (1988) or Loosmore and Ford (2006).

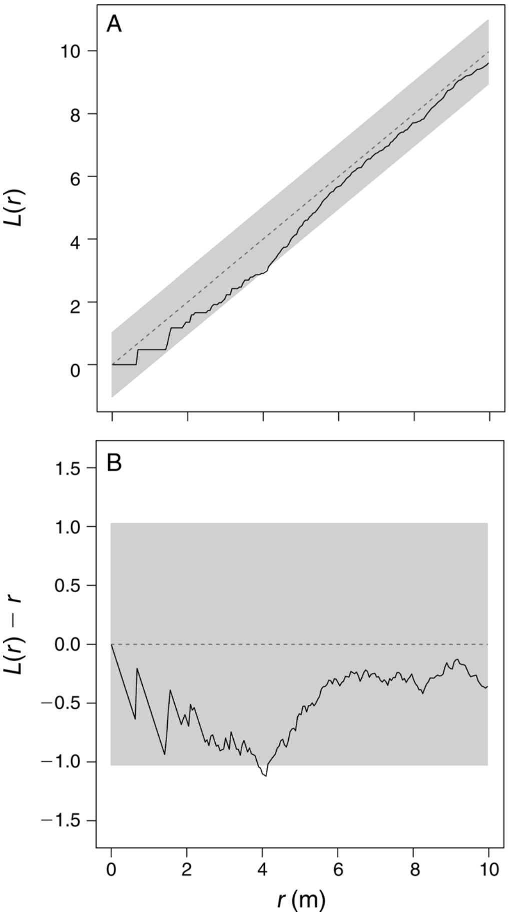

FIG. 3. Analysis of spatial pattern using a global envelope (zone of constant width). Figure components are defined as in Fig. 2. The gray region has constant width and encloses the L functions of 19 simulations of CSR.

The global envelope in Fig. 3 is based on 19 simulated point patterns. The centered L function of the data point pattern wanders outside the global envelope; this is statistically significant at the level $1/(1+19) = 0.05$, often reported in terms of the P value as $P < 0.05$. The global envelope test is statistically valid; the critique of Loosmore and Ford does not apply. Further explanation is given in *Monte Carlo tests and envelopes: Monte Carlo tests based on functional summary statistics*.

Confidence envelopes

Many writers describe the simulation envelope as a confidence envelope, confidence interval, or similar (Kenkel 1988:1020, Loosmore and Ford 2006:1926).

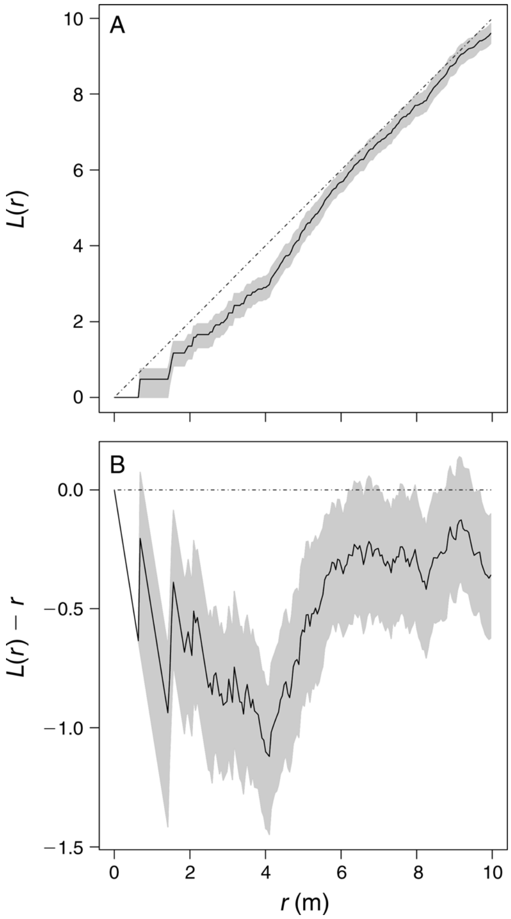

FIG. 4. Pointwise 95% confidence bands for (A) the L function and (B) the centered L function of the pattern in Fig. 1, using the bootstrap method of Loh (2008). The dot-dash line is the expected value under CSR.

Such usages are not standard statistical terminology and may be a source of confusion.

In brief, a confidence interval is designed to contain the true value of the target quantity with a specified degree of confidence. A confidence interval is usually centered around an estimated value of the target quantity. By contrast, an acceptance interval (or non-rejection interval) represents a statistical hypothesis test; it is the range of values that are not significantly different from the null value. An acceptance interval is usually centered around the hypothesized value of the target quantity.

True confidence intervals or confidence bands for the K function can be constructed using several well-established methods. Fig. 4 shows a pointwise 95% confidence band for the true K function of the natural process that generated the data in Fig. 1. The confidence band was obtained using a bootstrap method proposed by Loh (2008). Note that the bands are centered around the K function of the data. Such confidence bands might be useful in some applications of statistical ecology; they may be a better alternative to formal hypothesis testing (Yates 1951, Cox 1977).

MONTE CARLO TESTS AND ENVELOPES

All statistical tests described in this paper are based on the same underlying rationale, namely that of a Monte Carlo test (Barnard 1963, Hope 1968), which we now describe.

Fundamentals of Monte Carlo tests

There are two essential requirements of any Monte Carlo test. First, we must reduce the observed data to a single numerical value, the test statistic t_{obs}. There is considerable latitude in how we do this. In a forest survey, if the observed data are the spatial coordinates of trees, we could take t_{obs} to be any single number that is a summary of the spatial locations, for example, the Clark and Evans (1954) statistic, the mean nearest-neighbor distance, or the value $K(4)$ of the K function at a fixed distance of 4 m. (However, note that the K function itself is not a single number summary, since it depends on the variable r.)

Secondly we must be able to generate simulated random outcomes under the null hypothesis. These are synthetic (computer-generated) sets of observations, similar to the original set of observations, but generated under the assumption that the null hypothesis is true. To perform the Monte Carlo test, we generate m sets of simulated random observations, where typically m is 19, 39, or 99.

For example, in a forest survey, if the null hypothesis is that the tree locations are CSR, then we must be able to generate a random point pattern according to that (spatial) model. Suppose that we first generate $m = 19$ such patterns independently. Then we reduce each pattern to a single number summary, using the same technique as was applied to the original observations, yielding $t_1, t_2, \ldots, t_{19}$, and we determine the largest of these values $t_{\max}$ (for a one-sided test).

In this case, the Monte Carlo test procedure is that if the value t_{obs} for the observed data is larger than the maximum of the simulated values $t_{\max}$, then reject the null hypothesis at a significance level of $\alpha = 0.05$ (often reported in terms of the P value as $P < 0.05$, although in this case the P value is exactly equal to 0.05).

The rationale of the Monte Carlo test is as follows. Assume the null hypothesis is true (and that any parameters of the null model are known). Then the original data and the 19 simulated patterns are statistically equivalent, so the test statistic value t_{obs} calculated for the original data, and the test statistic values $t_1, t_2, \ldots, t_{19}$ calculated for the simulated patterns, are statistically equivalent. By symmetry, there is a 1:20 chance that the test statistic value t_{obs} is the largest of these 20 numbers, that is, that t_{obs} is larger than (each of) the other 19 values $t_1, t_2, \ldots, t_{19}$. Hence the result is statistically significant at level $\alpha = 1/20$ or $\alpha = 0.05$.

There are several modifications of this basic test. Instead of a one-sided test that rejects the null hypothesis when t_{obs} is large, we could have a two-sided test that rejects the null hypothesis when t_{obs} is either the largest or smallest of the $m + 1$ numbers. This has a significance level of $\alpha = 2/(m + 1)$. If we want the standard significance level of 0.05, then we would typically choose $m = 39$ simulations giving a significance level of $2/40 = 0.05$.

Instead of taking the maximum and minimum of the simulated values, a Monte Carlo test can be performed using the kth largest and kth smallest values out of m simulations, where k is a chosen rank. For a one-sided test, we reject the null hypothesis if the observed value t_{obs} is larger than the kth largest simulated value. This test has significance level $\alpha = k/(m + 1)$. For a two-sided test, we reject the null hypothesis if the observed value t_{obs} is either larger than the kth largest simulated value, or smaller than the kth smallest simulated value. This test has a significance level of $\alpha = 2k/(m + 1)$.

Many choices of m and k will give a test of significance level $\alpha = 0.05$. Examples include a one-sided test using the fifth largest simulated value out of 99 simulations, and a two-sided test using the fifth largest and fifth smallest simulated values out of 199 simulations. Researchers are free to make their own choices of m and k with an eye to computational cost, standards of evidence, performance, and other factors. This flexibility may explain the apparent inconsistencies in Table 1 of Loosmore and Ford (2006).

If we prefer to compute a P value instead of specifying a fixed significance level α, the test procedure is as follows. Count the number of simulated values t_i that are greater than the observed value t_{obs}. If there are j such values then the P value is $(j + 1)/(m + 1)$ for a one-sided test, or $2 \min(j + 1, m + 1 - j)/(m + 1)$ for a two-sided test, where min(a, b) denotes the minimum (the smaller) of the two numbers a and b.

Monte Carlo tests based on functional summary statistics

Monte Carlo tests for spatial pattern, based on simulation envelopes, were pioneered by Ripley (1977) and their statistical rationale was set out in detail by Besag and Diggle (1977) and Ripley (1979, 1981; see also Marriott 1979).

The test may be performed using any distance-based summary function for the point pattern. Typical choices would be the K function, the L function, or the nearest-neighbor distance function G (Diggle 2003). The letter H stands for any summary function of our choice. To perform the test, we first calculate the H function estimate for the observed data, $H_{\mathrm{obs}}(r)$. Next we generate m random simulated point patterns, and obtain their H function estimates, $H_1(r), \ldots, H_m(r)$.

The Monte Carlo test principle in *Monte Carlo tests and envelopes: Fundamentals of Monte Carlo tests* cannot be applied to the function $H(r)$ directly, because functions cannot be ranked from smallest to largest. We need to reduce the function $H(r)$ to a single number, T, which will serve as the test statistic. Three possible strategies are described here.

Pointwise test.—One strategy is to take T as the value of the H function at a fixed distance. For example, $T = H(4)$ would be the numerical value of the H function at the prespecified distance of 4 m. This reduces the data to a single number T.

The two-sided Monte Carlo test (in the simplest case) rejects the null hypothesis if the observed value $t_{\mathrm{obs}} = H_{\mathrm{obs}}(4)$ lies outside the range of all the simulated values $t_1, \ldots, t_m$, that is $H_1(4), \ldots, H_m(4)$. This rule is equivalent to plotting the envelope of the simulated H functions $H_1(r), \ldots, H_m(r)$, and rejecting the null hypothesis if the data H function $H_{\mathrm{obs}}(r)$ lies outside this envelope at the prespecified distance $r = 4$ m. In the more general case, the upper and lower boundaries of the envelope are, for each value of r, the kth largest and kth smallest of the simulated values of $H(r)$.

Thus if the distance value r were fixed in advance of performing the analysis, for example if prior information indicates that (under the alternative hypothesis) seedlings are likely to compete over distances up to $r = 4$ m, then a valid test of statistical significance is to reject the null hypothesis if the H function lies outside the simulation envelope at distance $r = 4$ m. Admittedly, such situations are somewhat artificial.

Different values of the distance variable r correspond to different Monte Carlo tests. Plotting the pointwise envelope over a range of r values (see Fig. 2) displays the outcomes of these different tests. Each test is valid if performed on its own, but it is generally not valid to perform all the tests and to combine their results in some arbitrary fashion. In particular, it is not valid to scan the plot searching for values of r where the empirical function $H_{\mathrm{obs}}(r)$ falls outside the envelope. This is a problem of multiple testing or multiple comparison (Hochberg and Tamhane 1987, Shaffer 1995, Hsu 1996).

However, plotting the envelope over a range of r values enables us to assess, for different distance values of r, what the result of the test would have been if we had chosen that distance value to perform the test. This is a useful diagnostic, since it indicates the sensitivity of the test outcome to the choice of the distance value of r. The use of such significance traces is standard practice in other fields of statistics (Bowman and Azzalini 1997, Chandler and Scott 2011). However, we concede that the interpretation of the pointwise envelope is fraught with danger in the wrong hands.

Global or MAD test.—An alternative choice of test statistic is the maximum deviation between the H function of the observed data and the theoretical (expected) H function of the null model.

In some cases, the theoretical value of the H function under the null model is known. For example under CSR, the K function takes the form $K(r) = \pi r^2$ and so $L(r) = r$. In such cases, we may take the test statistic T to be the maximum deviation, in absolute value, between $H(r)$

and its theoretical value $H_{\text{theo}}(r)$ under the null model, where the maximum is taken over the range of distances from 0 to R m, where R is a chosen upper limit on the interaction distance (to be discussed in *Null models with parameters: Tests of goodness of fit of a clustered model*). That is, we choose

$$T = \max_{0 \le r \le R} |H(r) - H_{\text{theo}}(r)|. \qquad (1)$$

Then T is the maximum vertical separation between the graphs of $H(r)$ and $H_{\text{theo}}(r)$ over the chosen range of distances. This procedure reduces the data to a single number T called the maximum absolute deviation (MAD).

The one-sided Monte Carlo test procedure (in the simplest case) rejects the null hypothesis if t_{obs} is greater than the maximum $t_{\max}$ of all the simulated values $t_1, \ldots, t_m$. This rule is equivalent to plotting an envelope of constant width $t_{\max}$ centered on the theoretical curve $H_{\text{theo}}(r)$, and rejecting the null hypothesis if the observed H function $H_{\text{obs}}(r)$ ever wanders outside this envelope. This is a global envelope test or MAD test, with significance level $\alpha = 1/(m+1)$. See Fig. 3 for an example with $m = 19$ so that $\alpha = 0.05$. In the general case, the one-sided Monte Carlo test rejects the null hypothesis if t_{obs} is greater than the kth largest of the simulated values, and this has significance level $\alpha = k/(m+1)$.

If the theoretical value $H_{\text{theo}}(r)$ is not known for every r, then it could be estimated from a separate set of simulations of the null model (Diggle 1983), which guarantees the basic requirement of symmetry. Alternatively, using only a single set of simulations, we can replace $H_{\text{theo}}(r)$ with

$$\bar{\bar{H}}(r) = \frac{1}{m+1}(H_1(r) + \ldots + H_m(r) + H_{\text{obs}}(r)) \qquad (2)$$

the average of all the simulated and observed H functions. This preserves symmetry and ensures that the test has significance level $1/(m+1)$.

Test proposed by Loosmore and Ford.—As an alternative to envelope tests, Loosmore and Ford (2006) propose a Monte Carlo test based on the test statistic

$$T = \int_0^R (H(r) - H_{\text{theo}}(r))^2 \, dr \qquad (3)$$

the integral of the squared difference between the H function and its theoretical value under the null hypothesis. The formula given in Loosmore and Ford (2006: Eq. 3) is different, but is equivalent to our Eq. 3, as shown in the Appendix.

Tests of this kind were previously proposed by Diggle (1986:122, 2003:12) and Cressie (1991:667), so we shall call this the Diggle-Cressie-Loosmore-Ford (DCLF) test. If the theoretical value H_{theo} has to be estimated then Eq. 3 is replaced by

$$T = \int_0^R (H(r) - \bar{\bar{H}}(r))^2 \, dr \qquad (4)$$

where again $\bar{\bar{H}}$ is the average of the simulated and observed H functions as in Eq. 2. This is not precisely as advocated by Loosmore and Ford (2006), but is algebraically equivalent, as shown in the Appendix.

At face value, this test does not appear to have a graphical representation in terms of simulation envelopes. However, such an interpretation does exist. The test statistic T in Eqs. 3 or 4 depends on the choice of the upper limit on distances, R. For any specified value of R, suppose $t_{\text{obs}}(R)$ is the observed value of the test statistic and $t_i(R)$ is the ith simulated value. Then the DCLF test at significance level $\alpha = 1/(m+1)$ based on the interval [0, R] is to reject the null hypothesis if $t_{\text{obs}}(R) > t_{\max}(R)$ where $t_{\max}(R) = \max_i t_i(R)$ is the largest simulated value. Thus, by plotting $t_{\text{obs}}(R)$ and $t_{\max}(R)$ against R, we can represent the outcome of the DCLF test for each R. We call this the envelope representation of the DCLF test.

Fig. 5 shows the envelope representation for the DCLF test applied to the data of Fig. 1 using the L function. Fig. 5A shows the outcome based on the same 19 simulations as were used to generate Fig. 3. The DCLF test statistic $t_{\text{obs}}(R)$ and the Monte Carlo critical value $\max_i t_i(R)$ are plotted against R. The DCLF test rejects the null hypothesis at the 0.05 level when R is between 1 and 15 m, but not when R is greater than 15 m.

The same principle applies when we take the kth largest of the simulated values. Fig. 5B is based on $m = 1999$ simulations and rank $k = 100$, giving the same significance level $\alpha = 0.05$. The DCLF test statistic $t_{\text{obs}}(R)$ is plotted against R, and the gray shading shows the acceptance region (non-rejection region) delimited by the Monte Carlo critical value, the kth largest of the m values $t_1(R), \ldots, t_m(R)$. Based on this larger suite of simulations, the DCLF test now rejects the null hypothesis at the 0.05 level for all values of R.

Choice of test.—The three strategies outlined above are equally valid. The choices of summary function H and test statistic T materially affect the performance of the test, but not its basic validity. It will often be sensible to transform the summary function H (for example, transforming the K function to the L function) or to weight the values of H according to their variances under the null hypothesis (Cressie 1991:642) in order to improve performance. Factors that affect performance are discussed later in this paper.

Null Models with Parameters

An important caveat about Monte Carlo tests is that they "are strictly invalid, and probably conservative, if parameters have been estimated from the data" (Diggle 2003:89). This seems to have received little attention in applications.

To be more precise, the problem arises when the null hypothesis is a model depending on a parameter or parameters θ that must be estimated from the data (known as a composite hypothesis). The usual procedure is to fit the model to the observed point pattern, obtaining an estimate $\hat{\theta}$ of the parameter value, and then to generate the simulated point patterns from the model using this fitted parameter value $\hat{\theta}$. But this violates the essential requirement of Monte Carlo testing, that the observed and simulated point patterns should be statistically equivalent if the null hypothesis is true. Under the procedure just described, the simulated point patterns have been generated from the null model with parameter value $\hat{\theta}$, while (if the null hypothesis is true) the observed point pattern came from the null model with unknown parameter value θ. Since the estimate $\hat{\theta}$ is usually not exactly equal to the true parameter value θ, the simulated and observed point patterns do not come from the same random process, so the Monte Carlo test is invalid.

This effect often causes Monte Carlo tests to be conservative. A test is called conservative if the true significance level (the true probability of Type I error) is smaller than the reported significance level, or the significance level that we quote in reporting the outcome of the test. A conservative test may conclude that the data are not statistically significant when in fact they should be declared statistically significant.

Tests of CSR

This problem arises even in the simplest case of testing CSR. The null hypothesis of CSR has one free parameter, namely the intensity λ (the average number of points per unit area). Typically we estimate λ from our observed data by $\hat{\lambda} = n_{\text{obs}}/A$ where n_{obs} is the number of points in the observed point pattern and A is the area of the survey region. Then the simulated point patterns are generated by CSR with intensity $\hat{\lambda}$. Hence, any Monte Carlo test based on these simulations is strictly invalid and probably conservative.

To investigate whether conservative Monte Carlo tests are a substantial problem, we evaluated the true significance level of the DCLF Monte Carlo test of the null hypothesis of CSR with nominal significance level $\alpha = 0.05$ based on 19 simulations. Results are shown in Table 1 and explained here.

The test was repeated 100 000 times using a supercomputing cluster. For each replicate of the test, a synthetic data set X was generated according to CSR in a 100-m square, with intensity $\lambda = 0.005$ points per square m, giving a mean of 50 points. For the row labeled "unconditional," the test was performed in the usual way, first estimating the intensity from the observed pattern by $\hat{\lambda} = n(X)/A$ (where $A = 10^4$ m^2 is the window area) and then generating 19 simulations of CSR with intensity $\hat{\lambda}$. Summary statistics were Ripley's K and its transformed version L, the nearest-neighbor distance function G and its transformed version $G\dagger$, explained in *Performance: Factors that affect power*. The domain of integration for the DCLF test statistic was (0, 25) m for K and L and (0, 10) m for G and $G\dagger$, based on standard conventions (Ripley 1981, Diggle 2003).

Entries in the table show the proportion of rejections of the null hypothesis at nominal significance level $\alpha = 0.05$, counted in 100 000 applications of the test. The standard error of these estimates is roughly $\sqrt{0.05 \times 0.95/100\,000} = 0.0007$. The "unconditional" row of Table 1 shows that the usual Monte Carlo test is quite conservative, with true significance as low as 0.025 against a nominal significance of 0.05. Using a test statistic that is more sensitive to the point process intensity (e.g., using G rather than K) makes the test more conservative.

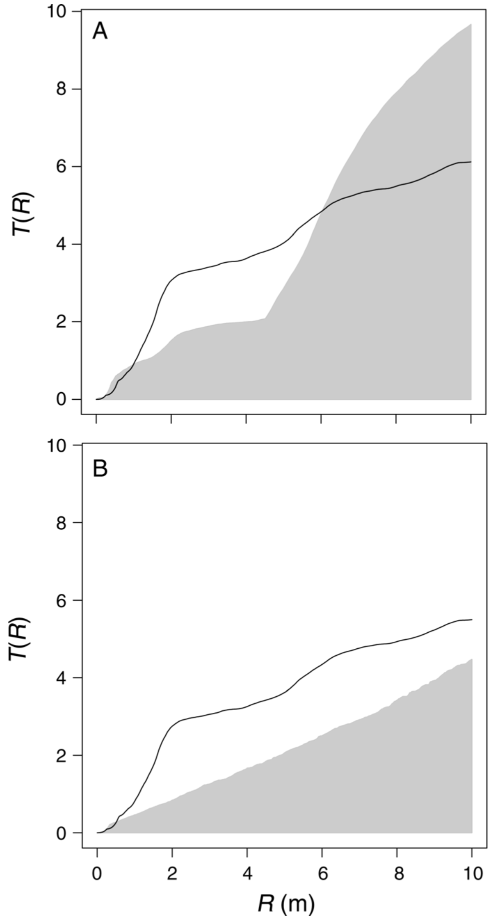

FIG. 5. Envelope representation for the Diggle-Cressie-Loosmore-Ford (DCLF) test of CSR with $\alpha = 0.05$, applied to the data of Fig. 1 and considering the full range of distance values from 0 m to 25 m. DCLF test statistic T (solid lines) and Monte Carlo acceptance/non-rejection region (shaded) are plotted as a function of the length R of the distance interval. Panel (A) is based on 19 simulations of CSR; critical value is the maximum of the simulated values. Panel (B) is based on 1999 simulations of CSR; critical value is the 100th largest of the simulated values.

In the case of CSR, a solution to this problem is to hold the number of points fixed. We generate the simulated patterns with the same number of points as the observed pattern. This exploits a special property of CSR that, if the number of points is known, the locations of the points are independent and uniformly distributed, whatever the value of the intensity. That is, conditional on the observed number of points, the

Table 1. Estimated actual significance level (probability of Type I error) of Diggle–Cressie–Loosmore–Ford (DCLF) test with nominal significance 0.05 in three applications.

		Summary function			
Model	Simulations	*K*	*L*	*G*	*G*†
CSR	unconditional	0.0433	0.0390	0.0239	0.0257
CSR	conditional	0.0498	0.0498	0.0504	0.0502
Cluster	unconditional	0.0115	0.0103	0.0018	0.0028

Notes: CSR is complete spatial randomness with a mean of 50 points in the window. Cluster is a Matérn Cluster process with a mean of 250 points in the window. *K* is Ripley's *K* function; *L* is Besag's transformation of the *K* function; *G* is the nearest-neighbor distance distribution function; and *G*† is the variance-stabilized version of *G*.

spatial distribution of the points does not depend on the intensity parameter. It follows that the Monte Carlo test of CSR, when conditional on the number of points, is exact (non-conservative).

For the row labeled "conditional" in Table 1, the test was performed conditionally on the observed number of points $n(X)$, so that 19 simulated random patterns were generated containing $n = n(X)$ independent uniformly distributed random points. The true significance levels shown in the table are equal to the nominal value 0.05 within sampling error.

Tests of goodness of fit of a clustered model

In the more realistic scenario where the null hypothesis is a model involving interpoint interaction, requiring several parameters to be fitted, the problem of conservatism may be greater and yet harder to handle.

As an example, we took a clustered point process model defined by Matérn (1960:46–47; see Diggle 2003:64–67). Parents form a homogeneous Poisson process (CSR) with intensity κ. Each parent, independently, produces a random number of offspring, according to a Poisson distribution with mean μ. Each offspring, again independently, is placed randomly and uniformly within a circle of radius R centered on the parent point. The parents are then discarded, and the offspring constitute the cluster process.

For this model, Table 1 shows the true probability of a Type I error for the DCLF test of goodness of fit. The table is based on 50 000 replicates of the test. For each replicate, a synthetic realization of the cluster process was generated with parent intensity $\kappa = 0.005$, $\mu = 5$, and $R = 14$ m in a 100-m square. To perform the test on a given data set, the model was fitted to the data using the method of minimum contrast (Pfanzagl 1969, Diggle and Gratton 1984) and 19 simulations were generated from this fitted model. Table 1 shows that this test is extremely conservative.

The correct handling of P values for composite null hypotheses is regarded as an unresolved research problem in statistical inference (Brooks et al. 1997, Bayarri and Berger 2000, and Robins et al. 2000). Various strategies have been suggested, including adjusting the test statistic so that its mean value is less sensitive to the parameter (Robins et al. 2000), using a summary function that is unrelated to the model fitting procedure (Diggle 2003:89), and adjusting the P value itself by performing additional simulations (Brooks et al. 1997, Dao and Genton 2014).

A conservative test may be tolerable in some applications, since it effectively applies a standard of statistical significance that is more stringent than intended. Our experiments suggest that Monte Carlo tests using the K and G functions are likely to be very conservative when model parameters must be estimated. The current usage of Monte Carlo tests (envelope or deviation tests) in spatial statistics is defensible for some purposes, such as exploratory data analysis. However, for goodness-of-fit testing, a very conservative test is problematic, since a non-significant outcome (where the null hypothesis is not rejected) is often misinterpreted as confirmation of the fitted model. For definitive formal inference and for goodness-of-fit tests, it would be wise to adjust the P values or the test statistic as discussed above.

Performance

The performance of a test is measured by its power (the probability of making the correct decision if the null hypothesis is false). Here we discuss various strategies that can improve the power of a Monte Carlo test, and we measure the power of tests that are based on summary functions.

Factors that affect power

The power of a test is defined as the probability of rejecting the null hypothesis when the null hypothesis is false and a specified alternative hypothesis is true (i.e., when the data were actually generated in some other specified way). The test power depends on this alternative hypothesis and can be regarded as a measure of the sensitivity of the test to the specified alternative. A test may have strong power against one alternative and weak power against another alternative.

All the choices involved in designing a Monte Carlo test will affect the power of the test. The power typically increases if we increase the number of simulations m, for a fixed level of significance α. The power is affected by the choice of summary function H and the test statistic T.

The choice of summary function is pivotal in determining the sensitivity of the test to different types of spatial pattern. For example, an envelope test based on the nearest-neighbor distance distribution G is usually very sensitive to the presence of inhibition between points, but insensitive to clustering. In extreme cases, two different spatial point process models may have exactly the same summary function. Baddeley and Silverman (1984) constructed a point process that has the same K function as CSR, but is manifestly different from CSR. A test of CSR based on the K function has no utility against this alternative.

The power of the test depends partly on the sampling variability of the summary function H. For the K function, pointwise envelopes tend to have a funnel shape, because the mean and variance of $K(r)$ for a random point pattern are approximately proportional to r^2. This means that the MAD test statistic (Eq. 1) and the DCLF test statistic (Eq. 4) tend to be more influenced by fluctuations of $K(r)$ occurring at larger distances r. In a discussion of Ripley (1977), Besag (1977) proposed transforming $K(r)$ to $L(r) = \sqrt{K(r)/\pi}$ before applying the MAD test. This transformation stabilizes the variance, making $L(r)$ approximately constant as a function of r. Variance stabilization substantially improves the power of the MAD test and of the DCLF test.

The nearest-neighbor distance distribution function G and the empty space function F (Ripley 1981, Diggle 2003, Illian et al. 2008) have greatest variability for intermediate values of r. The variance of $G(r)$ for random point patterns is approximately proportional to $G(r) \times (1 - G(r))$. The appropriate variance-stabilizing transformation is due to Fisher (1915; see Aitkin and Clayton 1980) and involves replacing $G(r)$ by $G\dagger(r) = \arcsin(\sqrt{G(r)})$. Similar comments apply to F.

Measurements of power

We have compared the power of the DCLF test and the MAD test, based on different summary functions, in a series of simulation experiments. Following is a summary of our findings; more detail will be reported in a forthcoming paper (Baddeley et al., *unpublished manuscript*). A detailed description of experiments to measure the power of a test has been given by Kornak et al. (2006).

In all our experiments, the null hypothesis was CSR, while a range of alternative hypotheses were studied. For each choice of alternative hypothesis, we measured the power of the test by generating 1000 simulated realizations of the alternative hypothesis, performing the test on each simulated pattern, and counting the proportion of correct outcomes of the test. For comparison, both the DCLF and MAD tests were performed, using different choices of summary statistics and different choices of the length of the interval R. The estimated power was plotted against R.

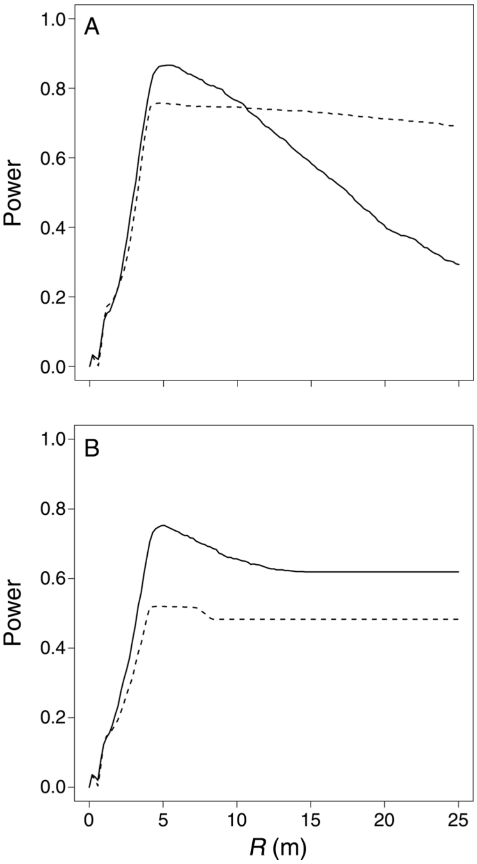

FIG. 6. Performance curve for the DCLF test (solid lines) and the maximum absolute deviation (MAD) test (dashed lines) for detecting spatial inhibition. The power is plotted against the length of the interval of distance values used in the test. The null hypothesis is CSR and the alternative hypothesis is an inhibitory point process with interaction range $s = 4$ m. Panel (A) is a test based on the L function. Panel (B) test is a test based on the variance-stabilized G function.

Two examples are shown in Figs. 6 and 7. Fig. 6 is the power curve for an alternative hypothesis that is an inhibitory (regular) point process, namely the Strauss process (Strauss 1975, Kelly and Ripley 1976), with intensity parameter $\beta = 0.025$, interaction parameter $\gamma = 0.5$ and interaction range $s = 4$ m. (For further information, see Diggle 2003:75, Illian et al. 2008:141, 147). Fig. 7 is the power curve when the alternative is the cluster point process of Matérn (1960:46–47), as described in *Null models with parameters: Tests of goodness of fit of a clustered model.*

We found that the power of each test is maximized when the interval length R is slightly larger than the range of spatial interaction. Additionally, the power of a test based on the L function is greater than the power of the corresponding test based on the variance-stabilized G function $G\dagger(r) = \arcsin\sqrt{G(r)}$.

If the alternative hypothesis is a point process with inhibition at distance s, and we use the L function, then for $R > s$ the power of the DCLF test declines sharply as

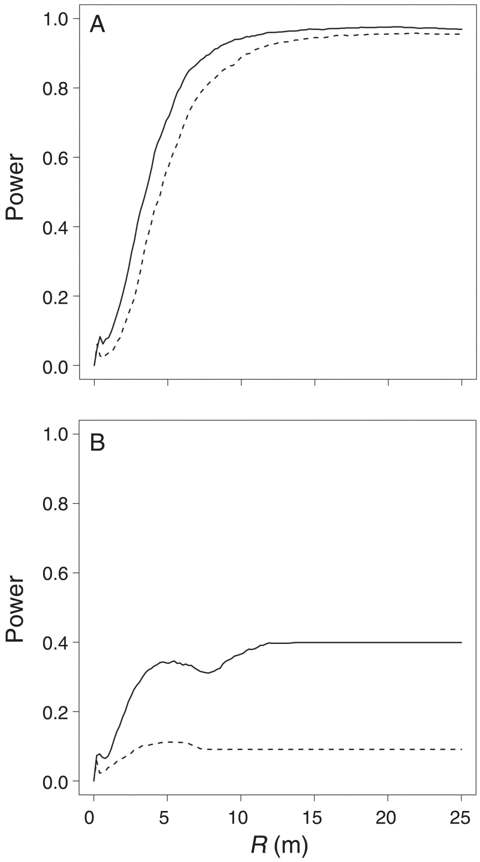

FIG. 7. Performance curve for the DCLF test (solid lines) and the MAD test (dashed lines) for detecting spatial clustering. The null hypothesis is CSR and the alternative hypothesis is the clustered process of Matérn (1960) with cluster radius $s = 14$ m. Panel (A) is a test based on the L function. Panel (B) is a test based on the variance-stabilized G function.

R increases, while the power of the MAD test is (essentially) constant, as shown in the left panel of Fig. 6. Thus, the DCLF test is more powerful than the MAD test when R is close to s, but less powerful when R is large.

If we use instead the variance-stabilized G function with the same alternative hypothesis, then the DCLF is always more powerful than the MAD test, as shown in the right panel of Fig. 6. The power of each test rises sharply from $R = 0$ to $R = s$, then declines slightly before reaching a plateau.

If the alternative hypothesis is a point process with clustering at distance s, and we use either the L function or variance-stabilized G function, then for $R > s$ the power of both tests is (essentially) constant, and the DCLF test is more powerful, as shown in Fig. 7.

Our power analysis showed that the DCLF test is more powerful than the MAD test in most cases; optimal power is achieved when the interval of distance values used for the test is slightly larger than the range of spatial interaction; and power is improved by variance-stabilization. An exception to the assertion that DCLF tests are more powerful than MAD tests occurs when the data are spatially inhibited and the interval of distance values is much greater than the range of spatial interaction, as shown in the left panel of Fig. 6. We expect the findings of this study to hold more widely, because they are consistent with the general pattern found in classical goodness-of-fit testing (Stephens 1986). The DCLF and MAD tests are analogous, respectively, to the Cramér-von Mises and Kolmogorov-Smirnov tests of goodness of fit, and the former is usually more powerful than the latter.

Consequently our recommendation is to use the DCLF test, provided the range of spatial interaction is known approximately (and the interval length is chosen accordingly). If there is no information about the range of spatial interaction, then it may be prudent to use the MAD test, choosing the interval length to be as large as practicable.

SCALES OF INTERACTION

Summary functions like Ripley's K function convey information across a range of spatial scales. This is an important motivation for using empirical functions, rather than simple summary statistics, and for displaying them graphically.

Many researchers use a graph of the K function or L function to infer the scale of spatial interaction in a point pattern. This scale is often estimated by reading off the position where the empirical function lies furthest outside the simulation envelope. For example, in Fig. 2, the L function envelopes are breached at distances between about 3 and 5 m; this would be taken as an estimate of the scale of spatial interaction. This estimate is consistent with the true interaction range, 4 m, in that simple synthetic example.

Loosmore and Ford (2006) criticize such approaches on the grounds that simulation envelopes are invalid and that the summary functions K, L, G, and F are cumulative. "Results at any distance reflect both the instantaneous value at that distance as well as the combined results from all smaller distances. Results for an observed pattern could, therefore, lie outside the envelope at a distance where the instantaneous value was not different than the specified model" (Loosmore and Ford 2006:1927).

In our opinion, Loosmore and Ford (2006) reach the correct conclusion here, but for the wrong reasons. Firstly, as we have shown earlier, there is nothing invalid about envelope tests correctly applied. More subtly, it is not obvious that the cumulative nature of the summary functions K, L, G, and F prevents them from correctly identifying the scale of interaction.

Consider, for example, the Gibbs hard-core point process in which points are forbidden to occur closer than a specified range s whose value unambiguously defines the scale of interaction of the process (Illian et al. 2008). In this case, both the K function and the L function show their greatest deviation from CSR at

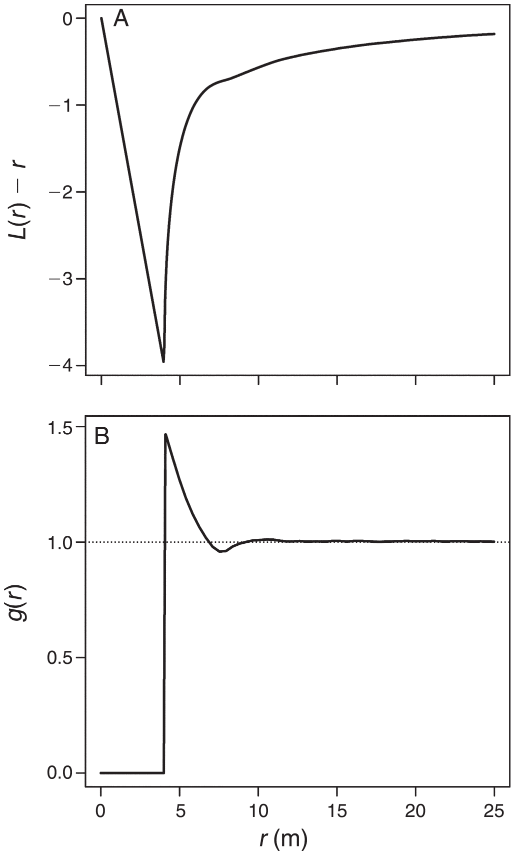

FIG. 8. (A) Centered L function and (B) pair correlation function $g(r)$ of the hard-core process with range $s = 4$ m and base intensity $\beta = 0.06$ in a 100-m square, estimated by simulating 10 000 realizations. The 95% confidence interval in panel (A) is so narrow that it is not visible in this graphic. The dot-dash line is the expected value of the pair correlation for CSR.

distance s. Fig. 8A illustrates this by plotting $L(r) - r$ against r, with $L(r)$ estimated by simulating 10 000 realizations of a hard-core process with range $s = 4$ m.

If we were to accept the objection of Loosmore and Ford (2006), we might be tempted to identify the scale of interaction from the behavior of the non-cumulative counterpart of the K function, namely the pair correlation function $g(r) = (2\pi r)^{-1}K'(r)$. Under CSR, $g(r) = 1$. Fig. 8B plots $g(r)$ against r, again estimated from 10 000 simulated realizations. The shark-fin shape of the pair correlation at distances between 4 and 7 m, the trough at 7.5 m, and the subsequent small peak at 10 m, might naively suggest the existence of multiple scales of interaction in the underlying process; in fact, the peaks and troughs in the plot beyond 4 m are simply echoes of a single scale of interaction. The parameter values in this example ensure that the points are very tightly packed, subject to the hard-core restriction. Each point is very likely to have a neighbor lying just beyond the critical distance $s = 4$ m, and this gives rise to the shark-fin shape. Close neighbors inhibit the presence of other neighbors, giving rise to the trough at 7.5 m. The essential problem here is that correlation is not causation.

A second example is provided by a cluster process, identical to the model of Matérn described in *Null models with parameters: Tests of goodness of fit of a clustered model*, except that each offspring point is displaced relative to its parent according to a bivariate. Normal distribution centered on the parent point, with standard deviation (σ) on each coordinate axis. This is usually called the modified Thomas process (Thomas 1949, Diggle et al. 1976). Its K function and pair correlation function are respectively (Diggle 2003, Illian et al. 2008)

$$K(r) = \pi r^2 + \kappa^{-1}\{1 - \exp(-0.25r^2/\sigma^2)\}$$

$$g(r) = 1 + (4\sigma^2\pi\kappa)^{-1}\exp(-0.25r^2/\sigma^2).$$

In common with the hard-core process, this cluster process embodies a single scale of interaction, as a consequence of the property that only offspring of the same parent interact. Unlike the hard-core process however, it is not obvious how we might assign a numerical value to the scale. One reasonable suggestion would be to use some property of the distribution of the distance between two offspring of the same parent. This depends only on σ. Another might be some property of the distribution of the maximum distance between any two offspring of the same parent. This depends on both σ and μ. We cannot think of any justification for involving the parameter κ, yet this clearly features in the expressions for $K(r)$ and $g(r)$, whereas μ does not. Also, $K(r) - \pi r^2$ and $g(r) - 1$ are, respectively, increasing and decreasing functions of r, so in neither case would it be sensible to identify the scale of the process as the distance at which the function in question deviates maximally from its theoretical value under CSR. The maximum deviation of $L(r)$ from its value under CSR does occur at a finite, nonzero value of r but, as Table 2 illustrates, this value depends on both σ and κ.

In both of our examples, we have considered possible definitions of a scale of interaction only in terms of properties of the underlying stochastic process, not of any observed point pattern. We would argue that this is the correct way to think about the issue, since the objective of any data analysis is to understand what

TABLE 2. The distance, r, at which $L(r)$ deviates maximally from its value under CSR, in a cluster process with parent intensity κ and cluster standard deviation σ.

	σ		
κ	0.01	0.02	0.05
2	0.039	0.070	0.149
5	0.036	0.065	0.137
10	0.034	0.061	0.130

natural processes may or may not have generated the data. In summary, the notion of a scale of interaction is useful for heuristics, but can only be quantified precisely within the confines of a declared family of parametric models. It must be expressible as some function of the model parameters, and can only be estimated by first estimating those parameters.

Conclusions

Statistical tests of spatial pattern based on simulation envelopes, and those based on measures of deviation have the same statistical rationale, namely the Monte Carlo test principle. Both are valid statistical tests under appropriate conditions (including conditions on the way they are applied and interpreted). Simulation envelopes can be of several different kinds, including pointwise envelopes (Fig. 2) and global envelopes (Fig. 3), all of which are statistically valid when used appropriately. Pointwise envelopes carry a higher risk of misinterpretation or misuse, due to the problem of multiple testing.

If model parameters must be estimated from the data, then both envelope tests and deviation tests are strictly invalid, and probably conservative, so that significance will be under-reported. In the examples studied, this effect was quite substantial.

The test advocated by Loosmore and Ford (2006) generally performs better than competing methods, and is recommended, provided there is some prior information about the range of spatial interaction. If no such information is available, the global envelope test (maximum deviation test) should be used.

Envelopes are not confidence intervals. True confidence intervals for the K function of a spatial pattern can be constructed using bootstrap techniques. Confidence intervals may be preferable to significance tests in many investigations. Envelopes and other graphical devices are useful for investigation of data but cannot be used directly to identify scales of spatial interaction without a parametric model.

Acknowledgments

We thank N. Cressie, D. Ford, and A. Nield for their helpful and insightful comments on an earlier version of the paper. A. Baddeley and G. Nair were supported by the Australian Research Council (Discovery Grants DP 130102322, DP 130104470, and an ARC Discovery Outstanding Researcher Award). P. Diggle was supported by a methodology research grant from the UK Medical Research Council (G0902153). The majority of the research was conducted during G. Nair's sabbatical with CSIRO Mathematics Informatics and Statistics. Access to a supercomputing cluster was provided by CSIRO Computational and Simulation Sciences. All graphs and tables in this paper were generated using the spatstat package (Baddeley and Turner 2005) in the R language. The code is provided in the Supplement.

Literature Cited

Aitkin, M., and D. Clayton. 1980. The fitting of exponential, Weibull, and extreme value distributions to complex censored survival data using GLIM. Applied Statistics 29:156–163.

Baddeley, A. 2010. Modelling strategies. Pages 339–369 *in* A. E. Gelfand, P. J. Diggle, M. Fuentes, and P. Guttorp, editors. Handbook of spatial statistics. CRC Press, Boca Raton, Florida, USA.

Baddeley, A. J., and B. W. Silverman. 1984. A cautionary example on the use of second-order methods for analyzing point patterns. Biometrics 40:1089–1094.

Baddeley, A., and R. Turner. 2005. Spatstat: an R package for analyzing spatial point patterns. Journal of Statistical Software 12:1–42.

Barnard, G. 1963. Contribution to the discussion of M.S. Bartlett: The spectral analysis of point processes. Journal of the Royal Statistical Society B 25:294.

Bayarri, M. J., and J. O. Berger. 2000. *P* values for composite null models. Journal of the American Statistical Association 95:1127–1142.

Besag, J. E. 1977. Contribution to the discussion of the paper by Ripley (1977). Journal of the Royal Statistical Society B 39:193–195.

Besag, J., and P. J. Diggle. 1977. Simple Monte Carlo tests for spatial pattern. Applied Statistics 26:327–333.

Bowman, A. W., and A. Azzalini. 1997. Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations. Oxford University Press, Oxford, UK.

Brooks, S. P., B. J. T. Morgan, M. S. Ridout, and S. E. Pack. 1997. Finite mixture models for proportions. Biometrics 53:1097–1115.

Chandler, R. E., and E. M. Scott. 2011. Statistical methods for trend detection and analysis in the environmental sciences. John Wiley and Sons, Chichester, UK.

Clark, P. J., and F. C. Evans. 1954. Distance to nearest neighbor as a measure of spatial relationships in populations. Ecology 35:445–453.

Cox, D. R. 1977. The role of significance tests. Scandinavian Journal of Statistics 4:49–70.

Cressie, N. A. C. 1991. Statistics for spatial data. John Wiley and Sons, New York, New York, USA.

Dao, N. A., and M. Genton. 2014. A Monte Carlo adjusted goodness-of-fit test for parametric models describing spatial point patterns. Journal of Computational and Graphical Statistics 23:497–517.

Diggle, P. J. 1983. Statistical analysis of spatial point patterns. Academic Press, London, UK.

Diggle, P. J. 1986. Displaced amacrine cells in the retina of a rabbit: analysis of a bivariate spatial point pattern. Journal of Neuroscience Methods 18:115–125.

Diggle, P. J. 2003. Statistical analysis of spatial point patterns. 2nd Edition. Hodder Arnold, London, UK.

Diggle, P. J., J. Besag, and J. T. Gleaves. 1976. Statistical analysis of spatial point patterns by means of distance methods. Biometrics 32:659–667.

Diggle, P. J., and R. J. Gratton. 1984. Monte Carlo methods of inference for implicit statistical models (with discussion). Journal of the Royal Statistical Society B 46:193–227.

Fisher, R. A. 1915. Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. Biometrika 10:507–521.

Grabarnik, P., M. Myllymäki, and D. Stoyan. 2011. Correct testing of mark independence for marked point patterns. Ecological Modelling 222:3888–3894.

Hochberg, Y., and A. Tamhane. 1987. Multiple comparison procedures. John Wiley and Sons, New York, New York, USA.

Hope, A. C. A. 1968. A simplified Monte Carlo significance test procedure. Journal of the Royal Statistical Society B 30:582–598.

Hsu, J. C. 1996. Multiple comparisons: theory and methods. Chapman and Hall, London, UK.

Illian, J., A. Penttinen, H. Stoyan, and D. Stoyan. 2008. Statistical analysis and modelling of spatial point patterns. John Wiley and Sons, Chichester, UK.

Kelly, F. P., and B. D. Ripley. 1976. A note on Strauss' model for clustering. Biometrika 63:357–360.

Kenkel, N. C. 1988. Pattern of self-thinning in jack pine: testing the random mortality hypothesis. Ecology 69:1017–1024.

Kornak, J., M. E. Irwin, and N. Cressie. 2006. Spatial point process models of defensive strategies: detecting changes. Statistical Inference for Stochastic Processes 9:31–46.

Law, R., J. Illian, D. F. R. P. Burslem, G. Gratzer, C. V. S. Gunatilleke, and I. A. U. N. Gunatilleke. 2009. Ecological information from spatial patterns of plants: insights from point process theory. Journal of Ecology 97:616–628.

Loh, J. M. 2008. A fast and valid spatial bootstrap for correlation functions. Astrophysical Journal 681:726–734.

Loosmore, N. B., and E. D. Ford. 2006. Statistical inference using the *G* or *K* point pattern spatial statistics. Ecology 87:1925–1931.

Marriott, F. H. C. 1979. Barnard's Monte Carlo tests: how many simulations? Applied Statistics 28:75–77.

Matérn, B. 1960. Spatial variation: stochastic models and their application to some problems in forest surveys and other sampling investigations. Meddelanden från Statens Skogsforskningsinstitut 49(5):1–144.

Pfanzagl, J. 1969. On the measurability and consistency of minimum contrast estimates. Metrika 14:249–276.

Ripley, B. D. 1977. Modelling spatial patterns (with discussion). Journal of the Royal Statistical Society B 39:172–212.

Ripley, B. D. 1979. Tests of randomness for spatial point patterns. Journal of the Royal Statistical Society B 41:368–374.

Ripley, B. D. 1981. Spatial statistics. John Wiley and Sons, New York, New York, USA.

Robins, J. M., A. van der Vaart, and V. Ventura. 2000. Asymptotic distribution of *P* values in composite null models. Journal of the American Statistical Association 95(452):1143–1156.

Shaffer, J. P. 1995. Multiple hypothesis testing. Annual Review of Psychology 46:561–584.

Stephens, M. A. 1986. Tests based on EDF statistics. Pages 97–193 *in* R. B. D'Agostino and M. A. Stephens, editors. Goodness-of-fit techniques. Volume 68. Statistics: textbooks and monographs. Marcel Dekker, New York, New York, USA.

Strauss, D. J. 1975. A model for clustering. Biometrika 63:467–475.

Strong, D. R., Jr. 1980. Null hypotheses in ecology. Synthese 43:271–285.

Thomas, M. 1949. A generalization of Poisson's binomial limit for use in ecology. Biometrika 36:18–25.

Yates, F. 1951. The influence of statistical methods for research workers on the development of the science of statistics. Journal of the American Statistical Association 46:19–34.

Supplemental Material

Appendix

Notes on Loosmore and Ford's test statistic (*Ecological Archives* M084-017-A1).

Supplement

R code for tests of spatial pattern (*Ecological Archives* M084-017-S1).